Research Square

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

Pose Estimation Method for Orchard Fruit Picking under Missing Depth Information Scenarios

Gang Peng
Huazhong University of Science and Technology
Jiaqi Yang
M202473734@hust.edu.cn
Huazhong University of Science and Technology
Chen Lai
Huazhong University of Science and Technology

Research Article

Keywords: orchard fruit harvesting, pose estimation, deformable attention, feature enhancement, decoupling modeling strategy

Posted Date: November 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11626454/v1

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

Additional Declarations: No competing interests reported.

Pose Estimation Method for Orchard Fruit Picking under Missing Depth Information Scenarios

Gang Peng[1,2,3], Jiaqi Yang[1,2,3,*], Chen Lai[1,2,3]

1 School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, China

2 Key Laboratory of Image Information Processing and Intelligent Control, Ministry of Education, Wuhan 430074, China

3 Engineering Research Center of Autonomous Intelligent Unmanned System, Ministry of Education, Wuhan 430074, China

***Corresponding author: Jiaqi Yang**

Email: M202473734@hust.edu.cn

1037 Luoyu Road, Hongshan District, Wuhan 430074, China

Additional author contact: Gang Peng, penggang@hust.edu.cn

Categories (7), (6)

MSC 2020: 68T45; 68T40

Acknowledgments

This work was supported by the Hubei Province Core Technology Application Research Project for Agricultural Machinery Equipment (No. HBSNYT202213).

Abstract:

A fruit 6D pose estimation method relying solely on monocular RGB images is proposed to address the problem of missing depth information in outdoor orchard scenes. The method first extracts multi-scale features and target bounding box information from images using an object detection network, and then introduces a deformable attention mechanism and explicit position priors into the Transformer decoder to further enhance the capability of modeling the spatial structure of the target and focusing on key regions. In the pose regression stage, a translation/rotation decoupling modeling strategy based on six-dimensional rotation representation and the pinhole camera model is designed, and a two-stage regression head combining initial prediction and residual iterative refinement is adopted to achieve high-precision and robust estimation from coarse to fine. Experimental results show that the method outperforms existing mainstream methods on public datasets and self-collected orchard datasets, and exhibits high accuracy, robustness, and generalization capability in complex scenes such as weak texture, occlusion, and illumination variation, providing reliable support for visual perception of agricultural harvesting robots in complex environments.

Keywords: orchard fruit harvesting; pose estimation; deformable attention; feature enhancement; decoupling modeling strategy

1. Introduction

At present, the deep integration of agriculture and robotics technology is increasingly becoming an important direction for promoting the modernization of agriculture. Influenced by the continuous reduction of rural labor and the growing demand for high-efficiency and high-precision agricultural operations, the traditional labor-intensive agricultural model is facing severe challenges. Meanwhile, the agricultural environment is highly unstructured and dynamic, imposing greater requirements on perception and manipulation systems. With the rapid development of smart agriculture [1], perception technology, information processing, intelligent

decision-making, and automated execution have been widely applied in multiple stages of farming, management, harvesting, transportation, and livestock production [2–4], significantly enhancing agricultural intelligence and operational efficiency. By introducing robotic systems and intelligent technologies, agricultural production can more efficiently and precisely address issues such as labor shortages and complex working conditions, demonstrating tremendous application potential and industrial value. As an essential carrier of intelligent agriculture, agricultural robots are gradually becoming a key force driving technological innovation and the construction of modern agricultural systems.

In fruit-picking robotic systems, the manipulator, as the core execution mechanism, relies heavily on environmental perception accuracy and stability—particularly on the precision of fruit pose estimation. Accurate 6D pose estimation not only improves the success rate of single picking operations but also plays a crucial role in task planning for continuous multi-target operations, serving as a bridge between visual perception and motion control. However, traditional methods that depend on RGB-D cameras to obtain depth information are often unreliable in outdoor orchards. Factors such as strong sunlight, cluttered backgrounds, and sensor power constraints can lead to severe noise or complete loss of depth maps. Moreover, fruit surfaces are often low-textured, and significant variations in size and shape exist across different species, making it difficult for model-based pose estimation methods to maintain generalization accuracy. This instability in perception severely restricts the applicability of existing approaches in real-world picking scenarios.

Recent studies on fruit-handling devices and sensing further indicate that pose estimation is closely related to the downstream execution quality of harvesting robots. For example, tactile sensing in electric gripper actuators has been used to predict fruit hardness for nondestructive grasping [5], while a telescopic-link mechanical gripper for fruit harvesting highlights the need for pose-aware end-effector alignment during picking [6]. These studies suggest that accurate fruit pose estimates can improve not only fruit localization, but also grasping force regulation and stem-cutting trajectory planning.

Although stereo vision is widely used in indoor environments, it also faces challenges in orchards: fruit surfaces with simple and repetitive textures cause disparity-matching difficulties; significant size variations among different fruit species make it hard to select a suitable stereo baseline for both near and distant targets; and maintaining stable calibration and synchronization under outdoor conditions is challenging. In contrast, monocular cameras offer advantages such as structural simplicity, low cost, and low power consumption, making them easier to deploy on orchard robotic platforms. Combined with learning-based geometric modeling methods, monocular vision can maintain high robustness under varying illumination, weak textures, and partial occlusions, thus presenting a more promising solution to the problem of missing depth information. Kim et al. [7] developed a tomato harvesting robotic system based on 6D pose estimation, and Sun et al. [8] estimated citrus poses from single RGB images for automated harvesting, demonstrating that monocular RGB information alone can provide useful geometric cues for fruit pose estimation in orchard or greenhouse environments. However, these methods do not explicitly address severe occlusion and missing depth information in more complex orchard scenes. Therefore, achieving accurate 6D fruit pose estimation solely from RGB images has become a key research direction and challenge in the field. In recent years, deep learning-based 6D pose estimation methods have developed rapidly, exploring various input modalities, network architectures, and data processing strategies. For RGB-D settings, DenseFusion[9] further demonstrated the effectiveness of pixel-level color-depth feature fusion and iterative pose refinement, but its reliance on valid depth maps also highlights the necessity of RGB-only pose estimation when depth information is missing in outdoor orchards.

From the perspective of robotic measurement and engineering implementation, vision-based pose measurement is valued for its non-contact, low-cost, and wide-range characteristics. Recent monocular pose measurement and calibration studies have shown that refined projective geometry, homography decomposition, bundle adjustment, and calibrated pinhole projection can improve pose accuracy in robotic systems [10,11]. In addition, image-based under-canopy navigation demonstrates that

visual perception remains an effective alternative when GNSS, depth sensing, or other environmental cues become unreliable in crop environments [12].

Regarding network architecture, convolutional neural networks (CNNs), known for their strong local modeling capability and fast convergence, have been widely used in early 6D pose estimation tasks, such as PoseCNN. With the introduction of Transformer structures [13], researchers discovered its superior ability to model long-range dependencies and spatial context, making it an emerging dominant paradigm in recent pose estimation studies. Recent studies published in Measurement Science and Technology further highlight the importance of combining local correspondences, image matching, and monocular visual constraints for robust pose measurement: Wu et al. [14] proposed P2P-Pose for relative camera pose regression guided by image matching, while Tang et al. [15] presented a monocular vision-based 3D pose tracking method using 3D boundary and contour segments. These studies confirm the value of integrating geometric priors with image-level feature learning for accurate pose measurement under constrained sensing conditions. Xiang et al. [16] pioneered PoseCNN by introducing a symmetry-aware loss function that directly regresses 3D translation and rotation parameters through CNNs. Li et al. [17] incorporated object category priors to discretize the continuous pose space, reformulating 6D pose estimation as a classification problem.

Model-based methods that integrate 3D object geometry have demonstrated distinct advantages. Kehl et al. [18] proposed a two-stage framework that first predicts an initial pose and then refines it. Wang et al. [19] embedded the PnP algorithm into the end-to-end trainable GDR-Net, achieving deep integration of geometric constraints and learning. Li et al. [17] further introduced an iterative refinement method that progressively corrects pose estimates by comparing rendered and input images, while Labbe et al. [20] extended this strategy to multi-view scenarios to enhance robustness. Billings et al. [21] designed SilhoNet, which predicts object silhouettes from 3D models and infers 3D translation vectors from 2D bounding boxes, while introducing corrections for rotational ambiguities in symmetric objects. However, the reliance on object symmetry limits its applicability. Notably, Amini et al. [22] recently introduced

Transformer architectures into this domain, extending the Detection Transformer with translation and rotation heads to achieve end-to-end multi-object pose estimation. Although their method employs symmetry-based loss functions, it still depends on object 3D models as prior knowledge.

This paper focuses on the problem of fruit 6D pose perception under depth-missing conditions in complex orchard environments and proposes a depth-free visual pose estimation method. The main contributions are as follows:

(1) A monocular RGB-only 6D pose estimation framework is proposed. Based on multi-scale features and bounding boxes provided by an object detector, deformable attention and explicit positional priors are introduced into the Transformer decoder to enhance spatial modeling and key-region focus.

(2) In the pose regression stage, a decoupled translation/rotation modeling strategy is designed based on a 6D rotation representation and the pinhole camera model. A two-stage regression head combining initial prediction and residual iterative refinement achieves coarse-to-fine, high-precision estimation.

(3) Comparative experiments on YCB-Video and Orange-Pose datasets, along with extensive ablation studies, validate the effectiveness and contribution of each module.

2. Framework of the pose estimation method

The network architecture of this study, as shown in Fig. 1, consists of three main components: an object detection network, a feature enhancement module, and a pose estimation regression head. This cascaded design allows the integration of any pretrained object detection network, enabling joint optimization of detection and pose estimation.

To balance real-time performance and accuracy, YOLOv8 is selected as the object detection backbone. The input image is fed into the YOLOv8 network, which outputs multi-scale feature maps and bounding box parameters (x_c, y_c, w, h), where (x_c, y_c) denotes the center pixel coordinates of the detected object, and w, h represent the width and height of the bounding box, respectively. The multi-scale feature maps and

bounding box information generated by the detection network provide essential semantic and spatial priors for subsequent pose estimation.

After obtaining the multi-scale feature maps, a feature enhancement module based on the Deformable Transformer [23] is introduced to further refine the detected features. The core objective of this module is to enhance the spatial representation and global modeling capabilities of the features. By leveraging the self-attention mechanism, the network establishes semantic associations between the target and the global scene, models contextual relationships among multiple objects, and thereby improves target discriminability and localization robustness.

Finally, a decoupled pose estimation strategy is adopted. In the designed pose estimation head, translation and rotation of the target are modeled and predicted independently to better accommodate their distinct feature distributions and regression complexities.

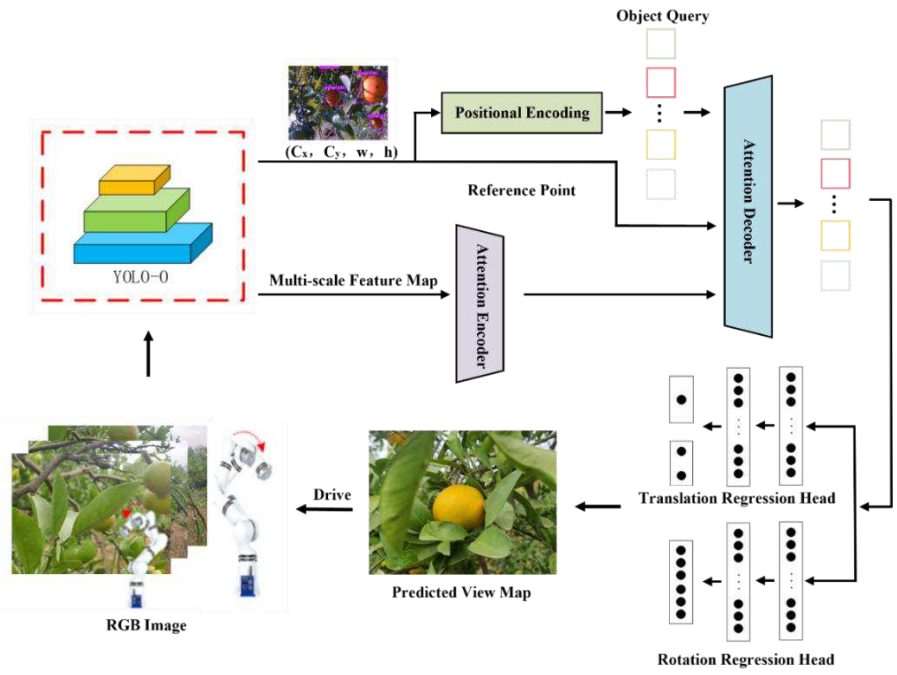

Fig.1 Framework of pose estimation method

3. Feature enhancement module based on deformable attention

The Transformer network structure used in this study is shown in Fig. 2. Compared

with the original Transformer architecture, this structure maintains global modeling capability while incorporating two key improvements tailored for visual pose estimation tasks.

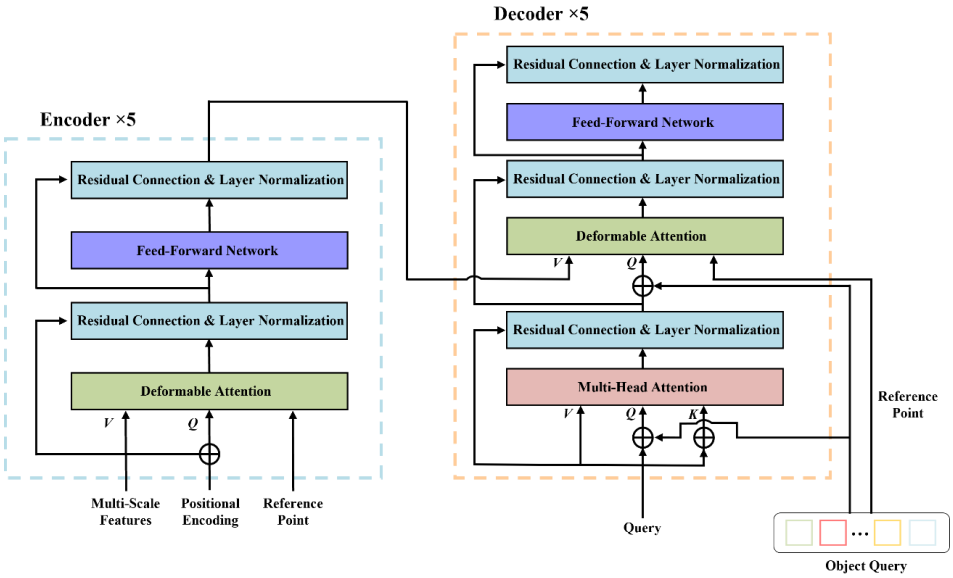

Fig.2 Transformer architecture with deformable attention

The first improvement is the introduction of a deformable attention mechanism. Deformable attention effectively combines the spatially sparse, position-aware capability of deformable convolution with the global modeling power of the Transformer, balancing flexible feature selection with contextual awareness. Specifically, for each query position, instead of computing attention over the entire feature map, the mechanism dynamically selects several key reference points in the neighborhood as sampling locations to perform local attention. This significantly reduces the computation between each query and all keys, thereby greatly lowering overall computational complexity. The deformable attention module, as shown in Fig. 3, efficiently handles multi-scale feature maps, fully extracting relevant information across different scales. Moreover, deformable attention exhibits favorable training characteristics: by focusing spatially on semantically relevant regions, it mitigates the problem of overly dispersed attention weights during the early training stages of the Transformer, accelerating model convergence. Importantly, its computational complexity scales linearly with the number of sampling points rather than with the spatial dimensions of the feature map, overcoming the efficiency bottleneck of standard

attention when processing high-resolution images and enhancing scalability and practicality in visual tasks.

The second improvement is the introduction of explicit positional priors, replacing learnable object queries with bounding box information to strengthen the model's capacity for modeling object location. Specifically, for each detected object, the center coordinates, width, and height of the bounding box are normalized and position-encoded to generate the object query embedding, ensuring effective information propagation among different queries.

Unlike traditional DETR [24], which initializes query vectors randomly, this study adopts a detection-prior-based query generation mechanism. The bounding box parameters are normalized and then passed through a pose encoder to produce query vectors with geometric representation capabilities. These vectors enable semantic interaction among multiple objects via an inter-query attention mechanism. The hidden dimension of the Transformer is $d_h = 2n_pL$, where n_p is the number of queries per object (set to 4 in this study) and $L = 32$.

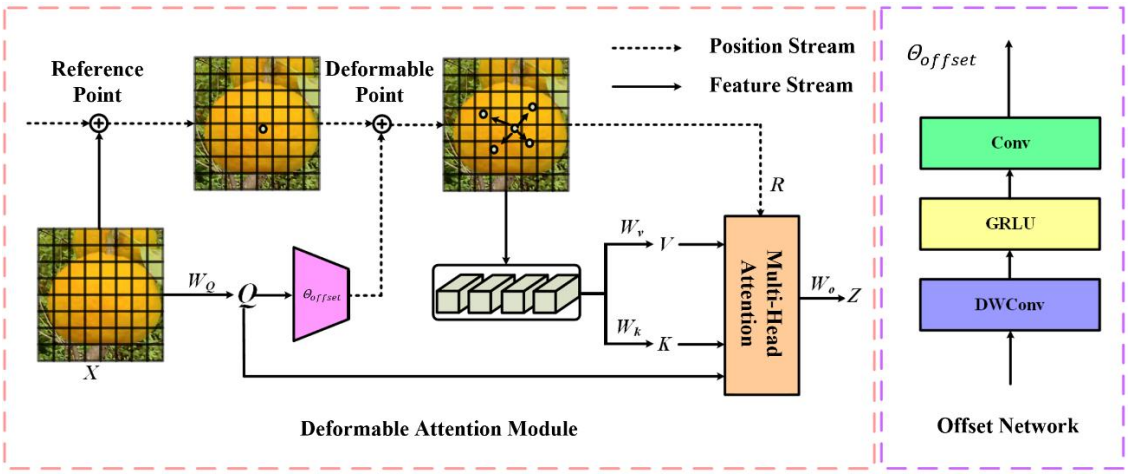

Fig.3 Structure of deformable attention module

4. Pose regression method

After obtaining the feature representations enhanced by the feature enhancement module, these refined features are further used for 6D pose estimation of the target. Considering the significant differences in the feature characteristics of rotation and translation, a targeted decoupled modeling strategy is designed to improve the accuracy

and robustness of pose prediction. Specifically, rotation estimation relies more on the object's intrinsic appearance and local shape variations, such as edge contours and surface textures, which generally exhibit strong local discriminability. In contrast, translation estimation focuses on the object's overall spatial position in the image, including the relative layout and scale features between the object and its environment. Unified modeling of both often fails to fully capture their respective feature representations and may even introduce mutual interference.

To address this, a decoupled modeling strategy is employed, designing separate rotation and translation branches for independent modeling and prediction of the two tasks, thereby fully leveraging the advantages of each feature dimension. In the decoder stage, the three-dimensional rotation vector and spatial translation vector are regressed separately, effectively enhancing the overall accuracy, stability, and generalization capability of 6D pose estimation.

4.1 Pose regression strategy

Rotation estimation relies on the object's intrinsic appearance and local shape variations, such as edge contours and surface textures, which generally possess strong local discriminability. Translation estimation, on the other hand, focuses on the object's overall spatial position in the image, including the relative layout and scale features between the object and its environment. Therefore, unified modeling of both often fails to fully capture their respective feature representations and may even introduce mutual interference. To address this, a decoupled modeling strategy is adopted, with separate rotation and translation branches designed for independent modeling and prediction of the two tasks.

To overcome the limitations of traditional rotation representations, this study introduces a six-dimensional rotation representation based on high-dimensional vectors [25]. Compared with conventional methods, the six-dimensional representation offers stronger continuity, a simpler form, and has been proven to achieve superior convergence performance and prediction accuracy in multiple pose estimation tasks.

Rotation is represented by two three-dimensional vectors, corresponding to two

orthogonal directions in the first three columns of the rotation matrix, thereby implicitly constructing the orthogonal structure of the rotation matrix. Specifically, the model regresses a six-dimensional vector $r_{6d}(r_1, r_2, r_3, r_4, r_5, r_6)$ through the rotation estimation head. From this, two three-dimensional vectors $v_1(r_1, r_2, r_3)$ and $v_2(r_4, r_5, r_6)$ are extracted, which serve as the basis to construct the standard orthonormal frame e_1, e_2, e_3, as formulated below:

$$e_1 = \Phi(v_1) \tag{1}$$

$$e_2 = \Phi(e_1 \times e_2) \tag{2}$$

$$e_3 = e_1 \times e_2 \tag{3}$$

Here, $\Phi(\cdot)$denotes vector normalization, and × represents the vector cross product. Using this orthonormal basis, the rotation matrix $\boldsymbol{R}$ is constructed as follows:

$$\boldsymbol{R} = [{e_1}^T, {e_2}^T, {e_3}^T] \tag{4}$$

To more accurately quantify the angular difference between the predicted and ground-truth rotations, the geodesic loss is employed as the optimization objective in the rotation estimation task, mathematically expressed as:

$$L_{\text{rot_6d}} = arccos\left\{\frac{tr\left(\boldsymbol{R}_{gt}^T R_{pred}\right) - 1}{2}\right\} \tag{5}$$

Here, $tr(\cdot)$ denotes the matrix trace, $\boldsymbol{R}_{gt}^T$ represents the ground-truth rotation matrix, and $\boldsymbol{R}_{pred}$ denotes the predicted rotation matrix.

For translation estimation, since convolutional neural networks (CNNs) are translation-invariant and excel at extracting local semantic features but have limited capability in modeling the absolute spatial position of objects in the input image, directly regressing the 3D translation vector often yields limited accuracy. Therefore, instead of end-to-end regression of the object's translation vector $\boldsymbol{T}(t_x, t_y, t_z)^{\mathrm{T}}$, a decoupled modeling strategy is adopted to construct a more reasonable translation prediction pipeline. Specifically, considering the clear perspective projection relationship between the object's 2D image projection (image center) and its actual 3D spatial position, the relationship can be modeled using the pinhole camera model, expressed as:

$$\boldsymbol{t}_z(p_x, p_y, 1)^T = \boldsymbol{K}(t_x, t_y, t_z)^T \tag{6}$$

Here, $\boldsymbol{K}$ denotes the camera intrinsic matrix, and (p_x, p_y) represents the projection coordinates of the object's 3D center on the image plane.

As illustrated in Fig. 4, the 3D translation T is first decoupled into the object's 2D projection coordinates (p_x, p_y) on the image and its corresponding depth t_z. To enhance the network's capability in modeling object positions, the 2D pixel coordinates are further decomposed into a reference point (c_x, c_y) and the offsets relative to this reference point $(\Delta x, \Delta y)$.

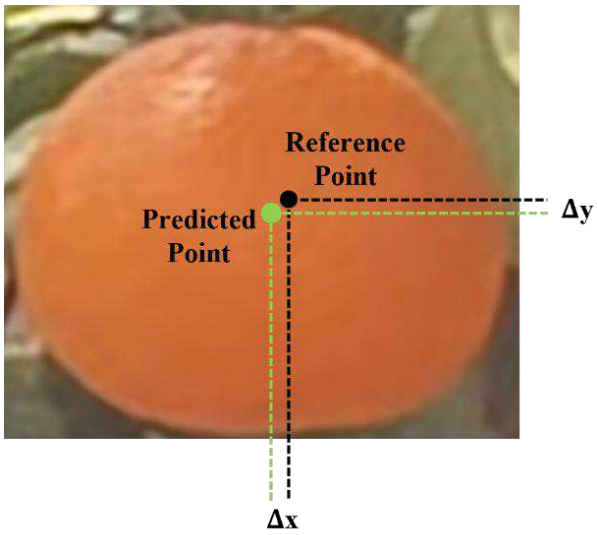

Fig.4 Illustration of translation prediction method

The translation branch of the network only needs to predict the two-dimensional offset of the target relative to the reference point and the target depth value, thereby obtaining the three-dimensional translation vector of the target. Combined with the camera intrinsic parameters, the translation vector of the target in three-dimensional space can be recovered through back-projection, which can be expressed as:

$$t_x = \frac{(c_x + \Delta x - p_x)}{f_x} \tag{7}$$

$$t_y = \frac{(c_y + \Delta y - p_y)}{f_y} \tag{8}$$

$$t_z = t_z \tag{9}$$

Here, f_x,f_y,p_x,p_y can be obtained from the camera's intrinsic parameters.

To effectively supervise the translation prediction, the L2 norm is adopted as the

loss function during the 2D projection coordinate prediction stage to measure the error between the predicted and ground-truth projection coordinates. Its mathematical expression is as follows:

$$L_{trans_1} = \left\|(t_x, t_y)_{pred}, (t_x, t_y)_{gt}\right\|_2 \tag{10}$$

Here, $(t_x, t_y)_{pred}$ and $(t_x, t_y)_{gt}$ denote the predicted and ground-truth 2D translation coordinates, respectively.

In the depth prediction stage, considering that the scene may contain a certain degree of occlusion, illumination variation, or noise interference, directly using the L2 loss could make the model overly sensitive to outliers, thereby affecting overall training stability. Therefore, the L1 absolute error is adopted as the loss function for depth estimation to enhance the model's robustness. Its mathematical expression is as follows:

$$L_{trans_2} = \left\|t_{z_{pred}}, t_{z_{gt}}\right\|_1 \tag{11}$$

Here, $t_{z_{pred}}$ and $t_{z_{gt}}$ denote the predicted and ground-truth depth values of the target, respectively.

4.2 Pose estimation regression head

When integrating the above strategy into the network architecture, the model is required to regress both the six-dimensional rotation representation r_{6D} and the translation parameters $\boldsymbol{t} = (t_x, t_y, t_z)$.

Inspired by EfficientPose [26], this study designs an iterative refinement strategy to progressively approach the target's final 6D pose, aiming to improve the accuracy and stability of pose estimation. This strategy performs multiple refinement regressions by introducing feature residuals based on the initial prediction, achieving more precise pose estimates. The rotation estimation head shares the same network structure as the translation estimation head, as illustrated in Fig. 5. Taking the rotation prediction head as an example, the method involves the following three steps: First, after obtaining the global feature representations $F_i \in \mathbb{R}^{C\times H\times W}, i \in \{1,2,3\}$ from the Transformer encoder, the model performs an initial rotation pose estimation to obtain the preliminary

rotation feature r_{pre}. Specifically, the spatial coordinates (x, y) of F_i are first normalized and then mapped to the range$[-1,1]$, generating two coordinate feature channels containing spatial position information. Subsequently, this positional information is concatenated with the encoder's global feature F_i along the channel dimension to integrate spatial structure with semantic information.

The fused features are sequentially processed through a standard convolution layer followed by three depthwise separable convolution layers to progressively extract more discriminative pose-related features. After each convolution layer, group normalization and the Swish activation function are applied to enhance the model's nonlinear representation capability and training stability. The final output, the preliminary rotation feature representation $\boldsymbol{r}_{pre}$, serves as the input for subsequent refinement and prediction, computed as follows:

$$r_{pre} = DSConv^3\left(Conv\left(C\left(F_i, coord(F_i)\right)\right)\right) \tag{12}$$

Here, $DSConv(\cdot)$ denotes the depthwise separable convolution, $Conv(\cdot)$ denotes the standard convolution, $C(\cdot)$ represents the concatenation operation, and $coord(\cdot)$ indicates the spatial coordinate feature extraction operation.

Then, an additional depthwise separable convolution layer is used to compress the feature channels to 6, producing the initial rotation prediction vector $\boldsymbol{r}_{init}$. The process can be expressed as:

$$r_{init} = DSConv\left(r_{pre}\right) \tag{13}$$

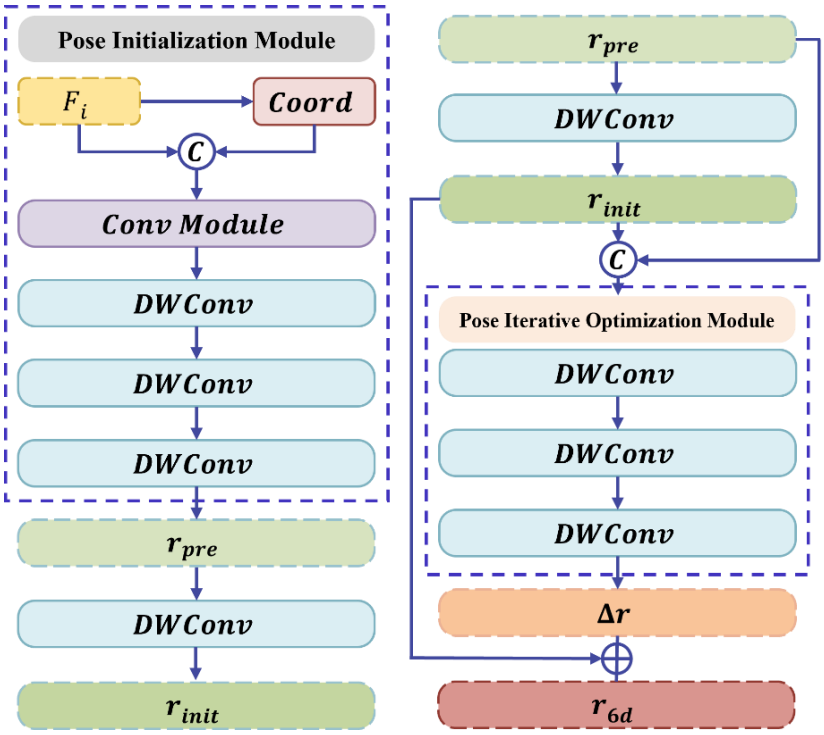

Fig.5 Network structure of pose estimation regression head

Finally, the initial prediction vector r_{pre} and the initial rotation estimate r_{init} are concatenated along the channel dimension and fed into the pose iterative refinement module. This module consists of two depthwise separable convolution layers, each followed by group normalization and the Swish activation function to enhance nonlinear representation and extract higher-order rotation features. The features are then passed through an additional depthwise separable convolution layer for channel compression, producing the residual vector of the rotation feature Δr, which is used to further refine the rotation estimation:

$$\Delta r = DSConv^3\left(C\left(r_{pre}, r_{init}\right)\right) \tag{14}$$

The residual vector of the rotation feature Δr is added to the initial rotation estimate r_{init} to obtain the final optimized 6D rotation prediction $\boldsymbol{r}_{6d}$.

For translation estimation, to further improve prediction accuracy, the reference point coordinates are not simply used as the object's centroid. Instead, a more flexible modeling strategy is adopted. Specifically, the model introduces two independent estimation branches to predict the object's 2D pixel offsets $(\Delta x, \Delta y)$ and depth value t_z, enabling decoupled modeling of the 3D translation vector. Each branch shares the

same structure as the rotation estimation head, combining standard and depthwise separable convolutions, along with group normalization and activation functions for feature refinement. This design allows the model to more accurately capture the object's spatial position variations and maintain robust translation estimation performance even in orchard scenarios with weak textures or complex backgrounds.

5. Experimental design and results analysis

To evaluate the performance of the proposed framework, extensive experiments were conducted on the publicly available YCB-Video benchmark dataset and the self-collected Orange-Pose dataset, assessing the method under conditions without depth images. Training was performed on an Ubuntu 20.04 system using an NVIDIA Ampere GPU, with PyTorch v1.10 and CUDA 10.2. The YCB-Video dataset is a widely used benchmark for 6D pose estimation, featuring complex lighting, severe occlusions, and multiple overlapping objects, making it suitable for testing algorithm robustness. The experiments include comparisons with existing methods and ablation studies on the contributions of various model components. Additionally, real-world picking experiments were conducted using a self-built intelligent orchard harvesting robot, and the results were summarized and analyzed.

5.1 Experimental platform

An intelligent harvesting robot hardware experimental platform was established for data collection, visual perception, and robotic arm picking tests. The platform is equipped with two RGB-D cameras and two robotic arm picking systems, with the gripper and scissors independently designed and 3D-printed. Hand-eye calibration between the robotic arms and cameras has been completed in the laboratory.

5.2 Ablation study

To validate the effectiveness of each module in this study, ablation experiments were first conducted on the YCB-V dataset. Different model variants were designed targeting key components of the pose estimation framework. The specific model

variants are as follows:

Baseline：The complete pose estimation network, including object detection, the deformable Transformer-based feature enhancement module, and the pose prediction head.

Trans-A：Based on the Baseline model, the bounding box center is removed as the Transformer reference point and replaced with a self-learned reference point by the Transformer.

Trans-B：Building on Trans-A, the bounding box information is further removed from the Transformer query embedding, using query embeddings self-learned by the Transformer network.

Trans-C：Based on the Baseline model, the Transformer feature enhancement module is removed, and the pose regression is performed directly with the prediction head, without introducing the Transformer structure.

Evaluation metrics include the commonly used ADD and ADD-S metrics for 6D pose estimation and the area under the accuracy curve (AUC) for quantitative assessment. For asymmetric objects, the ADD metric is used: the predicted pose is applied to the object model point cloud, and the mean distance error is calculated between corresponding points under the ground-truth pose. For symmetric objects, the improved ADD-S metric is adopted: for each point in the predicted model, the nearest distance to the set of points in the ground-truth model is computed, and the average over all points is taken as the error. Accuracy is computed and plotted, with a larger AUC indicating higher overall precision. The detailed performance comparison of the ablation experiments is presented in Table 1.

Tab.1 Results of ablation experiments

Model	AUC of ADD-S	AUC of ADD	Avg. T. Error [cm]	Avg. Rot. Error [°]
Baseline(Ours)	**92.4**	**80.3**	**2.15**	**24.03**
Trans-A	**87.3**	**66.5**	**2.78**	**38.31**
Trans-B	**86.9**	**64.3**	**3.62**	**36.39**
Trans-C	**61.7**	**50.9**	**5.21**	**36.98**

Comparing the Baseline model with Trans-A, it is observed that using the bounding box center as the reference point significantly improves pose estimation performance, particularly increasing the AUC of the ADD metric by 13.8%. This indicates that using the bounding box center effectively helps the Transformer network focus more accurately on the target region. Next, comparing Trans-A with Trans-B, the performance difference is minimal, suggesting that bounding box prior information has little impact on improving pose estimation accuracy. However, during training, it was noted that under identical conditions, Trans-B requires significantly longer training time than Trans-A. Introducing bounding box priors as query embeddings accelerates model convergence.

Comparing the Baseline model with Trans-C, the ADD and ADD-S AUC values for Trans-C dropped sharply by 30.7% and 29.4%, respectively, with translation and rotation errors increasing by 3.86 cm and 12.95°, confirming that the deformable Transformer-based feature enhancement module effectively integrates local features and global information to improve pose estimation accuracy.

Using the bounding box center as the reference point provides the Transformer with explicit spatial priors, guiding the attention mechanism to focus on the target region and reducing sensitivity to background interference. This explains why the Baseline model shows a significant improvement over Trans-A in ADD performance. In the comparison between Trans-A and Trans-B, although explicit bounding box embedding does not lead to notable precision gains, it accelerates convergence during training, indicating that this module primarily enhances training stability and efficiency rather than final prediction accuracy. Furthermore, comparing Trans-C with the Baseline shows a substantial performance drop after removing the deformable Transformer, particularly in translation and rotation errors, demonstrating that the Transformer enhancement module plays a crucial role in modeling global semantics and multi-scale feature interactions, effectively mitigating feature loss under occlusion and weak texture conditions.

Overall, the proposed pose estimation method based on bounding box priors and deformable Transformers achieves higher pose prediction accuracy while maintaining

computational efficiency, offering improved stability and generalization, and is suitable for multi-object 6D pose estimation tasks in complex scenarios.

5.3 Comparison with classical methods

To further evaluate the performance of the pose estimation method, comparative experiments were conducted under the same training environment and hyperparameter settings with current mainstream RGB-only pose estimation models, including PoseCNN, SilhoNet, and GDR-Net. The comparison results are shown in Table 2.

Compared with other methods, the proposed approach achieves better performance in position error, indicating its ability to more accurately predict the 3D location of the target. However, in terms of rotation error, the method is slightly inferior to SilhoNet, primarily because SilhoNet emphasizes leveraging the geometric symmetry of objects to reduce rotation estimation impact, resulting in more stable rotation predictions. The algorithm achieves an average position error of approximately 2.15 cm and a rotation error of 24.03°. While this rotation error may be relatively high for high-precision grasping tasks, its impact on orchard harvesting tasks is relatively minor, as fruits naturally grow mostly in a downward vertical orientation, reducing reliance on rotation information. The significant advantage of the proposed method in position estimation enhances its practical value for orchard harvesting applications.

Tab.2 Comparative experimental results of mainstream pose estimation methods

Method	Translation Error[cm]				Rotation Error[degree]			
	PoseCNN	SilhoNet	GDR-Net	Ours	PoseCNN	SilhoNet	GDR-Net	Ours
master chef can	3.29	3.02	5.40	2.36	50.70	1.21	48.16	75.34
cracker box	4.02	5.24	5.09	3.04	19.69	19.86	25.30	20.56
sugar box	3.06	2.10	5.23	1.52	9.29	12.28	20.10	5.40
soup can	3.02	2.40	6.60	1.52	18.23	1.91	30.23	45.34
mustard bottle	1.72	1.65	2.61	1.89	9.94	5.78	9.16	26.45
bowl	5.23	3.30	3.81	1.66	75.53	49.95	69.60	24.79
mug	4.00	2.61	2.77	1.85	19.44	18.14	7.75	22.67
power drill	4.59	6.77	4.80	2.39	9.91	30.54	29.27	5.23
wood block	6.34	5.59	6.33	4.65	23.63	25.52	79.75	6.04

Method	Translation Error[cm]				Rotation Error[degree]			
	PoseCNN	SilhoNet	GDR-Net	Ours	PoseCNN	SilhoNet	GDR-Net	Ours
scissors	6.40	9.91	3.69	3.72	43.98	155.53	87.35	15.78
large marker	3.89	3.24	4.45	2.85	92.44	10.44	80.28	23.67
large clamp	9.79	6.27	7.46	2.23	38.12	3.54	12.01	4.97
extra lamp	8.36	4.86	5.83	3.10	34.18	29.18	52.39	24.67
foam brick	2.48	3.98	2.36	3.52	22.67	13.84	4.41	32.35
tuna fish can	2.41	1.57	5.59	1.84	32.80	1.46	30.53	61.27
pudding box	3.69	7.15	1.78	1.51	10.20	20.95	6.74	6.34
gelatin box	2.49	1.09	1.67	1.65	5.25	12.52	7.88	6.45
potted meat can	3.65	4.30	5.14	1.75	28.67	7.27	18.24	8.93
banana	2.43	4.12	2.02	2.05	15.48	16.29	9.16	22.67
pitcher base	4.43	1.31	4.89	1.55	11.98	6.64	30.17	7.04
bleach cleanser	4.86	3.60	7.52	2.57	20.85	51.28	43.85	22.86

5.4 Real-World harvesting experiments

To evaluate the algorithm's performance in real-world orchard harvesting, comparative experiments were conducted using an actual dual-arm harvesting robot equipped with the proposed algorithm, as shown in Fig. 6. Each robotic arm is equipped with an RGB-D camera and independently runs the algorithm. The performance was compared with benchmark methods SilhoNet and GDR-Net. For each method, 30 picking attempts were conducted. Mature oranges were manually suspended at the same positions to ensure that each method faced identical initial conditions, with the robotic arms able to reach and grasp the fruit. For each attempt, the detection success (whether the fruit was recognized and its pose estimated) and the picking success (whether the gripper cut the stem and retrieved the fruit) were recorded.



Fig.6 Intelligent dual-arm picking robot

The experimental results are shown in Table 3. As observed, the proposed method successfully detected the target in all 30 attempts, significantly outperforming GDR-Net and SilhoNet, indicating a high detection accuracy. This success is attributed to the cascade design of our method, which allows integration of any pre-trained object detection network and enables joint optimization of detection and pose estimation.

Tab.3 Comparative results of picking experiments

Method	Number of Successful Detections	Detection Success Rate	Number of Successful Pickings	Picking Success Rate
SilhoNet	27	90%	8	26.67%
GDR-Net	29	96.67%	14	46.67%
Ours	**30**	**100%**	**17**	**56.67%**

In terms of picking success, the proposed method also performs well, achieving a success rate of 56.67%, far exceeding the other two methods. These results demonstrate that our method provides higher localization accuracy, significantly improving the gripper's success rate in aligning with the fruit stem. The algorithm maintains strong stability under complex backgrounds and varying lighting conditions, offering reliable support for outdoor harvesting tasks.

5.5 Visualization of pose estimation results

To provide a more intuitive demonstration of the pose estimation performance of the proposed method on the YCB-V dataset, the predicted 3D object poses were projected onto the 2D image plane, as shown in Fig. 7. The figure presents four

representative complex scenarios, including object stacking and occlusion, showing the pose predictions of the proposed method, the predictions from GDR-Net, and the ground-truth poses.



Fig.7 Visualization comparison of pose estimation on YCB-V dataset

From the figure, it can be observed that for objects such as the “large clamp,” “foam brick,” and “cookie box,” GDR-Net exhibits noticeable deviations in pose estimation, primarily due to its weaker regression capability under occlusion. In contrast, the proposed method maintains high pose estimation accuracy across various complex scenarios, demonstrating strong robustness and generalization ability, further validating its effectiveness in practical applications.

Figure 8 illustrates the pose estimation results of the proposed method on the self-collected Orange-Pose dataset. To visually represent the predicted orange poses, the estimated results were projected onto the 2D image plane. Due to the lack of a standard 3D model for oranges, the target object’s spatial pose is simplified using a cuboid representation. In this representation, the plane defined by the red wireframe corresponds to the normal plane of the Z-axis, while the yellow and purple wireframes form the normal plane of the Y-axis. The X-axis direction is determined according to the right-hand rule.



Fig.8 Visualization of pose estimation on Orange-Pose dataset

6. Conclusion

For outdoor orchard harvesting scenarios with missing depth information, this paper proposes a 6D fruit pose estimation method tailored for depth-deficient environments. The approach constructs an end-to-end cascade architecture that organically combines a lightweight object detection network with a deformable Transformer-based feature enhancement module, leveraging bounding box information to provide explicit spatial priors and significantly improving the model's ability to capture key regions and spatial context. Building on this, a rotation–translation decoupled regression structure is designed, coupled with an iterative refinement mechanism, effectively addressing issues of pose regression accuracy and convergence instability under depthless conditions. Experimental results on the YCB-Video and the self-collected Orange-Pose datasets demonstrate the method's excellent performance, providing stable and accurate pose estimates even in the absence of depth information. Furthermore, real-world harvesting experiments confirm that the approach enhances robotic grasping success rates, meeting the engineering requirements of orchard harvesting tasks. This work provides a feasible and efficient technical pathway for high-precision, robust fruit pose estimation in complex agricultural environments lacking depth data, laying a solid perceptual foundation for the intelligent development of agricultural harvesting robots.

References

[1] Zhao, C., Fan, B., Li, J., et al. Progress, challenges, and trends of agricultural robotics [J]. Smart Agriculture , 2023, 5(4): 1–15. DOI:10.12133/j.smartag.SA202312030

[2] Arad, B., Balendonck, J., Barth, R., Ben-Shahar, O., Edan, Y., Hellström, T., Hemming, J., Kurtser, P., Ringdahl, O., Tielen, T., & van Tuijl, B. Development of a sweet pepper harvesting robot. Journal of Field Robotics, 37(6), 1027–1039, 2020. https://doi.org/10.1002/rob.21937

[3] Bac, C. W., Hemming, J., Henten, E. J. (2014). Harvesting robots for high-value crops: State-of-the-art review and challenges ahead. Journal of Field Robotics, 31(6), 888–911.

[4] Zhang, K., Lammers, K., Chu, P., Li, Z., & Lu, R. An automated apple harvesting robot—From system design to field evaluation. Journal of Field Robotics, 41(7), 2384–2400, 2024. https://doi.org/10.1002/rob.22268

[5] Li, S., Sun, W., Liang, Q., Liu, C., & Liu, J. Assessing fruit hardness in robot hands using electric gripper actuators with tactile sensors [J]. Sensors and Actuators A: Physical, 2024, 365: 114843. DOI:10.1016/j.sna.2023.114843

[6] Rybak, L., Carbone, G., Mohan, S., Gaponenko, E., Malyshev, D., & Voloshkin, A. New design and construction of a mechanical gripping device with a telescopic link of a fruit harvesting robot [J]. Robotica, 2025, 43(1): 69-85. DOI:10.1017/S0263574724001127

[7] Kim, J. Y., Pyo, H. R., Jang, I., Kang, J., Ju, B. K., & Ko, K. E. (2022). Tomato harvesting robotic system based on Deep-ToMaToS: Deep learning network using transformation loss for 6D pose estimation of maturity classified tomatoes with side-stem. Computers and Electronics in Agriculture, 201, 107300.

[8] Sun, Q., Zhong, M., Chai, X., Zeng, Z., Yin, H., Zhou, G., & Sun, T. (2023). Citrus pose estimation from an RGB image for automated harvesting. Computers and Electronics in Agriculture, 211, 108022.

[9] C. Wang et al., "DenseFusion: 6D Object Pose Estimation by Iterative Dense Fusion," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019, pp. 3338-3347, doi: 10.1109/CVPR.2019.00346.

[10] Zhang, H., Miao, Y., Chen, Q., Li, Y., Bai, S., & Park, G. Monocular vision-based kinematic calibration for a 6-DoF serial robot using perspective phase-based pose estimation [J]. Measurement: Journal of the International Measurement Confederation, 2026, 263: 120188. DOI:10.1016/j.measurement.2025.120188

[11] Yang, B., Cen, X., Xie, L., & Yin, M. Automatic pose measurement of robotic drilling system based on zoom monocular vision [J]. Advanced Engineering Informatics, 2025, 65: 103121. DOI:10.1016/j.aei.2025.103121

[12] Calera, E. S., Oliveira, G. C. D., Araujo, G. L., Filho, J. I. F., Toschi, L., Hernandes, A. C., Velasquez, A. E. B., Gasparino, M. V., Chowdhary, G., Higuti, V. A. H., & Becker, M. Under-canopy navigation for an agricultural rover based on image data [J]. Journal of Intelligent & Robotic Systems, 2023, 108(2): 29. DOI:10.1007/s10846-023-01849-8

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30

[14] Wu, W., Huang, C., Shen, Y., & Ma, X. P2P-Pose: point to point relative camera pose regression with image matching. Measurement Science and Technology, 37(5), 056302, 2026. https://doi.org/10.1088/1361-6501/ae3d58

[15] Tang, H., Liu, C., Wang, Q., & Hu, W. Monocular-vision-based non-cooperative spacecraft tracking for close-proximity missions. Measurement Science and Technology, 36(11), 116310, 2025. https://doi.org/10.1088/1361-6501/ae1a06

[16] XIANG Y, SCHMIDT T, NARAYANAN V, et al. PoseCNN: A Convolutional Neural Network for 6D Object Pose Estimation in Cluttered Scenes[C/OL]//Robotics: Science and Systems XIV. 2018.

[17] Y. Li, G. Wang, X. Ji, Y. Xiang, and D. Fox. Deepim: Deep iterative matching for 6d pose estimation. In Proceedings of the European Conference on Computer Vision (ECCV), pages 683–698, 2018.

[18] W. Kehl, F. Manhardt, F. Tombari, S. Ilic, and N. Navab. Ssd-6d: Making rgb-based 3d detection and 6d pose estimation great again. In Proceedings of the IEEE International Conference on Computer Vision, pages 1521–1529, 2017.

[19] WANG G, MANHARDT F, TOMBARI F, et al. GDR-Net: Geometry-Guided Direct Regression Network for Monocular 6D Object Pose Estimation[C/OL]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA. 2021.

[20] Y. Labb´e, J. Carpentier, M. Aubry, and J. Sivic. Cosypose: Consistent multi-view multi-object 6d pose estimation. In European Conference on Computer Vision, pages 574–591. Springer, 2020.

[21] G. Billings and M. Johnson-Roberson. Silhonet: An rgb method for 6d object pose estimation. IEEE Robotics and Automation Letters, 4(4):3727–3734, 2019.

[22] A. Amini, A. S. Periyasamy, and S. Behnke. T6d-direct: Transformers for multi-object 6d pose direct regression. In DAGM German Conference on Pattern Recognition, pages 530–544. Springer, 2021.

[23] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, J. Dai. Deformable DETR: Deformable Transformers for End-to-End Object Detection[EB/OL]. arXiv, 2021

[24] CARION N, MASSA F, SYNNAEVE G, et al. End-to-End Object Detection with Transformers[C]//Computer Vision – ECCV 2020, 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I. Cham, Switzerland: Springer, 2020: 213–229.

[25] Y. Zhou, C. Barnes, J. Lu, J. Yang, H. Li. On the continuity of rotation representations in neural networks[C], in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019: 5745-5753

[26] Y. Bukschat, M. Vetter. EfficientPose: An efficient, accurate and scalable end-to-end 6D multi object pose estimation approach[EB/OL]. arXiv, 2020.

Statements and Declarations

Funding

This work was supported by the Hubei Province Core Technology Application Research Project for Agricultural Machinery Equipment (No. HBSNYT202213).

Competing interests

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.

Author contributions

Gang Peng provided the research ideas and resources. Jiaqi Yang edited and checked the manuscript. Chen Lai implemented and completed the work.

Data availability

The data and materials supporting the findings of this study are available from the corresponding author upon reasonable request.

AI assistance

OpenAI Codex was used to assist with manuscript formatting and the preparation of submission materials. The authors are responsible for the content of the final submitted manuscript.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- video.mp4