Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

MSA-MVSNet: A Cross-Scale Collaborative Attention-Based Multi-View Reconstruction Network for Orchard Tree 3D Reconstruction with Instance Segmentation for Fruit Counting

Hui Li

3023763876@qq.com

Huzhou University

Jun Zhang

ZheJiang Academy of Agricultural Sciences

Jianhua Hong

Huzhou University

Fuzhi Ke

Zhejiang Citrus Research Institute

Chu Zhang

Huzhou University

Zhixin Li

ZheJiang Academy of Agricultural Sciences

Research Article

Keywords: Multi-View Stereo, 3D Reconstruction, Semantic Segmentation, Fruit Counting, Attention Mechanism, Smart Agriculture

Posted Date: March 5th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8759462/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

MSA-MVSNet: A Cross-Scale Collaborative Attention-Based Multi-View Reconstruction Network for Orchard Tree 3D Reconstruction with Instance Segmentation for Fruit Counting

Hui Li*[1] , Jun Zhang[2] , Jianhua Hong[1] , Fuzhi Ke[3] , Chu Zhang[1], Zhixin Li[2]

Abstract:
To address the issues of detail loss and matching difficulties in fruit tree 3D reconstruction caused by complex branch–leaf morphology, fruit occlusion, and illumination variations, this paper proposes an end-to-end cross-scale collaborative attention multi-view stereo network, termed MSA-MVSNet, for high-quality 3D reconstruction of orchard trees, while integrating semantic segmentation for fruit counting. A multi-scale feature enhancement module is designed to adaptively fuse deep semantic features and shallow fine-grained details through a spatial–channel collaborative attention mechanism, thereby enhancing the network's capability to represent multi-scale structures such as trunks, branches, and leaves. Multi-branch dilated convolutions are introduced to enlarge the receptive field, and deformable convolutions are incorporated to adaptively capture the irregular geometric shapes of fruits, improving modeling robustness. In addition, a feature matching transformer is introduced to strengthen long-range global contextual correlations within and across images via intra-attention and inter-attention mechanisms, thereby improving matching stability in low-texture and repetitive-texture regions.To validate the effectiveness of the proposed method, experiments are conducted on self-collected real orchard dataset and public benchmark datasets. The results demonstrate that MSA-MVSNet outperforms baseline models by 8.2% in terms of 3D reconstruction quality. Finally, by combining depth filtering with the semantic segmentation results of YOLOv11-Seg, a semantic-guided fruit reconstruction and counting framework is constructed. This framework achieves an overall counting F1-score of 92.8% on the self-collected dataset with varying scene sparsity and 93.5% on the public Fuji-sfm dataset, demonstrating its effectiveness and generalization capability.

Keywords: Multi-View Stereo, 3D Reconstruction, Semantic Segmentation, Fruit Counting, Attention Mechanism, Smart Agriculture

Introduction

With the rapid transition of traditional orchard management toward smart agriculture, there is a growing demand for general-purpose research systems capable of target perception, intelligent reasoning, and autonomous operation in complex outdoor environments. Fruit trees are characterized by complex spatial structures, and accurate acquisition of their structural parameters forms the foundation for orchard maintenance, scientific planning, and effective management, providing essential data support and scientific evidence for precision agricultural monitoring (Kok et al., 2023).

In orchard resource monitoring, traditional field surveys require manual entry into orchards to directly measure and count fruit size, leaf diameter, trunk circumference, and fruit quantity, which is time-consuming and labor-intensive. (Jin et al., 2021)。In contrast, non-contact sensing technologies such as depth cameras

Hui Li
3023763876@qq.com
Jun Zhang
hunterzju@163.com
Jianhua Hong
2715222819@qq.com
Fuzhi Ke
kefz@qq.com
Zhixin Li
zxli0429@163.com
Chu Zhang
chuzh@zjhu.edu.cn
1 School of Information Engineering, Huzhou University, Huzhou 313000,China
2 Food Science Research Institute, Zhejiang Academy of Agricultural Sciences, Hangzhou 310021,China
3 Zhejiang Citrus Research Institute, Zhejiang Academy of Agricultural Sciences, Taizhou 318026,China

and LiDAR exhibit significant advantages in plant phenotyping acquisition and analysis.(Chunyan et al., 2015)。Currently, high-precision 3D sensors and multi-view image-based 3D reconstruction techniques have been applied to fruit tree modeling. Laser scanners and depth cameras are representative high-precision depth sensing devices. Although laser scanners can acquire highly accurate 3D data, their high cost and sensitivity to environmental conditions limit their applicability mainly to indoor environments or fruit trees with relatively simple structures, making them unsuitable for complex outdoor orchards. Depth cameras, on the other hand, typically rely on traditional Iterative Closest Point (ICP) algorithms(Besl & McKay, 1992), which exhibit limitations when modeling objects with complex geometries.

In recent years, advances in computer vision and image processing have provided new technical pathways for fruit tree 3D reconstruction.(D. Yang et al., 2024) constructed a multi-camera image acquisition system to achieve monocular vision-based 3D reconstruction. however, the fixed camera orientations constrain the spatial coverage and make the system susceptible to installation errors. To address such issues , (Remondino et al., 2014) proposed a multi-view geometry-based approach that estimates 3D point positions using Structure from Motion (Schonberger & Frahm, 2016) and multi-view stereo matching, enabling relatively accurate reconstruction of fruit tree structures. Compared with multi-camera monocular systems, this approach offers greater flexibility and generality. Nevertheless, it requires substantial time and manual effort in feature extraction, matching, camera calibration, and 3D point computation, and is mainly applicable to scenes with strong feature continuity, resulting in limited generalization capability. To overcome the limitations of traditional SfM and MVS methods in complex environments—namely insufficient accuracy and low computational efficiency—deep learning-based stereo reconstruction networks represented by MVSNet (Yao et al., 2018) have been proposed. These methods adopt an end-to-end design that integrates feature extraction, cost volume construction, cost aggregation, and depth regression into a unified network, enabling high-accuracy depth estimation and dense reconstruction from multi-view images. While improving accuracy, generalization, and automation, MVSNet-based approaches significantly reduce manual intervention. In orchard environments, such methods facilitate unmanned intelligent operations by directly recovering 3D point clouds from given inputs.

Although deep learning-based multi-view stereo reconstruction networks have achieved notable progress across various domains, directly applying them to fruit tree 3D reconstruction remains challenging. Fruit trees exhibit complex geometric structures with diverse branch–leaf morphologies, dense fruit distributions, and varying illumination conditions and occlusions. Existing algorithms often lack effective cross-scale information fusion mechanisms during feature extraction, making it difficult to simultaneously preserve local details and global structures. As a result, large-scale tree reconstructions frequently suffer from missing fine details such as branches and leaves, while excessive emphasis on small-scale features may cause trunk discontinuities, ultimately degrading reconstruction accuracy and parameter reliability. Moreover, uneven illumination and occlusion lead to texture degradation and weakened cross-view consistency, increasing matching errors and resulting in unstable depth estimation, noisy point clouds, and local structural defects,

thereby reducing reconstruction completeness and accuracy.(Zeng et al., 2025)。In addition, abundant non-target background elements increase the risk of false predictions. Therefore, designing a deep reconstruction model with strong multi-scale perception and feature attention tailored to orchard tree structures and outdoor lighting conditions is a key challenge for improving fruit tree 3D reconstruction quality(Wan et al., 2025).

To address the above challenges, this study proposes an end-to-end network for orchard fruit tree 3D reconstruction, termed MSA-MVSNet , and integrates YOLOv11-Seg (Khanam & Hussain, 2024) to construct a fruit counting analysis framework. MSA-MVSNet is built upon Cas-MVSNet (Gu et al., 2020) , and incorporates an enhanced geometry-aware feature pyramid network, referred to as the Enhanced Feature and Geometry-aware Feature Pyramid Network (EFG-FPN). In the encoder stage, multi-branch dilated convolution structures(F. Yu & Koltun, 2016) are introduced to extract multi-scale features, enlarge the receptive field, and enhance contextual awareness. In the decoder stage, deformable convolutional networks (R. Wang et al., 2021) are integrated to adaptively adjust sampling locations, enabling effective capture of irregular structures such as branches and leaves. At the feature skip connections, a Multi-Scale Feature Enhancement (MSFE) module is designed, which employs a Spatial–Channel Synergistic Attention mechanism(Si et al., 2025) to adaptively reweight features and achieve effective multi-scale feature fusion. To further strengthen long-range global information association and matching across images, a Feature Matching Transformer(Ding et al., 2021) is introduced after feature extraction to enhance deep feature interaction. The main contributions of this study are summarized as follows:

1. An end-to-end network, termed MSA-MVSNet, is proposed for high-quality 3D reconstruction of orchard fruit trees.
2. To address the limitations of Cas-MVSNet in simultaneously modeling multi-scale structures such as tree trunks and canopies, a Multi-Scale Feature Enhancement (MSFE) module is designed, combined with dilated convolutions and deformable convolutions to enhance multi-scale feature extraction and fusion.
3. A Feature Matching Transformer (FMT) is introduced to achieve robust long-range global context aggregation both within individual images and across multiple views.
4. Extensive experiments are conducted on a self-collected real orchard fruit tree dataset and public benchmark datasets, demonstrating the effectiveness of MSA-MVSNet for orchard 3D reconstruction tasks.
5. MSA-MVSNet is integrated with the YOLOv11-Seg semantic segmentation algorithm to construct a framework for fruit counting and spatial analysis in orchard environments.

2.Materials and methodology

2.1Dataset Details

This study employs a handheld smartphone to capture multi-view images of fruit trees and constructs an orchard dataset comprising multiple sets of multi-view image sequences. Data acquisition was conducted from October to November during the fruit ripening period. To ensure scene realism and environmental diversity, image collection was performed between 08:00 and 17:00, thereby reducing the impact of natural illumination variations in orchard environments. The distance between the camera and the fruit trees was controlled within

a range of 50–100 cm. To comprehensively cover the three-dimensional structure of each fruit tree, images were captured at multiple height levels, as illustrated in Fig. 1(a), with the angular difference between adjacent views maintained within 30°.As shown in Fig. 1(b), the Structure-from-Motion (SfM)-based COLMAP pipeline(Schonberger & Frahm, 2016) was used for camera pose estimation and image undistortion, and the data were subsequently converted into the MVSNet input format(Yao et al., 2018) using COLMAP2MVSNet. Images of each individual fruit tree were organized as an independent scene. Since acquiring high-confidence fruit tree reconstruction datasets with ground-truth geometry requires high-precision and costly sensors, and such datasets are currently not publicly available, this study follows recent stereo reconstruction practices by leveraging widely used public MVS datasets, including DTU(Jensen et al., 2014), Tanks and Temples(Knapitsch et al., 2017), and BlendedMVS(Yao et al., 2020), for network training, evaluation, and fine-tuning. Specifically, the model is trained and tested on the DTU dataset, its generalization capability is validated on the intermediate scenes of the Tanks and Temples dataset, and BlendedMVS is used for fine-tuning to better adapt the model to orchard fruit tree reconstruction.

To support subsequent fruit recognition and counting tasks, two fruit segmentation datasets were constructed, each containing 1,000 images of citrus and apple fruits, respectively. To enhance the robustness and generalization capability of the segmentation model, data augmentation strategies such as contrast adjustment and brightness variation were applied, expanding each fruit category to 1,500 images for training and validation of the semantic segmentation model. Furthermore, to comprehensively evaluate the performance of the proposed fruit counting framework that integrates 3D reconstruction and semantic segmentation, a large-scale public Fuji-sfm dataset(Gené-Mola et al., 2020) was also introduced for counting performance assessment.

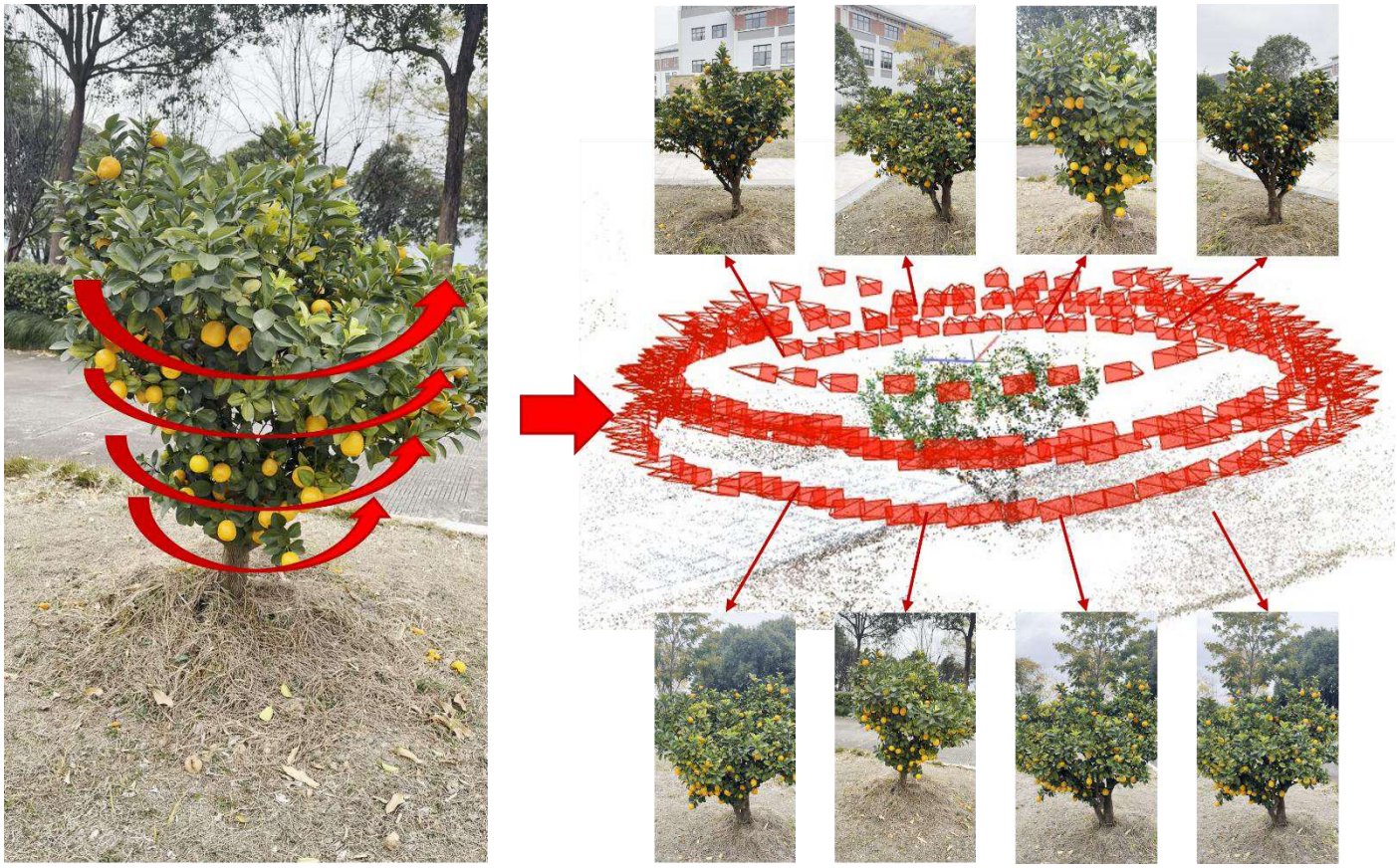

Fig. 1 (a)Simplified Image Acquisition Diagram;(b) Camera Pose Estimation Diagram

2.2 MSA-MVSNet architecture

Multi-View Stereo (MVS) estimates depth maps aligned with the reference view by leveraging a reference image, multiple source images, and their corresponding intrinsic and extrinsic camera parameters. The predicted depth maps are subsequently filtered and fused to generate a dense reconstructed point cloud. The overall reconstruction pipeline is illustrated in Fig. 2, and the main steps are summarized as follows:1.acquisition of multi-view images and camera parameters, which are divided into a reference image and source images;2.extraction of multi-scale feature maps from each view using a Feature Pyramid Network (FPN), followed by cost volume construction through geometric transformation and similarity measurement;3.regularization of the cost volume using a 3D convolutional neural network (3DCNN) to produce a probability volume that represents the confidence of depth hypotheses;4.depth regression to obtain depth maps for each view, which are then fused to generate a high-density point cloud model.

The proposed MSA-MVSNet multi-view stereo reconstruction model follows a framework similar to mainstream MVS methods and adopts a cascaded architecture to estimate depth maps in a coarse-to-fine manner across multiple stages. At each stage, the depth prediction from the previous stage is used to guide the construction and regularization of the subsequent cost volume, enabling consistent multi-view matching and progressive depth refinement. In the feature extraction stage, MSA-MVSNet employs an Enhanced Feature and Geometry-aware FPN (EFG-FPN) to extract multi-scale features at three resolution levels. The encoder consists of three Multi-Scale Residual Convolution Blocks (MSRCB), which use parallel dilated convolutions

to capture multi-scale information and residual connections to enhance feature propagation and gradient stability. The decoder is composed of three Feature Adaptive Modeling (FAM) blocks, which utilize deformable convolutions to achieve adaptive feature alignment and geometric structure modeling, thereby facilitating smooth and seamless propagation of local features. At the skip connections of the decoder, feature maps from different scales are processed by the Multi-Scale Feature Enhancement (MSFE) module. The MSFE module fuses multi-scale convolutional features and incorporates Spatial–Channel Synergistic Attention (SCSA) to strengthen effective multi-scale feature integration and enhance responses in key regions. Prior to cost volume construction, a Feature Matching Transformer (FMT) is applied to the low-resolution features output by the FAM to aggregate global contextual information. Specifically, intra-attention is used to model dependencies within each single view, while inter-attention aggregates global information between the reference image and source images, thereby improving multi-view feature matching accuracy and the robustness of depth estimation. Finally, MSA-MVSNet constructs the cost volume using differentiable homography, regularizes it to obtain a probability volume of depth hypotheses, and generates continuous depth maps via a soft-argmin regression strategy, achieving high-precision depth estimation.

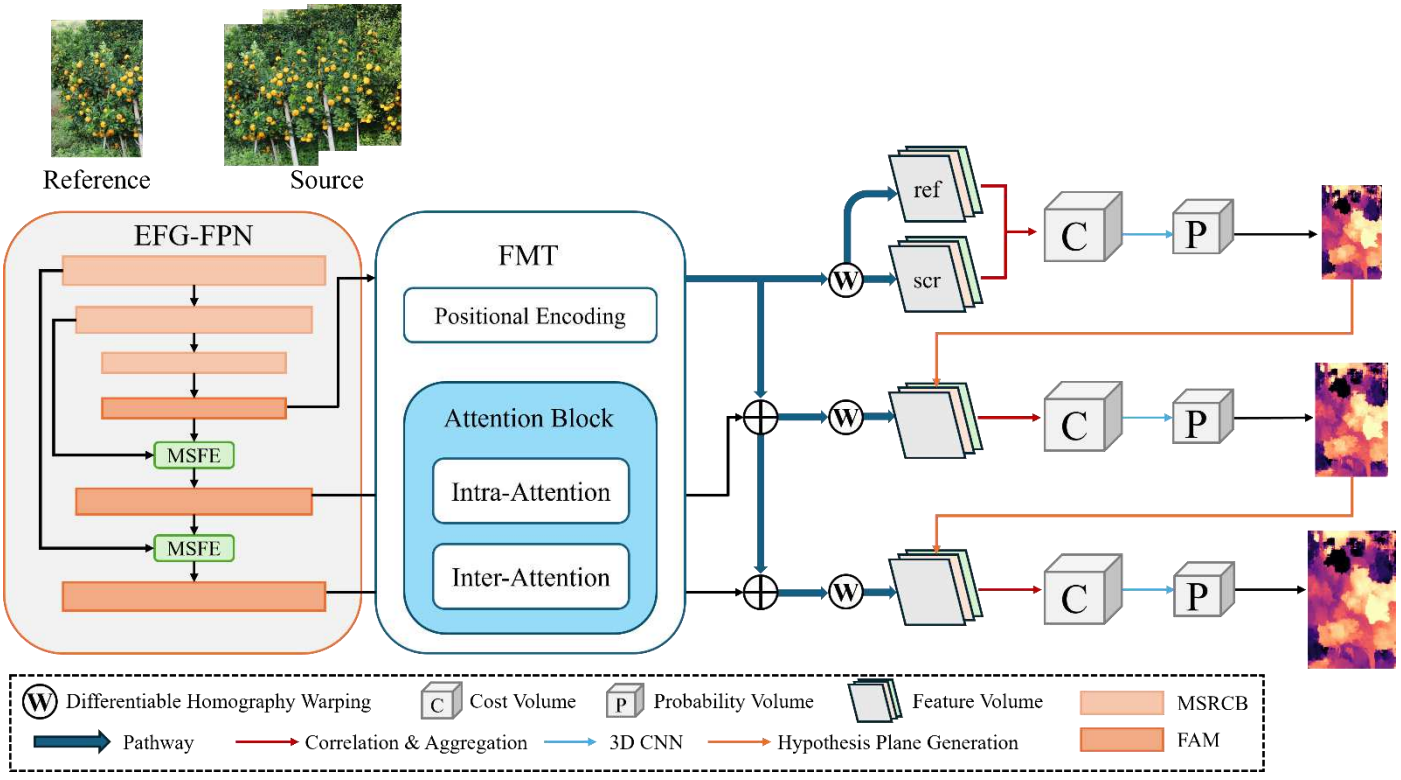

Fig. 2 Overall structure of MSA-MVSNet

2.2.1 Enhanced Feature and Geometry-aware Feature Pyramid Networks

Fruit tree canopies exhibit highly complex and dense branch–leaf structures, and orchard environments are often affected by illumination variations, occlusions, and other disturbances. Although conventional feature pyramid networks can extract multi-level features at different scales, fixed-size convolution kernels have limited receptive field flexibility and struggle to simultaneously capture fine-grained details and high-level semantic information. This limitation often leads to detail loss and blurred boundaries in regions with

complex branch–leaf structures. Moreover, simple feature concatenation operations are insufficient for effective cross-scale feature fusion, which constrains the model's ability to represent multi-scale geometric structures. To address these challenges, this study proposes an Enhanced Feature and Geometry-aware Feature Pyramid Network (EFG-FPN),which strengthens multi-scale feature representation and enables efficient fusion of shallow detailed information with deep semantic information.

In the encoder, multiple Multi-Scale Residual Convolution Blocks (MSRCB) are employed to replace standard convolutional structures, achieving comprehensive feature extraction and stable gradient propagation through multi-layer convolutional operations and residual connections. Each MSRCB consists of a standard convolution branch and dilated convolution branches with different dilation rates $r \in \{2,3\}$. Given an input image $X \in \mathbb{R}^{H\times W\times 3}$,the MSRCB operation can be expressed as:

$$F_0^{(i)} = \mathrm{Conv2D}_{k_i\times k_i}(X), F_0^{(i)} \in \mathbb{R}^{H\times W\times C}, i = 0,1,2,\ k_i \in \{3,5,5\} \tag{1}$$

Where $F_0^{(i)} \in \mathbb{R}^{H\times W\times C}$ denotes the output of the i-th standard convolution layer, and k_i represents the kernel size of the corresponding layer. Subsequently, each feature map is processed by two dilated convolution branches with different dilation rates $r \in \{2,3\}$:

$$F_1 = \mathrm{Conv}_{3\times 3}\left(Atrous\mathrm{Conv2D}_{3\times 3}^{r=2}\left(F_0^{(i)}\right)\right), F_2 = \mathrm{Conv}_{3\times 3}\left(Atrous\mathrm{Conv2D}_{3\times 3}^{r=3}\left(F_0^{(i)}\right)\right) \tag{2}$$

where $\mathrm{AtrousConv2D}_{3\times 3}^{r}(\cdot)$ denotes a dilated convolution. (x, y) denotes the pixel coordinates of the input feature map, and $W(p,q)$represents the convolution kernel weights. The operation can be expressed as:

$$\mathrm{AtrousConv2D}_{3\times 3}^{r}(F)(x,y) = \sum_i \sum_j W(p,q)\cdot F(x + r\cdot p, y + r\cdot q) \tag{3}$$

By introducing sampling intervals, dilated convolutions effectively enlarge the receptive field without increasing kernel size or reducing feature map resolution, enabling multi-scale feature perception. Unlike pooling operations, dilated convolutions preserve pixel-level spatial resolution while expanding the receptive field, maintaining fine-grained spatial information with lower computational cost and no additional parameter overhead. This design enhances the model's adaptability to spatial scale variations while preserving detailed canopy structures, providing high-precision geometric feature inputs for the subsequent decoder. Within each MSRCB, features extracted from different branches are concatenated along the channel dimension and fused through a 3×3convolution, followed by residual addition with the input feature map. This structure captures both local details and global contextual information, achieving efficient feature integration and enhancement, as shown in:

$$\begin{aligned} F_{\mathrm{cat}} &= [F_1, F_2], \in \mathbb{R}^{H\times W\times 2C} \\ F_{\mathrm{out}} &= F_0^{(i)} + \mathrm{Conv}_{3\times 3}(F_{\mathrm{cat}}) \end{aligned} \tag{4}$$

In the decoder, the EFG-FPN architecture enables adaptive fusion and progressive upsampling of multi-

scale features. Specifically, at two skip-connection stages, deep features undergo channel compression and are spatially aligned with shallow features via bilinear interpolation, followed by weighted integration through the Multi-Scale Feature Enhancement (MSFE) module to achieve adaptive cross-layer feature fusion. The fused features are then processed by the Feature Adaptive Modeling Block (FAM). The FAM module consists of a 3×3standard convolution and multi-level DCNv2 deformable convolutions. The standard convolution generates offset fields and modulation masks, where offsets control the positional displacement of convolution sampling points and masks assign adaptive contribution weights to each sampling location. During convolution, kernels sample features at non-uniform positions determined by the offsets and aggregate them using mask-based weighting, effectively enhancing salient features while suppressing background noise. The output is further processed by batch normalization and nonlinear activation to obtain structurally enhanced feature representations. This mechanism allows convolution kernels to dynamically adjust their receptive fields according to the geometric characteristics of the input features, adapting to the complex spatial structures and local deformations of branches and leaves while preserving spatial resolution and fine-grained details. The offset and mask parameters are optimized end-to-end during network training without introducing additional manually designed parameters, thereby improving robustness and geometric deformation adaptability. By strengthening the geometric representation of local features, the FAM module enables a smooth transition from convolution-based local structures to Transformer-based modeling(Han et al., 2021), providing high-quality inputs for subsequent Feature Matching Transformer (FMT) modules. Through multi-scale and multi-view feature alignment and spatially adaptive adjustment, the features output by FAM exhibit improved semantic consistency and spatial alignment, establishing a stable foundation for global context modeling in FMT and achieving a natural integration of local features and global semantics.

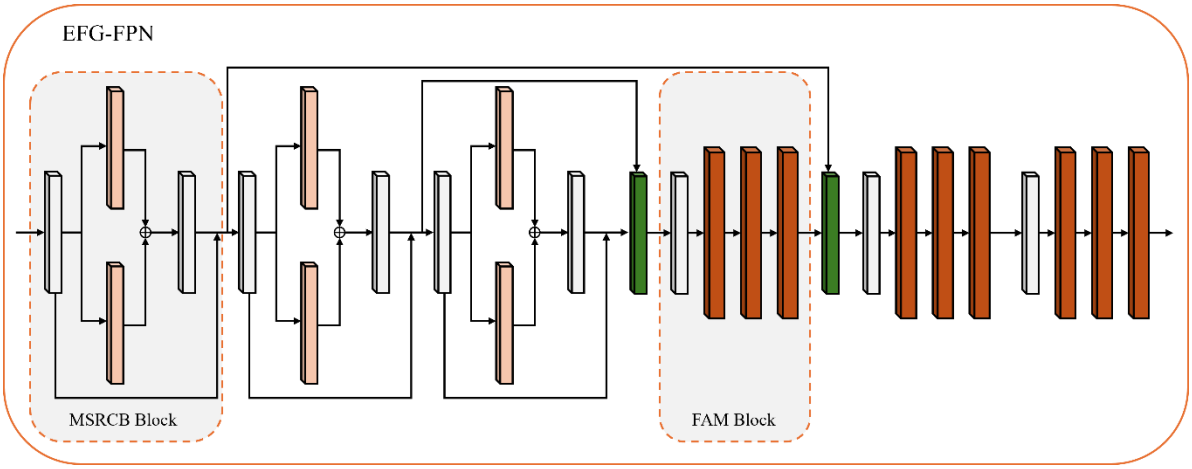

Fig. 3 Enhanced Feature and Geometry-aware FPN

2.2.2Multi-Scale Feature Enhancement module

Traditional feature pyramid decoders commonly rely on simple feature concatenation for feature fusion. However, this strategy tends to introduce noise and redundant information from shallow features into the fused representation, while the high-level semantic information contained in deep features is not sufficiently

emphasized. To overcome the limitations of conventional feature fusion methods in multi-scale feature representation, this study designs a Multi-Scale Feature Enhancement (MSFE) module. This section provides an in-depth analysis of the MSFE module from the perspectives of architectural design and the incorporation of attention mechanisms.

The MSFE module is deployed at two skip-connection stages of the EFG-FPN and consists of one multi-scale convolution branch and two attention branches, enabling effective fusion of shallow and deep features. Before feeding features into the MSFE module, deep features are first compressed along the channel dimension and spatially aligned with shallow features using bilinear interpolation. The MSFE module takes two input feature streams: a shallow feature map $F_s \in \mathbb{R}^{C\times H\times W}$, which contains rich fine-grained detail information, and a deep feature map $F_d \in \mathbb{R}^{C\times H\times W}$, which encodes high-level semantic information.In the multi-scale convolution branch, the two feature maps are concatenated along the channel dimension and passed through a 1×1convolution to reduce the channel dimensionality. This operation can be expressed as:

$$F_{\text{cat}} = \text{Concat}_{\text{channel}}(F_s, F_d) \in \mathbb{R}^{H_s\times W_{s\times}(C_s+C_d)} \tag{5}$$

$$F_{\text{proj}} = P_{1\times 1}(F_{\text{cat}}) \in \mathbb{R}^{H_s\times W_{s\times}C_p} \tag{6}$$

where $P_{1\times 1}(\cdot)$denotes the 1×1convolutional projection. The compressed feature map is then evenly divided into four parallel branches, with each branch containing one quarter of the total channels, which can be expressed as:

$$\text{Split}(F_{proj}) = \left[F_c^{(1)}, F_c^{(2)}, F_c^{(3)}, F_c^{(4)}\right], F_c^{(i)} \in \mathbb{R}^{H_s\times W_s\times Cp/4}, i = 1,2,3,4 \tag{7}$$

The first three branches employ depthwise separable convolutions (Nascimento et al., 2019) with different kernel sizes (3×3, 5×5, and 7×7) to extract multi-scale features. Correspondingly, the padding sizes are set to 1, 2, and 3 to maintain consistent spatial dimensions of the feature maps. The fourth branch preserves the original features to realize an identity mapping:

$$F_{\text{ds}}^{(i)} = \text{DSConv}_{k_i\times k_i}\left(F_c^{(i)}\right), k_i \in \{3,5,7\}, i = 1,2,3,$$
$$F_{\text{ds}}^{(4)} = F_c^{(4)} \tag{8}$$

At the end of the multi-scale convolution branch, the outputs of the four branches are concatenated along the channel dimension and fused across channels via a 1×1 convolution. A Sigmoid activation function is then applied to generate the corresponding attention weight map:

$$F_{\text{ms}} = \sigma\left(P_{1\times 1}\left(\left[F_{\text{ds}}^{(1)}, F_{\text{ds}}^{(2)}, F_{\text{ds}}^{(3)}, F_{\text{ds}}^{(4)}\right]\right)\right) \in \mathbb{R}^{C_p\times H_s\times W_s} \tag{9}$$

where $\sigma(\cdot)$denotes the Sigmoid activation function.

In the attention branches, to further enhance the modeling capability of spatial and channel information during feature fusion, a Spatial–Channel Collaborative Attention (SCSA) mechanism is introduced. The shallow and deep feature maps, $F_s$and F_d, are separately fed into the SCSA module for feature reweighting.

Within the SCSA module, the input feature map is divided into four groups along the channel dimension. Each group is processed using 1D depthwise separable convolutions with different kernel sizes (3, 5, 7and 9) to capture multi-scale spatial contextual information. This design achieves multi-scale spatial modeling while maintaining computational efficiency. After group normalization and nonlinear activation, the convolution outputs generate spatial attention weight maps along the horizontal and vertical directions, which can be expressed as:

$$F_{\text{spa}} = F \odot A_h \odot A_w, \qquad A_h, A_w \in \mathbb{R}^{C\times H\times 1}, \mathbb{R}^{C\times 1\times W} \tag{10}$$

where $\odot$ denotes the element-wise multiplication operation, and $A_h$and $A_w$represent the spatial attention weight maps generated along the horizontal and vertical directions, respectively, which are used to reweight the feature map Fin the spatial domain. After obtaining the spatial attention weights, the SCSA module further constructs a channel attention mechanism to capture semantic dependencies across different channels. Specifically, the module employs 1×1group convolutions to generate the query (Q), key (K), and value (V) features, and applies a multi-head self-attention mechanism to compute channel-wise correlations. The channel affinity matrix is obtained via dot-product operations between Q and K, followed by normalization and scaling. This matrix is then used to weight and aggregate V, producing a channel-refined feature representation. Finally, global average pooling across the channel dimension and a Sigmoid activation are applied to generate channel attention weights, enabling adaptive enhancement of channel-wise features:

$$A_c = \sigma\left(\text{AvgPool}\big(\text{MHSA}(Q,K,V)\big)\right) \in \mathbb{R}^{C\times 1\times 1}$$
$$F_{\text{sca}} = F_{\text{spa}} \odot A_c \tag{11}$$

where $\text{AvgPool}(\cdot)$denotes the channel-wise average pooling operation, $\text{MHSA}(\cdot)$represents the multi-head self-attention mechanism, and A_c is the channel attention weight map. The collaborative use of spatial and channel attention enables joint optimization of spatial structures and semantic information. With low computational overhead, the SCSA module is capable of capturing both local geometric details and global contextual relationships, thereby providing more stable and discriminative feature representations for subsequent multi-scale feature fusion.

During the feature fusion stage, the MSFE module performs element-wise multiplication between the attention weights generated by the multi-scale convolution branch and the shallow and deep features reweighted by the SCSA attention branches, respectively. These enhanced features are then combined with the original input features through residual connections to achieve feature enhancement and information compensation. The fused features are further refined by a 3×3convolution to enforce spatial consistency. Ultimately, the MSFE module outputs a fused feature map that contains rich fine-grained details and strong semantic representations. While preserving shallow-level detail information, it effectively strengthens deep semantic expressions, thereby improving the model's capability to represent target features and enhancing structural reconstruction accuracy in complex fruit tree canopy regions, as shown in Eq. (12).

$$F_{fuse} = F_{ms} \odot F_s^{sca} + F_{ms} \odot F_d^{sca} + F_s + F_d$$
$$F_{out} = ReLU(BN(Conv_{3\times3}(F_{fuse}))) \quad (12)$$

where F_{ms} denotes the attention weight map generated by the multi-scale convolution branch, F_s^{sca} and F_d^{sca} represent the shallow and deep features reweighted by the SCSA attention mechanism, respectively, and F_s and F_d correspond to the original input shallow and deep feature maps.

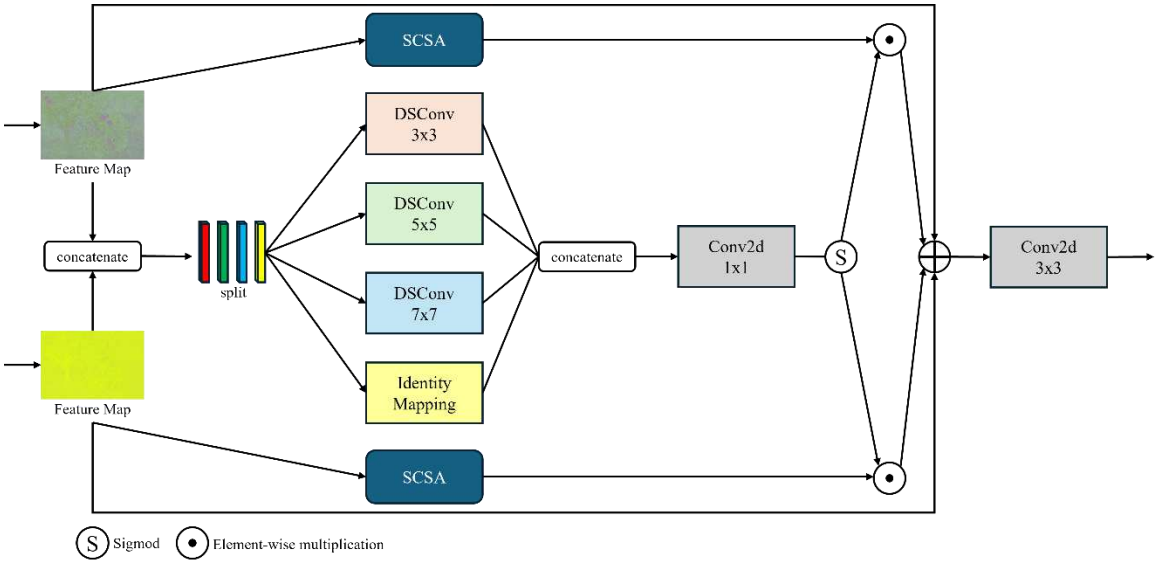

Fig. 4 Multi-Scale Feature Enhancement module

2.2.3 Feature Matching Transformer

In multi-view stereo (MVS) reconstruction, the computation of matching costs is typically based directly on extracted features to construct the cost volume, without explicitly considering potential inter-image correspondences or global contextual information. To enhance the robustness and cross-view consistency of feature matching in complex fruit tree scenes, this study introduces the Feature Matching Transformer (FMT) module, which explicitly models global contextual dependencies in the MVS task.

The core idea of FMT is to reformulate the multi-view matching problem as a “one-to-many” feature matching task. Through intra-attention and inter-attention mechanisms, the module captures long-range dependencies both within a single image and across multiple images. Convolutional networks, limited by their local receptive fields, often fail to capture global semantic information across regions, leading to ambiguous matches in low-texture, repetitive-texture, or non-Lambertian areas. To address this, the FMT module leverages the Transformer’s attention mechanism to establish global pixel-level dependencies at the feature level. Its core structure consists of an intra-attention layer and an inter-attention layer:1.Intra-attention models global semantic dependencies within a single view. By adaptively weighting feature interactions, it aggregates contextual information and enhances feature consistency.2.Inter-attention establishes cross-view feature interactions between the reference view and source views. It extracts complementary geometric information from projections across viewpoints, enforcing spatial constraints and global alignment among multiple views. Specifically, FMT first flattens the feature maps of the reference image and its source images, adding

positional encodings to preserve spatial location information. A multi-head linear attention mechanism is applied to compute feature correlation weights. Compared with standard Transformers, which have quadratic complexity, the linear attention design significantly reduces both computation and memory costs, making it feasible to perform global matching on high-resolution feature maps.

During the matching stage, FMT updates only the source view features while keeping the reference view features fixed, ensuring consistency of the matching target and maintaining the "one-to-many" feature correspondence characteristic of multi-view stereo. By introducing FMT, the network can learn feature correlations over a wider contextual range, substantially improving matching performance in occluded, low-texture, and complex geometric regions. This provides more discriminative global feature representations for subsequent cost volume construction and depth estimation.

2.3 Plant Fruit Counting via Semantic Segmentation and Multi-View Stereo Fusion

To achieve high-precision recognition and spatial counting of fruit in orchards, this study proposes a fruit counting method that integrates semantic segmentation with multi-view stereo (MVS) reconstruction. The method takes multi-view image sequences as input and leverages the depth estimation, segmentation masks, and camera pose information from each view to achieve semantically consistent 3D reconstruction and precise counting of fruit in 3D space.Specifically, the YOLOv11-Seg semantic segmentation model is employed to perform pixel-level segmentation of fruit instances in each view, separating fruit foreground from the background. YOLOv11-Seg maintains the fast inference advantage of the YOLO series while enhancing semantic segmentation accuracy through an improved feature pyramid structure and a refined segmentation head. For each input image, the model outputs a set of detected fruit instances, including bounding box coordinates, class confidence scores, and pixel-level segmentation masks, providing semantic guidance for subsequent depth fusion.Depth maps for each view are predicted using MSA-MVSNet, providing per-pixel depth information in the camera coordinate system. During depth fusion, semantic filtering is applied, the binary masks generated by semantic segmentation are used to retain only the depth values corresponding to fruit foreground pixels, while background pixels are treated as invalid and filtered out. This step removes background noise that could interfere with reconstruction. Using the camera parameters, the foreground depth points from all views are projected into a unified world coordinate system, achieving multi-view, semantically consistent 3D point cloud fusion.

Fruit counting is performed during the depth fusion stage using an incremental tracking mechanism based on distance thresholds. For each detected fruit instance, the 3D points within the corresponding segmented region are used to calculate the centroid, representing the spatial center of that fruit, which can be expressed as:

$$\overline{P} = \frac{1}{n}\sum_{i=1}^{n} P_i \tag{13}$$

Here, n denotes the number of 3D points belonging to a fruit, and P_i represents the coordinates of each point. The 3D fruit centroids are matched to historical fruit centers in the database using the Euclidean distance

$$d = \| \bar{P}_{\text{current}} - \bar{P}_{\text{historical}} \|_2.$$

Specifically, for the first viewpoint, the computed fruit centroids are assigned unique IDs and stored in the fruit database. For subsequent viewpoints, each current fruit centroid is compared against all centroids in the database using the Euclidean distance. If the distance is smaller than a predefined threshold θ_{match}, the centroid is considered a match to an existing fruit and retains its original ID; otherwise, it is treated as a newly detected fruit, added to the database, and assigned a new tracking ID. Finally, the counting results are output, and a 3D point cloud containing only fruit points is saved.

3.1Implementation details

To comprehensively evaluate the performance of MSA-MVSNet, this study conducted both quantitative and qualitative experiments on a public dataset as well as a self-collected orchard dataset. The experimental setup for training and testing is described as follows: multi-view image data were captured using the main camera of an iPhone 13 (focal length 26 mm, aperture f/1.6, 12 MP). Model computations were performed on an NVIDIA RTX 4070 GPU (12 GB VRAM) with an i5-13600KF CPU and 64 GB RAM. The software environment was based on PyTorch 1.11.0, TorchVision 0.12.0, and CUDA 11.3.

MSA-MVSNet was trained on the DTU dataset, which contains 27,097 samples. During training, the input consisted of 3-view images, while 5-view images were used for testing (Yao et al., 2018). The total number of depth hypotheses was set to 192, with a depth interval scaling factor of 1.06. Input images were resized to a resolution of 512 × 640. For coarse-to-fine regularization, depth hypotheses were sampled within a range of 425 mm to 935 mm. Plane-sweep depth estimation was conducted in three stages, with the number of depth hypotheses per stage set to 48, 32, and 8, and depth interval ratios set to 4, 2, and 1, respectively.

A multi-stage depth supervision strategy was adopted, with stage-wise depth loss weights set to 0.5, 1.0, and 2.0, and the loss computed only over valid pixels. The Adam optimizer was used, and the learning rate was dynamically adjusted using a cosine annealing schedule, starting at 0.001 and decayed to half of the previous value at the 10th, 12th, and 14th epochs.

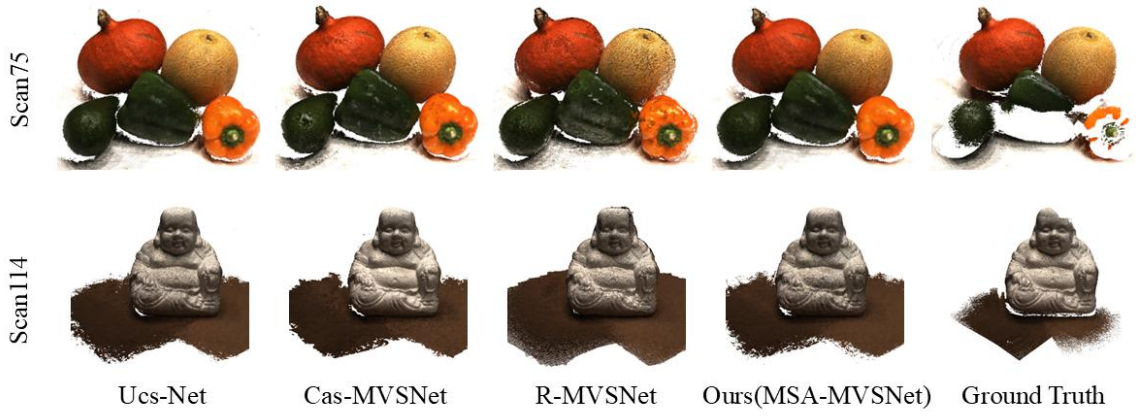

Fig. 5 Qualitative comparison of MSA-MVSNet and other methods on the DTU dataset.

3.2Results with public datasets

3.2.1 Evaluation on DTU dataset

MSA-MVSNet was evaluated on the DTU benchmark with a maximum input image resolution of $1152 \times 864$for inference, and reconstruction results were analyzed using the official evaluation metrics. As shown in Fig. 5, thanks to the collaborative effect of EFG-FPN and FMT, MSA-MVSNet demonstrates superior visual quality in terms of both geometric completeness and surface detail reconstruction, producing point clouds with richer details and more complete structural integrity. By comparison, UCS-Net (Cheng et al., 2020)and Cas-MVSNet (Gu et al., 2020) exhibit noticeable noise along object boundaries and fragmented local structures. R-MVSNet (Yao et al., 2019) improves overall structural consistency but fails to accurately capture local details, showing a degree of over-smoothing. In contrast, MSA-MVSNet can suppress background noise effectively while preserving sharp object edges and fine details in complex geometric regions. Its reconstructed results visually align more closely with the ground truth, validating the effectiveness of the multi-scale collaborative attention mechanisms in multi-view feature fusion. To further substantiate the qualitative observations, the point cloud quality was quantitatively evaluated using the official MATLAB code provided by DTU. The results are summarized in Table 1. The official metrics, Accuracy and Completeness, are used to quantitatively assess the reconstruction quality of the generated point cloud (Yao et al., 2018). Accuracy measures the average absolute distance from the reconstructed point cloud to the ground-truth points, while Completeness evaluates the opposite perspective. These metrics are computed as follows:

$$\text{Accuracy}(R,G) = \frac{1}{|R|}\sum_{r\in R} \min_{g\in G} \| r-g \| \tag{15}$$

$$\text{Completeness}\ (G,R) = \frac{1}{|G|}\sum_{g\in G} \min_{r\in R} \| g-r \| \tag{16}$$

In Equations (15) and (16), R denotes the set of reconstructed point cloud points, G denotes the set of ground-truth points, g is a point in G, and r is a point in R.$|\ G\ |$ represents the number of points in G, and $\| g-r \|$ is the Euclidean distance between g and r.The Overall metric represents the average of Accuracy and Completeness, reflecting the model's overall performance, as expressed in Equation (17).As shown in Table 1, MSA-MVSNet outperforms competing methods in Accuracy, Completeness, and Overall, further confirming its superior performance in both geometric precision and reconstruction completeness.

$$\text{Overall}(R,G) = \frac{\text{Accuracy}(R,G) + \text{Completeness}(G,R)}{2} \tag{17}$$

Table 1 Quantitative results for the DTU evaluation set

Methods	Acc.(mm)↓	Comp.(mm)↓	Overall(mm)↓
Gipuma(Galliani et al., 2016)	0.283	0.873	0.578
MVSNet(Yao et al., 2018)	0.396	0.527	0.462

R-MVSNet(Yao et al., 2019)	0.383	0.452	0.417
P-MVSNet(Luo et al., 2019)	0.406	0.434	0.420
MVSCRF(Xue et al., 2019)	0.371	0.426	0.398
PointMVSNet(R. Chen et al., 2019)	0.342	0.411	0.376
Fast-MVSNet(Z. Yu & Gao, 2020)	0.336	0.403	0.370
Cas-MVSNet(Gu et al., 2020)	0.325	0.385	0.355
CVP-MVSNet(J. Yang et al., 2020)	0.296	0.406	0.351
PVA-MVSNet(Yi et al., 2020)	0.372	0.350	0.361
Vis-Net(Zhang et al., 2020)	0.369	0.361	0.365
MVSNet++(P.-H. Chen et al., 2020)	0.407	0.345	0.376
UCS-Net(Cheng et al., 2020)	0.338	0.349	0.344
D²HC-RMVSNet(Yan et al., 2020)	0.395	0.378	0.386
DeepFusion(Li et al., 2022)	0.357	0.502	0.429
AA-RMVSNet(Wei et al., 2021)	0.376	0.339	0.357
PatchmatchNet(F. Wang et al., 2021)	0.427	0.277	0.352
Ours	0.335	0.317	0.326

3.2.2 Evaluation on tanks and temples dataset

To evaluate the generalization ability of the model, we tested a version trained solely on the DTU dataset without any fine-tuning on the Tanks and Temples (T&T) intermediate scenes (Knapitsch et al., 2017). Official metrics—including Precision, Recall, and F1 score—were used to compare MSA-MVSNet with state-of-the-art methods.To ensure an objective and rigorous evaluation, all networks were tested under consistent conditions regarding camera parameters, depth ranges, and the number of input views. In this evaluation, let G denote the ground-truth point set and R denote the reconstructed point set to be evaluated. The distance from a reconstructed point $r \in R$ to the ground truth is defined as:

$$d(r, G) = \min_{g \in G} |r - g| \tag{18}$$

These distances can be aggregated to define the Precision of the **reconstruction** P for any given distance threshold τ:

$$P(\tau) = \frac{100}{|R|} \sum_{r \in R} \prod [d(r, G) < \tau] \tag{19}$$

Here, $\prod$ denotes the Iverson bracket. $P(\tau)$ is defined on the interval $[0, 100]$ and is interpreted as a percentage. Similarly, for a ground-truth point $g \in G$, the distance to the reconstructed point set is defined as:

$$d(g, R) = \min_{r \in R} |g - r| \tag{20}$$

At a distance threshold τ, the reconstruction Recall R is defined as:

$$R(\tau) = \frac{100}{|G|} \sum_{g \in G} \prod [d(g, R) < \tau] \tag{21}$$

Precision and Recall can be combined into a summary measure, the F-score:

$$F(\tau) = \frac{2P(\tau)R(\tau)}{P(\tau) + R(\tau)} \tag{22}$$

At a distance threshold τ, the F-score is defined as the harmonic mean of precision and recall at that threshold. After testing, the reconstructed point clouds were uploaded to the Tanks and Temples website to compute F-scores for all scenes as well as the average. As shown in Table 2, MSA-MVSNet achieved superior performance among state-of-the-art methods, attaining an average F-score of 53.58. we fine-tuned the model trained on the DTU dataset using BlendedMVS, a comprehensive synthetic dataset specifically designed for MVS training. BlendedMVS includes a wide variety of outdoor objects and scenes, and fine-tuning on this dataset significantly improves reconstruction accuracy for orchard fruit trees. The model was trained on BlendedMVS for 16 epochs with a batch size of 1 using the same training strategy. After fine-tuning, inference was performed again on the same T&T evaluation set, and the reconstructed results were uploaded to the T&T official website for evaluation. The results, shown in Table 3, indicate that both precision and recall improved after fine-tuning, Precision increased by 2.72%, Recall by 1.46%, and the F-score by 3.09%. These improvements demonstrate that the fine-tuned model is better suited for outdoor scenarios. The fine-tuned MSA-MVSNet is therefore regarded as the final model.

Table 2 Reconstruction quality on T&T evaluation set

Methods	Intermediate								
	Mean	Family	Francis	Horse	L.H.	M60	Panther	P.G.	Train
Colmap(Schonberger & Frahm, 2016)	42.14	50.41	22.25	25.63	56.43	44.83	46.97	48.53	42.04
MVSNet (Yao et al., 2018)	43.48	55.99	28.55	25.07	50.79	53.96	50.86	47.90	34.69
R-MVSNet (Yao et al., 2019)	48.40	69.96	46.65	32.59	42.95	51.88	48.80	52.00	42.38
Point-MVSNet (R. Chen et al., 2019)	48.27	61.79	41.15	34.20	50.79	51.97	50.85	52.38	43.06
PatchmatchNet (F. Wang et al., 2021)	53.15	66.99	52.64	43.24	54.87	52.87	49.54	54.21	50.81
UCS-Net (Cheng et al., 2020)	54.83	76.09	53.16	43.03	54.00	55.60	51.49	57.38	47.89
Cas-MVSNet(Gu et al., 2020)	56.42	76.36	58.45	46.20	55.53	56.11	54.02	58.17	46.56
Ours(Before fine-tuning)	53.58	72.57	47.13	43.64	52.97	53.47	53.92	57.53	47.41

Table 3 Reconstruction quality after fine-tuning experiment on T&T

	Precision(%)	Recall(%)	F-score(%)
Before fine-tuning	42.55	74.20	53.58
After fine-tuning	45.27	75.66	56.67

Table 4 Quantitative analysis various modules on the DTU evaluation dataset in terms of overall performance (Overall)

Method configuration			Acc.(mm)	Comp.(mm)	Overall.(mm)
EFG-FPN	MSFE	FMT			
√	×	×	0.353	0.349	0.351
√	√	×	0.341	0.325	0.333
×	√	×	0.346	0.332	0.339
√	√	√	0.335	0.317	0.326

3.3 Experimental results on the fruit tree dataset

As described in Section 2.1, for each orchard scene, video sequences were captured using a multi-level, height-surrounding approach, and frames were extracted from the videos using FFmpeg to generate multi-view images. These multi-view fruit tree images were input to COLMAP for camera pose estimation, and the Colmap2MVSNet tool was used to convert the data into the required format, producing camera intrinsics, extrinsics, and view correspondences for each perspective, ultimately constructing a complete orchard scene dataset.

Due to the difficulty of obtaining high-precision ground-truth 3D point clouds for orchard scenes, the 3D reconstructions in this study were generated using the model trained and fine-tuned on the DTU and BlendedMVS datasets. To evaluate the reconstruction accuracy of MSA-MVSNet on orchard point clouds, Cas-MVSNet was also used for comparison. Both quantitative and qualitative analyses were performed to systematically assess reconstruction performance.

For quantitative evaluation, 75 orchard scenes were used. Corresponding points between point clouds were computed to calculate the mean and variance of point-to-point distances. To ensure objective comparison, point clouds were first aligned via coordinate transformation. Then, for each pair of corresponding points, the Euclidean distance was computed, and the mean and variance were obtained using Equation (23) (Zeng et al., 2025).Across the 75 test scenes, the distribution of point-to-point means and variances is shown in Fig. 6. The quantitative analysis demonstrates that MSA-MVSNet achieves high reconstruction accuracy and stability in most scenes: over 80% of scenes have mean and variance concentrated around 0.035 mm and 0.042 mm², respectively. This pattern indicates a high alignment of corresponding points, confirming that both algorithms achieve excellent reconstruction consistency and similarity under typical conditions.However, in a few challenging scenes (IDs 10, 41, and 62), error metrics exhibited significant spikes, with mean and variance increasing sharply. These fluctuations indicate that while MSA-MVSNet performs well overall, it encounters difficulties in scenes with severe occlusions, large weak-texture areas, or highly repetitive structures, leading to localized surface deviations or inconsistencies. This suggests that although the overall reconstruction quality is high, further improvement is needed to enhance the model's generalization to extreme or complex scenarios. Future work will focus on improving the model's adaptability and robustness to challenging scene features.For qualitative evaluation, typical orchard scenes were examined. In single-tree reconstructions (Fig. 7), Cas-

MVSNet exhibits clear structural deficiencies: sparse and broken canopy regions, distorted surface textures, and loss of detail. In contrast, MSA-MVSNet (Fig. 7a) preserves fine details: leaf textures are precise, branch layering is distinct, branch directions and bifurcations are accurately reconstructed, fruit shapes and spatial distributions are consistent, and the overall canopy is full, continuous, and naturally contoured.To assess reconstruction performance in large-scale orchards, the Fuji-sfm dataset was used for further comparison. In structured, vertical-wall orchards (Fig. 7, lower panels), differences between the models are more pronounced. MSA-MVSNet maintains detailed orchard structure, coherent branches, controlled background noise, and good retention of weak-texture areas. Cas-MVSNet, however, shows more surface noise, occasional tree breakages, and weaker overall structural consistency and continuity.

$$\begin{cases} d^i = \text{sqrt}\left(\left(a_x^i - b_x^i\right)^2 + \left(a_y^i - b_y^i\right)^2 + \left(a_z^i - b_z^i\right)^2\right) \\ \text{mean} = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} d^i \\ \text{var} = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} \left(d^i - \text{mean}\right)^2 \end{cases} \tag{23}$$

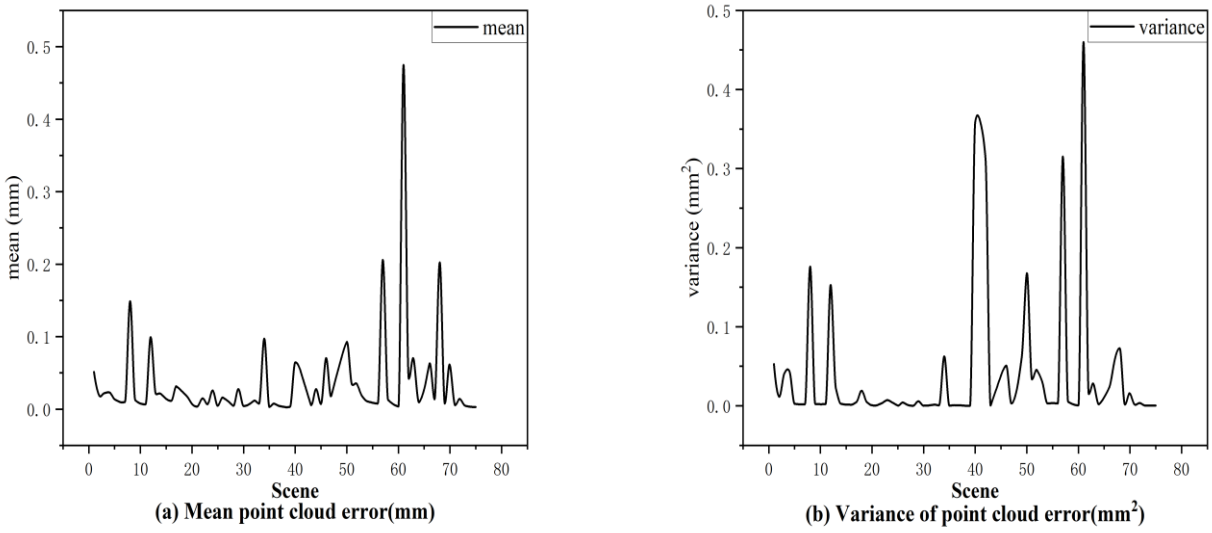

Fig. 6 (a) Mean distribution of point cloud error, (b) Variance distribution of point cloud error.



(a)MSA-MVSNet (b)Cas-MVSNet

Fig. 7 3D reconstruction results of fruit trees by MSA-MVSNet and Cas-MVSNet

3.3.1 Semantic-guided 3D Fruit Reconstruction for Counting

After comprehensively validating the reconstruction performance of the algorithm, this study introduces a reconstruction and counting framework that integrates MSA-MVSNet with YOLO-Seg (Khanam & Hussain, 2024) to achieve semantically guided fruit point cloud reconstruction. During the depth fusion stage, binary masks generated from semantic segmentation for each view are used to retain reliable pixels corresponding to fruit regions while filtering out background, branches, leaves, and other non-target pixels. Combined with an incremental tracking mechanism, the framework associates fruits across multiple views and performs spatial counting.To evaluate the framework's performance in real orchard environments, experiments were conducted on the FUJI vertical-wall orchard dataset and a self-built orchard dataset. YOLOv11-Seg was selected as the segmentation model. It maintains high efficiency while achieving precise segmentation, and its improved feature pyramid and segmentation head design produce fine-grained pixel-level masks, providing critical quality assurance for subsequent fruit reconstruction and counting.To further enhance the model's robustness in complex orchard scenarios, the dataset was systematically constructed to account for varied weather conditions (sunny, overcast, cloudy) and different canopy sparsity levels, covering diverse lighting, occlusion, and target-scale variations. This improves the model's ability to detect occluded and small fruits. On the self-built dataset, YOLOv11m-Seg achieved an average precision of 0.94, with single-image inference taking only 8.2 ms, demonstrating high efficiency and accuracy sufficient for real-time multi-view processing, laying a solid foundation for subsequent 3D reconstruction and counting.As illustrated in Fig. 8, in well-lit scenes with clearly visible fruits (left panel), the model accurately detects and localizes prominent foreground fruits with high confidence. In more challenging scenes with low light, strong contrast variations, and heavy leaf occlusions (right panel), the model still successfully detects and segments fruits, including small and partially occluded targets. Although detection confidence decreases for these difficult samples, the model is able to

identify them consistently, validating the effectiveness of training on a dataset with diverse conditions.



Fig. 8 Scene recognition and segmentation effect of fruit trees

In the semantically guided fruit reconstruction and counting experiments, fruit semantic segmentation is integrated into the depth map filtering stage to achieve fruit point cloud reconstruction guided by semantic information. The overall processing workflow of the framework is illustrated in Fig. 9.

To evaluate performance under varying fruit densities, six representative orchard scenes were selected from the self-built dataset for fruit counting experiments. To assess the generalization ability of the method, additional testing was performed on the public Fuji-sfm dataset, which features a structured planting environment and serves as a standard benchmark for evaluating fruit tree reconstruction methods.The fruit point clouds obtained from semantic-guided reconstruction are shown in Fig. 10. In high-density orchard scenes ((Fuji), (d), and (f)), severe occlusions and extremely high fruit densities pose significant challenges for 3D reconstruction and segmentation algorithms. Due to occlusion, some fruits are not fully segmented, resulting in fragmented or incomplete point clouds. These incomplete point clouds, combined with residual background points and leaf noise, produce noticeable spurious points. This indicates that during the depth map filtering stage, non-target information such as background and leaves was not fully removed, and the remaining noise may interfere with subsequent centroid-based counting, potentially causing false positives or missed detections.To quantitatively evaluate the performance of the counting framework, precision (P_r), recall (R_e), and F1 score were adopted as evaluation metrics (Shao et al., 2024).

$$P_r = \frac{T_P}{T_P + F_P} \times 100\% \tag{24}$$

$$R_e = \frac{T_P}{T_P + F_N} \times 100\% \tag{25}$$

$$F1 = \frac{2P_r R_e}{P_r + R_e} \tag{26}$$

Here, T_P denotes the number of correctly detected fruits, F_N the number of missed fruits, and F_P the number

of falsely detected fruits. Table 5 presents the fruit counting results for all test scenes. The proposed semantic-guided 3D reconstruction and counting framework achieves an overall precision of 95.3%, recall of 90.4%, and a combined F1 score of 92.8% across various scenarios.In sparse, lightly occluded scenes (e.g., scene c), the algorithm achieves 100% in all evaluation metrics. In heavily occluded, high-density scenes ((d) and (f)), precision remains above 95%, while recall is slightly lower due to undetected fruits. On the independent public benchmark Fuji-sfm dataset, the method attains 95.0% precision and 92.1% recall, with recall outperforming several high-density scenes from the self-built dataset, demonstrating that the proposed approach has strong generalization capability.

Table 5 Counting results of fruit number

Type	Real number	Measured number	T_P	F_P	F_N	$R_e/\%$	$P_r/\%$	$F1/\%$
a	60	57	54	3	6	90.0	94.7	92.3
b	47	42	40	2	7	85.1	95.2	89.9
c	15	15	15	0	0	100.0	100.0	100.0
d	304	270	260	10	44	85.5	96.3	90.5
e	52	47	45	2	7	86.5	95.7	90.9
f	183	168	160	8	23	87.4	95.2	91.2
Fuji	1455	1410	1340	70	45	92.1	95.0	93.5
Total	2116	2009	1914	95	202	90.4	95.3	92.8

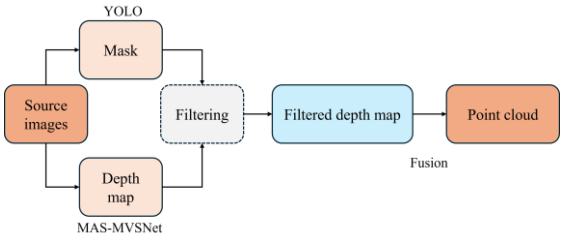

Fig. 9 Semantic-guided fruit reconstruction counting framework processing flowchart

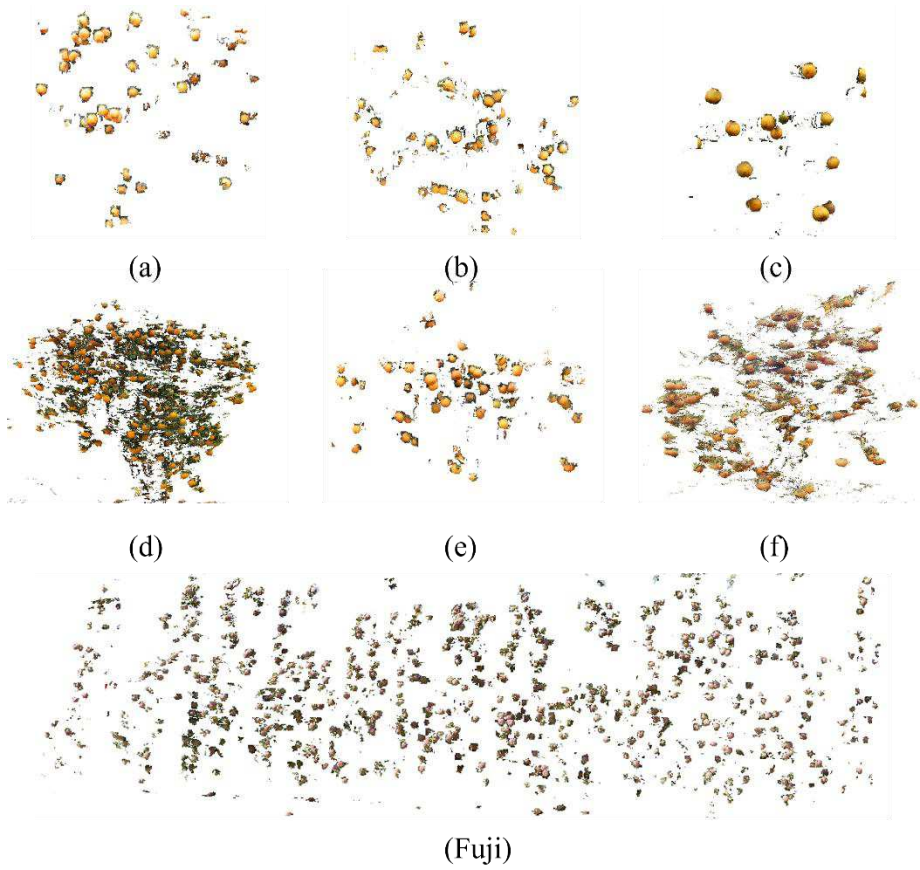

Fig. 10 Semantic-guided three-dimensional reconstruction of fruits

3.4 Ablation experiment

To evaluate the effectiveness of each module in the network, ablation experiments were conducted on the DTU validation set to quantify the contributions of EFG-FPN, MSFE, and FMT, as summarized in Table 4.In the first experiment, a baseline model containing only the EFG-FPN module was constructed. The use of MSRCB and FAM enhanced the perception of complex branch and leaf geometries, achieving an overall error of 0.351 mm. Although the enhanced local feature extraction was effective, the model still lacked global context modeling and cross-view matching capabilities.The second experiment introduced the MSFE module at the skip connections of EFG-FPN. With its multi-scale convolution branch and spatial-channel collaborative attention (SCSA), MSFE enabled adaptive weighted fusion of shallow details and deep semantic features, significantly improving feature discriminability. Incorporating MSFE increased overall performance to 0.333 mm.In the third experiment, only the MSFE module was used (without EFG-FPN), resulting in an overall error of 0.339 mm. While this outperformed the standalone EFG-FPN baseline, it was still inferior to the combination of EFG-FPN and MSFE, demonstrating the complementarity between the geometry-aware base features provided by EFG-FPN and the feature enhancement of MSFE. Their synergy enables more effective feature encoding.Finally, the fourth experiment incorporated the FMT module. By leveraging intra-attention and inter-attention, FMT explicitly modeled long-range dependencies within each view and global correspondences across views at the feature level. This effectively mitigated matching errors in low-texture

and occluded regions. With FMT, the model achieved the best performance, reducing the overall error to 0.326 mm, while both accuracy and completeness reached their optimal values.

These ablation results clearly demonstrate the incremental contribution of each module,EFG-FPN provides a robust foundation for geometry-aware feature extraction,MSFE achieves efficient multi-scale feature fusion and enhancement,FMT addresses the core challenge of cross-view global matching through global attention.The modules work progressively and synergistically, producing reconstruction accuracy superior to any partial combination of the modules.

4. Discussion

This study proposes MSA-MVSNet for 3D reconstruction of fruit tree scenes. The algorithm integrates mechanisms such as multi-scale feature enhancement, deformable convolutions, and a feature matching transformer, effectively improving the reconstruction quality of complex canopy structures in orchard environments. Evaluations on both public datasets and real-world orchard scenarios demonstrate that the proposed method achieves advantages in overall accuracy and detail completeness, validating the effectiveness of the design. However, point cloud error analysis indicates that in a few challenging scenes with extreme occlusions or large low-texture regions, reconstruction quality can still fluctuate.

The algorithm has relatively high computational complexity, which may affect reconstruction efficiency when processing higher-resolution images or a larger number of views. Future work will focus on enhancing the model's robustness to extreme scenarios and on algorithm lightweighting and acceleration optimization. It is worth noting that while depth fusion methods like Gipuma are commonly used to enhance the completeness of MVSNet-based reconstructions, they can also introduce some accuracy loss. The choice of post-processing strategy and its impact on final point cloud quality still requires careful consideration in practical applications.

5.Conclusion

This study proposes a cross-scale collaborative attention-based multi-view reconstruction network, MSA-MVSNet, for high-quality 3D reconstruction of fruit tree scenes. Experimental results show that the proposed method outperforms mainstream baseline models on public datasets. Its effectiveness stems primarily from three core designs: EFG-FPN, which enhances the model's perception of irregular structures; MSFE, which facilitates effective cross-scale feature fusion; and FMT, which improves the robustness of global feature matching.In real orchard scenarios, the reconstructed point clouds demonstrate clear advantages in detail completeness and structural coherence. By integrating YOLOv11-Seg into a semantic-guided counting framework, the system can perform fruit counting, providing a practical solution for yield estimation and automated orchard management. However, the method may still exhibit counting deviations in challenging scenarios, such as densely packed or heavily occluded citrus trees, and computational efficiency remains a concern.

Future work will focus on enhancing model robustness in challenging conditions, optimizing computational efficiency, and further exploring end-to-end joint optimization of reconstruction and

segmentation tasks.

Funding Declaration

This research was supported by the earmarked fund for China Agriculture Research System (CARS-26-05BY).

Declaration

Conflict of interest All authors declare no conflict of interest.

Corresponding Author Email

Hui Li

Email: 3023763876@qq.com

Author Information

Hui Li
School of Information Engineering, Huzhou University
Huzhou, China

Jun Zhang
Food Science Research Institute, Zhejiang Academy of Agricultural Sciences
Hangzhou, China

Jianhua Hong
School of Information Engineering, Huzhou University
Huzhou, China

Fuzhi Ke
Zhejiang Citrus Research Institute, Zhejiang Academy of Agricultural Sciences
Taizhou, China

Zhixin Li
Food Science Research Institute, Zhejiang Academy of Agricultural Sciences
Hangzhou China

Chu Zhang
School of Information Engineering, Huzhou University
Huzhou, China

References

Besl, P. J., & McKay, N. D. (1992). Method for registration of 3-D shapes. *Sensor Fusion IV: Control Paradigms and Data Structures*, *1611*, 586–606. https://www.spiedigitallibrary.org/conference-proceedings-of-spie/1611/1/Method-for-registration-of-3-D-shapes/10.1117/12.57955.short

Chen, P.-H., Yang, H.-C., Chen, K.-W., & Chen, Y.-S. (2020). MVSNet++: Learning depth-based attention pyramid features for multi-view stereo. *IEEE Transactions on Image Processing*, *29*, 7261–7273.

Chen, R., Han, S., Xu, J., & Su, H. (2019). Point-based multi-view stereo network. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1538–1547. http://openaccess.thecvf.com/content_ICCV_2019/html/Chen_Point-Based_Multi-View_Stereo_Network_ICCV_2019_paper.html

Cheng, S., Xu, Z., Zhu, S., Li, Z., Li, L. E., Ramamoorthi, R., & Su, H. (2020). Deep stereo using adaptive thin volume representation with uncertainty awareness. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2524–2534. http://openaccess.thecvf.com/content_CVPR_2020/html/Cheng_Deep_Stereo_Using_Adaptive_Thin_Volume_Representation_With_Uncertainty_Awareness_CVPR_2020_paper.html

Chunyan, M., Lihua, Z., Hong, S., & Wei, Y. (2015). Research on 3D reconstruction of fruit tree and fruit recognition and location method based

on RGB-D camera. *Nongye Jixie Xuebao/Transactions of the Chinese Society of Agricultural Machinery*. https://www.nyjxxb.net/index.php/journal/article/view/373

Ding, Y., Yuan, W., Zhu, Q., Zhang, H., Liu, X., Wang, Y., & Liu, X. (2021). *TransMVSNet: Global Context-aware Multi-view Stereo Network with Transformers* (No. arXiv:2111.14600). arXiv. https://doi.org/10.48550/arXiv.2111.14600

Galliani, S., Lasinger, K., & Schindler, K. (2016). Gipuma: Massively parallel multi-view stereo reconstruction. *Publikationen Der Deutschen Gesellschaft Für Photogrammetrie, Fernerkundung Und Geoinformation e. V*, *25*(361–369), 2.

Gené-Mola, J., Sanz-Cortiella, R., Rosell-Polo, J. R., Morros, J.-R., Ruiz-Hidalgo, J., Vilaplana, V., & Gregorio, E. (2020). Fuji-SfM dataset: A collection of annotated images and point clouds for Fuji apple detection and location using structure-from-motion photogrammetry. *Data in Brief*, *30*, 105591.

Gu, X., Fan, Z., Zhu, S., Dai, Z., Tan, F., & Tan, P. (2020). Cascade Cost Volume for High-Resolution Multi-View Stereo and Stereo Matching. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2492–2501. https://doi.org/10.1109/CVPR42600.2020.00257

Han, K., Xiao, A., Wu, E., Guo, J., Xu, C., & Wang, Y. (2021). Transformer in transformer. *Advances in Neural Information Processing Systems*, *34*, 15908–15919.

Jensen, R., Dahl, A., Vogiatzis, G., Tola, E., & Aanæs, H. (2014). Large scale multi-view stereopsis evaluation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 406–413. https://www.cv-foundation.org/openaccess/content_cvpr_2014/html/Jensen_Large_Scale_Multi-view_2014_CVPR_paper.html

Jin, S., Sun, X., Wu, F., Su, Y., Li, Y., Song, S., Xu, K., Ma, Q., Baret, F., & Jiang, D. (2021). Lidar sheds new light on plant phenomics for plant breeding and management: Recent advances and future prospects. *ISPRS Journal of Photogrammetry and Remote Sensing*, *171*, 202–223.

Khanam, R., & Hussain, M. (2024). *YOLOv11: An Overview of the Key Architectural Enhancements* (No. arXiv:2410.17725). arXiv. https://doi.org/10.48550/arXiv.2410.17725

Knapitsch, A., Park, J., Zhou, Q.-Y., & Koltun, V. (2017). Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (ToG)*, *36*(4), 1–13.

Kok, E., Wang, X., & Chen, C. (2023). Obscured tree branches segmentation and 3D reconstruction using deep learning and geometrical constraints. *Computers and Electronics in Agriculture*, *210*, 107884.

Li, Y., Yu, A. W., Meng, T., Caine, B., Ngiam, J., Peng, D., Shen, J., Lu, Y., Zhou, D., & Le, Q. V. (2022). Deepfusion: Lidar-camera deep fusion for multi-modal 3d object detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17182–17191. http://openaccess.thecvf.com/content/CVPR2022/html/Li_DeepFusion_Lidar-Camera_Deep_Fusion_for_Multi-Modal_3D_Object_Detection_CVPR_2022_paper.html

Luo, K., Guan, T., Ju, L., Huang, H., & Luo, Y. (2019). P-mvsnet: Learning patch-wise matching confidence aggregation for multi-view stereo. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10452–10461. http://openaccess.thecvf.com/content_ICCV_2019/html/Luo_P-MVSNet_Learning_Patch-Wise_Matching_Confidence_Aggregation_for_Multi-View_Stereo_ICCV_2019_paper.html

Nascimento, M. G. do, Fawcett, R., & Prisacariu, V. A. (2019). Dsconv: Efficient convolution operator. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5148–5157. http://openaccess.thecvf.com/content_ICCV_2019/html/do_Nascimento_DSConv_Efficient_Convolution_Operator_ICCV_2019_paper.html

Remondino, F., Spera, M. G., Nocerino, E., Menna, F., & Nex, F. (2014). State of the art in high density image matching. *The Photogrammetric Record*, *29*(146), 144–166. https://doi.org/10.1111/phor.12063

Schonberger, J. L., & Frahm, J.-M. (2016). Structure-from-motion revisited. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4104–4113. https://www.cv-foundation.org/openaccess/content_cvpr_2016/html/Schonberger_Structure-From-Motion_Revisited_CVPR_2016_paper.html

Shao, H., Li, X., Wang, F., Sun, L., Wang, C., & Hu, Y. (2024). Feasibility study on fruit parameter estimation based on hyperspectral LiDAR point cloud. *Journal of Food Measurement and Characterization*, *18*(8), 7185–7197. https://doi.org/10.1007/s11694-024-02729-x

Si, Y., Xu, H., Zhu, X., Zhang, W., Dong, Y., Chen, Y., & Li, H. (2025). SCSA: Exploring the synergistic effects between spatial and channel attention. *Neurocomputing*, *634*, 129866.

Wan, H., Zeng, X., Fan, Z., Chen, Q., Zhang, K., & Zhang, H. (2025). ET-PatchNet: A low-memory, efficient model for Multi-view Stereo with a case study on the 3D reconstruction of fruit tree branches. *Computers and Electronics in Agriculture*, *237*, 110459. https://doi.org/10.1016/j.compag.2025.110459

Wang, F., Galliani, S., Vogel, C., Speciale, P., & Pollefeys, M. (2021). Patchmatchnet: Learned multi-view patchmatch stereo. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14194–14203. http://openaccess.thecvf.com/content/CVPR2021/html/Wang_PatchmatchNet_Learned_Multi-View_Patchmatch_Stereo_CVPR_2021_paper.html

Wang, R., Shivanna, R., Cheng, D., Jain, S., Lin, D., Hong, L., & Chi, E. (2021). Dcn v2: Improved deep & cross network and practical lessons for web-scale learning to rank systems. *Proceedings of the Web Conference 2021*, 1785–1797. https://dl.acm.org/doi/abs/10.1145/3442381.3450078

Wei, Z., Zhu, Q., Min, C., Chen, Y., & Wang, G. (2021). Aa-rmvsnet: Adaptive aggregation recurrent multi-view stereo network. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6187–6196. http://openaccess.thecvf.com/content/ICCV2021/html/Wei_AA-RMVSNet_Adaptive_Aggregation_Recurrent_Multi-View_Stereo_Network_ICCV_2021_paper.html

Xue, Y., Chen, J., Wan, W., Huang, Y., Yu, C., Li, T., & Bao, J. (2019). Mvscrf: Learning multi-view stereo with conditional random fields. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4312–4321. http://openaccess.thecvf.com/content_ICCV_2019/html/Xue_MVSCRF_Learning_Multi-View_Stereo_With_Conditional_Random_Fields_ICCV_2019_paper.html

Yan, J., Wei, Z., Yi, H., Ding, M., Zhang, R., Chen, Y., Wang, G., & Tai, Y.-W. (2020). Dense hybrid recurrent multi-view stereo net with dynamic consistency checking. *European Conference on Computer Vision*, 674–689. https://link.springer.com/chapter/10.1007/978-3-030-58548-8_39

Yang, D., Yang, H., Liu, D., & Wang, X. (2024). Research on automatic 3D reconstruction of plant phenotype based on Multi-View images.

Computers and Electronics in Agriculture, *220*, 108866.

Yang, J., Mao, W., Alvarez, J. M., & Liu, M. (2020). Cost volume pyramid based depth inference for multi-view stereo. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4877–4886. http://openaccess.thecvf.com/content_CVPR_2020/html/Yang_Cost_Volume_Pyramid_Based_Depth_Inference_for_Multi-View_Stereo_CVPR_2020_paper.html

Yao, Y., Luo, Z., Li, S., Fang, T., & Quan, L. (2018). Mvsnet: Depth inference for unstructured multi-view stereo. *Proceedings of the European Conference on Computer Vision (ECCV)*, 767–783. http://openaccess.thecvf.com/content_ECCV_2018/html/Yao_Yao_MVSNet_Depth_Inference_ECCV_2018_paper.html

Yao, Y., Luo, Z., Li, S., Shen, T., Fang, T., & Quan, L. (2019). Recurrent mvsnet for high-resolution multi-view stereo depth inference. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5525–5534. http://openaccess.thecvf.com/content_CVPR_2019/html/Yao_Recurrent_MVSNet_for_High-Resolution_Multi-View_Stereo_Depth_Inference_CVPR_2019_paper.html

Yao, Y., Luo, Z., Li, S., Zhang, J., Ren, Y., Zhou, L., Fang, T., & Quan, L. (2020). Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1790–1799. http://openaccess.thecvf.com/content_CVPR_2020/html/Yao_BlendedMVS_A_Large-Scale_Dataset_for_Generalized_Multi-View_Stereo_Networks_CVPR_2020_paper.html

Yi, H., Wei, Z., Ding, M., Zhang, R., Chen, Y., Wang, G., & Tai, Y.-W. (2020). Pyramid Multi-view Stereo Net with Self-adaptive View Aggregation. In A. Vedaldi, H. Bischof, T. Brox, & J.-M. Frahm (Eds.), *Computer Vision – ECCV 2020* (Vol. 12354, pp. 766–782). Springer International Publishing. https://doi.org/10.1007/978-3-030-58545-7_44

Yu, F., & Koltun, V. (2016). *Multi-Scale Context Aggregation by Dilated Convolutions* (No. arXiv:1511.07122). arXiv. https://doi.org/10.48550/arXiv.1511.07122

Yu, Z., & Gao, S. (2020). Fast-mvsnet: Sparse-to-dense multi-view stereo with learned propagation and gauss-newton refinement. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1949–1958. http://openaccess.thecvf.com/content_CVPR_2020/html/Yu_Fast-MVSNet_Sparse-to-Dense_Multi-View_Stereo_With_Learned_Propagation_and_Gauss-Newton_Refinement_CVPR_2020_paper.html

Zeng, X., Wan, H., Fan, Z., Yu, X., & Guo, H. (2025). MT-MVSNet: A lightweight and highly accurate convolutional neural network based on mobile transformer for 3D reconstruction of orchard fruit tree branches. *Expert Systems with Applications*, *268*, 126220. https://doi.org/10.1016/j.eswa.2024.126220

Zhang, J., Yao, Y., Li, S., Luo, Z., & Fang, T. (2020). *Visibility-aware Multi-view Stereo Network* (No. arXiv:2008.07928). arXiv. https://doi.org/10.48550/arXiv.2008.07928