IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Tomato-Nerf: Advancing Tomato Model Reconstruction with Improved Neural Radiance Fields

Xiajun Zheng[1], Xinyi Ai[1], Hao Qin[1], Jiacheng Rong[1], Zhiqin Zhang[1], Yan Yang[1], Ting Yuan[1], Wei Li[1]

[1]College of Engineering, China Agricultural University, Beijing 100083, China

Corresponding author: Ting Yuan E-mail address: yuanting122@cau.edu.cn
Wei Li E-mail address: 2337485740@qq.com

This work was supported in National Key Research and Development Program of China, NK202315020107.

ABSTRACT The real-time simulation of large-scale agricultural operations will offer farmers data-driven and physically consistent decision support, facilitated by predictive digital twins. To construct a predictive digital twin, the initial step involves 3D reconstruction of plant geometry. In this paper, a high-resolution, accurate 3D reconstruction of tomato plants, Tomato-NeRF, is proposed, which is specially used for three-dimensional reconstruction of tomato plants. Our approach used a modular design to integrate ideas from their research paper into Tomato-NeRF. By using hash encoding to map coordinates to trainable feature vectors, we balance quality, memory usage, and performance in NeRF training. The proposal sampler targets key regions for rendering, and customized loss functions are designed to optimize specific tasks. The effectiveness of our approach is demonstrated by the ability to generate high-resolution geometric models from phone camera data. Comparative results show that Tomato-NeRF has significant advantages over Instant-NGP and MipNeRF in the tomato plant reconstruction task. The data acquisition method is simpler and more efficient than other reconstruction methods, providing a practical solution for real-time agricultural simulations.

INDEX TERMS 3D reconstruction, Agricultural automation, Deep learning, NeRF, Multi-view imaging

I. INTRODUCTION

Real-time simulation and predictive digital twins have become important tools to support decision-making in large-scale agricultural operations[1-6]. These technologies provide farmers with data-driven and physically consistent insights to optimize farming practices and increase productivity. A key aspect of building predictive digital twins is the accurate three-dimensional reconstruction of plant geometry. Accurate 3D reconstruction of tomato plants offers several significant benefits[7-13]. First, it can simulate large-scale agricultural operations in real time, providing farmers with data-driven decision support and physically consistent models. Second, three-dimensional reconstruction of tomato plants provides valuable visual analysis and understanding. By visualizing and analyzing reconstructed tomato plant models, researchers and agricultural experts can gain insight into the plant's morphology, structure, and geometry. In addition, the reconstructed 3D model helps in agricultural research and optimization. Scientists can use these models to study the effects of growing environment, light conditions and climate factors on tomato growth and yield.

However, traditional methods of geometric reconstruction of plants often involve expensive setups using LiDAR or destructive imaging methods in controlled environments[14]. These approaches present challenges in terms of cost, data collection, and scalability. With the development of computer vision, researchers have recognized the great potential of 3D reconstruction methods and have proposed many related algorithms. Among these methods, the multi-view reconstruction technique only needs to use the camera to capture 2D images of the plant scene to reconstruct the 3D structure. Compared with precision instrument measurement methods, these technologies have lower equipment acquisition cost, easy operation and wide applicability. As a result, they have received a great deal of attention from researchers and have been studied extensively. For example, Rose et al.[15] reconstructed the three-dimensional structure of tomato canopy using open source software. They further analyzed the correlation between the obtained phenotypic parameters and those obtained by the laser scanner. This method allowed them to study the relationship between visual reconstruction and the data obtained through laser scanning. COLMAP[16] is a multifunctional pipeline that combines structure-from-motion (SfM) and multi-view Stereo (MVS) algorithms. It provides various functions for reconstructing ordered and unordered image collections. COLMAP provides a graphical and command line interface that makes it accessible and flexible for multi-view reconstruction. Dandrifosse et al.[17] proposed a method to reconstruct the three-dimensional structure of wheat canopy using multi-resolution images. Zhu et al.[18] proposed a method based on low-cost 3D reconstruction technology to analyze phenotypic development of soybean plants throughout their growth period. Multi-view images were constructed by digital cameras from different angles, image features were extracted by SURF algorithm, and stereoscopic feature point matching was realized by RANSAC algorithm. The multi-view reconstruction method based on stereo matching needs to shoot images from different angles at the same time, which is a tedious process of data acquisition and needs to deal with algorithm details such as image matching.

However, recent advances in 3D computer vision, particularly Neural radiation Fields (NeRF), have revolutionized 3D scene reconstruction[19]. NeRF and its variants have recently been explored using volume rendering methods to learn neural radiation fields for new view synthesis[20]. NeRF allows the generation of detailed and realistic visualizations that approach the quality of actual photos. Its introduction opens up new possibilities for capturing and reconstructing 3D scenes with impressive visual fidelity[21]. As with traditional multi-view 3D reconstruction methods, NeRF takes 2 images captured from different viewpoints as input data. However, NeRF takes a novel approach, directly utilizing MLP to learn the characteristics of a scene. It implicitly stores scene information in a neural network and uses volume rendering technology to generate scene images from multiple viewpoints. This implicit representation allows NeRF to capture fine detail and achieve high quality rendering. It breaks through the limitation of traditional explicit geometric model in multi-view reconstruction. When NeRF was first introduced, its main application was synthesizing new views, with a focus on ensuring visual consistency between the generated image and the reference image. However, the limitation of NeRF is that its relatively limited focus is on precisely reconstructing the geometry of the object being seen, where precision is Paramount. Therefore, the original NeRF method was unable to meet the strict requirements of geometric reconstruction accuracy required for plant reconstruction. In addition, NeRF's mlp structure is very expensive to train and evaluate, each job is designed for its own specific task, different network structure, training a NeRF model can take several hours. Subsequent advances, such as nvidia's proposed instant-NGP implementation of "5s training a NeRF"[22] received researchers' attention; VolSDF[23], NeuS[24], and Geo-NeuS[25], address these limitations by combining signature distance function (SDF) representations to describe the surface of a reconstructed object. Müller et al. presented Instant-NGP, which realizes end-to-end 3D scene modeling and rendering directly from RGB images. Key technological innovations include the integration of multi-resolution Hash coding to improve rendering efficiency, and the use of self-attention mechanisms to enhance network structure. Instant-NGP achieved faster results and higher reconstruction quality than other advanced neurorendering technologies of the time. Xu et al.[26] proposed a point-NeRF based NeRF model, using MVSnet to get the initial neural Point cloud scene, and then using the neural point cloud and neural features to build a point-based radiation field rendering scene. These improved methods complement the volume rendering techniques used in NeRF, improve the accuracy of fitting the surface of the constructed object, and increase the speed of NeRF training. As a result, these advances significantly improve the ability of models to learn to reconstruct the geometry of objects. These new implicit surface reconstruction methods can effectively meet the requirement of reconstruction accuracy for complex 3D reconstruction tasks, including plant phenotype construction.

In this work, we propose a high-resolution, accurate 3D reconstruction of tomato plants to support data-driven agricultural decision-making, called Tomato-Nerf. We demonstrate the effectiveness of our approach by presenting a high-resolution geometric model obtained from mobile phone camera data. In addition, we discuss the potential applications of these reconstructions, particularly in the development of predictive digital twins for tomato plants, which enable real-time decision support and optimization for tomato cultivation. Overall, our research contributes to the advancement of precision agriculture by introducing innovative methods for three-dimensional reconstruction of tomato plants. The use of neural radiation fields for accurate and efficient plant geometric reconstruction has great potential for real-time simulation, predictive digital twins, and data-driven decision-making in agriculture.

II. Data Acquisition and Ground Truth Comparison

The focus of this study is **on Israeli red tomato grown** in the intelligent greenhouse of Beijing Hongfu Group, under standardized planting conditions that result in high yield and quality as shown in Fig. 1.

In order to train NeRF models and obtain high-fidelity tomato plant meshes, we collected multiple sets of tomato plant photos in the form of 2D images. We filmed video clips of tomato plants using a smartphone camera. Specifically, we used an iPhone 13 Pro with a resolution of 4K and frame rate of 30fps. The iPhone 13 Pro's main camera is equipped with a 12-megapixel sensor featuring 1.9-micron pixels, paired with a 26mm equivalent f/1.5 aperture lens. This setup provides enhanced low-light performance and improved image quality. The camera was positioned at a constant height while circling around the plants to ensure comprehensive coverage. Simple frame extraction was performed on the recorded videos to obtain 2D images of the scene. Every 5th frame was extracted, resulting in high resolution 4K images being obtained from each video set of a tomato plant. We further processed these images using the COLMAP library, which is a powerful computer vision tool for reconstructing 3D scenes and generating dense point clouds, mesh reconstructions, and camera pose estimations by leveraging a large collection of 2D images to recover the geometric structure of a 3D scene. Its core capabilities are structure-from-motion (SfM) and multiview geometry. SIFT and other algorithms are used to detect and describe the local features of all images, and the feature points between different images are matched to obtain the feature matching relationship. Based on robust matching results, the pose of each image is preliminarily estimated by PnP algorithm.Add more images iteratively, optimize camera parameters and 3D point position through bundle adjustment, and get sparse point cloud and accurate camera pose. MVS algorithm is used to match images in stereo and generate dense point cloud. The camera parameters and images of COLMAP output are saved in json format required by NeRF as model training data. With the camera poses and RGB values of each pixel obtained from the images, we assembled a dataset suitable for training our NeRF model. The dataset includes spatial location (x, y, z) and viewing direction (θ, φ) for each image, along with corresponding RGB values.

The entire data collection process, from video recording to dataset creation, took approximately 4 hours. To evaluate the fidelity of the NeRF rendered mesh, lidar scan data of a synthetic tomato plant was also acquired as ground truth reference. At the beginning of this study, XX, YY and other structured light 3D scanners were used to obtain tomato point clouds. However, the experimental results found that due to the soft tissue light transmission of tomatoes and high surface gloss, high-quality 3D reconstruction results could not be obtained using traditional scanners. The scanning point cloud has serious fragmentation, data omission and noise. To solve this problem, the Freescan Combo 3D scanner was used for data acquisition, as shown in Fig. 2. The device integrates structured light and single-point laser ranging technology, scanning speeds of up to 2 million points per second. Experiments show that the whole three-dimensional point cloud of tomato can be obtained quickly with very high precision by using this device.

Each LiDAR scan was completed at 1/4 and 1x resolution. Due to occlusion between individual tomato fruits, multiple scans were necessary to capture the complete structure. After scanning the overall structure, we performed targeted scans of individual fruits and aligned the point clouds from multiple scans to generate a complete 3D point cloud of the plant. To ensure data integrity, we used the 3D point cloud processing software CloudCompare and applied a Statistical Outlier Removal (SOR) filter with 50 nearest neighbors to eliminate duplicate points.

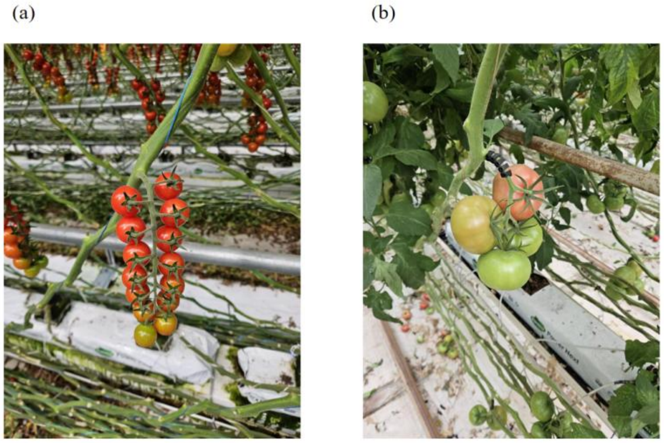


FIGURE 1. **Greenhouse growing environment for tomatoes**

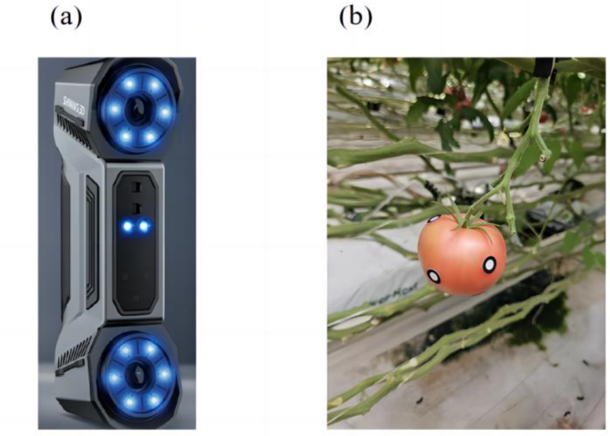


FIGURE 2. The Scanner (a) and the sample of scan tomatoes (b)

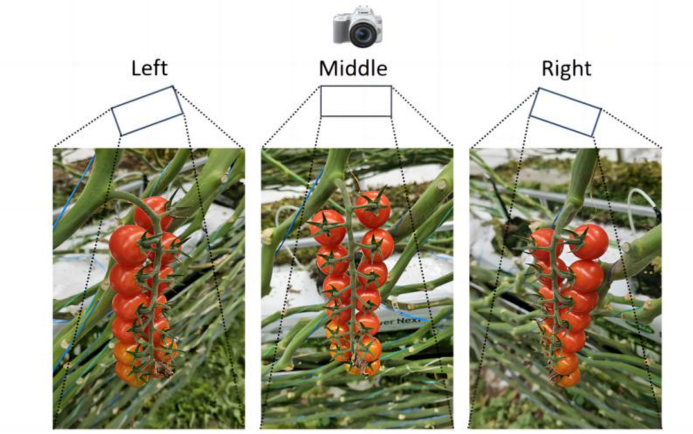


FIGURE 3. The data collection process

By comparing the NeRF rendered mesh with the lidar exported point cloud, we established a reliable ground truth dataset to evaluate the accuracy and fidelity of our reconstructed tomato plant model. The combined dataset and lidar ground truth data laid the foundation for training and evaluating our tomato plant reconstruction method based on NeRF.

III. METHOD

NeRF(Neural Radiance Fields) is a novel method for representing 3D scenes based on neural networks. The core idea is to use Multi-Layer Perceptron (MLP) to learn the color value (RGB) and Density of each spatial position in the scene, and construct a continuous five-dimensional radiation Field function.

Specifically, the NeRF system inputs are the coordinates (x,y,z) of any three-dimensional space position and the Viewing Direction of the camera, and the color (RGB) and Density of that position are mapped by the neural network. During the rendering process, multiple sample points are uniformly sampled on the Camera Ray, and the (x,y,z) coordinates and viewing direction of each point are entered into the MLP to obtain the color and density. Based on these discrete samples, Volume Rendering integrals are then applied to calculate the RGB value of the light, that is, the color of the pixels in the image. The training of NeRF requires multi-view image data and corresponding camera parameters such as internal and external parameters. By optimizing the network parameters to fit the image prediction and ground truth, the model can learn the three-dimensional geometry and reflection characteristics of the scene. NeRF is trained to render never-before-seen perspectives in high quality and can be used in a variety of 3D modeling and rendering tasks.

Compared with traditional grid-based or voxel-based 3D reconstruction and rendering technologies, NeRF can directly learn continuous 3D scene representation from 2D image data alone, and realize efficient 3D modeling and illumination synthesis through deep learning. Moreover, Mip-NeRF proposed the idea of using cone sampling to better deal with anti-aliasing effects[27, 28]. It uses hierarchical coding to improve the network's understanding of sampling range size. This method can generate a more continuous and smooth image. Instant-NGP is another influential general-purpose neural renderer that can reconstruct complex 3D scenes directly from RGB images. Its technological innovations include the use of multi-resolution Hash coding to improve efficiency and self-attention mechanisms to enhance the network. This method has fast rendering speed and high quality reconstruction effect.

This paper presents Tomato-NeRF model, and the core idea is to use multi-layer perceptrons (MLPS) to learn detailed information about the color (RGB) and density of spatial locations in each 3D scene, as shown in Fig. 4 and Fig. 5. The input includes the coordinates (x, y, z) of any 3D space point and the camera's viewing direction. The neural network does this by mapping these inputs to corresponding color and density values. The rendering process involves uniformly selecting multiple sample points on the camera rays, entering their coordinates and viewing Angle directions into the MLP for color and density prediction, and finally obtaining the RGB value of the synthesized image through volume rendering integration. The Tomato-NeRF model uses the hash coding technique, which allows for memory-efficient high-resolution neural rendering. Hash coding maps input 3D coordinates to high-dimensional random vectors through hash functions, enhancing the ability of the network to represent complex 3D scenes. It improves the sensitivity of the network to coordinate information, and makes use of the local invariance of the hash coding space to produce more generalized coordinate features.

Additionally, Tomato-NeRF utilizes a proposal sampler strategy, focusing on critical visible surface areas to enhance reconstruction quality. Sampling is based on the scene density function, using the Hash MLP network for a balance between speed and quality. Concatenating multiple density functions enables cascading sampling and rendering for a finer distribution. The core ideas include using an auxiliary network for predicting sampling density, sampling according to the predicted distribution, using the main network for color and density prediction, and compositing perspectives through volume rendering.

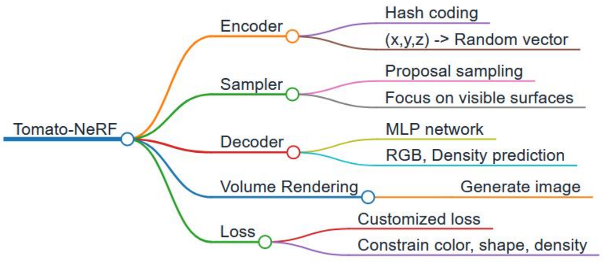


Fig. 4. **Structure diagram of tomato-NeRF**

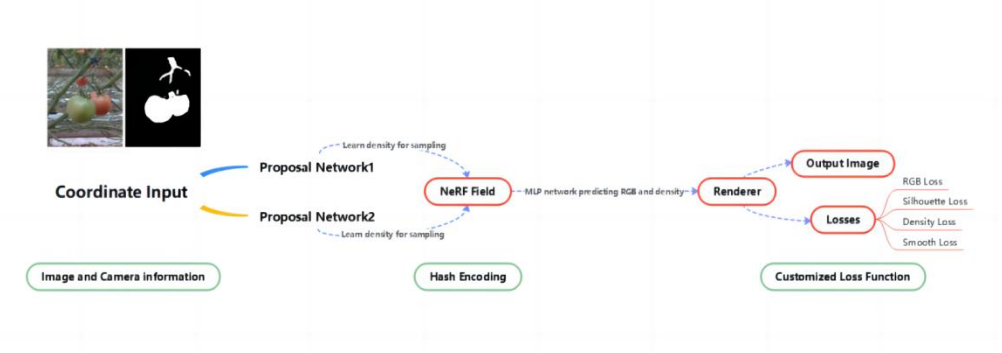


Fig. 5. **The flow chart of tomato-NeRF**

In terms of the loss function design, Tomato-NeRF customizes it by enhancing RGB loss weight for improved color, incorporating Silhouette Loss for shelter information, retaining Density MSE Loss for geometry fitting, and adding a Smoothness regular term. These constraints ensure key attributes like color, shape, and density are maintained while simplifying the training process by removing less impactful loss items for tomatoes.

A. Hash Encoding

Tomato-NeRF was inspired by Instant-NGP design, using a memory-efficient neural rendering scheme based on hash coding. Hash coding enhances the ability of neural networks to represent complex 3D scene structures by mapping input 3D coordinates to high-dimensional random vectors through hash functions[22]. This technique improves the sensitivity of the model to the coordinate input and uses the local invariance of the hash coordinate space to produce more generalized coordinate features, as shown in Fig. 6.

The spatial coordinate input (x,y,z) is encoded by a hash function into a high-dimensional random vector. These encoded vectors are fed into the multi-layer perceptron (MLP) network together with the perspective direction vector. The MLP network learns the mapping of coordinates to the corresponding RGB color and density values. Multiple spatial locations are sampled along each camera light and mapped through this encoding and MLP pipeline. Finally, volume rendering integrates the MLP output along the light to synthesize a new perspective of the image.

The function of the hash function is to encode the spatial coordinates into a high-dimensional random potential space, so that the adjacent positions map to similar code vectors and inject positional priors. Under the conditions of this encoding,MLP networks model complex mappings from coordinates to color and density values to represent 3D scenes. Learning from the coding space facilitates fitting more complex and high-frequency functions rather than direct coordinate input.

The trainable features are the F-dimensional vectors and arranged into an L-grid, where L represents the number of resolution features and T represents the number of feature vectors in each hash grid. Hash grid coding steps are as follows: Given the input coordinates, find the surrounding voxels at the L-resolution level and hash the vertices of these grids; Using hash vertices as keys to find trainable F-dimensional eigenvectors; According to the position of the coordinates in space, the eigenvector of the input coordinates is matched by linear interpolation; The feature vectors in each grid are associated with any other parameters, such as Angle of view direction, illumination, etc; The final vector is fed into the neural network to predict the RGB and density outputs. This encoding structure makes a trade-off between quality, memory, and performance. The main parameters that can be adjusted are the size of the hash table (T), the size of the feature vector (F), and the number of resolutions (L).

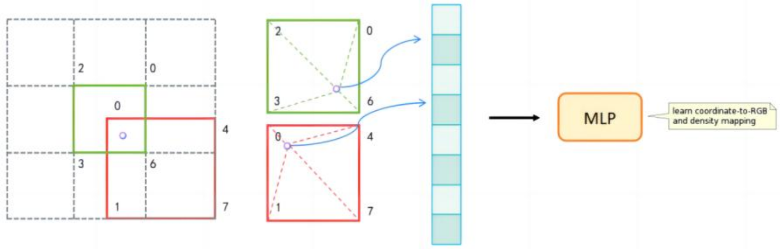


Fig. 6. **The illustration of the hash encoding**

B. Proposal Sampler

The proposal sampler scheme strategically concentrates samples on the scene region that holds the utmost significance in shaping the final rendering effect, typically the foremost visible surface[29]. This targeted sampling approach markedly enhances the overall quality of the reconstruction process. The introduced learning sampling method relies on the density function of the scene and boasts versatility in its implementation.

The results show that the compact multi-layer perceptron (MLP) using hash coding achieves an effective balance between accuracy and efficiency, and verifies its effectiveness in optimizing sampling. For further improvement, a significant enhancement connects multiple density functions, facilitating more detailed sampling and thus improving reconstruction accuracy. This augmentation contributes to a more nuanced and accurate sampling strategy, further elevating the precision of the reconstruction outcomes.

In essence, the proposed samplers include:

An auxiliary network that predicts the sampling density per location from coordinates;

Random sampling based on the predicted density coordinates has more samples in the high-density region;

Feed the sample into the main network to predict color and density values;

The predicted volume of the sample is fused to present the final image;

Combined optimization of end-to-end density prediction and rendering network enables learning samples to be concentrated in significant areas.

In summary, the method strategically assigns computations to key scene regions by predicting sample density and volume fusion samples of concentrated areas. The joint training process allows for a gradual concentration of samples in the most important places, thereby improving the quality of reconstruction.

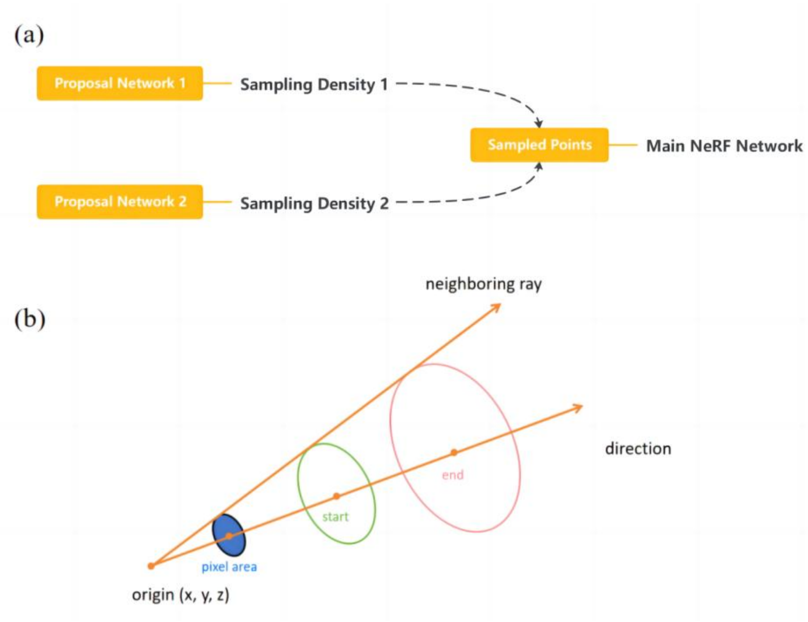


Fig. 7. The proposal network structure diagram(a); Sample representations(b)

C. Loss Function

Within the framework of the Tomato-NeRF model, we have meticulously devised a loss function tailored to the specific intricacies of tomato reconstruction. The formulation of this loss function adheres to two fundamental principles: firstly, a deliberate emphasis on optimizing pivotal color and shape information; secondly, a strategic removal of extraneous elements to streamline and simplify the training process.In particular, our approach incorporates an RGB Loss with augmented weights, strategically implemented to intensify color fidelity. Additionally, the inclusion of Silhouette Loss, in the form of binary cross entropy, introduces crucial tomato shading information. Concurrently, we preserve the Mean Squared Error (MSE) Loss pertaining to predicted density and truth-sampling density to enforce constraints on scene geometry information.

In contrast to the general NeRF framework, certain loss components, such as those pertaining to normal direction and sample consistency, have been judiciously excluded due to their limited discernible impact on the training outcomes. These simplifications have proven instrumental in significantly accelerating the training speed while channeling the model's expressiveness towards the paramount aspects of color and shape, which are deemed more critical for the nuanced task of tomato reconstruction.

The amalgamation of these diverse loss types within Tomato-NeRF ensures a comprehensive constraint on color, shape, and density. The incorporation of a consistent prior further augments the model's capacity to generate high-quality tomato reconstruction results, underscoring.

$$L = w_1 * L_{rgb} + w_2 * L_{sil} + w_3 * L_{den} + w_4 * L_{smooth} \quad (1)$$

$$L_{rgb} = \frac{1}{N}\sum_{i=1}^{N} || I_i^{pred} - I_i^{gt} ||^2 \quad (2)$$

$$L_{sil} = \frac{1}{M}\sum_{j=1}^{M} || S_j^{pred} - S_j^{gt} ||^2 \quad (3)$$

$$L_{den} = \frac{1}{K}\sum_{k=1}^{K} (D_k^{pred} - D_k^{gt})^2 \quad (4)$$

$$L_{smooth} = \frac{1}{K}\sum_{k=1}^{K} || \nabla D_k^{pred} ||_2 \quad (5)$$

With this customized Loss design, Tomato-NeRF can efficiently produce high-quality tomato 3D reconstruction results. Our exploration validates the importance and effectiveness of NeRF model Loss design for specific scenarios.

IV. Evaluation and discussion

Unlike object detection algorithms which use recall, precision and mAP@0.5:0.95 as evaluation metrics, neural radiance fields (NeRF) models employ Peak Signal-to-Noise Ratio (PSNR), SSIM and LPIPS as assessment criteria[30]. PSNR, SSIM, and LPIPS are commonly used image quality assessment metrics, each with its own advantages and limitations: PSNR is the ratio of maximum signal power to noise power, expressed in dB. The calculation of PSNR requires both the original image before noise destruction and the image after noise destruction. The mean square error (MSE) of the noise image is used to calculate the noise power component. For grayscale images with dimensions of m x n pixels, MSE is calculated as follows:

$$MSE = \frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}[I(i,j) - K(i,j)]^2 \quad (6)$$

$$PSNR = 10 \cdot \log_{10}(\frac{MAX_I^2}{MSE}) \quad (7)$$

$$PSNR(dB) = 20\log_{10}(MAX_I) - 10\log_{10}(MSE) \quad (8)$$

By combining brightness, contrast and structure comparison, SSIM is closer to subjective evaluation. It better represents the sensitivity of the human visual system to structural information. However, SSIM involves more complex calculations than PSNR. SSIM is a measure of structural similarity between two images based on brightness, contrast and structure of the image. In contrast to PSNR, this indicator takes into account the specifics of human visual perception i and provides a quality assessment function closer to that of the average person. It is calculated by comparing local patches (rectangular Windows) of pixels in the image. Its calculation formula is as follows:

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \quad (9)$$

Where, x and y are two image blocks of size N × N, μ_x and μ_y are the average pixel values of x and y, σ_x and σ_y are the variances of x and y, σ_{xy} is the covariance of x and y, and c_1 and c_2 are two constants to avoid instability when the denominator approaches zero. The value of the SSIM ranges from -1 to 1. 1 indicates that the SSIM is completely similar, and -1 indicates that the SSIM is completely different. LPIPS extracts deep features using a pretrained VGG network and computes distances in feature space to evaluate perceptual similarity. By leveraging deep learning, LPIPS further improves consistency with human perception. However, it requires a pretrained CNN, making it more computationally expensive. Lower LPIPS values denote higher perceptual similarity and better quality.

A. Model Training

An experiment was conducted in the intelligent greenhouse of Beijing Hongfu Group.To validate our proposed model, we used an accurate 3D scanning system to capture ground-based true point cloud data for 5 groups of tomato samples and 5 groups of cherry tomato samples. We then recorded video footage of the tomato scene by bypassing the sample and following the data collection protocol described in our method. The video data is preprocessed and provided as input to the Tomato-NeRF model for training and inference. In addition, we constructed a multi-scale dataset following the approach of Mip-NeRF to examine the model's capability for anti-aliasing and multi-scale reasoning. This dataset was generated by downsampling each image in the original dataset by factors of 2, 4, and 8, while adjusting the intrinsic camera parameters accordingly. These downsampled images were combined with the original high-resolution images to form the multi-scale dataset. Based on the principles of projective geometry, this simulates rendering the scene with the camera positioned at 2, 4 and 8 times the original distance. There are 720 original images and they are downsampled by factors of 2, 4, and 8. 7/8 of the images are used for training, and the remaining 1/8 of the images are uniformly spaced and used for evaluation. The model training device is an HP laptop with 16GB RAM, Intel(R) Core(TM) i7-10870H CPU @ 2.20GHz, NVIDIA GeForce RTX 2070 (8G). During training on this dataset, we weighted the loss contribution of each pixel by the area it occupied in the original high-resolution image. This balances the impact of the few low-resolution pixels and the many high-resolution pixels. Quantitative evaluation was performed by averaging each error metric across all four scales. To benchmark the performance of Tomato-NeRF, we performed comparative experiments with the universal NeRF-rendering systems Mip-NeRF and Instant-NGP using the same dataset.

B. Evaluation

As shown in Fig. 9, the NeRF model outputs the RGB color and density value of each pixel, where the density value contains the depth information. By combining these predicted density values with known camera parameters, point clouds representing the 3D scene geometry can be acquired through depth interpolation on novel view RGB images.

The evaluation of the point cloud predicted by Tomato-NeRF uses the CloudCompare open source software. The real point cloud and Tomato-NeRF predicted point cloud were imported for comprehensive comparative analysis. In order to improve the accuracy, both the predicted point cloud and the real point cloud are normalized.Then, the quality of the predicted point cloud is evaluated by calculating the average distance and standard deviation between the two sets of point clouds. A quantitative point cloud analysis is then performed to measure the accuracy of the Tomato-NeRF reconstruction compared to ground-based real data.

Specifically, we utilize CloudCompare software to visualize the point cloud reconstruction results of different methods.Through the color mapping reconstruction error code for the RGB values, green said error is low, the orange and red error is higher. It can be observed that the vast majority of the Tomato-NeRF point cloud appears dark green, indicating that its reconstruction fidelity is significantly higher than other methods, and it has the ability to retain complex details. This is consistent with the quantitative experiments presented in the paper. Through a visual examination of Fig. 10 a unique pattern emerges that highlights Tomato-NeRF's unique advantage in point cloud representation. The alternative method shows large orange-red areas in the rendered point cloud, emphasizing a large deviation from ground reality. In contrast, the point cloud generated by Tomato-NeRF is mostly green, indicating a smaller error.

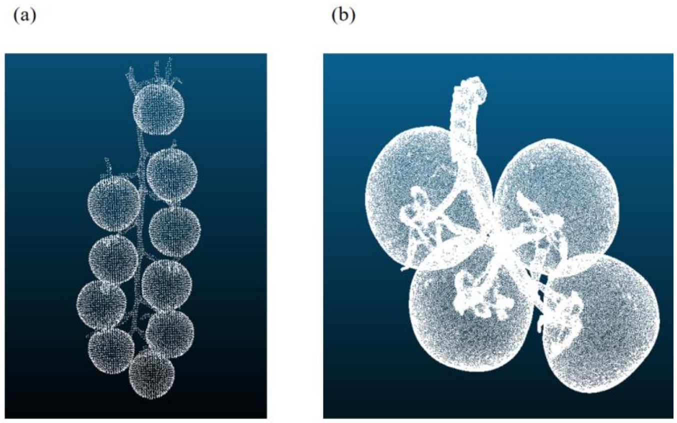


Fig. 8. **A real tomato model obtained by the scanner**

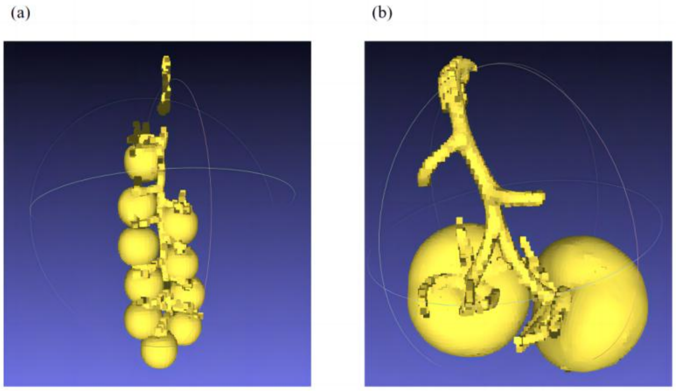


Fig. 9. **The plant point cloud extracted from the predicted model**

To sum up, the point cloud visualization in Fig. 10 provides intuitive visual evidence, which confirms the excellence of Tomato-NeRF in 3D scene representation and detail preservation, and further validates the accuracy advantage reflected by the quantitative indicators in this study.

The quantitative evaluation results are shown in Table 1. The average distance and standard deviation between the predicted point cloud generated by Tomato-NeRF and the reference LiDAR scan are 5.2mm and 2.1mm, respectively. These findings show that Tomato-NeRF has remarkable similarity and accuracy in the modeling of Tomato's intrinsic complex geometric structure. In contrast, Tomato-NeRF showed substantial improvement in quantitative results compared to the baseline methods Instant-NGP and Mip-NeRF. The mean distances of Instant-NGP and Mip-NeRF are 15.7mm and 10.3mm, respectively, and the standard deviations are 5.2mm and 3.1mm, respectively. This highlights tomato - NeRF's outstanding performance in the Tomato reconstruction task. The lower mean distance and standard deviation values affirm Tomato-NeRF's accuracy in faithfully capturing the three-dimensional complexity of the Tomato. Similarly, for cherry tomatoes, our Tomato-NeRF method obtains an average distance of 4.8mm and a standard deviation of 1.9mm. This confirms the power of the method in modeling and reconstructing the complex geometry of cherry tomatoes. Compared with Instant-NGP(mean distance 14.2mm, standard deviation 4.7mm) and Mip-NeRF(mean distance 9.5mm, standard deviation 2.8mm), Tomato-NeRF achieves significant improvement in quantitative indicators, thus reflecting its excellent reconstruction performance in cherry tomato scenes. Compared to the baseline method, the quantitative evaluation of Tomato-NeRF emphasizes its ability to generate highly accurate and detailed geometric representations of tomatoes. The normalization process and rigorous analysis help improve the accuracy of Tomato-NeRF, making it an invaluable tool for powerful and reliable 3D reconstruction of complex organic structures.

In addition, we computed the PSNR, SSIM and LPIPS between Tomato-NeRF rendered images and ground truth images. As shown in Table 2, Tomato-NeRF achieved a PSNR of 27.55, SSIM of 0.894 and LPIPS of 0.148. In comparison, Instant-NGP obtained a PSNR of 26.54, SSIM of 0.849 and LPIPS of 0.182, while Mip-NeRF scored 27.28 in PSNR, 0.889 in SSIM and 0.176 in LPIPS. Tomato-NeRF outperformed the general-purpose frameworks across all metrics. Across all metrics, Tomato-NeRF outperformed the general-purpose frameworks.Similarly, for cherry Tomato scenes, our Tomato-NeRF method also achieves excellent image reconstruction quality. It achieved a PSNR of 28.12, SSIM of 0.912, and LPIPS of 0.124, significantly outpacing Instant-NGP(PSNR 26.87, SSIM 0.883, LPIPS 0.159) and Mip-NeRF(PSNR 27.92, SSIM 0.903, LPIPS 0.149) performance. This is a testament to the Tomato-NeRF framework's excellent rendering ability in complex cherry tomato scenes. Considering many target indexes such as PSNR, SSIM and LPIPS, our method is superior to the existing general framework and has reached a new level of image quality in the reconstruction of cherry tomato. The results demonstrate that Tomato-NeRF can also realistically render the surface textures and color details to produce novel views closer to the ground truth images. This further verifies both the reconstruction and rendering capacities of Tomato-NeRF.

The biggest practical trade-offs between these methods are time versus space. Tomato-NeRF achieved remarkable accuracy in modeling complex geometric structures of tomatoes, with an average distance of 5.2mm and a standard deviation of 2.1mm between the predicted point cloud and the reference LiDAR scan. This indicates a high degree of fidelity in the 3D reconstruction process. Tomato-NeRF is a substantial improvement over the baseline methods Instant-NGP and Mip-NeRF. For example, in the reconstruction of tomatoes, Tomato-NeRF obtained an average distance of 4.8mm and a standard deviation of 1.9mm, significantly outperforming Instant-NGP and Mip-NeRF. The study emphasizes the ability of Tomato-NeRF to generate highly accurate and detailed geometric representations of tomatoes. This is crucial for applications that require accurate 3D models, such as agricultural monitoring and robotic harvesting. The evaluation of PSNR, SSIM, and LPIPS metrics further highlights Tomato-NeRF's superior image reconstruction quality. For cherry tomato scenes, Tomato-NeRF achieved a PSNR of 28.12, SSIM of 0.912, and LPIPS of 0.124, outperforming both Instant-NGP and Mip-NeRF. The method proved to be effective for both common tomatoes and cherry tomatoes, demonstrating its universality and robustness on different tomato varieties with different geometric complexity. Overall, the study highlights Tomato-NeRF's robustness and effectiveness in achieving high-precision 3D reconstructions and realistic renderings, making it a valuable tool for agricultural and other applications that require detailed 3D models.

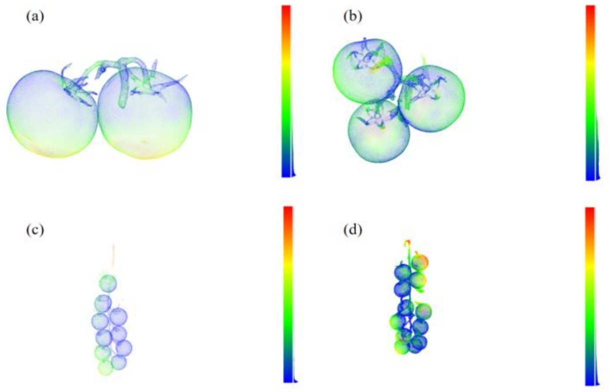


Fig. 10. Color-based error plot of extracted point cloud for tomato

TABLE I
POINT CLOUD RECONSTRUCTION QUALITY COMPARISON

	Tomato	
Method	Average Distance	Standard Deviation
Tomato-NeRF	5.2	2.1
Instant-NGP	15.7	5.2
Mip-NeRF	10.3	3.1
	Cherry Tomato	
Method	Average Distance	Standard Deviation
Tomato-NeRF	4.8	1.9
Instant-NGP	14.2	4.7
Mip-NeRF	9.5	2.8

TABLE II
QUANTITATIVE COMPARISON. WE REPORT LPIPS (LOWER IS BETTER) AND PSNR/SSIM (HIGHER IS BETTER)

	Tomato		
Method	PSNR↑	SSIM↑	LPIPS↓
Tomato-NeRF	27.55	0.894	0.148
Instant-NGP	26.54	0.849	0.182
Mip-NeRF	27.28	0.889	0.176
	Cherry Tomato		
Method	PSNR↑	SSIM↑	LPIPS↓
Tomato-NeRF	28.12	0.912	0.124
Instant-NGP	26.87	0.883	0.159
Mip-NeRF	27.92	0.903	0.149

V. Conclusion

Developing predictive digital twins for plants can lay the foundation for real-time simulation, predictive digital twins, and data-driven decision-making in agricultural practices. This paper introduces a multi-view reconstruction method for tomato plants, Tomato-NeRF, which is specially tailored for tomato reconstruction. It significantly improves the accuracy of Tomato complex geometry modeling, and the average distance between the predicted point cloud and the reference liDAR scan is 5.2mm, and the standard deviation is 2.1mm. The high fidelity of this 3D reconstruction is a substantial improvement over baseline methods such as Instant-NGP and Mip-NeRF. Specifically, for cherry tomatoes, the mean distance of Tomato-NeRF is 4.8mm and the standard deviation is 1.9mm, which is significantly better than other methods. Tomato-nerf also perform well in image reconstruction quality, with a PSNR of 28.12,SSIM of 0.912, and LPIPS of 0.124 for Cherry Tomatoes scenes. This method is very practical because of its simple data acquisition and the ability to create high-resolution geometric models from mobile phone camera data. Future research could further explore the integration of the Tomato-NeRF model with advanced IOT technology to achieve a more intelligent and responsive decision support system in the agricultural field. By deploying sensors in agricultural environments, we can get a rich stream of data in real time, covering plant growth, soil quality, meteorological conditions and more.

ACKNOWLEDGMENT

This work was supported in National Key Research and Development Program of China, NK202315020107.

The greenhouses were provided by the Zhongnong Futong Technology Park site in Beijing, China, and Hongfu Agricultural Tomato Production Park site in Beijing, China. All authors contributed to the work.

We declare that we have no financial and personal relationships with other people or organizations that can inappropriately influence our work, and there is no professional or other personal interest of any nature or kind in any product, or service.

References

[1].Sreedevi, T.R. and M.B. Santosh Kumar. Digital Twin in Smart Farming: A Categorical Literature Review and Exploring Possibilities in Hydroponics. in Advanced Computing and Communication Technologies for High Performance Applications (ACCTHPA). 2020. Cochin, India: IEEE.

[2].Slob, N. and W. Hurst, Digital Twins and Industry 4.0 Technologies for Agricultural Greenhouses. Smart Cities, 2022. 5(3): p. 1179-1192.

[3].Purcell, W., T. Neubauer and K. Mallinger, Digital Twins in agriculture: challenges and opportunities for environmental sustainability. Current opinion in environmental sustainability, 2023. 61: p. 101252.

[4].Angin, P.A.M.H., A digital twin for smart farming. J. Wirel. Mob. Networks Ubiquitous Comput. Dependable Appl., 2020. 4(11): p. 77-96.

[5].Pylianidis, C., S. Osinga and I.N. Athanasiadis, Introducing digital twins to agriculture. Computers and Electronics in Agriculture, 2021. 184: p. 105942.

[6].Cesco, S., et al., Smart agriculture and digital twins: Applications and challenges in a vision of sustainability. European Journal of Agronomy, 2023. 146: p. 126809.

[7].Sampaio, G.S., L.A. Silva and M. Marengoni, 3D Reconstruction of Non-Rigid Plants and Sensor Data Fusion for Agriculture Phenotyping. Sensors, 2021. 21(12): p. 4115.

[8].Ghandar, A., et al., A Decision Support System for Urban Agriculture Using Digital Twin: A Case Study With Aquaponics. IEEE Access, 2021. 9: p. 35691-35708.

[9].Xiang, L. and D. Wang, A review of three-dimensional vision techniques in food and agriculture applications. Smart Agricultural Technology, 2023. 5: p. 100259.

[10].Wang, T., et al., Applications of machine vision in agricultural robot navigation: A review. Computers and Electronics in Agriculture, 2022. 198: p. 107085.

[11].Fracarolli, J.A., et al., Computer vision applied to food and agricultural products. REVISTA CIÊNCIA AGRONÔMICA, 2020. 51(5).

[12].Dhanya, V.G., et al., Deep learning based computer vision approaches for smart agricultural applications. Artificial Intelligence in Agriculture, 2022. 6: p. 211-229.

[13].He, L. and J. Schupp, Sensing and Automation in Pruning of Apple Trees: A Review. Agronomy, 2018. 8(10): p. 211.

[14].Jignasu, A.E.A., Plant Geometry Reconstruction From Field Data Using Neural Radiance Fields. 2nd AAAI Workshop on AI for Agriculture and Food Systems, 2023.

[15].Rose, J., S. Paulus and H. Kuhlmann, Accuracy Analysis of a Multi-View Stereo Approach for Phenotyping of Tomato Plants at the Organ Level. Sensors, 2015. 15(5): p. 9651-9665.

[16].Leibe, B., et al., Pixelwise View Selection for Unstructured Multi-View Stereo. 2016, Springer International Publishing AG: Switzerland. p. 501-518.

[17].Dandrifosse, S., et al., Imaging Wheat Canopy Through Stereo Vision: Overcoming the Challenges of the Laboratory to Field Transition for Morphological Features Extraction. Frontiers in Plant Science, 2020. 11.

[18].Zhu, R., et al., Analysing the phenotype development of soybean plants using low-cost 3D reconstruction. Scientific reports, 2020. 10(1): p. 7055-7055.

[19].Mildenhall, B., et al., NeRF. Communications of the ACM, 2022. 65(1): p. 99-106.

[20].Šlapak, E., et al., Neural radiance fields in the industrial and robotics domain: applications, research opportunities and use cases. 2023, Cornell University Library, arXiv.org: Ithaca.

[21].Tretschk, E.A.T.V., Non-Rigid Neural Radiance Fields: Reconstruction and Novel View Synthesis of a Dynamic Scene From Monocular Video. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: p. 12959-12970.

[22].Müller, T., et al., Instant neural graphics primitives with a multiresolution hash encoding. ACM transactions on graphics, 2022. 41(4): p. 1-15.

[23].Yariv, L., et al., Volume Rendering of Neural Implicit Surfaces. 2021, Cornell University Library, arXiv.org: Ithaca.

[24].Wang, P., et al., NeuS: Learning Neural Implicit Surfaces by Volume Rendering for Multi-view Reconstruction. 2023, Cornell University Library, arXiv.org: Ithaca.

[25].Fu, Q., et al., Geo-Neus: Geometry-Consistent Neural Implicit Surfaces Learning for Multi-view Reconstruction. 2022, Cornell University Library, arXiv.org: Ithaca.

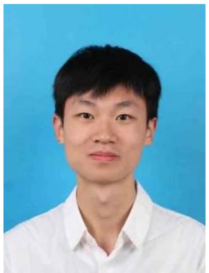
[26].Xu, Q., et al., Point-NeRF: Point-based Neural Radiance Fields. 2022, Cornell University Library, arXiv.org: Ithaca.

[27].Barron, J.T., et al., Mip-NeRF 360: Unbounded Anti-Aliased Neural Radiance Fields. 2021, Cornell University Library, arXiv.org: Ithaca.

[28].Barron, J.T.B.M., Mip-NeRF: A Multiscale Representation for Anti-Aliasing Neural Radiance Fields. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: p. 5855-5864.

[29].Tancik, M.E.W.E., Nerfstudio: A Modular Framework for Neural Radiance Field Development. ACM SIGGRAPH 2023 Conference Proceedings, 2023: p. 1-12.

[30].Hore , A. and D. Ziou. Image Quality Metrics: PSNR vs. SSIM. in 2010 20th international conference on pattern recognition. 2010: IEEE.



Xiajun Zheng received the B.S. of Engineering degree in Marine Technology from Dalian University of Technology, Dalian, China, in 2022. He is currently studying for a master's degree in Mechanical Engineering at China Agricultural University. His research interests include automation technology for smart greenhouses, agri-robotics and robotic control.

Xinyi Ai graduated from Beijing Information Science and Technology University with a Bachelor's degree in Mechanical Design manufacture and Automation Major in 2023. She is currently studying for a master's degree in mechanical engineering at China Agricultural University. Her research interests include agricultural robots and computer vision technology

Hao Qin graduated from Nanjing Agricultural University with a Bachelor's degree in Mechanical Design manufacture and Automation Major in 2022. He is currently studying for a master's degree in mechanical engineering at China Agricultural University. His research interests include intelligent equipment and agricultural robotics.

Jiacheng Rong is currently a Ph.D. student in Mechanical Manufacturing and Automation at China Agricultural University's College of Engineering. He obtained his M.S. degree in mechanical engineering from Soochow University in 2022, and his current research interests are focused on robotic visual perception and control.

Zhiqin Zhang received B.S. of Engineering degree in Mechanical Design, Manufacture and Automation from Shandong Agricultural University, Shandong, China, in 2021. She is currently studying for a master's degree in Mechanical Manufacturing and Automation at China Agricultural University. Her research interests include intelligent greenhouse inspection technology and agricultural robotics.

Yan Yang received B.S. of Engineering degree in agricultural mechanization and automation from Northeast Agricultural University, Harbin, China, in 2022. He is currently studying for a master's degree in Mechanical Engineering at China Agricultural University. His research interests include harvesting robot.

Ting Yuan was born in Zhoushan City, Zhejiang Province, China in 1981. He received the B.S. and Ph.D. degrees in mechanical engineering from China Agricultural University in 2004 and in 2012. From 2012 to 2016, he was a Lecturer in the Agricultural Robot Laboratory. Since 2017, he has been an Associate Professor with the Department of Mechanical Design and Manufacturing, College of Engineering, China Agricultural University. His research interests include machine vision, agricultural robot and intelligent agricultural equipment. He is the author of one book, nearly 30 articles, and more than 30 inventions.

Wei Li received Ph.D. in Vehicle Engineering, China Agricultural University, 2004. She is currently a Professor with college of engineering, China Agricultural University. She has charged over 20 scientific research projects such as National Key Technology Research and Development Program, National High Technology Research and Development Program and published more than 100 articles, declared nearly 60 national patents. Her research interests include Agricultural Robot and Agriculture Intelligent Equipment, Machine Vision and Computer Vision Inspection and Mechanical Manufacturing and Automation.