Computers and Electronics in Agriculture 251 (2026) 111939

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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

Original papers

Binocular-depth-guided dual-view camera-LiDAR fusion for kiwifruit pose estimation in robotic harvesting

Shiao Niu [a], Leilei He [a], Qilin Gong [a], Qichen Huang [a], Ruichao Luo [a], Rui Li [a,*], Bao Zhou [b], Xiancun Zhou [b], Zhenchao Wu [c], Jinyong Chen [d], Xiaoxu Sun [d], Anastasia Grecheneva [e], Shen Hin Lim [f], Longsheng Fu [a,b,c,*]

[a] College of Mechanical and Electronic Engineering, Northwest A&F University, Yangling, Shaanxi 712100, China
[b] School of Electronic Information and Artificial Intelligence, West Anhui University, Lu'an, Anhui 237012, China
[c] School of Engineering, Anhui Agricultural University, Anhui 230036, Hefei, China
[d] Zhengzhou Fruit Research Institute, Chinese Academy of Agricultural Sciences, Zhengzhou, Henan 450009, China
[e] Russian State Agrarian University - Moscow Timiryazev Agricultural Academy, Moscow 127434, Russia
[f] School of Engineering, University of Waikato, Hamilton 3240, New Zealand

ARTICLE INFO

Keywords:
Kiwifruit harvesting
Pose estimation
Camera-LiDAR fusion
Dual-view perception
Guided filtering

ABSTRACT

Accurate detection and pose estimation of kiwifruit are essential for robotic harvesting in modern orchards, which remains challenging due to complex lighting conditions and dense canopy in natural environments. In this study, a dual-view camera-LiDAR fusion methodology based on Depth-Iterative Canopy Filtering (DICF) method was proposed for precise pose estimation of kiwifruits. The proposed methodology aimed to overcome limitations of existing camera-based fruit pose estimation research by enhancing robustness under variable lighting and complex canopy conditions. The methodology integrated a lightweight end-to-end stereo matching network with the DICF, leveraging binocular disparity maps to guide LiDAR point-cloud adhesion noise suppression. YOLO11m was employed for kiwifruit detection, dual-view fusion increased detection accuracy of occluded fruits to 92.8%, representing a 28.5% improvement over single-view detection. Experiments conducted in commercial orchards demonstrated that the methodology achieved a mean absolute localization error of 6.54 mm and an orientation angle error of 7.70°. Moreover, filtering time was reduced to 1.12 s for point clouds containing 700,000 points, compared to 3.56 s required by the TSPFA, meeting the requirements of robotic harvesting. This study established a novel and efficient methodology for kiwifruit pose estimation and provided a foundation for end-to-end multimodal sensor fusion in future agricultural robotics.

1. Introduction

Accurate detection and pose estimation of kiwifruits are crucial for robotic harvesting. In modern orchards, kiwifruits are commonly cultivated under standardized steel-framed pergola trellis systems where branches extend horizontally and fruits grow vertically downward (Liu et al., 2024), as shown in Fig. 1. This provides a relatively organized environment for robotic harvesting. For kiwifruit detection, deep learning-based models such as You Only Look Once version 8 (YOLOv8) have achieved high accuracy, reaching 95.9% mean Average Precision (*mAP*@50) in orchards (Zhao et al., 2024). However, while these methods effectively address the problem of kiwifruit detection, they represent only the initial stage of robotic perception. The subsequent process of pose estimation, which involves accurately determining picking points and orientation of each kiwifruit to ensure stable picking, remains complex and insufficiently explored. Reliable pose estimation is essential for improving precision, efficiency, and safety of kiwifruit harvesting robots.

Camera-based methods are widely applied to estimate fruit pose because of their capacity for cost-effective depth acquisition. For the localization component of pose estimation, Jing et al. (2024) combined a binocular camera with end-to-end LaC-Gwc Net stereo matching network to achieve a Mean Absolute Error (*MAE*) of 4.0 mm in the z-axis for kiwifruit calyx in standardized orchards. However, achieving high localization accuracy alone is insufficient for optimal harvesting efficiency. Fang et al. (2023) found that the force required to pick kiwifruits

* Corresponding authors.
E-mail addresses: ruili1216@nwafu.edu.cn (R. Li), fulsh@nwafu.edu.cn (L. Fu).

https://doi.org/10.1016/j.compag.2026.111939
Received 9 December 2025; Received in revised form 17 May 2026; Accepted 20 May 2026
Available online 10 June 2026
0168-1699/© 2026 Elsevier B.V. All rights are reserved, including those for text and data mining, AI training, and similar technologies.



Fig. 1. Planting pattern for kiwifruit, where the fruits grow vertically downwards from horizontal branches. Average vertical clearance between the canopy base and ground surface measured 1.7 m, with inter-row spacing maintained at 3 m.

at a fruit-stem angle of 60° was reduced by over 70% compared to picking at 180°. This necessitates estimating orientation of kiwifruits to optimize picking angles to minimize fruit damage and enhancing harvesting efficiency. Current fruit orientation estimation research mainly relies on point clouds obtained from Red-Green-Blue-Depth (RGB-D) cameras. Jang and Hwang (2024) estimated tomato orientation from RGB-D point clouds by connecting fruit and sepal centers, achieving a *MAE* of 6.8° ± 3.2° under greenhouse conditions. Luo et al. (2022) estimated orientation of grape clusters in orchards by analyzing geometric structure of RGB-D point clouds, achieving a mean error of 22.2° ± 18.0°. Although accurate fruit orientation estimation has been achieved indoors, constraints such as occlusion and dynamic lighting often make camera-based methods unreliable within complex field environments.

Light Detection and Ranging (LiDAR) provides accurate three-dimensional (3D) information with high robustness in complex environments. Its ranging principle based on emitted laser pulses ensures high precision and robustness in outdoor environments. LiDAR capability has been extensively validated in autonomous driving (Goberville et al., 2020; Karim et al., 2024), and its application in agriculture has been increasing (Chen et al., 2024). Zhang et al. (2024) confirmed the high precision of LiDAR for plant phenotyping based on a skeleton extraction algorithm, with a root mean square error of 1.659 cm in leaf length when deriving structural traits of cigar plants. Pin et al. (2025) demonstrated the robustness of LiDAR in agricultural navigation by integrating it with a density-based clustering and navigation line extraction algorithm, achieving an average lateral error of 0.059 m at 1 m/s despite weed interference. Despite these strengths, lack of semantic information in LiDAR point clouds hinders fruit detection accuracy compared to camera images, thereby limiting its application in fruit harvesting.

LiDAR offers precise depth but lacks semantic information (Elhousni and Huang, 2020), whereas cameras provide rich semantic information but limited depth accuracy in complex environments (Wang, 2021). Thus, camera-LiDAR fusion methods have potential to enable effective harvesting of kiwifruit. Kang and Wang (2023) improved apple point-cloud segmentation on PointNet++ by integrating RGB information into LiDAR point clouds, raising mean intersection over union from 0.73 to 0.88. Kang et al. (2022) applied You Only Look At CoefficienTs (YOLACT) instance segmentation network to extract apple pixels and projected LiDAR point clouds onto RGB images to obtain depth information of corresponding pixels, achieving localization Standard Deviation (SD) of 2.6 mm under varying illumination and distance conditions. Aforementioned research validates the feasibility of camera-LiDAR fusion-based fruit localization methods in agricultural applications. However, these studies only achieved accurate localization of picking points, lacking orientation estimation, which requires more precise raw point-cloud information.

In kiwifruit orchards, LiDAR-derived point clouds are often degraded by Point-Cloud Adhesion (PCA) phenomenon, which substantially compromises accuracy of pose estimation based on camera-LiDAR fusion. The phenomenon arises when diverging laser beams strike fruit edges, causing scattered reflections that generate false points between kiwifruits and surrounding leaves. Jing et al. (2024) developed a Two-Stage Partition Filtering Algorithm (TSPFA) by combining Otsu with Statistical Outlier Removal (SOR) algorithms to eliminate PCA in kiwifruit orchards (Rusu and Cousins, 2011). However, SOR algorithm suffers from high computational cost when applied to large-scale point clouds, limiting its suitability for real-time applications in robotic harvesting. Yang et al. (2019) reported that guiding LiDAR point-cloud filtering with binocular depth values effectively removed long-tail artifacts caused by moving objects such as vehicles, increasing percentage of correct pixels from 72.9% to 98.0% after two iterations. Although the long-tail artifacts differ from PCA in context, both phenomena originate from spurious point generation near depth discontinuities. Therefore, applying the guided filtering method for severe PCA mitigation in kiwifruit orchards is expected to improve the accuracy of camera-LiDAR fusion pose estimation.

A kiwifruit pose estimation methodology based on dual-view camera-LiDAR fusion and Depth-Iterative Canopy Filtering (DICF) methods was proposed to provide accurate fruit localization and orientation estimation for kiwifruit harvesting robots. The DICF incorporates guided filtering while effectively circumventing the substantial computational costs associated with the TSPFA. The methodology employs a lightweight end-to-end stereo matching network to guide PCA removal, resulting in accurate depth maps. By integrating the depth maps with YOLO-detected object maps, 3D positions and orientations of kiwifruits are calculated based on growth characteristics. Furthermore, dual-view kiwifruit pose estimation results are combined to provide more comprehensive perceptual information. The proposed methodology effectively addresses the challenges of achieving robust and accurate kiwifruit pose estimation in complex orchard environments, while simultaneously eliminating PCA without incurring substantial computational costs.

2. Materials and methods

In this study, LiDAR-derived point clouds served as the primary depth source. Left and right images captured by binocular cameras were processed through a deep learning-based stereo matching network

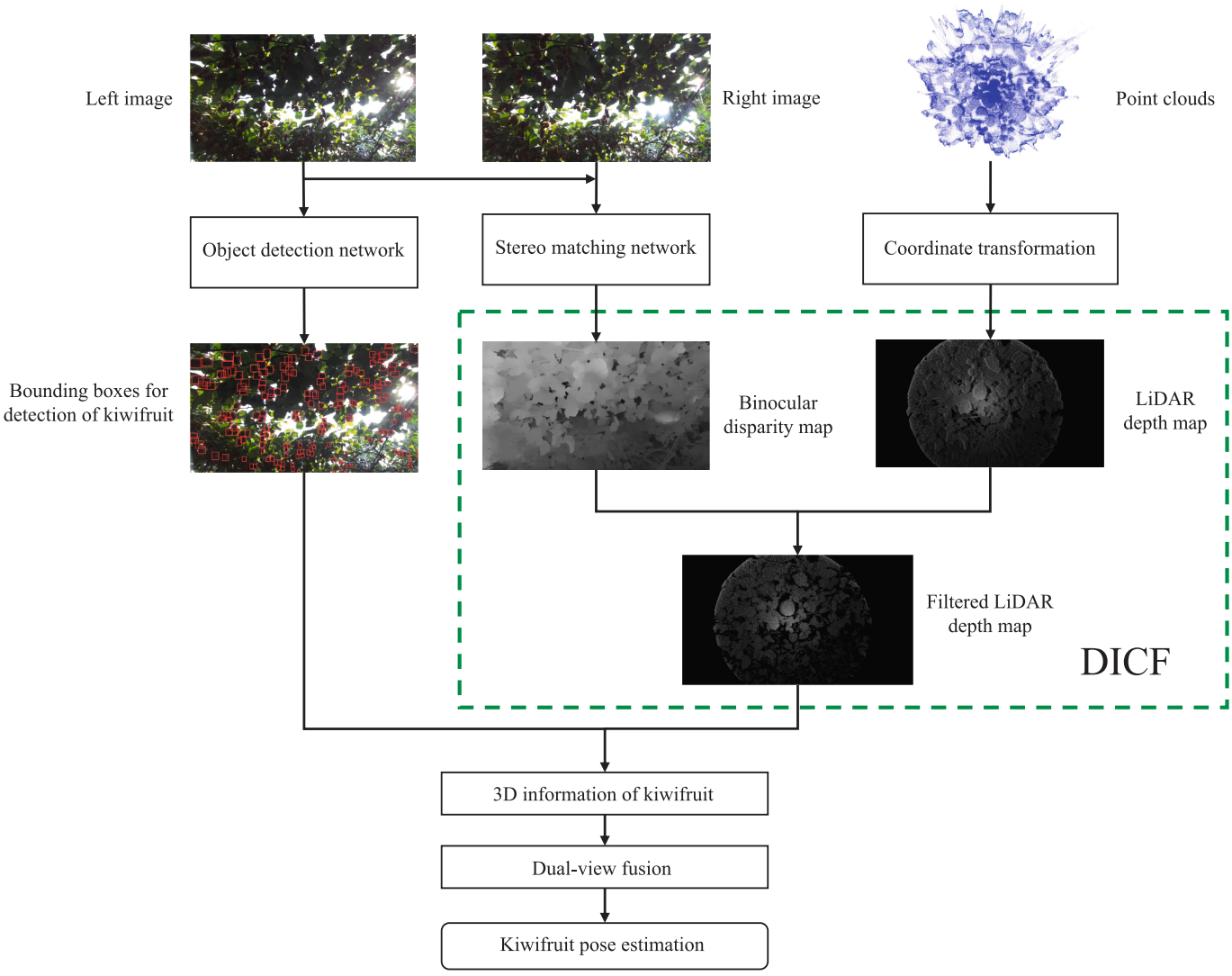


Fig. 2. Pipeline of kiwifruit pose estimation based on object detection network and the DICF.

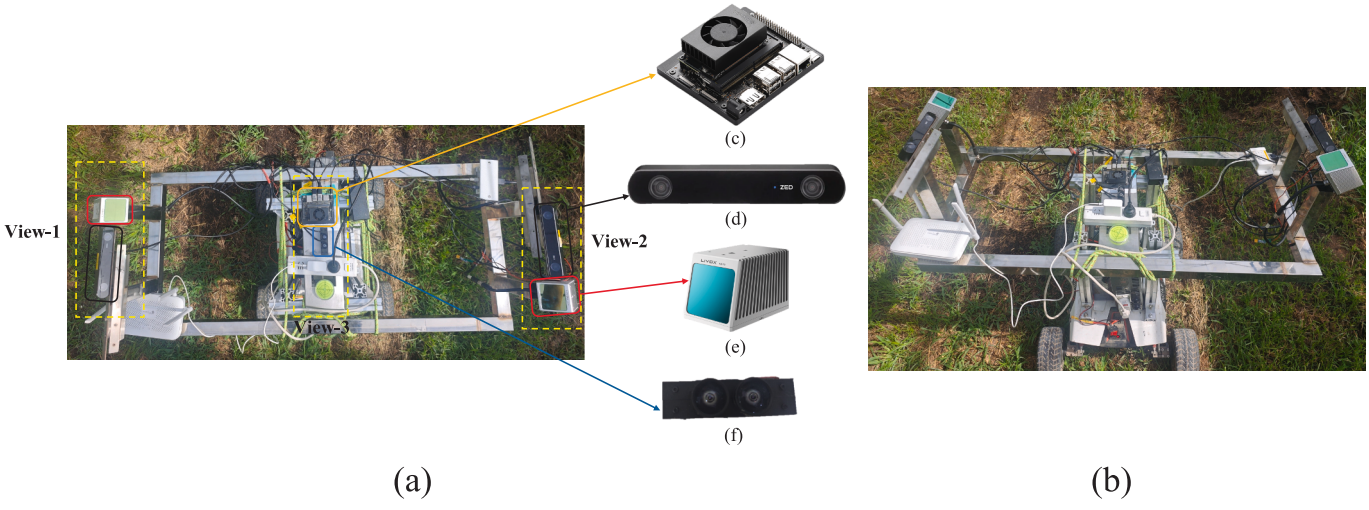


Fig. 3. (a) Top and (b) side views of data acquisition platform. The platform comprising (c) Jetson Xavier NX-based synchronization and acquisition equipment, (d) two ZED 2i binocular cameras, (e) two Livox Mid70 LiDARs and (f) a custom-assembled SH500 binocular camera, where marked boxes of the same color indicate identical sensor types. View-1 and View-2, separated by a horizontal distance of 1.8 m, were installed 1.0 m below kiwifruit canopy plane and oriented upward at a 45° inclination, while View-3 was positioned approximately 1.3 m from the canopy and aligned horizontally with ground.

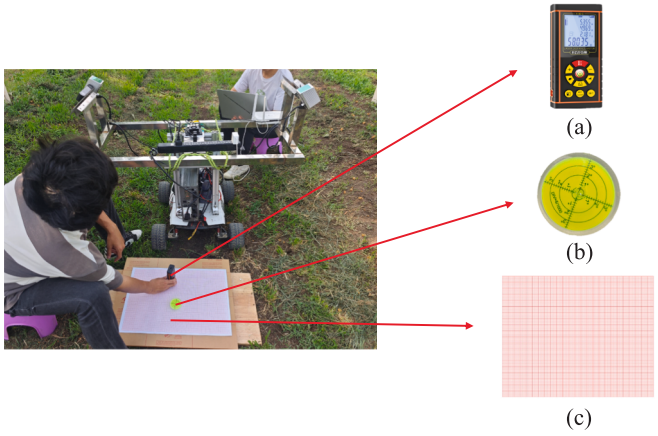


Fig. 4. Ground truth of kiwifruit pose in kiwifruit orchard is acquired using a special designed experimental platform, which consisted of (a) a laser rangefinder, (b) a level gauge, and (c) coordinate paper.

(referred to GuideNet in this study) to obtain a binocular disparity map. After sensor calibration and coordinate transformation, the LiDAR point clouds were projected onto image plane to generate a LiDAR depth map. The disparity map was introduced into the DICF as guidance to suppress PCA in LiDAR depth map, thereby generating a filtered LiDAR depth map. Furthermore, object detection was performed on the left image to identify kiwifruits. The detection results were combined with the filtered LiDAR depth map to acquire 3D spatial information of kiwifruits. Finally, dual-view fusion was employed to enhance detection robustness and integrate richer surface details for kiwifruit pose estimation. The complete processing pipeline is shown in Fig. 2.

2.1. Data acquisition and ground-truth measuring

Data acquisition was conducted during harvest period in September 2024 across two commercial orchards in Shaanxi Province, China, namely Yangling International Kiwifruit Innovation and Entrepreneurship Park (108°06′ E, 34°31′ N) and Aiki Farm (108°09′ E, 34°24′ N), where kiwifruit cultivar 'Xu Xiang' was grown. Data acquisition platform consisted of a custom-assembled SH500 binocular camera (JIERUIWEITONG, ShenZhen, China) together with two identical sensor assemblies, each comprising a ZED 2i binocular camera (Stereolabs, San Francisco, USA) and a Livox Mid70 LiDAR (Livox, ShenZhen, China). During data collection, all sensors remained fixed relative to the platform, as shown in Fig. 3. The two sensor assemblies corresponded to View-1 and View-2 representing left and right observation perspectives respectively. In the selected commercial orchards, the inter-row spacing is approximately 3.0 m with an average canopy clearance of 1.7 m. Taking this into account, the horizontal distance between View-1 and View-2 was set to 1.8 m to ensure safe intra-row navigation. These two lateral views were installed 1.0 m below kiwifruit canopy plane and oriented upwards at an inclination angle of 45° from horizontal plane. The SH500 binocular camera, which defined View-3, was positioned approximately 1.3 m below the canopy and aligned horizontally with ground. Its left lens served as origin of central horizontal reference coordinate system for alignment and fusion of the View-1 and View-2. Binocular images were recorded in 'PNG' format and LiDAR point clouds were stored in 'Rosbag' format at different time periods (from 9 AM to 7 PM), with 200 sets collected for each of the two lateral views.

Ground truth of kiwifruit pose was obtained within orchard by a custom designed experimental platform. The platform comprised a laser rangefinder (VCHON H-40, JinYun, China; random distance error of 1 mm), a vernier caliper, a level gauge, coordinate paper, and four micro height adjusting holders. It was positioned approximately 1.5 m below canopy to measure coordinates of kiwifruit calyx and fruit-stem junction as shown in Fig. 4. Prior to measurement, the platform was precisely leveled using the level gauge and micro height adjusting holders. Depth of the calyx was obtained using the laser rangefinder, while 2D horizontal coordinates were determined utilizing the coordinate paper. Since the fruit-stem junction cannot be directly observed from below, its depth was determined in two steps. First, depth coordinate of a visible surface point located directly beneath the fruit-stem junction was measured using the laser rangefinder. Then, the vernier caliper was used to measure vertical distance between the surface point and the fruit-stem junction. By adding the distance to the depth of surface point, depth coordinate of fruit-stem junction was derived. Horizontal coordinate of the fruit-stem junction was identical to the surface point. This procedure was repeated in different regions of kiwifruit orchard at three different time periods (from 9 to 11 AM, 12 to 3 PM, 4 to 7 PM). For each period, 10 kiwifruits were measured, obtaining a total of 30 ground-truth pose samples. Although the sample size of 30 ground-truth poses might be limited, these samples still provide insights into the accuracy of the proposed pose estimation methodology.

2.2. Sensor calibration

A total of 30 sets of calibration-board images and point clouds were collected indoors from different distances and angles to calibrate camera intrinsic matrices and sensor-to-sensor extrinsic matrices. Details of the calibration setup are shown in Fig. 5. Binocular cameras were calibrated using Stereo Camera Calibration Toolbox in MATLAB 2024b (https://www.mathworks.com), obtaining intrinsic matrices, distortion

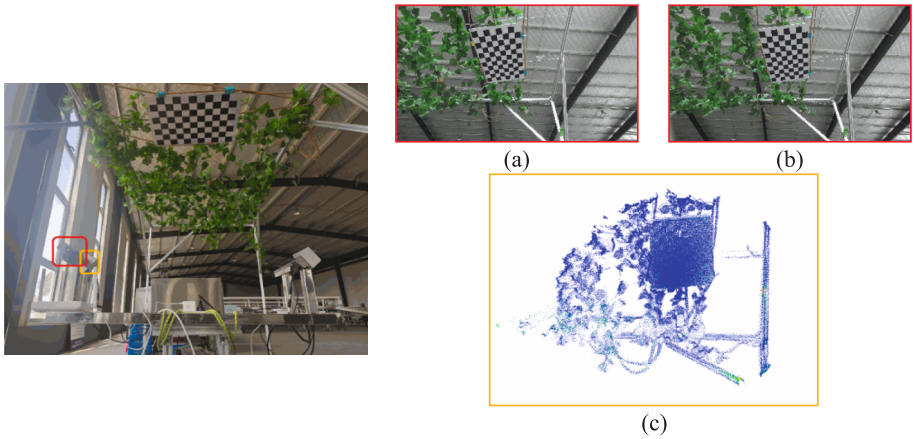


Fig. 5. Indoor setup for calibration of sensors. Representative (a) left and (b) right lens images captured by ZED 2i binocular camera correspond to red marked box, while (c) point clouds acquired by Livox Mid70 LiDAR corresponds to yellow marked box.



Fig. 6. Occlusion-aware annotation strategy, where occlusion of 50% or more and occlusion of less than 50% are indicated by (a) red and (b) blue boxes, respectively.

coefficients and baseline. Since the ZED SDK was used to capture images for the ZED 2i camera (https://www.stereolabs.com/en-hk/developers/release), the images had already undergone distortion and epipolar rectification, additional distortion calibration was not required. Extrinsic calibration of multi-sensor system was conducted using two open-source tools, calibration tool livox_camera_lidar_calibration (https://github.com/Livox-SDK/livox_camera_lidar_calibration) and point-cloud processing software CloudCompare (https://github.com/CloudCompare/CloudCompare). The extrinsic calibration consisted of intra-group and inter-group calibration, which determined spatial transformation relating ZED 2i binocular camera to LiDAR within each lateral view, and cross-view transformations relating the two lateral-view LiDARs to centrally mounted binocular camera, respectively.

2.3. Kiwifruit detection

YOLO-based object detection networks have been proven effective for kiwifruit detection in harvesting scenarios. Among existing YOLO-based models, YOLOv8m provides a strong balance between detection accuracy and computational efficiency (Jing et al., 2024). Two recent architectural iterations (YOLO11, YOLOv12) were compared with YOLOv8m model to identify optimal model configurations for real-time kiwifruit harvesting systems (Khanam and Hussain, 2024; Tian et al., 2025a).

2.3.1. Dataset construction for kiwifruit detection

Manual annotation of kiwifruits in images was conducted using LabelImg (https://github.com/HumanSignal/labelImg), with bounding boxes labeled as 'kiwi'. For occlusion under annotation, when occlusion exceeded or equaled 50% coverage, kiwifruits were enclosed within a single bounding box, as shown in Fig. 6a. Conversely, when occluded area was less than 50% of kiwifruit total size, kiwifruits were annotated separately, as shown in Fig. 6b. A total of 400 images were labeled to construct Object Detection Dataset (OD-Dataset). The OD-Dataset was initially divided into training, validation, and testing sets at a ratio of 7:2:1, respectively. Through data augmentation including image flipping and brightness adjustment, the OD-Dataset was expanded to 1,200 images.

2.3.2. Kiwifruit detection network training and testing

Kiwifruit detection was conducted through training YOLO-based networks within PyTorch framework. Training platform comprised an AMD R7-5800X quad-core CPU operating at 3.80 GHz, a GeForce GTX 3080 Ti GPU with 10,240 CUDA cores, and 12 GB memory, running Windows 10 64-bit operating system. Consistent training parameters included an input image size of 640 × 640 pixels, a batch size of 8, and a commencing learning rate of 0.01. Optimization process leveraged Stochastic Gradient Descent (SGD) with a momentum coefficient of 0.937 and weight decay of 0.0005. Training was executed through 300 complete epochs. Since kiwifruit detection in field environments typically relies on portable computing devices rather than server hardware, testing platform differed from the training platform. Testing of kiwifruit detection models was performed on a laptop configured an NVIDIA GeForce RTX 4060 GPU with 8 GB VRAM, an Intel Ultra 9 185H processor, and Windows 11 operating system, maintaining identical software configurations to the training environment. Input size of image during the test phase was 640 × 640 pixels.

2.4. Filtering of canopy point clouds acquired from LiDAR

Canopy point clouds acquired from Livox LiDAR were processed through point-cloud preprocessing, stereo matching with GuideNet, and filtering with the DICF. The DICF consisted of Binocular-Depth-Guided Filtering (BDGF) approach and two iterations of GuideNet refinement. The BDGF produced depth maps that were used as pseudo-ground truth to refine GuideNet in two iterations, generating an orchard-specific stereo matching model. During inference, however, the trained model was fixed, and only the BDGF was employed to filter LiDAR depth maps for subsequent kiwifruit pose estimation.

2.4.1. Point-cloud preprocessing

In point-cloud preprocessing stage, LiDAR point clouds were converted from 'Rosbag' format to 'PCD' format to facilitate filtering. In original 'Rosbag' data, an eight-bit tag parameter indicated probability of each point being an energy or spatial noise point. Only points with lower four bits equal to zero were retained during conversion to reduce outliers. The retained points were then projected into image coordinate system using intrinsic matrix **K** and extrinsic matrix **T** obtained from calibration in Section 2.2, generating an initial LiDAR depth map, as

Table 1
Iterative training process and dataset composition of GuideNet.

Stage	Model	Training Dataset	
		RGB images	Disparity maps
Pre-training	GuideNet-p	SceneFlow	SceneFlow
First refinement	GuideNet-i1	Orchard-specific	Orchard-specific (filtered by the BDGF using the GuideNet-p)
Second refinement	GuideNet-i2	Orchard-specific	Orchard-specific (filtered by the BDGF using the GuideNet-i1)

shown in Eq. (1).

$$Z_c \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \mathrm{K}\,\mathrm{T} \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix} \tag{1}$$

Where X_w, Y_w, and Z_w denote coordinates in LiDAR coordinate system with units of mm; **K** denotes intrinsic matrix of camera left lens; **T** represents extrinsic matrix between the camera left lens and LiDAR; Variables u and v represent pixel coordinates in image coordinate system expressed in pixels; Z_c denotes depth value in camera coordinate system with units of mm.

2.4.2. Selection of end-to-end stereo matching GuideNet

Disparity map derived from an end-to-end stereo matching network was used to filter out point-cloud noise such as PCA. Multi-scale structural guidance and cascade disparity refinement stereo matching network proposed by Jing (2025) was utilized as GuideNet. The network employed a cascaded iterative architecture and a warping-based cost volume, providing a lightweight algorithm with robust generalization across complex agricultural scenarios. Depth value at each pixel of binocular depth map was obtained from disparity map generated by GuideNet combined with baseline B, as expressed in Eq. (2).

$$Z_c = \frac{f \times B}{d_c} \tag{2}$$

Where B represents baseline of binocular camera with units in mm; f represents focal length of binocular camera left lens with units in pixels; d_c represents disparity value with units in pixels.

2.4.3. Training of GuideNet

Training of GuideNet consisted of a pre-training and iterative training stages. In the pre-training stage, GuideNet underwent training on large-scale synthetic stereo dataset SceneFlow (https://lmb.informatik.uni-freiburg.de/resources/datasets/ SceneFlowDatasets) to obtain a pre-trained model of GuideNet (GuideNet-p). The GuideNet-p was then used to guide LiDAR depth map filtering and served as the initialization for two iterations of orchard-specific training, producing the first-iteration model (GuideNet-i1) and second-iteration model (GuideNet-i2). Orchard-specific dataset used in this process contained 400 pairs of RGB images and LiDAR disparity maps, which were divided into training and validation subsets with a ratio of 8:2. While the RGB images remained consistent across iterations, the disparity maps were updated through the BDGF guided by the GuideNet models from previous stage, as shown in Table 1. Training hardware and environment were consistent with those described in Section 2.3.2, with a batch size of 2 for 100 epochs. Learning rate was initialized at 0.001 and reduced to 0.0001 at epoch 50. Optimization process leveraged Adaptive Moment Estimation (Adam) with β_1 of 0.9 and β_2 of 0.999.

2.4.4. Point-cloud filtering based on the DICF

The DICF consists of the BDGF and an iterative GuideNet refinement process with two training iterations. The BDGF features a two-stage point-cloud filtering pipeline. In the first stage, a Consistency Filter (CF) compared depth maps obtained from GuideNet with those projected from LiDAR according to Eq. (3). A threshold of 30 mm, determined from 100 fruit pixels across 10 stereo-LiDAR pairs, was used to

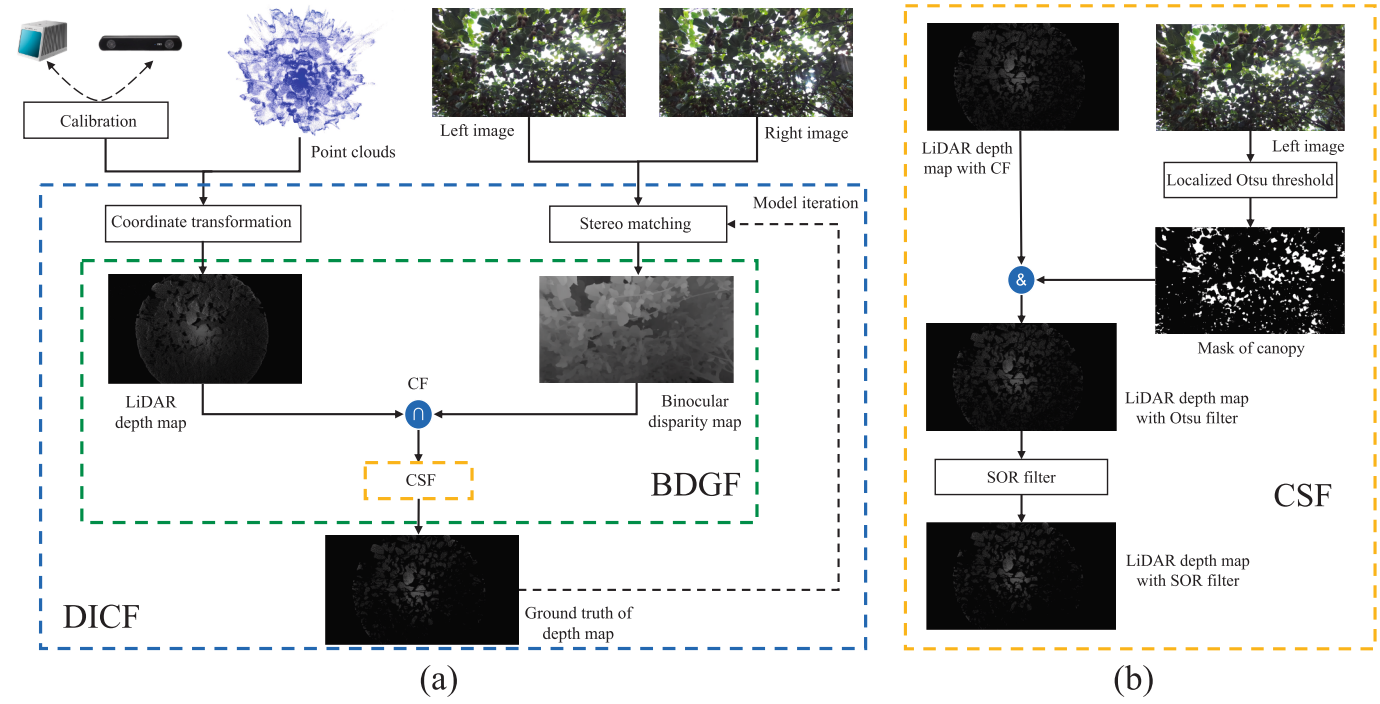


Fig. 7. Flowchart of point clouds filtering by the DICF. (a) Stereo matching iteration guided by the BDGF. LiDAR depth map is filtered using the CF and CSF, and filtered depth maps are employed as ground truth to iteratively refine GuideNet model twice, ultimately generating a model well-suited for kiwifruit orchard scenarios; (b) Flowchart of the CSF. Left image is segmented using Otsu thresholding algorithm to generate a mask, which was then combined with the CF-filtered LiDAR depth map through a logical AND operation to remove background depth values for obtaining ground-truth depth map. Subsequently, SOR algorithm is applied to produce precision-enhanced depth maps.

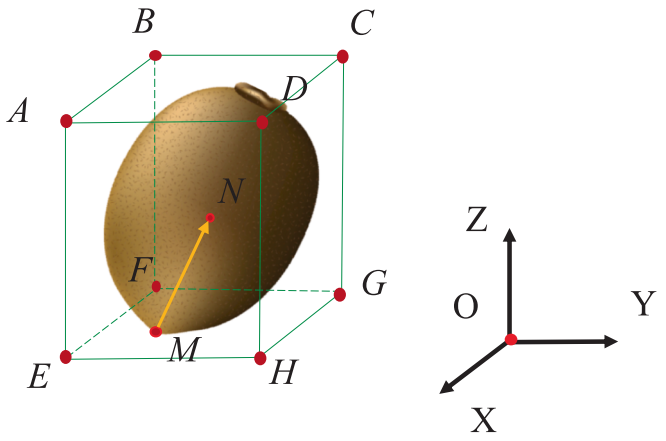


Fig. 8. Schematic illustration of kiwifruit pose estimation. Points *A* to *H* represent eight vertices of 3D bounding box surrounding the fruit. Calyx reference point *M* is determined by identifying the minimum z-coordinate position within kiwifruit point clouds. Geometric centroid *N* is derived from arithmetic mean coordinates of all bounding box vertices. Orientation vector $\overrightarrow{MN}$ extending from the calyx to the centroid, characterizes principal axis of fruit orientation.

suppress PCA. In the second stage, a Canopy-Specific Filter (CSF) further removed residual background projections by exploiting grayscale differences between canopy and sky regions in left RGB image, and the resulting mask was combined with the CF-filtered maps through logical AND operations. Filtered-out pixels were marked as invalid and excluded from subsequent network loss computation to avoid introducing incorrect supervisory signals. The refined depth map was restored to 3D using Eq. (1), and residual outliers were eliminated with SOR algorithm using a neighborhood size of 6 and a standard deviation multiplier of 0.5. Final point clouds were then re-projected to obtain a precision-enhanced depth map. Complete workflow of the DICF is illustrated in Fig. 7a, and detailed process of the CSF is shown in Fig. 7b. The BDGF-filtered depth maps were further employed for iterative training of GuideNet. The first training iteration generated orchard-specific model named GuideNet-i1, and the second training iteration refined the GuideNet-i1 to obtain GuideNet-i2. GuideNet-i2 was subsequently integrated into the BDGF to filter canopy point clouds for kiwifruit pose estimation without additional refinement. The proposed DICF was implemented utilizing Python version 3.9.19, Open3D version 0.18.0, and OpenCV version 4.10.0.84, with hardware configuration remaining identical to the setup described in Section 2.3.2.

$$Z_l \cap Z_c = \begin{cases} Z_l & |Z_l - Z_c| \leqslant \delta_d \\ 0 & |Z_l - Z_c| > \delta_d \end{cases} \tag{3}$$

Where Z_l represents depth value in projected LiDAR depth map; δ_d denotes depth threshold beyond which pixel points are filtered out.

2.5. Kiwifruit pose estimation

2.5.1. Single-view camera-LiDAR kiwifruit pose estimation

Single-view pose estimation integrated object detection with depth-based point-cloud segmentation. Object detection and depth maps were generated separately for the View-1 and View-2. LiDAR depth maps were first processed using the BDGF and the iteratively trained GuideNet-i2 to obtain refined depth maps. After detection network inference, pixel coordinates of kiwifruits in RGB images were combined with depth values of the refined depth maps to obtain fruit depth values. These depth values were then projected into LiDAR coordinate system through intrinsic and extrinsic matrices to generate point clouds. The resulting point clouds were segmented using a Region-Growing (RG) algorithm that incorporated prior knowledge of kiwifruit cultivar 'Xu Xiang', which includes a mean fruit length of 70 mm, a diameter of 50 mm, and a sphericity value of 0.80, to reduce interference from non-target canopy elements. Geometric centroid of each detection bounding box was first computed. 3D point at the depth-map pixel nearest the centroid was selected as initial RG seed, with angular and Euclidean thresholds of 100° and 30 mm, respectively. RG algorithm was then applied to these growth seed points to segment kiwifruit point-cloud clusters. The segmented point clouds were transformed into central camera coordinate system using inter-group extrinsic matrix. Spatial extrema (x_{min}, x_{max}, y_{min}, y_{max}, z_{min}, z_{max}) were extracted to construct 3D bounding boxes encompassing individual kiwifruit specimens.

Considering natural downward vertical growth characteristic of kiwifruits in the orchard, the central camera coordinate system was used to provide a horizontal reference coordinate system. Within the horizontal coordinate system, the spatial point with the minimum z-coordinate in each fruit point clouds were designated as calyx reference point *M*. Centroid *N* of 3D bounding box was determined by calculating geometric mean of its eight vertex coordinates (*A*, *B*, *C*, *D*, *E*, *F*, *G*, *H*). Orientation vector $\overrightarrow{MN}$ was then derived as principal axis extending from *M* to *N*, with calculation formulas for *M*, *N*, and $\overrightarrow{MN}$ provided in Eqs. (4)-(6), and the corresponding schematic illustrated in Fig. 8. Euler angles α, β, and γ represent orientation angles of fruit relative to X, Y, and Z axes, respectively, with their calculation formulas provided in Eq. (7). Fruits exhibiting any Euler angle exceeding 40° were categorized as obstructed targets and excluded from harvesting consideration to avoid end-effector damage.

$$M = (x_M, y_M, z_M) = \underset{(x,y,z)\in P_f}{\operatorname{argmin}} z \tag{4}$$

$$N = (x_N, y_N, z_N) = \frac{1}{8}\sum_{i=1}^{8} V_i \tag{5}$$

$$\overrightarrow{MN} = \begin{bmatrix} x_N - x_M \\ y_N - y_M \\ z_N - z_M \end{bmatrix} \tag{6}$$

Where P_f denotes set of 3D points belonging to fruit *f*; V_i represents 3D coordinates of the *i*th vertex in 3D detection bounding box for fruit *f*; (x_M, y_M, z_M) and (x_N, y_N, z_N) represent spatial coordinates of *M* and *N* in horizontal coordinate system, respectively.

$$\begin{aligned} \alpha &= \arccos\left(\frac{x_N - x_M}{\|\overrightarrow{MN}\|}\right) \\ \beta &= \arccos\left(\frac{y_N - y_M}{\|\overrightarrow{MN}\|}\right) \\ \gamma &= \arccos\left(\frac{z_N - z_M}{\|\overrightarrow{MN}\|}\right) \end{aligned} \tag{7}$$

Where $\|\overrightarrow{MN}\|$ is Euclidean length of $\overrightarrow{MN}$.

2.5.2. Fusion of dual-view pose estimation results

Dual-view camera-LiDAR fusion methodology was employed to mitigate fruit omission and structural occlusion problems in single-view perception, thereby enhancing kiwifruit detection accuracy and pose estimation precision. The View-1 and View-2 independently detected kiwifruits within its own coordinate system, generating 3D detection bounding boxes for individual fruits. During dual-view fusion, point clouds and vertices of each box obtained from the View-1 and View-2 were transformed into central horizontal camera coordinate system using inter-group extrinsic matrices. A centroid-distance-based 3D bounding box merging strategy was then adopted to determine and integrate the same kiwifruit detected from both sides. Two bounding boxes were merged when their inter-centroid distance was less than 50% of the box longest axis length. After identifying corresponding bounding boxes from both views for merging, coordinate extrema (x_{min}, x_{max}, y_{min},

Table 2
Summary of training hyper-parameters and key algorithmic parameters in the proposed dual-view camera-LiDAR fusion methodology.

Module	Stage	Parameter	Value
YOLO	Training	Learning rate	0.01
		Epoch	300
		Batch size	8
		Image size	640 × 640 pixels
		SGD momentum coefficient	0.937
		SGD weight decay	0.0005
	Inference	Image size	640 × 640 pixels
GuideNet	Training	Learning rate	0.001 → 0.0001 (epoch 50)
		Epoch	100
		Batch size	2
		Image size	2208 × 1216 pixels
		Adam β_1 coefficient	0.9
		Adam β_2 coefficient	0.999
	Inference	Image size	2208 × 1216 pixels
DICF	CF	Logical AND threshold	30 mm
	CSF	Neighborhood size	6
		Standard deviation multiplier	0.5
Pose estimation	RG	Angular threshold	100°
		Euclidean distance threshold	30 mm

Note. CF, Consistency Filter; CSF, Canopy-Specific Filter; RG, Region Growing.

y_{max}, z_{min}, z_{max}) of each box were first calculated in the central camera coordinate system. The vertices of the updated box were then obtained by aggregating these extrema from both boxes to define the smallest bounding box enclosing them. Based on these recalculated vertices, the fused calyx point M' and centroid N' of kiwifruit were determined according to Eq. (4) and Eq. (5), enabling subsequent dual-view-fused fruit pose estimation through Eq. (6) and Eq. (7). This complete pose estimation pipeline was implemented utilizing identical hardware and software environments detailed in Section 2.4.4.

Effective detection area for kiwifruit pose estimation formed a unified perceptive workspace directly above the platform after spatial fusion of the View-1 and View-2. In transverse direction, overlapping perception width reached approximately 1.5 m. Given the 3.0 m inter-row spacing, a round-trip navigation strategy, moving forward along one side of the tree row and returning along the opposite side, was required to achieve complete perceptual coverage of entire canopy. Longitudinally, this effective detection area extended approximately 5.5 m along each tree row.

2.6. Summary of parameter configurations

For clearer presentation of parameter configurations and hyper-parameter settings, Table 2 summarizes the principal parameters used in the proposed method.

2.7. Evaluation metrics

2.7.1. Detection accuracy

Performance of kiwifruit detection models trained on YOLO-series networks was evaluated using *mAP*, formally defined through Eq. (8) and Eq. (9).

$$AP = \int_0^1 P(R)dR \tag{8}$$

$$mAP = \frac{1}{|C|}\sum_{c\in C} AP_c \tag{9}$$

Where $P(R)$ denotes precision-recall curve, which is defined in Eq. (10) and Eq. (11); C represents set of object categories (in this study, single class 'kiwifruit'); AP_c indicates *AP* for category c.

$$P = \frac{TP}{TP + FP} \tag{10}$$

$$R = \frac{TP}{TP + FN} \tag{11}$$

Where *TP*, *FP* and *FN* denote three types of test results in dataset, namely true positive, false positive and false negative, respectively.

Detection Rate (*DR*) calculated through Eq. (12) was employed to calculate ratio of correctly detected kiwifruit to ground truth in harvestable area.

$$DR = \frac{n}{n_t} \times 100\% \tag{12}$$

Where n represents mean kiwifruits of method detects; n_t denotes mean kiwifruits of human counts.

2.7.2. Depth estimation accuracy

Performance of GuideNet and the DICF was assessed using End-Point Error (*EPE*) and *bad*-3 metrics. *EPE* measures mean disparity error per pixel, whereas *bad*-3 indicates percentage of pixels with disparity errors exceeding three pixels. *EPE* and *bad*-3 were calculated through Eq. (13) and Eq. (14), respectively.

$$EPE = \frac{1}{N}\sum_{(x,y)} \left| d_{pre}(x,y) - d_{gt}(x,y) \right| \tag{13}$$

$$bad-3 = \frac{1}{N}\sum_{(x,y)} \left(\left| d_{pre}(x,y) - d_{gt}(x,y) \right| > 3 \right) \times 100\% \tag{14}$$

Where N denotes total number of valid pixels; $d_{gt}(x, y)$ and $d_{pre}(x, y)$ correspond to ground truth and predicted disparity values, respectively.

2.7.3. Pose estimation accuracy

Pose estimation performance for kiwifruits was evaluated through localization accuracy and orientation accuracy metrics. Localization accuracy was defined as *MAE* between estimated calyx and its ground truth coordinates along X, Y, and Z axes in centrally camera coordinate system, calculated using Eq. (15). Orientation accuracy quantifies average angular deviation between estimated and true orientation vectors, where each orientation vector was defined by axis connecting calyx to fruit centroid, with angular error computed via Eq. (16).

$$MAE = \begin{cases} \frac{1}{N}\sum_{i=1}^{N} \left| x_{est}^{(i)} - x_{gt}^{(i)} \right| & \text{(x - axis)} \\ \frac{1}{N}\sum_{i=1}^{N} \left| y_{est}^{(i)} - y_{gt}^{(i)} \right| & \text{(y - axis)} \\ \frac{1}{N}\sum_{i=1}^{N} \left| z_{est}^{(i)} - z_{gt}^{(i)} \right| & \text{(z - axis)} \end{cases} \tag{15}$$

$$\theta_{err} = \begin{cases} \frac{1}{N}\sum_{i=1}^{N} \left| \alpha_{est}^{(i)} - \alpha_{gt}^{(i)} \right| & \text{(x - axis)} \\ \frac{1}{N}\sum_{i=1}^{N} \left| \beta_{est}^{(i)} - \beta_{gt}^{(i)} \right| & \text{(y - axis)} \\ \frac{1}{N}\sum_{i=1}^{N} \left| \gamma_{est}^{(i)} - \gamma_{gt}^{(i)} \right| & \text{(z - axis)} \end{cases} \tag{16}$$

Where N denotes total number of samples; superscript i indexes the ith sample; ($x(i)$ est, $y(i)$ est, $z(i)$ est), ($x(i)$ gt, $y(i)$ gt, $z(i)$ gt) correspond to estimated and ground-truth coordinates of calyx, respectively; ($\alpha(i)$ est, $\beta(i)$ est, $\gamma(i)$ est), ($\alpha(i)$ gt, $\beta(i)$ gt, $\gamma(i)$ gt) indicate estimated Euler angles relative to coordinate axes and their ground-truth counterparts, respectively.

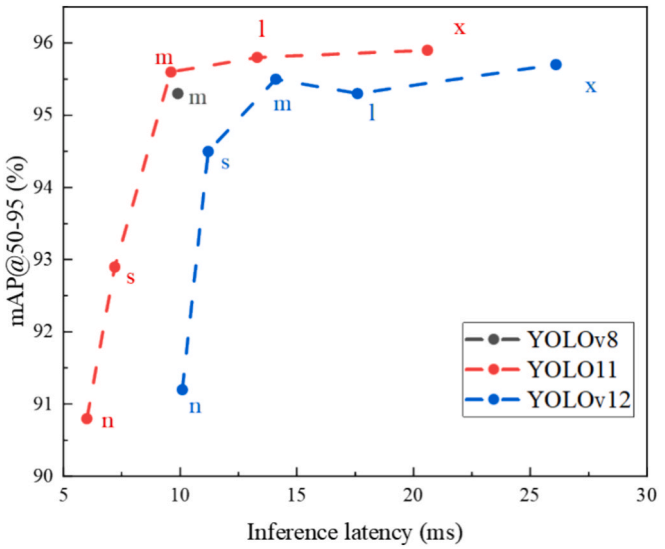


Fig. 9. Speed-accuracy relationships among YOLO-series architectures trained on OD-Dataset.

Table 3
Detection results of different YOLO models on OD-Dataset.

Models	Scale	*mAP*@50–95 (%)	*Inference latency* (ms)
YOLOv8	m (50.90 MB)	95.3	9.9
YOLO11	n (5.36 MB)	90.8	**6.0**
	s (18.73 MB)	92.9	7.2
	m (39.58 MB)	95.6	9.6
	l (50.02 MB)	95.8	13.3
	x (111.73 MB)	**95.9**	20.6
YOLOv12	n (5.38 MB)	91.2	10.1
	s (18.44 MB)	94.5	11.2
	m (39.74 MB)	95.5	14.1
	l (53.36 MB)	95.3	17.6
	x (118.85 MB)	95.7	26.1

3. Results and discussion

3.1. Performance of object detection models

Both YOLO11 and YOLOv12 demonstrated slight improvements in detecting kiwifruits compared to YOLOv8. Considering high *mAP*@50 obtained during single-class kiwifruit detection trials, more rigorous *mAP*@50–95 was employed for performance assessment. Fig. 9 illustrates speed-accuracy relationships among various YOLO-series architectures trained on OD-Dataset, with corresponding metrics presented in Table 3. Since built-in FlashAttention memory acceleration module in YOLOv12 not only substantially reduces inference latency gap between YOLOv12 and YOLO11 models but also imposes stricter hardware requirements (Dao et al., 2022), it was excluded from comparative analysis. Under identical model scales, YOLO11 exhibited superior inference speed relative to YOLOv12, as shown in Fig. 9 and Table 3, attributed to increased computational complexity caused by attention-centric modules of YOLOv12 (Tian et al., 2025a). Introduction of the attention-centric modules did not substantially enhance *mAP*@50–95 for kiwifruit detection tasks with distinctive morphological features, while YOLO11 offered greater advantages for this task. Similar observations have also been reported in other fruit-detection studies, where fruits with substantial pixel occupancy and distinctive morphological characteristics can be reliably extracted using computationally efficient architectures (Wu et al., 2025).

YOLO11m achieved satisfactory performance for detecting kiwifruits in field orchards. YOLO11n and YOLO11x achieved optimal performance in inference speed of 6.0 ms and detection accuracy of 95.9% *mAP*@50–95, respectively. Notably, while YOLO11n and YOLO11s demonstrated high inference speed, they exhibited noticeable differences in detection accuracy compared to YOLO11m, with *mAP*@50–95 values lower by 4.8% and 2.7%, respectively. Conversely, YOLO11m achieved a *mAP*@50–95 of 95.6% while demonstrating a 27.8% and 53.4% improvement in inference speed per image compared to YOLO11l and YOLO11x, respectively. These comparisons showed a speed-accuracy tradeoff, as YOLO11m avoided accuracy degradation of smaller models while maintaining substantially faster inference than larger models. Considering dual requirements of real-time performance and precision for harvesting robots (Paul et al., 2024), YOLO11m was selected as the object detection model.

3.2. Generalization performance evaluation of GuideNet

GuideNet exhibited satisfactory efficiency-accuracy balance, maintaining high precision while attaining the highest inference speed, thereby demonstrating robust generalization performance across diverse environments. Generalization capability of GuideNet was validated through comparative experiments with three established and three state-of-the-art (SOTA) stereo matching networks, including PSMNet (Chang and Chen, 2018), GwcNet (Guo et al., 2019), CFNet (Shen et al., 2021), IGEVNet (Xu et al., 2023), Stereobase (Guo et al., 2024), LightStereo (Guo et al., 2025). All networks were pre-trained on SceneFlow synthetic dataset to optimize initial parameters, using the hardware configuration described in Section 2.3.2. Without fine-tuning, all models were assessed on real-world KITTI2012 and KITTI2015 datasets for generalization evaluation. IGEVNet exhibited superior generalization capabilities, achieving SOTA accuracy with an *EPE* of 0.98 pixels and a *bad*-3 of 5.30% on KITTI2012, alongside 1.17 pixels in *EPE* and 5.35% in *bad*-3 on KITTI2015, as demonstrated in Table 4. The GuideNet-p achieved comparable accuracy with an *EPE* of 1.09 pixels and *bad*-3 error of 5.52% on KITTI2012, while matching KITTI2015 performance at 1.17 pixels in *EPE* and 5.35% in *bad*-3. Notably, it attained an inference latency of approximately 40 ms per image with resolution of 1248 × 384 pixels, which is substantially lower

Table 4
Experimental results of generalization capabilities across different models.

Research	Model	KITTI2012			KITTI2015		
		EPE (pixel)	*bad*-3 (%)	*Inference latency* (ms)	*EPE* (pixel)	*bad*-3 (%)	*Inference latency* (ms)
Chang and Chen (2018)	PSMNet	4.68	31.34	156.09	5.99	32.44	140.98
Guo et al. (2019)	GwcNet	2.76	23.75	63.85	3.58	25.46	61.10
Shen et al. (2021)	CFNet	2.27	14.29	63.26	2.89	12.35	60.16
Xu et al. (2023)	IGEVNet	**0.98**	**5.30**	3935.54	**1.17**	**5.35**	3885.30
Guo et al. (2024)	Stereobase	1.02	5.47	3493.76	1.20	5.71	3203.99
Guo et al. (2025)	LightStereo-S	1.87	12.26	85.37	1.96	9.29	81.61
Guo et al. (2025)	LightStereo-M	1.23	7.56	124.85	1.40	6.80	124.73
Guo et al. (2025)	LightStereo-L	1.23	7.43	150.40	1.28	6.63	155.21
Ours	**GuideNet-p**	1.09	5.52	**42.71**	1.23	5.87	**44.52**

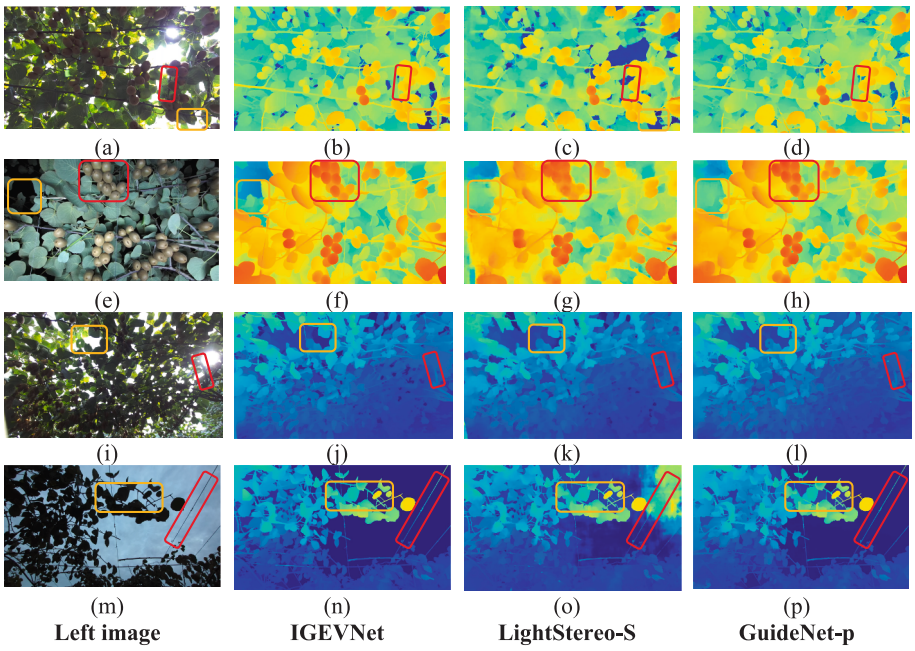


Fig. 10. Comparative performance analysis of three stereo matching networks in multi-scenario kiwifruit orchard. Left image depicts RGB image acquired by left lens of ZED 2i camera, while remaining panels illustrate disparity pseudo-color maps produced through stereo matching network in (a) front-lit daylight observation condition, (e) supplemental lighting-assisted nighttime scenario, (i) oblique daylight observation condition, and (m) low-light scenario. Red and yellow boxes highlight regions with substantial feature disparities and background disparity errors, respectively.

Table 5
Filtering strategies for different datasets.

Filtering strategies Datasets	CF	CSF	Iterative refinement
KS-Dataset	✓	✓	✓
NF-Dataset	×	×	×
GF-Dataset	✓	✓	×
CF-Dataset	✓	×	×
CSF-Dataset	×	✓	×
CFI-Dataset	✓	×	✓

than that of other network models. The comparative analysis demonstrated that GuideNet achieved substantial improvements in inference latency compared to most stereo matching networks while maintaining generalization performance.

The GuideNet-p trained by SceneFlow dataset maintained robust generalization performance across various kiwifruit orchard scenarios. The generalization performance of the GuideNet-p was validated through comparative experiments with IGEVNet, selected for its SOTA accuracy in Table 4, and LightStereo-S, chosen for its low inference latency, across four representative orchard scenarios featuring different lighting conditions and observation angles, as shown in Fig. 10. Red and yellow boxes in Fig. 10, highlight regions with substantial feature disparities and background disparity errors, respectively. Strong feature extraction capability of the GuideNet-p in kiwifruit orchard scenarios was demonstrated by correctly matching thin branch structures even when they were partially truncated by fruits or leaves, as illustrated by the regions marked by red boxes in Fig. 10b-d. Despite fluctuations in illumination critically impacting stereo matching localization, the GuideNet-p maintained robust performance under both overexposure conditions and low-light scenarios, as evidenced by its effective matching of fine branches and wires in the red-boxed regions of Fig. 10j-l and Fig. 10n-p.

The GuideNet-p exhibited sufficiently reliable stereo matching performance to serve as effective guidance for the DICF. Its performance was demonstrated in kiwifruit boundary segmentation, as evidenced by clear separation of individual fruit boundaries within fruit clusters shown by the red boxes in Fig. 10f-h. The segmentation enabled binocular depth to reliably filter PCA along fruit boundaries. However, regions highlighted by yellow boxes in Fig. 10 revealed inconsistent disparity in low-texture backgrounds such as sky. These inconsistencies arise primarily from two factors. First, lack of distinctive texture itself creates ambiguous correspondences. Second, strong cost aggregation and large effective receptive field result in over-smoothing of disparity near depth discontinuities, which amplifies matching errors at sharp transitions between the fruit and sky (Khamis et al., 2018). Since GuideNet is used solely to guide point-cloud filtering, such errors in background areas do not affect kiwifruit pose estimation.

3.3. Component-wise performance validation of the DICF

Effectiveness of each module in the DICF was systematically validated through ablation experiments of six distinct datasets, each dataset constructed with a different filtering strategy to generate varied disparity maps. Every dataset contained 400 sets of left–right image pairs, and corresponding filtered disparity maps. Kiwi-Stereo Dataset (KS-Dataset) was constructed through comprehensive implementation of the DICF featuring the BDGF and two-iteration refinement. None-Filtered dataset (NF-Dataset) comprised disparity maps obtained through direct conversion of raw point cloud data without noise suppression mechanisms. Component contributions were examined by constructing four partially processed dataset variants. Consistency-Filtered Dataset (CF-Dataset) and Canopy-Specific Filtered Dataset (CSF-Dataset), each applied only their respective filtering strategies, the CF and CSF, without iterative refinement. Consistency-Filtered Iterative Dataset (CFI-Dataset) combined the CF with two-iteration optimization but did not apply the CSF. Guided-Filtered Dataset (GF-Dataset) implemented the BDGF without iterative refinement. The filtering strategies employed for the six datasets are shown in Table 5. Classical network PSMNet was also introduced for ablation experiments to prevent self-evaluation bias, since the DICF employs GuideNet, with the same training equipment and environment as in Section 2.3.2. Effect of each part in the DICF on preserving true depth was further verified by

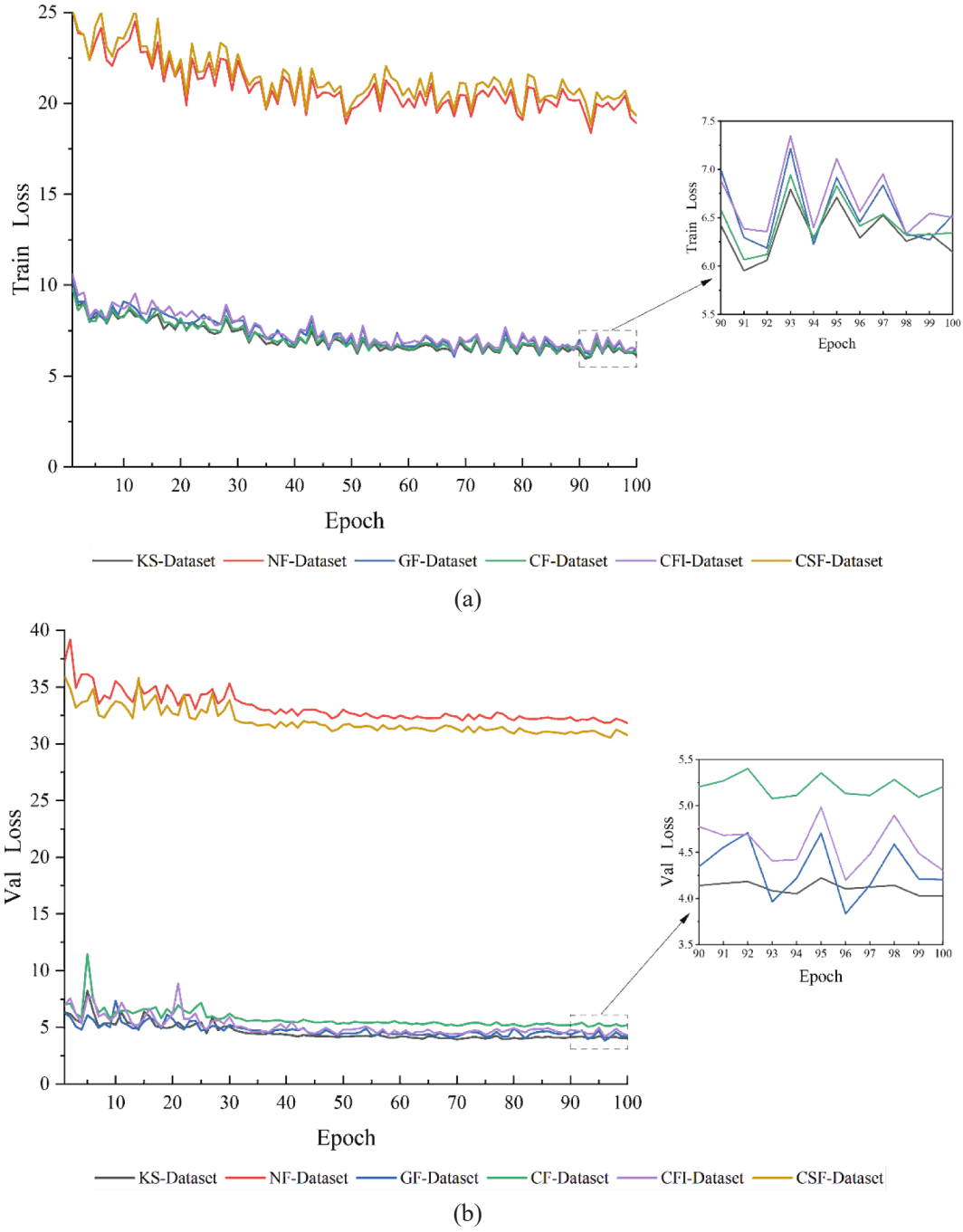


Fig. 11. Training performance of the PSMNet model across six datasets is shown by (a) training loss curves for distinct training subsets and (b) validation loss curves during model optimization.

Table 6
PSMNet performance on different datasets.

Evaluation metrics Datasets	*EPE* (pixel)	*bad*-3 (%)
KS-Dataset	**2.97**	**4.02**
NF-Dataset	9.64	31.83
GF-Dataset	3.01	4.20
CF-Dataset	3.49	5.21
CSF-Dataset	8.32	30.76
CFI-Dataset	3.09	4.30

comparing convergence of PSMNet training on different datasets, whose training and validation set loss curves are shown in Fig. 11.

KS-Dataset achieved superior convergence performance during network validation, with the lowest *EPE* and *bad*-3 among all datasets. These results indicated that LiDAR point clouds denoising, PCA removal, and redundant disparity value elimination substantially improved point-cloud depth accuracy and disparity map quality. Conversely, NF-Dataset and CSF-Dataset produced the highest *EPE* and *bad*-3, as shown in Fig. 11b and Table 6, primarily attributed to unprocessed LiDAR point clouds containing substantial noise and PCA that introduced substantial errors in disparity ground truth and critically impeded network optimization. Although CF-Dataset exhibited moderate improvement over the NF-Dataset and the CSF-Dataset, three to five seconds LiDAR depth

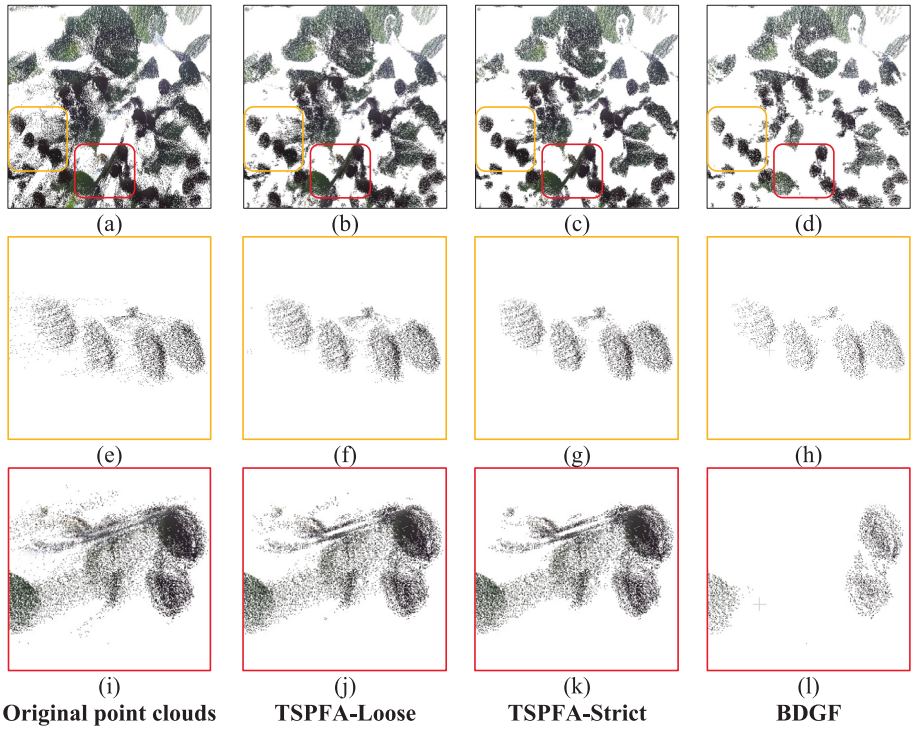


Fig. 12. Comparison of filtering performance among different algorithms within the same kiwifruit orchard environment. SOR parameters were set to a neighborhood of 6 and a standard deviation multiplier of 0.5 for TSPFA-Loose, a neighborhood of 30 and a standard deviation multiplier of 0.3 for TSPFA-Strict, a neighborhood of 6 and a standard deviation multiplier of 0.5 for the DICF. Yellow and red boxes highlight key kiwifruit point-cloud comparison areas, respectively.

Table 7
Computational performance comparison between TSPFA and the BDGF.

Point cloud filtering approaches	Processing time (s)
TSPFA-Loose	2.34
TSPFA-Strict	3.56
BDGF	**1.12**

integration time led to overestimated fruit regions due to wind-driven canopy movement, exceeding actual fruit geometries (Yun et al., 2016). The geometric discrepancy caused spatial misalignment between disparity maps and corresponding left image regions, ultimately inducing feature extraction inaccuracies in end-to-end stereo matching architectures. Improved *EPE* and *bad*-3 results in CFI-Dataset relative to the CF-Dataset demonstrated that iterative refinement strategies effectively mitigated LiDAR noise impact on disparity maps (Yang et al., 2019), producing stereo matching models with improved alignment to kiwifruit outdoor orchards. Notably, enhanced performance of both GF-Dataset and KS-Dataset relative to NF-Dataset indicated that the BDGF effectively mitigated PCA, delivering more accurate and stable point clouds.

3.4. Evaluation of filtering performance of the BDGF

The BDGF demonstrated outstanding filtering performance in kiwifruit orchards, effectively suppressing PCA even in high-density point-cloud regions containing closely spaced objects. SOR-based TSPFA (Jing et al., 2024) was first configured with commonly adopted initial

Table 8
Fruit detection accuracy from different views, where the specific viewing configurations representing View-1, View-2, and the fusion view are explicitly illustrated in Fig. 13.

Different views	*DR* (%)
View-1	61.8
View-2	66.8
Fusion view	**92.8**

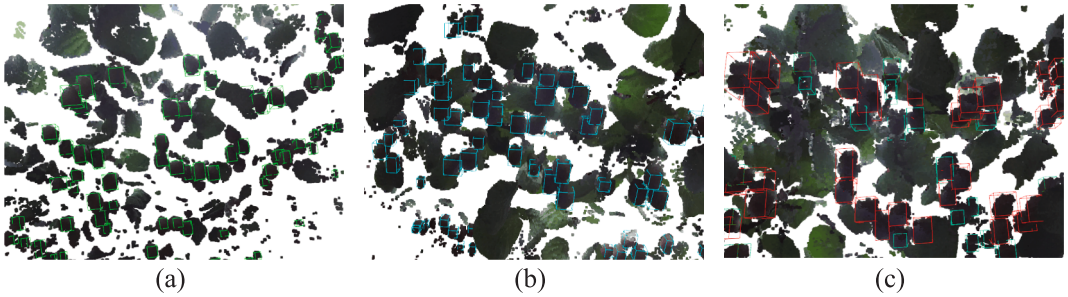


Fig. 13. Visualization of kiwifruit identification at 1 m beneath canopy from (a) View-1, (b) View-2 and (c) fusion view. 3D bounding boxes indicate detected kiwifruits, where green boxes represent detections from the View-1, blue boxes indicate detections from the View-2, and red boxes denote fruits detected through fusion of the View-1 and View-2.

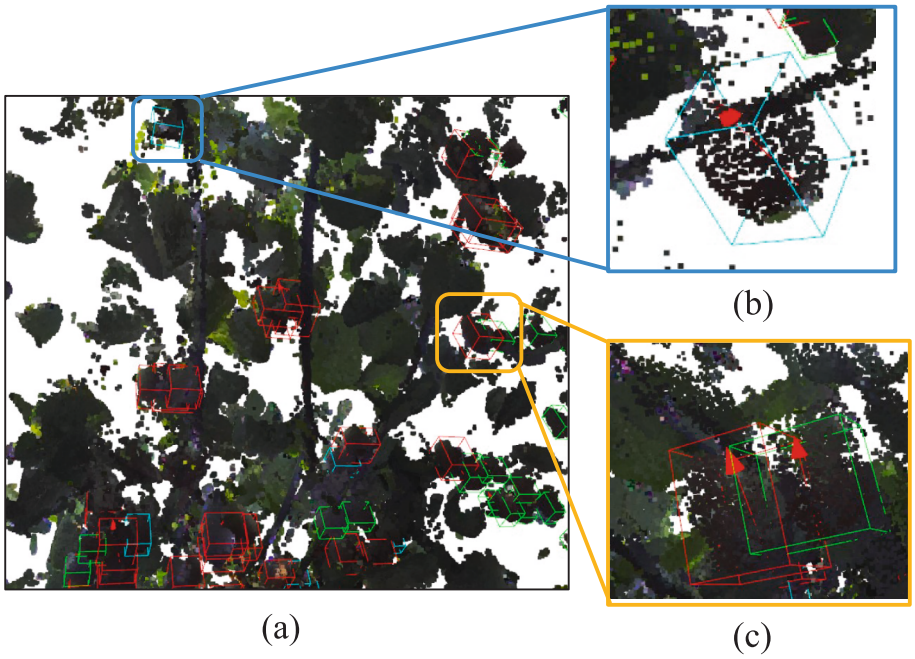


Fig. 14. Visualization of kiwifruit pose estimation within dual-view sensing coverage with (a) global and (b), (c) close-range results.

Table 9
Kiwifruit pose estimation accuracy at different time periods.

Time periods	Axis	*MAE* (mm)	θ_{err} (°)
9 to 11 AM	X	7.26	8.29
	Y	7.87	9.67
	Z	5.78	7.67
12 to 3 PM	X	7.56	6.74
	Y	7.64	8.61
	Z	4.61	8.34
4 to 7 PM	X	4.63	5.02
	Y	6.52	9.95
	Z	6.97	5.02

parameters, a neighborhood of 6 and a standard deviation multiplier of 0.5 for TSPFA-Loose, which determine local neighborhood size and filtering strictness, respectively. For comparison, TSPFA was further employed with a neighborhood of 30 and a standard deviation multiplier of 0.3 for TSPFA-Strict, parameters empirically determined to be more effective for suppressing PCA than TSPFA-Loose. Then these two variants were compared with the BDGF, as shown in Fig. 12. In yellow-highlighted areas of Fig. 12, where fruits were sparsely distributed, all approaches were able to reduce PCA between kiwifruit clusters while preserving fruit geometry, with stricter SOR parameters producing more effective suppression. In contrast, in red-highlighted high-density regions of Fig. 12, TSPFA failed to remove PCA due to close spatial spacing between kiwifruits and leaves, whereas the BDGF was still able to eliminate PCA. The limitation of TSPFA arises from its reliance solely on point-cloud structural information, while the BDGF incorporates binocular-vision depth values to guide edge-aware filtering, thereby achieving more reliable PCA suppression in dense orchard scenarios.

The BDGF achieved notably high computational efficiency in point-cloud filtering within kiwifruit field environments. Compared with TSPFA-Loose, the BDGF reduced processing time by 52.1% under a point-cloud size of 700,000, as shown in Table 7. The improvement was attributed to the utilization of binocular vision in the BDGF, which reduced point-cloud dimensionality by projecting data onto 2D images. Through the BDGF, redundant PCA was effectively removed during down-sampling, substantially decreasing overall point cloud scale and consequently reducing processing time in the CSF stage (Song et al., 2025).

3.5. Performance of kiwifruit pose estimation

Dual-view detection, which reduced kiwifruit occlusion in single-view perception, served as a prerequisite for accurate pose estimation. Visualizations of single-view and dual-view detections from approximately 1 m beneath the canopy are demonstrated in Fig. 13, wherein bounding boxes represent detected kiwifruits, with different colors corresponding to detections from distinct sensor view combinations. As shown in Fig. 13c, fruits visible from both views were enclosed in red boxes, while fruits visible only from View-1 and View-2 were marked in green and blue boxes, respectively. Low performance of single-view detection primarily resulted from inclined imaging method adopted to

Table 10
Results from previous studies on fruit pose estimation.

Research	Localization method	Object	Environment	Localization error (mm)	Orientation error (°)
Jang and Hwang (2024)	RGB-D	Tomato	Indoor	–	6.79
Jiang et al. (2025)	RGB-D	Apple	Orchard	*MAE* = 14	–
Zhao et al. (2025)	RGB-D	Cucumber	Orchard	*MAE* = 6.18	6.25
Jing et al. (2024)	Binocular	Kiwifruit	Orchard	*MAE* = 4	–
Kang et al. (2022)	Camera-LiDAR	Apple	Orchard	*SD* = 2.56	–
Ours	Camera-LiDAR + the DICF	Kiwifruit	Orchard	*MAE* = 6.54	7.70

Note. –, not specified; *MAE*, Mean Absolute Error; *SD*, Standard Deviation.

maximize kiwifruit surface geometry completeness. Severe occlusions that substantially hindered fruit detection was produced by clustered growth characteristics of kiwifruit and overlapping canopies (Tian et al., 2025b). By detecting fruits from two views with large angular differences, the dual-view detection effectively mitigated the occlusion that occurs when the scene was observed from a single view. Efficacy of pose estimation is contingent on accuracy of detection and pose estimation. Detection accuracy within pickable area was evaluated using 20 randomly sampled sets. The *DR* of dual-view fusion reached 92.8%, which corresponds to an absolute increase of 28.5% compared with the average *DR* of 64.3% obtained from single-view detection, as shown in Table 8.

Proposed dual-view camera-LiDAR fusion-based kiwifruit poses estimation methodology, which integrated YOLO11m and the DICF, achieved satisfactory accuracy and robustness. Visualization of kiwifruit pose estimation is shown in Fig. 14. The methodology was employed at different times of day to estimate kiwifruit poses. Experimental results demonstrated that the methodology maintained reliable accuracy under varying lighting conditions. *MAE* in X, Y, and Z axes, as well as θ_{err}, were within 8 mm and 10°, respectively, as demonstrated in Table 9. *MAE* variations across different times were mainly due to changes in lighting conditions, which degraded image quality and consequently affected pose estimation accuracy.

Additionally, the proposed methodology was evaluated against existing fruit pose estimation methods within agricultural scenarios, achieving competitive performance. As shown in Table 10, RGB-D camera-based methods obtained acceptable results indoors. For example, Jang and Hwang (2024) achieved an orientation estimation error of 6.79° for tomatoes. However, this depth-sensing principle exhibits high sensitivity to uneven illumination, which compromises accuracy and robustness under field conditions (Jiang et al., 2025). Similar research achieved a *MAE* of 6.18 mm and an orientation error of 6.25° for cucumbers under field conditions, but remains fundamentally limited when applied to spherical fruits such as kiwifruits (Zhao et al., 2025). Binocular cameras provide relatively accurate depth estimation in the field but struggle to reconstruct texture-less fruit surfaces due to disparity-based computational constraints, which severely restricts pose estimation (Jing et al., 2024). In contrast, camera-LiDAR fusion methods capture detailed fruit surface geometry while demonstrating superior robustness under field illumination, as evidenced by Kang et al. (2022), who achieved a *SD* of 2.56 mm in localizing apples across distances from 0.5 m to 1.8 m under varying illumination. However, in outdoor orchards, PCA limits the reliability of fruit pose estimation. Proposed binocular-depth-guided camera-LiDAR pose estimation methodology effectively suppresses PCA within kiwifruit orchards, achieving reliable and robust kiwifruit pose estimation with a *MAE* of 6.54 mm and an orientation error of 7.70°.

There were various reasons for kiwifruit detection and pose estimation errors. Firstly, inherent complexity of kiwifruit orchard canopies and clustered fruit morphology inevitably resulted in some fruits remaining undetectable even with dual-view detection. Additionally, fusing camera-LiDAR data and integrating multi-view observations relied on calibrating intrinsic and extrinsic matrices, where inevitable calibration inaccuracies propagated into localization errors. Furthermore, defining the lowest point of fruit point clouds as calyx reference introduced systematic deviations when central camera coordinate systems were not perfectly horizontal due to uneven orchard terrain, resulting in discrepancies between the estimated and actual calyx positions. These factors collectively led to discrepancies between predicted and actual fruit poses.

4. Conclusions

This study introduces a binocular-depth-guided point-cloud filtering methodology for camera-LiDAR fusion, enabling reliable fruit pose estimation in kiwifruit orchards. The fusion ensures robust depth perception under varying lighting conditions. The proposed DICF effectively alleviates PCA within kiwifruit canopy. Building on the DICF, a dual-view stereo matching dataset (KS-Dataset) was constructed through two iterations, achieving an *EPE* of 2.94 pixels on PSMNet. The BDGF substantially reduces computational costs by projecting 3D point clouds onto 2D images for dimensionality reduction and down sampling during filtering, which enables a processing time of only 1.12 s for a point-cloud size of 700,000 points. In field settings, the dual-view detection substantially enhances detection accuracy of occluded fruits, attaining a *DR* of 92.8%. The assumption that vector from the lowest to centroid point in fruit point-cloud cluster ensures accurate kiwifruit pose estimation. An *MAE* of 6.54 mm was achieved for calyx localization, and the orientation angle error was 7.70°, meeting accuracy requirements for robotic kiwifruit harvesting. This study provides a novel methodology for kiwifruit pose estimation in orchards, delivering precise and robust results for kiwifruit harvesting. Subsequently, future research will concentrate on end-to-end camera-LiDAR fusion methods, alongside exploring fine-grained morphological analysis to differentiate defective kiwifruits based on surface shape, thereby enhancing applicability in autonomous harvesting robots.

Declaration of generative AI and AI-assisted technologies in the writing process.

During the preparation of this work, the authors used ChatGPT for grammar checking to improve the readability of article. After using these servers, the authors reviewed and edited the content as needed and takes full responsibility for the content of the published article.

CRediT authorship contribution statement

Shiao Niu: Writing – original draft, Methodology, Investigation, Data curation. **Leilei He:** Writing – review & editing, Methodology, Data curation, Conceptualization. **Qilin Gong:** Writing – review & editing, Methodology, Conceptualization. **Qichen Huang:** Writing – review & editing, Methodology. **Ruichao Luo:** Writing – review & editing, Methodology, Conceptualization. **Rui Li:** Writing – review & editing, Methodology. **Bao Zhou:** Writing – review & editing, Methodology. **Xiancun Zhou:** Methodology, Conceptualization. **Zhenchao Wu:** Methodology, Conceptualization. **Jinyong Chen:** Methodology, Conceptualization. **Xiaoxu Sun:** Methodology, Conceptualization. **Anastasia Grecheneva:** Writing – review & editing, Methodology, Conceptualization. **Shen Hin Lim:** Writing – review & editing, Methodology, Conceptualization. **Longsheng Fu:** Writing – review & editing, Supervision, Methodology, Conceptualization.

Declaration of competing interest

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

Acknowledgements

This work was partially supported by National Natural Science Foundation of China (32501786, 32572215); Chinese Universities Scientific Fund (2452026021); National Foreign Expert Project, Ministry of Human Resources and Social Security, China (H20240238, Y20240046); Key Research and Development Program of Shaanxi, China (2024NC-YBXM-195); Central Public-interest Scientific Institution Basal Research Fund, China (1610192023105); National Key Research and Development Program, China (2022YFD1600700); Jiangxi Provincial Key Laboratory of Modern Agricultural Equipment (Grant Number 20242BCC32127).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.

org/10.1016/j.compag.2026.111939.

Data availability

Data is available at https://github.com/fu3lab/Kiwifruit_Camera-LiDAR

References

Chang, J.-R., Chen, Y.-S., 2018. Pyramid Stereo Matching Network. In: In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5410–5418. https://doi.org/10.48550/arXiv.1803.08669.

Chen, J., Han, B., Liu, D., 2024. 3D Lidar Technology for Slag Flushing of BF in Auto Crane. In: In: 2024 8th International Conference on Robotics and Automation Sciences (ICRAS), pp. 101–105. https://doi.org/10.1109/ICRAS62427.2024.10654486.

Dao, T., Fu, D.Y., Ermon, S., Rudra, A., Ré, C., 2022. FlashAttention: Fast and Memory-Efficient exact attention with IO-Awareness. Advances in Neural Information Processing Systems (NeurIPS). 35, 16344–16359. https://doi.org/10.48550/arXiv.2205.14135.

Elhousni, M., Huang, X., 2020. A Survey on 3D LiDAR Localization for Autonomous Vehicles. In: In: 2020 IEEE Intelligent Vehicles Symposium (IV), pp. 1879–1884. https://doi.org/10.48550/arXiv.2006.00648.

Fang, W., Wu, Z., Li, W., Sun, X., Mao, W., Li, R., Majeed, Y., Fu, L., 2023. Fruit detachment force of multiple varieties kiwifruit with different fruit-stem angles for designing universal robotic picking end-effector. Comput. Electron. Agric. 213, 108225. https://doi.org/10.1016/j.compag.2023.108225.

Goberville, N., El-Yabroudi, M., Omwanas, M., Rojas, J., Meyer, R., Asher, Z., Abdel-Qader, I., 2020. Analysis of LiDAR and Camera Data in Real-World Weather Conditions for Autonomous Vehicle Operations. SAE Int. J. Adv. Curr. Pract. Mobil. 2, 2428–2434. https://doi.org/10.4271/2020-01-0093.

Guo, X., Yang, K., Yang, W., Wang, X., Li, H., 2019. Group-Wise Correlation Stereo Network. In: In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3273–3282. https://doi.org/10.1109/CVPR.2019.00339.

Guo, X., Zhang, C., Lu, J., Duan, Y., Wang, Y., Yang, T., Zhu, Z., Chen, L., 2024. OpenStereo: A Comprehensive Benchmark for Stereo Matching and Strong Baseline. arXiv preprint, arXiv: 2312.00343. Doi: 10.48550/arXiv.2312.00343.

Guo, X., Zhang, C., Zhang, Y., Zheng, W., Nie, D., Poggi, M., Chen, L., 2025. LightStereo: Channel boost is all you need for Efficient 2D cost Aggregation. In: In: 2025 IEEE International Conference on Robotics and Automation (ICRA), pp. 8738–8744. https://doi.org/10.48550/arXiv.2406.19833.

Jang, M., Hwang, Y., 2024. Tomato pose estimation using the association of tomato body and sepal. Comput. Electron. Agric. 221, 108961. https://doi.org/10.1016/j.compag.2024.108961.

Jiang, Z., Zhou, J., Han, H., Wang, Y., 2025. An apple fruit localization system based on accurate and flexible hand-eye pose acquisition for robotic harvesting. Comput. Electron. Agric. 229, 109705. https://doi.org/10.1016/j.compag.2024.109705.

Jing, X., Jiang, H., Niu, S., Zhang, H., Gilbert Murengami, B., Wu, Z., Li, R., Zhou, C., Ye, H., Chen, J., Majeed, Y., Fu, L., 2024. End-to-end stereo matching network with two-stage partition filtering for full-resolution depth estimation and precise localization of kiwifruit for robotic harvesting. Comput. Electron. Agric. 225, 109333. https://doi.org/10.1016/j.compag.2024.109333.

Jing, X., 2025. Dual-arm kiwifruit harvesting robot based on binocular inter-row perception and optimized stereoscopic operation. Master's Thesis, Northwest a&f University, Shaanxi, China. https://doi.org/10.27409/d.cnki.gxbnu.2025.002120.

Kang, H., Wang, X., 2023. Semantic segmentation of fruits on multi-sensor fused data in natural orchards. Comput. Electron. Agric. 204, 107569. https://doi.org/10.1016/j.compag.2022.107569.

Kang, H., Wang, X., Chen, C., 2022. Accurate fruit localisation using high resolution LiDAR-camera fusion and instance segmentation. Comput. Electron. Agric. 203, 107450. https://doi.org/10.1016/j.compag.2022.107450.

Karim, M.R., Reza, M.N., Jin, H., Haque, M.A., Lee, K.-H., Sung, J., Chung, S.-O., 2024. Application of LiDAR Sensors for Crop and Working Environment Recognition in Agriculture: a Review. Remote Sens. 16, 4623. https://doi.org/10.3390/rs16244623.

Khamis, S., Fanello, S., Rhemann, C., Kowdle, A., Valentin, J., Izadi, S., 2018. StereoNet: Guided Hierarchical Refinement for Real-Time Edge-Aware Depth Prediction. In: In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 573–590. https://doi.org/10.1007/978-3-030-01267-0_35.

Khanam, R., Hussain, M., 2024. YOLOv11: An Overview of the Key Architectural Enhancements. arXiv preprint, arXiv: 2410.17725. Doi: 10.48550/arXiv.2410.17725.

Liu, X., Jing, X., Jiang, H., Younas, S., Wei, R., Dang, H., Wu, Z., Fu, L., 2024. Performance evaluation of newly released cameras for fruit detection and localization in complex kiwifruit orchard environments. J. Field Robot. 41, 881–894. https://doi.org/10.1002/rob.22297.

Luo, L., Yin, W., Ning, Z., Wang, J., Wei, H., Chen, W., Lu, Q., 2022. In-field pose estimation of grape clusters with combined point cloud segmentation and geometric analysis. Comput. Electron. Agric. 200, 107197. https://doi.org/10.1016/j.compag.2022.107197.

Paul, A., Machavaram, R., Ambuj, K., Nagar, D., 2024. Smart solutions for capsicum Harvesting: Unleashing the power of YOLO for Detection, Segmentation, growth stage Classification, counting, and real-time mobile identification. Comput. Electron. Agric. 219, 108832. https://doi.org/10.1016/j.compag.2024.108832.

Pin, J., Guo, T., Xv, M., Zou, X., Hu, W., 2025. Fast extraction of navigation line and crop position based on LiDAR for cabbage crops. Artif. Intell. Agric. 15, 686–695. https://doi.org/10.1016/j.aiia.2025.03.007.

Rusu, R.B., Cousins, S., 2011. 3D is here: Point Cloud Library (PCL). In: In: IEEE International Conference on Robotics and Automation (ICRA), pp. 1–4. https://doi.org/10.1109/ICRA.2011.5980567.

Shen, Z., Dai, Y., Rao, Z., 2021. CFNet: Cascade and Fused cost volume for Robust Stereo Matching. In: In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 13906–13915. https://doi.org/10.48550/arXiv.2104.04314.

Song, Z., Yang, L., Xu, S., Liu, L., Xu, D., Jia, C., Jia, F., Wang, L., 2025. GraphBEV: Towards Robust BEV Feature Alignment for Multi-modal 3D Object Detection. In: In: European Conference on Computer Vision (ECCV), pp. 347–366. https://doi.org/10.1007/978-3-031-73347-5_20.

Tian, Y., Ye, Q., Doermann, D., 2025a. YOLOv12: Attention-Centric Real-Time Object Detectors. arXiv preprint, arXiv: 2502.12524. Doi: 10.48550/arXiv.2502.12524.

Tian, Z., Xue, X., Yang, Y., Ma, W., 2025b. Spatial clustering analysis and group-based path planning optimization for robotic kiwifruit harvesting. Smart Agric. Technol. 12, 101508. https://doi.org/10.1016/j.atech.2025.101508.

Wang, P., 2021. Research on Comparison of LiDAR and Camera in Autonomous Driving. J. Phys. Conf. Ser. 2093, 012032. https://doi.org/10.1088/1742-6596/2093/1/012032.

Wu, N., Wu, J., Wang, Z., Zhao, Y., Xu, X., Wang, Y., Skobelev, P., Mi, Y., 2025. Maturity detection and counting of blueberries in real orchards using a 1novel STF-YOLO model integrated with ByteTrack algorithm. Front. Plant Sci. 16, 1682024. https://doi.org/10.3389/fpls.2025.1682024.

Xu, G., Wang, X., Ding, X., Yang, X., 2023. Iterative Geometry Encoding volume for Stereo Matching. In: In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 21919–21928. https://doi.org/10.1109/CVPR52729.2023.02099.

Yang, G., Song, X., Huang, C., Deng, Z., Shi, J., Zhou, B., 2019. DrivingStereo: a Large-Scale Dataset for Stereo Matching in Autonomous Driving Scenarios. In: In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 899–908. https://doi.org/10.1109/CVPR.2019.00099.

Yun, T., An, F., Li, W., Sun, Y., Cao, L., Xue, L., 2016. A Novel Approach for Retrieving tree Leaf Area from Ground-based LiDAR. Remote Sens. 8, 942. https://doi.org/10.3390/rs8110942.

Zhang, Q., Chen, Z., Zhou, Z., Wang, L., Liao, Q., Yang, C., Yang, J., 2024. 3D terrestrial LiDAR for obtaining phenotypic information of cigar tobacco plants. Comput. Electron. Agric. 226, 109424. https://doi.org/10.1016/j.compag.2024.109424.

Zhao, D., Pan, R., Wang, Y., Yue, Y., Li, J., Dong, X., Shan, W., 2024. Research on kiwifruit recognition based on improved YOLOv8. In: In: Proceedings of the 3rd International Conference on Computer, Artificial Intelligence and Control Engineering (CAICE), pp. 698–704. https://doi.org/10.1145/3672758.3672875.

Zhao, G., Dong, S., Wen, J., Ban, Y., Zhang, X., 2025. Selective fruit harvesting prediction and 6D pose estimation based on YOLOv7 multi-parameter recognition. Comput. Electron. Agric. 229, 109815. https://doi.org/10.1016/j.compag.2024.109815.