Plant Geometry Reconstruction From Field Data Using Neural Radiance Fields

Anushrut Jignasu[1], Ethan Herron[1], Talukder Jubery[1], James Afful[1], Aditya Balu[1], Baskar Ganapathysubramanian[1], Soumik Sarkar[1], Adarsh Krishnamurthy[1]*

[1]Iowa State University
* adarsh@iastate.edu

Abstract

Real-time simulations of large-scale farming operations would provide farmers with data-driven and physics-consistent decision support. These real-time farming simulations could be accomplished using predictive digital twins. Predictive digital twins of biological entities allow for a virtual simulation of real-life processes for various environmental conditions, thus paving the way for a comprehensive understanding of various biological responses. One of the first steps in constructing a predictive digital twin is the 3D reconstruction of plant geometry. While traditional approaches for the reconstruction of plant geometry exist, they require a very expensive setup using a LIDAR or destructive imaging of the plant in a controlled environment. Neural approaches for 3D scene reconstruction have alleviated the data collection burden associated with traditional 3D reconstruction methods. In this work, we demonstrate the ability to generate a 3D reconstruction (mesh) of a maize plant by leveraging a recent work in 3D computer vision, Neural Radiance Fields (NeRFs), which uses data collected from a mobile phone camera. Our approach aims to generate high-resolution geometric models for several downstream tasks, such as developing a predictive digital twin.

Introduction

The agricultural industry can gain immensely by injecting more high-quality information into decision-making for farming, leading to crop production increases. One of the most effective ways to make such informed decisions is through a self-contained feedback process accomplished by creating a digital twin of the field. A digital twin of farmland would enable a farmer to digitally simulate the impacts of an action (such as applying pesticides, nitrogen, herbicides, etc.) on the final yield and soil compositions. In working towards this goal of a robust digital twin of large-scale farming operations, there are a few key roadblocks, one of which is the digital representation of crops. Accurate geometry representation is a challenging problem. This issue is exacerbated for real-world data owing to disparate sources of information. Thus, the need for a robust geometry representation is a top priority. Recent advancements in 3D computer vision, specifically using implicit neural representations, have made it relatively easy to reconstruct meshes from various input data (images, point clouds, distance fields, voxels). The ability of implicit functions to be decoupled from resolution-based constraints allows for rapid processing of the input data. Additionally, voxels, meshes, and point cloud representations are not memory efficient, given their dependency on resolution. In this work, we leverage Neural Radiance Fields (NeRF) (Mildenhall et al. 2020a) to represent a 3D scene using images and generate a 3D reconstruction of a maize plant. In computer graphics, accurately modeling light for viewing tasks has been a well-studied problem. Traditional radiance fields describe color and density for every point and viewing direction in a given scene. We can leverage this concept to synthesize views for a given set of images when combined with neural networks, specifically MLPs (Multi-layer Perceptrons).

We focus on the geometry reconstruction of slender, complex, and flexible structures observed in plants (maize plants in our case). We adopt the Mip-NeRF 360 framework to generate a 3D mesh of a maize plant. This framework offers anti-aliasing attributes, which excel at generating 3D reconstructions from 2D images taken in a 360-degree manner around a central point in the scene. This central point in our work is the maize plant. NeRF methods can generate 3D reconstructions with relatively small amounts of data, allowing the use of sparsely sampled images. In this case, it is a video taken from a mobile phone. Our work improves current 3D reconstruction methods in two ways. First, we propose a simpler and cheaper data collection process that an inexperienced user can perform with a mobile phone capable of recording video. This dramatically eases the burden typically associated with in-field data collection compared to traditional methods relying on LiDAR scanners. Second, we implement an implicit neural method for surface reconstruction based on the NeRF scene.

Our key contributions are:

1. We capture and generate pose information for a scene with a maize plant captured using a mobile phone, relieving the need for costly point cloud capture equipment.
2. Using captured images, we obtain a Neural Radiance Field (NeRF), an implicit representation of the scene, where we can synthesize any novel view.
3. Extract a dense point cloud from predicted depth images.

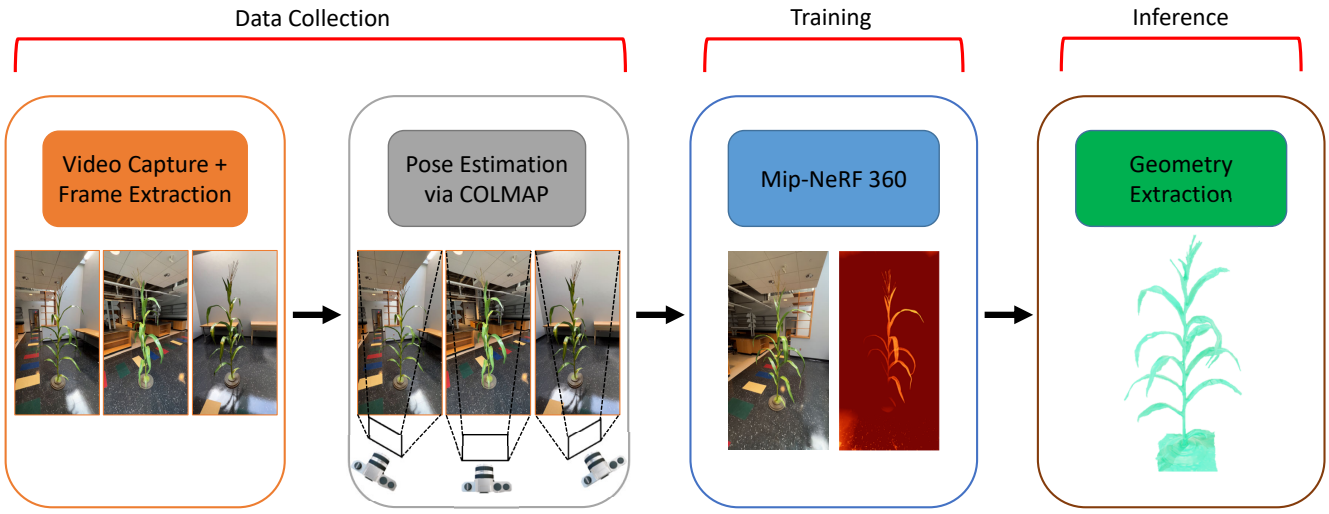

Figure 1: Outline of our method. We begin with the data collection process, followed by training using Mip-NeRF 360 architecture and geometry extraction.

Related Work

Neural counterparts of implicit representations (parameterizing the input as a continuous function that maps the domain of the input to a quantity of interest), commonly called Implicit Neural Representations, have gained immense traction in recent years and have been used for various applications ranging from solving a boundary value problem for surface reconstruction (Sitzmann et al. 2020a), shape reconstruction (Chen and Zhang 2019; Chibane, Alldieck, and Pons-Moll 2020; Tagliasacchi, Zhang, and Cohen-Or 2009), and representing 3D scenes (Mildenhall et al. 2020b; Yu et al. 2021). Voxel carving is a well-developed method for 3D visual hull reconstruction (Kutulakos and Seitz 1999; Schultze et al. 2012). It has been successfully applied for the 3D reconstruction of plants, particularly in the context of plant phenotyping (Tross et al. 2021; Gaillard et al. 2020). However, multiple cameras are needed to capture images of the plant from different view angles, which makes the method data intensive.

3D shape reconstruction is a challenging problem in computer graphics and 3D computer vision. Recent methods have demonstrated the ability to learn a continuous representation that maps *xyz* coordinates to a signed distance field (Atzmon and Lipman 2020; Williams et al. 2021; Sitzmann et al. 2020a) over an unknown 3D geometry represented by a point cloud. One such approach, called DeepSDF (Park et al. 2019), uses a neural network to approximate the continuous signed distance function of a 3D geometry. The DeepSDF network predicts the signed distance of a randomly sampled 3D coordinate and the corresponding nearest point on the 3D geometry's surface. Once fully trained, the network can create a mesh of the 3D geometry. Point cloud representations are tricky to work with owing to their sparse nature, which can lead to problems using data-hungry neural networks. Additionally, when learning the signed distance function over an unknown 3D geometry, we cannot optimize a neural network in a supervised manner, i.e., no ground truth values are available. To circumvent this issue, (Sitzmann et al. 2020b) proposes to optimize the neural network representing a continuous signed distance field of the 3D geometry solely by observing the derivatives of the neural network itself. The network is then optimized to solve an Eikonal boundary value problem where the norm of the spatial gradients must be equal to 1. In practice, the neural network cannot solve this Eikonal boundary value problem over the entire domain. Instead, the network is limited to only approximating a solution to the boundary value problem, reducing the fidelity of the final signed distance function. Additionally, there are nontrivial convergence issues with even arriving at lower fidelity representations of the 3D geometry, given the difficulty of approximating this first-order nonlinear wave equation.

View synthesis is a long-studied problem in the graphics and vision community. The aim is to generate novel photorealistic views given a set of input views. Recent neural view synthesis approaches have demonstrated the ability to compress a continuous representation of a volume into the weights of a neural network (Mildenhall et al. 2020b; Barron et al. 2022; Yu et al. 2021). Early volumetric approaches (Kutulakos and Seitz 2000; Seitz and Dyer 1999; Szeliski and Golland 1998) were based on assigning RGB values to voxels but were not scalable to higher resolution imagery. On the other hand, Neural Radiance Fields (NeRF), (Mildenhall et al. 2020a), demonstrate the ability to compress entire 3D scenes into the weights of a neural network with only a series of 2D RGB images. The NeRF framework is composed of a single Multilayer Perceptron (MLP) queried multiple times using a coarse to fine distillation scheme. First, a ray from each pixel in the 2D image is cast into the 3D scene and sampled at a coarse set of points on the ray. At each of the sampled points along the ray, the MLP conducts a forward pass taking a 5D vector consisting of the

spatial location (x, y, z) and viewing direction (θ, ϕ) as input and approximating the color (RGB) and volume density (σ). Points along this ray cast into the 3D scene are resampled according to the outputs at the coarsely sampled points. This resampling allows the MLP to sample higher-density areas of the 3D scene, thus minimizing the computation on "less occupied areas" of the scene. The NeRF framework leverages positional encodings to incorporate additional information, which exposes the MLP to higher frequency features in the scene. This amounts to upsampling each cartesian coordinate through an additional non-parameterized function before being used as input to the MLP. Once a NeRF model is trained to optimality on a given scene, we can render the 3D scene by using the MLP to predict novel views of the scene, essentially continuously interpolating between 2D images of the scene used during training.

Neural Radiance Fields

In the original NeRF implementation, scenes are sampled by casting a single ray for each pixel in the image. Because of this, distant features in the scene may be blurred during rendering. An immediate solution to this issue is a supersampling-based approach for the scene where we cast several rays for each pixel in the image. This solution would be extremely computationally expensive and, therefore, infeasible given the NeRF MLP is queried multiple times for several points along each ray. Mip-NeRF, (Barron et al. 2021a), addresses this issue without using the computationally expensive supersampling solution. In Mip-NeRF, instead of casting an infinitesimally small ray, as done in NeRF, they propose to cast a cone from each pixel. In practice, the Mip-NeRF MLP is then learning a distribution of values rather than distinct values on a ray. More specifically, the cone cast from each pixel is sliced several times, perpendicular to the direction of the cone cast; the Mip-NeRF MLP is queried to predict the distribution of values over the cross-section of the cone (effectively, a frustum) at each slice along the cone. Instead of the positional encoding used in NeRF, an integrated positional encoding scheme is used, which accounts for the volume of the cone being cast at the sampled slice. This allows the neural network to operate on additional information about the size and shape of each cone rather than strictly the centroid as used in the original NeRF implementation.

Mip-NeRF 360 (Barron et al. 2021b), the method used in this work, extends the Mip-NeRF framework to excel at target scenes where the camera rotates 360 degrees about a central point. First, Mip-NeRF 360 adds a scene and ray parameterization to address unbounded 3D scenes. If NeRF or Mip-NeRF is used with an unbounded 3D scene, it tends to predict blurry and opaque backgrounds due to the sparsity in sampling points that are distant from the camera. The scene and ray parameterization used in Mip-NeRF 360 uses a contraction operator, which, in practice, is similar to an Extended Kalman filter. This contraction operator provides a smooth bound for the projected ray to lie in. The second addition in Mip-NeRF 360 uses two different MLPs for a coarse-to-fine sampling scheme. The coarse network, referred to as the proposal MLP, predicts the volumetric density, represented as a vector of weights. These weights are used to sample high-density intervals for the NeRF MLP to evaluate. Both networks are randomly initialized and trained in tandem.

Methods

An outline of our method is shown in Fig. 1. We begin by describing the high-resolution input image collection process, followed by the training procedure, and point cloud extraction from depth images.

Data Collection

The data required to train a NeRF model is multiple scenes, 2D images, of a given 3D scene. In this work, we collect data to train the NeRF model to obtain a high-fidelity mesh of a maize plant by taking a video on a mobile phone. The video is taken using an iPhone 13 Pro with 4k resolution at 30fps, held at a constant height while circling the plant. To obtain 2D images of the scene from the video data, all that is required is a simple frame extraction operation. We obtained 173 4K resolution images from the video by extracting every 5th frame. These images were then passed through the COLMAP library (Schönberger et al. 2016; Schönberger and Frahm 2016) to get the camera poses for each of the images of the 3D scene. The COLMAP library is a Structure-from-Motion library that takes a series of images and defines the rotation and translation matrix to go from one camera angle to the camera angle in the subsequent 2D image. At this point, we have a working dataset to train Mip-NeRF 360 on since we have the spatial locations (x, y, z) and viewing direction (θ, ϕ) of each image, as well as the RGB values for each pixel from the image itself. The data collection pipeline to a fully curated dataset took approximately 4 hours. To assess the fidelity of the NeRF-rendered mesh, we also collected LIDAR data of the synthetic maize plant to serve as a ground truth comparison. We use a Faro Focus 5 Terrestrial Laser Scanner (TLS) to capture LiDAR data of the synthetic maize plant at 5 points, and each scan with color took 2 minutes 50 seconds at a resolution of $\frac{1}{4}$, and 1x quality. We imported the scans into the SCENE Software for automatic registration and generated a 3D point cloud. The point cloud data was then exported into CloudCompare (CloudCompare 2022), a 3D point cloud processing software, to remove duplicate points using the SOR(Statistical Outlier Removal) filter with 50-nearest neighbors.

Training

In this work, we utilize Mip-NeRF 360's (Barron et al. 2022) codebase. A cone is cast for each pixel, followed by weight initialization for each interval. These weights are updated by a course-to-fine distillation process that aids in anti-aliasing. The training architecture consists of two different MLPs, the proposal MLP and the NeRF MLP. The former consists of 4 layers and 256 hidden units, and the latter consists of 8 layers with 1024 hidden units. Both utilize ReLU activation functions and a softplus activation for density prediction (since density prediction is based on a multi-variate gaussian distribution). We use 150000 training iterations with a sampling grid resolution of 128×128.

(a) Ground truth views.

(b) Predicted views.

Figure 2: (a) Ground truth views used for training, and (b) predicted novel views (top row) and their corresponding depth maps (bottom row).

Figure 3: Multiple views of the plant point cloud extracted from the predicted depth images.

Point Cloud Extraction

The output of every NeRF pipeline is a collection of RGB and density values for each pixel. These density values are stored as depth information and can be utilized to obtain a point cloud of the geometry. Given that we already know the intrinsic camera parameters and have access to novel views and their corresponding depth predictions, we interpolate the depth value for each pixel in the RGB image to obtain the point cloud.

Results and Discussion

Results for novel scenes, as well as the corresponding depth maps, interpolated by the fully trained Mip-NeRF 360 model, are shown in Fig. 2b. The original ground truth images from the dataset are also included to exemplify the high-fidelity scene reconstruction capability of NeRF.

To quantify our reconstructed point cloud, we import the ground truth point cloud and the predicted point cloud into CloudCompare software (CloudCompare 2022). It is an open-source software for computing point cloud metrics. We generate the metrics using point cloud-point cloud compare functionality. The quantitative error metrics are shown in Table 1 and a color-based error plot in Fig. 4. We can see that the average distance for the extracted point cloud is small and further reaffirms the ability of the Mip-NeRF 360's architecture to generate accurate depth predictions.

Table 1: Quantitative comparison between the extracted point cloud and the ground truth point cloud of a maize plant.

Metric	**Value (mm)**
Plant Height	2200.0
Average Distance	17.5
Standard Deviation	12.1

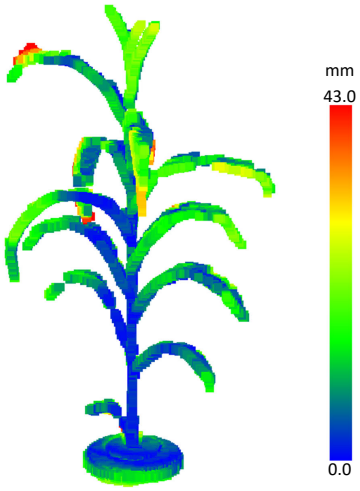

Figure 4: Color-based error plot of extracted point cloud for maize plant.

We empirically investigated the minimum number of views needed to accurately reconstruct the NeRF. We found that COLMAP fails to converge with less than 90 views. We address this challenge by shooting a video at 30 frames per second. All views are taken in a forward-facing manner, i.e., the object of interest is always in front of the camera in all frames. The training time for our scene is $\sim$18.5 hours, with an additional hour for rendering the final scene consisting of RGB values and depth maps. We train on NVIDIA's A100 GPU with 80 GB VRAM. Generalization to outdoor environments has previously been shown to not be problematic, and we expect it to work with most plant geometries, provided enough forward-facing views are captured in a 360-degree manner. We mark mesh extraction from Mip-NeRF 360 as future work.

Conclusion

Predictive digital twins provide us with the ability to simulate and predict plant responses to a variety of environmental conditions. Given the rising threat of climate change, the need for accessible data-driven farming techniques is greater than ever. The most effective data-driven techniques would be real-time simulations by digital twins of entire fields.

The goal of this work is to demonstrate the drastically improved efficacy of generating high-fidelity digital twins of plants from easily captured data, i.e., video data from a mobile phone. This is in stark contrast to the current methods, such as LiDAR point cloud collection, which is time-consuming, intensive, and extremely expensive. Neural Radiance Fields, on the other hand, are notably easier to generate, given the only required data may be captured on any mobile phone. The economical advantages of using Neural Radiance Fields for mesh generation are extremely important milestones in democratizing data-driven agriculture.

Acknowledgments

This work is supported by the National Science Foundation (NSF) and the National Institute of Food and Agriculture (USDA-NIFA) as part of the AI Institute for Resilient Agriculture (AIIRA), Award No. 2021-67021-35329.

References

Atzmon, M.; and Lipman, Y. 2020. Sal: Sign agnostic learning of shapes from raw data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2565–2574.

Barron, J. T.; Mildenhall, B.; Tancik, M.; Hedman, P.; Martin-Brualla, R.; and Srinivasan, P. P. 2021a. Mip-NeRF: A Multiscale Representation for Anti-Aliasing Neural Radiance Fields.

Barron, J. T.; Mildenhall, B.; Verbin, D.; Srinivasan, P. P.; and Hedman, P. 2021b. Mip-NeRF 360: Unbounded Anti-Aliased Neural Radiance Fields.

Barron, J. T.; Mildenhall, B.; Verbin, D.; Srinivasan, P. P.; and Hedman, P. 2022. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5470–5479.

Chen, Z.; and Zhang, H. 2019. Learning implicit fields for generative shape modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5939–5948.

Chibane, J.; Alldieck, T.; and Pons-Moll, G. 2020. Implicit functions in feature space for 3d shape reconstruction and completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6970–6981.

CloudCompare. 2022. CloudCompare (version 2.11.3).

Gaillard, M.; Miao, C.; Schnable, J. C.; and Benes, B. 2020. Voxel carving-based 3D reconstruction of sorghum identifies genetic determinants of light interception efficiency. *Plant direct*, 4(10): e00255.

Kutulakos, K. N.; and Seitz, S. M. 1999. A theory of shape by space carving. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 1, 307–314. IEEE.

Kutulakos, K. N.; and Seitz, S. M. 2000. A theory of shape by space carving. *International journal of computer vision*, 38(3): 199–218.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2020a. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis.

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

Park, J. J.; Florence, P.; Straub, J.; Newcombe, R.; and Lovegrove, S. 2019. DeepSDF: Learning Continuous Signed Distance Functions for Shape Representation.

Schönberger, J. L.; and Frahm, J.-M. 2016. Structure-from-Motion Revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Schönberger, J. L.; Zheng, E.; Pollefeys, M.; and Frahm, J.-M. 2016. Pixelwise View Selection for Unstructured Multi-View Stereo. In *European Conference on Computer Vision (ECCV)*.

Schultze, B.; Witt, M.; Schubert, K. E.; Hurley, R. F.; Bashkirov, V.; Schulte, R. W.; and Gomez, E. 2012. Space carving and filtered back-projection as preconditioners for proton computed tomography reconstruction. In *2012 IEEE Nuclear Science Symposium and Medical Imaging Conference Record (NSS/MIC)*, 4335–4340. IEEE.

Seitz, S. M.; and Dyer, C. R. 1999. Photorealistic scene reconstruction by voxel coloring. *International Journal of Computer Vision*, 35(2): 151–173.

Sitzmann, V.; Martel, J.; Bergman, A.; Lindell, D.; and Wetzstein, G. 2020a. Implicit neural representations with periodic activation functions. *Advances in Neural Information Processing Systems*, 33.

Sitzmann, V.; Martel, J. N. P.; Bergman, A. W.; Lindell, D. B.; and Wetzstein, G. 2020b. Implicit Neural Representations with Periodic Activation Functions.

Szeliski, R.; and Golland, P. 1998. Stereo matching with transparency and matting. In *Sixth International Conference on Computer Vision (IEEE Cat. No. 98CH36271)*, 517–524. IEEE.

Tagliasacchi, A.; Zhang, H.; and Cohen-Or, D. 2009. Curve skeleton extraction from incomplete point cloud. In *ACM SIGGRAPH*, 1–9. ACM.

Tross, M. C.; Gaillard, M.; Zwiener, M.; Miao, C.; Grove, R. J.; Li, B.; Benes, B.; and Schnable, J. C. 2021. 3D reconstruction identifies loci linked to variation in angle of individual sorghum leaves. *PeerJ*, 9: e12628.

Williams, F.; Trager, M.; Bruna, J.; and Zorin, D. 2021. Neural Splines: Fitting 3D Surfaces With Infinitely-Wide Neural Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9949–9958.

Yu, A.; Ye, V.; Tancik, M.; and Kanazawa, A. 2021. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4578–4587.