Computers and Electronics in Agriculture 225 (2024) 109333

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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

End-to-end stereo matching network with two-stage partition filtering for full-resolution depth estimation and precise localization of kiwifruit for robotic harvesting

Xudong Jing [a], Hanhui Jiang [a], Shiao Niu [a], Haosen Zhang [a], Bryan Gilbert Murengami [a], Zhenchao Wu [a], Rui Li [a], Chengquan Zhou [b], Hongbao Ye [b,*], Jinyong Chen [c,*], Yaqoob Majeed [g], Longsheng Fu [a,d,e,f,*]

[a] College of Mechanical and Electronic Engineering, Northwest A&F University, Yangling, Shaanxi 712100, China
[b] Key Laboratory of Agricultural Equipment for Hilly and Mountainous Areas in Southeastern China (Co-construction by Ministry and Province), Ministry of Agriculture and Rural Affairs, Hangzhou, Zhejiang 310021, China
[c] Zhengzhou Fruit Research Institute, Chinese Academy of Agricultural Sciences, Zhengzhou, Henan 450009, China
[d] Key Laboratory of Agricultural Internet of Things, Ministry of Agriculture and Rural Affairs, Yangling, Shaanxi 712100, China
[e] Shaanxi Key Laboratory of Agricultural Information Perception and Intelligent Service, Yangling, Shaanxi 712100, China
[f] Northwest A&F University Shenzhen Research Institute, Shenzhen, Guangdong 518000, China
[g] Department of Biological & Agricultural Engineering, Texas A&M University, Dallas, Texas 75252, United States

ARTICLE INFO

Keywords:
Two-stage partition filtering
Kiwifruit localization
LaC-Gwc Net
YOLOv8
Robotic harvesting

ABSTRACT

Full-resolution depth estimation within operational space of robotic arms and accurate localization of kiwifruits is very important for automated harvesting. Depth estimation is expected to be accurate and full-resolution while current depth estimation methods are susceptible to depth missing due to occlusion and uneven illumination. And depth estimation mostly focuses on fruit localization, while obstacles such as branches and wires, which can affect harvesting strategy, have not been considered. This paper localized kiwifruits based on bounding boxes output by YOLOv8m and full-resolution depth from an end-to-end stereo matching network, i.e., LaC-Gwc Net, which was trained after generating a stereo matching dataset by proposing a two-stage partition filtering algorithm. Results showed that LaC-Gwc Net achieved an end-point error (*EPE*) of 3.8 pixels, which means that accurate depth estimation can also be achieved for thin obstacles such as the branches and the wires. Additionally, YOLOv8m obtained acceptable results in detecting kiwifruits and their calyxes, reaching mean average precision (*mAP*) of 93.1% and detection speed of 7.0 ms. The methodology obtained only kiwifruit localization error of 4.0 mm on the Z-axis, which meets requirements of robotic harvesting. Furthermore, this study considered the localization of obstacles in kiwifruit orchards, providing high-precision full-resolution depth estimation for agricultural harvesting robots.

1. Introduction

Global kiwifruit planting area exceeded 2.8×10^5 ha, with production of 4.5×10^6 tons in 2022 (UN Food & Agriculture Organization, 2024). The manual harvesting of kiwifruit accounts for more than 25% of the annual production cost and is labor-intensive (Suo et al., 2021). Therefore, kiwifruit harvesting robots are crucial for addressing labor shortages and improving the efficiency of fruit harvesting. Robotic systems for kiwifruit harvesting represent a significant advancement in agricultural automation, addressing labor shortages and improving efficiency in fruit harvesting. Advanced technologies, such as artificial intelligence and automation, was embedded into robotic systems for harvesting in complex orchards. The software components of the robotic systems are primarily composed of perception, planning, and control. Perception serves as the initial and critical step in robotic harvesting, directly influencing the precision and efficiency of fruit harvesting (Yang et al., 2023). Therefore, it is very important to develop high-precision depth estimation methods for harvesting robots, which can provide the accurate location of target objects.

In orchards, accurate depth estimation has been deemed crucial to

* Corresponding authors.
E-mail addresses: yhb2008@zaas.ac.cn (H. Ye), chenjinyong@caas.cn (J. Chen), fulsh@nwafu.edu.cn (L. Fu).

https://doi.org/10.1016/j.compag.2024.109333
Received 3 February 2024; Received in revised form 18 June 2024; Accepted 8 August 2024
Available online 20 August 2024

providing harvesting robots with precise localization of target objects. Depth is typically acquired using RGB-D cameras, including structured light, Time-of-Flight (ToF), and stereo vision. Some studies have employed structured light cameras or ToF cameras for fruit localization (Au et al., 2023b; Lin et al., 2019; Wang et al., 2022; Xiong et al., 2020). These cameras measure depth by emitting a specific light source and capturing its reflection, but they are prone to interference from extrinsic light sources, which can result in depth missing (Popovic et al., 2021). Moreover, the mainstream consumer-grade cameras of the aforementioned types usually have a small field of view (FoV), limiting operational space (Li et al., 2023c). Other researches demonstrated potential of binocular cameras in fruit localization in the complex orchard (Hsieh et al., 2021; Mejia et al., 2023). Binocular cameras estimate depth by performing stereo matching and calculating disparity between the left and right images. They can easily achieve a large FoV, enabling them to cover extensive workspaces, which is crucial for the estimation of harvesting robots.

Despite the advancements in stereo vision for localization, challenges such as occlusion and uneven illumination in orchards still limit their accuracy and reliability. Ling et al. (2019) applied binocular camera to locate tomatoes, which reported localization errors of 5.0–10.0 mm. Gao et al. (2021) employed a Speed Up Robust Features matching algorithm to obtain three-dimensional (3D) coordinates of dragon fruit, with an average localization error of 6 mm. Hou et al. (2022) improved depth estimation for oranges by extracting their complete shape, which achieved localization accuracy of 4.0 mm. Additionally, Williams et al. (2019) and Au et al. (2023a) demonstrated the use of binocular camera on a kiwifruit harvesting robot and the practice of systematic calibration of binocular camera for kiwifruit localization, respectively, with the maximum error of localization reaching up to 8.2 mm. While these studies achieved favorable results in fruit localization, most of them focused solely on fruit localization and did not consider potential obstacles within the operational area. Additionally, these studies were largely affected by occlusion and uneven illumination, resulting in stereo matching failures and significant depth estimation errors (Mirbod et al., 2023).

Deep learning-based stereo matching is a potential solution for these challenges, offering enhanced accuracy in depth estimation by learning from a plethora of complex visual data. Li et al. (2023b) applied template matching with parallel polar line constraint to apple images detected by Faster R-CNN to fulfill binocular localization, which achieved an average localization error of 5.1 mm. Zhao et al. (2023) utilized YOLOv5s and template matching for apple binocular localization, which reported mean absolute percentage errors of 6.0% to 13.7%. Gao et al. (2024a) employed YOLOv5x for detecting kiwifruits and performed stereo matching to calculate their 3D coordinates, achieving localization precision of 4.8 mm. These methods reached high-precision fruit localization through binocular depth estimation by deep learning-based object detection or segmentation. However, they lack the capability to locate other objects in orchards, especially obstacles. Therefore, full-resolution depth is required, which includes not only accurately fruit localization but also perceiving and locating obstacles to effectively harvest fruits.

End-to-end stereo matching networks can achieve high-precision full-resolution depth estimation, thereby improving operational efficiency of harvesting robots in kiwifruit orchards. Because high-precision full-resolution depth can accurately locate fruits and their surrounding obstacles, it can also effectively reduce depth missing and errors which enables accurate fruit localization by harvesting robots and effective avoidance or handling of potential obstacles. It requires only left and right images from binocular camera to directly generate disparity map, to further calculating depth map. It can effectively handle depth errors or missing caused by changes in occlusion and uneven illumination through deep learning (Poggi et al., 2022). The precision of the aforementioned algorithms is experiencing consistent enhancement. Current mainstream algorithms, such as PSM-Net and Gwc-Net, achieved high accuracy in public datasets, with average error of only 0.6 pixels (Chang and Chen, 2018; Guo et al., 2019; Laga et al., 2022). This demonstrates that these algorithms can not only perceive full-resolution depth but also have significantly improved localization accuracy. Consequently, depth errors and missing caused by occlusion and uneven illumination in kiwifruit orchards are likely to be addressed. End-to-end stereo matching network hold the potential to provide high-precision full-resolution depth, offering essential support for the operation of kiwifruit harvesting robots.

This study presents a novel full-resolution depth estimation methodology tailored in kiwifruit orchard for harvesting robots. The overall work of this paper mainly includes dataset acquisition, the construction of datasets for end-to-end stereo matching network and object detection, and the localization of kiwifruit by combining stereo matching and object detection, as shown in Fig. 1. Left and right images were captured by ZED 2i binocular camera in the modern kiwifruit orchards. A novel methodology for depth estimation with end-to-end stereo matching network is implemented. This involves inputting the left and right images of the ZED 2i binocular camera into end-to-end stereo matching network to estimate disparity, calculating the full-resolution depth of the left image captured by the ZED 2i binocular camera mounted on kiwifruit harvesting robot. Kiwifruits in the left image were detected and located with YOLOv8m (Jocher et al., 2023). The final step integrates the results of kiwifruit detection in the left image using YOLOv8m with the aligned left lens depth estimated by LaC-Gwc Net to obtain the localization of 3D coordinates of kiwifruit.

2. Materials and methods

This study aims to develop a methodology for detecting and locating kiwifruit for robotic harvesting in orchards. The pipeline for kiwifruit detection and localization is shown in Fig. 2. Initially, in orchards, left and right images were captured by ZED 2i binocular camera mounted on kiwifruit harvesting robot. Hereinafter, unless specifically stated otherwise, the terms “left image” and “right image” refer to left and right images captured by ZED 2i binocular camera. Left and right images are then input into an end-to-end stereo matching network, which provides a disparity map aligned with the left image. Hereinafter, unless specifically noted, the term “disparity map” refers to disparity map aligned with left image. Concurrently, the left image is processed through object detection network to produce bounding boxes for kiwifruit and calyx, which combined with the disparity map to determine 3D coordinates of kiwifruit and calyx.

2.1. Data acquisition and coordinates measuring

Three-dimensional coordinates of kiwifruit in a modern orchard acquired through an end-to-end stereo matching network. Data acquisition process is shown in Fig. 3. The stereo matching dataset and object detection dataset for kiwifruit were collected at two locations in Shaanxi Province of China in September 2023, i.e. Baiheng Organic Kiwifruit Orchard in Xianyang (108°09′ E, 34°38′ N) and Kiwifruit Experimental Orchard of Northwest A&F University in Baoji (107°39′ E, 34°27′ N). The planting pattern for kiwifruit, where the fruit grows vertically downwards, as shown in Fig. 4. The canopy height is approximately 1.7 m and the plant spacing is about 3.0 m.

ZED 2i binocular camera (Stereolabs, San Francisco, USA) and Livox Mid70 LiDAR (Livox, ShenZhen, China) with the random distance error of 2 cm were mounted on a self-designed vehicle’s sensor bracket to collect data, as shown in Fig. 5. The relative positions of the binocular camera and the LiDAR remained constant. Sensors were placed about one meter below kiwifruit canopy, collecting data vertically upwards and tilted at 30 degrees. Without artificial lighting at different times, the binocular camera and the LiDAR captured left and right images (in ‘png’ format) and point clouds (in ‘lvx’ format), totaling 400 groups.

A total of 55 groups of calibration board data at various distances and

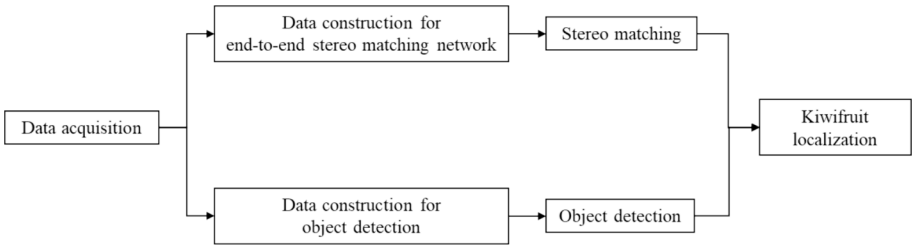

Fig. 1. Overall work of end-to-end stereo matching network with two-stage partition filtering for full-resolution depth estimation and localization of kiwifruit.

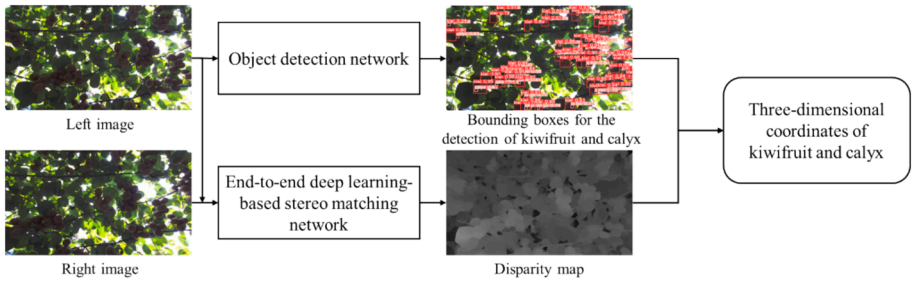

Fig. 2. Pipeline of kiwifruit localization based on end-to-end stereo matching network and object detection network. Insert the left and right images into an end-to-end stereo matching network to obtain a disparity map. Then, use object detection network to obtain bounding boxes of kiwifruit and calyx in the left image. Subsequently, combine the disparity map and bounding boxes to calculate the 3D coordinates of kiwifruit and calyx.

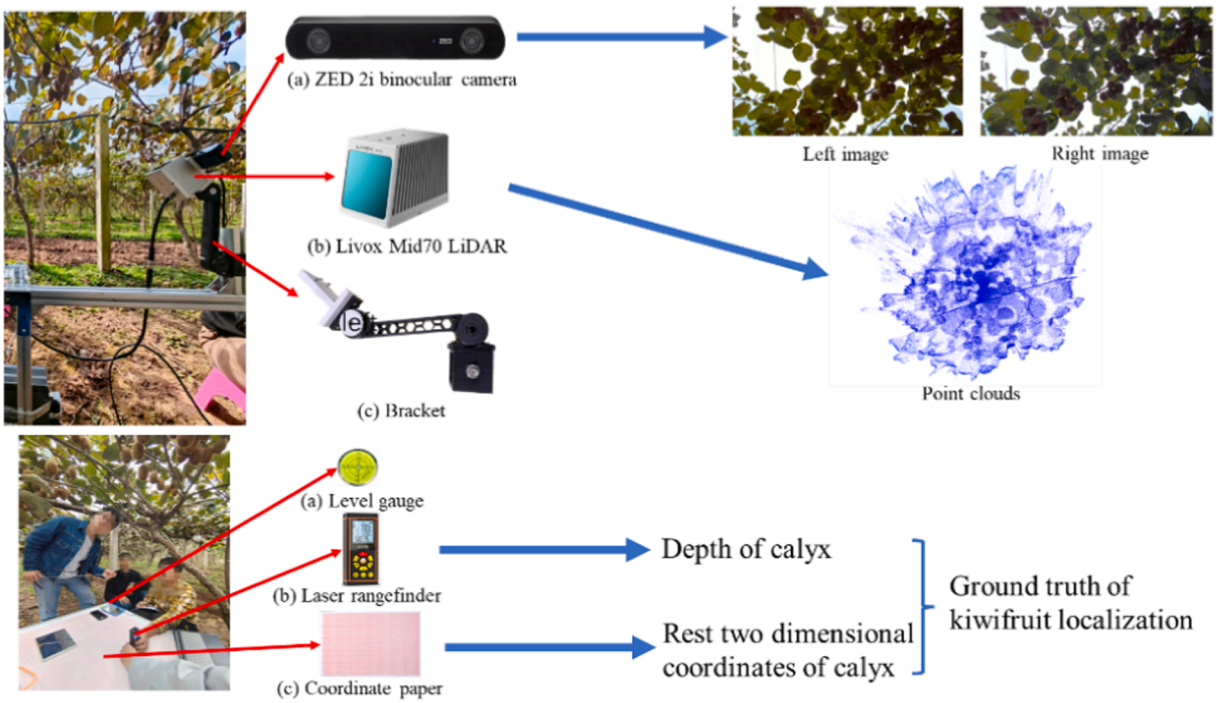

Fig. 3. Flowchart of data acquisition. Left and right images were captured using ZED2i binocular camera, and point cloud was obtained using LiDAR. The stereo camera and LiDAR were mounted on bracket, serving as the data foundation for end-to-end stereo matching network and object detection. The ground truth for kiwifruit localization was obtained by measuring depth with laser rangefinder and the rest two coordinates with coordinate paper.

angles were collected indoors to calibrate intrinsic parameters and positional relationship of the binocular camera and the LiDAR, as shown in Fig. 6. Stereo Camera Calibrator toolbox in MATLAB 2022b was utilized to calibrate the binocular camera, obtaining the intrinsic parameters K_1 and K_2 of the left and right lenses of the binocular camera. Additionally, extrinsic parameters of the binocular camera were obtained, which include rotation matrix R_{stereo}, translation vector t_{stereo}, and baseline B. Intrinsic parameter matrix K_i, consisting of focal length (f_x, f_y), principal point (C_x, C_y), is shown in Eq. (1). Livox_camera_lidar_calibration tool (https://github.com/Livox-SDK/livox_camera_lidar_calibration) was applied for joint calibration of left lens of the binocular camera and the LiDAR to obtain their extrinsic parameters. Subsequently, through manual adjustment, precise extrinsic parameters of the LiDAR and left lens of the binocular camera were obtained. These extrinsic parameters include rotation matrix R_{L2C} and translation vector t_{L2C}.

Fig. 4. The planting pattern for kiwifruit, where the fruit grows vertically downwards. The canopy height is approximately 1.7 m and the plant spacing is about 3.0 m.

$$K_i = \begin{bmatrix} f_x & 0 & C_x \\ 0 & f_y & C_y \\ 0 & 0 & 0 \end{bmatrix} \quad (1)$$

Where i equal to 1 represents the left lens of the binocular camera, while 2 represents the right lens of the binocular camera.

Finally, in the two orchards, an experimental platform with a laser rangefinder (VCHON H-40, JinYun, China, the random distance error of 1 mm), coordinate paper and four level gauges was designed to measure 3D coordinates of calyx, which was regarded as ground truth of corresponding kiwifruit localization, as shown in Fig. 7. The platform was about one meter below kiwifruit canopy and was precisely leveled using the four level gauges before calyx measurement. Ten kiwifruits were randomly selected to measure 3D coordinates of their calyx. For each calyx, the laser rangefinder acquired its depth and moved on the coordinate paper to obtain its rest two dimensional coordinates. Lastly, the platform was moved to another random area to repeat these steps. In each orchard, five areas were selected, and coordinates of 100 calyxes were recorded.

2.2. Stereo matching

An end-to-end stereo matching network was chosen to acquire a full-resolution disparity map for kiwifruit robotic harvesting. When using the end-to-end stereo matching network, only left and right images are needed as input to get the disparity map. However, during training, high-precision disparity map is required as ground truth to supervise the training of end-to-end stereo matching network.

2.2.1. Selection of end-to-end stereo matching network

LaC-Gwc Net (Liu et al., 2022) that employed Local Similarity Pattern extraction to paired feature was adopted for stereo matching in this study. Additionally, in disparity refinement stage, a Cost Self-Reassembling disparity refinement method was used to handle areas difficult for traditional stereo matching, such as disparity discontinuities and occluded regions.

2.2.2. Dataset construction by two-stage partition filtering algorithm

In dataset of LaC-Gwc Net, each group of data includes the left and right images following epipolar constraint, along with high-precision disparity map. Therefore, this study proposed a Two-Stage Partition Filtering Algorithm (TSPFA) to ensure accuracy of disparity map. Dataset construction process is shown in Fig. 8. Depth obtained from the LiDAR was applied to generate high-precision disparity map as ground truth. However, the LiDAR can exhibit noise or tailing phenomena (Putra et al., 2023), and extremely complex boundaries of objects in kiwifruit orchards further exacerbate these issues. Therefore, filtering point cloud data obtained from the LiDAR is necessary to ensure accuracy of the depth as much as possible.

The TSPFA was designated to eliminate noise and tailing effects to obtain high-precision disparity map as accurately as possible. The first stage involves filtering point clouds before coordinate transformation, while the second stage involves filtering image after generating the disparity map after coordinate transformation.

The LiDAR applied a non-repetitive scanning mode, resulting in inhomogeneous distribution of point cloud, which in turn increases the difficulty of identifying and processing noise and outliers. When employing Statistical Outlier Removal filter to eliminate outliers and noise, the uniform neighborhood size and standard deviation parameters fails to effectively remove all noise and outliers. Therefore, based on the density characteristics of the point cloud, the point cloud was

Fig. 5. Self-designed data collection vehicle equipped with the binocular camera and the LiDAR. In it, ZED 2i binocular camera and Livox Mid70 LiDAR were mounted on the same sensor bracket to ensure their constant relative position. (a) ZED 2i binocular camera; (b) Livox Mid70 LiDAR; (c) bracket that allows for the adjustment of the sensor's orientation within a range of 0 to 180 degrees.

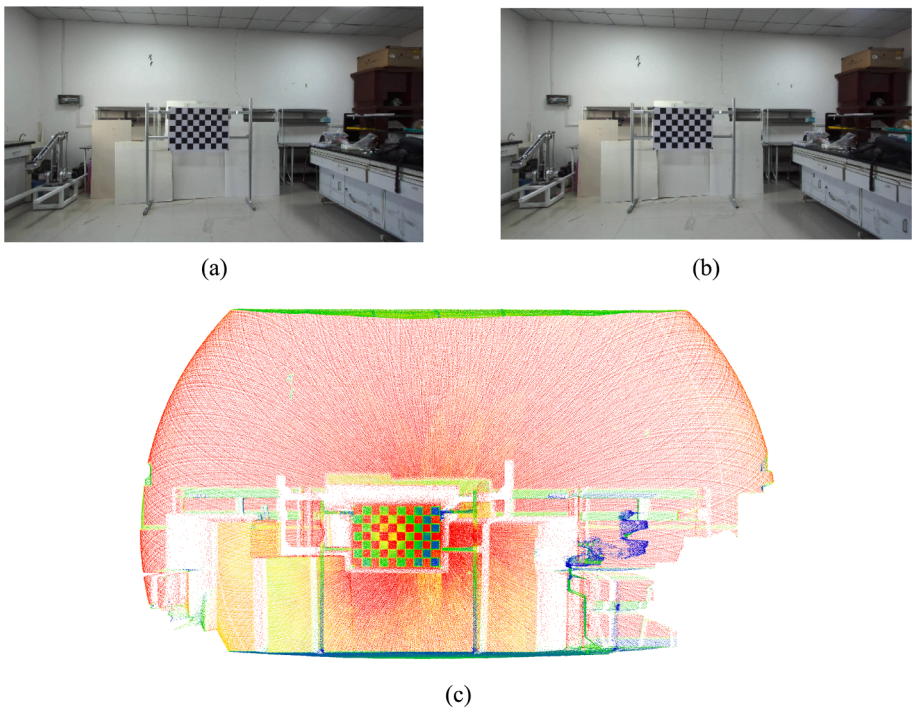

Fig. 6. Examples of calibration board data captured by ZED 2i binocular camera and Livox Mid70 LiDAR. (a) Image of the calibration board taken by the left lens of the binocular camera; (b) Image of the calibration board taken by the right lens of the binocular camera; (c) Point cloud of the calibration board captured by the LiDAR.

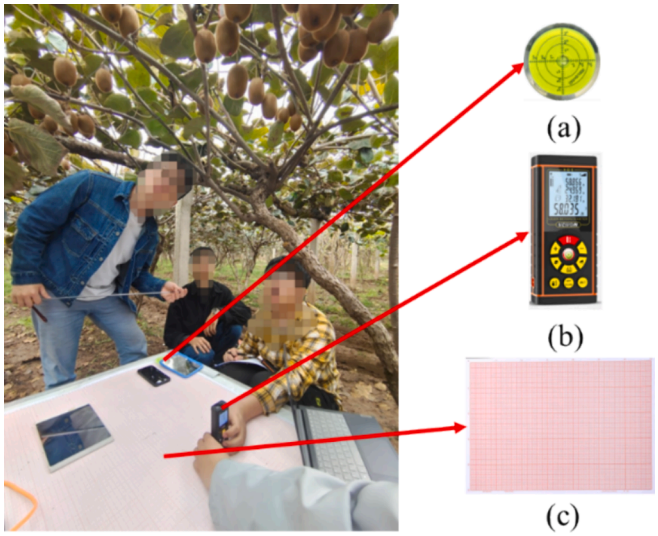

Fig. 7. A experimental platform to measure 3D coordinates of calyx as ground truth of corresponding kiwifruit localization. (a) Level gauge; (b) Laser rangefinder; (c) coordinate paper.

divided into a central region (within 0.3 m of the LiDAR) and an edge region due to the dense centers and sparse edges. For the edge region, SOR filter was applied twice with neighborhood size of 6 and standard deviation multiplier of 0.5. In the central area, SOR filter was applied three times with neighborhood size of 50 and standard deviation multiplier of 0.5 (Yun et al., 2021). The first stage of this filtering method can eliminate most of tailing phenomena. The next step is to convert depth from the LiDAR into disparity map.

The LiDAR's world coordinates were transformed into the depth map in pixels coordinate system to obtain the disparity map. Intrinsic parameters of the left lens of the binocular camera, along with extrinsic parameters of the LiDAR and left lens of the binocular camera, are utilized for this coordinate system transformation of Eq. (2). This provides the depth map of left image, which is then relied to calculate the disparity map using Eq. (3) with the extrinsic parameters of the binocular camera.

$$z_c \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = K_1 [R_{L2C} | t_{L2C}] \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix} \tag{2}$$

Where X_w, Y_w, and Z_w are 3D coordinates in the world coordinate system, i.e., point cloud coordinates obtained by the LiDAR, with units in mm. And u, v are corresponding points in pixel coordinate system (u, v) with units in pixels, and z_c is depth value in camera coordinate system, with units in mm.

$$d_c = \frac{f \times B}{z_c} \tag{3}$$

Where z_c represents depth with units in mm, B represents baseline of the binocular camera with units in mm, f represents focal length of the binocular camera's left lens with units in pixels, and d_c represents the disparity value with units in pixels.

The generated disparity maps were secondly filtered to further eliminate the noise caused by the trailing phenomenon. The binocular camera captures left and right images from below, focusing only on the kiwifruit canopy and sky. Based on this characteristic, where only the kiwifruit canopy should have point cloud, the disparity map is filtered to further eliminate the effect of tailing. Sky and kiwifruit canopy show clear distinction in grayscale image, but due to complex lighting conditions in the orchards, using single threshold is inadequate for effectively separating sky from canopy. Therefore, local Otsu threshold method (Zhuang et al., 2018) was chosen to segment kiwifruit canopy in left image, creating mask of the canopy, which was applied to remove values from the disparity map that do not belong to the canopy, ensuring

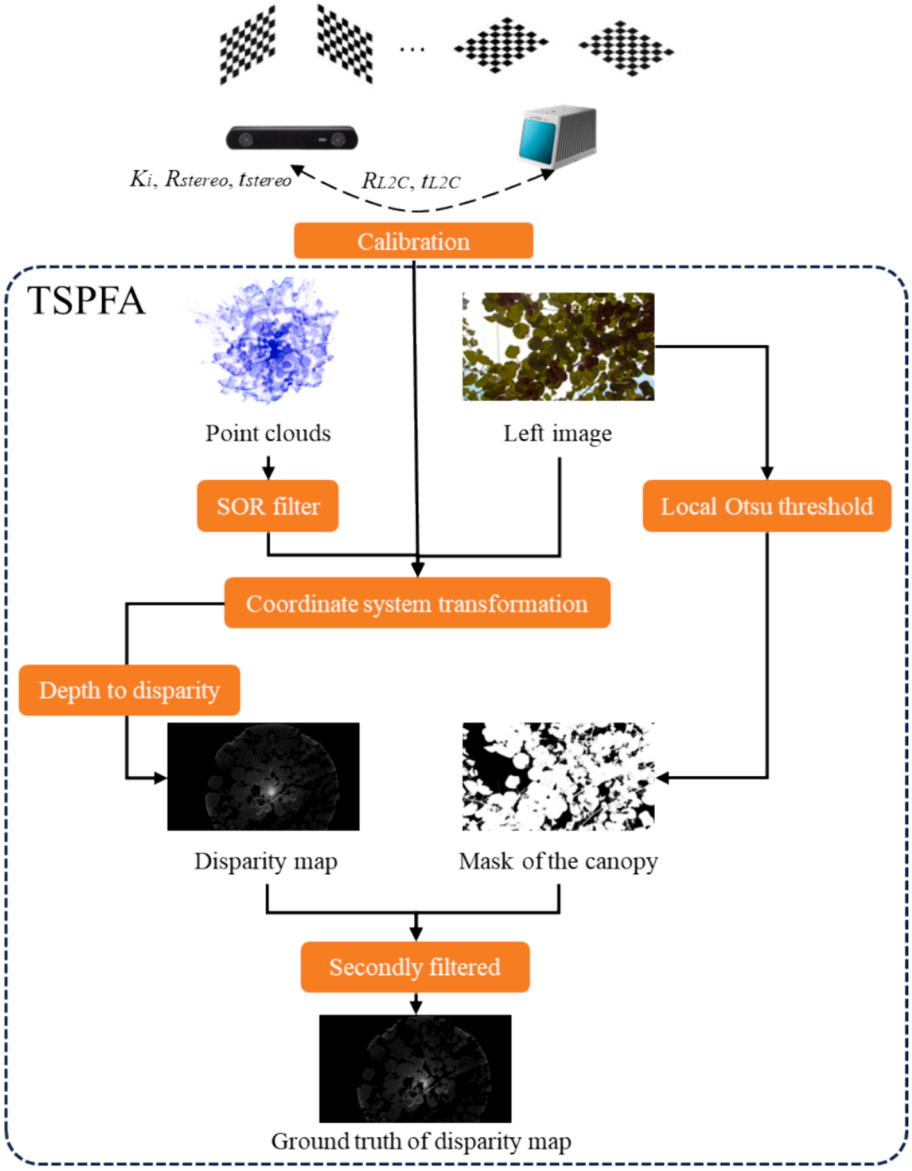

Fig. 8. Flowchart of dataset construction by Two-Stage Partition Filtering Algorithm (TSPFA). Point cloud undergoes first stage filtering using Statistical Outlier Removal (SOR) filtering. Then, based on the calibration results of binocular camera, point cloud filtered in the first stage is converted into a depth map through coordinate transformation. Subsequently, it is transformed into disparity map using extrinsic parameters of binocular camera. Mask generated from left image is then used to remove excess disparity values in the generated disparity map, resulting in the ground truth of disparity map.

reliability of ground truth.

Through the aforementioned the TSPFA, Dataset-KS was generated for training LaC-Gwc Net. It has 400 groups data which included the left and right images, and disparity map (as shown in Fig. 9). What's more, three other stereo matching datasets were produced to verify the effectiveness of the TSPFA: one without filtering (Dataset-NF), one with only the first stage of filtering (Dataset-FF), and one with only the second stage of filtering (Dataset-SF). Details of the end-to-end stereo matching datasets are shown in Table 1. These datasets were divided into training, validation, and test sets in a ratio of 7:2:1.

2.2.3. *Training of LaC-Gwc Net*

The model was trained on the PyTorch framework, version 2.0.1, on a desktop computer equipped with Intel Xeon Gold 5128 32-core

Fig. 9. An example of one group data in the Dataset-KS. (a) Left image; (b) Right image; (c) Disparity map.

Table 1
Details of the end-to-end stereo matching datasets.

Datasets	Filtering stages included	Resolution (pixels)	Number of groups
Dataset-KS	Two-stage filtering	1280 × 720	400
Dataset-NF	No filtering		
Dataset-FF	First stage filtering only		
Dataset-SF	Second stage filtering only		

processor (2.3 GHz), dual Nvidia GeForce GTX 4090 24G GPUs, and 128 GB of memory, running on a 64-bit Windows 10 system. The software toolkit includes CUDA 12.1, cuDNN 8.8.1, Python 3.9, and OpenCV 4.8. The end-to-end stereo matching network input size during training was 512 × 256, with a batch size of 8. The learning rate and the number of iterations were set to 0.001, and the number of iterations was set to 300 for the training process.

2.3. Object detection

You Only Look Once version 8 (YOLOv8) series demonstrates significant potential in object detection. Within the YOLOv8 series, an appropriate object detection model was selected for identifying kiwifruit and its calyx to optimize the picking operation. Consequently, a comparative analysis was conducted on four variants: You Only Look Once version 8 nano (YOLOv8n), You Only Look Once version 8 small (YOLOv8s), You Only Look Once version 8 medium (YOLOv8m), and You Only Look Once version 8 large (YOLOv8l).

2.3.1. Dataset construction for object detection

Detecting kiwifruits and their calyxes is a crucial step before locating them. Kiwifruits hang from vines, and kiwifruit harvesting robots usually operate beneath the canopy, with the binocular camera capturing kiwifruits from below. From this angle, calyxes (usually located at the center of the bottom of kiwifruits) can be easily observed. Therefore, bounding boxes of the calyxes of kiwifruit can be applied for more precise localization of kiwifruits. However, due to factors such as occlusion, the calyxes may not always be fully captured in FoV. In such cases, it's necessary to also label the kiwifruits itself, so that the bounding box for the kiwifruits can be utilized for localization when the calyxes were not visible.

Training YOLOv8 series requires images and their corresponding labels which can share the same dataset. The labels were manually annotated using LabelImg. In the images, the kiwifruits and calyxes were labeled as "kiwi" and "calyx", respectively, with bounding boxes, as shown in the Fig. 10. Labels were saved in 'xml' format and subsequently converted to 'txt' format for YOLOv8 series. A dataset named Dataset-KD was created by labelling 800 images of the kiwifruits and calyxes for their detection which was then divided into training, validation, and test sets in a ratio of 7:2:1.

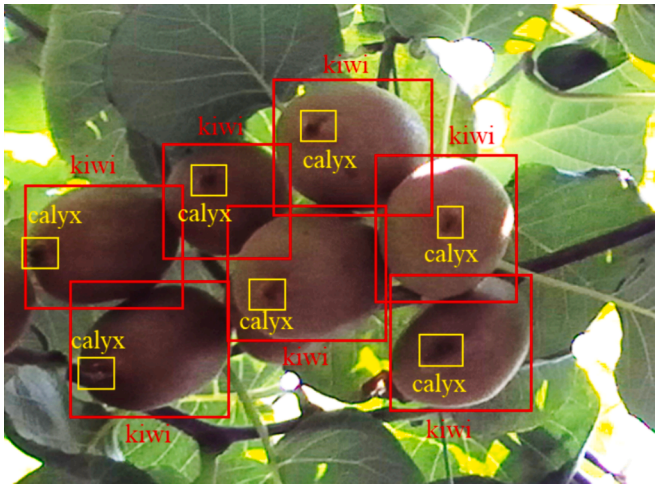

Fig. 10. Labeling examples of kiwifruit and calyx, where kiwifruit is labeled as "kiwi" using red boxes, and calyx is labeled as "calyx" with yellow boxes.

2.3.2. Network training for YOLOv8 series

The training environment of YOLOv8 series is maintained consistently. YOLOv8 series were selected for training the detection of the kiwifruits and calyxes on the PyTorch framework. YOLOv8 series input size was 640 × 640 pixels, with a batch size of 16. Stochastic gradient descent was used for training, with a momentum of 0.937 and a weight decay of 0.0005. The initial learning rate for YOLOv8 series was set at 0.01. The number of iterations was set to 350 for the training process.

2.4. Fruit localization

2.4.1. Depth map acquisition

By inputting left and right images (resolution of 1280 × 720 pixels) into LaC-Gwc Net, the disparity map can be obtained. The disparity map was converted to the depth map using Eq. (4).

$$z_{\mathrm{pre}} = \frac{f \times B}{d_{\mathrm{pre}}} \tag{4}$$

Where z_{pre} represents the predicted depth value, and d_{pre} represents the disparity value predicted by LaC-Gwc Net.

2.4.2. Calculation of calyx 3D coordinates

Firstly, a model exhibiting a balanced performance in terms of detection accuracy and speed from the YOLOv8 series has been selected for obtaining the bounding boxes of calyx. And then center of bounding boxes were calculated as two-dimensional coordinates of calyx in pixels coordinate system. Using these two-dimensional coordinates, depth was indexed in depth map, obtaining the (u, v) coordinates' depth z_{pre} in the pixel coordinate system. Then, using Eqs. (5), (6) and (7) these pixel coordinates can be converted into world coordinates to obtain 3D coordinates of calyx.

$$X_{\mathrm{w-pre}} = \frac{(u - C_{\mathrm{x}}) \times z_{\mathrm{pre}}}{f_{\mathrm{x}}} \tag{5}$$

$$Y_{\mathrm{w-pre}} = \frac{(v - C_{\mathrm{y}}) \times z_{\mathrm{pre}}}{f_{\mathrm{y}}} \tag{6}$$

$$Z_{\mathrm{w-pre}} = z_{\mathrm{pre}} \tag{7}$$

Where $X_{\mathrm{w\text{-}pre}}$, $Y_{\mathrm{w\text{-}pre}}$, and $Z_{\mathrm{w\text{-}pre}}$ represent the predicted 3D coordinates of the kiwifruit in the world coordinate system, respectively.

2.5. Evaluation metrics

The performance of Lac-Gwc Net was evaluated using end-point error (*EPE*) and *bad*-3 (Gao et al., 2024b). *EPE* is the per-pixel average disparity error and *bad*-3 is the fraction of pixels with errors larger than three. The *EPE* and *bad*-3 can be calculated through Eqs. (8) and (9), respectively.

$$EPE = \frac{1}{N}\sum_{(x,y)} \left| d_{\mathrm{pre}}(x,y) - d_{\mathrm{gt}}(x,y) \right| \tag{8}$$

$$bad-3 = \frac{1}{N}\sum_{(x,y)} \left(\left| d_{\mathrm{pre}}(x,y) - d_{\mathrm{gt}}(x,y) \right| > 3 \right) \times 100\% \tag{9}$$

Where N is the number of efficient pixels in the disparity map, $d_{\mathrm{pre}}(x, y)$ and $d_{\mathrm{gt}}(x,y)$ denote the predicted disparity and the ground truth, respectively.

The mean average precision (*mAP*) was adopted to evaluate the detection performance of YOLOv8 series, which defined by Eqs. (10) and (11).

$$mAP = \frac{1}{c}\sum_{i=0}^{c} AP_i \tag{10}$$

$$AP_i = \int_0^1 P(R_i)\mathrm{d}R_i \tag{11}$$

Where c is the number of labeled classes, which is 2 in this study; AP_i is the average precision of a certain class, which is the area under *P-R* curve of the i^{th} class; *P* and *R* are precision and recall, which are defined in Eqs. (12) and (13), respectively. The *TP*, *TN*, *FP* and *FN* are indicated four types of detected datasets: true positive, true negative, false positive and false negative according to integration of ground truth and predicted class, respectively.

$$P = \frac{TP}{TP + FP} \tag{12}$$

$$R = \frac{TP}{TP + FN} \tag{13}$$

The performance of kiwifruit localization was assessed by localization accuracy. These evaluation indicators and the analysis of corresponding results can further reflect performance comparison of binocular localization. In terms of localization accuracy, it refers to mean absolute error in coordinate-axis (MAE_{axis}), where the axis represents X-axis or Y-axis or Z-axis of the world coordinate. The definition of MAE_{axis} is shown in Eq. (14).

$$MAE_{axis} = \frac{1}{N}\sum_i (CV_i - GT_i) \tag{14}$$

Where, CV_i and GT_i represent the calculated value and ground truth of the i^{th} calyx, respectively. N represents the total number of the ground truth of calyx localization.

3. Results and discussion

3.1. Calibration results of the binocular camera and the LiDAR

The results of the joint calibration of the binocular camera and the LiDAR are acceptable for converting the LiDAR point cloud into the ground truth of disparity map, which were employed for training the end-to-end stereo matching network. Joint calibration results of the binocular camera and the LiDAR are a critical aspect in determining accuracy of disparity map, as precise calibration parameters decide whether disparity is correct. Joint calibration of the binocular camera and the LiDAR was carried out in two steps. The first step was the calibration of the intrinsic and extrinsic parameters of the binocular camera. The results are shown in Table 2. After calibration, the average reprojection error is 0.1454 mm, which is acceptable for depth and disparity conversion. Next was calibration of extrinsic parameters of the left lens of the binocular camera and the LiDAR, as shown in Table 3, with average reprojection error of 1.0395 pixels in horizontal direction and 1.0308 pixels in vertical direction, which is acceptable for coordinate transformation between the left lens of the binocular camera and the LiDAR.

Table 2
Intrinsic and extrinsic parameters of ZED 2i binocular camera.

Parameters	Value
Intrinsic parameter matrix of left camera K_1	$\begin{bmatrix} 518.843 & 0 & 540.017 \\ 0 & 614.084 & 359.271 \\ 0 & 0 & 0 \end{bmatrix}$
Intrinsic parameter matrix of right camera K_2	$\begin{bmatrix} 518.767 & 0 & 541.496 \\ 0 & 615.583 & 363.337 \\ 0 & 0 & 0 \end{bmatrix}$
Rotation matrix R_{stereo}	$\begin{bmatrix} 1 & 0.000473 & -0.003134 \\ -0.000487 & 1 & -0.004598 \\ 0.003132 & 0.004600 & 0 \end{bmatrix}$
Translation vector t_{stereo}	[-119.671, 3.842, 2.025]

Table 3
Extrinsic parameters of the left lens of ZED 2i binocular camera and Livox Mid 70 LiDAR.

Extrinsic parameters	Value
R_{L2C}	$\begin{bmatrix} 0.006440 & -0.999978 & -0.001625 \\ -0.002682 & 0.001607 & -0.999995 \\ 0.999976 & 0.006444 & -0.002671 \end{bmatrix}$
t_{L2C}	[0.053854, 0.055226, −0.003569]

3.2. Performance of LaC-Gwc Net using the TSPFA on full-resolution disparity map

LaC-Gwc Net was trained on four different datasets, and the comparison of losses indicates that the model converges best on Dataset-KS. This excellent convergence suggests that Dataset-KS provides high quality training data. LaC-Gwc Net was trained on four datasets at different processing stages: Dataset-NF, Dataset-FF, Dataset-SF, and Dataset-KS. Loss curve in Fig. 11 shows that Dataset-KS exhibited best convergence during network training. Dataset-NF and Dataset-SF showed the worst convergence. This indicates that absence of denoising and tailing removal in first stage of the LiDAR point cloud processing leaded to a significant number of errors in the ground truth of disparity,

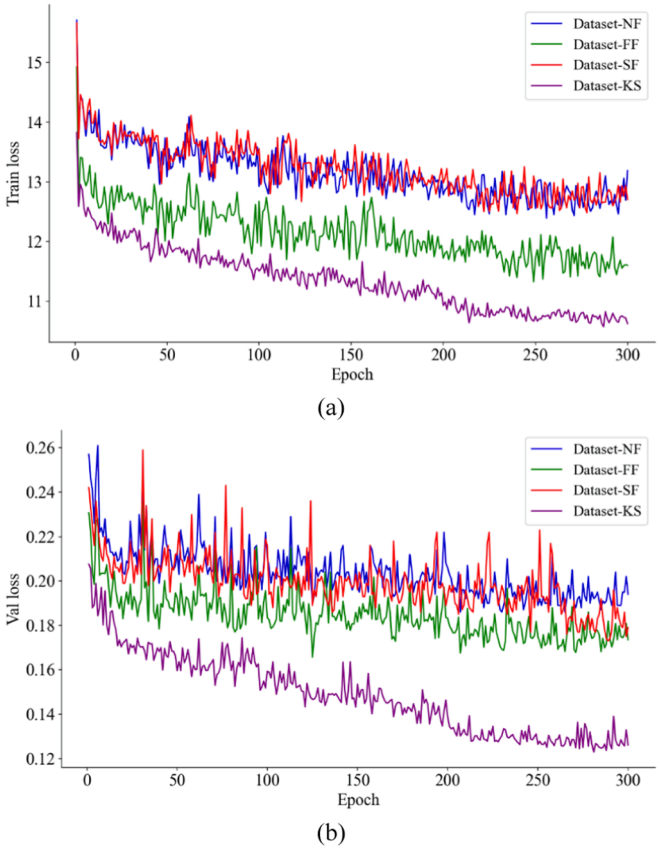

Fig. 11. Loss curves on different datasets. (a) Train loss curves on different datasets; (b) Val loss curves on different datasets.

hindering network's convergence. While Dataset-SF showed convergence, the generated disparity regions are larger than the actual target objects. This is due to that the LiDAR took 3 ~ 5 s to generate point cloud that applied for obtaining disparity regions. The target objects have larger disparity regions than actual regions if they were shake by wind during this time. This leads to the failure of the disparity map to align with part areas of the left image, thereby causing the end-to-end stereo matching network to be unable to extract key features. The TSPFA effectively reduced the impact of noise from the LiDAR and ensured the alignment of the disparity map with the left image, providing a high-quality dataset for network feature learning.

Furthermore, Lac-Gwc Net trained using Dataset-KS achieved the lowest *EPE* and *bad*-3 compared with that using other three datasets, demonstrating excellent performance in terms of matching. As shown in Table 4, Lac-Gwc Net trained using Dataset-KS performed best, with *EPE* of 3.8 pixels and *bad*-3 at 11.7%, which is significantly better compared to models trained on the other three datasets. Specifically, *bad*-3 was at least 4.2% ahead and improved by at least 26.4%, showing considerable advantage. Additionally, LaC-Gwc Net only took average of 250.0 ms to process images with resolution of 1080 × 720 pixels. The model is capable of achieving rapid and precise full-resolution stereo matching in kiwifruit orchards, laying the foundation for the perception of kiwifruit harvesting robots.

3.3. Performance of LaC-Gwc Net with the TSPFA on thin obstacles

LaC-Gwc Net, trained on Dataset-KS, demonstrated satisfactory results in stereo matching of normal objects, even thin obstacles, such as wires and branches. As shown in Fig. 12, the position of the wire was indicated by a yellow bounding box, while branch was marked with a red bounding box. Analysis of the disparity map predicted by Lac-Gwc Net trained on different datasets reveals that only the model trained on Dataset-KS can effectively match the disparity of the wire. Lac-Gwc Net trained on Dataset-KS also showed the distinct edges results in predicting the edges of the disparity map for branches, while Lac-Gwc Net trained on other three datasets performed poorly in this regard, exhibiting blurred or missing boundaries of thin obstacles. Additionally, many studies have successfully utilized deep learning methods for stereo matching of fruits based on the bounding boxes of object detection networks (Tang et al., 2023; Li et al., 2023b). However, these methods were limited to matching objects after detecting them and cannot obtain the disparity values of the full-resolution image, hence failing to localize obstacles. Traditional stereo matching algorithms can match the full-resolution image. However, they are plagued with numerous matching failures, leading to significant disparities, especially for thin obstacles, making their localization difficult (Jafari et al., 2019; Niknejad et al., 2023). The movement path of harvesting arms is affected by the thin obstacles inevitably, making their stereo matching equally critical. LaC-Gwc Net, an end-to-end stereo matching network, greatly ameliorates this situation, particularly to thin obstacles when trained on Dataset-KS using the TSPFA.

LaC-Gwc Net, trained on Dataset-KS, performs well with thin obstacles was due to using the TSPFA. Dataset-KS allows the end-to-end stereo matching network to extract key features more effectively, thereby enhancing its ability to match thin obstacles. In the first stage of filtering, the main goal is to remove outliers and noise points from the LiDAR data, ensuring the accuracy and authenticity of the point cloud to provide a reliable foundation for subsequent disparity map processing. The focus of the second-stage filtering is on utilizing RGB images to further optimize the data. By ensuring that only the areas with kiwifruit canopy have disparity values, the authenticity of the disparity map is further enhanced. This approach ensures that background areas like the sky do not have disparity values. Through the TSPFA, the quality of the dataset is significantly improved, which is essential for the end-to-end stereo matching network to effectively extract key features and accurately match thin obstacles.

Table 4
Stereo matching results of LaC-Gwc Net on four datasets.

Evaluation metrics	Dataset-NF	Dataset-FF	Dataset-SF	Dataset-KS
EPE (pixels)	5.5	5.2	5.5	3.8
bad-3 (%)	17.0	15.9	16.9	11.7

3.4. Performance of object detection models

YOLOv8m achieved satisfactory performance in detecting kiwifruits and calyxes in modern kiwifruit orchards. The performance of object detection was tested using the Dataset-KD test set in two orchards. YOLOv8m processed images with resolution of 1280 × 720 pixels in average of only 7.0 ms, indicating that object detection can quickly detect kiwifruits and their calyxes. Several advanced the models of object detection were tested, with results as shown in Table 5. Among them, YOLOv8l had the highest *mAP*, reaching 93.5%, while YOLOv8n had the lowest at only 71.8%, difference of 21.7%. It was also found that YOLOv8m has similar accuracy to YOLOv8l but with faster detection speed. Many researches have chosen YOLO series of object detection models for kiwifruit detection and achieved good results, but comparison revealed that YOLOv8m has good *mAP* and detection speed (Xia et al., 2023; Gao et al., 2024a). Therefore, YOLOv8m has been selected to detect kiwifruit and their calyxes for localization.

To explore factors affecting detection accuracy of kiwifruit, it can be found from Table 5 that *AP* of calyx gave a larger influence on *mAP* than that of kiwifruit. It is observed that there is not a significant difference among the models of object detection for kiwifruits, with only a 0.8% variation. However, there was considerable gap in *mAP* of calyx, with the highest being 87.7% for YOLOv8l and the lowest being 45.1% for YOLOv8n, difference of 42.6%. It is clear that in the detection of kiwifruits and their calyxes, the detection of the calyx is the most important factor influencing the *mAP* (Williams et al., 2019). This is because the calyx is a small target, with relatively low pixels representation in the image, less prominent features, and difficulty in feature extraction, necessitating the use of more complex network of object detection for feature extraction (Mahaur et al., 2023; Li et al., 2023a). Conversely, kiwifruit has higher pixels ratio in images and more distinct features, allowing less complex network of object detection to effectively extract features.

3.5. Performance of kiwifruit localization

The methodology using LaC-Gwc Net that trained on Dataset-KS and YOLOv8m provided an acceptable accuracy in kiwifruit localization. *MAE* for kiwifruit localization along the X, Y, and Z axis are 4.2 mm, 6.7 mm, and 4.0 mm, respectively. Considering that the average length of kiwifruit is over 60.0 mm and the average diameter is over 50 mm (Wang et al., 2024), localization results obtained by the depth estimation of LaC-Gwc Net trained on Dataset-KS, combined with YOLOv8m detection, are acceptable for kiwifruit harvesting.

The reasons for localization errors in kiwifruit in this study are varied. The first reason for localization error of kiwifruit is due to the conversion of disparity into depth, which requires extrinsic parameters of the binocular camera. These parameters are obtained through calibration using calibration board. Inevitable deviation during calibration led to incorrect depth measurements for kiwifruit. The second reason is that depth errors or missing can occur due to obstacles, resulting in disparity errors and kiwifruit localization errors. The third reason is that sometimes YOLOv8m fails to detect the calyxes of kiwifruit. In these cases, the central point of the bounding boxes of kiwifruit was utilized for localization, while the ground truth of kiwifruit localization was obtained from the calyx, leading to discrepancies. These are the main reasons for localization errors, which can cause the predicted position of the kiwifruit to differ from the ground truth.

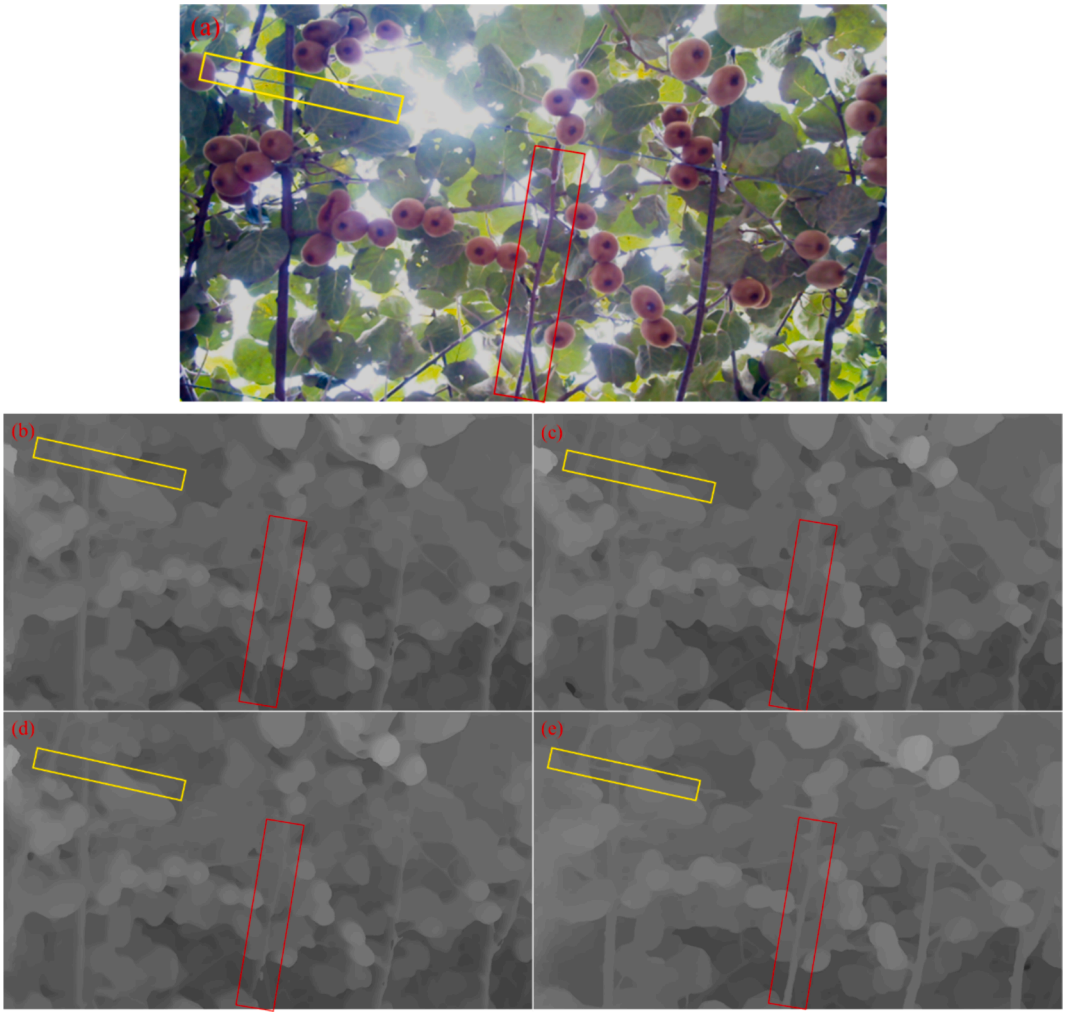

Fig. 12. Predicted disparity map from Lac-Gwc Net trained on different datasets, with the yellow bounding box in the image representing wire and the red bounding box representing the branch of kiwifruit trees. (a) Left image; (b) Predicted disparity map obtained by training Lac-Gwc Net on Dataset-NF; (c) Predicted disparity map obtained by training Lac-Gwc Net on Dataset-FF; (d) Predicted disparity map obtained by training Lac-Gwc Net on Dataset-SF; (e) Predicted disparity map obtained by training Lac-Gwc Net on Dataset-KS. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 5
Detection results of four YOLOv8 models on the Dataset-KD.

Models	*AP* (%)		*mAP* (%)	Detection speed (ms / image)
	kiwi	calyx		
YOLOv8n	98.5	45.1	71.8	4.2
YOLOv8s	99.1	82.7	90.9	6.0
YOLOv8m	99.3	86.9	93.1	7.0
YOLOv8l	99.3	87.7	93.5	9.5

Additionally, the study conducted comparisons with other methods for kiwifruit localization, finding that the results still achieving similar results. As show in Table 6, traditional stereo matching methods (Song, 2021; Liu et al., 2024) were highly sensitive to occlusion and uneven illumination, which often results in depth missing and errors. In contrast, deep learning techniques have been introduced (Gao et al., 2024a) to enhance the localization precision of kiwifruit stereo cameras. However, this method only obtained the depth of detected objects, leaving other parts unknown and not providing comprehensive depth information for harvesting robots. Other types of RGB-D cameras have been employed for kiwifruit localization(Liu, 2020; Liu et al., 2024). While these RGB-D cameras demonstrated acceptable accuracy, their

Table 6
Results from previous studies on kiwifruit localization.

	Localization methods	Object	Environments	Localization error (mm)
Liu (2020)	Structured light	Kiwifruit	Indoor	3.6
Song (2021)	Classical binocular	Kiwifruit	Orchard	10.4
Gao (2024a)	Improved binocular	Kiwifruit and calyx	Orchard	4.8
Liu (2024)	ToF	Kiwifruit and calyx	Orchard	18.9
	Structured light			18.8
	Classical binocular			30.0
Our method	End-to-end stereo matching	Kiwifruit and calyx	Orchard	4.0

increased sensitivity to light often resulted in depth missing, thereby complicating accurate positioning and obstacle avoidance. In contrast, our method achieves high-precision full-resolution depth estimation.

4. Conclusions

This study proposes a methodology for full-resolution depth estimation and precise localization of kiwifruit by combining LaC-Gwc Net and YOLOv8m. The proposed TSPFA can effectively reduce the noise of the LiDAR and generate accurately aligned disparity map, constructing a stereo matching dataset of kiwifruit orchards, i.e., Dataset-KS, enabling the end-to-end stereo matching network to effectively learn features. LaC-Gwc Net trained with Dataset-KS achieved *EPE* of 3.8 pixels, which is beneficial for thin obstacles. YOLOv8m excels in detecting kiwifruit and their small-sized calyxes, achieving a notable *mAP* of 93.1% with a detection speed of only 7.0 ms, benefiting from increased network complexity. The use of LaC-Gwc Net and YOLOv8m in the field achieved high-precision localization of kiwifruits. The localization error on the Z-axis was only 4.0 mm, fully meeting the operational accuracy requirements of kiwifruit harvesting robots. This study provides a new methodology to locate fruit in orchards, offering high-precision full-resolution depth for harvesting operations. Next, we will focus on researching lightweight high-precision full-resolution stereo matching methods for better apply in harvesting robots.

CRediT authorship contribution statement

Xudong Jing: Writing – original draft, Methodology, Investigation, Data curation. **Hanhui Jiang:** Writing – review & editing, Software, Investigation. **Shiao Niu:** Writing – review & editing, Software. **Haosen Zhang:** Writing – review & editing, Software. **Bryan Gilbert Murengami:** Data curation, Writing – review & editing. **Zhenchao Wu:** Writing – review & editing, Methodology, Investigation. **Rui Li:** Writing – review & editing, Methodology. **Chengquan Zhou:** Writing – review & editing, Methodology. **Hongbao Ye:** Writing – review & editing, Methodology, Conceptualization. **Jinyong Chen:** Methodology, Writing – review & editing. **Yaqoob Majeed:** Writing – review & editing, Methodology. **Longsheng Fu:** Writing – review & editing, Supervision, Methodology, Data curation, 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.

Data availability

https://github.com/fu3lab/Kiwifruit_stereo_camera_image.

Acknowledgements

This work was partially supported by the National Natural Science Foundation of China (32371999); Key Research and Development Program of Shaanxi, China (2024NC-YBXM-195, 2023JBGS-21); Open Project of Key Laboratory of Agricultural Equipment for Hilly and Mountainous Areas in Southeastern China (Co-construction by Ministry and Province), Ministry of Agriculture and Rural Affairs, China (QSKF2023002); National Foreign Expert Project, Ministry of Science and Technology, China (QN2022172006L, DL2022172003L).

References

Au, C., Lim, S., Duke, M., Kuang, Y., Redstall, M., Ting, C., 2023a. Integration of stereo vision system calibration and kinematic calibration for an autonomous kiwifruit harvesting system. Int. J. Intell. Robot. Appl. 7, 350–369. https://doi.org/10.1007/s41315-022-00263-x.

Au, W., Zhou, H., Liu, T., Kok, E., Wang, X., Wang, M., Chen, C., 2023b. The monash apple retrieving system: A review on system intelligence and apple harvesting performance. Comput. Electron. Agric. 213, 108164 https://doi.org/10.1016/j.compag.2023.108164.

Chang, J., Chen, Y., 2018. PCW-Net: Pyramid combination and warping cost volume for stereo matching, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5410–5418. doi: 10.48550/arXiv.1803.08669.

Gao, K., Gui, C., Wang, J., Miu, H., 2021. Research on recognition and positioning technology of dragon fruit based on binocular vision. Lect. Notes Data Eng. Commun. Technol. 88, 1257–1264. https://doi.org/10.1007/978-3-030-70665-4_136.

Gao, C., Jiang, H., Liu, X., Li, H., Wu, Z., Sun, X., He, L., Mao, W., Majeed, Y., Li, R., Fu, L., 2024a. Improved binocular localization of kiwifruit in orchard based on fruit and calyx detection using YOLOv5x for robotic picking. Comput. Electron. Agric. 217, 108621 https://doi.org/10.1016/j.compag.2024.108621.

Gao, Y., Wang, Q., Rao, X., Xie, L., Ying, Y., 2024b. OrangeStereo: A navel orange stereo matching network for 3D surface reconstruction. Comput. Electron. Agric. 217, 108626 https://doi.org/10.1016/j.compag.2024.108626.

Guo, X., Yang, K., Yang, W., Wang, X., Li, H., 2019. Group-wise correlation stereo network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3273–3282. doi: 10.1109/CVPR.2019.00339.

Hou, C., Zhang, X., Tang, Y., Zhuang, J., Tan, Z., Huang, H., Chen, W., Wei, S., He, Y., Luo, S., 2022. Detection and localization of citrus fruit based on improved You Only Look Once v5s and binocular vision in the orchard. Front. Plant Sci. 13, 972445 https://doi.org/10.3389/fpls.2022.972445.

Hsieh, K., Huang, B., Hsiao, K., Tuan, Y., Shih, F., Hsieh, L., Chen, S., Yang, I., 2021. Fruit maturity and location identification of beef tomato using R-CNN and binocular imaging technology. J. Food Meas. Charact. 15, 5170–5180. https://doi.org/10.1007/s11694-021-01074-7.

Jafari, A., Khojastehpour, M., Emadi, B., 2019. Disparity map computation of tree using stereo vision system and effects of canopy shapes and foliage density. Comput. Electron. Agric. 156, 627–644. https://doi.org/10.1016/j.compag.2018.12.022.

Jocher, G., Chaurasia, A., Qiu, J., 2023. Ultralytics YOLO (Version 8.0.0). https://github.com/ultralytics/ultralytics.

Laga, H., Jospin, L., Boussaid, F., Bennamoun, M., 2022. A survey on deep learning techniques for stereo-based depth estimation. IEEE Trans. Pattern Anal. Mach. Intell. 44, 1738–1764. https://doi.org/10.1109/TPAMI.2020.3032602.

Li, T., Fang, W., Zhao, G., Gao, F., Wu, Z., Li, R., Fu, L., Dhupia, J., 2023b. An improved binocular localization method for apple based on fruit detection using deep learning. Inf. Process. Agric. 10, 276–287. https://doi.org/10.1016/j.inpa.2021.12.003.

Li, K., Gong, W., Shi, Y., Li, L., He, Z., Ding, X., Wang, Y., Ma, L., Hao, W., Yang, Z., Cui, Y., 2023a. Predicting positions and orientations of individual kiwifruit flowers and clusters in natural environments. Comput. Electron. Agric. 211, 108039 https://doi.org/10.1016/j.compag.2023.108039.

Li, T., Xie, F., Zhao, Z., Zhao, H., Guo, X., Feng, Q., 2023c. A multi-arm robot system for efficient apple harvesting: Perception, task plan and control. Comput. Electron. Agric. 211, 107979 https://doi.org/10.1016/j.compag.2023.107979.

Lin, G., Tang, Y., Zou, X., Li, J., Xiong, J., 2019. In-field citrus detection and localisation based on RGB-D image analysis. Biosyst. Eng. 186, 34–44. https://doi.org/10.1016/j.biosystemseng.2019.06.019.

Ling, X., Zhao, Y., Gong, L., Liu, C., Wang, T., 2019. Dual-arm cooperation and implementing for robotic harvesting tomato using binocular vision. Rob. Auton. Syst. 114, 134–143. https://doi.org/10.1016/j.robot.2019.01.019.

Liu, B., Yu, H., Long, Y., 2022. Local similarity pattern and cost self-reassembling for deep stereo matching networks, in: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 1647–1655. doi: 10.48550/arXiv.2112.01011.

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. F. Robot 41 (4), 881–894. https://doi.org/10.1002/rob.22297.

Liu, Z., 2020. Kiwifruit detection and localization methods based on multi-source information fusion. Master Thesis, Northwest A&F University, Shaanxi, China. doi: 10.27409/d.cnki.gxbnu.2020.000944.

Mahaur, B., Mishra, K., Kumar, A., 2023. An improved lightweight small object detection framework applied to real-time autonomous driving. Expert Syst. Appl. 234, 121036 https://doi.org/10.1016/j.eswa.2023.121036.

Mejia, G., Montes, A., Flores, G., 2023. Strawberry localization in a ridge planting with an autonomous rover. Eng. Appl. Artif. Intell. 119, 105810 https://doi.org/10.1016/j.engappai.2022.105810.

Mirbod, O., Choi, D., Heinemann, P., Marini, R., He, L., 2023. On-tree apple fruit size estimation using stereo vision with deep learning-based occlusion handling. Biosyst. Eng. 226, 27–42. https://doi.org/10.1016/j.biosystemseng.2022.12.008.

Niknejad, N., Bidese, R., Bao, Y., Payn, K., Zheng, J., 2023. Phenotyping of architecture traits of loblolly pine trees using stereo machine vision and deep learning: Stem diameter, branch angle, and branch diameter. Comput. Electron. Agric. 211, 107999 https://doi.org/10.1016/j.compag.2023.107999.

Poggi, M., Tosi, F., Batsos, K., Mordohai, P., Mattoccia, S., 2022. On the synergies between machine learning and binocular stereo for depth estimation from Images: A survey. IEEE Trans. Pattern Anal. Mach. Intell. 44, 5314–5334. https://doi.org/10.1109/TPAMI.2021.3070917.

Popovic, M., Thomas, F., Papatheodorou, S., Funk, N., Vidal, T., Leutenegger, S., 2021. Volumetric occupancy mapping with probabilistic depth completion for robotic navigation. IEEE Robot. Autom. Lett. 6, 5072–5079. https://doi.org/10.1109/LRA.2021.3070308.

Putra, O., Riansyah, M., Rahmanti, F., Priyadi, A., Wulandari, D., Ogata, K., Yuniarno, E., Purnomo, M., 2023. Enhancing LiDAR-based object recognition through a novel

denoising and modified GDANet framework. IEEE Access 12, 7285–7297. https://doi.org/10.1109/access.2023.3347033.

Song, Z., 2021. Kiwifruit canopy image segmentation and multi-classes fruit localization methods based on deep learning. Master Thesis, Northwest A&F University, Shaanxi, China. doi: 10.27409/d.cnki.gxbnu.2021.000573.

Suo, R., Gao, F., Zhou, Z., Fu, L., Song, Z., Dhupia, J., Li, R., Cui, Y., 2021. Improved multi-classes kiwifruit detection in orchard to avoid collisions during robotic picking. Comput. Electron. Agric. 182, 106052 https://doi.org/10.1016/j.compag.2021.106052.

Tang, Y., Zhou, H., Wang, H., Zhang, Y., 2023. Fruit detection and positioning technology for a Camellia oleifera C. Abel orchard based on improved YOLOv4-tiny model and binocular stereo vision. Expert Syst. Appl. 211, 118573 https://doi.org/10.1016/j.eswa.2022.118573.

UN Food & Agriculture Organization. 2024. Production/Yield quantities of kiwi fruit in World. Retrieved 2024-06-08, from https://www.fao.org/faostat/zh/#data/QCL/visualize.

Wang, J., Cai, X., Zeng, S., Zhang, Z., Chi, Q., Guo, W., 2024. Effect of forchlorfenuron and thidiazuron on kiwifruits' internal qualities, optical properties and their relationship during growth. Spectrochim. Acta Part A Mol. Biomol. Spectrosc. 308, 123749 https://doi.org/10.1016/j.saa.2023.123749.

Wang, X., Kang, H., Zhou, H., Au, W., Chen, C., 2022. Geometry-aware fruit grasping estimation for robotic harvesting in apple orchards. Comput. Electron. Agric. 193, 106716 https://doi.org/10.1016/j.compag.2022.106716.

Williams, H., Jones, M., Nejati, M., Seabright, M., Bell, J., Penhall, N., Barnett, J., Duke, M., Scarfe, A., Ahn, H., Lim, J., MacDonald, B., 2019. Robotic kiwifruit harvesting using machine vision, convolutional neural networks, and robotic arms. Biosyst. Eng. 181, 140–156. https://doi.org/10.1016/j.biosystemseng.2019.03.007.

Xia, Y., Nguyen, M., Yan, W., 2023. A real-time kiwifruit detection based on improved YOLOv7. Lect. Notes Comput. Sci 48–61. https://doi.org/10.1007/978-3-031-25825-1_4.

Xiong, Y., Ge, Y., Grimstad, L., From, P., 2020. An autonomous strawberry-harvesting robot: Design, development, integration, and field evaluation. J. F. Robot. 37 (2), 202–224. https://doi.org/10.1002/rob.21889.

Yang, Y., Han, Y., Li, S., Yang, Y., Zhang, M., Li, H., 2023. Vision based fruit recognition and positioning technology for harvesting robots. Comput. Electron. Agric. 213, 108258 https://doi.org/10.1016/j.compag.2023.108258.

Yun, T., Jiang, K., Li, G., Eichhorn, M., Fan, J., Liu, F., Chen, B., An, F., Cao, L., 2021. Individual tree crown segmentation from airborne LiDAR data using a novel Gaussian filter and energy function minimization-based approach. Remote Sens. Environ. 256, 112307 https://doi.org/10.1016/j.rse.2021.112307.

Zhao, G., Yang, R., Jing, X., Zhang, H., Wu, Z., Sun, X., Jiang, H., Li, R., Wei, X., Fountas, S., Zhang, H., Fu, L., 2023. Phenotyping of individual apple tree in modern orchard with novel smartphone-based heterogeneous binocular vision and YOLOv5s. Comput. Electron. Agric. 209, 107814 https://doi.org/10.1016/j.compag.2023.107814.

Zhuang, J., Luo, S., Hou, C., Tang, Y., He, Y., Xue, X., 2018. Detection of orchard citrus fruits using a monocular machine vision-based method for automatic fruit picking applications. Comput. Electron. Agric. 152, 64–73. https://doi.org/10.1016/j.compag.2018.07.004.