Computers and Electronics in Agriculture 220 (2024) 108848

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

Original papers

High-fidelity 3D reconstruction of plants using Neural Radiance Fields

Kewei Hu [a], Wei Ying [a], Yaoqiang Pan [a], Hanwen Kang [a,*], Chao Chen [b,*]

[a] College of Engineering, South China Agriculture University, China
[b] Department of Aerospace and Mechanical, Monash University, Australia

ARTICLE INFO

Dataset link: https://pan.baidu.com/s/1lBMlC8FFVrwOC7VRx4j9cw?pwd=39tc

Keywords:
Deep-learning
Robotics
NeRF
Phenotyping

ABSTRACT

Accurate reconstruction of plant phenotypes plays a key role in optimizing sustainable farming practices in the field of Precision Agriculture (PA). Currently, optical sensor-based approaches dominate the field, but the need for high-fidelity 3D reconstruction of crops and plants in unstructured agricultural environments remains challenging. Recently, a promising development has emerged in the form of Neural Radiance Fields (NeRF), a novel method that utilizes neural density fields. This technology has shown impressive performance in various novel vision synthesis tasks, but has remained relatively unexplored in the agricultural context. In our study, we focus on two fundamental tasks within plant phenotyping: (1) the synthesis of 2D novel-view images and (2) the 3D reconstruction of crop and plant models. We explore the world of NeRF, in particular two state-of-the-art (SOTA) methods: Instant-NGP, which excels in generating high-quality images with impressive training and inference speed, and Instant-NSR, which improves the reconstructed geometry by incorporating the Signed Distance Function (SDF) during training. In particular, we present a novel plant phenotype dataset comprising real plant images from production environments. This dataset is a first-of-its-kind initiative aimed at comprehensively exploring the advantages and limitations of NeRF in agricultural contexts. Our experimental results show that NeRF demonstrates commendable performance in the synthesis of novel-view images and is able to achieve reconstruction results that are competitive with Reality Capture, a leading commercial software for 3D Multi-View Stereo (MVS)-based reconstruction. Moreover, our study also highlights certain drawbacks of NeRF, including relatively slow training speeds, performance limitations in cases of insufficient sampling, and challenges in obtaining geometry quality in complex setups. In conclusion, NeRF introduces a new paradigm in plant phenotyping, providing a powerful tool capable of generating multiple representations, such as multi-view images, point cloud and mesh, from a single process.

1. Introduction

In recent years, integration of emerging sensors and Artificial Intelligence (AI) has revolutionized PA, significantly enhancing the efficiency, effectiveness, and productivity of breeding and primary production in agriculture industry (Sishodia et al., 2020). Unpredictable threats such as climate, soil characteristics, insect pests, and etc. are the main challenges to maintaining and guaranteeing crop yields (Fu et al., 2020). This has given rise to the increasing importance of monitoring plant growth through the comprehensive analysis of plant phenotyping (Feng et al., 2021). Phenomics studies a variety of phenotypic plant traits, such as growth, tolerance, yield, plant height, leaf area index, and etc. (Asaari et al., 2019). Traditional methods for manual phenotypic measurement and analysis were labor-intensive, time-consuming, and often destructive (Li et al., 2020; Feng et al., 2021; Furbank and Tester, 2011). Thus, emerging sensor technologies have been widely used to achieve non-invasive and high-throughput plant phenotyping (Rebetzke et al., 2019). The most current research shows that optical sensors dominate the detection system (Wang et al., 2023; Zhou et al., 2022) and various types of two-dimensional (2D) and three-dimensional (3D) imaging systems can directly measure morphological traits of plants, including colors of seeds, leaves, canopies, fruits, and roots, shapes and sizes of seeds, sizes, numbers, areas, textures, angles, architectures, and total volumes of canopies, leaves, and roots, and volumes sizes, shapes, numbers, and spatial distributions of fruits (Zhang and Zhang, 2018).

Despite the fact that 2D imaging systems deploy red, green, blue (RGB) camera to measure morphological traits (color, shape, size, and texture) of plants at affordable prices, these methods are limited by the dimensionality of the data and therefore cannot express the geometric form of the plant (Zhang and Zhang, 2018). 3D imaging systems enables tracking exact geometry and measurement of plant traits like plant height, plant width, root volume, root surface area, leaf size, leaf width,

* Corresponding authors.
E-mail addresses: hanwen.kang@outlook.com (H. Kang), Chao.Chen@monash.edu (C. Chen).

https://doi.org/10.1016/j.compag.2024.108848
Received 9 November 2023; Received in revised form 18 January 2024; Accepted 15 March 2024
Available online 27 March 2024

Fig. 1. High-fidelity 2D imaging (a) and 3D imaging (b) plant phenotypes from NeRF.

stem angle and projected canopy area. As a result, the 3D imaging system can keep track of the actual growth status of the plant at the organ level, which is almost impossible for the 2D imaging system (Zhang and Zhang, 2018; Feng et al., 2021). However, current phenotyping methods still encounter the following challenges: (1) Traditional multi-view RGB image phenotyping methods cannot generate novel views from the acquired data, resulting in time-consuming data collection and the potential omission of important details. (2) And MVS-based 3D reconstruction methods involve numerous individual steps, resulting in a lengthy process and reduced overall robustness. (3) Finally, point cloud-based approaches that rely on depth sensors may encounter problems such as bleeding points, especially at the boundaries between the foreground and background, leading to the expansion of foreground objects and significant errors in the point cloud at the edges.

Recently, an emerging deep learning rendering method, NeRF (Mildenhall et al., 2021), has shown superior performance in 2D image rendering. Its underlying mechanism can also learn the 3D geometry information from the rendering, which opens a new path for in-situ phenotyping in agricultural applications and brings significant advantages. For example, it requires only 2D images and is capable of generating high-quality images of novel views that are not merely interpolated, but are true inferences of the underlying scene geometry. This capability is valuable for plant phenotyping, where it is neither feasible nor efficient to manually capture all possible views of a plant or crop. Besides, although the main function of NeRF is implicit scene representation and view synthesis, its density information is stored in the Multi-Layer Perceptron (MLP), which provides an important insight and basis for the extraction of geometry. Based on these considerations, research on NeRF will be a key bridge between 2D imaging and 3D imaging, two types of plant phenotyping acquisitions (as shown in Fig. 1), in order to establish a low-cost, high-throughput, non-invasive plant phenotyping system.

This study explores the in-situ high-fidelity phenotyping by NeRF methods. Our study utilizes the volume rendering technology to render novel view images from the Neural Radiation Field parameters of the scene, and combined with the Marching cube algorithm which can extract the explicit point cloud and mesh model from the implicit representation of the scene, the method achieves the collection of multi-source plant phenotypic data in a single data acquisition. Moreover, we establish a multi-view image dataset of different plants, based on which we investigate the performance of the latest NeRF models for 2D view synthesis as well as 3D geometry extraction in plant phenotype acquisition. Specifically, the contributions of this paper are as follows:

- A novel technology, NeRF, is explored to agricultural scenarios in this study, aiming at high-fidelity plant phenotyping in typical agriculture environments.
- A thorough investigation is conducted on the central tasks of extracting high-fidelity multi-view RGB images and intricate topological geometries using NeRF in actual agricultural scenarios.
- A comparison of several SOTA NeRF models in terms of generating new viewpoints and extracting geometric structures is provided, offering invaluable insights for subsequent research.

The rest of this paper is organized as follows. Section 2 surveys related work. Section 3 details the principle of volume density methods and provides a detailed description of the implementation of data acquisition and network design. The experiment results and discussion are presented in Section 4, followed by the conclusion in Section 6.

2. Related works

2.1. Plant phenotyping

Measuring and analyzing plant phenotypes can be used to establish predictive models to assess plant growth characteristics, which are important for precision agriculture as a decision-making tool (Feng et al., 2021). Therefore, it is crucial to investigate non-invasive, affordable and efficient methods for plant phenotyping (Zhao et al., 2023). In recent years, a large number of scholars have made many attempts to combine novel sensors with computer technology. Among them, 2D imaging, which studies plant traits such as color through multi-view RGB imaging, and 3D imaging (Kolhar and Jagtap, 2023; Zhang and Zhang, 2018), which focuses on geometry extraction, have become one of the most important research interests in the field because they serve the most fundamental and widely concerned morphological plant traits.

2.2. Traditional methods

RGB imaging from various perspectives serves distinct purposes in plant phenotyping and growth monitoring (Kumar and Domnic, 2019; Ubbens et al., 2018). Kang et al. detected the location of fruits by processing RGB images of apple trees through deep learning (Kang and Chen, 2020). Top-view RGB imaging systems are typically employed when examining rosette plants to extract growth rate data. These systems capture top-down RGB images of plants such as Arabidopsis (Arabidopsis thaliana) and tobacco (Nicotiana tabacum) to investigate growth rates under conditions of drought stress, chilling stress, and biotic stress (Jansen et al., 2009; Clauw et al., 2015). Plant growth analysis based on top-view images is impacted by challenges such as overlapping leaves and the nastic movement of foliage (Dellen et al., 2015). These obstacles become particularly pronounced when imaging is limited to a single perspective. Multi-view RGB images of cereals, including barley, wheat, rice, sorghum, and various pea field cultivars, are harnessed for the study of growth rates under conditions of drought stress, salt stress, cold stress, and nutrient deficiency (Humplík et al., 2015).

Geometry Extraction of complex unstructured agricultural scenes is the key prerequisite for quantitative extraction of plant metrics. A number of papers applied technologies, which can be divided into two main categories, to obtain the geometric representation of a scene (Paulus, 2019; Kang et al., 2022b). The first is the explicit method, represented by Light Detection and Ranging (LiDAR), while the other is the implicit method, with SDF as its representative.

Guo et al. utilized the Realsense D435i to capture continuous multi-view images of the cabbage and input the images into professional 3D reconstruction software called RealityCapture to create 3D point cloud data for calculating the target cabbage dimensions (Guo et al., 2023). Wu et al. developed a detachable and adjustable according to the size of the target shoot to acquire multi-view stereo (MVS) images and reconstructed 3D point clouds using MVS-based commercial software (Wu et al., 2020). The aforementioned studies calculate the internal parameters of the images, along with the external parameters between them, using feature matching in a series of unordered images. They then proceed to sequentially perform sparse point cloud reconstruction and dense point cloud generation. The quality of the results obtained through these methods is heavily reliant on the resolution and volume of image data, making the process time-consuming. For instance, Guo et al. required 5–8 min to capture 150 photographs and an additional 20 min or more to complete the reconstruction of a single cabbage. Kang et al. proposed a LiDAR-color fusion-based visual sensing and perception strategy for achieving precise scene comprehension and fruit localization in orchards (Kang et al., 2022a; Kang and Wang, 2023). While this method enhances the density of point cloud data and depth sensor accuracy, it remains costly and time-consuming to accumulate a sufficient number of point clouds. Eugene Kok et al. processed RGB data and depth information from a depth camera using a semantic segmentation network and deep learning skeletonization method to reconstruct spatial information of both visible and hidden branches from a single-view image (Kok et al., 2023). However, the algorithm can only reconstruct trees from a single viewpoint and is not suitable for trees with more complex geometries. Yang et al. developed a system for rapid 3D model reconstruction using RGB-D cameras and the point cloud self-registration method (Yang et al., 2022). Although they introduced a rotating table to obtain a complete point cloud, this method is only applicable to potted plants and not suitable for field conditions.

2.3. Learning-based modeling methods

Since the conventional reconstruction method represented a 3D scene explicitly using grids of voxels, point clouds, or meshes, the reconstructed 3D shapes were discrete and at a limited resolution. The novel implicit methods parameterize different kinds of features from the scene (for instance, density, color, occupancy probability, SDF value) as a continuous function approximated via an MLP network. Due to the high accuracy of MLP's function fitting, implicit representations of the scene are often accurate at arbitrary resolutions (Samavati and Soryani, 2023). IM-NET (Chen and Zhang, 2019) trains the network through deep learning using VAE+GAN, which replaced the traditional reconstruction method with a new implicit surface function decoder during the input single view implementation of 3D modeling, resulting in improved reconstruction effectiveness and efficiency. Occupancy Networks (Mescheder et al., 2019) predict binary occupancy rates by acquiring feature vectors and points in space, so that the Occupancy Networks were able to implicitly represent 3D surfaces as continuous decision boundaries for deep neural network classifiers, enabling efficient 3D structural coding through the use of continuous functions to model objects in space. Unlike the principle of Occupancy Networks, DeepSDF (Park et al., 2019) implicitly represents continuous 3D spatial surfaces by directly regressing the SDF. DeepSDF can represent more complex shapes without discrete errors and requires significantly less memory. This concept offers a promising avenue for further research on using neural networks to define implicit scene representations (Park et al., 2019). Nevertheless, these implicit representations require supervised learning with prior information about the shape of a 3D object, making it challenging to directly employ these technologies in real-world environments.

In contrast to these methods, a series of NeRF-based (Mildenhall et al., 2021; Barron et al., 2021; Hedman et al., 2021) approaches accomplish implicit 3D reconstruction of a scene without a geometric prior of the scene by supervising the network training only with the captured multi-view images. However, training the original NeRF model and rendering a novel view image is time-consuming, thus many scholars have proposed improved models based on the original NeRF theory to enhance both the quality of rendered images and the speed of model training. DS-NeRF (Roessle et al., 2022), which utilizes sparse point cloud reconstruction from SfM preprocessing as a dense depth prior to constrain NeRF optimization, which accelerates the initialization of the NeRF network but the convergence of the network still takes hours. Although the hash encoding used by Instant-NGP (Müller et al., 2022) allows NeRF to be trained in less than a few seconds, the geometry extracted from the learned density field contains significant noise due to the lack of geometric constraints in the 3D representation.

Our research addresses the existing limitations in plant phenotyping methods by exploring the application of NeRF technology. In an effort to bridge the gap between 2D and 3D imaging technologies, we aim to pioneer a novel approach to plant phenotyping. Our work contributes to the advancement of agricultural practices by providing a unique perspective and potential breakthroughs in plant phenotype analysis.

3. Materials and methods

To adapt NeRF effectively for agricultural scenarios, firstly to overcome the limitation of slow training speed of the original NeRF, our study introduces hash encoding of Instant-NGP (Müller et al., 2022) to accelerate the training process. To strengthen the geometric constraints within the model, we bring in a rendering method based on SDF (Wang et al., 2021; Zhao et al., 2022). In this section, we will overview the entire process, from image data collection and preparation for training to the architecture and training of the network model. Finally, we will give an overview of the experimental setup used in our study. And our framework of NeRF-based phenotyping system is shown in Fig. 2.

3.1. Data preparation

3.1.1. Image acquisition

We capture high-resolution plant images from various agricultural scenes using the GoPro Hero 11 action camera in Table 1. To reduce

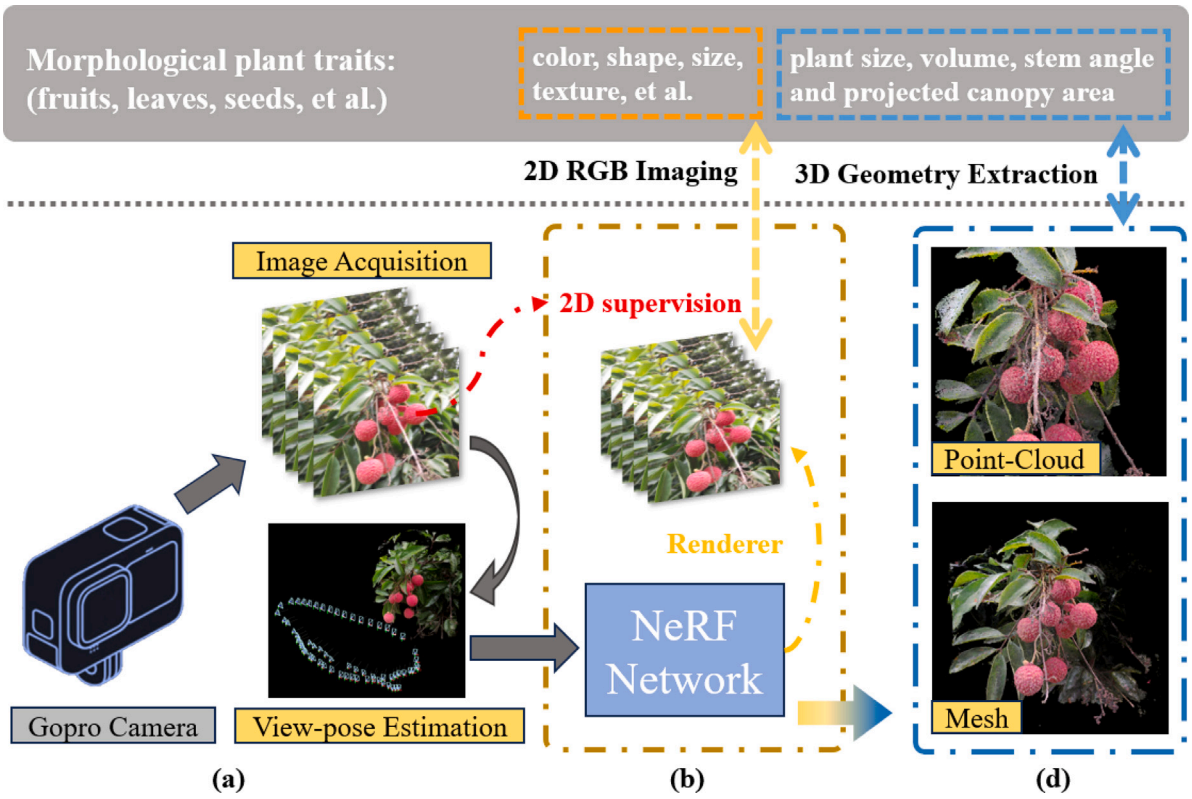

Fig. 2. Framework of phenotyping system via NeRF: (a) Data preparation, (b) Neural rendering (d) Geometry extraction.

Table 1
Camera parameters for GoProHERO11.

Parameter	Value
Name	GoProHERO11
Weight	149.00 g
FPS	120 fps
Resolution	3840 × 2160 pixels
Lens stabilization	Electronic
Battery life	90 min

motion blur and graphic quality issues, the camera is set to work at 120 fps in a 4 K resolution linear imaging mode. This setting allows us to collect image data at a rate of 120 frames per second with a resolution of 3840 × 2160 pixels. For single plants in Fig. 3(a), we aim to capture 360° all-around images to cover all details. For larger, more complex scenes in Fig. 3(b), we choose front views of regions of interest and took images from multiple angles.

3.1.2. Dataset generation for NeRF training

(1) View-pose estimation. COLMAP (Schonberger and Frahm, 2016), a SOTA Structure-from-Motion (SfM) and MVS pipeline, reconstructs 3D models from unordered image sets. The pipeline in Fig. 4 of SfM is used to estimate images' poses as a prior to supervise the training of the network.

For every image in the datasets, COLMAP detects and describes local features. A pairwise matching algorithm then associates keypoints based on their descriptors. Formally, if p_i and p_j are keypoints in images I_i and I_j respectively, their match is established based on:

$$D(p_i, p_j) = \|\mathrm{desc}(p_i) - \mathrm{desc}(p_j)\|_2, \tag{1}$$

where $\mathrm{desc}(p)$ returns the descriptor for keypoint p.

After feature matching, geometrically consistent matches are pinpointed by employing a fundamental or essential matrix. The essential matrix, denotes as E, captures the geometric relationship between two calibrated images. It is a matrix that relates corresponding points in one image to epipolar lines in the other image. Formally:

$$p_j^T E p_i = 0, \tag{2}$$

where p_i and p_j are corresponding points in homogeneous coordinates.

COLMAP utilizes an incremental approach to SfM. Starting with a pair of images with the largest number of geometrically consistent matches, the scene is incrementally expanded by registering additional images based on shared keypoints.

After initial camera pose estimation, COLMAP refines the camera parameters, 3D structure, and even image keypoints simultaneously using bundle adjustment. The objective function J being minimized is:

$$J = \sum_{i,j} w_{ij} \|p_j - \pi(P_i, X_j)\|^2. \tag{3}$$

Where w_{ij} is a visibility term, which is 1 if point X_i is visible in image I_i and 0 otherwise. π is the projection function. P_i is the projection matrix of the ith image. X_j is the 3D position of the jth point.

(2) Data format conversion.

To assist NeRF in processing datasets from different sources, these images, along with their correlated camera poses, camera parameters, are required to be converted into a certain format.

The Local Light Field Fusion (LLFF) format (Mildenhall et al., 2019) is designed for capturing real-world scenes using a series of photographs taken from various viewpoints. Distinct from traditional light field cameras, LLFF does not require specialized hardware but leverages conventional cameras. By moving the camera throughout the scene and capturing multiple images, a scene representation is generated. Given NeRF's objective to learn 3D representations of a scene from a series of

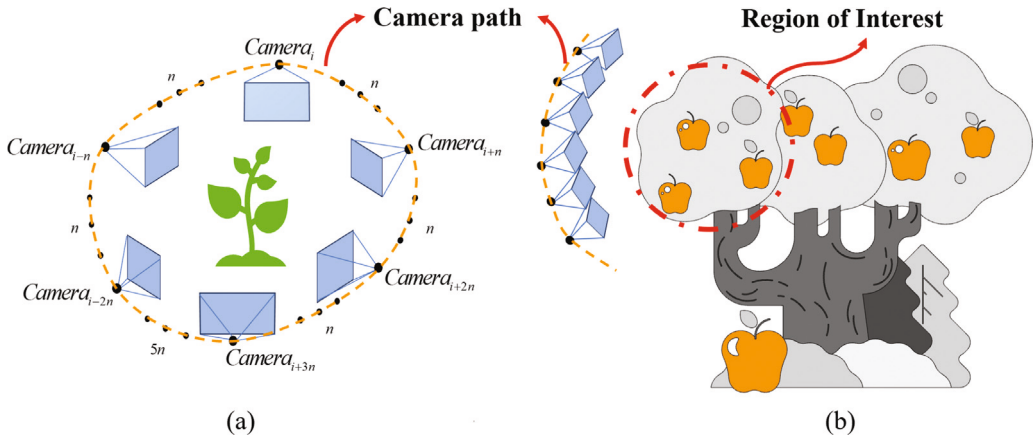

Fig. 3. (a) 360° image capturing, (b) front views capturing.

images, the LLFF datasets format naturally becomes an optimal choice for representing input data from real scenes.

A LLFF datasets typically comprises the following crucial components: (1) A set of images captured from different perspectives, (2) Camera intrinsic parameters, (3) Camera extrinsic parameters. Within the previous steps, we have obtained the (1) image sequences and (2) intrinsic parameters of cameras. To represent the output camera poses of COLMAP in the LLFF format, it is needed to invert the transformation from a world-to-camera format (COLMAP) to camera-to-world format for LLFF.

Specifically, COLMAP outputs a rotation matrix R and a translation vector t for each camera's pose $C = -R^T \cdot t$. For a rotation matrix, which is orthogonal, the inverse R^{-1} is equal to the transpose R^T. The translation in the world-to-camera format can be found by transforming the camera's position in the world coordinates using the inverse rotation:

$$t' = -R^T \cdot t, \tag{4}$$

Therefore, the LLFF camera-to-world transformation matrix is constructed as:

$$M = \begin{bmatrix} R^T & t' \\ 0 & 1 \end{bmatrix}. \tag{5}$$

3.2. Fundamental formulation of NeRF

Scene representation on volume rendering is essential for NeRF to generate novel view images and geometry. This section gives the formulations of the NeRF working mechanism in both novel-view rendering and geometry reconstruction.

3.2.1. Rendering novel views with NeRF

NeRF is first proposed to use volume rendering formula to achieve highly photorealistic view synthesis with implicit neural scene representation via MLP. Essentially, NeRF uses an MLP network H_Θ to describe the mapping (6) between density σ and directional emitted color $c = (r, g, b)$ of each point in a 3D scene with the spatial coordinates (x, y, z) and corresponding viewing direction vector $\mathbf{d}$.

$$H_\Theta(x, y, z, \mathbf{d}) \rightarrow (c, \sigma). \tag{6}$$

During the rendering process, as shown in Fig. 5, a ray $\mathbf{r}(t) = o + t\mathbf{d}$ is emitted into the 3D scene from a given camera position $o = (x_o, y_o, z_o)$, where t is the straight-line distance from the point on the ray to the camera's origin o and $\mathbf{d}$ is the 3D Cartesian unit vector representing viewing direction. As the ray travels, a sufficient number of spatial points of this ray are sampled, and the σ and c of each point are queried to the H_Θ. Finally, Eq. (7) from classical volume rendering (BP, 1984)

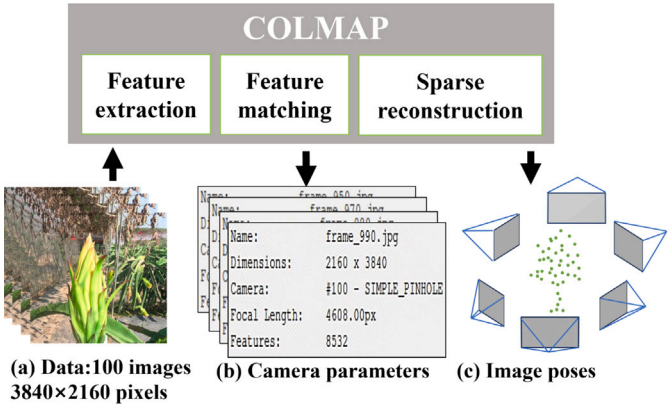

Fig. 4. Framework of data processing.

is used to accumulate all the sampled points and obtain the color value of the corresponding pixel on the image plane.

$$C(\mathbf{r}) = \int_{t_1}^{t_2} T(t) \cdot \sigma(\mathbf{r}(t)) \cdot \mathbf{c}(\mathbf{r}(t), \mathbf{d}) \cdot dt, \tag{7}$$

where $T(t)$ denotes the transmittance function, alternatively referred to as the accumulated density. This function quantifies the possibility that a ray traverses the distance from t_1 to t_2 without running into an obstruction, as described by the following equation:

$$T(t) = \exp(-\int_{t_1}^{t} \sigma(\mathbf{r}(u)) \cdot du), \tag{8}$$

For every pixel, a squared error photometric loss is employed for the optimization of the wight Θ of MLP. When applied across the entire image, this loss is represented as follows:

$$\mathcal{L}_{\text{NeRF}} = \sum_{r \in R} \|C(\mathbf{r}) - C_{gt}(\mathbf{r})\|^2. \tag{9}$$

where $C_{gt}(\mathbf{r})$ is the ground truth color of the training image pixels associated with the ray $\mathbf{r}$, and R is a batch of rays associated with the synthesized image.

3.2.2. Extraction geometry from NeRF

(1) Point-Cloud extraction from NeRF

While NeRF primarily focuses on the reconstruction and rendering of scenes from novel views, it inherently contains a wealth of 3D structural information, making it possible to extract point-cloud data. The continuous feature of NeRF's representation allows the inference of spatial geometries by observing changes in radiance and density along camera rays.

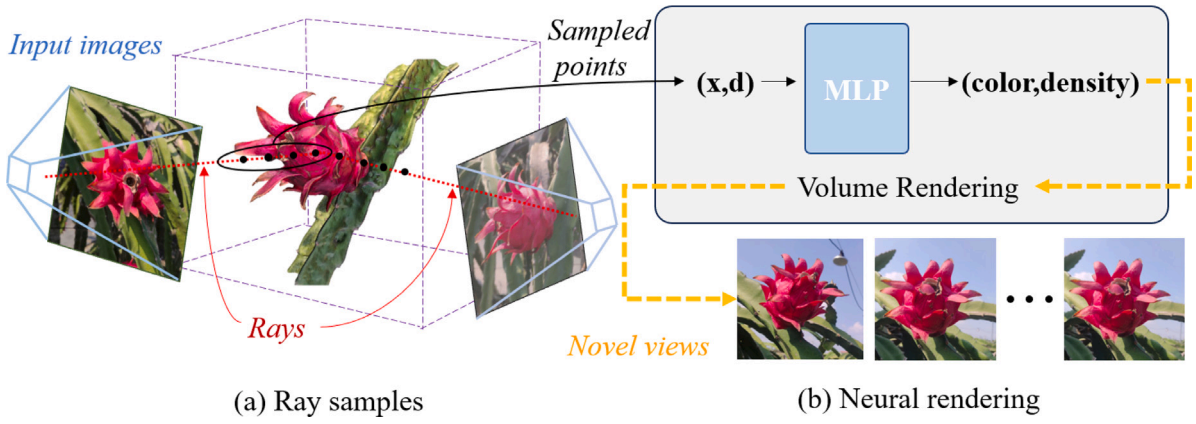

Fig. 5. Illustration of novel view synthesis based on volume rendering.

The crucial step in point-cloud extraction is depth estimation. By analyzing the transmittance function $T(t)$ in (8) along a ray $\mathbf{r}(t)$, one can observe the depth where the function experiences significant change, indicating the presence of a surface. The depth at which $T(t)$ sees a sharp decline is usually aligned with the surface of an object within the scene. Mathematically, this depth $t_{surface}$ for a ray $\mathbf{r}(t)$ can be pinpointed as the position where the transmittance's rate of change is most abrupt by minimizing the first order derivative of $T(t)$ relative to t:

$$t_{surface} \approx \arg\min_t \left(\frac{dT(t)}{dt} \right), \tag{10}$$

With the depth approximated, the next step is to calculate the corresponding 3D point on $\mathbf{r}(t) = o + t\mathbf{d}$:

$$\mathbf{p} = \mathbf{o} + t_{surface}\mathbf{d}, \tag{11}$$

where $\mathbf{p}$ represents the 3D point, $\mathbf{o}$ is the camera's origin, $\mathbf{d}$ signifies the ray direction and $t_{surface}$ represents the distance traveled by the ray as it crosses the surface.

Upon determining the 3D position, the color value at this position can be directly queried from NeRF (6):

$$\mathbf{color} = \mathbf{c}(\mathbf{p}, \mathbf{d}). \tag{12}$$

Repeating these steps for every pixel across one or multiple images generates a dense point-cloud. Each point within this cloud corresponds to a surface in the original scene, carrying a color that mirrors the appearance of that surface under the sampled viewing direction.

(2) Mesh extraction from NeRF

Given a predefined 3D region of interest, a set of spatial points $P = \{p_1, p_2, \ldots, p_n\}$ is generated via dense volumetric sampling. For each point $p_i \in P$, it is evaluated through the NeRF model to obtain The density values, $\sigma(p_i) = \text{NeRF}_\sigma(p_i)$, form the basis for surface extraction. The Marching Cubes (Lorensen and Cline, 1998) algorithm identifies the iso-surface by thresholding the density values:

$$M = \text{MarchingCubes}(P, \sigma_{\text{threshold}}), \tag{13}$$

where M is the resultant mesh and $\sigma_{threshold}$ is an optimal density value demarcating the object's boundary.

For every vertex v_j in mesh M, a viewing ray r_j is constructed and queried $\mathbf{c}(v_j) = \text{NeRF}_\mathbf{c}(v_j, r_j)$ in NeRF. Those radiance values derived from NeRF are mapped onto mesh M, assigning color to each vertex:

$$\text{Color}(v_j) = \mathbf{c}(v_j), \tag{14}$$

For a 2D texture representation, the vertex-colored mesh undergoes UV unwrapping (Sander et al., 2001). To minimize distortion, algorithms like Least Squares Conformal Mapping (LSCM) (Lévy et al., 2023) can be employed. The objective of LSCM is to minimize the conformal energy:

$$E(u, v) = \int_\Omega \left(|\nabla u|^2 + |\nabla v|^2 \right) dA. \tag{15}$$

Where (u, v) are the 2D texture coordinates for each vertex in M, Ω represents the object's surface, and dA is a differential area element on the mesh's surface, indicating that the energy is computed by integrating over the entire surface of the mesh.

3.3. Learning from NeRF

Multi-resolution hash encoding. To address the drawback of slow NeRF model training, Instant-NGP (Müller et al., 2022) makes use of a multi-resolution hashed positional encoding as additional learned features, the model could represent scenes accurately with tiny and efficient MLPs. In detail, Instant-NGP operates on the premise that the object to be reconstructed is enclosed within multi-resolution voxel grids. Each of these voxel grids at different resolutions is then correspondingly linked to a hash table, featuring a fixed-size array of adaptable feature vectors.

For any spatial point $\mathbf{x} \in \mathbb{R}^3$ within various resolution grids, it obtains the hash encoding $h^i(\mathbf{x}) \in \mathbb{R}^d$ (d is the dimension of a feature vector, $i = 1, \ldots, L$) corresponding to the respective level by trilinear interpolation. The hash encodings at all L levels are subsequently concatenated to form the multi-resolution hash encoding $h(\mathbf{x}) = \{h^i(\mathbf{x})\}_{i=1}^L \in \mathbb{R}^{L \times d}$.

Volume rendering of SDF. To precisely extract the geometric surface from NeRF's implicit representation, Neus (Wang et al., 2021) proposes to represent 3D scene as a SDF $f(\mathbf{x}) : \mathbb{R}^3 \to \mathbb{R}$ instead of NeRF's density field and introduce a ***S-density*** which can be represented as the logistic density distribution $\phi_b(f(\mathbf{x})) = be^{-bf(\mathbf{x})}/(1 + e^{-bf(\mathbf{x})})^2$, where b is a trainable hyper parameter and gradually increases to a large number as the network training converges. And the surface S can be extracted by the zero-level set of its SDF:

$$S = \left\{ \mathrm{x} \in \mathbb{R}^3 | f(\mathbf{x}) = 0 \right\}, \tag{16}$$

To train the neural SDF representation, Neus follows NeRF's volume rendering Eq. (7). Given a pixel, the renderer emitted a ray from this pixel as $\{\mathbf{p}(t) = \mathbf{o} + t\mathbf{v} | t \geq 0\}$, where $\mathbf{o}$ is origin of the ray and $\mathbf{v}$ is the ray direction and accumulate the colors along the ray by:

$$C(\mathbf{o}, \mathbf{v}) = \int_0^{+\infty} w(t) c(\mathbf{p}(t), \mathbf{v}) \mathrm{d}t, \tag{17}$$

where $C(\mathbf{o}, \mathbf{v})$ is the rendered color for this pixel, and $c(\mathbf{p}(t), \mathbf{v})$ the sampled colors along the ray. Especially, the weight $w(t)$ for point $\mathbf{p}(t)$ is rebuilt by unbiased and occlusion-aware properties to guarantee that

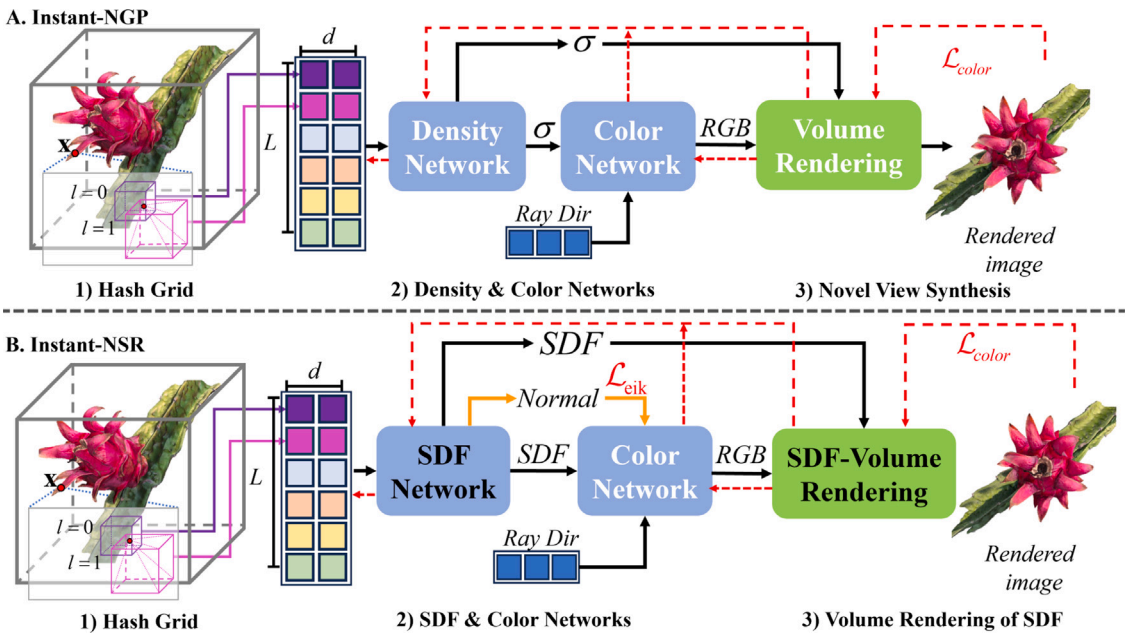

Fig. 6. Pipeline of Instant-NGP and Instant-NSR. **(A). Instant-NGP:** Given a 3D point **x**, the (1) hash grid corresponding to each level l in the voxel grid is interpolated to hash encoding, then the density and color values are predicted by the (2) MLPs of density and color, and the color of the pixel is calculated by (3) volumetric rendering. **(B). Instant-NSR:** Compared to the previous **Instant-NGP**, both the (2) MLPs and the (3) volume rendering are based on SDF and employ an extra normal regularization to strengthen the geometrical constraints in the network training.

the surface of an actual object contributes the most to the rendering result, that is:

$$w(t) = \frac{\phi_s(f(\mathbf{p}(t)))}{\int_0^{+\infty} \phi_s(f(\mathbf{p}(u)))\mathrm{d}u}. \tag{18}$$

Truncated SDF hash grids. To increase the stability of network training, here, we introduce a neural surface reconstruction method that accelerates training with hash encoding, Instant-NSR (Zhao et al., 2022). as shown in Fig. 6, this method is similar to Instant-NGP (Müller et al., 2022) in that it hash encodes points in spatial at the front-end of the neural network, but employs Neus's SDF architecture (Wang et al., 2021) in the neural network instead of the Instant-NGP's density architecture, and in the image renderer, also SDF-based volume rendering formulation (17) is used.

In addition, Instant-NSR uses Truncated SDF (TSDF) to skilfully solve the convergence problem caused by applying SDF representations to hash coding frameworks. Since original SDF-based methods utilize cumulative density distribution $\phi_b(f(\mathbf{x})) = be^{-bf(\mathbf{x})}/(1+e^{-bf(\mathbf{x})})^2$ to the compute $w(t)$ in Eq. (18), the term $-bf(\mathbf{x})$ will be a large positive number when b is increased, resulting in $e^{-bf(\mathbf{x})}$ closing to infinity. This numerical instability will cause the network to converge hardly during the training process. The characteristic of TSDF value between -1 to 1 can effectively prevent the occurrence of network divergence caused by numerical overflow. Therefore, we utilize the sigmoid function $\pi(\cdot)$ after the SDF output of the network to achieve the truncation effect of the TSDF, as below:

$$\pi(f(\mathbf{x})) = \frac{1-e^{-bf(\mathbf{x})}}{1+e^{-bf(\mathbf{x})}}. \tag{19}$$

Thus, we can now replace the formula $\phi_b(f(\mathbf{x})) = be^{-bf(\mathbf{x})}/(1+e^{-bf(\mathbf{x})})^2$ in Neus with $\phi_b(f(\mathbf{x})) = be^{-b\pi(f(\mathbf{x}))}/(1+e^{-b\pi(f(\mathbf{x}))})^2$.

3.4. Network training

In order to obtain an optimal representation of the scene via our neural network model, we employ a compound loss function. This loss function is constructed using two primary components: the rendering loss and the eikonal loss.

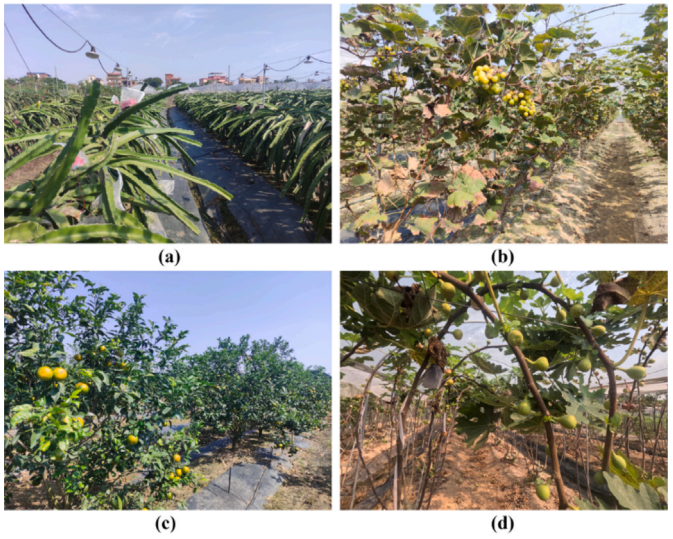

Fig. 7. Photographs of indoor and outdoor orchards: (a) pitahaya orchard, (b) grape orchard, (c) orange orchard, (d) fig orchard.

3.4.1. Rendering loss

The primary goal of our method is to produce high-quality renderings that closely match the ground truth images. Therefore, the rendering loss is crucial as it quantifies the discrepancy between the rendered images from the network and the actual images, and we generally apply this $\mathcal{L}_{color}$ in both Instant-NGP and Instant-NSR.

Given a set of ground truth images I_{gt} and the corresponding set of images rendered by the network I_{pred}, the rendering loss $\mathcal{L}_{color}$ is defined as:

$$\mathcal{L}_{color} = \frac{1}{N}\sum_{i=1}^{N} \| I_{gt}^{(i)} - I_{pred}^{(i)} \|_2^2, \tag{20}$$

where N is the total number of images in the datasets.

3.4.2. Eikonal loss

While the rendering loss $\mathcal{L}_{color}$ ensures the visual accuracy of the rendered images, the eikonal loss $\mathcal{L}_{eik}$ is used in Instant-NSR to ensure that the estimated SDF values conform to the properties of a true SDF following one of the primary properties that the gradient $\nabla f(\mathbf{x})$ of the SDF should have a magnitude of 1 everywhere.

Given an SDF represented by $f(\mathbf{x})$, the eikonal loss $\mathcal{L}_{eik}$ is given by:

$$\mathcal{L}_{eik} = \frac{1}{M}\sum_{j=1}^{M}\left(\|\nabla f(\mathbf{x}_j)\|_2 - 1\right)^2, \tag{21}$$

where M is the total number of sampled points from the scene, and $\nabla f(\mathbf{x}_j)$ is the gradient of the SDF for the jth point.

3.4.3. Total loss

Combining both losses, the total loss function L_{total} used to train the Instant-NSR network is:

$$L_{total} = \alpha L_{render} + \beta L_{eik}. \tag{22}$$

where α and β are weighting factors that balance the contribution of the two losses. These weights are hyperparameters and are chosen based on cross-validation to achieve the best performance on a validation set.

3.5. Implementation details

All neural network training and computational experiments are conducted on an Ubuntu-based platform. The hardware specifications of the system included an Intel Core i9-13900 CPU. For graphic processing and deep learning computations, we actually trained all of our models for plants using only a single NVIDIA 3090 graphics card, which has 24 Gigabytes of video memory. This high-end setup ensures the efficient execution of data-intensive processes and neural network operations.

4. Results

4.1. Experimental setup

In this study, we utilized a high-speed motion camera, GoPro Hero 11, to acquire the image datasets for our experiments, which were divided into three levels, L_1, L_2, and L_3, based on the geometrical distribution of important structures such as leaves and fruits of the plants in them. Firstly, we tested the performance of Instant-NGP in rendering the images in these datasets, and secondly, we tested two different geometrically expressed NeRF models, Instant-NGP based on the density field and Instant-NSR based on the SDF, for their ability to extract the geometrical models of the plants in these datasets.

4.2. Datasets

In this section, we collected image datasets of litchi at the Litchi Expo Park in Zengcheng District, Guangzhou, image datasets of bell peppers, tomatoes, and watermelons planted in greenhouses at the Baiyun Experimental Base of the Guangzhou Academy of Agricultural Sciences, and image datasets of grapes, pitahaya, pitahaya flowers, oranges, and figs at the Shangguo Ecological Picking Garden in Panyu District, Guangzhou, as show in Fig. 7. Furthermore, this study classified the scenes based on the interplay and occlusion among these key plant constituents. Three distinct levels were defined to represent these datasets: L_1, L_2, and L_3 in Fig. 8.

- L_1 represents scenes where the fruits, leaves and branches are clearly visible with minimal to no occlusion between them. In such scenarios, each component is clearly visible, making it an ideal representation of less dense plant geometry.
- L_2 represents scenes with a slightly denser configuration. Here, several fruits overlap each other and there is slight occlusion by the leaves. This level is moderately challenging and represents environments where components begin to intertwine.
- L_3 is indicative of the most complex and confused scenarios. In these scenes, fruit, leaves and branches are chaotically distributed and the geometric topology is highly complex. Such environments resemble dense plant canopies and thickets, where distinguishing individual components becomes particularly challenging.

4.3. Demonstration in 2D imaging of NeRF

This section demonstrates that the results of real-time rendering of our datasets in the quickest training NeRF model Instant-NGP (Müller et al., 2022). To explore the effect of the complexity of the scene on the NeRF rendering quality, we set the number of images in all datasets to 100 images, all of which were captured by GoPro cameras in 120 Hz, linear imaging mode, with a resolution of 3840 × 2160 pixels, according to the settings in Section 3.1.1.

The Peak Signal-to-Noise Ratio (PSNR) has been widely recognized as an essential metric for the quantitative evaluation of image quality. Predominantly used in the domain of image compression and reconstruction, its application has further expanded into the emerging realms of computer vision and neural graphics (Mildenhall et al., 2021). Higher PSNR values imply superior image fidelity, indicating a closer match to the reference. The foundation of PSNR lies in the Mean Squared Error (MSE), which quantifies the average squared discrepancies between the pixel values of the reference and the examined images. Formally, for an image of size $M \times N$, the MSE is defined as:

$$\text{MSE} = \frac{1}{M \times N}\sum_{i=1}^{M}\sum_{j=1}^{N}[I_{\text{reference}}(i,j) - I_{\text{examined}}(i,j)]^2, \tag{23}$$

where $I_{\text{reference}}$ and I_{examined} represent the pixel intensities of the reference and examined images, respectively. With the MSE in hand, the PSNR is calculated using:

$$\text{PSNR} = 10 \times \log_{10}\left(\frac{MAX_I^2}{\text{MSE}}\right). \tag{24}$$

where MAX_I signifies the maximum feasible pixel intensity for the image. (For instance, for a typical 8-bit grayscale image, MAX_I equals 255.)

To quantitatively evaluate these experimental results, we referenced the original NeRF paper's real-world dataset benchmark which recorded a PSNR of 26.50 dB in their Real Forward-Facing dataset. On this baseline, a dataset with a PSNR close to or higher than 26.50 dB indicates high reconstruction quality, and the opposite indicates that the dataset struggles to converge well.

Fig. 9 shows the training time of our full datasets in Instant-NGP and the PSNR for each scene. To demonstrate the power of NeRF to synthesize novel views, we have selected four views other than the training data used for the model, namely right, left, vertical and bottom views.

4.4. Demonstration of 3D geometry extraction form NeRF

This section details the experimental results of the geometry extraction from NeRF based on our plant datasets. First of all, we extract the point clouds and meshes of the plants from Instant-NGP according to the method introduced in Section 3.2.2, and demonstrate these results comprehensively in Fig. 10. And note that in this experiment, the mesh models generated by Reality capture is used as a reference for the geometric extraction results, because NeRF is supervised by 2D image data without the ground truth of 3D information. (Reality capture is a commercial MVS-based modeling software, which integrates the core methods of MVS as well as comprehensive steps,

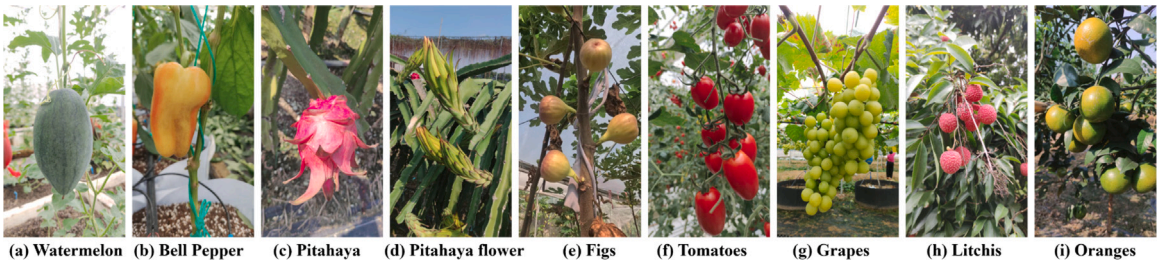

Fig. 8. Demonstration of our datasets with three progressive levels: L_1: (a), (b), (c); L_2: (d), (e), (f); L_3: (g), (h), (i).

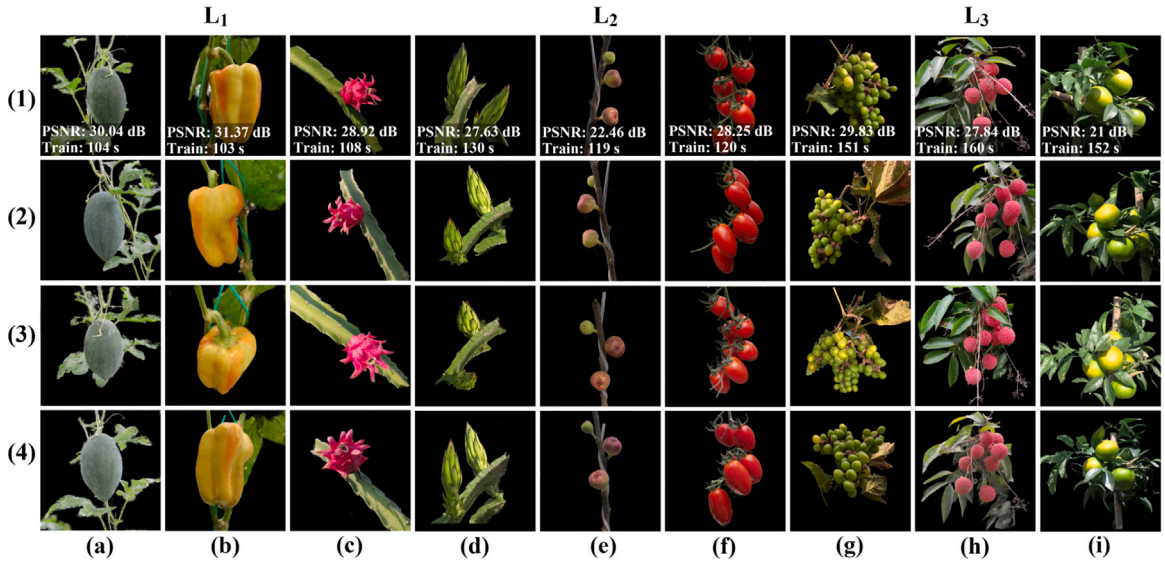

Fig. 9. Novel-view image rendered from Instant-NGP. (1) right-view, (2) left-view, (3) vertical-view, (4) bottom-view.

Table 2
Comparison of time and PSNR metrics for different datasets using Reality Capture (RC), Instant-NGP, and Instant-NSR.

	Dataset	RC	Instant-NGP		Instant-NSR	
	Method|Metric	Time	Time	PSNR	Time	PSNR
L_1	Watermelon	12 min	1.73 min	30.04	12.69 min	29.5
	Bell pepper	11 min	1.71 min	31.37	12.59 min	24.5
	Pitahaya	12 min	1.80 min	28.92	12.37 min	29.2
L_2	Pitahaya flower	15 min	2.16 min	27.63	12.88 min	26.8
	Figs	14 min	1.98 min	22.46	11.39 min	23.5
	Tomatoes	16 min	2 min	28.25	12.28 min	28.4
L_3	Grapes	22 min	2.51 min	29.83	13.46 min	25.5
	Litchis	25 min	2.66 min	27.84	13.58 min	27.5
	Oranges	21 min	2.53 min	29.89	13.47 min	20.5

including photo alignment, feature extraction, feature matching, camera viewpoint calculation, and 3D point cloud reconstruction Wu et al., 2020.)

Furthermore, Fig. 11 shows the comparison between the mesh of the plants extracted from Instant-NGP and Instant-NSR using the Marching cubes algorithm in Section 3.2.2. The goal of this controlled experiment is to explore the differences in geometric representation between the NeRF model based on the density architecture and the NeRF model based on the SDF architecture.

5. Discussion

In our comprehensive analysis of the experimental results, we observe distinctive challenges and variations across three datasets, L_1, L_2, and L_3, each representing different levels of complexity in agricultural scenes. These complexities are indicative of the unique features present in agricultural environments, such as natural diversity, occlusions, and surface complexities, as discussed below:

- L_1 **- Less Dense Plant Geometry:** L_1, featuring scenes with fruits like watermelon, bell pepper, and pitahaya, presents a scenario with relatively less dense plant geometry. The clear visibility of components is influenced by the predictable growth patterns and distinct characteristics of each plant. The challenge in this setting lies more in addressing the natural diversity and variances, as the well-separated components make occlusions and overlapping less pronounced. The unpredictability in growth patterns contributes to the need for generalization in 3D reconstruction.
- L_2 **- Moderate Density and Intertwining:** Moving to L_2, which includes pitahaya flower, figs, and tomatoes, there is a moderate increase in geometric density. The growth habit of the plants in this dataset results in overlapping and intertwining components.

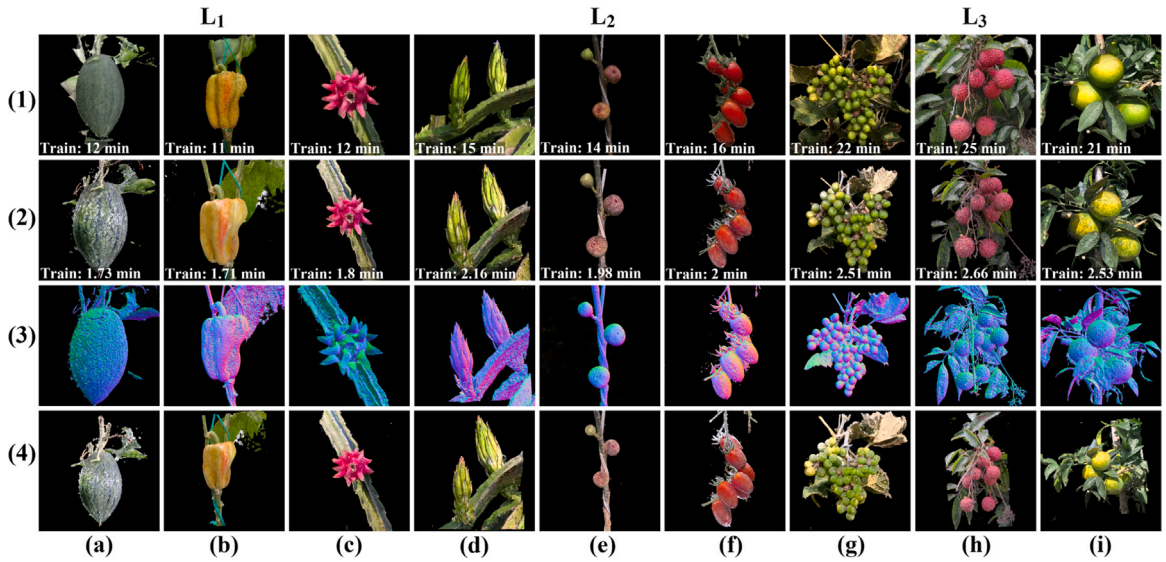

Fig. 10. 3D models extracted from L_1, L_2, L_3 datasets. (1) the mesh models extracted from Reality-Capture, (2) the textured mesh models, (3) the normal mapping models, and (4) the point clouds extracted from Instant-NGP.

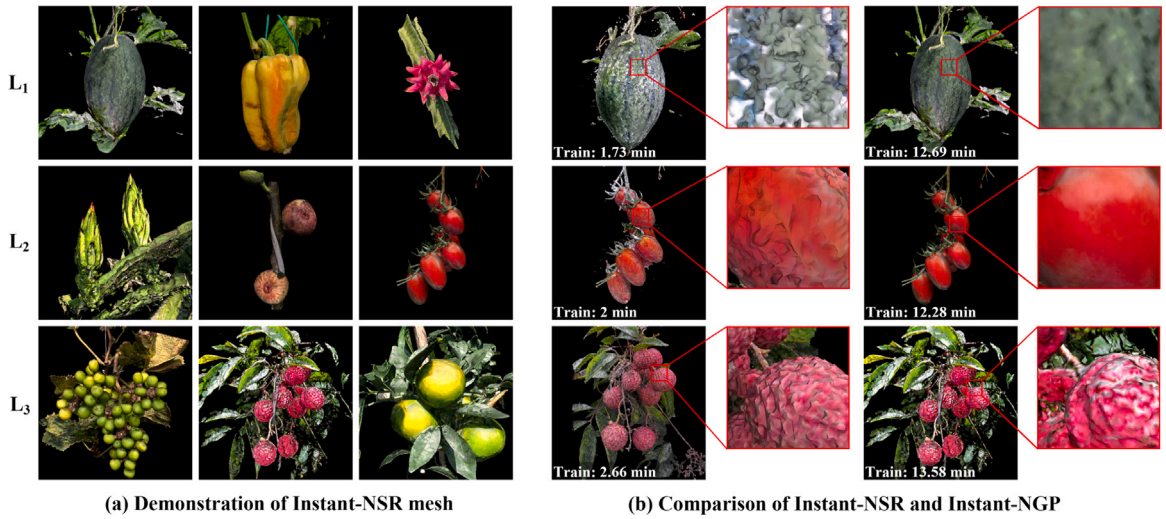

Fig. 11. (a) Demonstration of the meshes obtained via Instant-NSR, (b) Comparison of Instant-NGP (left) and Instant-NSR (right) in details on the model surface.

The challenges of occlusion and overlap become more apparent, demonstrating the impact of growth habit on visual blockage. In addition, dealing with surface complexity becomes critical, as the complex structures on leaves and fruits require advanced algorithms for accurate representation.

- L_3 **- Highly Complex and Disordered Environments:** In L_3, featuring grapes, litchis, and oranges, we explore the most complex and disordered agricultural scenarios. The unpredictability of growth patterns, coupled with significant overlap and entanglement, compounds the challenges. The dense leaves and randomly distributed components highlight the severe occlusions and overlaps, requiring advanced algorithms to navigate through the visual blockages. Capturing surface complexity becomes even more challenging in these highly complex environments.

5.1. Analysis of rendering results

In terms of efficiency in processing data, one of the key strengths of the Instant-NGP model lies in its efficiency. Across our datasets L_1, L_2, and L_3, the model consistently converges in under three minutes. Furthermore, once converged, the model is capable of real-time rendering from any given viewpoint, demonstrating its prowess in generating novel views. This efficiency exceeds that of traditional RGB camera sampling technologies for phenotype collection, demonstrating the potential of NeRF to enable more versatile and efficient ways of observing plant traits from different perspectives.

In terms of the quantitative metric PSNR for image rendering in Table 2, instant-NGP reaches the highest PSNR of 31.37 dB within two minutes for the three scenes in L_1. In L_2, both datasets (d) and (f) can achieve a PSNR of higher than 26.5 dB. In the scenes of Litchis and grapes in L_3, the training time is on average 50 s longer than that in L_1, but the PSNR also exceeds the baseline of 26.5 dB for all of them. Taking into account the class of the dataset and the quality of the image data, we can see that with simple geometry, instant-NGP can converge very quickly and get a high quality of new view rendering, while complex geometry will cause the network to converge slower.

However, the dataset (e) of figs and (i) of oranges show a very low level of PSNRs: 22.46 dB and 21 dB. Since the PSNR is an average representation of the difference between the rendered image and the ground truth image, we displayed the difference between the full rendered result containing the background and the plant subject and the actual image in Fig. 12. In this case **(1)** & **(2)**, the background in the rendering

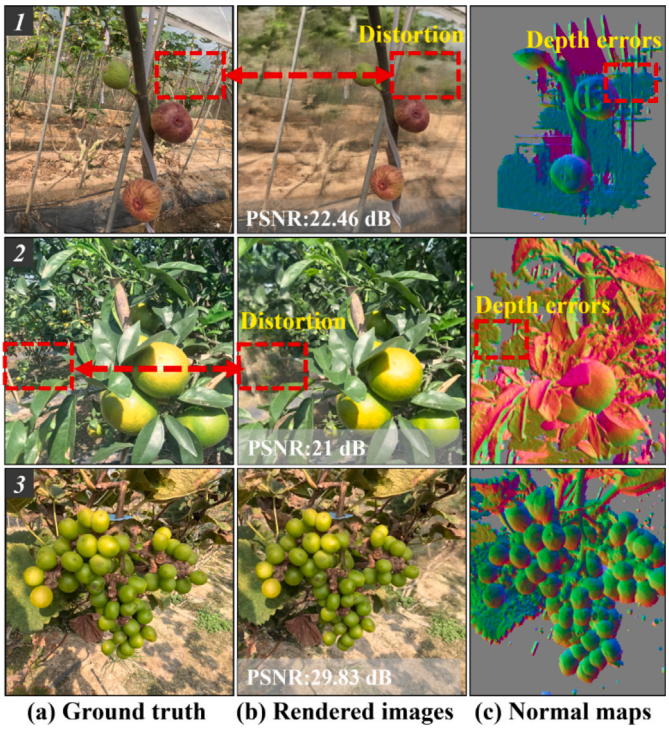

Fig. 12. Analysis of quantitative metric PSNR.

result is severely defocused, additionally, the background inside differs a lot from the real data, resulting in a low PSNR. On the contrary of **(3)**, the background and the subject are clearly distinguishable and not far from each other, thus giving a high PSNR. From this result, it can be observed that Instant-NGP works well for reconstructing plants in the center of the scene, but it cannot recover background objects in the distance, which may limit the application of NeRF in large-scale phenotype acquisition. In addressing the limitations observed in datasets (e) and (i) with low PSNR values, future improvements could involve incorporating additional depth priors. For instance, leveraging RGB-D cameras or sparse point clouds generated by COLMAP can provide depth information to supervise the training in distant scenes, mitigating the issue of blurred rendering results.

5.2. Analysis of geometry extraction results

In the comparison of Instant-NGP and Reality Capture, it is not hard to see that Instant-NGP has a faster modeling speed and at the meantime, it provides a variety of geometrical representations (point cloud, mesh, texture), and even capable of generating rendered images with arbitrary viewpoints. Compared with the traditional Imaging system, NeRF provides a new pattern of information acquisition, which is not to model the plants directly, but to store volumetric and color information in the neural network, and then convert the parameters of the neural network into the multi-source information needed for phenotyping by methods such as volume rendering.

However, the NeRF model based on volume density representation lacks geometric constraints to accurately represent the surface of an object, so when modeling plants with smooth surfaces such as watermelons and bell peppers, their meshes are distorted in surface details. The SDF-based NeRF models are excellent at representing the surface of an object with a smooth and continuous iso-surface, but in scenarios where the bump mapping is extremely complex and the SDF gradient varies drastically, the NeRF models converge very slowly. This emphasizes the potential for improving NeRF's geometric expression by incorporating additional geometric constraints, such as monitoring the gradient of SDF values.

6. Conclusion and future work

In this study, we investigated a novel NeRF-based approach for multi-source phenotypic information acquisition to achieve high-fidelity and high-throughput phenotypic reconstruction of a wide range of plants. First of all, our experiments across different agricultural datasets, L_1, L_2, and L_3, highlighted the diverse challenges inherent in agricultural scenes. These challenges include addressing natural diversity and variances, overcoming occlusions and overlapping due to plant growth habits, and accurately representing intricate surface complexities. These findings underscore the unique complexities of agricultural environments and emphasize the need for advanced technologies, such as NeRF, to navigate and overcome these challenges in 3D reconstruction.

In our experiments, Instant-NGP played a crucial role in accelerating NeRF network training, enabling the efficient modeling and inference of multiple geometric representations when compared to the traditional MVS reconstruction method. Additionally, NeRF demonstrated its capability to generate realistic novel view images through volume rendering. Our comprehensive evaluation of NeRF models, employing two architectures (volume density and SDF), revealed distinct strengths based on the nature of plant surfaces.

The experimental results indicated that the volume density-based NeRF model, exemplified by Instant-NGP, excelled in representing plants with uneven surfaces, such as litchi. On the other hand, the SDF-based model, represented by Instant-NRS, showcased superior performance in reconstructing smooth and continuous plant surfaces, as observed in watermelon and grape scenarios.

However, it is noteworthy that the dataset containing figs (e) and oranges (i) exhibited lower PSNR values (22.46 dB and 21 dB). The associated analysis revealed that Instant-NGP, while proficient in reconstructing plants in the center of the scene, faced challenges in recovering background objects in the distance. While this limitation does not significantly impact the three-dimensional reconstruction of the main plant subject, it becomes a noteworthy constraint in the context of large-scale plant phenotype extraction.

Future work will focus on improving the applicability of NeRF by enabling rapid and accurate modeling in scenarios with sparse views. Additionally, we will explore methods to enhance background processing while preserving the reconstruction quality of the subject. This exploration aims to advance the application of NeRF in acquiring large-scale plant phenotypes.

CRediT authorship contribution statement

Kewei Hu: Conceptualization, Data curation, Investigation, Methodology, Software, Validation, Writing – original draft. **Wei Ying:** Data curation, Validation, Visualization. **Yaoqiang Pan:** Data curation, Visualization, Writing – original draft. **Hanwen Kang:** Conceptualization, Supervision, Writing – review & editing. **Chao Chen:** Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare no conflict of interest.

Data availability

The subset of data used for training and evaluation in this study, as well as several examples of trained NeRF models, are available at https://pan.baidu.com/s/1lBMlC8FFVrwOC7VRx4j9cw?pwd=39tc.

References

Asaari, M.S.M., Mertens, S., Dhondt, S., Inzé, D., Wuyts, N., Scheunders, P., 2019. Analysis of hyperspectral images for detection of drought stress and recovery in maize plants in a high-throughput phenotyping platform. Comput. Electron. Agric. 162, 749–758.

Barron, J.T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., Srinivasan, P.P., 2021. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5855–5864.

BP, K.J.V.H., 1984. Ray tracing volume densities ACM SIGGRAPH comput. Graph 18 (3), 165.

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

Clauw, P., Coppens, F., De Beuf, K., Dhondt, S., Van Daele, T., Maleux, K., Storme, V., Clement, L., Gonzalez, N., Inzé, D., 2015. Leaf responses to mild drought stress in natural variants of arabidopsis. Plant Physiol. 167 (3), 800–816.

Dellen, B., Scharr, H., Torras, C., 2015. Growth signatures of rosette plants from time-lapse video. IEEE/ACM Trans. Comput. Biol. Bioinform. 12 (6), 1470–1478.

Feng, L., Chen, S., Zhang, C., Zhang, Y., He, Y., 2021. A comprehensive review on recent applications of unmanned aerial vehicle remote sensing with various sensors for high-throughput plant phenotyping. Comput. Electron. Agric. 182, 106033.

Fu, L., Gao, F., Wu, J., Li, R., Karkee, M., Zhang, Q., 2020. Application of consumer RGB-D cameras for fruit detection and localization in field: A critical review. Comput. Electron. Agric. 177, 105687.

Furbank, R.T., Tester, M., 2011. Phenomics–technologies to relieve the phenotyping bottleneck. Trends Plant Sci. 16 (12), 635–644.

Guo, R., Xie, J., Zhu, J., Cheng, R., Zhang, Y., Zhang, X., Gong, X., Zhang, R., Wang, H., Meng, F., 2023. Improved 3D point cloud segmentation for accurate phenotypic analysis of cabbage plants using deep learning and clustering algorithms. Comput. Electron. Agric. 211, 108014.

Hedman, P., Srinivasan, P.P., Mildenhall, B., Barron, J.T., Debevec, P., 2021. Baking neural radiance fields for real-time view synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5875–5884.

Humplík, J.F., Lazár, D., Husičková, A., Spíchal, L., 2015. Automated phenotyping of plant shoots using imaging methods for analysis of plant stress responses–a review. Plant Methods 11 (1), 1–10.

Jansen, M., Gilmer, F., Biskup, B., Nagel, K.A., Rascher, U., Fischbach, A., Briem, S., Dreissen, G., Tittmann, S., Braun, S., et al., 2009. Simultaneous phenotyping of leaf growth and chlorophyll fluorescence via GROWSCREEN FLUORO allows detection of stress tolerance in Arabidopsis thaliana and other rosette plants. Funct. Plant Biol. 36 (11), 902–914.

Kang, H., Chen, C., 2020. Fast implementation of real-time fruit detection in apple orchards using deep learning. Comput. Electron. Agric. 168, 105108.

Kang, H., Wang, X., 2023. Semantic segmentation of fruits on multi-sensor fused data in natural orchards. Comput. Electron. Agric. 204, 107569.

Kang, H., Wang, X., Chen, C., 2022a. Accurate fruit localisation for robotic harvesting using high resolution lidar-camera fusion. arXiv preprint arXiv:2205.00404.

Kang, H., Zang, Y., Wang, X., Chen, Y., 2022b. Uncertainty-driven spiral trajectory for robotic peg-in-hole assembly. IEEE Robot. Autom. Lett. 7 (3), 6661–6668.

Kok, E., Wang, X., Chen, C., 2023. Obscured tree branches segmentation and 3D reconstruction using deep learning and geometrical constraints. Comput. Electron. Agric. 210, 107884.

Kolhar, S., Jagtap, J., 2023. Plant trait estimation and classification studies in plant phenotyping using machine vision–A review. Inf. Process. Agric. 10 (1), 114–135.

Kumar, J.P., Domnic, S., 2019. Image based leaf segmentation and counting in rosette plants. Inf. Process. Agricult. 6 (2), 233–246.

Lévy, B., Petitjean, S., Ray, N., Maillot, J., 2023. Least squares conformal maps for automatic texture atlas generation. In: Seminal Graphics Papers: Pushing the Boundaries, Volume 2. pp. 193–202.

Li, Z., Guo, R., Li, M., Chen, Y., Li, G., 2020. A review of computer vision technologies for plant phenotyping. Comput. Electron. Agric. 176, 105672.

Lorensen, W.E., Cline, H.E., 1998. Marching cubes: A high resolution 3D surface construction algorithm. In: Seminal Graphics: Pioneering Efforts that Shaped the Field. pp. 347–353.

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

Mildenhall, B., Srinivasan, P.P., Ortiz-Cayon, R., Kalantari, N.K., Ramamoorthi, R., Ng, R., Kar, A., 2019. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. ACM Trans. Graph. 38 (4), 1–14.

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. Commun. ACM 65 (1), 99–106.

Müller, T., Evans, A., Schied, C., Keller, A., 2022. Instant neural graphics primitives with a multiresolution hash encoding. ACM Trans. Graph. (ToG) 41 (4), 1–15.

Park, J.J., Florence, P., Straub, J., Newcombe, R., Lovegrove, S., 2019. Deepsdf: Learning continuous signed distance functions for shape representation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 165–174.

Paulus, S., 2019. Measuring crops in 3D: using geometry for plant phenotyping. Plant Methods 15 (1), 1–13.

Rebetzke, G., Jimenez-Berni, J., Fischer, R., Deery, D., Smith, D., 2019. High-throughput phenotyping to enhance the use of crop genetic resources. Plant Sci. 282, 40–48.

Roessle, B., Barron, J.T., Mildenhall, B., Srinivasan, P.P., Nieß ner, M., 2022. Dense depth priors for neural radiance fields from sparse input views. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12892–12901.

Samavati, T., Soryani, M., 2023. Deep learning-based 3D reconstruction: A survey. Artif. Intell. Rev. 1–45.

Sander, P.V., Snyder, J., Gortler, S.J., Hoppe, H., 2001. Texture mapping progressive meshes. In: Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques. pp. 409–416.

Schonberger, J.L., Frahm, J.-M., 2016. Structure-from-motion revisited. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4104–4113.

Sishodia, R.P., Ray, R.L., Singh, S.K., 2020. Applications of remote sensing in precision agriculture: A review. Remote Sens. 12 (19), 3136.

Ubbens, J., Cieslak, M., Prusinkiewicz, P., Stavness, I., 2018. The use of plant models in deep learning: an application to leaf counting in rosette plants. Plant Methods 14, 1–10.

Wang, Y., Fan, J., Yu, S., Cai, S., Guo, X., Zhao, C., 2023. Research advance in phenotype detection robots for agriculture and forestry. Int. J. Agric. Biol. Eng. 16 (1), 14–25.

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.

Wu, S., Wen, W., Wang, Y., Fan, J., Wang, C., Gou, W., Guo, X., 2020. MVS-pheno: a portable and low-cost phenotyping platform for maize shoots using multiview stereo 3D reconstruction. Plant Phenomics 2020.

Yang, T., Ye, J., Zhou, S., Xu, A., Yin, J., 2022. 3D reconstruction method for tree seedlings based on point cloud self-registration. Comput. Electron. Agric. 200, 107210.

Zhang, Y., Zhang, N., 2018. Imaging technologies for plant high-throughput phenotyping: a review. Front. Agric. Sci. Eng. 5 (4), 406–419.

Zhao, F., Jiang, Y., Yao, K., Zhang, J., Wang, L., Dai, H., Zhong, Y., Zhang, Y., Wu, M., Xu, L., et al., 2022. Human performance modeling and rendering via neural animated mesh. ACM Trans. Graph. 41 (6), 1–17.

Zhao, G., Yang, R., Jing, X., Zhang, H., Wu, Z., Sun, X., Jiang, H., Li, R., Wei, X., Fountas, S., et al., 2023. Phenotyping of individual apple tree in modern orchard with novel smartphone-based heterogeneous binocular vision and YOLOv5s. Comput. Electron. Agric. 209, 107814.

Zhou, H., Wang, X., Au, W., Kang, H., Chen, C., 2022. Intelligent robots for fruit harvesting: Recent developments and future challenges. Precis. Agric. 23 (5), 1856–1907.