2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

Target-Aware Implicit Mapping for Agricultural Crop Inspection

Shane Kelly Alessandro Riccardi Elias Marks Federico Magistri
Tiziano Guadagnino Margarita Chli Cyrill Stachniss

***Abstract*— Crop inspection is a critical part of modern agricultural practices that helps farmers assess the current status of a field and then make crop management decisions. Current crop inspection methods are labour-intensive tasks, which makes them rather slow and expensive to apply. In this paper, we exploit recent advancements in implicit mapping to tackle the challenging context of agricultural environments to create dense maps of crop rows with high enough fidelity to be useful for automated crop inspection. Specifically, we map strawberry and sweet pepper crop rows using RGB images captured by a wheeled mobile field robot inside a greenhouse and then use this data to build 3D maps to document the development of plants and fruits. Our Target-Aware Implicit Mapping system (TAIM) uses a SLAM-based pose initialization strategy for robust pose convergence, an efficient information-guided training sample selection framework for faster loss reduction, and focuses on exploiting training samples for fruit regions of the scene, which are critical for crop inspection tasks, to create more accurate maps in less time.**

I. Introduction

Crop inspection is a key part of an informed agricultural operation, where farmers quantify critical characteristics of their fields and crops. Thorough crop inspection empowers optimal decision making, such as selecting harvest times and scheduling appropriate disease treatment plans [4]. However, traditional methods for crop inspection are manually intensive, which can make the procedure slow, expensive, and prone to human error. Even modernized methods rely on equipment such as terrestrial laser scanners, which are cumbersome, expensive, and require manual operation from specially-trained users. Researchers had success in using UAVs for automated crop inspection tasks such as disease diagnosis [7], yield estimation [2], weed detection [19], [18]. However, these methods rely on distant observations from UAVs high above the relevant crops, which can limit effectiveness. New methods for on-the-ground agricultural field navigation [1] and view planning [37] allow for closeup measurements of crops for automated inspection. Despite this progress, many on-the-ground inspection methods [8], [9], [25] generate inspection results from one-shot sensor measurements, rather than using all sensor measurements to create a refined crop model.

Fig. 1: Right: Our robot capturing images of crop rows inside a greenhouse. Left: Qualitative view synthesis results of our method, TAIM, compared to a baseline, BARF. We achieve higher-fidelity view synthesis, especially on fruit regions of the scene.

Recent works on neural rendering [24] and implicit mapping [31] have shown impressive results in using RGB images to model 3D scenes with high enough fidelity such that we believe these technologies could be useful in automated crop inspection tasks when paired with recent crop reconstruction approaches [20], [21], [22]. In this paper, we propose an implicit mapping system able to tackle the challenging context of agricultural field environments. Specifically, we attempt to make high-fidelity maps of strawberry and sweet pepper crops inside greenhouses from RGB images captured by a mobile robot, as shown in Fig. 1. These scenes have large amounts of visual repetition, are only seen from limited view points, require high precision to capture the small fruits, and contain noise from motion blur effects as the wheeled robot drives over uneven terrain [15].

The majority of existing works propose methods for implicit mapping systems on images without known poses, which removes a significant barrier for real-world applications. However, these pose learning strategies often are susceptible to convergence to local minima on complex scenes with repeated visual texture. Additionally, to the

Kelly and Chli are with the Vision for Robotics Lab, ETH Zurich. All other authors are with the University of Bonn. 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 Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy, EXC-2070 – 390732324 – PhenoRob, by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under STA 1051/5-1 within the FOR 5351 (AID4Crops) and by the Federal Ministry of Food and Agriculture (BMEL) based on a decision of the Parliament of the Federal Republic of Germany via the Federal Office for Agriculture and Food (BLE) under the innovation support programme under funding no 28DK108B20 (RegisTer).

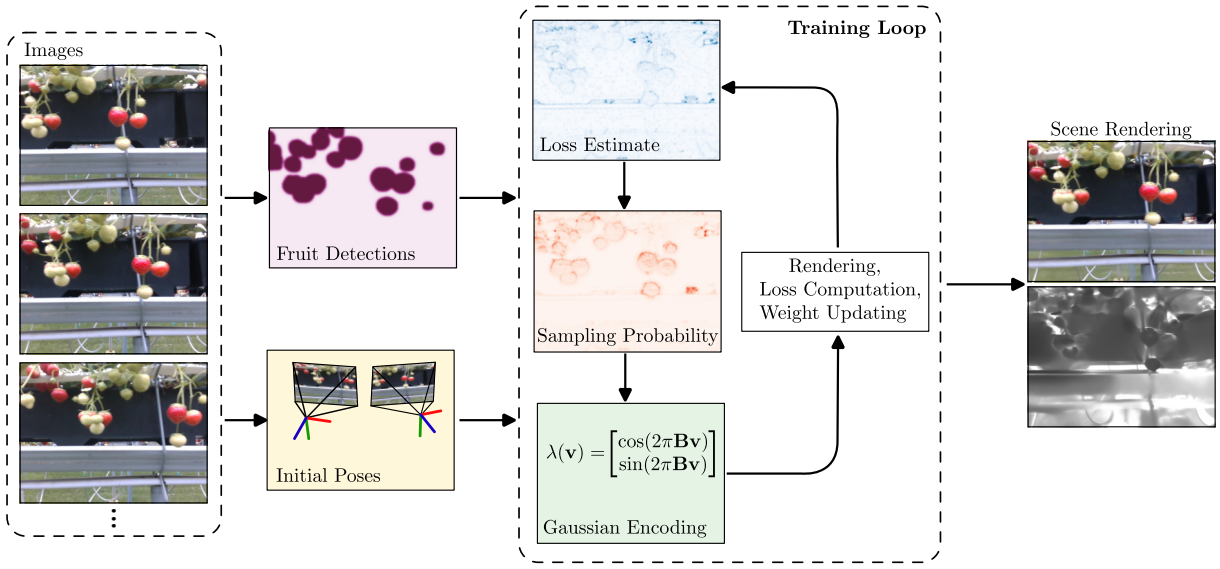

Fig. 2: A system overview of TAIM. All fruits within the input images are detected. A classical visual SLAM system is used to compute approximate initial image poses. In the training loop, a loss distribution over all images is estimated and used, in combination with the fruit detections, to compute a training sampling probability for each pixel, such that pixels inside fruit regions with high loss have the highest probability of being sampled during training. Training pixels are drawn from the sampling probability distribution, encoded using Gaussian positional encoding, and fed to the network.

best of our knowledge, no implicit mapping system uses application-specific knowledge to map certain regions of interest with higher fidelity than others, thereby creating a more accurate map of relevant regions in less time.

The main contribution of this paper is a method to implicitly map large and complex agricultural scenes with high fidelity in key fruit regions. During training, we use a fast method for estimating a dense loss over all training images and efficiently sample training pixels from high-loss regions. This results in learning more information per training sample and leads to convergence in fewer iterations. We also use application-specific knowledge to sample training pixels more densely from fruit regions of the scene, for agricultural crop inspection. Lastly, we use a classical SLAM system to initialize image poses and then continuously refine the poses during training. This reduces the chances that our pose estimates converge to local minima and increases convergence speed. Our experiments suggest that we have: (i) an efficient dense loss estimation framework over input images for focusing the training on high-loss regions and fruits regions of the scene, resulting in view synthesis with better quality, (ii) a system for sampling more densely from fruit regions of the input images during training that results in higher fidelity 3D fruit mapping, and (iii) a pose learning strategy that is initialized by a classical SLAM system and then refined in parallel with the scene during training to improve convergence accuracy and robustness.

II. Related Work

Most classical robot mapping systems use representations such as point clouds [12] or voxel grids [27] to represent the output 3D structure of the mapped scene. The unordered structure of a point cloud can represent a shortcoming to represent objects such as fruits and plants, which are highly spatially correlated. Voxel grids, on the other hand, are limited by the discretization to a hard-coded resolution and the prohibitive memory consumption when scaled to large scenes. In contrast, learning-based approaches that implicitly represent the scene do not lose accuracy due to discretization and efficiently embed the contents of the scene within the weights of the network.

Indirect mapping methods [5], [34] use keypoints as a sparse abstraction for stitching together dense map segments, but the keypoints can sometimes fail to be discriminative, such as in textureless scene regions, whereas neural networks can distinguish relevant features directly from image data for robust alignment. Direct methods [3], [6], [28] improve performance in textureless regions, but often still split pose estimation and scene estimation into two distinct processes, as popularized by Klein et al. [13], which can lead to less-coherent results compared to implicit approaches that use the same photometric loss to simultaneously learn scene contents and refine pose estimates. An alternative approach is the use of neural rendering to represent the 3D scene, so called implicit models. Mildenhall et al. [24] use MLPs to learn spatial contents of a scene from RGB images. Some recent works remove the known-pose requirement by learning image poses through a typical SLAM framework [31], [38], which unlock realtime mapping capabilities, but result in only approximate scene reconstructions which are often too over-smoothed for many high-precison applications.

Other recent systems use longer training times for higher fidelity map results while learning image poses from scratch [16], [35], [36], but are sensitive to repeated visual texture in the scene leading to pose convergence to local minima or make assumptions about the scene that limits real-world application of the systems. Some works make additional efforts to bring implicit mapping pipelines to fully

unconstrained real-world settings, such as accommodating input images taken in different lighting conditions [23], [32], but still often require multiple days of training to converge.

III. Our Approach - TAIM

Our system architecture is inspired by BARF [16], a recent and impressive implicit mapping pipeline proposed by Lin that adds pose learning capabilities to the original NeRF system [24]. It extends BARF and improves results in agricultural environments for crop inspection, as outlined in Fig. 2. Instead of initializing our pose estimates to the identity transform, we use a classical SLAM system for pose initialization to reduce the risk that the pose estimation converges to a local minimum. Using pose initialization also means that we do not need as wide of a basin of convergence for pose estimation, which allows us to replace BARF's coarse-to-fine positional encoding with Gaussian positional encoding [33], a more principled approach to the original NeRF positional encoding that leads to faster scene convergence that is still smooth enough for our initialized poses to converge accurately to the true poses. We replace BARF's uniform training pixel sampling strategy with an efficient loss-targeted strategy that samples more often from complex regions that are not yet mapped with high fidelity. We also target fruit regions of the scene during training sampling, such that those regions of the scene are mapped with higher fidelity than non-fruit regions.

A. Pose Initialization

Our crop row datasets have a large amount of repeated visual structure, such as clusters of strawberries, that often attract image pose estimates into local minima. The long and narrow shape of the crop rows also delays learning the scene since poses must follow the visual gradient over a large distance before accurate learning of the scene's contents can begin. To avoid these issues, we use a classical SLAM system to compute approximate poses of all our input images and use these approximate poses to initialize our system. We then further refine these initial pose estimates during training.

For pose initialization, we use ORB-SLAM [26], a well-known classical visual SLAM pipeline that uses ORB features [29] to co-register keyframes for tracking and has shown impressive results in many challenging environments.

B. Pose Learning

BARF removes the requirement for known image poses from NeRF by using coarse-to-fine positional encoding to learn image poses in parallel with the contents of the scene. Only low-frequency spatial information is learned early on in training, which creates smooth visual gradients over which image poses can converge using the same view-synthesis-based loss as the original NeRF system. Unlike BARF, TAIM removes the coarse-to-fine encoding and instead uses Gaussian positional encoding and initializes pose estimates from a classical SLAM system. In previous works, the Gaussian positional encoding strategy [33] has been shown to lead to faster convergence during training. Gaussian positional encoding requires encoding the input 3D coordinates with a matrix $\mathbf{B}$ of a chosen size and initialized by drawing values from a standard normal distribution multiplied by a chosen scale parameter. The objective for simultaneous pose and scene optimization, as formulated by BARF and used by our system, is

$$\min_{\mathbf{p}_{1:M},\boldsymbol{\Theta}} \sum_{i=1}^{M} \sum_{\mathbf{u}} \left\| \hat{\mathcal{I}}(\mathbf{u};\mathbf{p}_i,\boldsymbol{\Theta}) - \mathcal{I}_i(\mathbf{u}) \right\|_2^2, \tag{1}$$

where $\{\mathcal{I}_i\}_{i=1}^{M}$ and $\{\mathbf{p}_i\}_{i=1}^{M}$ are the images and associated poses, $\mathbf{u}$ is the set of all pixels in all training images, $\hat{\mathcal{I}}$ is the rendering function, and $\boldsymbol{\Theta}$ is the network parameters. The rendering function is differentiable and thus allows gradients to be traced end-to-end from the computed loss back to the pose and scene parameters, which are updated during training to encode the optimized poses of the input images and the high-fidelity spatial contents of the scene.

C. Loss-Targeted Sampling

The complexity of our scene is highly irregular, with a mix of extremely low-complexity regions, such as large solid-black plant boxes, and extremely high-complexity regions, such as fruits or vines. A simple uniform sampling strategy for selecting image pixels $\mathbf{s}$ to train on

$$\mathbf{s} \sim \mathcal{U}\left(\frac{1}{|\mathbf{u}|}\right) \tag{2}$$

results in inefficiently over-sampling from the low-complexity regions, and under-sampling from the high-complexity regions. To correct this problem, Sucar et al. [31] propose estimating an approximate loss distribution, L, over an 8×8 pixel grid in each training image, and then sampling training pixels proportionally to this distribution

$$\mathbf{s} \sim \frac{L}{|\mathbf{u}|}. \tag{3}$$

We found that this strategy still resulted in over-sampling from low-complexity regions that were in the same grid section as high-complexity regions, and vice-versa, due to the low resolution of the loss estimate. Additionally, we found this method to be prohibitively computationally expensive due to the large number of pixels being rendered on every training iteration just to estimate the loss distribution.

We propose a loss distribution estimation method that results in a pixel-level resolution loss estimate without needing any additional scene render computations. On the first training iteration, we sample uniformly over all training images, then set the loss estimate to a flat distribution of the average loss of all samples. On every subsequent training iteration we sample training pixels proportionally to a lightly blurred version of the loss estimate, render those samples as we normally would, and then update the loss estimate at those pixel locations to the loss values we just computed. Since we already compute loss during training, this system does not add additional computation. We sample from a blurred version of the loss estimate so that each pixel in the loss estimate can propagate local effects to nearby regions.

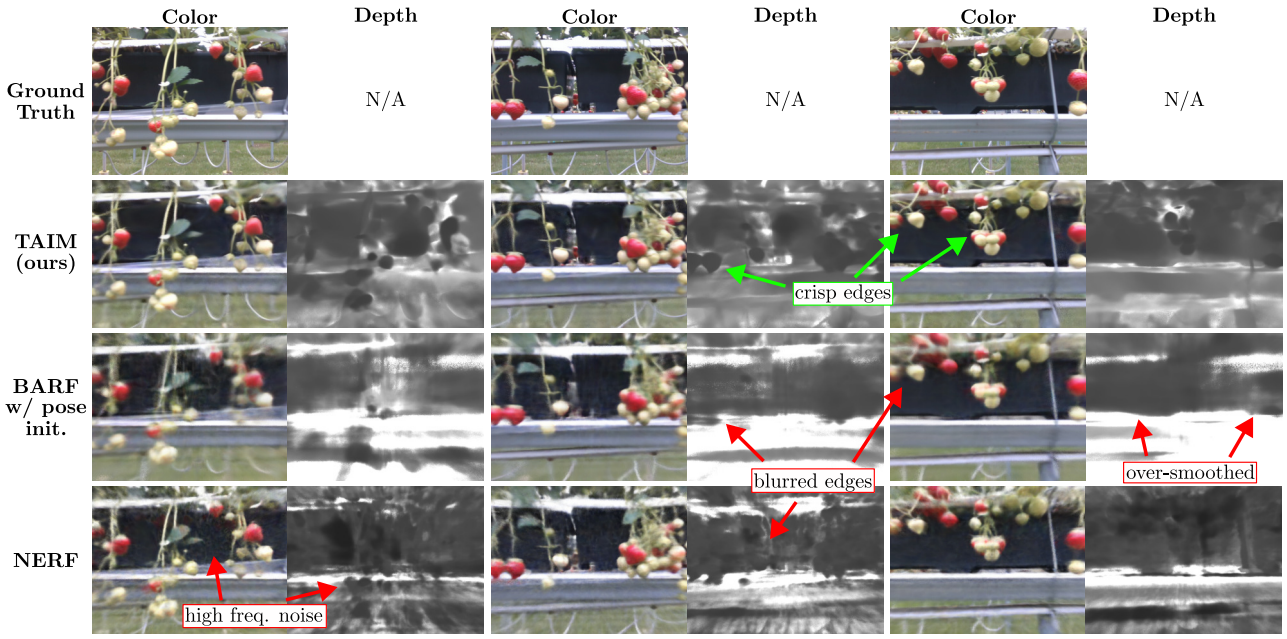

Fig. 3: Qualitative view synthesis results on our 3 meters strawberry dataset with over 200 fruits. During view synthesis, TAIM results have crisp edges, while BARF tends to over-blur, and NERF is susceptible to high-frequency noise.

In practice, the loss estimate can become stale in some extremely low-loss regions that are not sampled for many thousands of training iterations. Therefore, we use an exploration-exploitation framework to balance between sampling high-loss regions to learn the contents of the scene and sampling low-loss regions to update our loss estimate. We implement this by imposing a minimum probability for sampling any pixel, s_{min}, using a variable, α, that tunes the ratio between the highest sampling probability and the lowest sampling probability. In other words, $s_{\text{max}} = \alpha\, s_{\text{min}}$. Therefore, we sample training pixels such that

$$\mathbf{s} \sim \frac{L\,\frac{1-\alpha}{L_{\text{max}}} + \alpha}{|\mathbf{u}|}, \tag{4}$$

where L_{max} is the maximum value in the loss estimate. The numerator scales the loss estimate distribution, L, between α and $1-\alpha$ and the denominator normalizes it to a probability distribution. An example training image and the resulting blurred loss estimate in Fig. 2 show that complex scene regions, such as regions with large gradients, are clearly detected as having high loss and thus sampled more.

D. Fruit-Targeted Sampling

Fruits are a key focus of the crop inspection process so we aim to map them with higher fidelity than non-fruit regions of the scene. To accomplish this, we use a neural network for image segmentation to create a mask of all fruit pixels in our training images. We then slightly dilate this mask to ensure that all edges of the fruits are captured and then apply Gaussian blurring to create a smooth transition from fruit regions to non-fruit regions. This results in a distribution F over all pixels in our training images that describes how fruity each pixel is. An example training image and the resulting fruit mask are in Fig. 2. We scale F to be between 1 and β and then modify our loss-targeted sampling procedure to weigh each pixel's loss according to its fruityness

$$\mathbf{s} \sim \frac{\left(L\,\frac{1-\alpha}{L_{\text{max}}} + \alpha\right)\,(F\,(\beta-1)+1)}{|\mathbf{u}|}, \tag{5}$$

where β can be tuned to set how much more often fruit regions should be sampled compared to non-fruit regions. We found that increasing β improves the fidelity of fruit-regions of the map, though with increasingly diminishing returns, at the cost of lower non-fruit region fidelity.

Strawberry

	Whole Scene PSNR			Fruit Only PSNR		
# Training Iters.	10k	30k	50k	10k	30k	50k
NERF [24]	19.83	21.60	22.36	18.34	20.55	21.53
BARF [16] w/ pose	16.58	20.34	21.41	12.47	18.39	19.93
TAIM (ours)	**21.07**	**23.36**	**24.25**	**21.03**	**24.50**	**26.05**

Sweet Pepper

	Whole Scene PSNR			Fruit Only PSNR		
# Training Iters.	10k	30k	50k	10k	30k	50k
NERF [24]	17.51	18.66	19.15	19.44	20.86	21.19
BARF [16] w/ pose	14.37	16.94	17.73	15.22	18.50	19.40
TAIM (ours)	**17.61**	**18.93**	**19.44**	**20.24**	**21.91**	**22.73**

TABLE I: Quantitative view synthesis results. PSNR quantifies the similarity between rendered images and ground truth images.

	Whole Scene			Fruit Only		
	precision	recall	f-score	precision	recall	f-Score
NERF [24]	46.25	65.70	54.29	10.87	39.65	17.06
BARF [16] w/ pose	44.31	73.62	55.32	**16.36**	53.24	25.03
TAIM (ours)	**58.08**	**77.00**	**66.22**	16.34	**63.32**	**25.98**

TABLE II: Quantitative 3D scene reconstruction results.

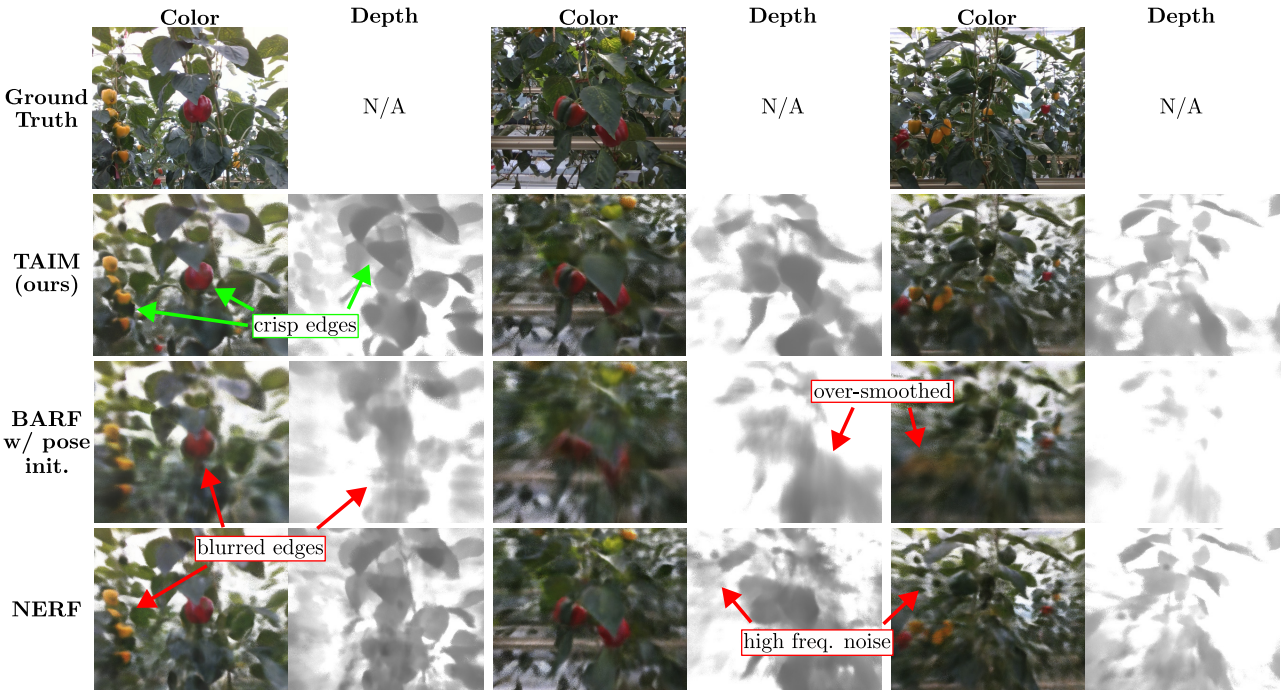

Fig. 4: Qualitative view synthesis results on 5 meters long sweet pepper dataset. TAIM synthesizes views with crisp and accurate edges, whereas BARF produces blurred edges and NERF renderings have a large amount of high-frequency noise.

IV. Experimental Evaluation

The main focus of this work is to robustly and accurately map strawberry and sweet pepper crop rows. The experiments presented here support our key claims, i.e., that we propose (i) a framework for focusing the training on high-loss regions of the scene, resulting in better quality view synthesis, (ii) a system for sampling most densely from fruit regions, resulting in higher fidelity 3D fruit mapping, and (iii) a classically-initialized pose learning strategy that is refined in parallel with the scene during training to improve convergence accuracy and robustness. Our experiments qualitatively and quantitatively evaluate our ability to synthesize 2D views and reconstruct 3D scenes compared to BARF and NERF.

A. Experimental Setup

We evaluate our implicit mapping pipeline on strawberry and sweet pepper crop rows inside a greenhouse near Bonn, Germany. For the sweet pepper, we used a part of the BUP20 dataset [8], [30] consisting of 200 images with over 50 prominent fruits. We collect our own strawberry dataset consisting of 146 images with over 200 prominent fruits. For the strawberry dataset, we additionally obtained a high-precision point cloud using a terrestrial laser scanner that we use as ground truth. Fruits within the ground truth strawberry point cloud are hand labeled such that points belonging to fruits can be isolated from the rest of the point cloud to compute fruit-only and whole-scene reconstruction metrics. Since BARF without pose initialization converges to extremely incorrect pose estimates, we compare TAIM to BARF with pose initialization in the following experiments.

B. Implementation Details

Similar to BARF, our MLP uses four, 256-node hidden layers with ReLU activation and we use the Adam optimizer to compute updates for both scene and pose parameters. We use a scene learning rate of 10^{-3}. Since our pose parameters are initialized with an approximate guess, we use a lower pose learning rate of 10^{-6}. Our Gaussian positional embedding matrix is 3×256 and is initialized using a scale of 12. For fruit detection in our training images, we use two Mask R-CNN networks [10] with the ResNet50 architecture [11] pretrained on the COCO dataset for object segmentation [17] and then refined to segment strawberries or sweet peppers using about 200 images each.

C. 2D View Synthesis

The first experiment shows that our loss estimation framework results in better quality of the synthesized views. We compare the ability of TAIM to synthesize views against the NERF and BARF baselines. We also qualitatively compare view synthesis results in Fig. 3 and Fig. 4, which show that we achieve higher synthesis quality on both color and depth renderings, especially in fruit regions of the scene, compared to BARF and NERF. TAIM generalizes well to both strawberries and peppers despite extreme differences in fruit size, texture, and color. We use peak signal to noise ratio (PSNR) to quantify the similarity between rendered images and ground truth images. Tab. I shows that TAIM achieves higher PSNR on whole-image and fruit-only regions at every evaluated training iteration interval compared to both NERF and BARF, due to its efficient sampling during training. At the final training iteration of the strawberry dataset, TAIM improves fruit-only PSNR by 21% and 31% compared to NERF and BARF, respectively. NERF is able to outperform BARF on these metrics due to BARF's coarse-to-fine encoding, which delays the schedule at which the network can learn high-frequency contents of the scene.

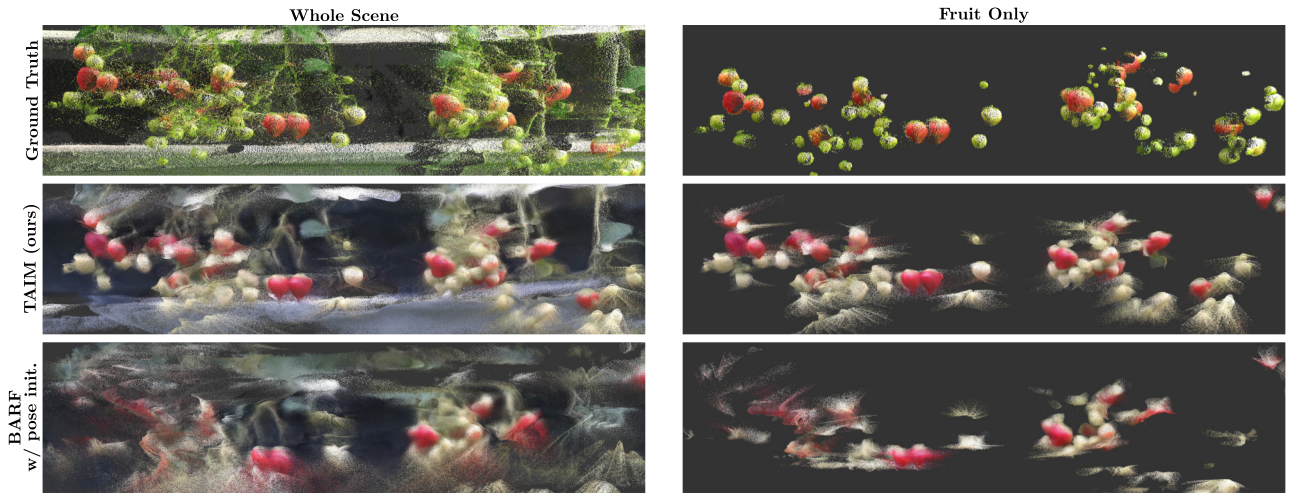

Fig. 5: Segments of the point clouds from TAIM and BARF compared to the ground truth. Relative to BARF, TAIM achieves a noticeably more well-defined reconstruction that better matches the ground truth in both the whole scene (left) and the fruit-only regions (right).

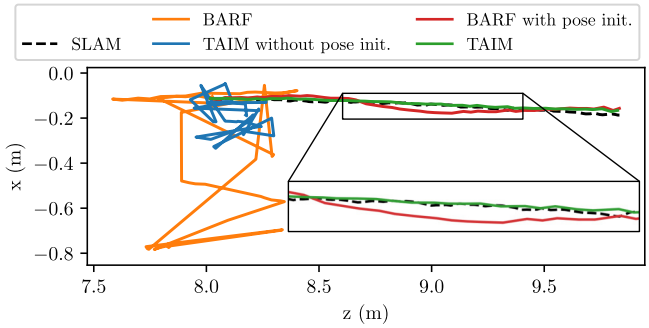

Fig. 6: Pose estimation of BARF and TAIM on the strawberry dataset with and without pose initialization from a SLAM system. Pose initialization is necessary for convergence to accurate poses.

D. 3D Scene Reconstruction

We evaluate the 3D mapping performances on the strawberry dataset by extracting a point cloud after training and comparing it to ground truth. We extract whole-scene point clouds by unprojecting each pixel of the RGB and depth renderings from every input training image into the scene and obtain fruit-only point clouds by filtering based on fruit detection on images. A qualitative comparison of cropped ground truth, TAIM, and BARF point clouds is shown in Fig. 5, which shows that TAIM generates clearer reconstructions with less blurring than BARF. For quantitative evaluation, following Knapitsch et al. [14], we compute precision, recall, and f-score on the point cloud reconstructions from NERF, BARF, and TAIM relative to the ground truth. As shown in Tab. II, compared to both BARF with pose initialization and NERF, TAIM improved the combined f-score metric on both the whole scene and the fruit-only scene. The achieved f-scores of NERF, BARF, and TAIM were 17.06, 25.03, and 25.98, respectively. BARF achieved a very slightly higher precision score on the fruit-only scene, which is sensitive to depth bleeding effects.

E. Pose Convergence

To show that pose initialization is necessary for our challenging agricultural scene, we qualitatively evaluate the pose estimation of BARF and TAIM with and without pose initialization on the strawberry dataset. Fig. 6 shows that BARF with pose initialization and TAIM have pose estimates that closely resemble the pose estimates provided by a classical SLAM system, which indicates that the pose estimates have converged to approximately correct values. However, the pose estimates for BARF and TAIM without pose initialization do not resemble the estimates from the SLAM system, indicating convergence to incorrect values. While the initial pose estimates are of good quality, we found that further pose refinement during training was still necessary to ensure crisp mapping without “doubled” edges.

V. Conclusion

In this paper, we presented a novel approach to implicit mapping in agricultural environments. Our system leverages a novel loss estimation method over all training images to efficiently focus training on high-information complex scene regions. Our method places additional training focus on fruit regions of the scene, a key part of agricultural crop inspection, to map them with higher fidelity than non-fruit regions. Additionally, our system uses a classical SLAM system to initialize pose estimates for each input image, which are then refined in parallel with scene mapping during training. We evaluated our system on data from strawberry and sweet pepper rows inside a greenhouse and compared its ability to do 2D view synthesis and 3D scene reconstruction against BARF, a state-of-the-art implicit mapping system, and its well-known predecessor, NERF. Our experiments suggest that TAIM produces higher-fidelity synthesized views and more accurate scene reconstructions than both BARF and NERF, especially in critical fruit regions of the scene. Potential areas for further improvement include the use of multi-camera platforms with different viewing angles for more accurate depth estimation, and targeting critical regions beyond just fruits during training, such as leaves and vines, whose intricate structure is challenging to map with high fidelity.

REFERENCES

[1] A. Ahmadi, M. Halstead, and C. McCool. Towards autonomous crop-agnostic visual navigation in arable fields. *arXiv preprint arXiv:2109.11936*, 2021.

[2] A. Barreto, P. Lottes, F.R.I. Yamati, S. Baumgarten, N.A. Wolf, C. Stachniss, A.K. Mahlein, and S. Paulus. Automatic uav-based counting of seedlings in sugar-beet field and extension to maize and strawberry. *Computers and Electronics in Agriculture*, 191:106493, 2021.

[3] M. Bloesch, M. Burri, S. Omari, M. Hutter, and R. Siegwart. Iterated extended kalman filter based visual-inertial odometry using direct photometric feedback. *Intl. Journal of Robotics Research (IJRR)*, 36(10):1053–1072, 2017.

[4] H. Bolley. Official field crop inspection. *Science*, 50(1287):193–199, 1919.

[5] 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.

[6] P. Geneva, K. Eckenhoff, W. Lee, Y. Yang, and G. Huang. Openvins: A research platform for visual-inertial estimation. In *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, pages 4666–4672, 2020.

[7] F. Görlich, E. Marks, A.K. Mahlein, K. König, P. Lottes, and C. Stachniss. Uav-based classification of cercospora leaf spot using rgb images. *Drones*, 5(2):34, 2021.

[8] M. Halstead, A. Ahmadi, C. Smitt, O. Schmittmann, and C. McCool. Crop agnostic monitoring driven by deep learning. *Frontiers in plant science*, 12, 2021.

[9] M. Halstead, C. McCool, S. Denman, T. Perez, and C. Fookes. Fruit quantity and quality estimation using a robotic vision system. *arXiv preprint arXiv:1801.05560*, 2018.

[10] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask R-CNN. In *Proc. of the IEEE Intl. Conf. on Computer Vision (ICCV)*, 2017.

[11] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[12] C. Kerl, J. Sturm, and D. Cremers. Dense visual slam for rgb-d cameras. In *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, pages 2100–2106, 2013.

[13] G. Klein and D. Murray. Parallel Tracking and Mapping for Small AR Workspaces. In *Proc. of the Intl. Symposium on Mixed and Augmented Reality (ISMAR)*, 2007.

[14] A. Knapitsch, J. Park, Q. Zhou, and V. Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Trans. on Graphics*, 36(4):1–13, 2017.

[15] J. Le Louëdec and G. Cielniak. 3d shape sensing and deep learning-based segmentation of strawberries. *Computers and Electronics in Agriculture*, 190:106374, 2021.

[16] C.H. Lin, W.C. Ma, A. Torralba, and S. Lucey. Barf: Bundle-adjusting neural radiance fields. In *Proc. of the IEEE/CVF Intl. Conf. on Computer Vision (ICCV)*, pages 5741–5751, 2021.

[17] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C.L. Zitnick. Microsoft COCO: Common Objects in Context. In *Proc. of the Europ. Conf. on Computer Vision (ECCV)*, pages 740–755, 2014.

[18] P. Lottes, J. Behley, A. Milioto, and C. Stachniss. Fully convolutional networks with sequential information for robust crop and weed detection in precision farming. *IEEE Robotics and Automation Letters (RA-L)*, 3:3097–3104, 2018.

[19] P. Lottes, N. Chebrolu, F. Liebisch, and C. Stachniss. UAV-based Field Monitoring for Precision Farming. In *25. Workshop Computer-Bildanalyse in der Landwirtschaft*, 2019.

[20] F. Magistri, N. Chebrolu, J. Behley, and C. Stachniss. Towards In-Field Phenotyping Exploiting Differentiable Rendering with Self-Consistency Loss. In *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, 2021.

[21] F. Magistri, E. Marks, S. Nagulavancha, I. Vizzo, T. Läbe, J. Behley, M. Halstead, C. McCool, and C. Stachniss. Contrastive 3d shape completion and reconstruction for agricultural robots using rgb-d frames. *IEEE Robotics and Automation Letters (RA-L)*, 7(4):10120–10127, 2022.

[22] E. Marks, F. Magistri, and C. Stachniss. Precise 3D Reconstruction of Plants from UAV Imagery Combining Bundle Adjustment and Template Matching. In *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, 2022.

[23] R. Martin-Brualla, N. Radwan, M.S. Sajjadi, J.T. Barron, A. Dosovitskiy, and D. Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 7210–7219, 2021.

[24] B. Mildenhall, P. Srinivasan, M. Tancik, J. Barron, R. Ramamoorthi, and R. Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[25] A. Milioto, P. Lottes, and C. Stachniss. Real-time Semantic Segmentation of Crop and Weed for Precision Agriculture Robots Leveraging Background Knowledge in CNNs. In *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, 2018.

[26] R. Mur-Artal and J. Tardós. ORB-SLAM2: An Open-Source SLAM System for Monocular, Stereo, and RGB-D Cameras. *IEEE Trans. on Robotics (TRO)*, 33(5):1255–1262, 2017.

[27] 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)*, pages 127–136, 2011.

[28] T. Qin, P. Li, and S. Shen. Vins-mono: A robust and versatile monocular visual-inertial state estimator. *IEEE Trans. on Robotics (TRO)*, 34(4):1004–1020, 2018.

[29] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski. Orb: an efficient alternative to sift or surf. In *Proc. of the IEEE Intl. Conf. on Computer Vision (ICCV)*, 2011.

[30] C. Smitt, M. Halstead, T. Zaenker, M. Bennewitz, and C. McCool. Pathobot: A robot for glasshouse crop phenotyping and intervention. In *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, 2021.

[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)*, pages 6229–6238, 2021.

[32] 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. *arXiv preprint arXiv:2202.05263*, 2022.

[33] M. Tancik, P. Srinivasan, B. Mildenhall, S. Fridovich-Keil, N. Raghavan, U. Singhal, R. Ramamoorthi, J. Barron, and R. Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in Neural Information Processing Systems*, 33:7537–7547, 2020.

[34] V. Usenko, N. Demmel, D. Schubert, J. Stückler, and D. Cremers. Visual-inertial mapping with non-linear factor recovery. *IEEE Robotics and Automation Letters (RA-L)*, 5(2):422–429, 2019.

[35] Z. Wang, S. Wu, W. Xie, M. Chen, and V.A. Prisacariu. Nerf–: Neural radiance fields without known camera parameters. *arXiv preprint arXiv:2102.07064*, 2021.

[36] L. Yen-Chen, P. Florence, J.T. Barron, A. Rodriguez, P. Isola, and T.Y. Lin. inerf: Inverting neural radiance fields for pose estimation. In *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, pages 1323–1330, 2021.

[37] T. Zaenker, C. Smitt, C. McCool, and M. Bennewitz. Viewpoint planning for fruit size and position estimation. In *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, pages 3271–3277.

[38] 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. *arXiv preprint arXiv:2112.12130*, 2021.