2950 Niles Road, St. Joseph, MI 49085-9659, USA
269.429.0300 fax 269.429.3852 hq@asabe.org www.asabe.org

An ASABE Meeting Presentation
DOI: https://doi.org/10.13031/aim.202301474
Paper Number: 2301474

Neural Implicit Representation-based Tree Crop Plant 3D Reconstruction

Tian Qiu[1], Jonathan Moon[2], Lailiang Cheng[3], Kaspar Kuehn[3], Yu Jiang[4]

[1]School of Electrical and Computer Engineering, Cornell University, Ithaca, NY

[2]Department of Computer Science, Cornell University, Ithaca, NY

[3]School of Integrative Plant Science, Cornell University, Ithaca, NY

[4]School of Integrative Plant Science, Cornell AgriTech, Cornell University, Geneva, NY

Written for presentation at the
2023 ASABE Annual International Meeting
Sponsored by ASABE
Omaha, Nebraska
July 9-12, 2023

ABSTRACT. *Tree architecture is crucial to both tree crop breeding and precision management and operations such as crop thinning and harvesting. While 3D imaging-based methods have been increasingly explored for tree architecture characterization, most efforts assume or rely on the decent quality of collected 3D images. Achieving decent 3D image quality is non-trivial and has been a long-lasting factor limiting the use of 3D imaging in agricultural applications. Neural representations that are capable of reconstructing scenes with unprecedented fidelity offer new opportunity for 3D reconstruction. The overarching goal of this study was to investigate neural implicit representations to the 3D reconstruction of potted apple trees in controlled and field conditions. Two potted apple trees with heavily pruned branches were selected for data acquisition and comparison of the reconstruction. Each potted apple tree was captured in the form of multi-view 2D images using a consumer grade mobile phone from 360-degree as well as in the form of point cloud using an industry-level laser scanner. While the collection of 2D images were used as the input for classic multi-view reconstruction and neural implicit 3D reconstruction, the processed point cloud from the laser scanner was regarded as the reference of the 3D geometry. Point cloud data generated by multi-view algorithms and neural implicit representations were visually assessed in terms of the correctness and structure completeness and compared with the reference point cloud. Preliminary experiments showed that neural implicit representations considerably outperformed classic multi-view algorithms for 3D reconstruction in both controlled and field conditions. The high-resolution reconstructed point cloud can be used for tree architecture characterization effectively.*

Keywords. *3D Computer Vision, Neural Implicit Representation, Point Cloud Reconstruction, Tree Fruit, Tree Architecture Characterization*

Introduction

The US apple industry is a significant contributor to the country's economy, with notable production and employment numbers. Based on data analyzed by USApple from the United State Department of Agriculture (USDA), it is expected that the total US apple production for the 2022/23 crop year will exceed 10.7 billion pounds, which is a 2.7% increase from last year (USApple, 2022). This production is valued at nearly $3.2 billion, primarily from fresh apple production. The industry also provides employment for over 300,000 individuals, including farmers, farm workers, and those involved in processing and distribution. Furthermore, apples are a crucial export crop, with export revenues of approximately $900 million in 2021. To fulfill the demand and sustain the industry, new cultivar breeding and precision management are two viable pathways that both require the thorough understanding and manipulation of tree morphology. Morphological traits such as tree height, trunk diameter, and branch diameter can serve as effective indicators of tree productivity, enabling growers to accurately assess crop potential and manage crop load with low costs.

Traditional methods for tree morphology characterization have relied on manual measurements and observations, which are often laborious and subject to human error and variability. The limitations of these methods have prompted researchers to explore non-invasive, versatile, and affordable optical sensing technologies for plant characterization. While 2D imaging-based systems have been intensively used to provide robust and reliable solutions for plant morphology characterization (Jiang & Li, 2020), the measurement accuracy was largely limited by occlusion and scene ambiguity due to a single perspective projection in 2D images. To overcome the limitations of 2D image-based methods, researchers have extensively studied and improved three-dimensional (3D) sensing technologies (Ge et al., 2021; Jin et al., 2018; Qiu et al., 2022; Sun et al., 2020; Sun et al., 2021; Westling et al., 2021). However, most efforts in this direction have assumed or relied on the decent quality of 3D representations such as mesh, point cloud, and voxels of objects. Achieving decent 3D representations of objects quality is non-trivial, and it has been a long-lasting factor limiting the use of 3D imaging technologies in agricultural applications.

Common 3D optical sensing technologies in agriculture include RGB-D sensors, multi-view systems and LiDAR scanners. Multi-view systems that utilize a collection of images taken from multiple angles around the object to generate a 3D reconstruction have provided promising results in many agriculture applications (Gené-Mola et al., 2020; Luo et al., 2022; Sun et al., 2020). One of the main advantages of multi-view systems is their affordability, as they often require only a standard camera or smartphone to capture images, whereas RGB-D sensors and LiDAR scanners usually have a much higher cost to offer high resolution and great accuracy in capturing fine details of plant morphology. Additionally, LiDAR scanners require a significant amount of time to capture and process data, whereas multi-view systems can rapidly capture images from multiple angles in a matter of seconds. Classic 3D multi-view reconstruction algorithms include Structure from Motion (SfM) (Schonberger & Frahm, 2016) and Multi-View Stereo (MVS) (Furukawa & Ponce, 2009; Galliani et al., 2015), which rely on image matching and triangulation to estimate the 3D structure of the scene. However, these methods have limitations due to the heavy dependence on the quality of correspondence matching, such as sensitivity to lighting conditions and the presence of occlusions, leading to occurrence of artifacts on textureless objects and inaccuracies in the reconstructed meshes or point clouds. Recently, learning-based MVS methods have gained increasing attention due to their capability of handling challenging scenarios. These methods apply deep neural networks for better pair-wise patch matching and cost volume regularization (Ma et al., 2022; Yao et al., 2018; Yao et al., 2019). A vast majority of these methods generate dense point cloud through depth-map fusion, thereby they are limited by discrete representations of the scene. On the other hand, neural implicit representations that learn to represent a scene as a continuous function have been actively explored to improve the geometry and appearance of 3D reconstructed objects given reasonably estimated camera parameters (Genova et al., 2020; Jiang et al., 2020; Kar et al., 2017; Park et al., 2019). These methods take advantage of the strong expressivity of deep neural networks and are capable of reconstructing scenes with unprecedented fidelity.

In principle, neural implicit methods represent the 3D scene as a continuous function by learning from discretely represented samples of the same scene captured under various lighting and viewing conditions. Neural radiance fields (NeRF) (Mildenhall et al., 2021) and neural surface models (Niemeyer et al., 2020; Yariv et al., 2020) are two representative neural implicit representations of 3D geometry and appearance. Radiance field is a function that describes the appearance of a 3D scene from a particular viewpoint. It specifies the radiance, or brightness, of each point in the scene as it is seen from that viewpoint. The radiance field in NeRF is represented as a neural network that takes in the 3D coordinates of a point and the viewing direction and outputs the corresponding radiance value. The 3D points were sampled along rays generated using the ray tracing rendering technique, and volume rendering (Max, 1995) was used to learn alpha-compositing of a radiance field along rays. To make the sampling more efficient, NeRF proposed a hierarchical sampling strategy by optimizing two neural networks simultaneously: one coarse and one fine. The points were uniformly sampled in the coarse network and were evaluated at each sampled location. Given the output of this coarse network, a more informed sampling of points was biased towards the relevant part of the volume. While NeRF has shown impressive results on novel view synthesis, there are critical assumptions for NeRF to work well. First, NeRF requires the camera parameters to be correctly estimated using classic photogrammetric methods (e.g., COLMAP (Schonberger & Frahm, 2016)), which is a strong assumption in terms of the scene of interests where there needed to have enough corresponding features between images. Second, NeRF assumes that the scene is static, meaning that the objects in the scene do not change position or appearance between images. Third, NeRF assumes that the light sources in the scene are known and fixed, meaning that they do not change position or intensity between images. Many following studies have been conducted to release such strict assumptions of NeRF (Lin et al., 2021; Martin-Brualla et al., 2021; Müller et al., 2022; Park et al., 2021; Q. Wang et al., 2021). The main focus of NeRF and related studies is to achieve photo-realistic rendering results. However, extracting surfaces from the predicted density is a challenging task, and finding an appropriate threshold is not straightforward. Consequently, the resulting geometry obtained is often unsatisfactory. Furthermore, sampling points along a ray for rendering pixel is done using a density function that is approximated from another network without any guarantee for correct approximation. Neural surface models define the surface implicitly by mapping a point in space to a signed distance function (SDF) that is implemented by a neural network (Park et al., 2019). The signed distance function represents the distance between a point in space and the surface of the object or scene. It is negative inside the surface and positive outside the surface, with zero indicating the surface itself. Several neural surface models have shown impressive rendering performance in 3D reconstruction (Niemeyer et al., 2020; Yariv et al., 2020). However, they require an appropriate network initialization since surface rendering techniques only provide gradient information locally where a surface intersects with a ray. Additional constraints such as pixel-wise mask supervision are necessary for converging to a valid surface. While obtaining accurate image masks is challenging especially in the agriculture scene, many studies that loosen the mask requirement have been conducted. UNISURF is a hybrid approach that gradually reduces the sampling region, encouraging a volumetric representation to converge to a surface (Oechsle et al., 2021; P. Wang et al., 2021; Yariv et al., 2021). Both VolSDF (Yariv et al., 2021) and NeuS (P. Wang et al., 2021) convert the SDF value into a density value and use regular volume rendering as in NeRF. VolSDF particularly represents the volume density as a transformed version of the SDF to the learned surface geometry, which provides a useful inductive bias allowing disentanglement of geometry (i.e., density) and radiance field. Furthermore, the transformed version of the SDF allows to bound the opacity approximation error leading to high fidelity sampling of the volume rendering integral.

Currently, neural representations have achieved remarkable success in general 3D object reconstruction in computer vision. However, these neural representations have been primarily developed and evaluated on standard datasets and benchmark tasks that do not necessarily reflect the challenges and complexities of real-world agricultural applications. Most existing computer vision datasets and benchmarks assume that images are taken in controlled environments with perfect lighting and unobstructed views of objects. While these assumptions make it easier to develop algorithms that perform well on the datasets, they don't reflect the realities of agricultural environments, where lighting conditions are often suboptimal, objects can be partially obscured by foliage or other objects, and cameras may have lower resolutions and poorer image quality. While there are more in-the-wild datasets being developed and recognized as well-regarded benchmarks, they do not necessarily cover diverse enough of agriculture examples: Many datasets focus on object recognition or segmentation in urban or indoor environments and may not include examples of agricultural objects or scenes.

The overall goal of this study is to investigate neural implicit representations to the 3D reconstruction of potted apple trees under controlled and field conditions. Specific objectives are to 1) collect multi-view image datasets in three scenes with a different level of background complexity using a consumer grade smartphone, 2) develop a framework for 3D reconstruction of potted apple trees using neural implicit representations with the focus on NeRF and VolSDF, and 3) compare the performance of neural implicit representations with classic multi-view 3D representation qualitatively and quantitatively.

Materials and Methods

Experimental Design and Data Acquisition

The data acquisition involved the use of a video recording taken by an individual using a mobile phone (i.e., iPhone 12 in this study), which captured footage of a potted apple tree. The potted apple tree was situated approximately 20 cm away from the camera and was recorded in a 360-degree fashion, thereby ensuring that all aspects of the tree were captured. Two potted apple trees with heavily pruned branches were studied to validate the 3D reconstruction framework in different scenes under controlled and field conditions. Each replicate was captured in three different scenes with different levels of background complexity by positioning the camera at varying angles in video recording. The camera was positioned downwards and laterally in a controlled condition to produce the images with clean and complex background. In the outdoor condition, the camera was positioned only laterally that leads to a complex background (Figure 1). The reference point cloud was collected by a 3D laser scanner (FARO Focus S350, FARO Technologies, USA) that was mounted on a tripod at approximately 80 cm from the ground to acquire colorized point cloud. The configuration of the laser scanner provides a resolution of 3.1 mm at 10 m. Two potted apple trees were scanned on two sides with one scan on each side, resulting in a max registration error of 1 mm.

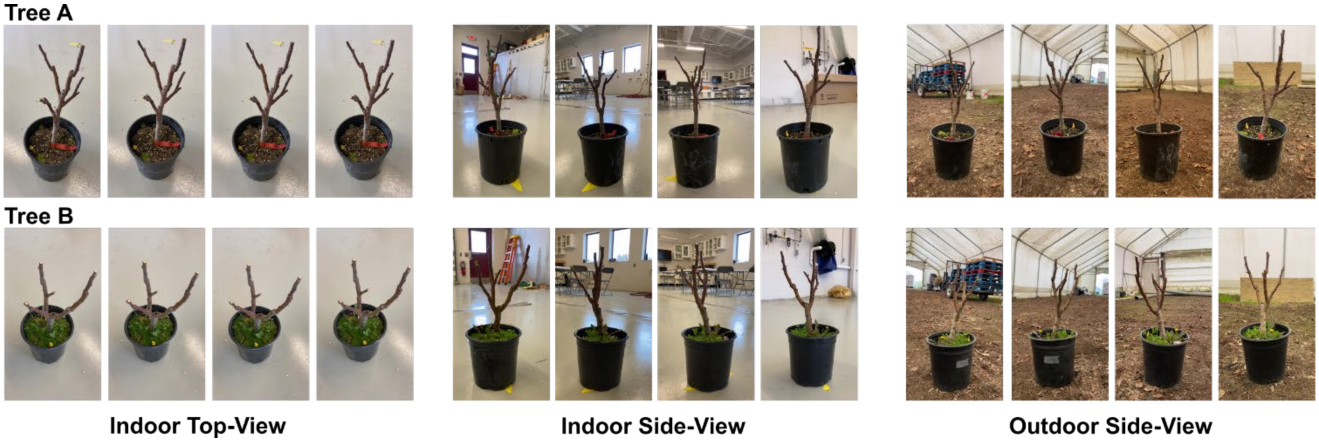

Figure 1. Representative 2D images of two replicates of potted apple trees used in this study from two different camera positions in both indoor and outdoor conditions.

3D Reconstruction Framework

Around 150 to 200 images were extracted from the video at a specified framerate using FFmpeg (Tomar, 2006). These multi-view images were then processed using the structure-from-motion (SfM) algorithm in COLMAP, which estimated the camera parameters and produced a sparse point cloud. The obtained camera parameters and multi-view images were used as input to train a neural radiance field (NeRF) model and a neural surface model (i.e., VolSDF). The well-trained models generated a high-quality 3D reconstruction of the potted apple tree, including realistic lighting and shading effects. To further improve the reconstruction quality, state-of-the-art Segment Anything Model (SAM) (Kirillov et al., 2023) can optionally be applied to multi-view images to generate image-wise foreground masks in a zero-shot setting. These masks can be integrated into the model training process as extra supervision. On the other hand, a dense point cloud was generated by the MVS algorithm in COLMAP for the comparison (Figure 2).

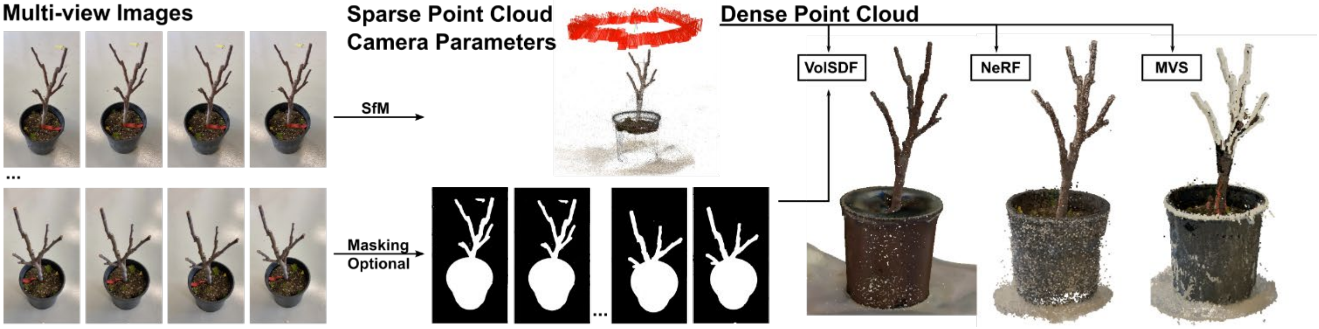

Figure 2. Workflow of the developed 3D reconstruction framework.

Classic Multi-view Reconstruction

COLMAP proposed a structure-from-motion (SfM) and multi-view stereo (MVS) pipeline that uses a multi-stage process to generate a dense 3D reconstruction from a collection of 2D images (Schonberger & Frahm, 2016). Initially, camera parameters are estimated using a feature-based matching algorithm and are optimized using a bundle adjustment algorithm. A sparse 3D point cloud is generated by triangulating the camera positions and the corresponding feature points. To generate a dense point cloud, the stereo-matching algorithm was used to produce a depth map that is then fused together to create a consistent and dense representation of the 3D space. The dense point cloud was further refined by the MVS algorithm that considers the visibility of each point in the scene from multiple viewpoints.

Neural Radiance Field

Neural Radiance Fields (NeRF) works by modeling the volumetric density and radiance of a 3D scene using a deep neural network (Figure 3). A neural radiance filed is a continuous mapping from a 3D location $x \in \mathbb{R}^3$ and a ray viewing direction $d \in \mathbb{R}^2$ to an RGB color $c \in [0,1]^3$ and volume density $\sigma \in \mathbb{R}^+$. The volume density reflects the rate that light is occluded at point x. Conditioning on the viewing direction d allows for modeling view-dependent effects such as specular reflections and improves reconstruction quality in case the Lambertian assumption is violated (Yariv et al., 2020). It can be formulated as,

$$[c,\sigma] = \mathrm{N}_\theta(\gamma_x(x), \gamma_d(d)) \tag{1}$$

where N is modeled as an MLP with learnable parameters θ, and $\gamma: \mathbb{R}^3 \rightarrow \mathbb{R}^N$ is a positional encoding of the input x required to capture high frequencies.

Given a camera pose $P = [R, T]$, each pixel coordinate $p \in \mathbb{R}^2$ determines a ray in the world coordinate system, whose origin is the camera center of projection $o = T$ and direction is defined as $d = RK^{-1}\bar{p}$ where K is the camera intrinsic and $\bar{p}$ is the homogeneous representation of p. A 3D point along the viewing ray associated with p can be expressed as $r(t) = o + td$. To render the color $C(r) \in [0,1]^3$ at pixel p, M discrete depth values t_m were sampled along the ray within the near and far plane $[t_n, t_f]$, and query N_θ at the associated 3D points. The corresponding predicted color and volume density values were then rendered from classic volume rendering as,

$$C(r) = \int_{t_n}^{t_f} T(t)\sigma\big(r(t)\big)c\big(r(t), d\big)dt, \text{ where } T(t) = \exp(-\int_{t_n}^{t} \sigma\big(r(s)\big)ds) \tag{2}$$

where $T(t)$ is the transparency function of the volume indicating the probability of a light particle succeeds traversing the segment $[o, r(t)]$ without bouncing off. Equation 7 was approximated using a numerical quadrature as,

$$C(r) = \sum_{i=1}^{N} T_i(1 - \exp(-\sigma_i\delta_i))c_i, \text{ where } T_i = \exp(-\sum_{j=1}^{i-1}\sigma_j\delta_j) \tag{3}$$

where $\delta_i = t_{i+1} - t_i$ is the distance between adjacent samples.

The loss function used in NeRF is simply the total squared error between the rendered and true pixel color for both the coarse and fine renderings:

$$\mathcal{L}_{RGB} = \sum_{r\in\mathcal{R}} (||C_c(r) - C(r)|| + ||C_f(r) - C(r)|| \tag{4}$$

where $||.||$ is the standard L2-norm and $\mathcal{R}$ is a set of rays, and $C_c(r)$, $C_f(r)$, and $C(r)$ are the RGB colors of ground truth, coarse volume prediction, and fine volume prediction for ray r , respectively.

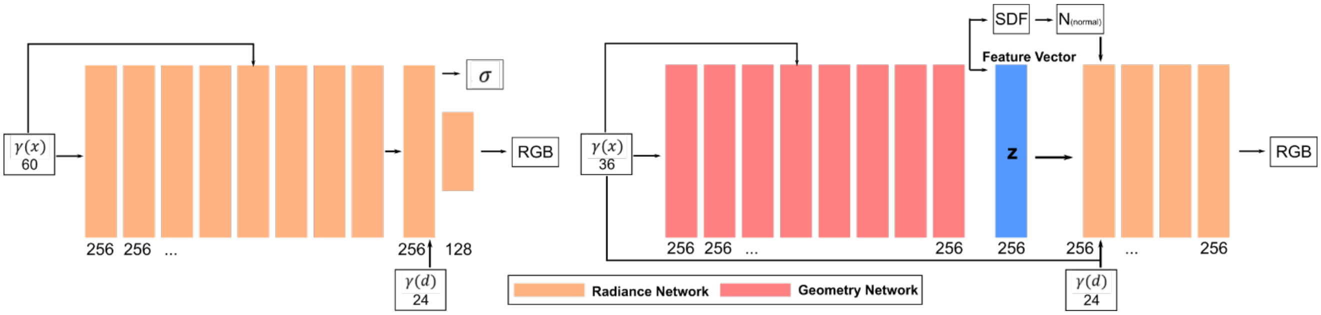

Figure 3. The architecture of NeRF (left) and VolSDF (right).

Neural Implicit Surface Model

Let the set $\Omega \subset \mathbb{R}^3$ represent the space occupied by some object in $\mathbb{R}^3$, and $\mathcal{M} = \partial\Omega$ its boundary surface. As a reminder, an SDF is a continuous function that, for a given spatial point, outputs the point's distance to the closet surface, whose sign encodes whether the point is inside (negative) or outside (positive) of the surface:

$$SDF(x) = s\!: x \in \mathbb{R}^3, s \in \mathbb{R}. \tag{5}$$

The underlying surface is implicitly represented by the iso-surface of $SDF(\mathcal{M}) = 0$. In other denotation, 1_Ω stands for the Ω indicator function and d_Ω for the SDF as follows,

$$1_\Omega = \begin{cases} 1, x \in \Omega \\ 0, otherwise \end{cases}, \text{ and } d_\Omega(x) = (-1)^{1_\Omega(x)} \min_{y\in\mathcal{M}} ||x - y|| \tag{6}$$

The proposed novel parameterization for volume density σ defined as transformed signed distance function is formulized as:

$$\sigma(x) = \alpha\Psi_\beta(-d_\Omega(x)) \tag{7}$$

where $\alpha, \beta > 0$ are learnable parameters, and Ψ_β is the Cumulative Distribution Function (CDF) of the Laplace distribution with zero mean and β scale (i.e., mean absolute deviation),

$$\Psi_\beta = \begin{cases} \dfrac{1}{2}\exp\left(\dfrac{s}{\beta}\right), s \le 0 \\ 1 - \dfrac{1}{2}\exp\left(-\dfrac{s}{\beta}\right), s > 0 \end{cases} \tag{8}$$

As can be readily checked from this definition, as β approach zero, the density σ converges to a scaled indicator function of Ω, that is $\sigma \to \alpha 1_\Omega$ for all points $x \in \Omega \backslash \mathcal{M}$.

The volume rendering depends on the level-set's normal (i.e., $n(t) = \nabla_r d_\Omega(r(t))$), which was motivated by the fact that Bidirectional Reflectance Distribution Functions (BRDFs) of common materials are often encoded with respect to the surface normal, facilitating disentanglement as done in surface rendering (Yariv et al., 2020).

$$C(r) = \int_0^\infty c(r(t), n(t), d)\tau(t)dt \tag{9}$$

where $c(r(t), n(t), d)$ is the radiance field and $\tau(t)$ is the Probability Density Function (PDF) of the opacity function $O(t)$ of the volume along the ray defined as:

$$\tau(t) = \frac{dO}{dt}(t) = \sigma\big(r(t)\big)T(t), \text{ where } O(t) = 1 - T(t) \tag{10}$$

and the integral in equation 8 was approximated similarly using a numerical quadrature as,

$$C(r) = \sum_{i=1}^{M-1} \tau(t_i)\delta_i c_i \tag{11}$$

Similar to the loss function in NeRF, VolSDF training loss consists of two terms: the RGB loss term and the Eikonal loss term that regularizes the geometry of the model (Gropp et al., 2020) as follows:

$$\mathcal{L} = \mathcal{L}_{RGB} + \lambda\mathcal{L}_{SDF}, \text{ where } \mathcal{L}_{SDF} = (||\nabla d_{\Omega}(z)|| - 1)^2 \tag{12}$$

rather than the standard L2-norm, L1-norm was used in equation 4. Notice there is only one network producing the pixel color in the $\mathcal{L}_{RGB}$ term in equation 12.

The beauty of density σ is that it facilitates a bound on the error of the transparency of the rendered volume, which is hard to devise for a generic MLP densities. Additionally, the error bound was mathematically proved to be bounded by ϵ that is a hyper-parameter with proper α and β that could be learned during training.

Implementation Details

NeRF and VolSDF are essentially multiple MLPs (Figure 3). NeRF has 8 layers of MLPs with ReLU activations and 256 channels per layer, and one additional layer outputs σ and a 256-dimensional feature vector. This feature vector is then concatenated with the camera ray's viewing direction and passed to one additional MLP with a ReLU activation and 128 channels) that output the view-dependent RGB color. VolSDF consists of two MLP networks, where the geometry network is identical to the first 8 layers of MLPs in NeRF. The geometry network outputs the SDF value and an extra feature vector of size 256. The geometry network was initialized as a unit sphere (Atzmon & Lipman, 2020). The radiance network has 4 layers of MLPs of width 256 and uses a Sigmoid activation to provide RGB color. The software packages employed for running NeRF and VolSDF are nerfstudio and sdfstudio. The default configuration of NeRF and VolSDF with the explicit specification of the near (i.e., 0.02 m) and far plane (i.e., 4 m) was used to train all models in this study. Both NeRF and VolSDF were trained for 100k iterations on a single NVIDIA A6000 GPU, with a typical training time of 4 hours for NeRF and 15 hours for VolSDF.

Performance Evaluation

2D Image Radiance Evaluation

The radiance prediction performance (i.e., novel view synthesis) of NeRF and VolSDF model was quantitatively evaluated using common metrics, including Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS). PSNR is a widely used metric that measures the quality of a reconstructed image by calculating the mean squared error (MSE) between the original and reconstructed images. SSIM, on the other hand, is a perceptual metric that compares the structural similarities between two images, taking into account features such as luminance, contrast, and structure. LPIPS is a recently proposed metric that uses a deep neural network to measure the similarity between two images and has been shown to better correlate with human perception than traditional metrics like PSNR and SSIM. Notice for PSNR and SSIM, a higher value suggests a better performance, but a lower LPIPS indicates a better performance.

3D Reconstruction Evaluation

The evaluation of 3D reconstruction performance in NeRF, VolSDF, and MVS in COLMAP involves both qualitative and quantitative assessment. Qualitative evaluation is conducted by visually assessing the reconstructed 3D point cloud in terms of the noise level, detail recovery, and structure completeness. This evaluation can help identify any visual artifacts, inconsistencies, or inaccuracies in the reconstructed models. The quantitative evaluation, on the other hand, involves calculating the chamfer distance between the reconstructed 3D models and the pseudo ground truth data. The chamfer distance is a widely used metric in 3D reconstruction evaluation and measures the distance between two sets of points. A lower chamfer distance indicates a higher quality reconstruction. The combination of qualitative and quantitative evaluation helps provide a comprehensive assessment of the 3D reconstruction performance of NeRF, VolSDF, and MVS in COLMAP.

Results

Neural Implicit Model Training and Evaluation

Both NeRF and VolSDF models achieved convergence during training and produced high-quality RGB images during evaluation, indicating the potential of neural implicit representations to learn both the geometry and appearance of the potted apple tree and its background (Figure 4 and 5). NeRF learned the radiance of the pot at the beginning of training and gradually captured the radiance of the branch from iteration 2500. As training progressed, NeRF learned radiance of finer-grained objects such as the color sticker and weeds in the pot due to the strong capacity of its MLP (Figure 3). VolSDF showed a similar pattern by first learning the pot and then capturing more details during the later stages of training. Qualitatively, the fine RGB images produced by NeRF contained sharper foreground details than those by VolSDF, indicating NeRF outperformed VolSDF in learning radiance of foreground objects. This is due to the specific MLP structure of NeRF, which encodes an implicit prior favoring a smooth surface reflectance function by treating scene position and viewing direction asymmetrically (Zhang et al., 2020). Furthermore, NeRF utilized a higher level of positional encoding (PE) for its scene position (i.e., 10) than VolSDF (i.e., 6), which improves details in rendering. On the other side, VolSDF uses the Eikonal loss that encourages d_Ω to approximate an SDF that enforces a more consistent radiance prediction at any point on the learned surface (Gropp et al., 2020). However, VolSDF presented a higher PSNR and SSIM value and a lower LPIPS value in evaluation than NeRF, implying VolSDF essentially had a better understanding of the radiance of the entire scene (i.e., foreground + background). NeRF had problems modeling the background when the scene was unbounded because of insufficient sampling in a Euclidean parameterization of 3D space (Zhang et al., 2020). Inverted sphere parameterization was proposed to address this issue by partitioning the scene space into two volumes: an inner volume for the foreground and an outer volume for the background and two separate NeRFs were used to model these two volumes. VolSDF used a predefined scene bounding sphere and modeled the background outside the sphere using the inverted sphere parameterization presented in NeRF++, resulting in higher resolution in the background.

In contrast to the process of explicit learning the radiance field, NeRF employed a simple thresholding strategy to produce depth maps. The depth of a single pixel was determined by thresholding the volume density of that pixel at a pre-defined value, rather than being explicitly learned by NeRF. This approach could result in ambiguity in the generated depth maps because NeRF could potentially hallucinate the depth at any distance along the camera ray as long as the accumulated radiance was correct. Consequently, the geometry of the scene was not guaranteed to be reasonable in NeRF since it lacked corresponding constraints. While this phenomenon was not evident in generated depth maps, it could be observed much more easily in the 3D reconstruction (Figure 6). In contrast, VolSDF explicitly learned an SDF function to represent the surface of the object and further integrated this SDF function in volume rendering, resulting in significantly more reliable normal and depth estimation. Moreover, since VolSDF disentangles the geometry and radiance, the geometry MLP parameters can be effectively and efficiently optimized, providing much cleaner and smoother depth maps compared with the depth maps generated by NeRF. Especially, NeRF's depth maps have particular inconsistency depth values around the boundaries of objects, which was reflected by the jagged shape and non-smooth color transition (Figure 5 Black Box) from foreground pixels to background pixels.

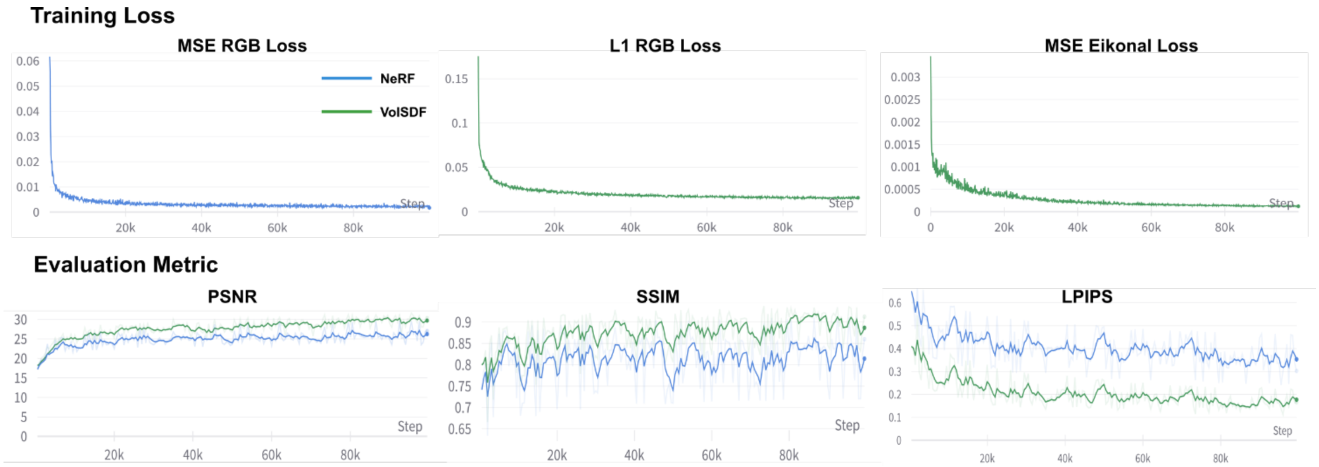

Figure 4. Training and evaluation curves of NeRF and VolSDF trained on the side-view images under the indoor condition. Notice curves were smoothed with a factor of 0.5 for better comparison.

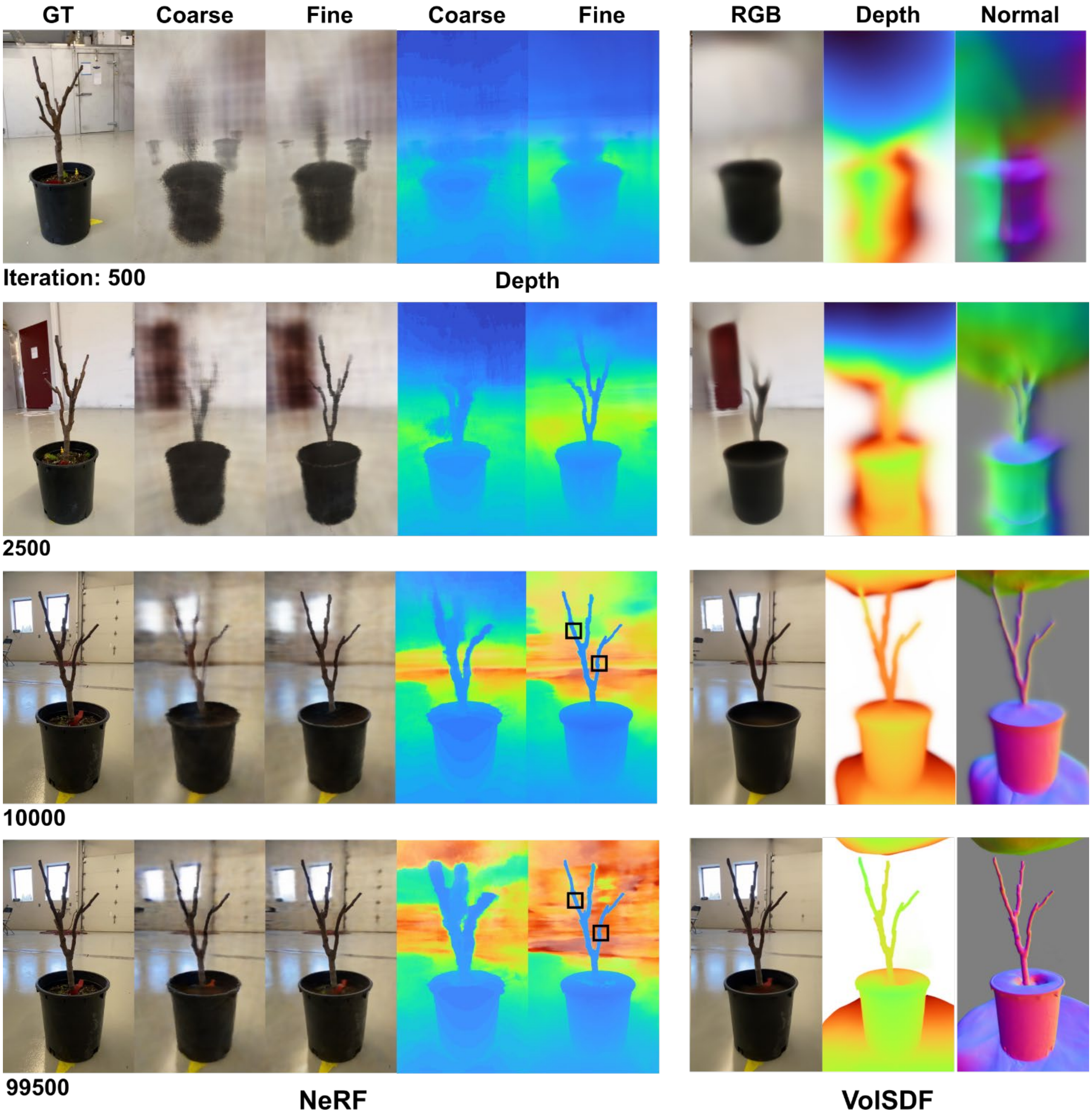

Figure 5. RGB images and depth maps generated by the NeRF and VolSDF models trained on the side-view images under the indoor condition in different iterations in evaluation. Notice the colormap of depth maps between NeRF and VolSDF is not identical due to the configuration of logging functions, so it is only reasonable to compare the value within a model.

Comparison of 3D Reconstruction

NeRF and VolSDF generated higher quality 3D point clouds from side-view images compared to MVS, demonstrating the superiority of neural implicit representation methods over classic multi-view reconstruction methods (Figure 6). While the 3D reconstruction from top-view images by VolSDF had some artifacts in the surrounding environment, the tree itself was well reconstructed. A more comprehensive analysis of these artifacts was provided in the next section. The improvement in overall reconstruction quality is attributed to the stronger expressivity and larger capacity of MLP in NeRF and VolSDF, which can capture more detailed information about the scene, such as surface textures and lighting effects. In contrast, MVS relies on image matching and triangulation, introducing errors and noise into the reconstructed point cloud, especially in complex geometries (Figure 6 B). The MVS-reconstructed point cloud has decent accuracy on the stem of the potted apple tree but presented considerable amount of noise (i.e., white points) on branches because branches are thinner and provide fewer corresponding features to match. While the NeRF-reconstructed point cloud contained a much smaller amount of noise, the radiance of points is inconsistent around the surface, and the surface itself is significantly rough and incomplete where points are not evenly scattered and there are gaps and holes. The inconsistent radiance of points was caused by the ambiguity of the volume density in NeRF. While NeRF's overall optimization objective is to minimize the accumulated radiance of pixels, it lacks enough regularization of radiance of single sample point, leading to insufficient constraint to guarantee the correct radiance of points in the generated point cloud. The roughness and incompleteness of the surface were a consequence of aforementioned missing geometry learning capabilities in NeRF. NeRF cannot learn any information related to the level set information of the object so NeRF's MLP does not encode the surface accordingly. In contrast, VolSDF demonstrated the best performance on the mesh and point cloud generation due to its capability of learning surface explicitly (Figure 7). The other advantage of VolSDF is its flexibility in managing the resolution of mesh and point cloud. However, VolSDF needs more images and more careful parameter tuning to capture the most fine-grained details such as the weeds in the pot.

The reconstructed point clouds by NeRF and VolSDF were found to be completer and more reliable than the reference point cloud by FARO, indicating that neural implicit representations have advantages in 3D reconstruction of complex plant architectures, especially when active 3D reconstruction methods such as laser scanning are not able to provide satisfactory results (Figure 6 and S1). The structure of the reference point cloud is significantly incomplete with many holes and ghost points due to the insufficient corresponding matching points in the point cloud registration. With the strong expressivity of neural implicit representations, they can potentially offer more opportunities for in-field tree reconstruction.

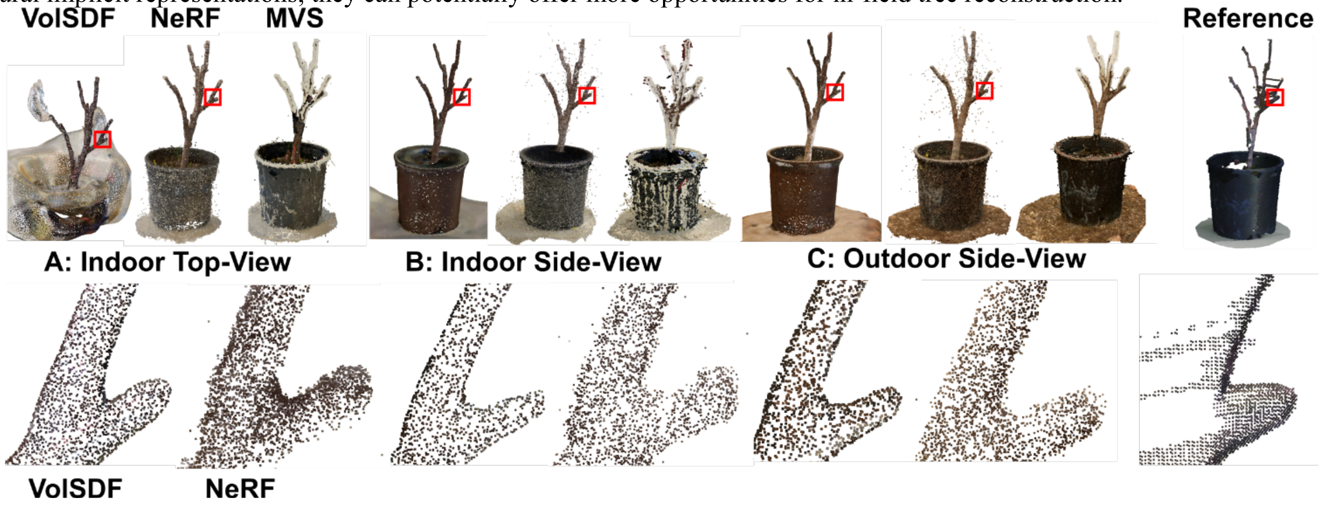

Figure 6. 3D reconstructed point cloud by COLMAP, NeRF, and VolSDF of Tree A and the reference point cloud

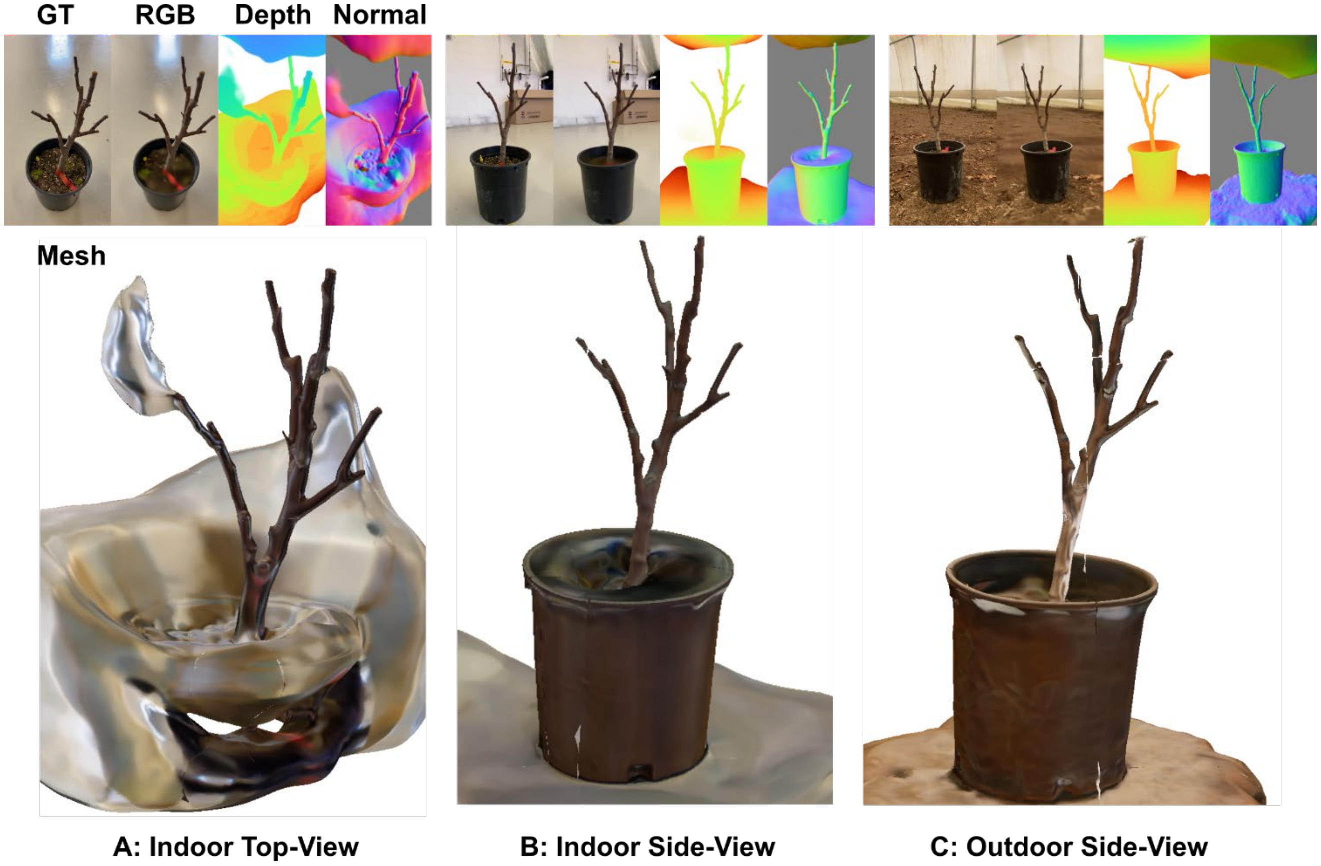

Figure 7. 2D RGB images, depth, and normal and 3D reconstructed mesh from VolSDF.

Scene Dependent Performance

The performance of NeRF and VolSDF model was scene dependent as expected, suggesting the certain constraints of neural implicit models on 3D reconstruction. For indoor environments, when the camera is positioned downwards, the background is typically ground, which presents a large contrast pixel distribution with the foreground potted apple tree, unlike the background when the camera is positioned upfront (Figure 1). Outdoor environments present more challenges due to the soil's color being similar to the branch (Figure 1). As a result, the clean background and high contrast pixel distribution are easier to learn for NeRF resulting in better radiance field reconstruction (Figure 8 Top). Consequently, the reconstructed point cloud using indoor top-view images from NeRF presented the best quality because of the benefits of the simple scene. Conversely, VolSDF achieved a better performance from the side-view images rather than the top-view images in indoor condition (Figure 8 Bottom). This is not because VolSDF does not favor a clean background and high contrast pixel distribution, rather, VolSDF was more likely to be affected by the specularity reflected from the ground when the camera is positioned downwards. The reason for this specularity-sensitivity is VolSDF learns the surface information and relies on the normal estimation to predict the radiance. Specularities are regions where the surface reflects light in a highly directional way, resulting in sudden changes in the surface normal. These sudden changes can cause large gradients in the radiance field, making the optimization process difficult to converge, thereby producing inaccurate estimated surface normal and leading to artifacts in the appearance (Figure 7 Top). The artifacts can be observed more easily from the learned geometry – depth and normal. Additionally, the downward camera position restricts the camera's field of view (FoV), which affects the quality of the 3D reconstruction for VolSDF (Figure 6 A and Figure 7 A). The quality of the reconstructed point cloud was significantly degraded due to the confounding effect of abnormal lighting and the limited FoV. This is because geometry is inherently harder to learn than appearance. While a radiance network can produce correct pixel radiance with simple loss regularization, optimizing geometry requires more robust supervision, such as additional images or pixel-wise masks.

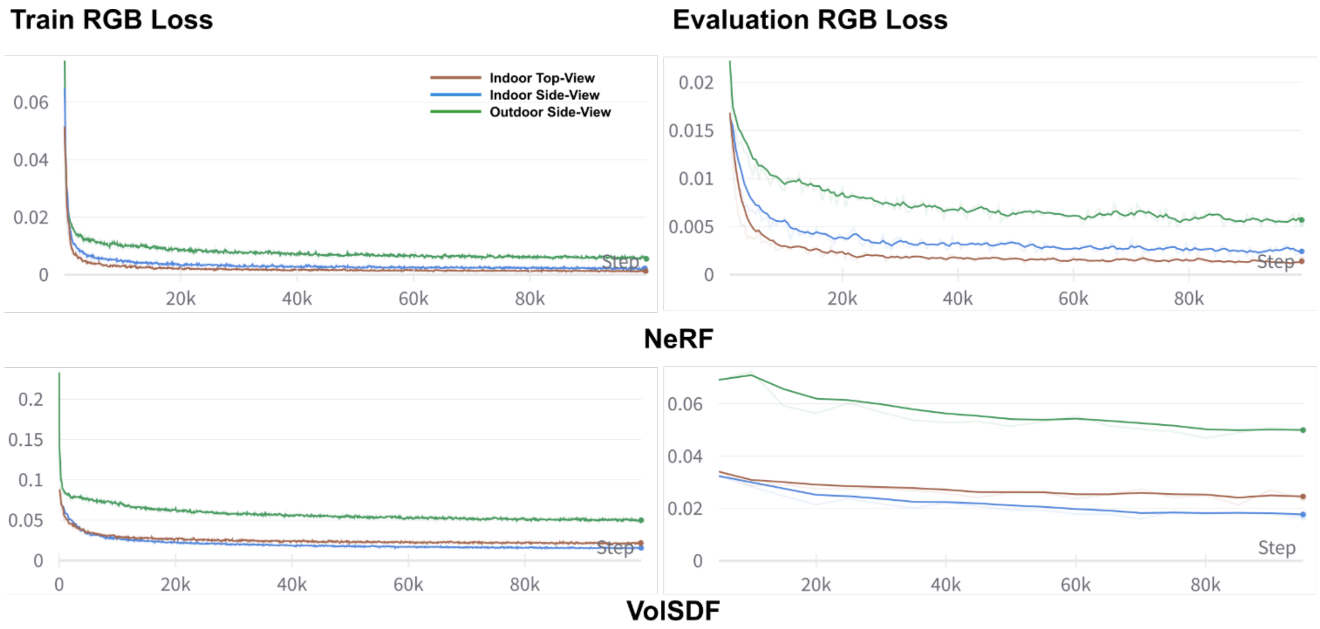

Figure 8. Cross scene radiance prediction performance comparison of NeRF.

Conclusion

In this study, the use of neural implicit representations for 3D reconstruction of potted apple trees was investigated for morphological characterization in indoor and outdoor conditions. The comparison between classic multi-view algorithms and neural implicit representation methods was conducted using two replicates of a heavily pruned potted apple tree, captured through 2D images and point cloud data. The point cloud was used as the pseudo ground truth for evaluating the correctness and structure completeness of the reconstructed point cloud data generated by both methods. The results demonstrated that neural implicit representations outperformed classic multi-view algorithms for 3D reconstruction in both indoor and outdoor conditions. The high-resolution reconstructed point cloud can be effectively used for tree architecture characterization. This study highlights the potential of neural implicit representations in 3D reconstruction for precision agriculture applications.

Acknowledgement

The present study was supported by the USDA NIFA Hatch project (accession No. 1025032) and USDA NIFA Specialty Crop Research Initiative (award No. 2020-51181-32197).

Reference

Atzmon, M., & Lipman, Y. (2020). Sal: Sign agnostic learning of shapes from raw data. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,

Furukawa, Y., & Ponce, J. (2009). Accurate, dense, and robust multiview stereopsis. *IEEE transactions on pattern analysis and machine intelligence*, *32*(8), 1362-1376.

Galliani, S., Lasinger, K., & Schindler, K. (2015). Massively parallel multiview stereopsis by surface normal diffusion. Proceedings of the IEEE International Conference on Computer Vision,

Ge, L., Zou, K., Zhou, H., Yu, X., Tan, Y., Zhang, C., & Li, W. (2021). Three dimensional apple tree organs classification and yield estimation algorithm based on multi-features fusion and support vector machine. *Information Processing in Agriculture*.

Gené-Mola, J., Sanz-Cortiella, R., Rosell-Polo, J. R., Morros, J.-R., Ruiz-Hidalgo, J., Vilaplana, V., & Gregorio, E. (2020). Fruit detection and 3D location using instance segmentation neural networks and structure-from-motion photogrammetry. *Computers and electronics in agriculture*, *169*, 105165.

Genova, K., Cole, F., Sud, A., Sarna, A., & Funkhouser, T. (2020). Local deep implicit functions for 3d shape. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,

Gropp, A., Yariv, L., Haim, N., Atzmon, M., & Lipman, Y. (2020). Implicit geometric regularization for learning shapes. *arXiv preprint arXiv:2002.10099*.

Jiang, C., Sud, A., Makadia, A., Huang, J., Nießner, M., & Funkhouser, T. (2020). Local implicit grid representations for 3d scenes. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,

Jiang, Y., & Li, C. (2020). Convolutional neural networks for image-based high-throughput plant phenotyping: a review. *Plant Phenomics*, *2020*.

Jin, S., Su, Y., Wu, F., Pang, S., Gao, S., Hu, T., Liu, J., & Guo, Q. (2018). Stem–leaf segmentation and phenotypic trait extraction of individual maize using terrestrial LiDAR data. *IEEE Transactions on Geoscience and Remote Sensing*, *57*(3), 1336-1346.

Kar, A., Häne, C., & Malik, J. (2017). Learning a multi-view stereo machine. *Advances in neural information processing systems*, *30*.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., & Lo, W.-Y. (2023). Segment anything. *arXiv preprint arXiv:2304.02643*.

Lin, C.-H., Ma, W.-C., Torralba, A., & Lucey, S. (2021). Barf: Bundle-adjusting neural radiance fields. Proceedings of the IEEE/CVF International Conference on Computer Vision,

Luo, L., Jiang, X., Yang, Y., Samy, E. R. A., Lefsrud, M., Hoyos-Villegas, V., & Sun, S. (2022). Eff-3DPSeg: 3D organ-level plant shoot segmentation using annotation-efficient point clouds. *arXiv preprint arXiv:2212.10263*.

Ma, Z., Teed, Z., & Deng, J. (2022). Multiview stereo with cascaded epipolar raft. Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXI,

Martin-Brualla, R., Radwan, N., Sajjadi, M. S., Barron, J. T., Dosovitskiy, A., & Duckworth, D. (2021). Nerf in the wild: Neural radiance fields for unconstrained photo collections. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,

Max, N. (1995). Optical models for direct volume rendering. *IEEE Transactions on Visualization and Computer Graphics*, *1*(2), 99-108.

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., & Ng, R. (2021). Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, *65*(1), 99-106.

Müller, T., Evans, A., Schied, C., & Keller, A. (2022). Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)*, *41*(4), 1-15.

Niemeyer, M., Mescheder, L., Oechsle, M., & Geiger, A. (2020). Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,

Oechsle, M., Peng, S., & Geiger, A. (2021). Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. Proceedings of the IEEE/CVF International Conference on Computer Vision,

Park, J. J., Florence, P., Straub, J., Newcombe, R., & Lovegrove, S. (2019). Deepsdf: Learning continuous signed distance functions for shape representation. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition,

Park, K., Sinha, U., Hedman, P., Barron, J. T., Bouaziz, S., Goldman, D. B., Martin-Brualla, R., & Seitz, S. M. (2021). Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *arXiv preprint arXiv:2106.13228*.

Qiu, T., Cheng, L., & Jiang, Y. (2022). 3D characterization of tree architecture for apple crop load estimation. 2022 ASABE Annual International Meeting,

Schonberger, J. L., & Frahm, J.-M. (2016). Structure-from-motion revisited. Proceedings of the IEEE conference on computer vision and pattern recognition,

Sun, S., Li, C., Chee, P. W., Paterson, A. H., Jiang, Y., Xu, R., Robertson, J. S., Adhikari, J., & Shehzad, T. (2020). Three-dimensional photogrammetric mapping of cotton bolls in situ based on point cloud segmentation and clustering. *ISPRS Journal of Photogrammetry and Remote Sensing*, *160*, 195-207.

Sun, S., Li, C., Chee, P. W., Paterson, A. H., Meng, C., Zhang, J., Ma, P., Robertson, J. S., & Adhikari, J. (2021). High resolution 3D terrestrial LiDAR for cotton plant main stalk and node detection. *Computers and electronics in agriculture*, *187*, 106276.

Tomar, S. (2006). Converting video formats with FFmpeg. *Linux journal*, *2006*(146), 10.

USApple. (2022). *Industry Outlook 2022*. https://usapple.org/wp-content/uploads/2022/08/USAPPLE-INDUSTRYOUTLOOK-2022.pdf

Wang, P., Liu, L., Liu, Y., Theobalt, C., Komura, T., & Wang, W. (2021). Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*.

Wang, Q., Wang, Z., Genova, K., Srinivasan, P. P., Zhou, H., Barron, J. T., Martin-Brualla, R., Snavely, N., & Funkhouser, T. (2021). Ibrnet: Learning multi-view image-based rendering. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,

Westling, F., Underwood, J., & Bryson, M. (2021). Graph-based methods for analyzing orchard tree structure using noisy point cloud data. *Computers and electronics in agriculture*, *187*, 106270.

Yao, Y., Luo, Z., Li, S., Fang, T., & Quan, L. (2018). Mvsnet: Depth inference for unstructured multi-view stereo.

Proceedings of the European conference on computer vision (ECCV),

Yao, Y., Luo, Z., Li, S., Shen, T., Fang, T., & Quan, L. (2019). Recurrent mvsnet for high-resolution multi-view stereo depth inference. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition,

Yariv, L., Gu, J., Kasten, Y., & Lipman, Y. (2021). Volume rendering of neural implicit surfaces. *Advances in neural information processing systems*, *34*, 4805-4815.

Yariv, L., Kasten, Y., Moran, D., Galun, M., Atzmon, M., Ronen, B., & Lipman, Y. (2020). Multiview neural surface reconstruction by disentangling geometry and appearance. *Advances in neural information processing systems*, *33*, 2492-2502.

Zhang, K., Riegler, G., Snavely, N., & Koltun, V. (2020). Nerf++: Analyzing and improving neural radiance fields. *arXiv preprint arXiv:2010.07492*.

Supplementary

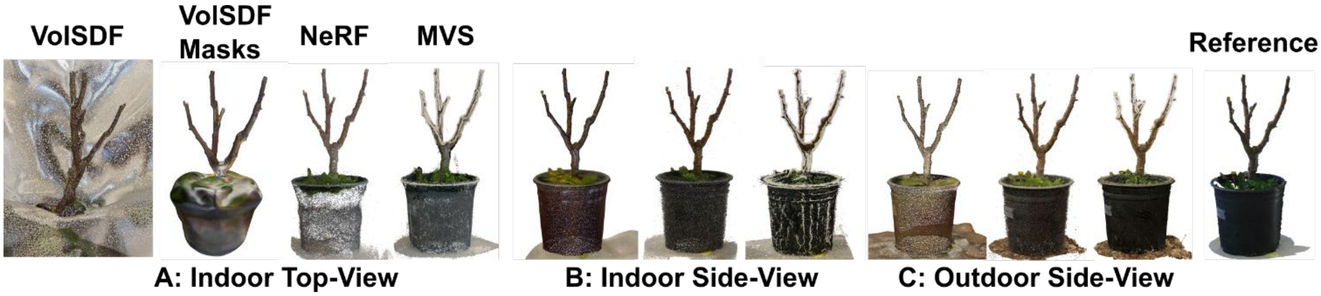

Figure S1. 3D reconstructed point cloud by the MVS algorithm in COLMAP, NeRF, and VolSDF of Tree B.