forests

MDPI

Article

CGAN-Based Forest Scene 3D Reconstruction from a Single Image

Yuan Li [1,2] and Jiangming Kan [1,2,*]

[1] School of Technology, Beijing Forestry University, Beijing 100083, China; muziyuan@bjfu.edu.cn
[2] Key Laboratory of State Forestry Administration on Forestry Equipment and Automation, Beijing 100083, China
* Correspondence: kanjm@bjfu.edu.cn

Abstract: Forest scene 3D reconstruction serves as the fundamental basis for crucial applications such as forest resource inventory, forestry 3D visualization, and the perceptual capabilities of intelligent forestry robots in operational environments. However, traditional 3D reconstruction methods like LiDAR present challenges primarily because of their lack of portability. Additionally, they encounter complexities related to feature point extraction and matching within multi-view stereo vision sensors. In this research, we propose a new method that not only reconstructs the forest environment but also performs a more detailed tree reconstruction in the scene using conditional generative adversarial networks (CGANs) based on a single RGB image. Firstly, we introduced a depth estimation network based on a CGAN. This network aims to reconstruct forest scenes from images and has demonstrated remarkable performance in accurately reconstructing intricate outdoor environments. Subsequently, we designed a new tree silhouette depth map to represent the tree's shape as derived from the tree prediction network. This network aims to accomplish a detailed 3D reconstruction of individual trees masked by instance segmentation. Our approach underwent validation using the Cityscapes and Make3D outdoor datasets and exhibited exceptional performance compared with state-of-the-art methods, such as GCNDepth. It achieved a relative error as low as 8% (with an absolute error of 1.76 cm) in estimating diameter at breast height (DBH). Remarkably, our method outperforms existing approaches for single-image reconstruction. It stands as a cost-effective and user-friendly alternative to conventional forest survey methods like LiDAR and SFM techniques. The significance of our method lies in its contribution to technical support, enabling the efficient and detailed utilization of 3D forest scene reconstruction for various applications.

Keywords: forest scene reconstruction; single image; point cloud; deep learning

Citation: Li, Y.; Kan, J. CGAN-Based Forest Scene 3D Reconstruction from a Single Image. *Forests* **2024**, *15*, 194. https://doi.org/10.3390/f15010194

Academic Editor: Huaiqing Zhang

Received: 16 December 2023
Revised: 7 January 2024
Accepted: 17 January 2024
Published: 18 January 2024

1. Introduction

Forests play a crucial role in our ecosystem, providing substantial contributions to maintaining ecological balance [1]. With the rapid development of artificial intelligence technologies such as computer vision and deep learning, forest scene 3D reconstruction has emerged as a significant research topic in forest resource inventory; it can be applied to forestry management and environmental analysis and has replaced traditional methods like "per-tree measurements" in forest resource surveys. Diameter at breast height (DBH) stands out as a fundamental measurement parameter of forest scene reconstruction. The implementation of automatic and precise measurement techniques for DBH holds the potential to significantly boost the efficiency of forest scene reconstruction [2].

Many forest scene reconstruction methods take dense point clouds as inputs, which are generated by LiDAR or photogrammetry [3]. Current acquisition systems such as UAVs integrate high-resolution cameras and LiDAR and terrestrial laser scanners, which allow us to capture multiple images from multiple view angles. LiDAR systems [4] constitute a prevalent technique in the reconstruction of forest scenes, facilitating the rapid and automated modeling of the three-dimensional structure of trees [5]. Under typical conditions,

most LiDAR systems emit laser pulses that bounce off an object or surface and are subsequently detected upon their return [6]. The LiDAR sensor calculates range measurements based on the time-in-flight between pulse emission and return [7]. These measurements enable the precise positioning of scanned objects in three-dimensional space [8], generating spatially accurate three-dimensional point clouds that faithfully represent the shapes of the objects [9]. Furthermore, technological developments in miniaturization have led to the invention of special types of terrestrial laser scanners (TLSs), such as backpack laser scanning (BLS) and even handheld solid-state LiDAR, which realizes the rapid and continuous acquisition of three-dimensional LiDAR point cloud data in the forest by artificially carrying LiDAR sensors [10]. Oveland et al. employed TLSs and BLS to capture the DBH of Norwegian spruce and Scots pine. The outcomes indicated that BLS effectively captures point cloud information from individual tree trunks, enabling the precise extraction of DBH [11]. However, there are evident shortcomings for LiDAR-based tasks in forest scene reconstruction. The success rate of three-dimensional reconstruction, such as DBH accuracy, is determined by the scanning density and precision of point clouds processed using three-dimensional laser scanning devices. Furthermore, the processing speed of LiDAR-based reconstruction has difficulty meeting real-time requirements, and the high cost of three-dimensional laser scanning devices hinders widespread adoption.

Hence, image-based methods, such as Structure from Motion (SfM) photogrammetry, offer a relatively cost-effective alternative to LiDAR [12]. Given its advantages in terms of affordability and operational efficiency, SfM photogrammetry holds promise in the reconstruction of forest scenes. This technique reconstructs a model by leveraging overlapping images captured from various viewpoints around an object or scene, employing the dense matching technique known as multi-view stereo (MVS) [13]. For instance, Tan et al. generated sparse point clouds based on SfM photogrammetry and employed 2D image segmentation to distinguish between foliage and woody components, determining the visible branch segments [14]. Guo et al. introduced a refined approach for reconstructing trees with foliage [15]. They utilized depth images reconstructed from multi-view inputs to provide guidance in the reconstruction process. However, the aforementioned image-based methods, all based on multi-images, reveal several limitations in forest scene reconstruction tasks. In forest environments with similar texture features, the extraction and matching of features become exceptionally challenging with the use of multiple overlapping images. Achieving the real-time and accurate reconstruction of forest scenes presents a considerable challenge under these circumstances.

In response to the bottlenecks associated with the aforementioned methods, many methods focusing on forest scene reconstruction from single images have begun to emerge. Forest scene reconstruction from a single RGB image poses a fundamentally ill-posed problem, but the pursuit of this approach could yield significant benefits given its broad applicability [16]. Tan et al. introduced a technique for reconstructing trees from individual images [17]. Subsequently, 2D strokes are employed to direct the synthesis of a 3D tree through a growth engine. Guénard et al. suggested creating a 3D plant model through an analysis-by-synthesis approach, combining data from a single image with a priori knowledge of the plant species [18].

The most similar to our approach is forest scene reconstruction from a single image with depth estimation. Depth estimation can be widely categorized into supervised and self-supervised techniques. Within the domain of single-image supervised depth estimation, diverse methodologies have been explored, encompassing end-to-end supervised learning [19], the fusion of local predictions [20], and non-parametric scene sampling [21]. It is imperative to note that full supervision necessitates the availability of the single image concomitant with ground truth depth data for each corresponding image. Recent advancements in single-image depth estimation have introduced enforced edge consistency [22] and the integration of a depth normalization layer as a smoothness term [23], demonstrating a marked superiority over stereo pair training methodologies. Self-supervised approaches operate on presumptions regarding material properties and appearance, often enforcing

brightness constancy between frames. Pivotal contributions in this domain include Lina Liu et al.'s [24] incorporation of domain separation to address illumination variations between day and night images, as well as Michael et al.'s [25] application of wavelet decomposition for the efficient generation of depth maps. Chen et al. predicted depth maps to facilitate forest scene reconstruction through the utilization of a single image, additionally providing forecasts for DBH [26]. Nevertheless, the current methods of forest scene 3D reconstruction from a single image still suffer from critical issues such as low reconstruction accuracy in the forest scene.

Addressing the practical challenges posed by the complexity of forest scene 3D reconstruction, we present a method that not only reconstructs the forest environment but also performs a more detailed tree reconstruction in the scene based on a conditional generative adversarial network (CGAN) using a single image. Given an input image, the proposed approach initially performs instance segmentation to obtain a tree image mask. Subsequently, it utilizes a network to derive depth information from the two-dimensional image, thereby obtaining three-dimensional spatial point cloud information for the trees depicted in the image. Additionally, a refinement network is introduced to obtain point cloud information for both the frontal and occluded parts of each tree. The contributions of our paper are summarized as follows:

- We propose a new method of forest scene 3D reconstruction based on CGANs from a single image and perform a more detailed reconstruction of the trees in the scene, which differs from the reconstruction accuracy of the scene.
- We propose an outdoor scene depth estimation network based on the CGAN structure that exhibits outstanding performance in reconstructing complex outdoor scenes.
- We achieve detailed 3D reconstruction of trees within the forest scene with a tree silhouette depth map. The maximum absolute error for single-image reconstruction is reduced to 1.76 cm.

2. Methods

In this section, we delve into the architectural details of the networks used in our method of tree point cloud reconstruction. Each network plays a crucial role in the overall pipeline, and their designs are optimized for accuracy and efficiency to address an ill-posed problem. We propose a method in this paper that uses a deep neural convolution network to reconstruct tree point clouds from a single image. The framework of this method is divided into four steps, as illustrated in Figure 1. The central element of the pipeline is a tree prediction network. This network is responsible for translating the input two-dimensional image into a three-dimensional tree point cloud. To train the network, we generate a large amount of synthetic tree point cloud data using a procedural modeling approach and create a series of single-model rendered images as the training dataset.

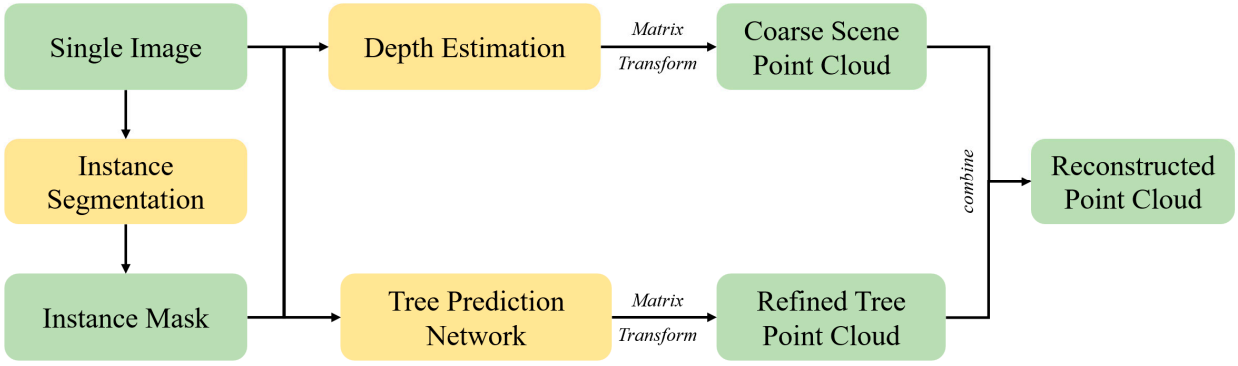

Figure 1. Pipeline of CGAN-based forest scene 3D reconstruction from a single image.

To reconstruct trees from a single image, our method first utilizes an instance segmentation model based on the Mask2Former neural network for image analysis, obtaining

tree instance image masks. This mask separates tree pixels (including leaves and branch structures) from other pixels (including the background and other objects). Next, our method uses the tree image mask obtained with the instance segmentation model along with the original image as inputs for a depth estimation network to predict point cloud information for the scene depth of the image. Additionally, we innovatively introduce a network structure based on Cycle-GAN for the front and back depth point cloud prediction of tree instances to obtain fine-grained tree point cloud models. Finally, we combine the scene depth point cloud information with segmentation mask information for fine-grained tree point clouds based on point cloud information to obtain the final reconstructed point cloud model. Our proposed method for reconstructing multiple-tree point clouds from a single image can be summarized in four main steps.

2.1. Instance Segmentation

Our instance segmentation method is responsible for isolating individual tree instances within the input image. Instance segmentation is a well-studied problem in computer imaging, and we leverage the advancements made in this area to achieve accurate tree instance segmentation. We employ a network called Mask2Former [27]. This network is a model designed for image segmentation, enabling the accurate extraction of tree masks, i.e., segmenting the input image into multiple pixel-level semantic categories. Compared with other common semantic segmentation models, Mask2Former adopts a transformer architecture to achieve image segmentation and demonstrates excellent performance on multiple image segmentation datasets.

However, training the Mask2Former network requires semantic mask information with pixel-level precision, which is often challenging to obtain. Therefore, this method renders tree models to generate synthetic datasets and constructs tree instance segmentation datasets by adjusting parameters such as lighting intensity, illumination angles, and camera position angles during image rendering. During the training of the instance segmentation network, this method applies data augmentation to the input images, including random variations in image color, brightness, and contrast. Furthermore, this method randomly applies Gaussian blur and horizontal flipping and crops the images with randomly sized bounding boxes, with a minimum image size of 512×512 pixels.

During the training process, our method utilizes the tree instance segmentation dataset to train Mask2Former until the network converges. To adapt the network to our specific task of tree instance segmentation, we finetune it on a large dataset of annotated forest images. Our custom dataset includes diverse tree species and various lighting conditions. Fine-tuning ensures that the network learns to segment trees accurately, even in challenging scenarios. Ultimately, this approach can obtain reliable mask information and separate plant instances from the images.

2.2. Depth Estimation (CDEN)

In our method, we propose a single-view depth estimation prediction network (CDEN) based on conditional generative adversarial networks (CGANs) [28] to address the transformation problem from the two-dimensional pixel image domain to the depth map domain. As illustrated in Figure 2, the generation network of this method takes the original image and instance segmentation mask information as network inputs. It processes this input through a series of modules within the generation network, including a down-sampling module based on attention mechanisms (SwinBlock), a residual module (ResBlock), and an up-sampling module (UpBlock). This network outputs a spatial depth representation. The discriminative network of this method takes the depth map as input and, after multiple convolution and activation operations, produces a probability value between 0 and 1, indicating the likelihood that each depth map input exists in the three-dimensional tree depth map domain. In the end, we convert the depth map into a 3D coarse scene point cloud using pre-calibrated intrinsic and extrinsic parameter matrices.

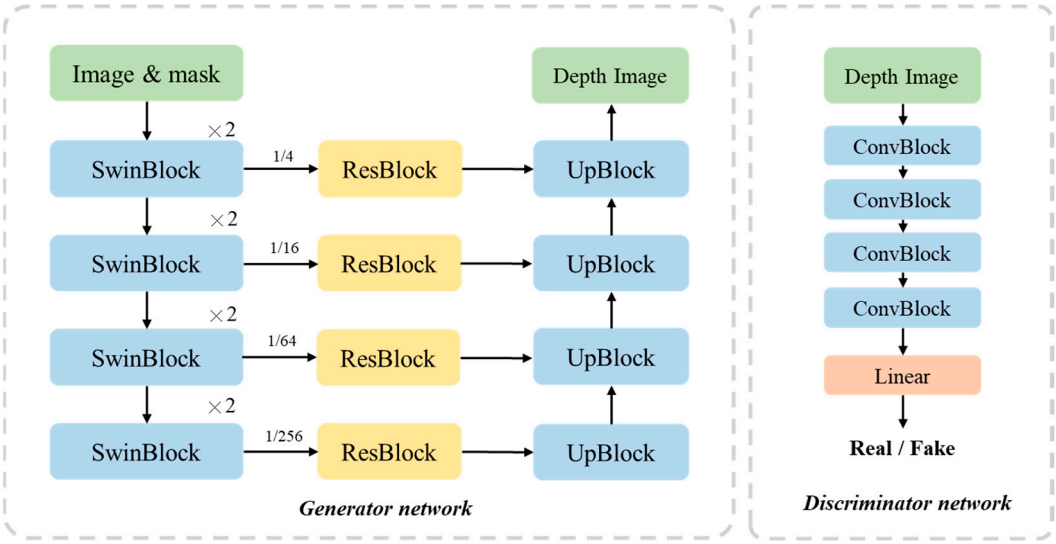

Figure 2. The architecture of the CDEN.

2.2.1. Generation Network

The generation network, as the core of the conditional generative adversarial network, translates two-dimensional pixel domain images into the depth map domain. The method employs an encoder–decoder structure to design the generation network. Specifically, Swin Transformer blocks are used as the down-sampling modules in the generation network's encoder. These modules perform down-sampling operations on the input image with a size of 512 × 512, ultimately achieving a latent tensor that contains information about the tree in the image. Additionally, residual modules are connected at down-sampling ratios of 4, 16, and 64, corresponding to the respective up-sampling tensors.

To implement the decoder, the method utilizes a series of up-sampling modules, each consisting of two sets of up-sampling layers and convolution layers. The up-sampling layers have a scale factor of 2, and all convolution layers use 1 × 1 filters with a stride of 1. Instance normalization is applied after the convolution layers, and Leaky ReLU is used as the activation function with a slope of 0.2 when the value is less than 0. The channel numbers for each layer in the decoder are 1024, 256, 64, 16, and 1, respectively. The decoder employs a Tanh activation function to compress the output into a range of −1 to 1.

2.2.2. Discriminative Network

During training, the discriminative network's objective is to distinguish between real depth maps and fake depth maps generated by the generation network. To achieve this, the method employs four two-dimensional convolution layers to construct the discriminative network. Batch normalization and Leaky ReLU activation functions are applied after the convolution layers, with a slope of 0.2 for the activation function when the value is less than 0. All convolution layers use 5 × 5 kernels with a stride of 4 and padding of 2. After the last convolution layer, a latent tensor of size 2 × 2 and 1024 dimensions is obtained. To map this latent tensor to a one-dimensional output, a fully connected layer and a Sigmoid activation function are used. The output of the discriminative network represents the probability of similarity between the output depth map and the real depth map, measuring the realism of the depth maps generated by the generation network. This process aims to transform the pixel image domain into the image depth domain.

2.2.3. Loss Function

Our method utilizes the generative adversarial network loss function to describe the competitive game between the generation network and the discriminative network. The generation network aims to generate fake data that resemble real data as closely as

possible, while the discriminative network tries to classify between real and fake samples. Specifically, the loss function can be defined as

$$L_{CGAN}(G, D) = E_{x,y}[logD(y|x)] - E_x[logD(G(x)|x)] \quad (1)$$

In this context, x represents the input image, and y corresponds to the corresponding depth map. The loss function comprises two primary components, mirroring the adversarial nature of the training process. The first component of the loss function pertains to the discriminator network, aiming to maximize its accuracy in discriminating between real data and generated data. This component seeks to enhance the discriminator's ability to distinguish between genuine and synthetic data effectively. The second component of the loss function is associated with a generator network, striving to minimize the disparity between the generated fake data and the real data. Throughout the training process, a competitive game unfolds: the generator seeks to minimize the loss function, while the discriminator endeavors to maximize it. This adversarial interaction between the generator and discriminator fosters a continuous improvement in their respective capabilities. Consequently, the gap between the generated synthetic data and real data is significantly narrowed.

In addition to the generative adversarial network (GAN) loss function, our method introduces a probability loss function to measure the pixel-level disparity between the predicted voxel probability values and the ground truth voxel probability values. The purpose of this loss function is to ensure the credibility of the generated voxel probability values. Unlike conventional classification problems, this method employs an L_1 distance regression loss to quantify the coherence among tree model voxels. Specifically,

$$L_1(G) = \sum_p \left\| d_p - \tilde{d_p} \right\|_1 \quad (2)$$

We use d_p and $\tilde{d_p}$ to represent both the real depth values and the predicted depth values, which results in a smoother generated depth map with improved continuity.

Taking into account the two aforementioned loss functions, the final loss function, L, can be expressed as follows:

$$L = L_{CGAN}(G, D) + \lambda L_1(G) \quad (3)$$

λ is a hyperparameter used to control the weight of the L_1 loss function in the overall loss, with λ set to 10. This loss function allows our method to simultaneously consider the authenticity and coherence of the generated depth map.

2.3. Tree Prediction Network

2.3.1. Tree Silhouette Depth Map

As shown in Figure 3, we define the nearest and farthest depths of the tree point cloud corresponding to the pixel at position (x, y) in a single-tree image as $z_{(1,x,y)}$ and $z_{(2,x,y)}$, respectively. We further define the frontmost depth, $R_{(x,y)}$, and the tree thickness $G_{(x,y)}$ as the tree silhouette depth map, which serves as the prediction output of the tree prediction network. The transformation formula can be expressed as follows:

$$\begin{cases} R_{(x,y)} = z_{(1,x,y)} \\ G_{(x,y)} = z_{(2,x,y)} - z_{(1,x,y)} \end{cases} \quad (4)$$

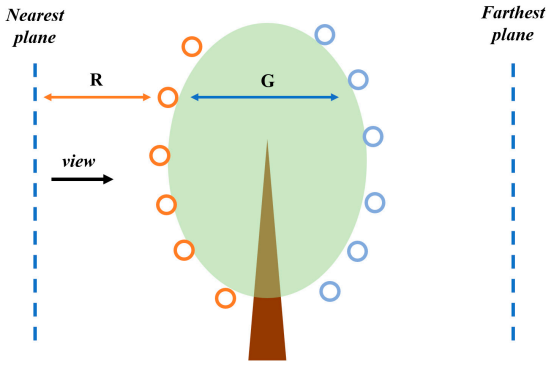

Figure 3. Tree silhouette depth map.

2.3.2. Network

In this paper, we employ a tree prediction network designed within the framework of Cycle-GAN [29] to predict a tree silhouette depth map. This ensures correspondence between RGB images and the tree silhouette depth map, enhancing the accuracy of predicting tree point clouds. The network details are illustrated in Figure 4. Cycle-GAN emphasizes cycle consistency, making the transformation process bidirectional. It converts images from domain $\boldsymbol{X}$ into domain $\boldsymbol{Y}$ and then back into domain $\boldsymbol{X}$, imposing constraints to produce results highly similar to the original images. This approach enhances the quality and stability of the transformation and exhibits robustness to variations and perturbations in input images to a certain extent. Finally, we convert the tree silhouette depth map into a 3D refined tree point cloud using pre-calibrated intrinsic and extrinsic parameter matrices.

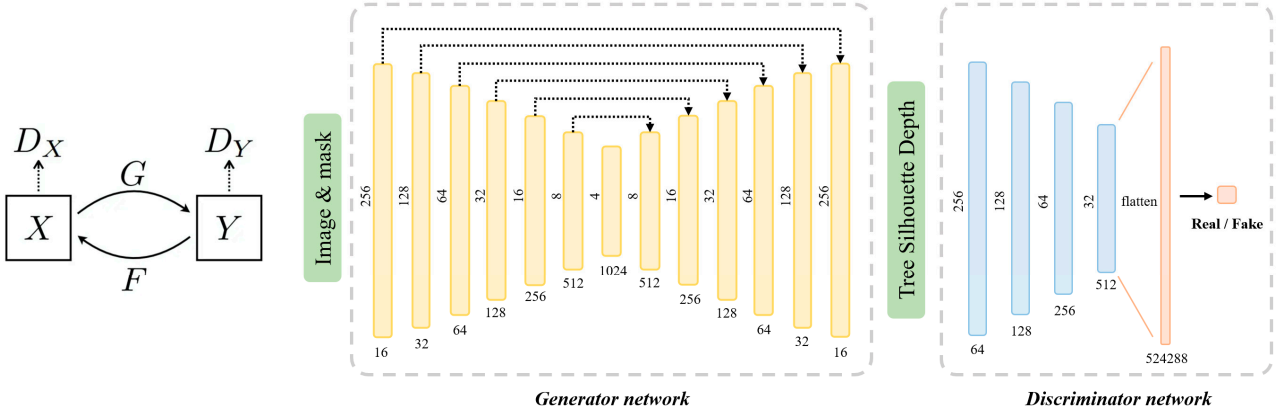

Figure 4. The framework of Cycle-GAN.

Generator Network. For our generator network, we employ a design based on the U-Net architecture [30]. We take a single image of size 512 × 512 × 4 and its corresponding mask information as inputs to the generator network. Through the utilization of a Res-net [31] down-sampling module, we perform down-sampling operations, effectively flattening the generated representation into a 1024 × 4 × 4 latent feature tensor. We consider the tensor's channel dimension the depth direction for voxel generation within the model. Following the down-sampling module, we incorporate batch normalization layers and Leaky ReLU activation layers. The channel dimensions after down-sampling are 16, 32, 64, 128, 256, 512, and 1024. In the decoder part, we employ seven 2D transpose convolution layers as the generator network's decoder, with output channel numbers of 512, 256, 128, 64, 32, 16, and 2. The size of the transpose convolution kernel is 4 × 4, with a stride of 2 and padding of 1. After the first six transpose convolution layers, batch normalization layers and Leaky ReLU activation layers are applied. The final layer is equipped with a Tanh activation function, resulting in a 512 × 512 × 2 tree silhouette depth map, which contains information about the depth of the tree point cloud in both the front and rear aspects.

Discriminator Network. We employ a series of 2D convolution layers. Following each 2D convolution layer in the discriminator network, batch normalization and Leaky ReLU are applied as the activation function, with a slope of 0.2 for activations less than 0. All convolution layers utilize 4×4 convolution kernels, a stride of 2, and padding of 1. The output channel numbers are 256, 128, 64, and 32, and the last convolution layer yields a $512 \times 32 \times 32$ tensor. To map this latent tensor to a one-dimensional output, our method employs a single fully connected layer and a Sigmoid activation function. The output of the discriminator network is utilized to measure the authenticity of the tree silhouette depth map generated by the generator network, facilitating the transformation between the image domain and the depth domain.

2.3.3. Loss Function

In this paper, we employ Cycle-GAN loss to measure the pixel-level disparity between the output tree silhouette depth map and the real image, aiming to enhance the quality and stability of the transformation. This is specifically expressed as

$$\begin{cases} Loss_{GAN} = E_y[logD_Y(y)] + E_x[log(1 - D_Y(G(x)))] \\ \qquad\qquad +E_x[logD_X(x)] + E_y[log(1 - D_X(F(y)))] \\ Loss_{cycle} = E_x[\|F(G(x)) - x\|_1] + E_y[\|G(F(y)) - y\|_1] \\ Loss_{identity} = E_x[\|F(x) - x\|_1] + E_y[\|G(y) - y\|_1] \end{cases} \tag{5}$$

In this context, where x represents the input image, y represents the tree silhouette depth map, and D and G represent the generator networks for the input image and tree silhouette depth map, we employ the same generator network structure in this study. The $Loss_{GAN}$ ensures the co-evolution of the generator network and discriminator network, thereby enabling the generator network to produce more realistic images. The $Loss_{cycle}$ ensures that the generator's output images differ in style but not in content from the input images. The role of $Loss_{identity}$ is primarily to preserve the hue. Therefore, the final loss function is expressed as follows:

$$Loss = Loss_{GAN} + Loss_{cycle} + Loss_{identity} \tag{6}$$

3. Experiment and Results

3.1. Implementation

3.1.1. Dataset

For the instance segmentation task, the majority of our training data consisted of manually annotated forest scene datasets. Additionally, we utilized rendered tree models to generate synthetic datasets, adjusting image rendering parameters such as lighting intensity, illumination angles, and camera positions to construct the tree instance segmentation dataset. During the training of the instance segmentation network, our approach employed data augmentation techniques on input images, including random variations in color, brightness, and contrast. Furthermore, we applied random Gaussian blur and horizontal flips and performed random-sized bounding box cropping on images, ensuring a minimum image size of 512×512 pixels.

For the training of the depth estimation network, we aimed to achieve robust performance on outdoor datasets. Therefore, we leveraged publicly available outdoor datasets such as Cityscapes [32] and Make3D [33]. Additionally, we collected RGB-D images from two distinct forest sampling sites in Beijing and Shandong. RGB images were used as inputs, while depth maps served as outputs. These sampling sites featured both pure forests and mixed forests, with varying vegetation growth conditions and tree trunk diameter distributions, ensuring dataset diversity.

For the training of the tree prediction network, we generated synthetic datasets using rendered tree models, adjusting image rendering parameters such as lighting intensity, illumination angles, and camera positions to construct the tree silhouette depth map

dataset. During the training of the instance segmentation network, data augmentation operations were applied to the input images. Moreover, we utilized densely captured 3D point cloud data, transforming it into two-dimensional tree silhouette depth maps through computer graphic coordinate transformations. Additionally, we rendered corresponding two-dimensional color images as inputs. Furthermore, we performed data augmentation on input images, including random variations in color, brightness, and contrast.

3.1.2. Training Details

The experiments in this paper were conducted using the Pytorch deep learning framework, and the training and inference were performed on the NVIDIA RTX 3090 GPU. The training of the instance segmentation model, depth estimation network, and tree prediction network utilized randomly selected training, testing, and validation sets in an 8:2:1 ratio, with image dimensions set at 512 × 512 pixels.

In the case of depth estimation, the generator network employed the Adam optimizer with a learning rate of 0.001, and the hyperparameters β_1 and β_2 were set to 0.9 and 0.999, respectively. The discriminator network used the SGD optimizer with a learning rate of 0.001, and the momentum hyperparameter was set to 0.99. A batch size of eight was employed, and after 1000 epochs of training on the dataset, the learning rate was reduced to 1×10^{-5}. The model converged after 2000 epochs, and the training process took a total of 30 h. The tree prediction network used the Adam optimizer with a learning rate of 0.001, and the hyperparameters β_1 and β_2 were set to 0.9 and 0.999, respectively. After 2000 epochs, the model converged, and the training process took a total of 20 h.

3.2. Evaluation Metrics

3.2.1. Depth Estimation Metrics

In order to assess the performance of single-image depth estimation networks, we employ standard evaluation metrics. Among these, absolute error (AbsRel) measures the absolute difference between predicted depth values and actual depth values, providing an overall accuracy assessment of depth estimation. Root mean square error (RMSE) calculates the square of the average difference between predicted and actual depth values and then takes the square root, offering a measure of the overall error in depth estimation. Square relative error (SqRel) first computes the relative error between predicted and true depth values and then squares these errors, emphasizing the impact of larger errors. Root mean square logarithmic error (RMSE-Log) first computes the square of the average of the logarithmic differences between the predicted and actual depth values and then takes the square root. This metric is commonly used to account for errors in smaller depth values in depth estimation.

As shown in Table 1, d and d^* represent the predicted and ground truth depth values, while D represents the set comprising all the predicted depth values for an image. These metrics are common choices for evaluating the performance of depth estimation algorithms because they provide crucial insights into the accuracy and quality of depth predictions. While AbsRel and SqRel focus on absolute and relative errors, RMSE and RMSE-Log offer more comprehensive error information.

Table 1. Depth estimation metrics.

Metric	Formula
AbsRel	$\frac{1}{\lvert D\rvert}\sum_{d\in D}\lvert d^* - d\rvert / d^*$
RMSE	$\sqrt{\frac{1}{\lvert D\rvert}\sum_{d\in D}\lVert d^* - d\rVert^2}$
SqRel	$\frac{1}{\lvert D\rvert}\sum_{d\in D}\lVert d^* - d\rVert^2 / d^*$
RMSE-Log	$\sqrt{\frac{1}{\lvert D\rvert}\sum_{d\in D}\lVert log d^* - log d\rVert^2}$

3.2.2. Forest Scene Reconstruction Metric

Constrained by hardware devices such as LiDAR, this study does not provide a direct assessment of the accuracy of the three-dimensional reconstruction results for forest scenes. However, the reliability of the three-dimensional reconstruction can be indirectly demonstrated by estimating the diameter at breast height (DBH) of standing trees within the three-dimensional scene. In this study, it is stipulated that the DBH of standing trees is measured using a DBH tape at a height of 1.3 m above the ground. The measured diameter, divided by π, is considered the true DBH value.

3.3. Reconstruction Results

The results of our forest scene reconstruction based on a single image are illustrated in Figure 5. The left side presents the reconstructed input image, while the right side displays the point cloud reconstruction effects. It is evident that the point cloud reconstruction exhibits good robustness, particularly in situations with varying intensities of image illumination. This section provides a detailed exposition of the experimental process, followed by comparisons and ablation studies after defining evaluation metrics.

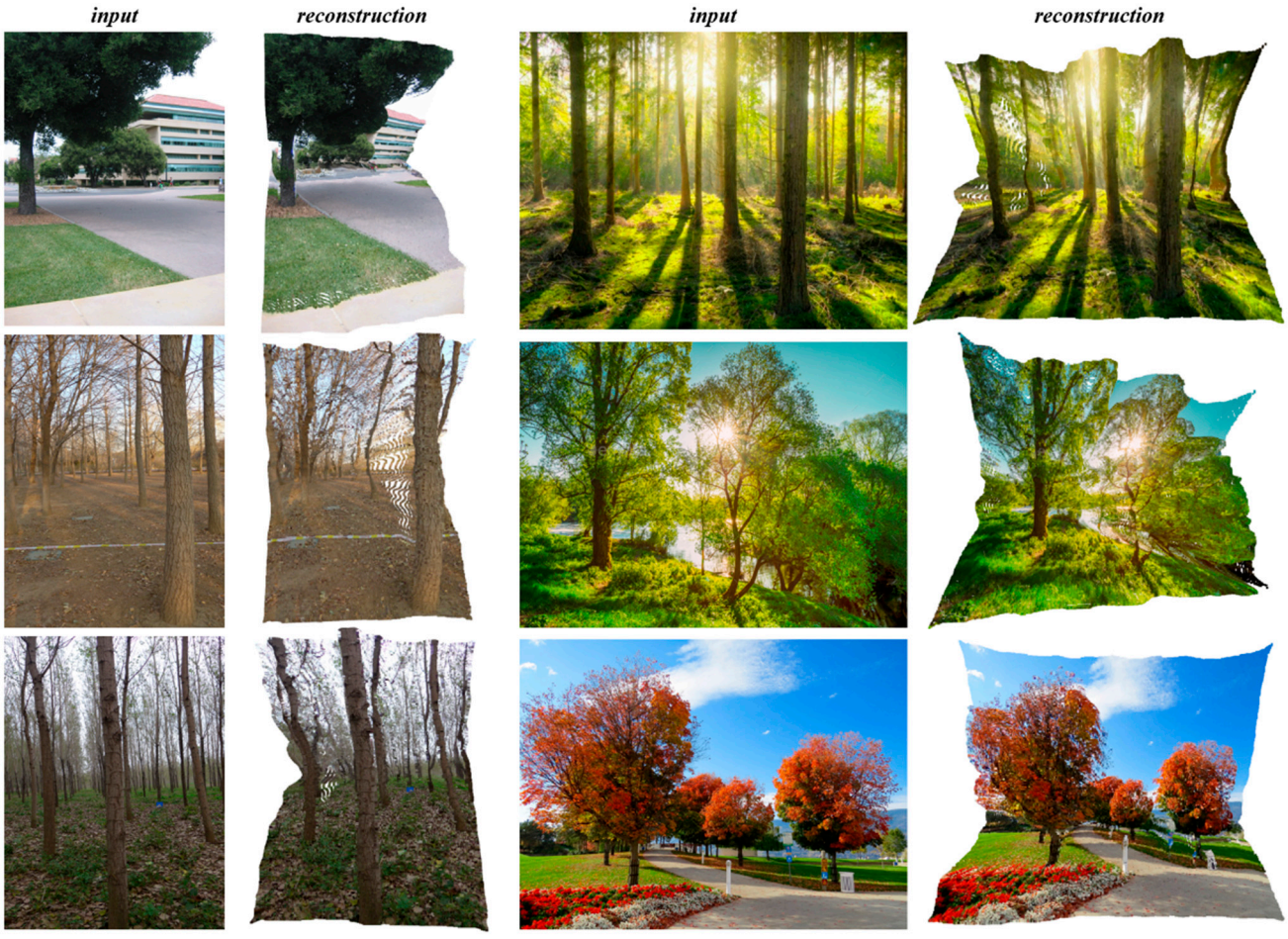

Figure 5. The results of our forest scene reconstruction method.

3.4. Comparison Study

3.4.1. Depth Estimation Comparison

The Make3D dataset comprises outdoor scene images captured using a custom 3D scanner. Because of early hardware limitations, the resolution of the images in this dataset is 2272 $\times$ 1704, while the corresponding depth maps have a resolution of 55 $\times$ 305. The depth information in this dataset is unreliable at long distances, necessitating the use of a

mask to filter out pixels with depth values greater than 70 m. The final training set was augmented to approximately 12,000 image pairs through offline data augmentation.

To train neural network models on the Make3D dataset, the predicted depth maps were first up-sampled to 512 × 512. The output results were then resized to the original image size. As shown in Figure 6, from left to right, we have the input single image, from Chen et al. [26], and our method. Our network's predictions enable a clearer distinction between trees in outdoor scenes compared with existing methods, and it exhibits better robustness in predicting under varying lighting conditions. As presented in Table 2, we conducted quantitative comparisons between our prediction method and currently effective methods. We resized the network output depth maps to match the size of the Make3D depth map ground truth using bilinear interpolation. From the table of evaluation metrics, it is evident that our approach outperforms other competitive methods in both global and local depth estimation.

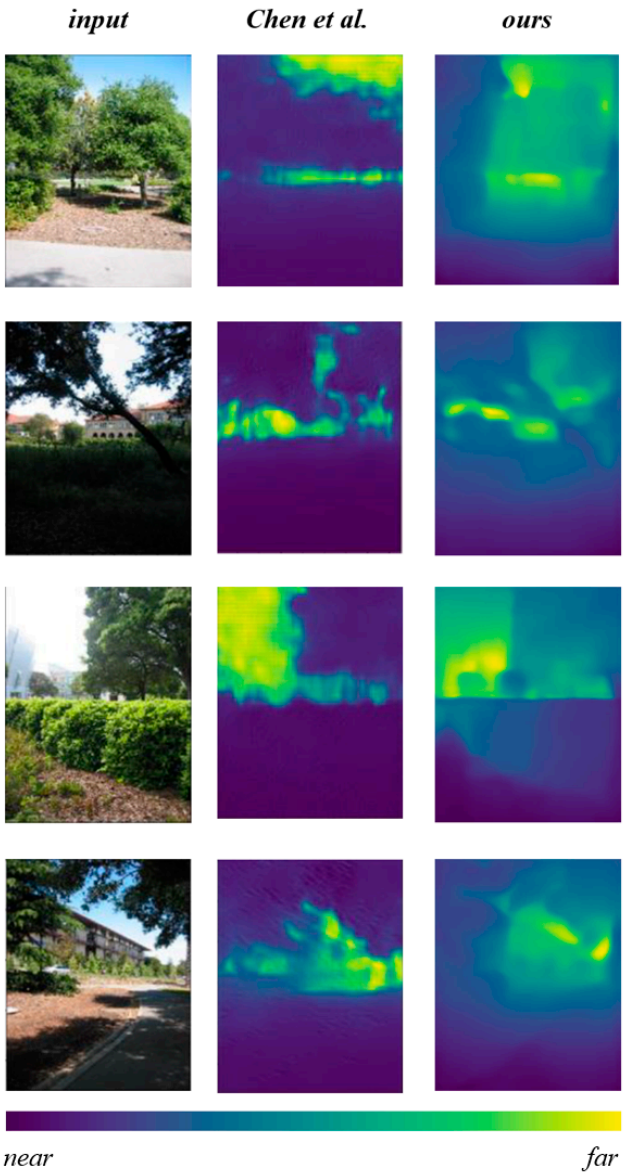

Figure 6. Depth estimation examples based on the Make3D dataset [26].

Table 2. Depth estimation examples based on the Make3D dataset.

Method	AbsRel	RMSE	SqRel	log10
GCNDepth [34]	0.424	6.757	3.075	0.107
SharinGAN [35]	0.377	8.388	4.901	0.225
Monodepth2 [36]	0.322	7.417	3.589	0.201
Chen et al. [26]	0.257	6.74	4.129	0.083
Ours	0.246	6.152	3.015	0.056

Furthermore, we conducted quantitative experiments using the Cityscapes dataset, comparing our depth estimation method with state-of-the-art depth estimation methods, as shown in Table 3. It is clear that, relative to other recent methods, our proposed depth estimation method demonstrates distinct advantages and exhibits strong generalization capabilities for measuring depth in outdoor environments.

Table 3. Depth estimation examples based on the Cityscapes dataset.

Method	AbsRel	RMSE	SqRel	log10
Laina et al. [37]	0.257	7.273	4.238	0.448
Xu et al. [38]	0.246	7.117	4.06	0.428
Zhang et al. [39]	0.234	7.104	3.776	0.416
Ours	0.221	6.918	3.451	0.403

3.4.2. Forest Scene Reconstruction Comparison

To validate the reconstruction effectiveness of our forest scene three-dimensional reconstruction method, we collected several color images from eight suburban parks, encompassing images of trees with different species and varying levels of leaf sparsity. Subsequently, we employed our single-view-based forest scene three-dimensional reconstruction method to reconstruct these images. The relevant results are presented in Figure 7, where, from left to right, we have the color input image, the depth prediction and reconstruction results using the method by Chen et al. [40], and the predicted depth map and three-dimensional reconstruction results obtained using our method. Our method is capable of performing three-dimensional reconstruction from single-view images without relying on depth information or additional perspective images. Moreover, it yields favorable reconstruction results in various complex environments while preserving the integrity of the input image information. The time taken for our method to perform three-dimensional reconstruction for each image is approximately 2.98 s.

Note that current forest scene reconstructions lack actual data as reference points. Tree diameter at breast height (DBH) is a directly measurable parameter closely linked to the actual shape and size of trees. Hence, if the reconstructed scene aligns well with the real environment, the information obtained from tree DBH measurements should correspond or closely match the features of the reconstructed trees. By comparing this data, the accuracy and reliability of the reconstruction process can be indirectly validated. We selected two forest sampling sites for different tree species located in Beijing and Shandong as our experimental locations. These sites encompass both pure forests and mixed forests, with varying vegetation growth conditions and tree diameter distributions, ensuring the diversity of our study locations. In each group of images, a single standing tree image was marked with a red rope at a height of 1.3 m above the ground. Approximately 5 to 6 standing tree images in each image set underwent multiple diameter measurements, captured from three different angles. As shown in Table 4, a total of 36 standing tree images from these two experimental locations were randomly tested in this experiment, yielding an average reconstruction error of 1.25 cm and a relative error of 7.49% in the best reconstruction results. We compared our method with state-of-the-art tree reconstruction methods based on computer vision, where that of Chen et al. [40] closely resembles our method based on a single image. Gao et al. [41] rely on multi-view SFM (Structure from

Motion). This indirectly validates the reliability of the three-dimensional reconstruction approach presented in this paper.

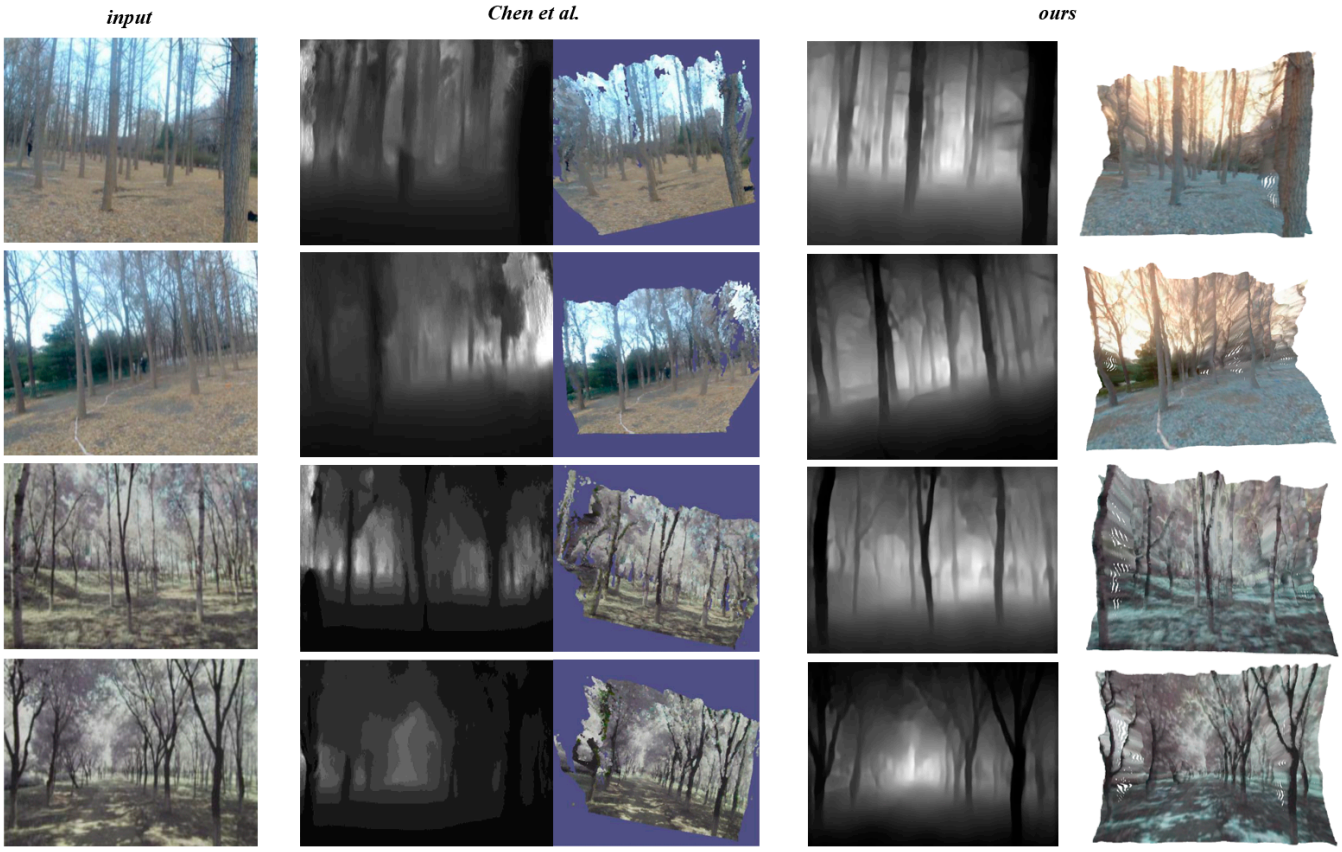

Figure 7. Depth estimation examples based on the Make3D dataset [40].

Table 4. The results of DBH reconstruction for standing trees.

Location	Mean DBH	Method	Min MAE (cm)	Max MAE (cm)	Mean MAE (cm)	Relative MAE
Beijing	20.1 cm	Chen et al. [40]	0.61	4.89	2.65	13.18%
		Gao et al. [41]	0.36	4.95	1.84	9.15%
		Ours	0.42	4.05	1.76	8.75%
Shandong	16.7 cm	Chen et al. [40]	0.49	3.75	2.36	14.12%
		Gao et al. [41]	0.29	3.64	1.48	8.86%
		Ours	0.33	3.56	1.25	7.49%

3.5. Ablation Study

We evaluate the enhancement effect of the tree prediction network in our pipeline on the single-image tree three-dimensional reconstruction task through ablation experiments. In this experiment, the quantitative evaluation metric used is the Mean Absolute Error (MAE) between the estimated DBH values after point cloud reconstruction and the actual measured ground truth values. The MAE loss is employed to assess the accuracy of the model's predicted DBH values compared with the true DBH values. The experimental data for ground truth are derived from the forest sampling site in Beijing, as previously mentioned, and mathematical model parameters for estimating DBH are obtained using random sample consensus (RANSAC). As shown in Table 5, in the experiment, the tree prediction network predicts the tree silhouette depth map, which contains depth information for both the front and back sides of tree trunks. This results in more accurate final point cloud data reconstruction.

Table 5. Comparative network ablation experiments.

Network	Mean MAE (cm)	Relative MAE
Depth Estimation	2.17	10.72%
Depth Estimation + Tree Prediction Network	1.76	8.75%

4. Conclusions

Our method presents a breakthrough in addressing the core objectives outlined in this study. The primary aim was to devise a portable and detailed 3D reconstruction technique for forest scenes, overcoming the limitations of traditional methods like LiDAR and addressing challenges related to feature extraction within multi-view stereo vision sensors. Our method successfully achieves these objectives by not only reconstructing entire forest environments but also providing intricate reconstructions of individual trees from a single RGB image. The introduction of a CGAN-based depth estimation network demonstrates exceptional performance, accurately capturing the complexity of outdoor environments. Moreover, the innovative tree silhouette depth map derived from our tree prediction network enables detailed 3D reconstructions of trees, effectively utilizing instance segmentation. Our method showcases remarkable performance with a relative error as low as 8% (absolute error: 1.76 cm) in estimating DBH. Our method not only fulfills but exceeds the set objectives, offering a transformative solution for detailed 3D forest scene reconstruction. Its implications extend beyond mere technological advancements, emphasizing its pivotal role in enabling efficient and comprehensive forest ecosystem analysis and management.

In future work, we intend to continue optimizing this method. We plan to extract richer information from two-dimensional images, such as tree branch orientation and crown color information, to further enhance the representation capabilities of single-image-based three-dimensional point cloud reconstruction. Additionally, we will simulate additional tree species information, expand the training dataset, and consider the reconstruction of more trees to better meet the requirements of practical applications.

Author Contributions: Conceptualization, J.K. and Y.L.; software, Y.L. and J.K.; investigation, Y.L. and J.K.; writing—original draft, Y.L. and J.K.; preparation, Y.L. and J.K.; writing—review and editing, Y.L. and J.K.; supervision, J.K.; project administration, J.K.; funding acquisition, J.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (No. 32071680).

Data Availability Statement: The data will be made available by the authors on request.

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

References

1. Dugesar, V.; Satish, K.V.; Pandey, M.K.; Srivastava, P.K.; Petropoulos, G.P.; Anand, A.; Behera, M.D. Impact of Environmental Gradients on Phenometrics of Major Forest Types of Kumaon Region of the Western Himalaya. *Forests* **2022**, *13*, 1973. [CrossRef]
2. Gollob, C.; Ritter, T.; Nothdurft, A. Forest inventory with long range and high-speed personal laser scanning (PLS) and simultaneous localization and mapping (SLAM) technology. *Remote Sens.* **2020**, *12*, 1509. [CrossRef]
3. Cárdenas-Donoso, J.L.; Ogayar, C.J.; Feito, F.R.; Jurado, J.M. Modeling of the 3D tree skeleton using real-world data: A survey. *IEEE Trans. Vis. Comput. Graph.* **2022**, *29*, 4920–4935. [CrossRef] [PubMed]
4. Hernandez-Santin, L.; Rudge, M.L.; Bartolo, R.E.; Erskine, P.D. Identifying species and monitoring understorey from UAS-derived data: A literature review and future directions. *Drones* **2019**, *3*, 9. [CrossRef]
5. Raumonen, P.; Kaasalainen, M.; Åkerblom, M.; Kaasalainen, S.; Kaartinen, H.; Vastaranta, M.; Holopainen, M.; Disney, M.; Lewis, P. Fast automatic precision tree models from terrestrial laser scanner data. *Remote Sens.* **2013**, *5*, 491–520. [CrossRef]
6. Tickle, P.K.; Lee, A.; Lucas, R.M.; Austin, J.; Witte, C. Quantifying Australian forest floristics and structure using small footprint LiDAR and large scale aerial photography. *For. Ecol. Manag.* **2006**, *223*, 379–394. [CrossRef]
7. Wallace, L.; Lucieer, A.; Malenovský, Z.; Turner, D.; Vopěnka, P. Assessment of forest structure using two UAV techniques: A comparison of airborne laser scanning and structure from motion (SfM) point clouds. *Forests* **2016**, *7*, 62. [CrossRef]

8. Davies, A.B.; Asner, G.P. Advances in animal ecology from 3D-LiDAR ecosystem mapping. *Trends Ecol. Evol.* **2014**, *29*, 681–691. [PubMed]
9. Guerra-Hernández, J.; Cosenza, D.N.; Rodriguez, L.C.E.; Silva, M.; Tomé, M.; Díaz-Varela, R.A.; González-Ferreiro, E. Comparison of ALS-and UAV (SfM)-derived high-density point clouds for individual tree detection in Eucalyptus plantations. *Int. J. Remote Sens.* **2018**, *39*, 5211–5235. [CrossRef]
10. Morgenroth, J.; Gómez, C. Assessment of tree structure using a 3D image analysis technique—A proof of concept. *Urban For. Urban Green.* **2014**, *13*, 198–203.
11. Oveland, I.; Hauglin, M.; Gobakken, T.; Næsset, E.; Maalen-Johansen, I. Automatic estimation of tree position and stem diameter using a moving terrestrial laser scanner. *Remote Sens.* **2017**, *9*, 350.
12. Karel, W.; Piermattei, L.; Wieser, M.; Wang, D.; Hollaus, M.; Pfeifer, N.; Surový, P.; Koreň, M.; Tomaštík, J.; Mokroš, M. Terrestrial photogrammetry for forest 3D modelling at the plot level. In Proceedings of the EGU General Assembly, Vienna, Austria, 8–13 April 2018; p. 12749.
13. Iglhaut, J.; Cabo, C.; Puliti, S.; Piermattei, L.; O'Connor, J.; Rosette, J. Structure from motion photogrammetry in forestry: A review. *Curr. For. Rep.* **2019**, *5*, 155–168. [CrossRef]
14. Tan, P.; Zeng, G.; Wang, J.; Kang, S.B.; Quan, L. Image-based tree modeling. In Proceedings of the ACM SIGGRAPH 2007 Papers, San Diego, CA, USA, 5–9 August 2007; p. 87-es.
15. Guo, J.; Xu, S.; Yan, D.M.; Cheng, Z.; Jaeger, M.; Zhang, X. Realistic procedural plant modeling from multiple view images. *IEEE Trans. Vis. Comput. Graph.* **2018**, *26*, 1372–1384.16. [PubMed]
16. Okura, F. 3D modeling and reconstruction of plants and trees: A cross-cutting review across computer graphics, vision, and plant phenotyping. *Breed. Sci.* **2022**, *72*, 31–47. [CrossRef] [PubMed]
17. Tan, P.; Fang, T.; Xiao, J.; Zhao, P.; Quan, L. Single image tree modeling. *ACM Trans. Graph. (TOG)* **2008**, *27*, 1–7. [CrossRef]
18. Guénard, J.; Morin, G.; Boudon, F.; Charvillat, V. Reconstructing plants in 3D from a single image using analysis-by-synthesis. In *Advances in Visual Computing, Proceedings of the 9th International Symposium, ISVC 2013, Rethymnon, Crete, Greece, 29–31 July 2013*; Springer: Berlin/Heidelberg, Germany, 2013; Proceedings, Part I 9; pp. 322–332.
19. Eigen, D.; Puhrsch, C.; Fergus, R. Depth map prediction from a single image using a multi-scale deep network. *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 2366–2374.
20. Hoiem, D.; Efros, A.A.; Hebert, M. Automatic photo pop-up. In Proceedings of the ACM SIGGRAPH 2005 Papers, Los Angeles, CA, USA, 31 July 4 August 2005; pp. 577–584.
21. Karsch, K.; Liu, C.; Kang, S.B. Depth extraction from video using non-parametric sampling. In *Computer Vision–ECCV 2012, Proceedings of the 12th European Conference on Computer Vision, Florence, Italy, 7–13 October 2012*; Springer: Berlin/Heidelberg, Germany, 2012; Proceedings, Part V 12; pp. 775–788.
22. Yang, Z.; Wang, P.; Wang, Y.; Xu, W.; Nevatia, R. Lego: Learning edge with geometry all at once by watching videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 225–234.
23. Godard, C.; Mac Aodha, O.; Brostow, G. Unsupervised monocular depth estimation with left-right consistency. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 270–279.
24. Liu, L.; Song, X.; Wang, M.; Liu, Y.; Zhang, L. Self-supervised monocular depth estimation for all day images using domain separation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 12737–12746.
25. Ramamonjisoa, M.; Firman, M.; Watson, J.; Lepetit, V.; Turmukhambetov, D. Single image depth prediction with wavelet decomposition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 11089–11098.
26. Chen, S.; Tang, M.; Dong, R.; Kan, J. Encoder–Decoder Structure Fusing Depth Information for Outdoor Semantic Segmentation. *Appl. Sci.* **2023**, *13*, 9924. [CrossRef]
27. Cheng, B.; Misra, I.; Schwing, A.G.; Kirillov, A.; Girdhar, R. Masked-attention mask transformer for universal image segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 July 2022; pp. 1290–1299.
28. Isola, P.; Zhu, J.Y.; Zhou, T.; Efros, A.A. Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1125–1134.
29. Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2223–2232.
30. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015, Proceedings of the 18th International Conference, Munich, Germany, 5–9 October 2015*; Springer International Publishing: Berlin/Heidelberg, Germany, 2015; Proceedings, Part III 18; pp. 234–241.
31. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
32. Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; Schiele, B. The cityscapes dataset for semantic urban scene understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 3213–3223.

33. Saxena, A.; Sun, M.; Ng, A.Y. Make3d: Learning 3d scene structure from a single still image. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *31*, 824–840. [CrossRef] [PubMed]
34. Masoumian, A.; Rashwan, H.A.; Abdulwahab, S.; Cristiano, J.; Asif, M.S.; Puig, D. Gcndepth: Self-supervised monocular depth estimation based on graph convolutional network. *Neurocomputing* **2023**, *517*, 81–92. [CrossRef]
35. Pnvr, K.; Zhou, H.; Jacobs, D. Sharingan: Combining synthetic and real data for unsupervised geometry estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 13974–13983.
36. Godard, C.; Mac Aodha, O.; Firman, M.; Brostow, G.J. Digging into Self-Supervised Monocular Depth Estimation. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 3827–3837.
37. Laina, I.; Rupprecht, C.; Belagiannis, V.; Tombari, F.; Navab, N. Deeper depth prediction with fully convolutional residual networks. In Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV), Stanford, CA, USA, 25–28 October 2016; pp. 239–248.
38. Xu, D.; Ouyang, W.; Wang, X.; Sebe, N. Pad-net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 675–684.
39. Zhang, Z.; Cui, Z.; Xu, C.; Jie, Z.; Li, X.; Yang, J. Joint task-recursive learning for semantic segmentation and depth estimation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 235–251.
40. Chen, S. Monocular Image Depth Estimation and Application in 3D Reconstruction of Forest Scene. Ph.D. Thesis, Beijing Forestry University, Beijing, China, 2021.
41. Gao, Q.; Kan, J. Automatic forest DBH measurement based on structure from motion photogrammetr. *Remote Sens.* **2022**, *14*, 2064. [CrossRef]