SiLVR: Scalable Lidar-Visual Reconstruction with Neural Radiance Fields for Robotic Inspection

Yifu Tao[1], Yash Bhalgat[2], Lanke Frank Tarimo Fu[1], Matias Mattamala[1], Nived Chebrolu[1], and Maurice Fallon[1]

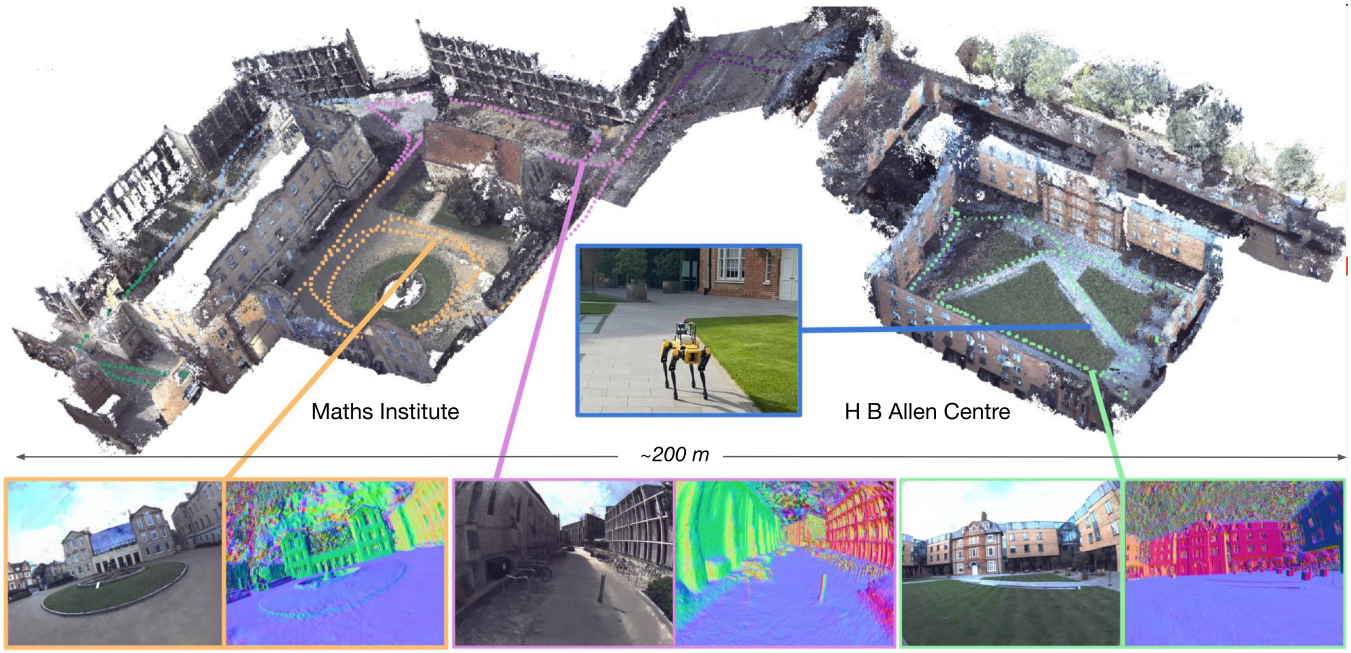

Fig. 1: Large-scale reconstruction consisting of 8 submaps of Maths Institute and H B Allen Centre in Oxford. The bottom row shows the novel views synthesised from the model and surface normals at three different locations. The trajectory of each submap is visualised in a different colour.

***Abstract*— We present a neural-field-based large-scale reconstruction system that fuses lidar and vision data to generate high-quality reconstructions that are geometrically accurate and capture photo-realistic textures. This system adapts the state-of-the-art neural radiance field (NeRF) representation to also incorporate lidar data which adds strong geometric constraints on the depth and surface normals. We exploit the trajectory from a real-time lidar SLAM system to bootstrap a Structure-from-Motion (SfM) procedure to both significantly reduce the computation time and to provide metric scale which is crucial for lidar depth loss. We use submapping to scale the system to large-scale environments captured over long trajectories. We demonstrate the reconstruction system with data from a multi-camera, lidar sensor suite onboard a legged robot, hand-held while scanning building scenes for 600 metres, and onboard an aerial robot surveying a multi-storey mock disaster site-building. Website: `https://ori.ox.ac.uk/labs/drs/nerf-mapping/`**

[1]Oxford Robotics Inst., Dept. of Eng. Science, Univ. of Oxford, UK. {yifu, fu, matias, nived, mfallon}@robots.ox.ac.uk
[2]Visual Geometry Group, Dept. of Eng. Science, Univ. of Oxford, UK. yashsb }@robots.ox.ac.uk.
This project has been partly funded by the Horizon Europe project Digiforest (101070405). Maurice Fallon is supported by a Royal Society University Research Fellowship. For the purpose of open access, the authors have applied a Creative Commons Attribution (CC BY) license to any Accepted Manuscript version arising.

I. INTRODUCTION

Dense 3D reconstruction is a task that underpins a range of robotic applications such as industrial inspection and autonomous navigation. Common sensors used for reconstruction include cameras and lidar. Camera-based reconstruction systems use techniques including Structure-from-Motion (SfM) and Multi-View Stereo (MVS) to produce dense textured reconstructions [1]. However, these systems rely on good lighting conditions as well as having observed the many view constraints. They also struggle with textureless areas. Lidar provides accurate geometric information at long range — as it directly measures distances to surfaces. This makes it desirable for large-scale outdoor environments, but the sensor measurement is usually much sparser than a camera image. It also does not capture colour which is important for applications such as virtual reality and 3D asset generation.

Classical reconstruction systems have used point clouds, occupancy maps, and sign-distance fields (SDF) as their internal 3D representation. Recently, neural radiance field (NeRF) [2] has gained popularity for visual reconstruction. With differentiable rendering, this approach optimises a continuous 3D representation by minimising the difference between the rendered image and reference camera images. It

can achieve state-of-the-art novel view synthesis quality.

As with traditional vision-based reconstruction methods, NeRF struggles to estimate accurate geometry in locations where there is limited multi-view input and little texture. Autonomous systems commonly encounter these situations - for example, reconstructing a wall with uniform colour when traveling directly towards it. This problem could be addressed by using lidar sensing to give accurate geometric measurements in these textureless objects. In addition, the use of lidar data can mitigate the need to capture multiple camera views. It is impractical for an inspection robot (e.g. The Spot quadruped on an industrial facility) to execute object-centric trajectories just to improve the visual reconstruction. This motivates the development of a reconstruction system that fuses these sensors.

In this work, we present a NeRF-based reconstruction system that integrates both lidar and visual information to generate accurate, textured 3D reconstruction which also provides photo-realistic novel view synthesis. Our method builds upon NeRF implementation [3] utilising hash encoding [4] that takes minutes to achieve photo-realistic rendering. This is extended with geometric constraints from lidar to improve reconstruction quality. The use of lidar enables depth measurements even from featureless areas, which cannot be obtained from SfM [5]. Surface normal can also be computed from a lidar scan, which is more robust than learning-based priors [6] which can suffer from input data distribution shift in real-world deployment.

This system is demonstrated using data from a perception sensor suite which contains three wide field-of-view cameras and a 3D lidar to enable robots to reconstruct in all directions. We take advantage of a lidar-inertial odometry and SLAM mapping system [7] as part of the pipeline. Experiments are presented using a drone, a legged robot, and a handheld device in industrial and urban environments.

In summary, our main contributions are:

- A dense textured 3D reconstruction system that achieves accurate geometry that's on-par with lidar, and photorealistic novel view synthesis
- Integration with a lidar SLAM system, so that the NeRF is trained with both depth and surface normal obtained from lidar data, and metric-scale trajectory with a reduction in computation time by 50% compared with up-to-scale offline Structure-from-Motion method commonly used in literature.
- A sub-mapping system that scales to large outdoor environments — with trajectories over 600 metres
- Evaluation of the system on real-world large-scale outdoor datasets captured from multiple robotic platforms.

II. RELATED WORKS

A. Large-scale 3D Reconstruction

Lidar is the dominant sensor for accurate 3D reconstruction of large-scale environments [8], [9], thanks to its accurate, long-range measurements. Volumetric lidar mapping relies on high-frequency odometry estimates from scan-matching and IMU measurements [7], [10]–[12]. Yet, lidar-based systems may yield partial reconstructions, especially when robots explore scenes and have limited field-of-view sensors. Visual cues can be used to densify lidar mapping [13].

Alternatively, SfM systems such as COLMAP [1] can generate large-scale textured reconstructions from images by first estimating camera poses using sparse feature points in a joint bundle adjustment process, followed by refinement with multi-view stereo algorithms [14]. However, they face challenges in low-texture areas, repetitive patterns, and feature matching across views can be problematic with changing lighting or non-Lambertian materials. Urban Radiance Fields [15] proposes using lidar sweeps along with RGB images to optimise a neural radiance field model that can be used for 3D surface extraction. Our work shares this approach, fusing both sensor modalities to generate precise geometry, overcoming limitations of vision-only approaches in low-texture areas, and being much denser than lidar-only reconstructions.

A common strategy to extend dense reconstructions to large-scale areas is through submapping [16]–[20]. These approaches partition the scene into individual local submaps which can incorporate the effect of loop closure corrections while still producing a consistent global map while achieving significantly lower runtime. Our work also adopts the submapping approach and partitions large-scale scenes into local maps (approximately 50x50m) using a globally-consistent lidar SLAM trajectory. This increases the representation capability and improves reconstruction, especially for thin objects.

B. Neural Field Representation

Neural Radiance Fields (NeRF) [2] proposed using a multilayer perceptron (MLP) to represent a continuous radiance field with differentiable volume rendering to reconstruct novel views. It implicitly induces multi-view consistency with geometric priors in the learned encodings. NeRF and its many variants used frequency encoding [21] to encode spatial coordinates, but these suffer from long training times, typically a few hours per scene. Alternative explicit representations of radiance fields, including dense voxel-grids with trainable per-vertex features [4], [22] and more recently 3D Gaussians [23] are shown to accelerate the training, at the cost of being more memory intensive. Octree or sparse-grid structures [4], [24] can reduce memory usage by pruning grid-features in empty space. Our work is built upon Nerfacto [3] which incorporates the main features from other NeRF works [4], [25], [26] that have been found to work well with real data.

While NeRFs excel at high-quality view synthesis, obtaining a 3D surface from these representations remains challenging, mainly due to the flexible volumetric representation being under-constrained by the limited multi-view inputs. One approach to improve the reconstruction is to impose depth regularisation [5], [15] or surface normal regularisation [6]. Another approach is to impose surface priors on the volumetric field and use representations such as Signed Distance Field (SDF) [27], [28] to enforce a surface reconstruction output, although the novel view synthesis

quality might be compromised [29] with this approach. Our method is built upon a volume density representation which is extended with depth [5] and surface normal [6] regularisations from lidar measurements instead of using SfM [5] or learnt priors [6]. In particular, it can significantly improve the reconstruction quality in texture-less areas with smooth surface.

Neural field representations have been used for lidar-base mapping [30], [31], showing promise in generating more complete and compact reconstructions than traditional methods. While these works also build upon implicit map representation, they do not use visual data for building the map. Our system uses visual information and multi-view geometry constraints, therefore can reconstruct regions outside lidar field-of view.

III. METHOD

We present a system for large-scale 3D reconstruction based on a NeRF representation tailored for robotic inspection tasks. We use a custom-designed sensor payload with a 3D lidar sensor, three fisheye cameras, and an IMU, which is suitable for use on various robot platforms. In Sec. III-B, we present our approach to fusing both lidar and vision information during the optimization phase to ensure high-quality reconstruction. Furthermore, we employ a submapping strategy to scale the approach to large areas described in Sec. III-D. An overview of the system is presented in Fig. 2.

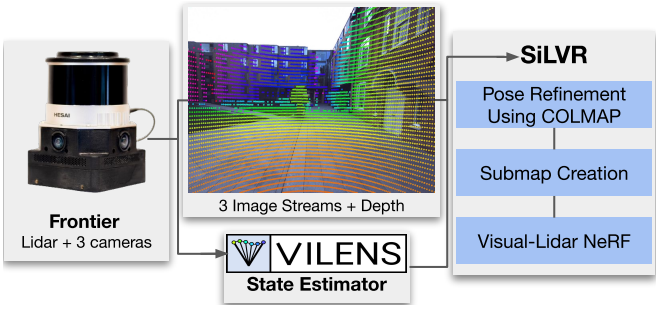

Fig. 2: System Overview: Frontier, our custom perception unit, has three fisheye colour cameras with an IMU and a 3D lidar. Our online state estimator's trajectory is refined with COLMAP and partitioned into submaps. The camera image, lidar depth, and normal image are used to train a NeRF to get the final 3D reconstruction.

A. NeRF-based Scene Representation

Our work builds on the differentiable volume rendering framework used for novel view synthesis [2], [32]–[34]. To render a novel view from a NeRF given a viewpoint, we cast rays from the camera origin along the viewing direction for each pixel u in the image plane, and render the pixel-colour by integrating over points sampled along the ray. This volume rendering integral is approximated using quadrature rule [35], [36] as $\hat{c}_u = \sum_{i=0}^{N} w_i c_i$, where

$$w_i = \exp\left(-\sum_{j=1}^{i-1} \delta_j \sigma_j\right)\left(1 - \exp\left(-\delta_i \sigma_i\right)\right). \tag{1}$$

Here, σ_i and c_i are the predicted density and color for the sampled 3D points and $\hat{c}_u$ is the rendered pixel color.

Our implementation is built on top of the Nerfacto method from Nerfstudio [3]. Nerfacto's rendering quality is comparable to state-of-the-art methods such as MipNeRF-360 [25] while achieving a substantial acceleration in reconstruction speed since it also incorporates efficient hash encoding from Instant-NGP [4]. We also use scene contraction proposed in [25] to improve memory efficiency and represent scenes with high-resolution content near the input camera locations. The contraction function non-linearly maps any point in space into a cube of side length 2, and represents the scene within this contracted space. Since there is large variation in exposure and lighting conditions, we use a per-frame appearance encoding for each image, similar to [15], [26].

B. Geometric Constraints from Lidar Measurements

3D reconstruction with NeRF becomes challenging when the surface has uniform texture and limited multi-view constraints. Lidar measurements are complementary as it can provide accurate measurements in such scenarios. In our work, we incorporate the lidar measurements in the NeRF optimization. Specifically, we impose a lidar-based depth regularisation by adding a depth loss [5] defined as the KL-Divergence between a normal distribution around the lidar depth-measurement $\mathbf{D}$ and the rendered ray distribution $h(t)$ from the NeRF model:

$$\mathcal{L}_{\text{Depth}} = \sum_{\mathbf{r}\in\mathcal{R}} \text{KL}[\mathcal{N}(\mathbf{D}, \hat{\sigma}) \| h(t)] \tag{2}$$

We also run a semantic segmentation network [37] to obtain a sky mask, and minimise the weights of these rays similar to [15].

While the depth loss improves 3D reconstruction, we found that the surface contains wavy artifacts in regions where it is expected to be smooth (see Fig. 5). To mitigate this, we compute the surface normal as the negative gradient of the NeRF's density field, and impose a further surface normal regularisation loss, inspired by [6]:

$$\mathcal{L}_{\text{normal}} = \sum_{\mathbf{r}\in\mathcal{R}} \|\hat{N}(\mathbf{r}) - \bar{N}(\mathbf{r})\|_1 + \left\|1 - \hat{N}(\mathbf{r})^\top \bar{N}(\mathbf{r})\right\|_1 \tag{3}$$

C. Bootstrapping Camera Poses from SLAM with scale

Obtaining accurate camera poses is crucial as the pose accuracy directly impacts the fidelity of the reconstructed model. A popular approach used in most NeRF works is to obtain camera poses using offline Structure-from-Motion methods such as COLMAP [1]. However, COLMAP has the following limitations: (1) long computation time, especially for large image collections collected over a long trajectory, and (2) inability to register all frames into one global map when there is limited visual overlap between the images. These issues limit its application in building a large-scale globally consistent map for robotic applications.

In our work, we use our lidar-inertial odometry and SLAM system VILENS [7]. While VILENS achieves state-of-the-art

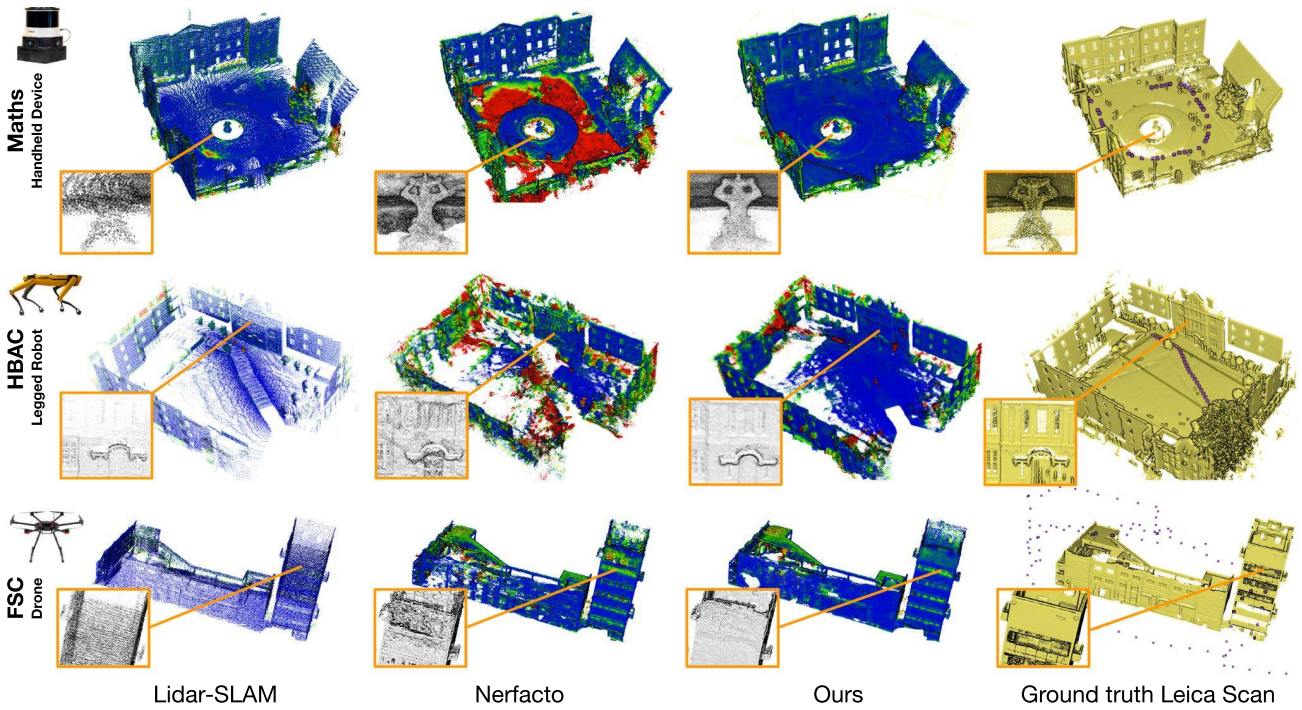

Fig. 3: Comparison of reconstruction quality of Lidar-SLAM, Nerfacto (vision-only) and our approach. Reconstructions are coloured with point-to-point distance to the ground truth with increasing error from blue (0m) to red (1m). The trajectory is shown in purple and overlaid on the ground truth scan captured using a Leica BLK360. The zoomed-in views show the difference in reconstruction quality. Overall, our approach is more complete w.r.t lidar-only reconstruction, and geometrically more consistent w.r.t vision-only reconstruction.

results for online motion tracking, we found that the camera poses obtained are less accurate than those of COLMAP. This results in *blurring* artifacts in the renderings of the NeRF model. Several works [20], [38] use noisy pose inputs and then jointly refine the poses in parallel with the NeRF optimization to generate better results. In our experiments with a collection of large number of images, we found that while this pose-refinement approach sometimes leads to slightly better PSNR (peak signal-to-noise ratio), the resulting rendering is still less sharp compared to using the COLMAP estimated poses, and the training process is usually unstable.

To overcome the above limitations, we use the SLAM poses as a pose prior and refine the trajectory using COLMAP. Specifically, we replace the COLMAP *mapper* with *point triangulator* which reads prior poses. This method has the advantage of being faster, as it converts the incremental Structure-from-Motion into a global bundle adjustment problem, and more importantly, results in COLMAP being able to merge all available images in a single map. For a mission spanning over 20 minutes, our COLMAP-with-prior pipeline achieves similar rendering quality, while only taking half the computation time compared to a fully offline COLMAP run. The computation time is similar to the robot's mission duration, making it more applicable for robotic applications.

After COLMAP computation, we rescale the trajectory using Sim(3) Umeyama alignment with lidar-slam trajectory, so that the final trajectory is metric. This is essential to use lidar measurements in III-B, since lidar depth is also metric.

D. Scaling NeRF with Submapping

Training a NeRF for large-scale scenes is challenging as a single NeRF model has limited representation power and hardware constraints such as RAM usage when loading thousands of images. We adopt a sub-mapping approach, and partition the COLMAP-refined SLAM trajectory into clusters using Spectral Clustering [39]. Different from [20], our representation is based on Instant-NGP [4] which is orders of magnitude faster than a Multilayer Perception. The submaps are trained in their local coordinate frames, and the final reconstruction are *Sim(3)* transformed to the world coordinate frame using their metric pose from III-C.

To generate 3D reconstructions from NeRF, we sample the rays used to train each model and render the colour and depth to generate a 3D point. When evaluating the 3D reconstruction, we found that submaps contain artifacts, especially around their boundary. One cause of the artifacts is that the boundary regions are observed sparsely with limited view constraints. To tackle this, we identify regions with low surface density and remove them when merging submap clouds to get the final reconstruction.

IV. EXPERIMENTAL RESULTS

A. Hardware and Datasets

We evaluate our methods on a custom perception unit called Frontier, a multi-camera lidar inertial device. It includes three 1.6 MP colour[1] fisheye Alphasense cameras on 3 sides of

[1]Raw Bayered images are processed using `https://github.com/leggedrobotics/raw_image_pipeline`

the device, with an IMU from Sevensense Robotics AG and a Hesai Pandar QT64 lidar. We used the Frontier to collect datasets on multiple platforms: a legged robot (Boston Dynamics Spot), a drone (DJI M600), and a hand-held device. Spot and handheld Frontier datasets were taken at the H B Allen Centre (HBAC) and Mathematical Institute (Maths Inst.) in Oxford. The DJI M600 drone was operated at the Fire Service College (FSC). In the FSC dataset, we only use the rectified front camera image as the drone propellers were visible in the left and right camera images.

We use VILENS [7], a lidar-inertial SLAM system online to provide a globally consistent trajectory and motion-corrected lidar measurements. The online SLAM trajectory is further optimised by COLMAP offline, which improves the visual reconstruction quality, as shown later in Tab. II. The lidar point clouds are projected as a depth image. The surface normal is computed from the lidar range image, and also projected as a normal image. We calibrate our cameras-to-IMU extrinsics with Kalibr [40] and cameras-to-lidar with [41]. When running COLMAP, we further optimise the intrinsics from Kalibr. For training the NeRF model, we used an Nvidia RTX 3080 Ti and one iterations takes 4096 rays.

B. Evaluation Metrics

To evaluate the geometry of the reconstruction, we report *Accuracy* and *Completeness* following the conventions of the DTU dataset [42]. Accuracy is measured as the distance from the reconstruction to the reference 3D model (ground truth) and encapsulates the reconstruction quality. Completeness is the distance from the point-wise reference to the reconstruction and shows how much of the surface is captured by the reconstruction. For the ground truth, we use the centimetre-accurate point cloud captured by a survey-grade Leica BLK360 laser scanner and register the lidar point clouds using transformation obtained from ICP, which is also used to register NeRF reconstructions.

We also evaluate the visual quality by reporting the Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) [43]. Note that our images have various exposure times, which lower the test PSNR even if the reconstructed image is photorealistic.

C. Evaluation of the 3D Reconstruction

We perform a quantitative evaluation of our method on real-world datasets captured by different robotic platforms. We compare the point cloud reconstructions generated with the following configurations:

1) Lidar-SLAM: lidar point cloud registered with SLAM poses refined by COLMAP
2) Nerfacto [3]: baseline method using only the images
3) SiLVR: Our method using photometric loss, depth loss, and surface normal loss

We sample the training rays and generate 10 million points from each submap for Nerfacto and SiLVR. For Nerfacto, we excluded the points belonging to the sky by computing a sky segmentation mask with further manual cropping. For lidar-slam, we only include points within the camera field-of-view. This omits points pointing backwards that may be occluded by the operator.

TABLE I: Evaluation of 3D Reconstruction Quality

Method	Accuracy↓ (m)	Completeness↓ (m)	PSNR↑ train	PSNR↑ test	SSIM↑ test
Maths Quad					
lidar-SLAM	**0.06**	0.15	/	/	/
Nerfacto mono	1.38	0.33	**31.3**	17.7	0.65
Nerfacto 3-cam	0.76	0.31	28.0	**21.1**	**0.72**
Ours mono	0.08	0.12	30.1	17.5	0.61
Ours 3-cam	0.08	**0.11**	27.3	20.5	0.71
Oxford HBAC					
lidar-SLAM	**0.05**	0.25	/	/	/
Nerfacto mono	0.49	5.40	**32.6**	19.5	0.65
Nerfacto 3-cam	0.28	0.40	29.8	20.6	**0.74**
Ours mono	0.30	4.60	31.0	**21.2**	**0.74**
Ours 3-cam	0.09	**0.18**	28.8	19.7	**0.74**
FSC					
lidar-SLAM	**0.08**	**0.08**	/	/	/
Nerfacto mono	0.14	0.11	**28.8**	**19.1**	**0.76**
Ours mono	0.11	0.09	27.7	**19.1**	0.75

NeRF models are trained for 10000 iterations for 5 minutes; When computing accuracy and completeness, we crop regions where there is a change or no overlap between the ground truth and lidar scans.

We summarise the quantitative results in Tab. I, and show 3D reconstructions in Fig. 3. Compared to Nerfacto, our method incorporates lidar measurements and has significantly better geometry which is shown in terms of accuracy and completeness. Nerfacto struggles when reconstructing the uniformly-coloured ground in Maths Inst., and the quad area in HBAC where the robot only walked forwards. Compared to lidar-SLAM, our method generally achieves more complete reconstruction since it uses dense visual information, while the accuracy (8-11cm) is nearly on-par with lidar-slam (6-8cm) and much better than Nerfacto (14-76cm).

Fig. 4: Comparison of reconstruction of HBAC building using the front camera only vs. using all the three cameras. The three-camera setup generates more complete and accurate reconstructions compared to using only a single front-facing camera. The multi-camera setting is important in robotic applications where it would be infeasible to actively scan the entire scene to obtain strong viewpoint constraints.

The advantage of our multi-camera sensor setup is demonstrated qualitatively in Fig. 4. Compared to the three-camera

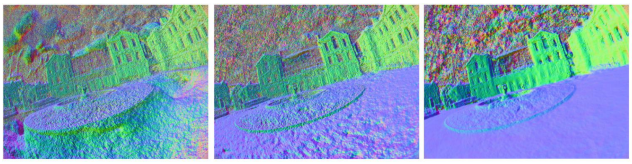

Fig. 5: Comparison of surface normal renderings of the Maths Institute. Incorporating normal constraints in addition to depth from lidar improves the smoothness of the reconstruction. Right: The smooth reconstruction of the ground portion highlights this improvement.

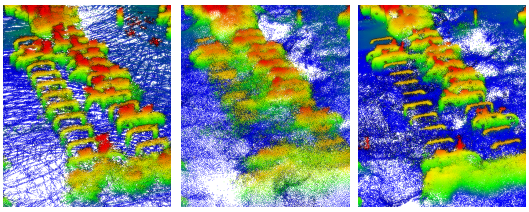

Fig. 6: Effect of submap size on reconstruction quality. A larger number of submaps for a given area results in better reconstruction on thin objects such as the bike racks on the right.

setup, using only the front-facing camera leads not only to an incomplete reconstruction but also worse geometry. Visual reconstruction with photometric loss is biased towards generating a good rendering only at the input viewing angle. The reconstruction using the front-only camera in Fig. 4 is trained with images viewing a shallow angle of the scene, and results in a poor geometric reconstruction when rendered from an unseen angle.

D. Effect of Lidar Surface Normal Loss

While the depth loss Eq. (2) provides geometric constraints, we observe that the resulting 3D reconstruction is not necessarily smooth for flat surfaces. The results in Fig. 5 demonstrate how the surface normal loss Eq. (3) further constrains the reconstruction geometry and improves the reconstruction quality. Nerfacto fails to estimate the depth of the ground due to its uniform texture. Using the lidar depth loss ensures ground reconstruction is at the right height, however, the surface is still not smooth. The surface normal loss furthers smoothens the surface and results in a higher-quality reconstruction.

TABLE II: Ablation: Effect of Bootstrapping SLAM Poses

Method	Features	Prior	Traj. Registered (%)	PSNR↑ Train	PSNR↑ Test	SSIM↑ Test	Time (s)
VILENS	/	/	100.0	23.0	17.4	0.64	Online
NeRF refined	/	/	100.0	23.2	17.9	0.65	Online
COLMAP Sequential	1024		57.6	25.9	19.1	0.71	3299.2
	1024	✓	100.0	26.2	**20.6**	**0.74**	1729.9
	8192		94.0	26.1	19.8	0.72	7850.0
	8192	✓	100.0	26.2	20.4	0.73	4448.4
COLMAP VocabTree	1024		54.7	26.2	19.0	0.71	4444.8
	1024	✓	100.0	26.3	20.4	0.73	1052.5
	8192		94.8	**26.6**	19.9	0.72	37067.5
	8192	✓	100.0	26.3	20.4	**0.74**	11015.3

Results evaluated on HBAC-Maths dataset with 3254 images and duration of 1270s. Models trained for 4000 iterations. PSNR and SSIM are evaluated after masking out the sky.

E. Effect of Bootstrapping SLAM Poses

We compare the performance of different strategies for computing poses: online SLAM poses, SLAM poses with NeRF pose refinement [3], SLAM poses with COLMAP [1] in different configurations, and COLMAP without any prior poses. For COLMAP, we tested different numbers of features extracted per image, and two different COLMAP feature matching algorithms: sequential matching with loop closures and Vocabulary Tree Matcher.

The results are summarised in Tab. II. For all COLMAP configurations, providing the SLAM prior poses not only accelerates pose computation, but also leads to better test rendering, compared to the offline COLMAP. Our SLAM prior poses also register all images in the trajectory, while when not provided, COLMAP only registers 55%-95% images. Extracting more visual features per image (from 1024 to 8192) leads to higher percentage of image registration and better visual reconstruction (PSNR and SSIM). This comes at the expense of a higher computation time, especially with the VocabTree matcher. Using the COLMAP Sequential Matcher is generally faster than Vocabulary Tree Matcher.

F. Submapping for Large-Scale Environments

We show a large-scale reconstruction of HBAC-Maths with the handheld Frontier using submapping in Fig. 1, as well as the trajectory for each submap. To demonstrate the advantage of submapping, we compare the 3D reconstruction and rendering quality using one NeRF model for the entire sequence versus NeRF models built with multiple submaps. We present the qualitative results in Fig. 6. The reconstruction of the bike racks in Maths Inst. is blurred when only using a single NeRF map due to its limited representation capability for storing all objects over a large area. While using only a dedicated submap for that local area, the reconstruction quality improves significantly as seen in Fig. 6 (right).

V. CONCLUSIONS

In summary, we proposed a large-scale 3D reconstruction system fusing both lidar and vision in a neural field via differentiable rendering. The proposed sensor fusion overcomes the limitations of individual sensors, namely the sparsity of the lidar and the fragility of vision in the presence of texture-less surface, and limited multi-view constraints. We demonstrate large-scale reconstruction results from real-world datasets collected from multiple robot platforms in conditions suited for inspection tasks.

ACKNOWLEDGMENT

The authors would like to thank Ren Komatsu for his help on software development, Tobit Flatscher for deploying spot robot, Rowan Border for drone data collection, and Sundara Tejaswi Digumarti for his helpful discussions.

REFERENCES

[1] J. L. Schönberger and J.-M. Frahm, "Structure-from-motion revisited," in *IEEE Int. Conf. Computer Vision and Pattern Recognition*, 2016.

[2] 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," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.

[3] M. Tancik, E. Weber, E. Ng, R. Li, B. Yi, J. Kerr, T. Wang, A. Kristoffersen, J. Austin, K. Salahi, A. Ahuja, D. McAllister, and A. Kanazawa, "Nerfstudio: A modular framework for neural radiance field development," in *SIGGRAPH*, 2023.

[4] T. Müller, A. Evans, C. Schied, and A. Keller, "Instant neural graphics primitives with a multiresolution hash encoding," *ACM Transactions on Graphics (TOG)*, vol. 41, no. 4, pp. 102:1–102:15, Jul. 2022.

[5] K. Deng, A. Liu, J.-Y. Zhu, and D. Ramanan, "Depth-supervised nerf: Fewer views and faster training for free," in *IEEE Int. Conf. Computer Vision and Pattern Recognition*, 2022, pp. 12 882–12 891.

[6] Z. Yu, S. Peng, M. Niemeyer, T. Sattler, and A. Geiger, "MonoSDF: Exploring monocular geometric cues for neural implicit surface reconstruction," *Conf. on Neural Information Processing Systems (NeurIPS)*, 2022.

[7] D. Wisth, M. Camurri, and M. Fallon, "VILENS: Visual, inertial, lidar, and leg odometry for all-terrain legged robots," *IEEE Trans. Robotics*, vol. 39, no. 1, pp. 309–326, 2023.

[8] J. Behley and C. Stachniss, "Efficient surfel-based slam using 3d laser range data in urban environments," in *Robotics: Science and Systems (RSS)*, 2018.

[9] J. Lin and F. Zhang, "R3LIVE: A robust, real-time, rgb-colored, lidar-inertial-visual tightly-coupled state estimation and mapping package," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2022, pp. 10 672–10 678.

[10] J. Zhang and S. Singh, "LOAM: Lidar odometry and mapping in real-time." in *Robotics: Science and Systems (RSS)*, vol. 2, no. 9. Berkeley, CA, 2014, pp. 1–9.

[11] S. Zhao, H. Zhang, P. Wang, L. Nogueira, and S. Scherer, "Super odometry: Imu-centric lidar-visual-inertial estimator for challenging environments," in *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 8729–8736.

[12] W. Xu, Y. Cai, D. He, J. Lin, and F. Zhang, "Fast-lio2: Fast direct lidar-inertial odometry," *IEEE Trans. Robotics*, vol. 38, no. 4, pp. 2053–2073, 2022.

[13] Y. Tao, M. Popović, Y. Wang, S. T. Digumarti, N. Chebrolu, and M. Fallon, "3d lidar reconstruction with probabilistic depth completion for robotic navigation," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 5339–5346.

[14] Y. Furukawa and J. Ponce, "Accurate, dense, and robust multiview stereopsis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 8, pp. 1362–1376, 2010.

[15] K. Rematas, A. Liu, P. P. Srinivasan, J. T. Barron, A. Tagliasacchi, T. Funkhouser, and V. Ferrari, "Urban radiance fields," in *IEEE Int. Conf. Computer Vision and Pattern Recognition*, 2022, pp. 12 932–12 942.

[16] M. Bosse, P. Newman, J. Leonard, M. Soika, W. Feiten, and S. Teller, "An atlas framework for scalable mapping," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, vol. 2, 2003, pp. 1899–1906 vol.2.

[17] B.-J. Ho, P. Sodhi, P. Teixeira, M. Hsiao, T. Kusnur, and M. Kaess, "Virtual occupancy grid map for submap-based pose graph slam and planning in 3d environments," in *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2018, pp. 2175–2182.

[18] V. Reijgwart, A. Millane, H. Oleynikova, R. Siegwart, C. Cadena, and J. Nieto, "Voxgraph: Globally consistent, volumetric mapping using signed distance function submaps," *IEEE Robotics and Automation Letters*, 2020.

[19] Y. Wang, M. Ramezani, M. Mattamala, S. T. Digumarti, and M. Fallon, "Strategies for large scale elastic and semantic lidar reconstruction," *Journal of Robotics and Autonomous Systems*, vol. 155, p. 104185, 2022.

[20] M. Tancik, V. Casser, X. Yan, S. Pradhan, B. Mildenhall, P. P. Srinivasan, J. T. Barron, and H. Kretzschmar, "Block-nerf: Scalable large scene neural view synthesis," in *IEEE Int. Conf. Computer Vision and Pattern Recognition*, 2022, pp. 8248–8258.

[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Conf. on Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.

[22] S. Fridovich-Keil, A. Yu, M. Tancik, Q. Chen, B. Recht, and A. Kanazawa, "Plenoxels: Radiance fields without neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5501–5510.

[23] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering," *ACM Transactions on Graphics*, vol. 42, no. 4, July 2023.

[24] A. Yu, R. Li, M. Tancik, H. Li, R. Ng, and A. Kanazawa, "PlenOctrees for real-time rendering of neural radiance fields," in *Intl. Conf. on Computer Vision (ICCV)*, 2021.

[25] J. T. Barron, B. Mildenhall, D. Verbin, P. P. Srinivasan, and P. Hedman, "Mip-NeRF 360: Unbounded anti-aliased neural radiance fields," *IEEE Int. Conf. Computer Vision and Pattern Recognition*, 2022.

[26] R. Martin-Brualla, N. Radwan, M. S. M. Sajjadi, J. T. Barron, A. Dosovitskiy, and D. Duckworth, "NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections," in *IEEE Int. Conf. Computer Vision and Pattern Recognition*, 2021.

[27] Z. Li, T. Müller, A. Evans, R. H. Taylor, M. Unberath, M.-Y. Liu, and C.-H. Lin, "Neuralangelo: High-fidelity neural surface reconstruction," in *IEEE Int. Conf. Computer Vision and Pattern Recognition*, 2023.

[28] L. Yariv, J. Gu, Y. Kasten, and Y. Lipman, "Volume rendering of neural implicit surfaces," *Conf. on Neural Information Processing Systems (NeurIPS)*, vol. 34, pp. 4805–4815, 2021.

[29] 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," *Conf. on Neural Information Processing Systems (NeurIPS)*, 2021.

[30] X. Zhong, Y. Pan, J. Behley, and C. Stachniss, "Shine-mapping: Large-scale 3d mapping using sparse hierarchical implicit neural representations," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2023.

[31] J. Deng, Q. Wu, X. Chen, S. Xia, Z. Sun, G. Liu, W. Yu, and L. Pei, "Nerf-loam: Neural implicit representation for large-scale incremental lidar odometry and mapping," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

[32] S. Lombardi, T. Simon, J. Saragih, G. Schwartz, A. Lehrmann, and Y. Sheikh, "Neural volumes: learning dynamic renderable volumes from images," *ACM Transactions on Graphics (TOG)*, vol. 38, no. 4, pp. 1–14, 2019.

[33] V. Sitzmann, J. Thies, F. Heide, M. Nießner, G. Wetzstein, and M. Zollhofer, "Deepvoxels: Learning persistent 3d feature embeddings," in *IEEE Int. Conf. Computer Vision and Pattern Recognition*, 2019, pp. 2437–2446.

[34] V. Sitzmann, M. Zollhöfer, and G. Wetzstein, "Scene representation networks: Continuous 3d-structure-aware neural scene representations," *Conf. on Neural Information Processing Systems (NeurIPS)*, vol. 32, 2019.

[35] N. Max, "Optical models for direct volume rendering," *IEEE Trans. on Visualization and Computer Graphics*, vol. 1, no. 2, pp. 99–108, 1995.

[36] J. T. Kajiya and B. P. Von Herzen, "Ray tracing volume densities," *SIGGRAPH*, vol. 18, no. 3, pp. 165–174, 1984.

[37] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, "Detectron2," https://github.com/facebookresearch/detectron2, 2019.

[38] D. Azinović, R. Martin-Brualla, D. B. Goldman, M. Nießner, and J. Thies, "Neural rgb-d surface reconstruction," in *IEEE Int. Conf. Computer Vision and Pattern Recognition*, 2022, pp. 6290–6301.

[39] U. Von Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, pp. 395–416, 2007.

[40] P. Furgale, J. Rehder, and R. Siegwart, "Unified temporal and spatial calibration for multi-sensor systems," in *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*. IEEE, 2013, pp. 1280–1286.

[41] L. F. T. Fu, N. Chebrolu, and M. Fallon, "Extrinsic calibration of camera to lidar using a differentiable checkerboard model," in *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2023.

[42] H. Aanæs, R. R. Jensen, G. Vogiatzis, E. Tola, and A. B. Dahl, "Large-scale data for multiple-view stereopsis," *International Journal of Computer Vision*, pp. 1–16, 2016.

[43] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.