applied sciences

MDPI

Article

Multi-View Jujube Tree Trunks Stereo Reconstruction Based on UAV Remote Sensing Imaging Acquisition System

Shunkang Ling [1], Jingbin Li [1,*], Longpeng Ding [1,*] and Nianyi Wang [2]

[1] College of Mechanical and Electrical Engineering, Shihezi University, Shihezi 832003, China; 20212009037@stu.shzu.edu.cn
[2] College of Software Engineering, Xi'an Jiaotong University, Xi'an 710049, China; ny.wang@stu.xjtu.edu.cn
* Correspondence: lijingbin@shzu.edu.cn (J.L.); dy2016@shzu.edu.cn (L.D.)

Abstract: High-quality agricultural multi-view stereo reconstruction technology is the key to precision and informatization in agriculture. Multi-view stereo reconstruction methods are an important part of 3D vision technology. In the multi-view stereo 3D reconstruction method based on deep learning, the effect of feature extraction directly affects the accuracy of reconstruction. Aiming at the actual problems in orchard fruit tree reconstruction, this paper designs an improved multi-view stereo structure based on the combination of remote sensing and artificial intelligence to realize the accurate reconstruction of jujube tree trunks. Firstly, an automatic key frame extraction method is proposed for the DSST target tracking algorithm to quickly recognize and extract high-quality data. Secondly, a composite U-Net feature extraction network is designed to enhance the reconstruction accuracy, while the DRE-Net feature extraction enhancement network improved by the parallel self-attention mechanism enhances the reconstruction completeness. Comparison tests show different levels of improvement on the Technical University of Denmark (DTU) dataset compared to other deep learning-based methods. Ablation test on the self-constructed dataset, the MVSNet + Co U-Net + DRE-Net_SA method proposed in this paper improves 20.4% in Accuracy, 12.8% in Completion, and 16.8% in Overall compared to the base model, which verifies the real effectiveness of the scheme.

Keywords: 3D reconstruction; deep learning; multi-view stereo; remote sensing; feature extraction

Citation: Ling, S.; Li, J.; Ding, L.; Wang, N. Multi-View Jujube Tree Trunks Stereo Reconstruction Based on UAV Remote Sensing Imaging Acquisition System. *Appl. Sci.* **2024**, *14*, 1364. https://doi.org/10.3390/app14041364

Academic Editor: Yutaka Ishibashi

Received: 8 January 2024
Revised: 4 February 2024
Accepted: 5 February 2024
Published: 7 February 2024

1. Introduction

In recent years, modern agricultural production has moved toward precision, and information technology integration has become an unstoppable trend [1]. The popularization of advanced image acquisition equipment, such as panoramic cameras and drones, has made it possible to acquire agricultural remote sensing data efficiently, and the dynamic development of remote sensing and artificial intelligence technologies has expanded new directions for sustainable agricultural applications [2]. Drones have potential applications not only in agricultural production but also in many other fields [3,4]. The forest and fruit industry plays an important role in agriculture and is not only vital to economic development but also contributes to ecological conservation and the rural economy. As the main forest and fruit industry in Xinjiang region, the development and management of jujube orchards are of strategic significance and have positive impacts on improving yield and quality, resource conservation and environmental protection, and farmers' income and economic benefits [5]. It is worth mentioning that TabNet (Tabular Neural Networks), as a method for dealing with large amounts of structured data involved, is commonly used in various scenarios in the agricultural field, especially in agricultural data analysis, crop prediction, and optimization [6–8]. In this paper, we mainly utilize remote sensing, artificial intelligence, and other technologies to extract image information from the trunk of jujube trees and carry out three-dimensional reconstruction work on them so as to provide the necessary working basis for the subsequent management work of jujube tree gardens

(e.g., tree pruning [9,10], volume estimation of jujube [11], pest and disease monitoring [12,13], orchard management [14], etc.).

Using remote sensing and AI in sustainable agriculture poses challenges. For instance, video sequences from orchard cameras generate a substantial amount of redundant data. The quick and accurate extraction of required target image data from these sequences directly impacts the accuracy of subsequent AI model processing. Failures in this extraction process may lead to incorrect management decisions, potentially resulting in significant economic losses [15]. This paper begins with an algorithmic approach, using the key frame extraction algorithm based on target tracking to quickly extract high-quality remote sensing data from the orchard, and combined with the feature extraction enhancement network technology based on deep learning so as to efficiently and accurately generate the required three-dimensional model of the jujube tree, to help implement the refinement of the management and support for decision-making, and improve the efficiency of agricultural production [16].

This paper focuses on the MVS (Multi-view Stereo) technique, which can generate a 3D model of a scene target by using a set of target images from different viewpoints with known camera parameters [17]. Traditional MVS methods can usually be categorized into four types, including algorithms based on point clouds, voxels, variable polygon meshes, and depth maps. Although these methods achieve good results in ideal Lambertian scenes, they often fail to produce good reconstruction results in texture-poor, texture-repeated, or illumination-variable situations [18].

With the rapid development of artificial intelligence-related technologies such as deep learning, 3D reconstruction technology has ushered in great improvements. The application of convolutional neural networks to extract features and regularize cost volume makes MVS reconstruction more powerful. The first deep learning-based 3D reconstruction network in this field, MVSNet, was proposed by Yao et al. [19] but suffers from the problem of poor reconstruction accuracy. Aiming at the problem that the MVSNet model using recurrent neural network does not effectively consider the contextual information, Yan et al. [20] proposed the D2HC-RMVSNet architecture, where existing 3D reconstruction methods expand the receptive field in the feature extraction part by downsampling, which causes feature loss and thus affects the reconstruction completeness, and cited a new feature extraction structure, DRE-Net, which enlarges the receptive field by inflated convolution and no longer performs downsampling [21,22]. Overall, MVS inherits the stereo geometry theoretical basis of stereo matching, and with the help of more image viewpoints, it effectively improves the influence of the occlusion problem and achieves a big improvement in both accuracy and generalization. However, the existing methods still have room for improvement in terms of feature extraction effect and matching customers in specific practical agricultural production scenarios.

Aiming at the above practical problems of production, this paper is dedicated to the use of remote sensing and artificial intelligence technology to realize the rapid and high-quality acquisition of fruit tree information and realize accurate reconstruction, and makes the following contributions:

1. An automatic key frame extraction algorithm based on the DSST target tracking algorithm was designed to realize efficient and high-quality remote sensing image data acquisition of a multi-angle view of a jujube tree in a large-scale jujube tree garden.
2. The composite U-Net network structure is proposed as a feature extraction network model to make the extracted features more accurate.
3. The parallel self-attention mechanism is designed to improve the feature-enhanced improvement network of DRE-Net, which utilizes the self-attention mechanism and residual connectivity to make the model more complete.
4. For the cost volume regularization link, a multi-scale cost volume information fusion network based on a hierarchical feature body decoder is designed so that the information within the cost volume can be passed layer by layer to improve the estimation quality of the depth map.

2. Methods

2.1. Feature Extraction Network Construction

The first step in multi-view stereo matching reconstruction is to perform feature extraction on the image. Feature matching refers to the process of searching for feature points with the same name in the images captured by multiple viewpoints and multiple sensors and matching them together accurately. The main process is usually divided into feature extraction, feature matching, and mismatch point rejection. The feature matching algorithm with rotational invariance and scale invariance makes it possible to effectively improve the matching accuracy, but because the matching process is easily affected by factors such as light, noise, imaging angle, etc., the generation of mismatched points is unavoidable, and it is necessary to improve the process of feature matching to obtain high-precision matching results [23]. In image 3D reconstruction technology, the extraction and matching of feature points are crucial, and the use of feature extraction and matching algorithms that consider both accuracy and efficiency is the key to realizing high-precision image 3D reconstruction [24].

In previous studies, the feature extraction steps are mostly processed in a serial time sequence, and there are problems such as feature loss due to downsampling to expand the sensory field, blurring of the edge structure of the final point cloud, and ineffective differentiation between the depth of the object and the background region, among other issues [25]. To address the above problems, the input multi-view image first undergoes a 2D convolution to capture simple features of the image (e.g., lines, color blocks, etc.) to help the model understand the image and provide raw information for 3D reconstruction. Next, another 2D convolution is performed so that the network can start learning to understand more complex features (e.g., shapes, textures, etc.) and use them to obtain depth information. Then, into the feature extraction network, this paper combines the composite U-Net and self-attention mechanism to improve the DRE-Net (Density Reception Expanded) structure, as shown in Figure 1. This structure can extract more semantic information and semantic segmentation.

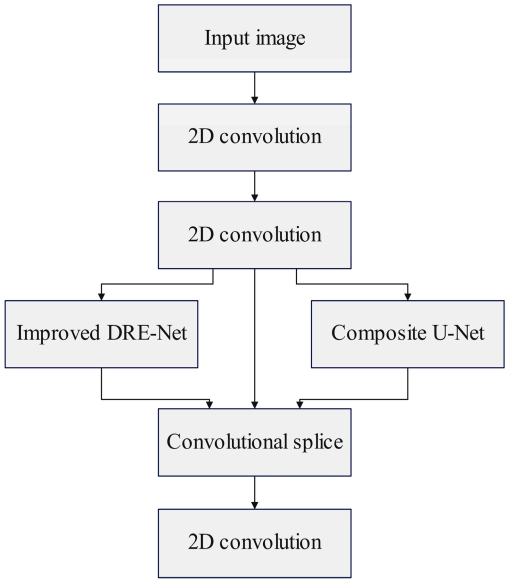

Figure 1. Feature extraction network structure.

2.1.1. Composite U-Net Feature Extraction Network

The U-Net network structure contains two main parts: Encoder and Decoder. The encoder is responsible for downsampling the input image, receiving the input image, and gradually reducing the size of the feature map through a series of convolutional layers and pooling layers to achieve the extraction of deep features in the image. However, in the process affected by the depth, the resolution of the image is reduced, and the number

of channels is increased. Although the most critical feature information in the image is obtained in this way, a lot of global detail information will be lost [26]. Therefore, a connection is established between the last layer of the encoder's feature map and the first layer of the decoder's feature map. The decoder receives the feature information transmitted from the encoder and gradually increases the size of the feature map through a series of convolutional layers and up-sampling operations to realize the superposition of the feature map fusion and thus increase the image detail information.

In the actual use of the scene, the orchard environment is complex fruit tree feature information, and the feature extraction network has higher requirements. To better extract the complete features of the input image so that the subsequent operation is more accurate, this paper introduces a U-Net network structure after the basic network so that after a convolution from coarse to fine convolution of its again convolution processing, the network structure as shown in Figure 2. The resolution is first reduced by maximum pooling operation and then gradually restored to the original resolution by up-sampling operation while keeping the number of channels unchanged. In the first half of the network, I.e., the encoding stage, the global features are extracted, and the input image is first subjected to a convolution operation with a convolution kernel size of 3 × 3, followed by four 2 × 2 convolutions and maximum pooling operations. In the up-sampling stage, the information of the feature maps with the same resolution located in the encoding and decoding structures, respectively, is concatenated and fused in the channel dimension, and after merging at the side, the convolution and up-sampling are continued to obtain the 32-channel feature maps with the same resolution as that of the original image, and then this is taken as an input to go through the network once more, and finally, three groups of feature maps with different sizes are obtained, and the number of channels corresponds to 8 channels in the order of the feature maps from the largest to the smallest, respectively. The corresponding channel numbers are 8, 16, and 32, and the resolution is changed to 1, 1/4, and 1/16 of the original image.

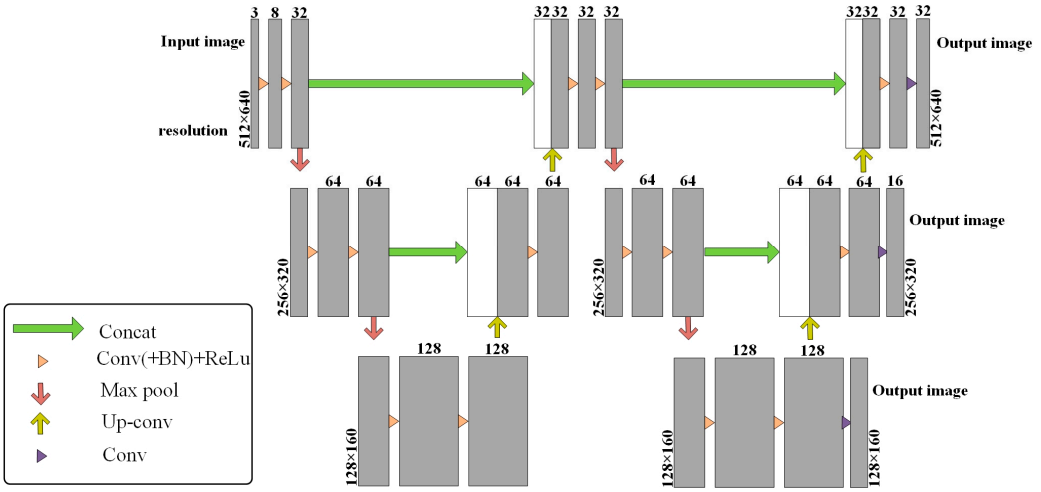

Figure 2. Composite U-Net feature extraction network architecture.

Such a composite U-Net feature extraction network model has the following characteristics:

- The tandem U-Net structure allows the network to consider information at multiple scales simultaneously, and the sensory field can be expanded gradually by stacking, which improves the understanding of the overall image structure.
- The overall network has a better immunity to the variation and noise of the input data, which helps to improve the network's generalization performance.
- The tandem U-Net structure usually achieves better performance in tasks with complex image features, especially in situations that require the fusion of global and local information. Complex image features, especially when global and local information

fusion is required, and this structure can be more flexible in dealing with targets of different sizes and complex scenes [27].

2.1.2. Feature Extraction Enhanced Network Module with Self-Attention

The Self-Attention mechanism allows the model to focus on different parts of the sequence when processing the sequence data instead of fixing the same weights for all parts, which is particularly helpful for dealing with long-distance dependencies. The self-attention mechanism can be used in the Transformer model, which is a deep-learning model that uses the self-attention mechanism to capture global information in the input sequence. Kitaev et al. [28] were the first to propose the transformer module for application in computer vision and showed through experimental results that the transformer module outperforms the convolutional neural network in the task of image restoration. Therefore, in this paper, we use the self-attentive transformer module to improve the DRE-Net network and apply it in the multi-view stereo matching 3D reconstruction feature extraction session to improve the image features through the self-attention mechanism, with the intention of enriching the semantics and improving the completeness of the reconstructed point cloud.

The structure of the self-attention module transformer used in this paper is shown in Figure 3. The structure first converts the image into an input in the form of image blocks and slides the convolution kernel in order to extract the image blocks, with the total number of image blocks $N = ((H - l)/s + 1) \times ((W - l)/s + 1)$, where the convolution kernel is of length and width l and the step size is s. Then, flattening is performed to get the input as $x' \in R^{N \times D}$, and the self-attention mechanism is applied to these blocks of images to be processed later. First, the image blocks are processed through the fully connected network layer to obtain (*key*) x_k, (*value*) x_v, (*query*) x_q in the self-attention mechanism, and the key and query are multiplied pointwise to obtain the weight matrix of the image block $x_w = x_q \cdot {x_k}^T$, where D' is the feature dimension of the D-transform.

$$x_k = x'W_k, W_k \in R^{D \times D'}, x_k \in R^{N \times D'} \quad (1)$$

$$x_q = x'W_q, W_q \in R^{D \times D'}, x_q \in R^{N \times D'} \quad (2)$$

$$x_v = x'W_v, W_v \in R^{D \times D'}, x_v \in R^{N \times D'} \quad (3)$$

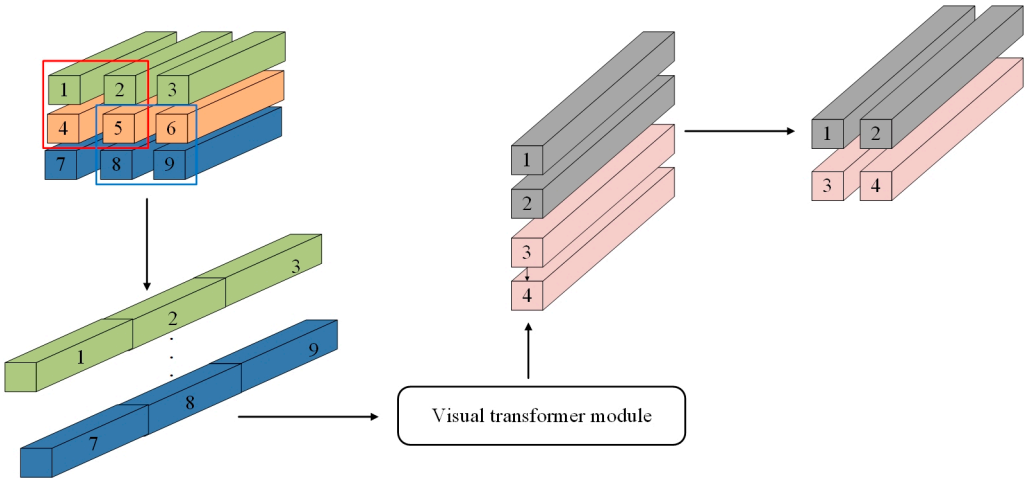

Figure 3. Structure of the self-attention mechanism module.

The weight matrix is normalized by the Softmax function and multiplied with (query) to get the result $x_{att} = x_w \cdot x_v$ of the image block after the action of the self-attention mechanism. All the image blocks are processed by the transformer module and the feature dimensions are changed to D', and finally the feature matrix $x_{att} \in R^{N \times D'}$ is transformed into the feature image resolution pattern to get $F \in R^{H \times W \times D'}$, where H, W, and D' are the length, width, and feature dimensions of the output.

Through the above process, the codec structure is constructed to learn the features in the image in the feature extraction session, and the important features in the image block are enhanced by considering the relationship between each pixel point in the image and the surrounding pixel points through the self-attentive transformer module [29].

2.1.3. Feature Extraction Enhanced Network with Improved DRE-Net

DRE-Net (Density Reception Expanded) is a deep learning network whose core idea is to use a recursive structure to capture complex patterns and hierarchies in images. In the recursive structure, the network processes data through multiple recursive layers, each of which can be considered as a deep learning model, and these layers are repeatedly applied to the input data, each time based on the previous output, and the network extracts and refines the features during the recursive process. The DRE-Net network extracts the semantic information within different receptive fields through convolution at different expansion scales, while spaced convolution operations can increase the receptive fields without affecting the resolution.

In this paper, three main improvements are made to the DRE-Net network:

1. The semantic features with expansion convolutional layers 2 and 3 are enhanced by the self-attention mechanism;
2. The convolutional layer with expansion scale four is deleted;
3. A symmetric coding and decoding network layer are added. The improved DRE-Net network structure is shown in Figure 4.

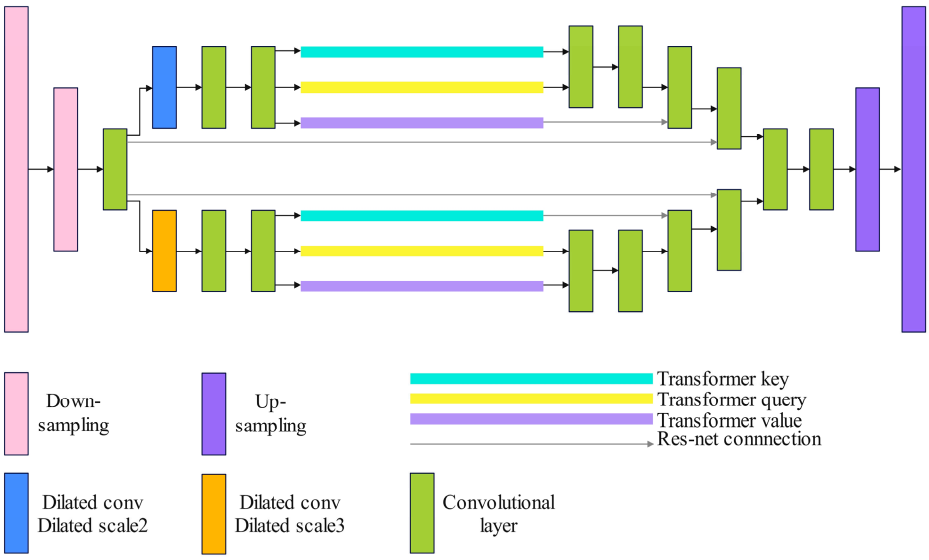

Figure 4. DRE-Net network architecture improved by self-attention mechanisms.

The process of the improved DRE-Net algorithm is as follows: firstly, select an uncomputed view in the dataset and downsample it twice with a step size of 2; secondly, convolve the downsampled results with an expansion scale of 2 and an expansion scale of 3, respectively; then, enhance the features by applying the self-attention mechanism and integrate the features by connecting them across the layers; then, upsample it twice with a step size of 2; and finally, store the computed feature maps.

2.2. Cost Volume Construction with Plane Scanning

The cost volume is a structure formed by the fusion of feature bodies from different perspectives, which mainly plays a role in measuring the similarity between the feature maps obtained by feature extraction and determining whether they match or not and the commonly used construction method is the plane scanning method. The process of constructing the cost volume is mainly divided into three steps:

1. Determine the discrete hypothetical depth planes, in which when the spatial resolution of the cost volume is fixed, the larger the number of hypothetical planes, the more accurate the result is accordingly, but it will increase the GPU memory and the running time.
2. Transform the feature maps extracted from each view into the hypothetical planes and construct the cost volume by using the microscale monoclinic transform, in which, at the depth d, the source image of the ith view is the same as the reference image. The reference image with monoclinicity $Hi(d)$ can be expressed as:

$$H_i(d) = K_i \cdot R_i \cdot \left(I - \frac{(t_0 - t_i) \cdot n_0^T}{d}\right) \cdot R_0^T \cdot K_0^T \quad (4)$$

where K_i, R_i, and t_i denote the internal reference, rotation, and translation of the source image camera of the ith viewpoint, n_0 denotes the principal axis of the reference camera, and I denotes the hypothetical planar interval, respectively.
3. The feature bodies of different viewpoints are formed into the feature body Fi by the univariate responsive change, and then all the feature bodies are aggregated into a single cost volume by the variance-based cost metric Fi, which is used to judge the similarity between N viewpoints and can be expressed as:

$$C = \frac{\sum_{i=1}^{N} \left(F_i - \overline{F}_i\right)^2}{N} \quad (5)$$

2.3. Multi-Scale Cost Volume Regularization

Since the just-constructed cost volumes may not satisfy the ideal Lambertian body structure on the surface due to mutual occlusion between fruit tree branches in the image or illumination, and there are network noise formation errors in the feature extraction network, etc., it is necessary to incorporate smoothness constraints on the initially constructed cost volumes in order to continue to estimate a more accurate depth map. The whole process is called cost volume regularization operation. In this paper, a hierarchical feature body-based decoder is used in the stereo matching problem, where the input feature body is divided into the output feature body and the output cost volume through 3D convolution and pyramid pooling [30]. This is performed by preprocessing the generated cost volumes before the regularization operation, separating the three cost volumes with different scale sizes, and fusing them into the cost volumes with a slightly larger scale layer, with the aim of gradually transferring the information stored in the bottom cost volumes to the top cost volumes to improve the accuracy of depth estimation.

To enhance the connection between different cost volumes, this paper designs a multi-scale cost volume information fusion network in the preprocessing stage of cost volume regularization. The role of this network is to improve the estimation quality of the depth map by separating the cost volumes generated in each layer and fusing them into the next layer so that the information on the small-scale cost volumes is fused into the cost volumes of the next layer, and the operation flow is shown in Figure 5.

After the initial processing of the cost volumes, the cost volumes C_f are obtained, and the next step is to perform a regularization operation on each cost volume C_f. The process is to input the cost volumes C_f into a 3D U-Net regularization network to extract the detailed information in the cost volume. Finally, after the 3D convolutional neural network, a single-channel probabilistic body will be obtained, whose role is to be used for predicting the depth value so that the generated probabilistic body can be used to predict the depth of individual pixels step by step as well as to measure the confidence of the depth prediction.

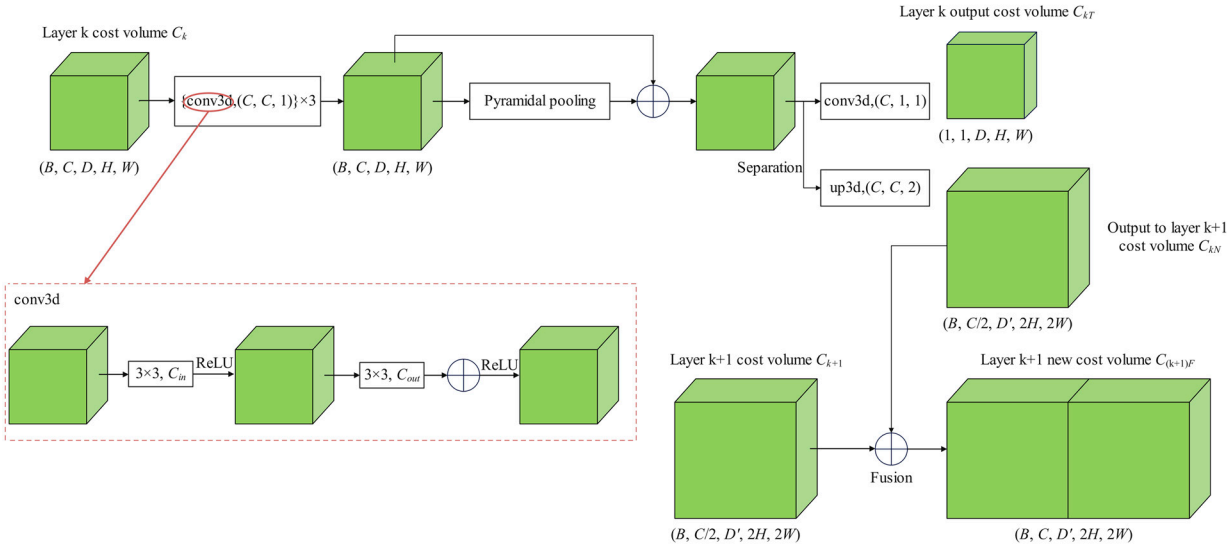

Figure 5. Schematic of multi-scale cost volumes fusion network.

3. Dataset

3.1. Image Capture

The data for this study were sampled on 20 March 2023 in the dwarf and densely planted jujube cultivation demonstration garden of the 224th Regiment, Kunyu City, 14th Division, Xinjiang Production and Construction Corps. The jujube plantation covers an area of about 103.3 km^2, and the planting mode is standardized: four rows are in a column team, the distance between rows is 2 m, and there is a 4m wide mechanized operation channel between each column team. The jujube trees are evenly arranged, and the overall straightness is good, providing favorable conditions for UAV route planning and data collection, as shown in Figure 6.

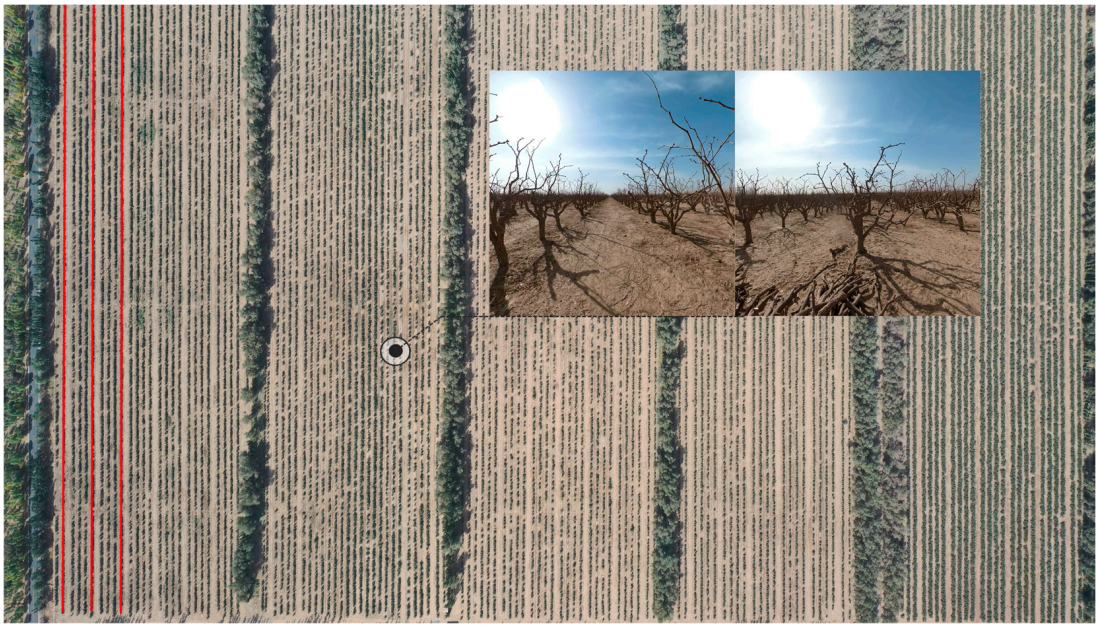

Figure 6. Schematic diagram of straightness of standardized jujube garden.

In the process of jujube tree video acquisition, the panoramic camera has distinct advantages and characteristics, so the UAV-mounted panoramic camera is used for data acquisition. The panoramic camera has a high-resolution and comprehensive perspective and can capture the multi-view image of each jujube tree at one time, which greatly improves the work efficiency; at the same time, it can also understand the spatial structure distribution of the jujube tree more comprehensively, which helps to accurately construct the jujube tree model.

The jujube tree image acquisition equipment used in this study is: The jujube tree image acquisition equipment used in this study is an Insta360 x2 panoramic camera, which has an aperture of F2.0 with an equivalent focal length of 7.2 mm, and the image settings are set to record at 30 fps at 5.7 K. The panoramic camera is suspended and mounted via a DJI Mavic3 drone in the form shown in image acquisition equipment of Figure 7. To ensure the consistency of data collection, the UAV was set to route flight mode, with a constant route height of 1.4 m (the height of the center of the jujube tree crown) and a path set to fly along the center of the jujube tree rows at a constant speed in a straight line. Due to the mobile shooting process, it will inevitably appear to motion blur, but hope that in the case of the date palm is clear enough to ensure the sampling efficiency. Firstly, we set the motion blur tolerance as the camera's moving distance in each frame is set to be less than 5% of the target size; secondly, based on the target blur length combined with the camera's frame rate, resolution, field of view angle hardware parameters can be calculated as the moving distance of the image in the photographic element, and then theoretically launched the range of moving speed. However, in the actual orchard environment, due to the light intensity and the stability of the UAV movement and other factors, to ensure the actual shooting effect, we have conducted a variety of speed attempts and finally set a reasonable flight speed of 1.2 m/s.

The panoramic camera has excellent advantages, but the panoramic lens will inevitably produce image distortion in physical imaging, forming wide-angle aberrations leading to 3D reconstruction failure. Therefore, we adopt the polar coordinate mapping model to eliminate the distortion [31]. The polar coordinate mapping model utilizes the idea of the polar coordinate system, which is to convert the original pixel point from a right-angle coordinate system (x, y) to a polar coordinate system (r, θ), where r is the distance from the origin (the center of the image) to the point, and θ is the angle from the origin to the point. During polar mapping, a pixel point on a panoramic image is determined by its coordinates in polar coordinates, which means that each pixel point is equivalent to a point radiating outward from the center, which helps us to adjust the shape of the image to overcome the problems generated by lens distortion.

The initial image data is the panoramic image. In order to get the multi-view image of jujube trees, it needs to be pre-processed. The process is as follows: the panoramic image is divided into two 180° wide-angle fisheye images according to the direction perpendicular to the rows of jujube trees, and then the two sections of the image will be corrected for aberrations using polar coordinate mapping and cropping around the deformation of the stretched portion; and finally, adjust the field-of-view to the height of the screen to be able to accommodate standardized cultivation of jujube trees, and obtain the jujube trees front and side view image data.

3.2. Multi-View Image Extraction

Due to the dwarfed and densely planted plantation in the area, the jujube trees are close to each other, and there are cases of mutual occlusion, scale change, and transient loss of the target field of view. Target detection is also an important part of efficiently and accurately collecting multi-view images of the jujube garden. In this paper, Accurate Scale Estimation for Robust Visual Tracking (DSST) is used to accurately track the target of jujube trees according to the difficulties of the actual task.

DSST algorithm is an improvement of the Kernelized Correlation Filter (KCF), which has the advantages of a simple algorithm, excellent performance, strong portability, etc. DSST algorithm introduces a multi-feature fusion mechanism, which combines HOG (Histogram of Orientation Gradient) features, CN (Color Name) features, and grayscale features, and adopts the way of adding scale filters to adapt to the scale change of the target during the tracking process. The DSST algorithm uses a Translation Filter to predict the location of the target in the next frame, a Scale Filter to adaptively estimate the target scale, and a Response Score is obtained through the correlation filter operation, with the maximum value being the current target location and scale [32].

The specific operational steps for dataset acquisition are as follows:

1. Use a position filter for position estimation of the tracked target. Obtain the correlation filter ft for a certain position through training and use the filter to predict the target position in the next frame. When all the image blocks in the d-1 frame are operated with the filter, the maximum value of the response score of the image block is the estimation of the new position of the target, and finally, the parameters of the ftd filter are updated, and read into the next frame, and repeatedly iteratively updated until the end of the tracking, and the schematic diagram of the training process is shown in Figure 8. In which the response at the center of the target is the largest, and then gradually decreases from the center to the surrounding area, and finally according to the location of the maximum value of the response, thus realizing the accurate positioning of the target.
2. Use a scale filter for scale estimation of the tracked target. To overcome the problem of target scale change of "near target is big, the far target is small", a one-dimensional scale filter is created to estimate the scale in the image, extract the features at the center position of the target, and calculate the training samples for updating the scale filter. In the actual tracking process of the orchard, the position transformation of the target in two consecutive frames is larger than the scale transformation. Therefore, the DSST algorithm firstly captures the position information of the target in the current frame through the position filter and, secondly, captures the scale information of the target by using the scale filter and the schematic diagram of the training process is shown in Figure 9. The response is largest at the center point and decreases to both ends, and the scale corresponding to the largest value of the response is the result of the final scale estimation of the target.
3. Use keyframe automatic extraction algorithm for target multi-view image extraction. Determine the fruit tree target to always keep in the center of the screen as the key frame selection conditions, on the basis of the DSST target tracking algorithm, realize the tracking and keeping of the target, and extract the multi-view image of a single jujube tree at equal intervals, and the same operation traverses all the jujube trees.
4. Combine the sensor data to construct the 3D reconstruction dataset. The automatic extraction algorithm extracts the target multi-view image data outside. The camera outputs focal length, position, and view angle information, and the scale filter outputs target scale information. According to the scale focal length and other information, crop the size to the target equal size. Traversal operation of all jujube trees to construct the 3D reconstruction dataset. The target multi-view image acquisition process is shown in Figure 7.

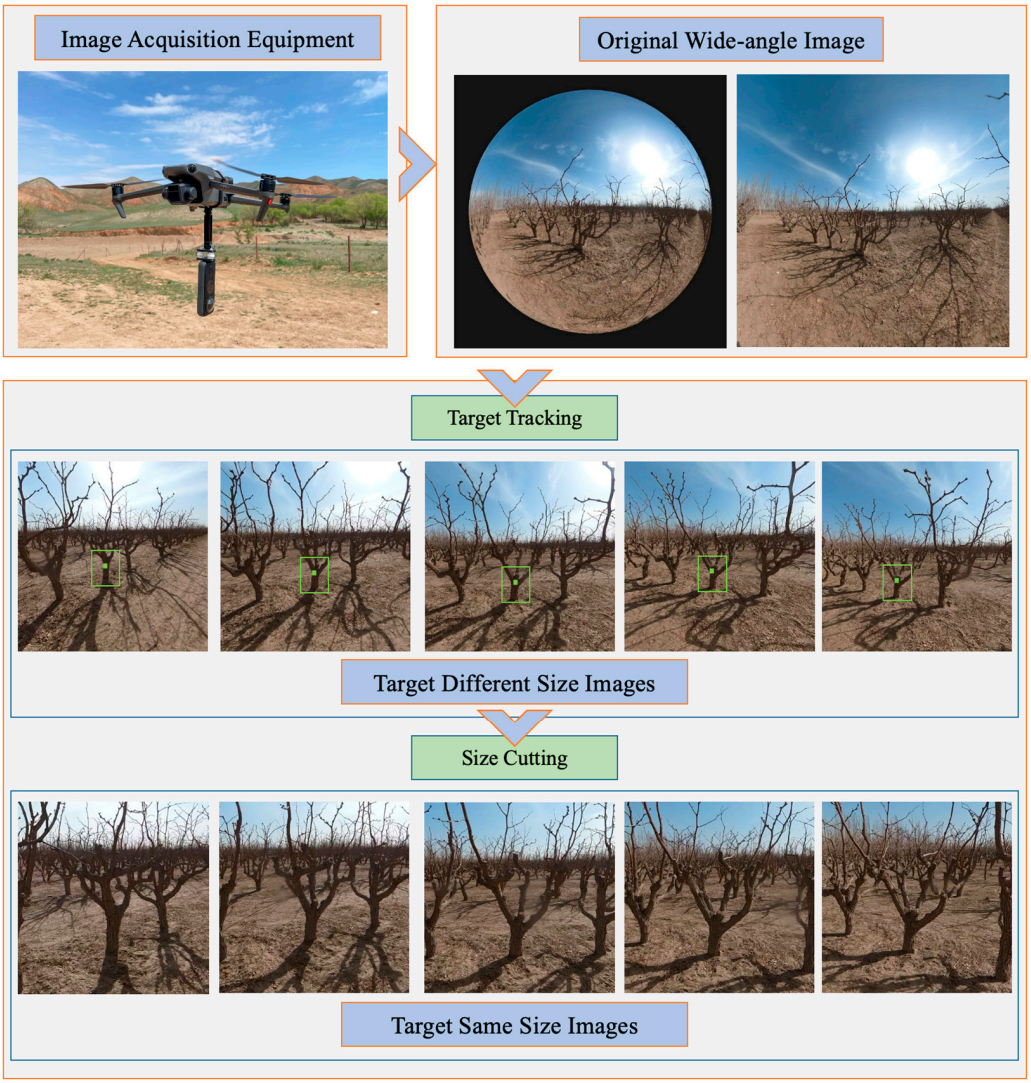

Figure 7. Multi-view data acquisition process.

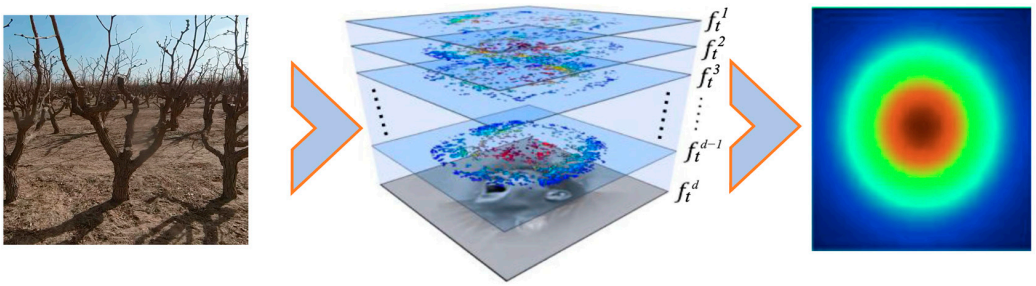

Figure 8. Translation filter training process.

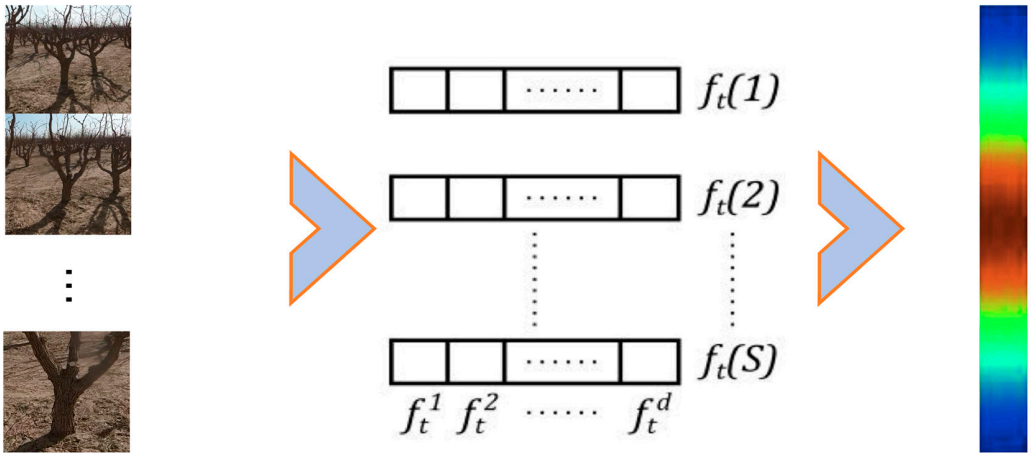

Figure 9. Scale filter training process.

4. Experiments

4.1. Experimental Settings

The experimental environment of this paper is: CPU is Intel I7-8700K (Intel, Santa Clara, CA, USA); RAM is 16G; GPU is NVIDIA GTX3090 (24G) (NVIDIA, Santa Clara, CA, USA), the operating system is Ubuntu20.04 system, Cuda is 11.0, and the experimental process mainly relies on the computing environment of GPU to run. In addition, Pytorch 1.7.2 was used as the algorithmic framework, and the algorithmic code was written through Python 3.8.0. At the same time, we use Anaconda multi-environment management software to configure the environment that the algorithm relies on, which can effectively avoid the problems that may be caused by the conflict between Cuda and Pytorch versions.

In this paper, the input image resolution is set to 512 × 640, and the input consists of a reference view and two neighboring source views. For the depth of the hypothesis layer, this paper is set to 48, 32, and 8. The interval is 4, 2, and 1 times the MVSNet interval, and the feature map resolution becomes 1/16, 1/4, and 1 of the original map. After many attempts, the learning rate is initialized to 0.001 in training, the learning rate is set to decay with the number of iterations to avoid overfitting, and the learning rate is proportional to the decay, the training period is 16 rounds, the decay step is 100,000, the decay factor 0.5.

4.2. Evaluation Indicators

In order to quantitatively assess the effectiveness of the method in this paper, the evaluation indexes in the experiments are Accuracy (Acc), Completion (Comp), and Overall, the units are millimeters (mm). Acc calculates the average distance from the point cloud reconstructed by the model to the corresponding points of the real point cloud; Comp calculates the average distance from the real point cloud to the point cloud reconstructed by the model; and Overall calculates the average value of Acc and Comp, which reflects the overall quality of the reconstruction results.

4.3. Comparative Analysis

Firstly, an image group consisting of a reference image and a source image is input, which is a feature extracted by the model. The feature map is divided into three different scales, large, medium, and small, through the composite U-Net feature extraction network and the improved DRE-Net feature enhancement network, and the small-scale feature map is processed by the attention mechanism to enhance its features; then the feature map extracted from each view is transformed to the hypothesis plane through the microsimplex responsive transform to construct the feature body, and the feature bodies of the different views are fused together to form a three-dimensional cost body; regularize the cost body to get a single-channel probabilistic body; then recover the depth map from the probabilistic

body, and finally optimize and fuse the depth map until all layers are constructed, and finally output to get the 3D model of the scene.

In order to compare the traditional 3D reconstruction algorithms, deep learning-based 3D reconstruction algorithms, and many other algorithms, this paper has carried out a comparison test with them, taking into account the cost of resource training, the dataset used for the comparison test is the 3D reconstruction universal dataset DTU [33], which is categorized in accordance with the training set, the validation set and the test set. There are 119 different scenes or objects in the dataset, which contains 79 scenes for training, 18 scenes for validation, and 22 scenes for testing. The test results are shown in Table 1.

Table 1. Comparison of comparative test results.

Methods	Acc	Comp	Overall
Gipuma [34]	0.283	0.873	0.578
CAMP [18]	0.835	0.554	0.695
COLMAP [35]	0.400	0.664	0.532
MVSNet(baseline)	0.396	0.527	0.462
R-MVSNet [19]	0.385	0.421	0.391
D2HC-MVSNet [20]	0.395	0.378	0.386
Point-MVSNet [36]	0.361	0.421	0.391
Cascade-MVSNet [37]	0.325	0.385	0.355
MVSNet + Co U-Net + DRE-Net_SA	0.354	0.338	0.346

From Table 1, we can see that compared with the traditional supervised 3D reconstruction methods, such as Gipuma [33], CAMP [34], COLMAP [18], etc., the method used in this paper ranks second in terms of accuracy, and the first one is the Gipuma method, which cannot be surpassed by other deep learning-based methods at present. However, deep learning-based methods can not only achieve significant results in reconstructing Comp and Overall but also have a huge advantage in terms of running time and memory consumption.

Comparing the unsupervised 3D reconstruction methods based on deep learning, such as MVSNet, D2HC-MVSNet [20], etc., the MVSNet + Co U-Net + DRE-Net_SA method proposed in this paper also shows significant improvement in Acc and Comp. Cascade-MVSNet achieves the best performance in Acc. However, it cannot simultaneously consider the reconstruction accuracy and completeness, and the model in this paper achieves the best performance in terms of reconstruction accuracy, completeness, and comprehensive evaluation indexes.

4.4. Ablation Analysis

To verify the validity of the progression scheme proposed in this paper, the ablation test is designed according to each improvement module. This test uses the same equipment and self-built datasets from field samples for training and testing to ensure comparability, and the comparison graph with the base model test is shown in Figure 10. Where Figure 10(a1–a3,b1–b3 and c1–c3) shows the localized zoomed in view of the results generated by each model.

The base model MVSNet's multi-view Figure 10(a1–a3), affected by the complex environment of the orchard, there is interference information in the trunks of the fruit trees, resulting in poor model integrity, and cannot well distinguish the main model structure.

Comparing the trunk detail maps generated by MVSNet + U-Net feature extraction network Figure 10(b1–b3) with Figure 10(c1–c3) of MVSNet + Co U-Net + DRE-Net_SA method, it can be seen that b1 is more than c1 the trunk portion of the tree is missing incomplete and the transition is not natural; Figure 10(c2) has a better effect than Figure 10(b2) the elimination of the edge and the background noise. Effective segmentation of the background interference; Figure 10(c3) than Figure 10(b3) more effective extraction of image detail features, in the texture sparse region rich semantic information, the reconstruction of the object edge of the smoothness and completeness have been greatly improved.

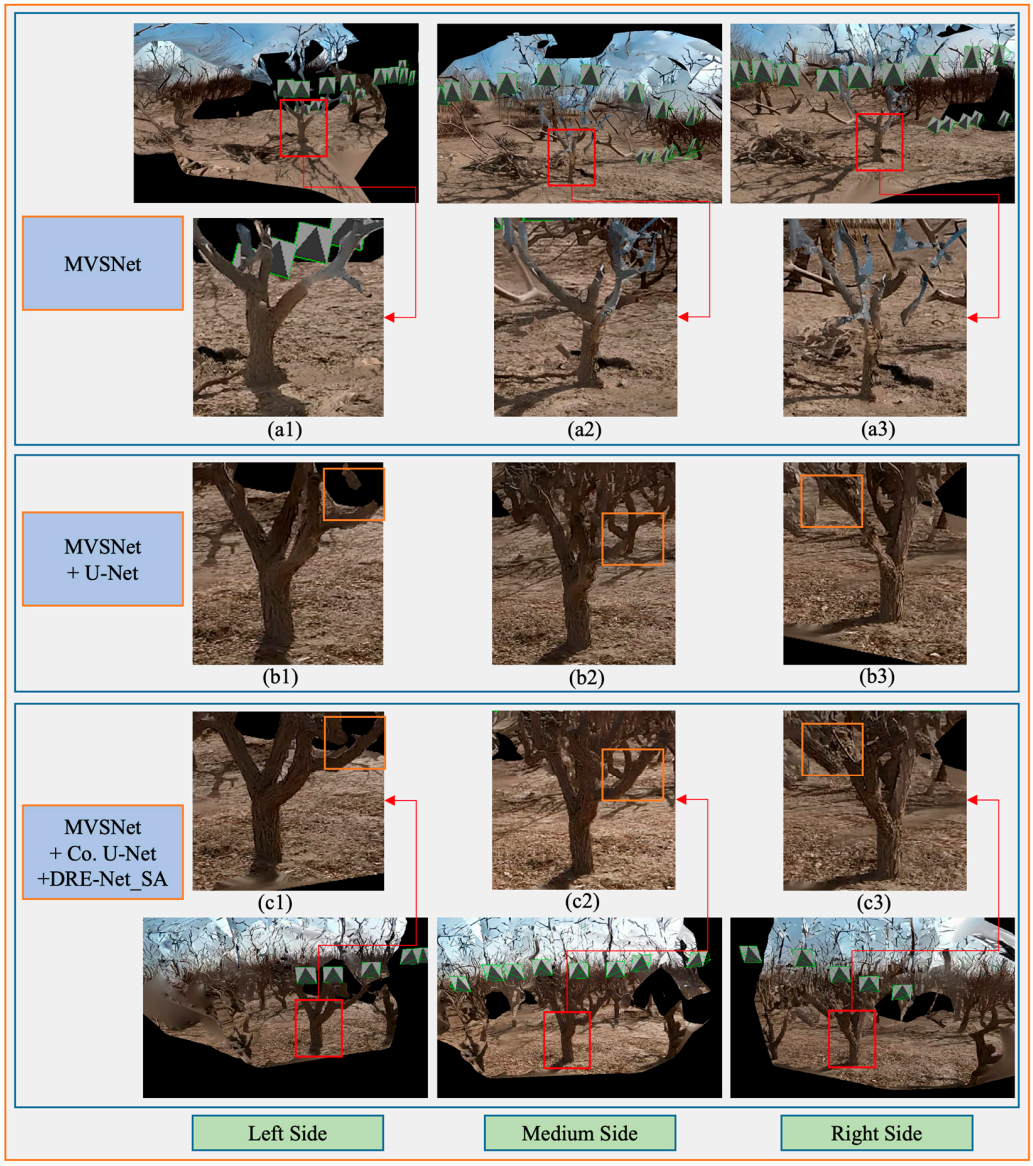

Figure 10. Comparison of reconstruction results of datasets by different models.

The data from the ablation test results are shown in Table 2, from which we can know by analysis.

Table 2. Comparison of ablation test results.

Methods	Acc	Comp	Overall
MVSNet(baseline)	0.560	0.538	0.549
MVSNet + U-Net	0.502	0.486	0.494
MVSNet + Co. U-Net	0.486	0.492	0.489
MVSNet + Co U-Net + DRE-Net_SA	0.446	0.469	0.457

Firstly, the composite U-Net feature extraction network structure used in feature extraction is tested, and the feature extraction network with a single U-Net structure is compared based on the base model. From the first and second rows, the performance of the base model has been optimized in each index after adding the U-Net feature extraction network, which proves that the base model underperforms in the complex orchard environment and demonstrates the feasibility of the improvement direction of the feature extraction network. As can be seen from the third line, compared with the single U-Net feature extraction

network, the use of a composite U-Net feature extraction network can try the model in the convolution process more times, deeper depth, and better semantic segmentation of the original image of the orchard, so that the performance of Acc and Overall metrics has improved. However, the performance of the Comp metrics decreased, and we analyzed the reasons that led to the performance degradation:

- Since the composite U-Net architecture is more efficient in processing specific features (e.g., edges, textures, etc.), it is not good at capturing the full details of the scene, which affects the completeness of the model's reconstruction of the objects.
- Composite U-Net may lead to a certain degree of overfitting because of its complex architecture and larger model capacity, especially when there is less data or insufficient diversity.

Secondly, to address the situation that the above improvement leads to a decrease in completeness, this paper improves the DRE-Net network structure by adding a self-attention mechanism to the feature extraction network in parallel. Relative to the third line from the fourth line, the model, after the enhancement of the feature network, does not degrade and improves the performance of the Comp metrics while ensuring a slight improvement in the performance of the Acc and Overall metrics. This is due to the characteristics of the architecture:

- The self-attention mechanism enables the network to pay more attention to the key information in the input data, which reduces the risk of over-adaptation to the training data by reducing the learning of unimportant features by the model and improving the reconstruction by highlighting the important features.
- DRE-Net enhances the feature transfer and learning capability through dense residual connections, which helps to retain detailed information from the input to the output. Detailed information and the network avoids the risk of overfitting due to the overly complex model architecture when dealing with deep-level features.
- The parallel use of DRE-Net on top of the composite U-Net improves the feature fusion process, and effective feature fusion helps to synthesize different levels of information, making the final reconstruction results more accurate and complete.

Therefore, in general, the MVSNet + Co U-Net + DRE-Net_SA method proposed in this paper has a large improvement in each performance index relative to the base model, with Acc improved by 20.4%, Comp by 12.8%, and Overall combined by 16.8%.

5. Conclusions

The development of modern cameras, drones, and other image acquisition equipment, as well as artificial intelligence remote sensing integration technology, has greatly contributed to the advancement of precision agriculture applications. In this paper, a solution is proposed for the use scenario of rapid 3D reconstruction of jujube trees in Xinjiang jujube garden, and the specific conclusions are as follows:

Proposed a feature extraction network model based on the composite U-Net network structure for feature extraction. Aiming at the specific image sampling process in the jujube garden, jujube trees are briefly obscured, scale change occurs, and the target field of view is briefly lost. The accurate target tracking of jujube trees is realized by using the translation filter and scale filter that introduces the multi-feature fusion mechanism. Combined with the key frame automatic extraction algorithm, the high-efficiency multi-view angle and equal-scale image acquisition of jujube tree trunks are realized.

Enhancement of feature extraction network using DRE-Net structure improved by self-attention mechanism. Aiming at the multi-view image of the jujube tree, the growth of branches is complicated, too much interference information, and the situation of feature extraction is not obvious. In the feature extraction network model based on a convolutional neural network, the most critical deep feature information in the image is effectively extracted by continuously performing convolution operations on the image. The U-Net-based composite improvement network model can retain richer detailed features while obtaining

feature maps with different scales, improving the characterization ability of features, expanding the range of sensory fields of features, and retaining the global information of the original image and perfecting the details through feature splicing and fusion, so that the subsequent depth estimation results are more accurate and complete.

This paper proposes a self-attention mechanism to improve the feature enhancement network of DRE-Net. Aiming at the decrease of completeness in the process of jujube trunk feature extraction. Using the characteristics of the self-attention mechanism to pay more attention to the key information in the input data and the effect of DRE-Net to avoid overfitting through dense residual connections, the feature extraction network is enhanced to extract more semantic information of the image and realize the performance improvement of the double indexes of accuracy and completeness.

In the link of cost volume regularization, a multi-scale cost volume information fusion network based on a hierarchical feature body decoder is designed to increase the interconnection between different layers of cost volumes so that the information within the cost volumes can be transferred layer by layer to improve the estimation quality of the depth map.

The algorithm proposed in this paper is compared with other traditional reconstruction and deep learning-based 3D reconstruction methods on the DTU dataset, and compared with the traditional algorithm, it performs better in terms of completeness, accuracy, and robustness; and compared with similar methods based on deep learning, the model in this paper performs more balanced. After the ablation test of the actual dataset of jujube trees, all modules of the improved model in this paper effectively improve the accuracy and completeness of the multi-view stereo matching 3D reconstruction, which achieves good results in practice and puts forward a feasible program reference for the integration of remote sensing and artificial intelligence into the existing agricultural system.

However, there is still room for further improvement and enhancement of the work in this paper. Due to the increase of convolutional scale, this leads to the increase of GPU's occupied memory and the increase of training cost. In future research work, we will adjust the network structure or utilize multi-GPU parallelism to reduce memory consumption and improve efficiency.

Author Contributions: Conceptualization, S.L. and N.W.; methodology, S.L. and N.W.; investigation, S.L. and L.D.; data curation, S.L.; writing—original draft preparation, S.L. and N.W.; writing—review and editing, S.L., J.L. and L.D.; funding acquisition, J.L. and L.D. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (52165037), and the Corps Regional Innovation Guidance Program (2021BB003).

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Acknowledgments: We would also like to thank all reviewers for their valuable comments.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Nie, J.; Wang, Y.; Li, Y.; Chao, X. Sustainable computing in smart agriculture: Survey and challenges. *Turk. J. Agric. For.* **2022**, *46*, 550–566. [CrossRef]
2. Ge, X.; Ding, J.; Jin, X.; Wang, J.; Chen, X.; Li, X.; Liu, J.; Xie, B. Estimating Agricultural Soil Moisture Content through UAV-Based Hyperspectral Images in the Arid Region. *Remote Sens.* **2021**, *13*, 1562. [CrossRef]
3. Liu, Z.; Song, Y.; Gao, S.; Wang, H. Review of Perspectives on Pantograph-Catenary Interaction Research for High-Speed Railways Operating at 400 km/h and above. *IEEE Trans. Transp. Electrif.* **2023**, 1. [CrossRef]
4. Jenie, Y.I.; van Kampen, E.J.; Ellerbroek, J.; Hoekstra, J.M. Safety assessment of a UAV CDR system in high density airspace using monte carlo simulations. *IEEE Trans. Intell. Transp. Syst.* **2018**, *19*, 2686–2695. [CrossRef]
5. Zhang, R.; Li, P.; Xu, L.; Zhong, S.; Wei, H. An integrated accounting system of quantity, quality and value for assessing cultivated land resource assets: A case study in Xinjiang, China. *Glob. Ecol. Conserv.* **2022**, *36*, e02115. [CrossRef]

6. Arik, S.Ö.; Pfister, T. Tabnet: Attentive interpretable tabular learning. In Proceedings of the AAAI Conference on Artificial Intelligence, Online, 2–9 February 2021; Volume 35, pp. 6679–6687.
7. de Zarzà, I.; de Curtò, J.; Calafate, C.T. Area Estimation of Forest Fires using TabNet with Transformers. *Procedia Comput. Sci.* **2023**, *255*, 553–563. [CrossRef]
8. Shah, C.; Du, Q.; Xu, Y. Enhanced TabNet: Attentive interpretable tabular learning for hyperspectral image classification. *Remote Sens.* **2022**, *14*, 716. [CrossRef]
9. Ma, B.; Du, J.; Wang, L.; Jiang, H.; Zhou, M. Automatic branch detection of jujube trees based on 3D reconstruction for dormant pruning using the deep learning-based method. *Comput. Electron. Agric.* **2021**, *190*, 106484. [CrossRef]
10. Li, Y.; Zhang, Z.; Wang, X.; Fu, W.; Li, J. Automatic reconstruction and modeling of dormant jujube trees using three-view image constraints for intelligent pruning applications. *Comput. Electron. Agric.* **2023**, *212*, 108149. [CrossRef]
11. Li, J.; Wu, M.; Li, H. 3D reconstruction and volume estimation of jujube using consumer-grade RGB-depth sensor. *IEEE Access* **2023**. [CrossRef]
12. Li, Y.; Ercisli, S. Data-efficient crop pest recognition based on KNN distance entropy. *Sustain. Comput. Inform. Syst.* **2023**, *38*, 100860. [CrossRef]
13. Yang, J.; Ma, S.; Li, Y.; Zhang, Z. Efficient data-driven crop pest identification based on edge distance-entropy for sustainable agriculture. *Sustainability* **2022**, *14*, 7825. [CrossRef]
14. Yang, Y.; Li, Y.; Yang, J.; Wen, J. Dissimilarity-based active learning for embedded weed identification. *Turk. J. Agric. For.* **2022**, *46*, 390–401. [CrossRef]
15. Chao, X.; Li, Y. Semisupervised few-shot remote sensing image classification based on KNN distance entropy. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 8798–8805. [CrossRef]
16. Huang, B.; Yi, H.; Huang, C.; He, Y.; Liu, J.; Liu, X. M3VSNet: Unsupervised multi-metric multi-view stereo network. In Proceedings of the 2021 IEEE International Conference on Image Processing (ICIP), Anchorage, AK, USA, 19–22 September 2021; pp. 3163–3167.
17. Yu, Z.; Gao, S. Fast-mvsnet: Sparse-to-dense multi-view stereo with learned propagation and gauss-newton refinement. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 1949–1958.
18. Campbell, N.D.; Vogiatzis, G.; Hernández, C.; Cipolla, R. Using multiple hypotheses to improve depth-maps for multi-view stereo. In Proceedings of the Computer Vision–ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, 12–18 October 2008; pp. 766–779.
19. Yao, Y.; Luo, Z.; Li, S.; Fang, T.; Quan, L. Mvsnet: Depth inference for unstructured multi-view stereo. In Proceedings of the European conference on computer vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 767–783.
20. Yan, J.; Wei, Z.; Yi, H.; Ding, M.; Zhang, R.; Chen, Y.; Wang, G.; Tai, Y.-W. Dense hybrid recurrent multi-view stereo net with dynamic consistency checking. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer International Publishing: Cham, Switzerland, 2020; pp. 674–689.
21. Huang, N.; Huang, Z.; Fu, C.; Zhou, H.; Xia, Y.; Li, W.; Xiong, X.; Cai, S. A Multiview Stereo Algorithm Based on Image Segmentation Guided Generation of Planar Prior for Textureless Regions of Artificial Scenes. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 3676–3696. [CrossRef]
22. Li, Y.; Chao, X. Distance-entropy: An effective indicator for selecting informative data. *Front. Plant Sci.* **2022**, *12*, 818895. [CrossRef] [PubMed]
23. Ji, M.; Gall, J.; Zheng, H.; Liu, Y.; Fang, L. Surfacenet: An end-to-end 3d neural network for multiview stereopsis. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2307–2315.
24. Huang, P.H.; Matzen, K.; Kopf, J.; Ahuja, N.; Huang, J.B. Deepmvs: Learning multi-view stereopsis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2821–2830.
25. Chang, D.; Božič, A.; Zhang, T.; Yan, Q.; Chen, Y.; Süsstrunk, S.; Nießner, M. RC-MVSNet: Unsupervised multi-view stereo with neural rendering. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; Springer Nature: Cham, Switzerland, 2022; pp. 665–680.
26. Li, Y.; Ercisli, S. Explainable human-in-the-loop healthcare image information quality assessment and selection. *CAAI Trans. Intell. Technol.* **2023**. [CrossRef]
27. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015; pp. 234–241.
28. Kitaev, N.; Kaiser, Ł.; Levskaya, A. Reformer: The efficient transformer. *arXiv* **2020**, arXiv:2001.04451.
29. Wu, Y.; He, K. Group normalization. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
30. Yang, G.; Manela, J.; Happold, M.; Ramanan, D. Hierarchical deep stereo matching on high-resolution images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5515–5524.
31. Conroy, T.L.; Moore, J.B. Resolution invariant surfaces for panoramic vision systems. In Proceedings of the Seventh IEEE International Conference on Computer Vision, Kerkyra, Greece, 20–27 September 1999; pp. 392–397.

32. Danelljan, M.; Shahbaz Khan, F.; Felsberg, M.; Van de Weijer, J. Adaptive color attributes for real-time visual tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 18–23 June 2014; pp. 1090–1097.
33. Aanæs, H.; Jensen, R.R.; Vogiatzis, G.; Tola, E.; Dahl, A.B. Large-scale data for multiple-view stereopsis. *Int. J. Comput. Vis.* **2016**, *120*, 153–168. [CrossRef]
34. Galliani, S.; Lasinger, K.; Schindler, K. Massively parallel multiview stereopsis by surface normal diffusion. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 873–881.
35. Schönberger, J.L.; Zheng, E.; Frahm, J.M.; Pollefeys, M. Pixelwise view selection for unstructured multi-view stereo. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; pp. 501–518.
36. Chen, R.; Han, S.; Xu, J.; Su, H. Point-based multi-view stereo network. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1538–1547.
37. Gu, X.; Fan, Z.; Zhu, S.; Dai, Z.; Tan, F.; Tan, P. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 2495–2504.