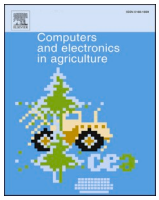
Computers and Electronics in Agriculture 217 (2024) 108629

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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



Original papers

Neural radiance fields for multi-scale constraint-free 3D reconstruction and rendering in orchard scenes

Jing Zhang [a,*], Xin Wang [b], Xindong Ni [b], Fangru Dong [a], Longrunmiao Tang [a], Jiahui Sun [a], Ye Wang [a]

[a] School of Management Engineering, Capital University of Economics and Business, Beijing 100070, China
[b] College of Engineering, China Agricultural University, Beijing 100083, China

ARTICLE INFO

Keywords:
Neural Radiance Fields
3D Reconstruction
3D Rendering
Digital Twin
Orchard

ABSTRACT

Efficiently, accurately, and realistically reconstructing large-scale 3D orchard scenes in a virtual world is an immensely challenging task. This complexity stems from the intricate and expansive of real orchard scenes. Traditional 3D reconstruction and rendering methods have encountered limitations in terms of modeling efficiency and computational costs, hindering the ability to provide users with immersive experiences. In response to these challenges, this study introduces a strategy for 3D scene reconstruction and rendering grounded in implicit neural representation: the NeRF-Ag model. Building upon the baseline NeRF, this model integrates a multi-resolution latent feature encoding technique, notably heightening training efficiency and elevating modeling precision. Furthermore, by means of environmental factor embedding, the model's robustness and practical applicability are further enhanced. The experimental outcomes illustrate that NeRF-Ag attains photo-realistic rendering outcomes across small, medium, and large scales. Moreover, it surpasses NeRF concerning the evaluation metrics of PSNR, SSIM, and LPIPS. Notably, the training speed of NeRF-Ag is roughly 39 times faster than NeRF. In 3D reconstruction tasks, NeRF-Ag showcases enhanced texture detail representation and higher modeling accuracy compared to the COLMAP-based 3D reconstruction method. Additionally, this study accomplishes free-viewpoint rendering of 3D scenes employing NeRF-Ag and provides evidence substantiating the connection between the quantity of training images and the precision of 3D rendering. The conclusions of this study will contribute to supporting and referencing the implementation of immersive visual interactive features within agricultural digital twin systems.

1. Introduction

Crops typically undergo a lengthy production cycle and incur significant human, material, and time costs throughout the process. Any error in the sowing, fertilizing, watering, or harvesting stages can result in irreparable losses. Nonetheless, the construction of a high-precision, digitized agricultural scenario model in the virtual world, enabling simulation, prediction, and optimization of the entire crop growth and development process at the physical level, will undoubtedly play a vital role in guiding real-world agricultural production (Ariesen-Verschuur et al., 2022). The introduction of the digital twin concept provides the direction to address the aforementioned issues. The primary task in achieving the integration of agriculture and digital twin technology is to accurately and realistically map real-world agricultural scenarios into the virtual world. Various virtualization methods aimed at agricultural objects have been proposed. For example, Moghadam et al. (2020) introduced the digital twin system AgScan3D + for phenotypic monitoring of fruit trees. The system utilizes 3D modeling in a virtual environment to simulate the canopy structure of fruit trees, facilitating the prediction of tree health and fruit quality. Ma et al. (2021) tackled the challenge of automated pruning for jujube trees by constructing a comprehensive 3D model, capable of generating key phenotypic information, including branch diameter and length. Li et al. (2022) developed a high-precision 3D model reconstruction system for cattle, enabling non-invasive detection of body size, weight, and physical condition, without subjecting the cattle to stress. This system establishes a dependable foundation for predicting beef production, quality, and ensuring traceability. However, it is worth noting that the majority of current agricultural virtualization research heavily depends on conventional 3D reconstruction and rendering techniques. These methods

* Corresponding author.
E-mail address: zhang-jing@cueb.edu.cn (J. Zhang).

https://doi.org/10.1016/j.compag.2024.108629
Received 11 August 2023; Received in revised form 4 December 2023; Accepted 8 January 2024
Available online 13 January 2024

have encountered limitations in modeling effectiveness, cost, and practicality, posing challenges in providing users with an immersive experience. Therefore, future endeavors should prioritize the exploration of advanced and innovative virtualization methods to further improve the effectiveness and user experience of agricultural virtualization technology.

1.1. Traditional 3D reconstruction and rendering

3D reconstruction and rendering techniques encompass the process of reverse engineering the real world to create corresponding 3D geometric models and compute lighting, textures, and additional information. Subsequently, these techniques render the 3D models into 2D images that faithfully depict real-world scenes from any given viewpoint (Ham et al., 2019). Traditional 3D reconstruction methods can be categorized into active and passive approaches. Active 3D reconstruction employs technologies like lidar (Rivera et al., 2023), structured light (Zhuo et al., 2022) and time-of-flight (ToF) (Jung et al., 2022) to scan the surface structure of objects. By analyzing the scan results, the desired 3D model is constructed. On the other hand, passive 3D reconstruction relies on visual sensors to capture image sequences. By integrating feature matching algorithms like Shape-from-Shading (SFS) (Cao et al., 2022), Speeded-Up Robust Features (SURF) (Ince, 2022), and Structure from Motion (SfM) (Gao et al., 2022), spatial coordinates of points are derived from the image sequences, contributing to the reconstruction of the desired 3D model. Meanwhile, deep learning-based 3D reconstruction methods like PointMVSNet (Chen et al., 2019), PF-Net (Huang et al., 2020), and Pix2Vox (Xie et al., 2019) exhibit substantial potential for integration with both active and passive 3D reconstruction paradigms, thus further enhancing the effectiveness of 3D model construction. However, regardless of being active or passive, traditional 3D reconstruction and rendering methods have long faced constraints due to the trade-off between accuracy and volume. Constructing high-precision 3D models frequently necessitates compromising on reconstruction efficiency and handling substantial volumes of data. Conversely, relatively simple and efficient reconstruction methods may compromise reliability, precision, and the final rendering quality. This challenge is particularly prominent in complex and extensive agricultural environments, including orchards or farmlands (Pylianidis et al., 2021). Constructing high-resolution and large-scale 3D scene models in these settings, while ensuring robustness and timeliness, becomes an essential requirement for the development of agricultural digital twin systems.

1.2. NeRF-based 3D reconstruction and rendering

NeRF (Neural Radiance Fields) (Mildenhall et al., 2021) present a novel approach to tackle the aforementioned challenges. Indeed, traditional 3D model representations like mesh, point cloud, and voxel are all regarded as “explicit” representations. These “explicit” representations are discrete, leading to issues like overlapping, blurring, and aliasing during the rendering process. Additionally, for large-scale scenes, substantial memory consumption may occur. In contrast, NeRF falls under the category of implicit neural representation methods, boasting the advantage of continuous implicit functions. This continuity enables NeRF to describe complex geometric shapes without significantly increasing storage space. Furthermore, NeRF can generate photo-realistic rendering results without requiring 3D feature supervision. Currently, several NeRF-based research works have further improved the effectiveness of 3D reconstruction and rendering. Models like NeRF++ (Zhang et al., 2020), Nerfstudio (Tancik et al., 2023), and Mip-NeRF (Barron et al., 2021) have advanced from various perspectives, including shape radiance ambiguity, proposal sampling, and anti-aliasing, enhancing the overall performance of NeRF-based approaches. Instant-NGP (Müller et al., 2022) and TensoRF (Chen et al., 2022) have achieved significant progress in accelerating the training process of NeRF models, reducing it from hours to seconds. This accomplishment enables real-time 3D scene modeling and rendering using NeRF. Additionally, in response to the challenges presented by large-scale open scenes, NeRF-W (Martin-Brualla et al., 2021) and Block-NeRF (Tancik et al., 2022) have proposed distinct solutions, establishing the groundwork for 3D reconstruction and rendering of complex scenes without limitations or constraints. However, it is worth noting that NeRF application in agricultural scenes is currently limited. The agricultural environment presents unique challenges, including large scale, high precision requirements and uncertainties. These factors undoubtedly challenge NeRF-based 3D reconstruction and rendering methods when applied in agriculture.

The primary contribution of this study as an early attempt of applying NeRF in the agricultural is proposing an improved version of NeRF, referred to as NeRF-Ag, for 3D reconstruction and rendering in multi-scale orchard scenes. The method achieves high-precision 3D mesh model generation and enables constraint-free, photo-realistic rendering. These advancements provide support and a reference basis for implementing immersive visual interactive features within agricultural digital twin systems.

The rest of this paper is organized as follows: the used data and proposed method is explained in detail in Section 2. Results and discussion are presented in Section 3 and 4. Finally, we conclude the paper in Section 5.

2. Materials and methods

2.1. Data acquisition and application

This study conducted data collection in a commercially operated strawberry orchard during the harvesting season. The orchard is located in the strawberry cultivation base of Daxing District, Beijing, China, and mainly employs greenhouse cultivation. A specific greenhouse within this orchard was chosen as the primary data collection scene. This greenhouse follows a short rear-slope construction, with a ridge height of approximately 2.8 m, a length of about 60 m, and a span of around 10 m. Unlike active 3D reconstruction methods, data collection in this study was carried out using only a Huawei smartphone. All image capture took place under natural lighting conditions, without any artificial intervention. The obtained RGB images have a resolution of 4608 × 3456 pixels. This approach highlights the cost-effectiveness, convenience, and practicality of the proposed 3D reconstruction and rendering method, positioning it as an advantageous technique.

In this study, we constructed a dataset for training and validating the NeRF neural network, as shown in Fig. 1. The dataset includes three subsets categorized by scale: small-scale, medium-scale, and large-scale. The large-scale dataset covers the entire strawberry orchard with captured images. The medium-scale dataset mainly captures individual rows of strawberries, while the small-scale dataset focuses on images of individual strawberry plants. Fig. 1.(a1), 1.(b1), and 1.(c1) illustrate camera poses and image capture ranges using sparse point cloud data for the three scale subsets. Corresponding RGB image capture examples from these positions are displayed in Fig. 1.(a2), 1.(b2), and 1.(c2). The dataset is categorized into small, medium, and large scales to evaluate the feasibility and effectiveness of the proposed 3D reconstruction and rendering method across different capture scales. The specific composition of the dataset is detailed in Table 1. Importantly, the actual number of images needed for NeRF model training could be significantly fewer than the total number of images in the training set. Detailed analysis results on this aspect will be presented in the experimental section.

2.2. Constructing a neural network as an implicit function for representing 3D objects

Deviating from traditional 3D representation forms using point

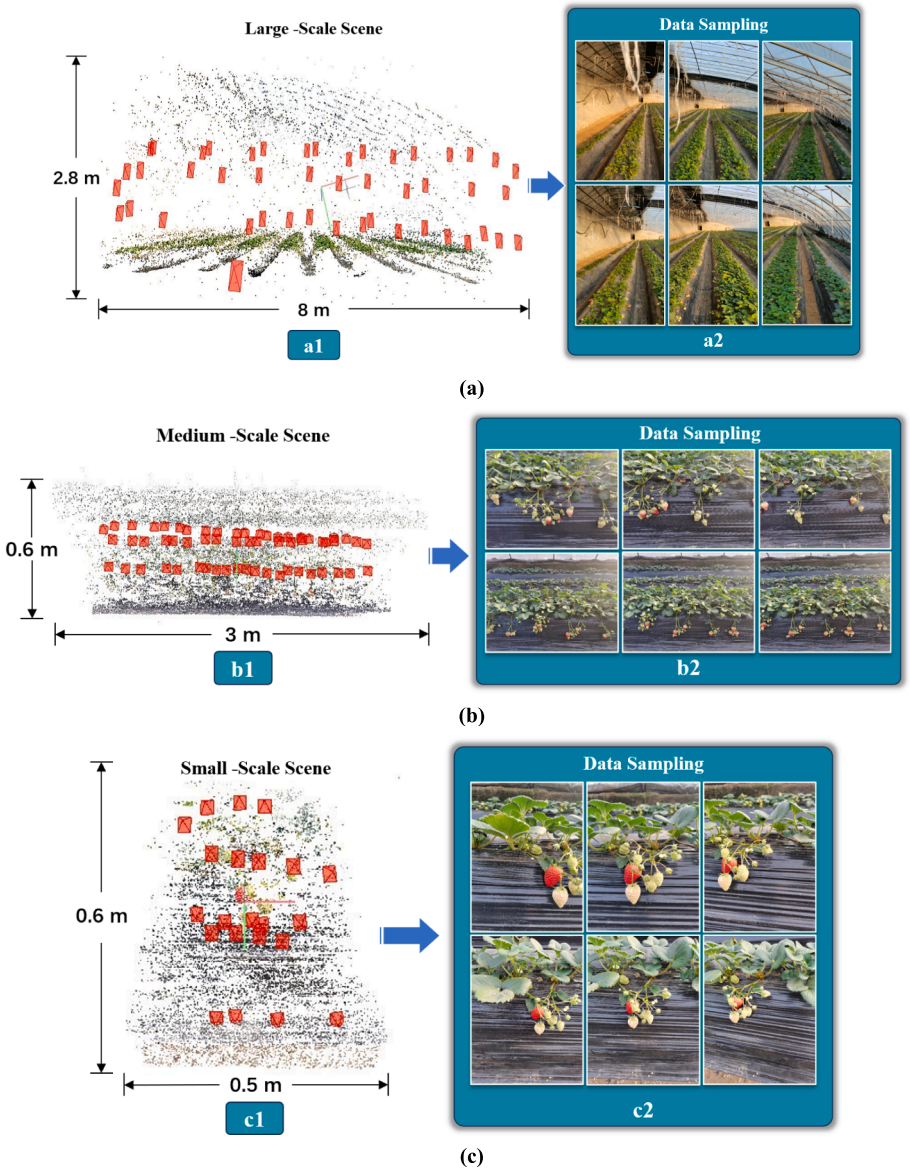


Fig. 1. Scenarios of data acquisition and showcase of acquired images. (a) Large -Scale scene; (b) Medium-Scale scene; (c) Small-Scale scene.

Table 1
Structure and composition of the dataset.

Delineation of sub-data sets	Training-set	Validation-set	Test-set
Small-Scale	50	10	20
Medium-Scale	200	50	100
Large-Scale	200	50	100

clouds or voxels, this study constructs a NeRF model that implicitly encodes the 3D orchard scene within a fully connected neural network. The specific pipeline of this NeRF model for representing 3D objects is shown in Fig. 2. In this context, the model's input consists of a 5-dimensional radiance field that combines (x, y, z) and (θ, Φ). The camera's shooting perspective (θ, Φ) corresponding to each image can be obtained from the camera extrinsic. If a ray $r(a)$ is emitted from this perspective and passes through point P on the image, entering the actual 3D scene, a

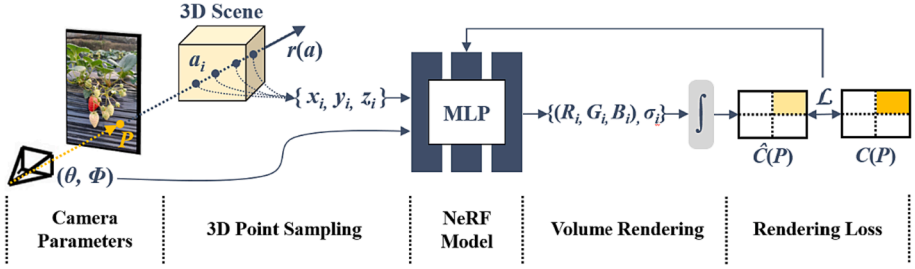


Fig. 2. NeRF-based 3D representation pipeline.

set of discrete points a_i is sampled within the 3D scene, with $r(a)$ considered as a reference. The positional coordinates of these sampled points are indicated as (x_i, y_i, z_i). Using this approach, any point within the 3D scene can be represented by a 5-dimensional vector (x, y, z, θ, Φ).

The NeRF model developed in this study is employed to predict the color (R_i, G_i, B_i) and voxel density (σ_i) values of different sampled points in 3D space projected onto 2D images, as shown in Fig. 2. The purpose of this is to utilize volume rendering algorithms for generating images from arbitrary viewpoints within 2D space. For neural network training, the conventional volume rendering equation is converted into a Riemann sum formulation (Equation (1). Here, $\hat{C}(P)$ represents the estimated RGB color value of pixel point P in the 2D image, computed by integrating rays $r(a)$ from the nearby sample point a_n to the distant sample point a_f. T symbolizes the transmittance from a_n to point a, while σ and c correspondingly indicate the voxel density value and RGB value of the ray at point a, acquired from the output of the MLP. Additionally, to discretize the volume rendering equation, the distance between a_n and a_f is partitioned into N small intervals. Consequently, the position of a_i along the ray can be approximated by employing the third equation in the set of Equation (1).

$$\begin{cases} \hat{C}(P) = \sum_{i=1}^{N} T_i(1 - \exp(-\sigma_i\delta_i))c_i \\ T_i = \exp\left(-\sum_{j-1}^{i-1}\sigma_i\delta_i\right) \\ a_i\ u\left[a_n + \dfrac{i-1}{N}(a_f - a_n), a_n + \dfrac{i}{N}(a_f - a_n)\right] \\ \delta_i = a_{i+1} - a_i \end{cases} \tag{1}$$

In summary, the rendering strategy of the NeRF model entails sampling every ray emitted from the camera at N points. This implies that extensive computations are carried out, even for empty or occluded areas that do not contribute to the rendering. To further optimize the computational load, this study utilizes a hierarchical sampling approach that divides the sampling process into "coarse" and "fine" stages, as illustrated in Equation (2).

$$\begin{cases} \hat{C}_c(P) = \sum_{i=1}^{N_c}\omega_i c_i \\ \omega_i = T_i(1 - exp(-\sigma_i\delta_i)) \end{cases} \tag{2}$$

Initially, a coarse sampling of N_c points is performed. Utilizing Equation (2), an estimated color value $\hat{C}_c(P)$ for point P is calculated, and normalization is applied to ω_i to determine the segmented constant probability density ($\widehat{\omega_i} = \omega_i / \sum_{j=1}^{N_c}\omega_j$). Subsequently, N_f points are acquired through inverse transform sampling and combined with the original N_c points for the final rendering. The hierarchical sampling technique leads to a reduced number of sampled points for NeRF model training, while also incorporating more meaningful color information. With the known RGB ground truth color value $C(P)$ on the 2D image for point P, it is utilized as a supervisory signal during NeRF model training. The model's loss function is expressed as depicted in Equation (3).

$$\mathscr{L} = \sum_{P\in R}[\ \|\hat{C}_c(P) - C(P)\|_2^2 + \|\ \hat{C}_f(P) - C(P)\|_2^2] \tag{3}$$

2.3. Strategies to enhance the NeRF model in actual orchard scenes (NeRF-Ag)

Efficiency and accuracy have consistently remained the foremost considerations in 3D reconstruction and rendering techniques, particularly when employed in actual orchard or farmland environments. Using orchard scenes as an example, time plays a crucial role across various growth stages of fruit cultivation, pruning, pollination, and harvesting. Delays in these agricultural activities can result in irreversible economic losses for commercial orchards. Traditional methods of creating high-precision 3D models for actual orchard scenes entail significant time and labor expenses. Moreover, virtual and simulation representations of fruit phenotypes and growth environments impose rigorous accuracy requirements on 3D model construction. Sparse and discrete point cloud objects exhibit various limitations in practical applications and may not meet the necessary demands. Therefore, as a response to the aforementioned challenges, this study puts forth targeted enhancement strategies to tackle the problems of low training efficiency and the excessive modeling constraints in the NeRF model. These strategies result in the creation of the NeRF-Ag model, as shown by its internal structure in Fig. 3.

Low training efficiency: For instance, training the baseline NeRF model for a basic indoor scene might demand approximately 12 h within a consumer-grade, single GPU resource environment. For more extensive orchard or farmland scenes, training could span several days, a timeline that does not correspond to the practical demands of agricultural production. Drawing inspiration from the Instant-NGP model, this study presents the innovative NeRF-Ag model. This model employs multi-resolution latent feature encoding to substitute the position encoding in the baseline NeRF, leading to a noteworthy reduction in the necessary training time, coupled with a heightened enhancement of 3D reconstruction and rendering precision.

Excessive modeling constraints: The baseline NeRF relies on fixed ambient lighting, uncomplicated modeling scenes, consistently uniform image capture viewpoints, and constant camera parameter settings as crucial prerequisites for attaining successful 3D reconstruction and rendering. However, in the context of expansive orchard or farmland scenes marked by substantial uncertainty, the ambient lighting undergoes real-time fluctuations, scenes become congested with less conspicuous features, and the perspectives and scales for image capture exhibit versatility and adaptability. To address these challenges, this study derives inspiration from NeRF-W and introduces an approach that embeds environmental factors.

2.3.1. Multi-Resolution latent feature encoding for 3D scenes

Training the NeRF model entails sampling in 3D space. However, the majority of 3D space regions do not align with actual scenes. Once sampled, this implies that the neural network assigns weights to the sampled point and proceeds with feature computation. Unfortunately, such computation squanders a substantial amount of time and resources. The NeRF-Ag model proposed in this paper employs hash encoding and spherical harmonics encoding techniques, mapping low-dimensional positional input and viewpoint information to a high-dimensional space, as depicted in Fig. 3. This encoding approach structurally stores latent features within a hash table, preventing the neural network not only from computing meaningless sampled points but also constraining the parameters of the MLP to the hash table's size, thereby notably

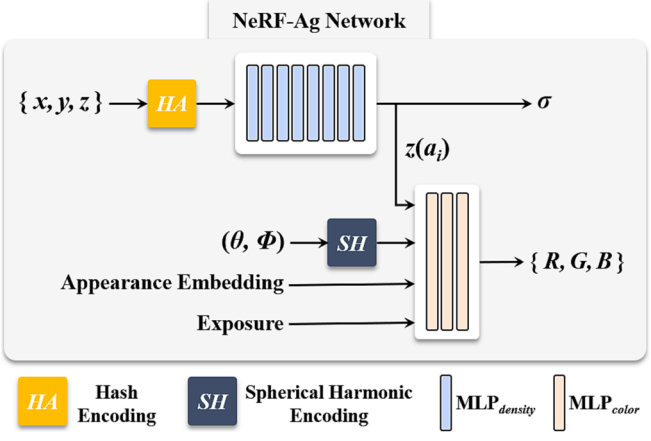


Fig. 3. Structural components within the NeRF-Ag model.

enhancing network training efficiency.

Inside the NeRF-Ag model, the entire neural network is bifurcated into two components: $MLP_{density}$ and MLP_{color}. The $MLP_{density}$ primarily undertakes the task of estimating voxel density (or opacity) value σ. This division is based on the theoretical concept that the object's σ in space is solely influenced by the observed position and remains unaffected by lighting conditions or viewing angles. Conversely, MLP_{color} is utilized to forecast color (RGB) values. The color of the imaging point hinges on the object's σ and the ray's angle, coinciding with MLP_{color}'s inputs. $MLP_{density}$ comprises 8 hidden layers, while MLP_{color} encompasses 3 hidden layers; each layer accommodates 256 neurons. Furthermore, the hash encoding process for latent features within the NeRF-Ag model parallels that of Instant-NGP. Yet, the distinguishing aspect pertains to the hash table, characterized by a length of 2^{19} and a width of 4. Given the extensive range of orchard scenes, the model predominantly incorporates 16 distinct resolution levels.

Due to the intricacies entailed in depicting latent feature encoding for position and viewpoint information within 3D space, this study transposes the procedures of 3D information hashing and spherical harmonics encoding to 2D space. Consequently, employing 2D images as input facilitates the visualization of the results of latent feature encoding. This method is depicted in Fig. 4. In this visualization, the 2D hash encoding is set up with 4 distinct resolution levels (*l1* to *l4*). Similarly, the spherical harmonics encoding extends to the 3rd order, and its encoding mechanism references the process of environment lighting reconstruction through spherical harmonics. Fig. 4 demonstrates that hash encoding significantly reduces computational costs and also more effectively captures high-frequency information features from the original image across varying resolution conditions. Meanwhile, spherical harmonics encoding adeptly extracts latent features connected to environmental lighting, thereby enhancing the overall model representation.

2.3.2. Environmental factor embedding

In real orchard scenarios, lighting varies due to shifting environmental conditions. Additionally, when capturing both global and localized images within the same orchard, the lighting conditions exhibit significant variations. Moreover, the utilization of various image capture devices can result in substantial disparities in exposure parameter configurations, which may include occurrences of overexposure. The baseline NeRF model is notably sensitive to these external factors. If the image quality within the training set fails to meet the necessary standards, the resulting reconstruction and rendering outcomes might be subpar or even unsuccessful. To tackle these challenges, the NeRF-Ag model incorporates both appearance information and exposure as two environmental factors.

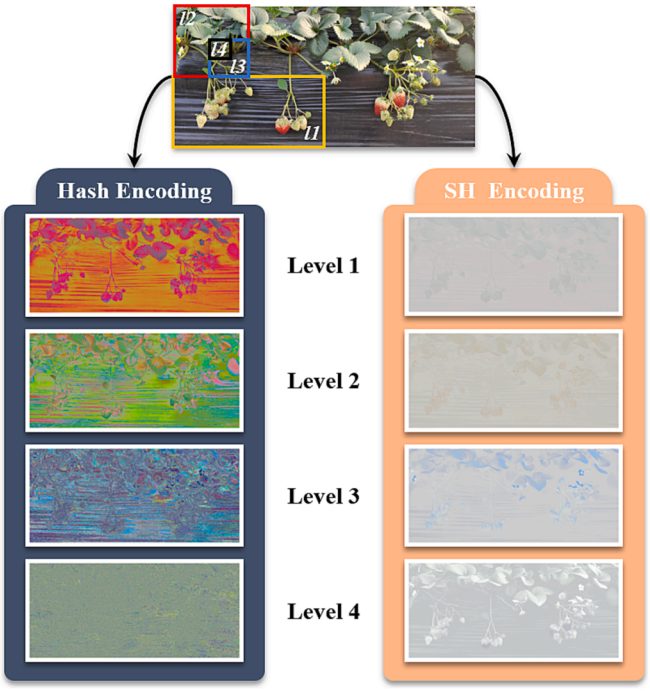


Fig. 4. Visualization of two latent feature encoding effects on 2D images.

$$z(a_i) = MLP_{density}([x_i, y_i, z_i]) \tag{4}$$

$$\widehat{C}(P) = MLP_{color}\left(z(a_i), \gamma(\theta, \Phi), l_i^{(t)}, \gamma_i^{PE}\right) \tag{5}$$

In Equation (5), the appearance information of the image i is represented by the symbol $l_i^{(t)}$, with t denoting the length of this vector. The $l_i^{(t)}$ vector is generated by a self-encoder named Generative Latent Optimization (GLO) (Bojanowski et al., 2017). Drawing inspiration from the concept of Generative Adversarial Networks (GANs) (Goodfellow et al., 2020), the $l_i^{(t)}$ vector can be viewed as a high-dimensional fusion of two components. One of these components represents the intrinsic features of the image, while the other component consists of random noise conforming to a normal distribution. Furthermore, variations in exposure levels significantly impact image quality. Thus, similar to the sinusoidal position encoding in Transformers (Vaswani et al., 2017), this paper adopts a mechanism for managing exposure levels. The resultant exposure embedding vector γ_i^{PE}, along with the $l_i^{(t)}$ vector, is fed into the MLP_{color}, as depicted in Equation (5). Embedding appearance information and exposure enables NeRF-Ag to capture the variations among diverse input images from a macroscopic viewpoint. This capability is pivotal in facilitating NeRF-Ag to attain comprehensive, multi-scale, and unconstrained 3D reconstruction and rendering, thereby enhancing its robustness and efficacy in intricate agricultural scenarios.

3. Results and discussion

3.1. Multi-Scale evaluation of 3D rendering performance in orchard scenes

The NeRF-Ag model was developed using Ubuntu 18.04 and PyTorch 2.0.1 with CUDA 11.7. Model training utilized a single consumer-grade GPU (GTX 1070 Ti). Comprehensive assessment of the NeRF-Ag model's practical 3D rendering performance was conducted using three typical image generation quality metrics: PSNR (Peak Signal to Noise Ratio), SSIM (Structural Similarity), and LPIPS (Learned Perceptual Image Patch Similarity). Among them, PSNR is commonly used to measure the pixel value difference between the reconstructed image and the Ground-Truth (GT) image, but it is not sensitive to the visual imaging quality perceived by the human eye. Its calculation method is shown in Equations (6) and (7). Equation (6) is employed to calculate the Mean Squared Error (MSE) between two images, A and B, of size $m \times n$, where A represents the GT, and B represents the rendering results of NeRF-Ag model. In Equation (7), MAX_I denotes the maximum possible pixel value.

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

$$PSNR = 10 \bullet lg\left(\frac{MAX_I^2}{MSE}\right) \tag{7}$$

Compared to PSNR, SSIM and LPIPS pay more attention to the perceptual differences in image perception by the human eye. Specifically, the SSIM models image distortion as a combination of three different factors: luminance, contrast, and structure. It uses the mean as an estimate of luminance, the standard deviation as an estimate of contrast, and covariance as a measure of structural similarity. The computational method is shown in Equation (8), where μ_A and μ_B are the means of images A and B, σ_A^2 and σ_B^2 are the variances of A and B, σ_{AB} is

the covariance between A and B, and c_1 and c_2 are constants positively correlated with the pixel value range. Additionally, the LPIPS index approximates human-perceived visual similarity by learning a neural network model, and its calculation process is detailed in (Zhang et al., 2018).

$$SSIM(A,B) = \frac{(2\mu_A\mu_B + c_1)(2\sigma_{AB} + c_2)}{(\mu_A^2 + \mu_B^2 + c_1)(\sigma_A^2 + \sigma_B^2 + c_2)} \tag{8}$$

Furthermore, a comparative analysis benchmarked the NeRF-Ag model against three representative NeRF models: baseline NeRF, TensoRF, and NeRF-W. This comparison aimed to elucidate NeRF-Ag's strengths and weaknesses in comparison to these baseline models. The results of these comparisons are presented in Table 2.

The results in Table 2 demonstrates that the proposed NeRF-Ag model outperforms the other three models, exhibiting superior performance across three scales: small, medium, and large. Specifically, in the small and medium scales, NeRF-Ag displays comparable rendering accuracy. However, for large-scale scene, the rendering accuracy of all models decreases relatively. This observation highlights the significant correlation between the 3D rendering accuracy of NeRF models and the employed image acquisition methodology. The inherent abundance of textures and details in the training data leads to improved rendering outcomes. This implication underscores the significance of capturing a substantial amount of detailed information during the initial phase of scene data collection, even for scenes of wide scope. This approach ensures the NeRF-Ag model's achievement of superior 3D rendering effects. Additionally, the study provides real-world 3D rendering results for distinct models, with a particular focus on accentuating differences in details, as depicted in Fig. 5.

The GT images in Fig. 5 were randomly chosen from the testing dataset, whereas the remaining images were generated directly by their respective models. In the case of NeRF models, images were generated to align with the viewpoints of the chosen images from the testing dataset. As the images within the testing dataset portray previously untrained viewpoints, contrasting the distinctions between these two image sets facilitates the assessment of NeRF models' 3D rendering efficacy. The resulting images possess a resolution of 1920 × 1080, albeit a few have been cropped to enhance their presentation. As evident from Fig. 5, the NeRF-Ag model attains photo-realistic rendering effects consistently across the three distinct scales. Particularly in rendering intricate textures, NeRF-Ag exhibits a capability that closely emulates the GT effects, especially within small-scale scenes. Furthermore, NeRF-Ag provides comparatively excellent rendering effects for fine details in the medium and large scales. The inclusion of appearance embedding factors leads the NeRF-W model to showcase rendering effects akin to those of NeRF-Ag. Meanwhile, images output by NeRF and TensoRF display noticeable discrepancies in clarity compared to the GT. Their generated fine-detail texture effects do not meet the desired benchmarks.

Besides rendering quality, the training and rendering efficiency of the NeRF-Ag model are also of paramount significance. This is particularly critical for expansive orchards or agricultural fields, where substantial training data may result in training durations spanning several days or more. This study performs a comparative analysis on the training and rendering efficiency of the aforementioned four NeRF models, and the outcomes are displayed in Table 3. Utilizing the previously mentioned experimental setup, all four models generate images with a consistent resolution of 1920 × 1080. Significantly, the NeRF-Ag model being proposed demonstrates a training speed nearly 39 times quicker than that of the baseline NeRF and NeRF-W models. While TensoRF exhibits superior training efficiency, it remains within the same order of magnitude as NeRF-Ag. Likewise, despite NeRF-Ag's larger sizes, its rendering efficiency surpasses that of the NeRF and NeRF-W models. This highlights the efficacy of the multi-resolution latent feature encoding method employed in this study to enhance the efficiency of model training. In practical application, TensoRF represents the scene as a four-dimensional tensor, differing significantly in model structure from conventional NeRF models. In conclusion, although potential for improvement exists in the training and rendering efficiency of the NeRF-Ag model, it currently fulfills the essential practical application demands for expansive scenes like orchards or agricultural fields, while upholding accuracy.

3.2. Evaluation of 3D reconstruction performance in orchard scenes

Serving as a representation technique for 3D objects, the NeRF-Ag model, functioning as an implicit function, possesses the capability not only to produce photo-realistic 2D images from new perspectives but also to directly reconstruct the 3D mesh structure data of the intended object. In contrast to point clouds, mesh structures encompass abundant topological and neighborhood information and can be rendered with ease, rendering them an essential element for intricate geometric operations within digital twins. In this study, the NeRF-Ag model was employed to accomplish 3D mesh model reconstruction and meticulous texture mapping across three distinct scales: small, medium, and large scenes. The results are depicted in Fig. 6.

Throughout the 3D reconstruction process, this study utilized a conventional incremental SfM method using COLMAP (Schonberger and Frahm, 2016) as a benchmark for a more comprehensive assessment of the 3D reconstruction efficacy of the NeRF-Ag model. As apparent from Fig. 6, NeRF-Ag effectively reconstructed the 3D mesh models of the respective scenes across various scales. Importantly, when scrutinizing the intricacies of the 3D models, it became evident that NeRF-Ag demonstrated superior accuracy in depth estimation in contrast to the COLMAP approach. Additionally, NeRF-Ag proficiently circumvented the creation of notable artifact regions surrounding object edges (Fig. 6a). Moreover, the size of the 3D mesh model for the medium-scale scene in this study is roughly 209 MB (from COLMAP), whereas the NeRF-Ag models for the three distinct scales share a size of 176 MB each. The size of high-precision 3D models generally grows as the scene expands (potentially reaching sizes exceeding a hundred GB). However, owing to the implicit 3D representation strategy embraced by NeRF-Ag, its size is exclusively governed by the count of model parameters in the absence of geometric computation requirements. This attribute signifies that for 3D reconstruction assignments in extensive scenes like orchards or agricultural fields, the NeRF-Ag model harbors significant potential for application.

To comprehensively evaluate the 3D reconstruction performance of the NeRF-Ag model, this study implemented oval rendering trajectories in 3D space for scenes of different scales: small, medium, and large. Furthermore, four extreme viewpoints were explicitly selected: upward, downward, leftward, and rightward. This configuration allowed NeRF-Ag to generate depth maps and the corresponding 3D rendered images

Table 2
Comparison of 3D rendering performance between different NeRF models in multi-scale.

NeRFs	Small-Scale			Medium-Scale			Large-Scale		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
NeRF	22.78	0.486	0.614	22.65	0.482	0.623	19.95	0.387	0.689
TensoRF	23.94	0.583	0.522	23.89	0.591	0.515	21.43	0.507	0.592
NeRF-W	24.86	0.627	0.434	24.72	0.619	0.458	22.36	0.551	0.528
NeRF-Ag (Ours)	**26.43**	**0.716**	**0.342**	**26.27**	**0.711**	**0.318**	**24.55**	**0.674**	**0.405**

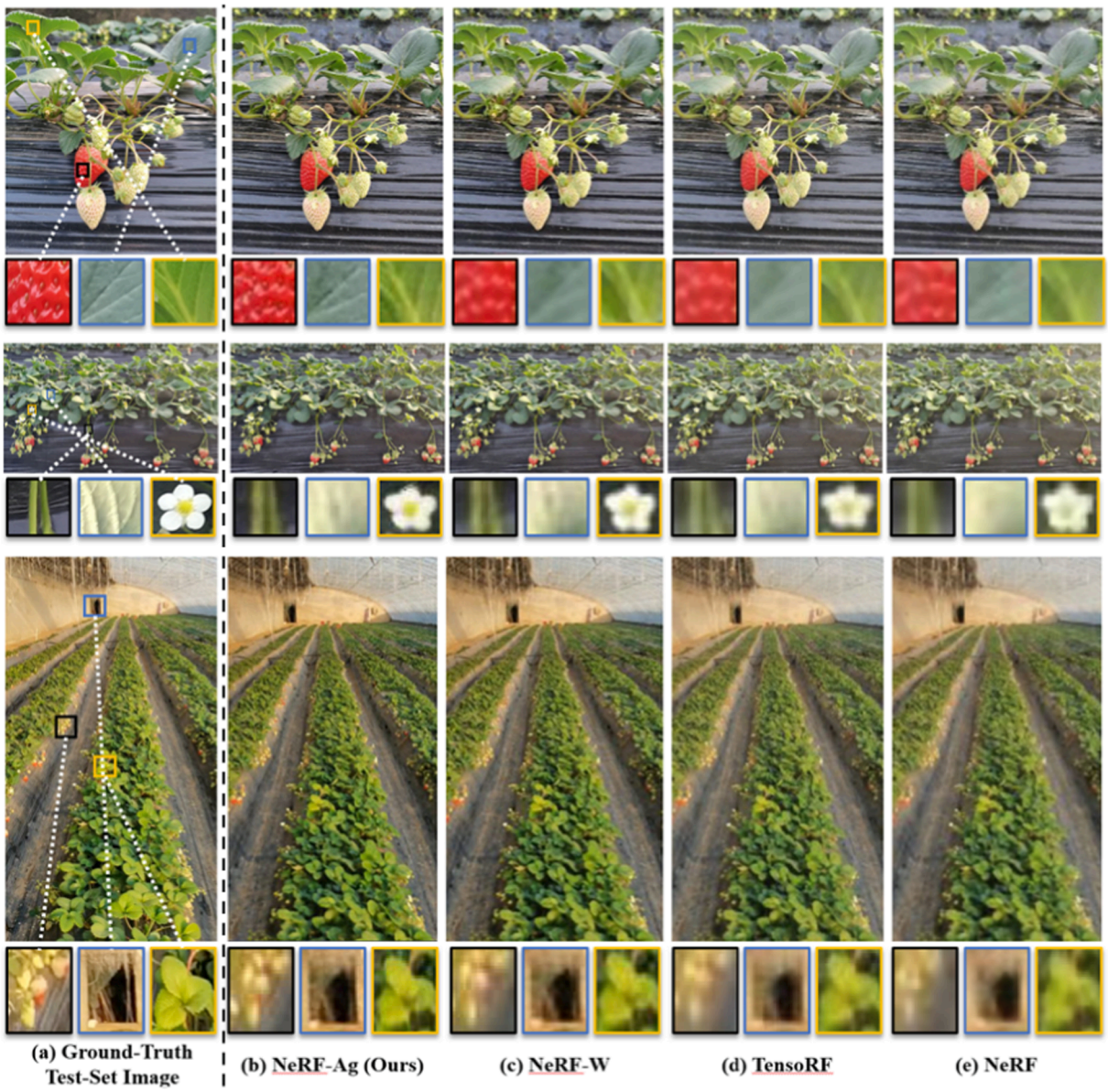


Fig. 5. Demonstration of 3D rendering details obtained from NeRFs at multi-scale.

Table 3
Comparison of training and rendering efficiency for different NeRFs.

NeRFs	Average Training Time	Average Rendering Time (FPS)	Model Size (MB)
NeRF	~12 h	0.02	14
TensoRF	11.4 *m*	0.06	42
NeRF-W	~12 h	0.02	17
NeRF-Ag (Ours)	18.7 *m*	0.05	176

from these distinct viewpoints, as illustrated in Fig. 7.

Expanding on the reconstruction of 3D scene models, this research further accomplished free-viewpoint rendering of the scenes across varied scales utilizing NeRF-Ag. The black dashed lines in Fig. 7 depict the rendering paths for the respective scenes, and the associated rendering videos will be subsequently made available. As apparent from Fig. 7, NeRF-Ag upholds photo-realistic rendering quality when viewed from the four extreme viewpoints: upward, downward, leftward, and rightward. Additionally, the generated depth maps demonstrate a high level of accuracy. This suggests that even in the absence of any 3D information as input, the NeRF-Ag model possesses the ability to generate 2D images from novel and untrained perspectives while precisely and comprehensively reconstructing the corresponding 3D scene. The extent of its 3D reconstruction relies exclusively on the coverage of the collected 2D images.

3.3. Ablation experiments

3.3.1. The correlation between the quantity of training images and 3D rendering performance

During the 3D reconstruction process for vast scenes such as orchards or agricultural fields, the initial data collection phase demands substantial time and human resources. Ensuring diminished reliance of the NeRF-Ag model on the quantity of training data, while retaining accuracy, holds paramount importance for practical applications. This study carried out ablation experiments to scrutinize the correlation between the number of training images and 3D rendering accuracy across various scales. The results of these experiments are depicted in Fig. 8. The ablation experiments encompassed training the baseline NeRF and

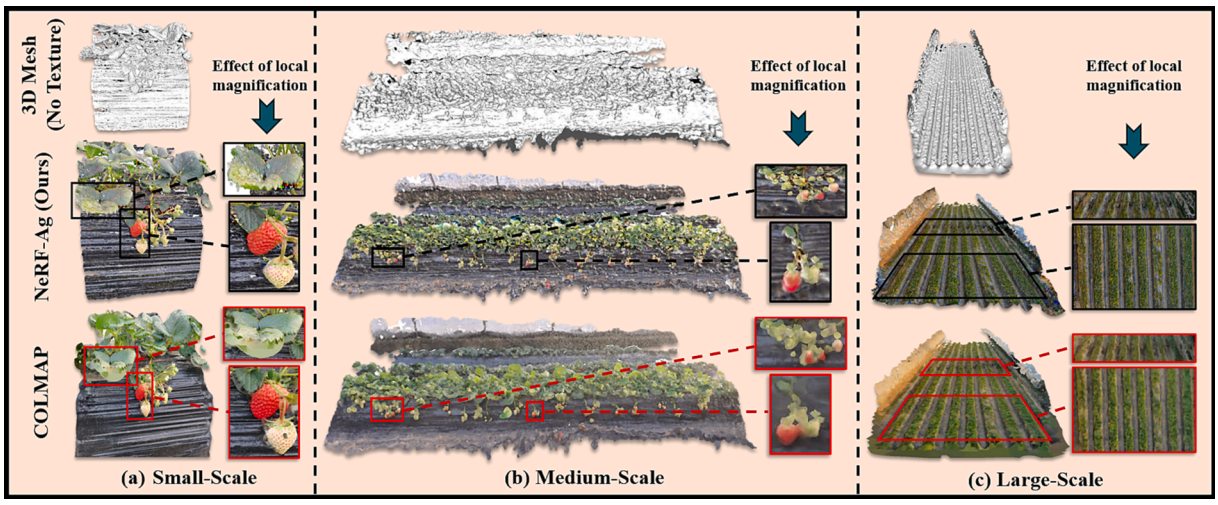


Fig. 6. Reconstruction of 3D mesh models at multi-scale.

NeRF-Ag model four times for each of the small, medium, and large scales, respectively. In each training iteration, the number of training images increased linearly. Following training, the 3D rendering accuracy of each model was assessed based on PSNR. Furthermore, owing to the diverse scales of the scenes, the necessary quantity of training images also varies. In the case of small-scale scenes, merely 10 images prove sufficient for training; however, this yields a considerable decline in the 3D rendering accuracy of the corresponding model. Employing 20 or 40 images to train models in small-scale scenes results in solely minimal disparities in rendering accuracy. This signifies that, while guaranteeing sufficient image coverage, creating a small-scale 3D scene utilizing the NeRF-Ag model mandates approximately 20 training images. In medium-scale and large-scale scenes, the precision of 3D rendering enhances with an augmented number of training images. Nevertheless, once the count of images reaches roughly 150, increasing the number of training images has a more subdued effect on the rendering accuracy of the NeRF-Ag model. This implies that the necessary quantity of training images for the NeRF-Ag model demonstrates characteristics of saturation. Training the baseline NeRF model with 50 images in medium and large scale resulted in PSNR values of only 10.77 and 13.24 for its rendered outputs, which proves inadequate for practical applications. In contrast, under data-limited conditions, NeRF-Ag demonstrates significantly better rendering performance compared to the NeRF model. This observation indirectly reflects the stronger practicality and robustness of NeRF-Ag model.

3.3.2. The correlation between environmental factor embedding and 3D rendering performance

While capturing 2D images in natural settings, the quality of acquired images is notably affected by lighting conditions and the equipment utilized for capturing. Especially in the case of expansive environments like orchards or agricultural fields, the unpredictability of environmental elements is substantial, posing a challenge to the maintenance of uniform image quality during the process of data collection. This study conducted relevant ablation experiments to validate the effectiveness of environmental factor embedding in the NeRF-Ag model. The experimental design is illustrated in Fig. 9. These experiments were conducted for small, medium, and large scenes. During the capture of images near viewpoint ①, exposure parameters were reduced, while exposure parameters were increased for images captured near viewpoint ②. Normal exposure parameters were used for capturing images from other perspectives. Specifically, for the small-scale scene training set, the data distribution included 5 underexposed and 5 overexposed images, along with 35 normally exposed images. In both the medium and large-scale scenes, the training set comprised 150 images, including 10 underexposed and 10 overexposed images for specific viewpoints ① and ②. This experiment involved training both the full NeRF-Ag model and the NeRF-Ag model without environmental factor embedding.

Through the output results of 3D rendering (Fig. 10), it is observed that the NeRF-Ag model without environmental factor embedding exhibits corresponding underexposure or overexposure in viewpoints ① and ②. This phenomenon is consistent across small, medium, and large scales. Conversely, under the same viewpoint, the full NeRF-Ag model produces relatively normal exposure results. This observation indicates that the proposed environmental factor embedding method in this paper further enhances the NeRF-Ag model's resistance to environmental interference and robustness.

4. Conclusions

To tackle issues including limited 3D reconstruction efficiency, substantial computational consumption, numerous constraints, and insufficient rendering accuracy within expansive scenes like orchards or agricultural fields, this study presents an inventive solution: the NeRF-Ag model, founded on implicit neural representation. Leveraging the basis of NeRF, this model integrates a multi-resolution latent feature encoding technique to amplify training efficiency and modeling accuracy. Additionally, environmental factor embedding techniques are employed to heighten the model's robustness and practicality. The following specific conclusions can be drawn from this study:

(1) Across the small, medium, and large scales, the NeRF-Ag model attains photo-realistic rendering outcomes, achieving image resolutions of 1920 × 1080. The NeRF-Ag model surpasses the baseline NeRF in regards to the evaluation metrics of PSNR, SSIM, and LPIPS. Furthermore, its training velocity is roughly 39 times quicker compared to that of the baseline NeRF.
(2) In comparison to the conventional 3D reconstruction technique grounded in COLMAP, NeRF-Ag demonstrates enhanced performance concerning texture intricacies and the precision of scene depth estimation. Moreover, this study accomplishes free-viewpoint rendering of 3D scenes encompassing various scales through the utilization of NeRF-Ag.
(3) By conducting pertinent ablation experiments, this study establishes a direct positive correlation between the quantity of training images and the 3D rendering precision of the NeRF-Ag model, manifesting saturation attributes. Besides, the technique of incorporating environmental factor embedding substantially bolsters the NeRF-Ag model's resilience against environmental

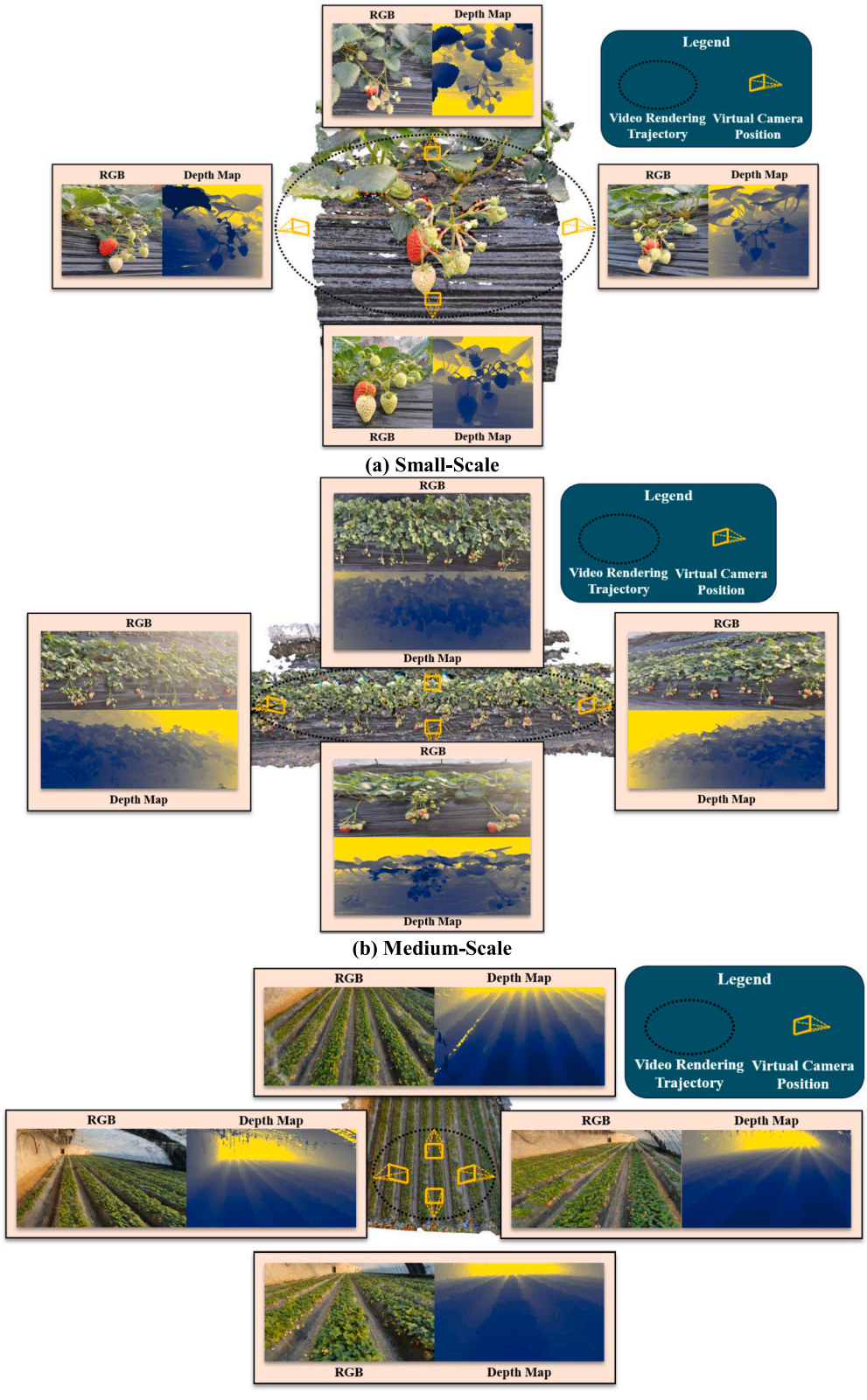


Fig. 7. NeRF-Ag based multi-scale, free-viewpoint 3D reconstruction and rendering.

disturbances, thereby providing a sturdy groundwork for its applicability in practical scenarios.

4.1. Limitations and future work

In medium-scale and large-scale scenes, ample potential for enhancing the 3D reconstruction of specific details and textures using

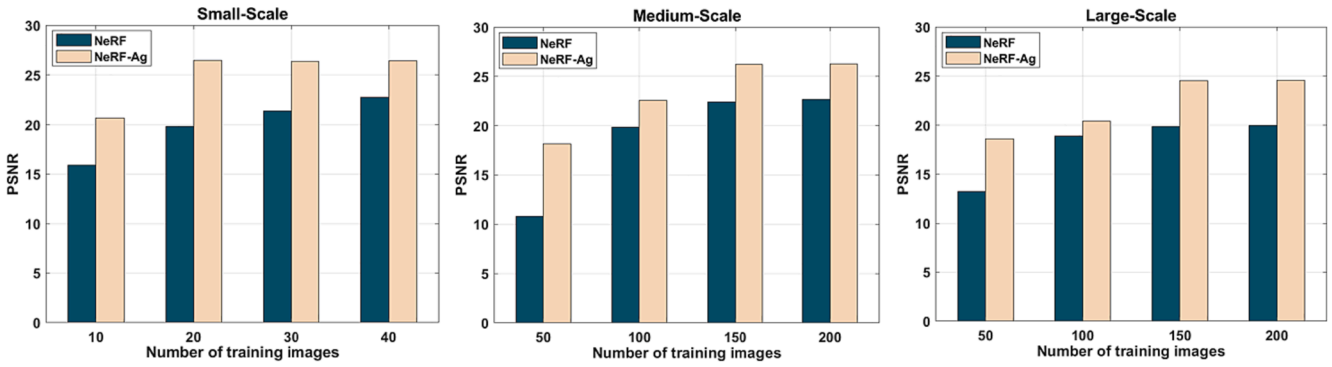


Fig. 8. Correlation analysis between the number of training images and 3D rendering accuracy at multi-scale.

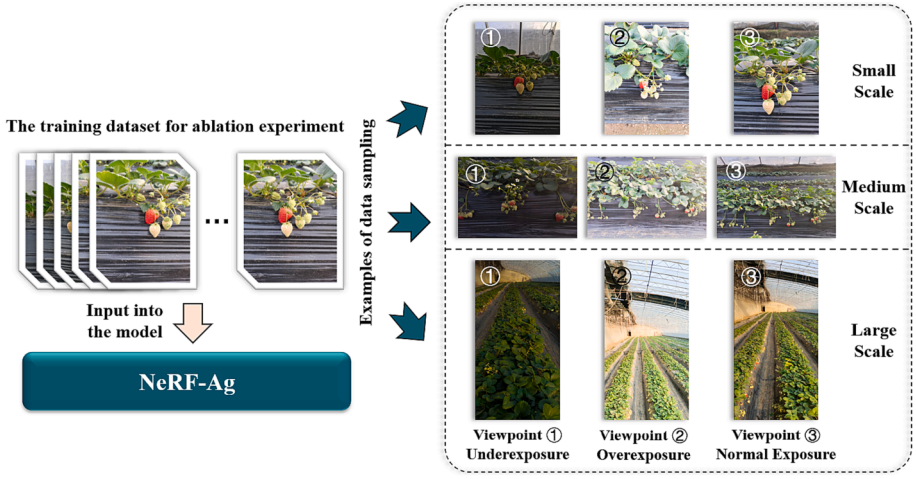


Fig. 9. Ablation experiment design for environmental factor embedding.

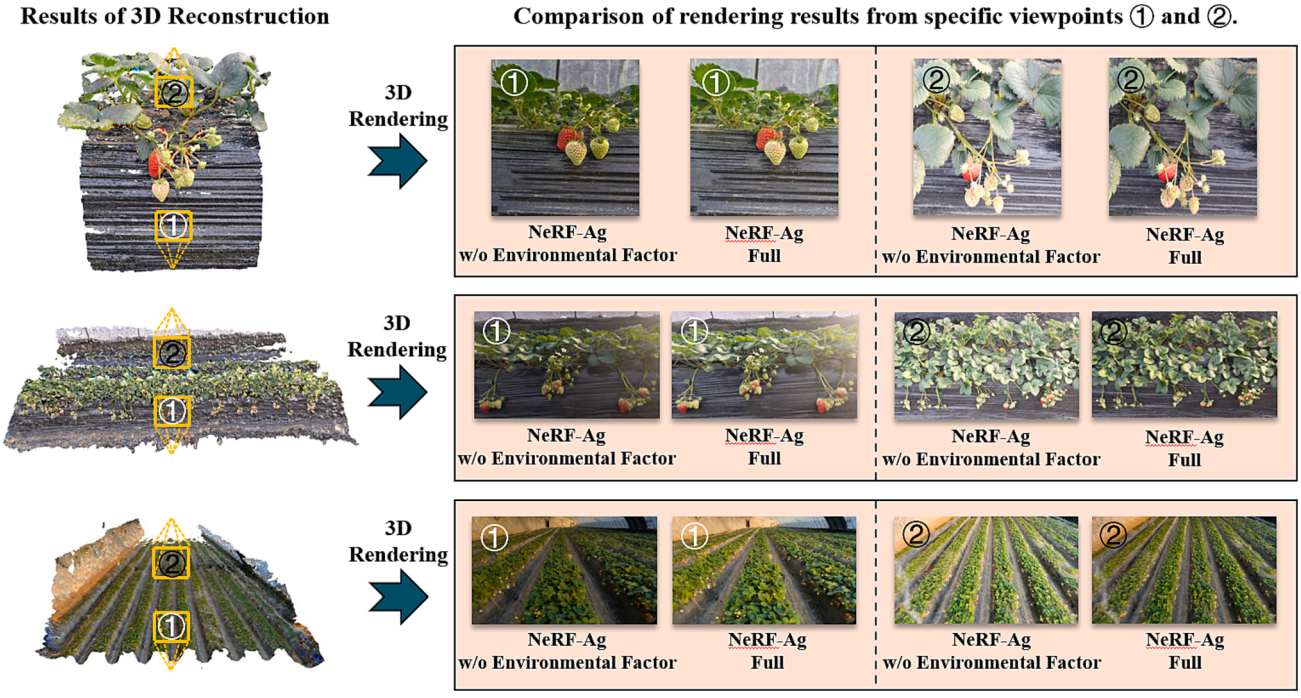


Fig. 10. Results of the ablation experiment on environmental factor embedding.

the NeRF-Ag model remains. Subsequent endeavors will be directed towards constructing NeRF models that integrate multi-scale and multi-modal information fusion, aiming to rectify the challenge of insufficient precision in detail reconstruction for vast scenes. Moreover, there exists the potential for further augmenting the training and rendering efficiency of the NeRF-Ag model. Therefore, refining the NeRF-Ag model and investigating real-time reconstruction and rendering techniques grounded in NeRFs will persist as pivotal avenues for prospective investigations.

CRediT authorship contribution statement

Jing Zhang: Conceptualization, Investigation, Methodology, Software, Writing – original draft, Writing – review & editing. **Xin Wang:** Investigation, Supervision. **Xindong Ni:** Investigation, Supervision. **Fangru Dong:** Data curation, Formal analysis, Validation. **Longrunmiao Tang:** Data curation, Formal analysis, Validation. **Jiahui Sun:** Data curation, Formal analysis, Validation. **Ye Wang:** Data curation, Formal analysis, Validation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This research was supported in part by the 2024 educational reform project of the Capital University of Economics and Business (CUEB) and the Beijing municipal universities basic scientific research operating expenses special funds (Grant No. 01892065119129). The authors would like to thank the KAIR lab at Berkeley AI Research (BAIR) and all subsequent contributors for developing and maintaining the Nerfstudio project.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compag.2024.108629.

References

Ariesen-Verschuur, N., Verdouw, C., Tekinerdogan, B., 2022. Digital Twins in greenhouse horticulture: a review. Comput. Electron. Agric. 199, 107183.
Barron, J.T., et al., 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.
Bojanowski, P., Joulin, A., Lopez-Paz, D. and Szlam, A., 2017. Optimizing the latent space of generative networks. arXiv preprint arXiv:1707.05776.
Cao, Z., et al., 2022. The algorithm of stereo vision and shape from shading based on endoscope imaging. Biomed. Signal Process. Control 76, 103658.
Chen, R., Han, S., Xu, J. and Su, H., 2019. Point-based multi-view stereo network, Proceedings of the IEEE/CVF international conference on computer vision, pp. 1538-1547.
Chen, A., Xu, Z., Geiger, A., Yu, J., Su, H., 2022. Tensorf: tensorial radiance fields. European Conference on Computer Vision. Springer 333–350.
Gao, L., Zhao, Y., Han, J., Liu, H., 2022. Research on multi-view 3D reconstruction technology based on SFM. Sensors 22 (12), 4366.
Goodfellow, I., et al., 2020. Generative adversarial networks. Commun. ACM 63 (11), 139–144.
Ham, H., Wesley, J., Hendra, H., 2019. Computer vision based 3D reconstruction: a review. International Journal of Electrical and Computer Engineering 9 (4), 2394.
Huang, Z., Yu, Y., Xu, J., Ni, F. and Le, X., 2020. Pf-net: Point fractal network for 3d point cloud completion, Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 7662-7670.
Ince, I.F., 2022. Robust image matching for information systems using randomly uniform distributed SURF features, applications of computational science in artificial intelligence. IGI Global 157–173.
Jung, S., Lee, Y.-S., Lee, Y., Lee, K., 2022. 3D reconstruction using 3D registration-based ToF-stereo fusion. Sensors 22 (21), 8369.
Li, J., et al., 2022. Multi-view real-time acquisition and 3D reconstruction of point clouds for beef cattle. Comput. Electron. Agric. 197, 106987.
Ma, B., Du, J., Wang, L., Jiang, H., Zhou, M., 2021. Automatic branch detection of jujube trees based on 3D reconstruction for dormant pruning using the deep learning-based method. Comput. Electron. Agric. 190, 106484.
Martin-Brualla, R., et al., 2021. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7210–7219.
Mildenhall, B., et al., 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. Commun. ACM 65 (1), 99–106.
Moghadam, P., Lowe, T., Edwards, E.J., 2020. Digital twin for the future of orchard production systems. Multidisciplinary Digital Publishing Institute Proceedings 36 (1), 92.
Müller, T., Evans, A., Schied, C., Keller, A., 2022. Instant neural graphics primitives with a multiresolution hash encoding. ACM Transactions on Graphics (ToG) 41 (4), 1–15.
Pylianidis, C., Osinga, S., Athanasiadis, I.N., 2021. Introducing digital twins to agriculture. Comput. Electron. Agric. 184, 105942.
Rivera, G., Porras, R., Florencia, R., Sánchez-Solís, J.P., 2023. LiDAR applications in precision agriculture for cultivating crops: a review of recent advances. Comput. Electron. Agric. 207, 107737.
Schonberger, J.L. and Frahm, J.-M., 2016. Structure-from-motion revisited, Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4104-4113.
Tancik, M., et al., 2022. Block-nerf: Scalable large scene neural view synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8248–8258.
Tancik, M. et al., 2023. Nerfstudio: A modular framework for neural radiance field development, ACM SIGGRAPH 2023 Conference Proceedings, pp. 1-12.
Vaswani, A. et al., 2017. Attention is all you need. Advances in neural information processing systems, 30.
Xie, H., Yao, H., Sun, X., Zhou, S., Zhang, S., 2019. Pix2vox: Context-aware 3d reconstruction from single and multi-view images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2690–2698.
Zhang, K., Riegler, G., Snavely, N. and Koltun, V., 2020. Nerf++: Analyzing and improving neural radiance fields. arXiv preprint arXiv:2010.07492.
Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O., 2018. The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 586–595.
Zhuo, Z., et al., 2022. 3D characterization of desiccation cracking in clayey soils using a structured light scanner. Eng. Geol. 299, 106566.