Rosetum3D: A Large-Scale 3D Vision Dataset from Preharvest Roses

Songyan Liu, Xipei Liu, Yujie Liu, Jiali Wu, and Xiaofei Liu*

School of Information Science & Engineering, Yunnan University, China

Abstract. The global rose cultivation industry has experienced continued expansion in recent years. There has been a set of value evaluation criteria covering multiple dimensions for roses. However, the evaluations are mainly conducted by human experts after harvesting, which is time-consuming and labor-intensive. 3D computer vision technologies are promising to automate the process before harvesting, but the lack of large-scale datasets with localization annotations that capture plant architecture hinders the progress. To bridge this gap, we propose **Rosetum3D**, which is a large-scale 3D vision dataset for preharvest roses. The dataset is constructed via occlusion-robust multi-view RGB-D capture protocols in commercial greenhouses. We provide fine-grained 2D localization labels using bounding boxes and botanically defined keypoints, and then obtain the 3D structures recovered through depth backprojection. Rosetum3D contains 21,114 images and 46,848 annotated rose objects. Models trained on Rosetum3D have achieved 2D/3D rose localization, which is a crucial step for automated preharvest quality grading and growth monitoring. Beyond localization, Rosetum3D serves as a benchmark for agricultural vision tasks, including 2D rose object detection, local feature matching, depth estimation, and instance reidentification. By enabling data-driven precision agriculture, Rosetum3D paves the way for robotic harvesting systems and AI-driven yield prediction in protected cultivation. The dataset are available at `https://huggingface.co/datasets/WaterMelon2333/Rosetum3D/tree/main`.

Keywords: Rosetum3D · Preharvest rose localization · Dataset for computer vision tasks

1 Introduction

The global rose industry has witnessed substantial expansion in recent years, precipitated by escalating demand for both ornamental and commercial applications. Nowadays, there has been a set of value evaluation criteria covering multiple dimensions for roses, including the number of petals, the shape of the flower, the length of the stem, etc. Meanwhile, the evaluation of these criteria is predominantly conducted by human experts after harvesting, which requires

* Corresponding author: lxfjl2016@163.com

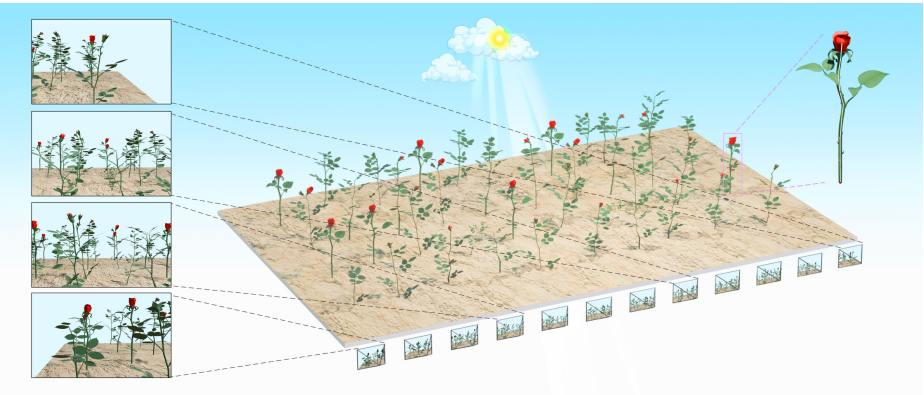

Fig. 1: Illustration of the collection procedure of Rosetum3D. The data collectors walk parallel to the plant rows at a constant speed and take RGB-D sequences of the preharvest roses. After collection, the annotators mark the bounding boxes and keypoints on the 2D images, and the annotations are backprojected to the 3D space to compute the 3D coordinates. The 3D keypoints are shown in the top-right corner.

more time and manpower. As a result, it is valuable to automate the process before harvesting, and 3D computer vision technologies provide a promising solution. However, existing flower datasets mainly focus on variety classification and 2D object detection [6,30], which is helpless for locating the preharvest roses in 3D space and measuring the length of the stem.

To address this gap, this paper introduces **Rosetum3D**, a dataset curated using structured-light RGB-D cameras deployed in operational rose greenhouses, the data collection process is as illustrated in Fig. 1 and some examples of images are shown in the Supplementary Materials. Structured-light sensors are prioritized over alternative 3D sensing modalities, such as LiDAR, owing to their cost-effectiveness and robustness under typical greenhouse lighting conditions. Multi-view RGB-D sequences are captured across a heterogeneous range of rose varieties and growth stages, ensuring comprehensive representation of variations in plant geometry and occlusion patterns.

Despite significant advances in 3D vision, existing public datasets, including KITTI [19], nuScenes [2], ScanNet [11], and SUN RGB-D [39], predominantly target autonomous driving or indoor scenes. In these contexts, LiDAR or photogrammetric 3D reconstructions facilitate direct annotations within point clouds or triangular meshes. Conversely, these methodologies are inadequate within rose cultivation settings as a consequence of severe inter-plant occlusions, the morphological fragility of petals prone to deformation, and highly variable lighting conditions. Consequently, the semantic annotations for Rosetum3D are produced via a hybrid 2D-to-3D pipeline. Initially, 2D bounding boxes and semantic keypoints, encompassing the corolla center, sepal-pedicel junction, stem-pedicel transition, and soil-emergence point, are manually labeled on RGB images. These 2D primitives are subsequently back-projected into 3D space uti-

lizing depth maps derived from synchronized RGB-D frames. This methodology circumvents the requirement for error-prone 3D reconstruction while maintaining geometric consistency across multiple viewpoints.

Parallel to the introduction of the Rosetum3D dataset, this study benchmarks existing computer vision models across several tasks, including 2D object detection, 2D keypoint localization, monocular depth estimation, local feature matching, and instance re-identification. Leveraging these tasks, this work proposes a novel model for 3D rose localization, which establishes a robust baseline for future research.

In summary, the contributions of this work are as follows:

- The introduction of Rosetum3D, the first large-scale 3D visual dataset specifically developed for rose cultivation scenarios, featuring synchronized multi-view RGB-D sequences and semantic 2D/3D labels (including bounding boxes and keypoints) for unharvested roses.
- A comprehensive empirical evaluation demonstrating that Rosetum3D establishes a novel benchmark for various vision tasks, spanning 2D object detection, 2D keypoint localization, local feature matching, plant re-identification, and depth estimation within occluded agricultural environments.
- The development of a novel method for identifying the 3D location of preharvested roses, leveraging fundamental vision models.

2 Related Work

2.1 3D Vision Dataset

The acquisition of accurate depth supervision remains a cornerstone of 3D computer vision, given that monocular images inherently lack intrinsic depth information. Consequently, 3D datasets typically integrate specialized sensors, such as LiDAR and RGB-D cameras, to capture high-fidelity ground-truth geometry. This reliance on specific sensors has precipitated the development of domain-specific dataset paradigms.

Indoor scene datasets primarily leverage RGB-D sensors to facilitate dense depth estimation. The NYU Depth Dataset V2 [38] provides aligned RGB-D frames, facilitating 3D segmentation and object recognition. Building upon these efforts, the SUN RGB-D Dataset [39] comprises over 10,000 RGB-D images annotated with 146,617 2D polygons and 58,657 3D bounding boxes, thereby supporting the classification, detection, and semantic segmentation of indoor scenes. ScanNet [11] represented a significant advancement in indoor reconstruction, featuring over 1,500 scenes and 2.5 million views, which support tasks such as camera pose estimation, surface reconstruction, and semantic instance segmentation. Its enhanced successor, ScanNet++ [47], provides further improvements in data quality and visual diversity, optimizing it for novel view synthesis and holistic 3D understanding. Additionally, complementary efforts include motion-capture-driven human datasets, such as Human3.6M [22], which provide precise 3D keypoints for articulated poses.

Outdoor scenes, particularly those focused on autonomous driving, predominantly utilize LiDAR for long-range depth acquisition. The seminal KITTI dataset [19], featuring multi-sensor data (stereo cameras, LiDAR, GPS) collected from vehicle-mounted platforms, functions as a standard benchmark for stereo matching, optical flow, odometry, and integrated 2D/3D object detection and tracking. Despite KITTI's foundational role, subsequent large-scale datasets have emerged; specifically, nuScenes [2,18], comprises 1,000 driving scenes across Boston and Singapore, incorporating 23-class 3D bounding boxes. Furthermore, PandaSet [43] provides over 48,000 camera images and 16,000 LiDAR sweeps with comprehensive 28-class detection and 37-class segmentation labels.

2.2 3D Vision in Agriculture

Contrasting with general indoor and outdoor environments, agricultural applications encounter unique challenges, namely small object scales, severe occlusions, and variable lighting, which necessitate diverse sensor strategies.

There have been several **3D vision datasets in agriculture scenes**. For example, Plant3D [7–9] contains a total of 714 3D laser scans of tomato, tobacco, sorghum, and Arabidopsis plants obtained within a 20-to-30-day developmental window. However, these 3D models lack chromatic information. The ROSE-X dataset [14] comprises 3D models of 11 rosebush plants acquired through X-ray computed tomography. Within these volumetric models, the voxels are categorized into three semantic classes: leaf, stem, and flower. Pheno4D [35] constitutes a dataset of 3D point clouds representing seven maize plants and seven tomato plants. These plants are scanned using a laser scanner at various growth stages, yielding a total of 244 point clouds. Of these, 126 specimens are manually annotated with semantic and instance-level labels. The Soybean-MVS dataset [41] diverges significantly from previous works in terms of data acquisition modality. The plants are imaged using an RGB camera within a controlled setup, subsequently generating point clouds through multi-view stereo (MVS) reconstruction. The Crops3D dataset [52] integrates terrestrial laser scanning with structured light, fused via structure-from-motion (SfM) and multi-view stereo (MVS), to reconstruct eight crop types under field conditions.

When it comes to the **recognition models**, Xiao et al. [44] utilize drone-based cross-orbit imaging to reconstruct organ-scale 3D models of crops spanning sugar beet and wheat. PlantGaussian [37] utilizes 3D Gaussian splatting for the spatio-temporal reconstruction of wheat and tobacco within both indoor and outdoor settings.

Despite these advances, large-scale 3D datasets remain scarce in the agricultural sector compared to general indoor and outdoor domains. Existing datasets and methodologies often focus on small-scale reconstruction or demonstrate limited efficacy in environments characterized by significant occlusion. Critically, there is a notable absence of large-scale, dedicated datasets for floriculture applications, particularly rose cultivation, where intricate botanical structures and dense foliage necessitate specialized 3D representation. This study addresses this

lacuna by introducing a large-scale, occlusion-robust dataset tailored to rose phenotyping and robotic harvesting.

3 Rosetum3D Dataset

This section provides a comprehensive overview of the Rosetum3D dataset, focusing on its construction, specifically the data collection protocol, annotation methodology, and associated statistics.

3.1 Data Collection

Data Collection Equipment The Orbbec Gemini 335L [31] or 336L [32] are employed to collect RGB-D sequences; these sensing units are stereo vision 3D cameras that integrate active and passive stereo technologies to facilitate operation under diverse indoor and outdoor conditions. The sensor is mounted to the chest via a shoulder strap and interfaced with an Android smartphone or tablet computer, as shown in Fig. 2, enabling synchronized depth and color acquisition at a frequency of 30 Hz. Both depth and color frames are captured at a resolution of 1280×800 pixels. Automatic white balance and auto exposure are enabled by default.

Calibration Accurate camera calibration is essential to reproduce the 3D locations of the roses reliably. Following the protocol established in ScanNet [11], a checkerboard pattern is utilized for calibration. The operator records an RGB-D sequence of the pattern placed on a large, flat surface, capturing views across various distances. Subsequently, a calibration procedure derived from existing methods [13, 42], is executed to obtain intrinsic parameters for both depth and color sensors, along with the extrinsic transformation matrix aligning depth to color.

Data Collection RGB-D data collection is conducted in operational rose greenhouses, adhering to a systematic protocol to ensure temporal consistency and the coverage of diverse growth stages. As illustrated in Fig. 1, operators maintain a trajectory parallel to the plant rows at a constant speed of 0.3–0.5 m/s, preserving a fixed distance of 0.8–1.2 m from the roses during data acquisition. The camera orientation is fixed at a tilt angle of downward from the horizontal plane to optimize the capture of both upper floral regions and lower stem structures.

3.2 Data Annotation

The annotation pipeline utilizes a structured approach modeled after human pose estimation keypoint frameworks, refined to characterize the specific morphology of preharvested roses. Each rose instance is subjected to manual 2D annotation followed by automated 3D reconstruction.

Fig. 2: Illustration of the data collection equipment. We fix the RGB-D sensor to the chest with a shoulder strap and attach it to an Android smartphone.

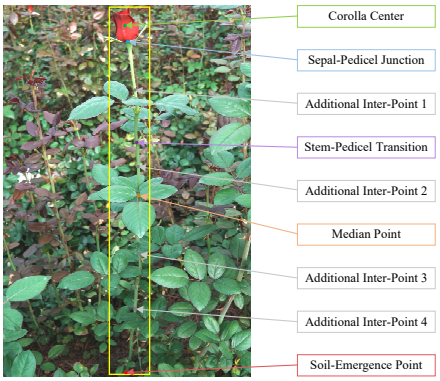

Fig. 3: Illustration of 2D Annotation. The main 5 points are as described in the main body. Besides those 5 points, we design 4 additional inter-points as follows: *Additional Inter-Point 1* is the midpoint between *Sepal-Pedicel Junction* and *Stem-Pedicel Transition*; *Additional Inter-Point 2* is the midpoint between *Stem-Pedicel Transition* and *Median Point*; *Additional Inter-Point 3* and *Additional Inter-Point 4* are the upper and lower third points between *Median Point* and *Soil-Emergence Point*.

2D Annotation Annotators initially define a bounding box enclosing the complete rose specimen within the RGB frames. Subsequently, the following five biologically significant keypoints are manually annotated:

- **Corolla Center**: The geometric centroid of the floral corolla, identifying the visual center of the flower.
- **Sepal-Pedicel Junction**: The specific point where the basal whorl of sepals converges at the proximal terminus of the pedicel, marking the flower-to-stem connection.
- **Stem-Pedicel Transition**: The locus where the primary stem diameter undergoes a significant reduction, transitioning to the flower stalk.
- **Median Point**: The spatial midpoint along the vertical axis of the complete above-ground botanical structure.
- **Soil-Emergence Point**: The precise coordinate where the plant stem initially penetrates the soil surface, also referred to as the crown interface.

In addition to these five primary landmarks, auxiliary points are annotated between distant keypoints to facilitate reliable spatial interpolation, as detailed in Fig. 3. During the annotation procedure, visually occluded points are flagged with an "invisible" attribute; consequently, such points are excluded from subsequent accuracy evaluations. Furthermore, rose objects with dimensions smaller

Table 1: Data Statistics of Annotation Validation.

Error(cm)	0.00 - 0.15	0.15 - 0.30	0.30 - 0.45
Num(instance)	85	45	7

than 32×32 pixels are excluded from annotation owing to the impracticality of keypoint measurement at minor scales, adhering to the established data quality protocol.

3D Reconstruction Utilizing synchronized depth maps from RGB-D frames, all 2D keypoints are projected back into 3D camera coordinates via standard perspective geometry:

$$\mathbf{P}_{3D} = \mathbf{K}^{-1} \cdot [u, v, 1]^T \cdot D(u, v), \tag{1}$$

where $\mathbf{K}$ denotes the camera intrinsic matrix, (u, v) the 2D keypoint position, and $D(u, v)$ the corresponding depth value.

Annotation Validation To verify that the inter-keypoint distances of rose stems, derived from 2D annotations and depth backprojection, correspond to physical measurements, supplementary RGB-D sequences are captured during data acquisition. Each sequence featured a calibration tape measure positioned parallel to the primary stem axis at distances ranging from 0.5 to 1.0 m from the camera, guaranteeing the clear visibility of metric markings. Representative examples of these RGB images are provided in Fig. 4. Adhering to the standard annotation protocol, two precise 10 cm interval ticks are annotated on each tape. The 3D Euclidean distance between these coordinates is computed via depth-based backprojection. Results demonstrate a mean absolute error (MAE) of 0.14 cm across 137 measurements; specific details are itemized in Table 1. This outcome validates the efficacy of the annotation framework in maintaining the precision of the Rosetum3D labels.

3.3 Data Statistics

In total, 20 scenes are collected, featuring a diverse array of rose varieties. The resulting dataset includes 46,848 rose objects characterized by 2D bounding boxes and comprehensive 2D/3D keypoint annotations. The dataset is partitioned into 17 scenes for training and 3 scenes for testing; comprehensive dataset statistics are itemized in Table 2. Furthermore, a comparative analysis between Rosetum3D and existing public 3D vision datasets is presented in Table 3.

4 Tasks and Benchmarks

Rosetum3D functions as a multi-task benchmark for agricultural perception challenges, encompassing: (1) 2D object detection, (2) 2D rose keypoint localization,

Fig. 4: Some samples of annotation validation with tape measures.

Table 2: Data Statistics of the Rosetum3D dataset.

Set	Scenes	Images	Objects	Keypoints	Visible-k	Occluded-k	Invisible-k
Train	17	17,924	40,759	366,831	281,779	1,758	83,294
Test	3	3,190	6,089	54,801	42,728	610	11,463
Total	20	21,114	46,848	421,632	324,507	2,368	94,757

Table 3: Comparison between Rosetum3D and Other Public 3D Vision Datasets, including NYU Depth V2 [38], SUN RGB-D [39], ScanNet [11], ScanNet++ [47], Human3.6M [22], KITTI [19], nuScenes [2], and PandaSet [43].

Dataset	Year	Scene Type	Size	Acquisition Method	Modality
NYU Depth V2	2012	Indoor scenes	0.5k scenes	RGB-D sensor	RGB-D
SUN RGB-D	2015	Indoor scenes	10k images	RGB-D sensor	RGB-D
ScanNet	2017	Indoor scenes	2,500k frames	RGB-D sensor	RGB-D
ScanNet++	2023	Indoor scenes	3,000k+ frames	RGB-D sensor	RGB-D
Human3.6M	2014	Human	3,600k frames	Multi-view RGB + MoCap	RGB + 3D keypoints
KITTI	2012	Vehicle (driving)	200k frames	Stereo RGB + LiDAR + GPS	RGB + LiDAR
nuScenes	2019	Vehicle (driving)	1,400k frames	RGB + LiDAR + Radar	RGB + LiDAR + Radar
PandaSet	2021	Vehicle (driving)	48k images & 16k LiDAR sweeps	RGB + LiDAR	RGB + LiDAR
Rosetum3D (ours)	2026	Rose scenes	20k+ images & 20k+ depth maps	RGB-D sensor	RGB-D + 3D keypoints

(3) depth estimation, (4) local feature matching, (5) instance re-identification, and (6) 3D rose keypoint localization. Regarding the last five tasks, experiments are performed utilizing existing methods, with comparative results tabulated accordingly. For the 3D rose keypoint localization task, we design two meth-

Table 4: Experimental 2D object detection results on the Rosetum3D dataset.

Algorithm	Backbone	AP	AP_{50}	AP_{75}	AP_M	AP_L
Faster R-CNN [34]	ResNet50	31.0	66.8	24.7	13.1	32.1
DETR [3]	ResNet50	33.0	66.5	28.8	14.2	34.3
YOLOv5 [28]	CSPDarkNet53	**35.0**	**67.1**	**32.4**	**19.1**	**35.9**
RT-DETR [50]	ResNet50	33.6	64.8	30.7	13.8	34.7

Table 5: Experimental 2D rose keypoint localization results on the Rosetum3D dataset. All top-down methods utilize a YOLOv5 detector, which achieves the best results in Table 4. DEKR† is an internal design inspired by recent decoupled keypoint regression strategies [20]. *Italic* indicates the best results without extra information. **Bold** indicates the best results by introducing extra information (the ground-truth detection boxes).

	Method	Backbone	Detector	AP	AP_{50}	AP_{75}	AR	AR_{50}	AR_{75}
Top-down	RTMPose [23]	CSPNeXt-L	YOLOv5	16.6	45.2	8.6	27.1	54.9	22.9
			Ground Truth	37.2	82.2	28.9	48.0	86.5	44.4
Top-down	ViTPose [45]	ViT-L	YOLOv5	21.9	48.5	17.2	32.0	56.9	31.2
			Ground Truth	**42.1**	**85.8**	**37.2**	53.6	**90.7**	53.7
Bottom-up	YOLOXPose [29]	CSPDarknet	-	*25.0*	*60.3*	*19.3*	***53.7***	*88.7*	***55.2***
Bottom-up	DEKR† + HRNet [40]	HRNet	-	24.9	57.8	18.6	43.9	81.2	41.8

ods in **Section 5**. Baselines are established using domain-relevant models, integrating both general state-of-the-art architectures and agriculturally-specialized networks, to assess performance under zero-shot and fine-tuning protocols.

4.1 2D Rose Object Detection

Several representative object detectors are evaluated on the Rosetum3D dataset to establish a comprehensive benchmark. Evaluated models include: (1) CNN-based detectors, such as Faster R-CNN [34] (two-stage detector), and YOLOv5 [28] (one-stage detector); and (2) Transformer-based detectors, notably DETR [3]. Experimental results are summarized in Table 4. Following the protocols of the MSCOCO dataset [27], evaluation indices are computed for objects of heterogeneous sizes. Notably, small rose objects are not annotated, given their significant distance from the sensor, which precludes precise keypoint measurement. Consequently, performance metrics AP_M and AP_L are reported in Table 4. Here, AP_M characterizes detection performance on medium-sized objects (area between 32×32 and 96×96 pixels), while AP_L assesses performance on large objects (area $> 96 \times 96$ pixels).

4.2 2D Rose Keypoint Localization

To establish a benchmark for 2D rose keypoint localization, the proposed dataset is evaluated using the MMPose framework [10]. Several high-performing archi-

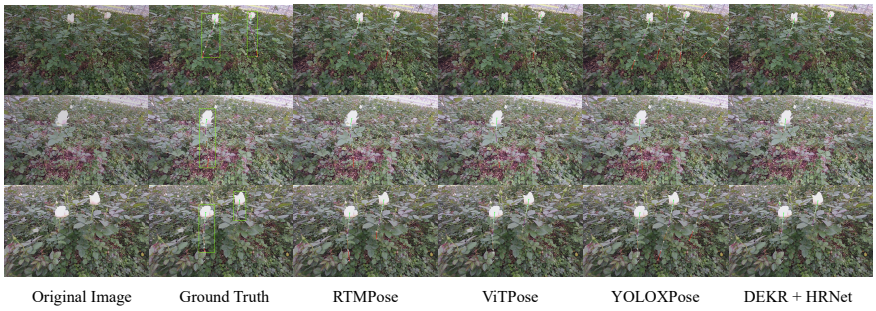

Fig. 5: Comparison of 2D Rose Localization Results. The evaluation encompasses two top-down approaches (RTMPose [23] and ViTPose [45]) and two bottom-up approaches (DEKR [20] + HRNet [40] and YOLOXPose [29])

tectures originally developed for human pose estimation were adapted for the rose target, including: (1) top-down approaches, such as RTMPose [23] and ViTPose [45]. Within the top-down pipeline, all detectors leverage YOLOv5 [28]. (2) bottom-up approaches, notably HRNet [40]. Experimental results are presented in Table 5. Qualitative comparison results are illustrated in Fig. 5.

Notably, the performance of top-down approaches appears inferior to that of bottom-up methodologies (e.g., 21.9 AP for VitPose vs. 24.9 AP for DEKR† + HRNet). This disparity is primarily attributable to detector-related constraints that limit top-down performance. Upon substituting detector results with ground-truth bounding boxes, the performance of top-down approaches exceeds that of bottom-up architectures significantly (42.1 AP vs. 24.9 AP).

4.3 Monocular Depth Estimation

Leveraging structured-light depth sensors during multi-scene data acquisition, this study curates 33,019 precisely aligned RGB-D image pairs to establish a comprehensive dataset for monocular depth estimation. To benchmark agricultural depth perception challenges, state-of-the-art monocular depth estimation methods are evaluated using dual protocols: zero-shot inference on pre-trained models and fine-tuning with the designated training split. Quantitative results across key metrics, including AbsRel, RMSE, and $\delta < 1.25$, are systematically compared in Table 6, supplemented by qualitative comparison results illustrated in Supplementary Materials. Experimental results demonstrate that the fine-tuning procedure yields significant performance enhancements.

4.4 Local Feature Matching

Each scene is represented using its RGB images as the primary data source; these are processed independently with COLMAP for 3D reconstruction, adhering to a methodology analogous to MegaDepth [24]. The resulting reconstructions

Table 6: The results of the monocular depth estimation task on the Rosetum3D dataset. For zero-shot evaluation, the models are trained on NYU Depth V2 [38] and tested on the Rosetum3D dataset without fine-tuning.

Method	Encoder	Lower is better ↓			Higher is better ↑		
		AbsRel	RMSE	$\log_{10}$	δ_1	δ_2	δ_3
Zero-shot							
IEBins [36]	Swin-Large	0.554	1.655	0.211	0.241	0.476	0.716
ZoeDepth [1]	ViT-L	0.591	2.186	0.270	0.266	0.468	0.618
Depth Anything [46]	ViT-L	**0.498**	**1.092**	**0.163**	**0.373**	**0.664**	**0.831**
With-Train							
DPT [33]	VIT-B	0.211	0.825	0.081	0.738	0.930	0.970
DepthFormer [25]	Swin-Large	0.181	**0.688**	0.067	0.797	**0.950**	**0.978**
BinsFormer [26]	Swin-Large	**0.165**	0.734	**0.066**	**0.822**	0.946	0.974

Table 7: The results of the local feature matching task on Rosetum3D-1800 with zero-shot evaluation. Measured in AUC (higher is better).

Method	AUC@ →	5° ↑	10° ↑	20° ↑
DeDoDe [16]		62.3	75.5	84.9
DKM [15]		71.1	80.6	87.5
PMatch [53]		70.7	80.9	**88.0**
RoMa [17]		**72.3**	**81.4**	87.8

serve as ground-truth references for feature matching. Leveraging this pipeline, Rosetum3D-1800 is established as a benchmark comprising 1,800 image pairs curated to facilitate the zero-shot evaluation of existing feature matching methods. Quantitative results are itemized in Table 7 and qualitative results are listed in Supplementary Materials. All models undergo zero-shot evaluation, preserving raw outputs to maintain unbiased performance assessment. Despite being comparable in scale to existing benchmarks, Rosetum3D-1800 presents distinct challenges as a result of the high similarity among local features inherent in agricultural environments.

4.5 Instance Re-Identification

Identity labels are assigned to a subset of individuals in Rosetum3D to evaluate the performance of existing person re-identification models within the proposed plant cultivation scenario; results are detailed in Table 8. A total of 1,923 individuals are annotated, comprising 36,712 images. The partitioning of training and test sets adheres to the protocol used in the 2D rose object detection task.

Table 8: The results of instance re-identification task. Given the limited number of images per identity, we report Recall@1 and Recall@3 instead of higher-rank metrics. "CAJ" denotes the CA-Jaccard distance. CC + CAJ is an unsupervised method.

Method	backbone	mAP (%)	Rank-1 (%)	Rank-3 (%)
OSNet [51]	OSNet	75.8	95.5	**97.9**
TransReID [21]	ViT-B	80.6	95.7	96.8
SOLIDER [4]	Swin-B	**80.7**	**96.0**	97.3
CC [12]+CAJ [5]	ResNet50	70.9	93.9	96.3

5 3D Rose Localization

The primary objective of Rosetum3D is to establish a customized 3D keypoint dataset for roses and to investigate methodologies for attaining precise 3D localization. Two methodologies are employed for 3D rose keypoint localization: **depth-inferencer-based localization** and **multi-frame 3D localization**.

5.1 Depth-Inferencer-Based Localization

Within the depth-inferencer-based localization framework, 3D keypoints are reconstructed by combining inferred 2D keypoints with depth estimates derived from a depth prediction model. Evaluation is performed against ground-truth 3D keypoint data, comprising manual annotations and corresponding ground-truth depth maps. A top-down keypoint detection model is utilized in conjunction with the DETR object detector; subsequently, keypoint prediction results are evaluated on a per-instance basis. Depth inference leverages both zero-shot and fine-tuned depth models.

In addition to utilizing predicted depth maps, this study evaluates results obtained using ground-truth depth maps acquired from RGB-D sensors. Performance evaluation utilizes the mean per joint position error (MPJPE) [48,49,54], a standard metric within human pose estimation (HPE). During the assessment, the predicted instance with the highest confidence score is matched with its corresponding ground-truth instance; quantitative results are presented in Table 9.

5.2 Multi-Frame 3D Localization

Under the multi-frame framework, 3D keypoint positions for both the current and neighboring frames are first recovered by integrating inferred 2D bounding boxes and keypoints with depth information. Subsequently, a re-identification (Re-ID) model is employed to match identical instances across frames based on 2D appearance cues. Finally, the matching results from adjacent frames are used as temporal constraints to optimize the 3D positions in the current frame.

Here, the framework undergoes the following optimization: Within a three-frame sequence, a temporal-consistency-based optimization strategy is adopted to enhance the stability and robustness of 3D keypoints in the intervening frame.

Table 9: 3D rose localization results obtained by integrating the 2D keypoint inference model with the depth estimation model. Measured in MPJPE (lower is better). DEKR† is an internal design inspired by recent decoupled keypoint regression strategies [20]. *Italic* indicates the best results without extra information. **Bold** indicates the best results by introducing extra information (the ground-truth depth map).

2D Keypoint Inferencer	Detector	Depth Inferencer	Encoder	MPJPE(mm) ↓
ViTPose [45]	DETR	Depth Anything	ViT-L	656.9
		BinsFormer	Swin-Large	647.2
		Ground Truth	-	556.8
RTMPose [23]	DETR	Depth Anything	ViT-L	650.6
		BinsFormer	Swin-Large	*644.5*
		Ground Truth	-	558.4
DEKR† + HRNet [40]	-	Depth Anything	ViT-L	669.4
		BinsFormer	Swin-Large	654.3
		Ground Truth	-	**522.8**
YOLOXPose [29]	-	Depth Anything	ViT-L	653.2
		BinsFormer	Swin-Large	645.1
		Ground Truth	-	589.3

Table 10: Comparison of whether to use the multi-frame 3D localization framework. Here we use OSNet-based [51] Re-ID model. MPJPE is reported, where lower values indicate better performance. ✓ and ✗ denote whether the multi-frame framework is used.

2D Keypoint Inferencer	Detector	Multi-Frame	MPJPE(mm) ↓
ViTPose [45]	DETR	✗	556.8
		✓	**539.1**
RTMPose [23]	DETR	✗	558.4
		✓	542.2

This strategy fuses the intervening frame measurement with the priors inferred from adjacent frames. Let the 3D keypoints of frames $t-1$, t, and $t+1$ be denoted by P_{t-1}, P_t, and P_{t+1}, respectively. Initially, a temporal prior for the intervening frame is constructed via linear interpolation:

$$P_{\text{interp}} = \frac{1}{2}\left(P_{t-1} + P_{t+1}\right). \tag{2}$$

Subsequently, the optimized 3D position for each keypoint is computed by weighted fusion between the intervening frame observation and the temporal prior.

The weight of the intervening frame observation is modulated by its 2D keypoint confidence (i.e., $w_d = W_{\text{depth}} \cdot \text{conf}$), while the temporal prior weight is defined as a constant $w_t = W_{\text{temp}}$. The final optimized 3D keypoints are formulated as:

$$P_t^* = \frac{w_d P_t + w_t P_{\text{interp}}}{w_d + w_t}. \tag{3}$$

This strategy effectively suppresses significant depth errors precipitated by inaccurate 2D localization or absent depth measurements, consequently enhancing the temporal smoothness and accuracy of the estimated 3D keypoints.

Compared with the Depth-inferencer-based Localization method, the proposed multi-frame framework achieves improved accuracy under the same 2D inference model; however, it does not attain the best overall performance, as shown in Table 10. This approach is only applicable when a top-down model is adopted as the 2D inferencer. Under our data acquisition setup (i.e., a camera moving horizontally at a constant speed), the optimization process does not account for errors introduced by changes in camera pose.

6 Conclusion

This paper introduces Rosetum3D, a pioneering large-scale RGB-D dataset designed to overcome the critical data scarcity in floriculture automation by capturing occlusion-heavy rose cultivation scenes across diverse varieties and growth stages using structured-light depth sensors, which enables biologically relevant annotations through our hybrid 2D-to-3D framework. We first manually label 2D bounding boxes and several semantic keypoints per rose and subsequently leverage depth maps to reconstruct their 3D positions, achieving accurate measurements. The Rosetum3D dataset establishes the first agricultural benchmark for occlusion-invariant vision tasks, including 2D object detection, 2D/3D rose localization, local feature matching, depth estimation, and instance re-identification. We propose two different frameworks for 3D rose localization; however, the accuracy errors remain relatively large, and the methods exhibit certain limitations. We believe that Rosetum3D can boost the intelligence and automation of agriculture. We will release Rosetum3D openly to bridge the computer vision algorithms and agriculture.

Acknowledgements

We gratefully acknowledge the Strategic Investment Department of Kunming Industrial Development Investment Co., Ltd and Anning Rose World Flower Industry Co., Ltd. for providing the site for our data collection. This work is supported by Yunnan Provincial Research Foundation for Basic Research, China [grant number 202401AT070427, 202401AT070442] and the 17th Graduate Student Scientific Research and Innovation Project of Yunnan University [grant number KC-252511855].

References

1. Bhat, S.F., Birkl, R., Wofk, D., Wonka, P., Müller, M.: Zoedepth: Zero-shot transfer by combining relative and metric depth (2023), https://arxiv.org/abs/2302.12288

2. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11621–11631 (2020)
3. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Proceedings of the European Conference on Computer Vision. pp. 213–229 (2020)
4. Chen, W., Xu, X., Jia, J., Luo, H., Wang, Y., Wang, F., Jin, R., Sun, X.: Beyond appearance: a semantic controllable self-supervised learning framework for human-centric visual tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15050–15061 (2023)
5. Chen, Y., Fan, Z., Chen, Z., Zhu, Y.: Ca-jaccard: Camera-aware jaccard distance for person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17532–17541 (2024)
6. Choudhury, S.D., Guha, S., Das, A., Das, A.K., Samal, A., Awada, T.: Flowerphenonet: Automated flower detection from multi-view image sequences using deep neural networks for temporal plant phenotyping analysis. Remote Sensing **14**(24), 6252 (2022)
7. Conn, A., Chandrasekhar, A., van Rongen, M., Leyser, O., Chory, J., Navlakha, S.: Network trade-offs and homeostasis in arabidopsis shoot architectures. PLOS Computational Biology **15**(9), 1–19 (2019)
8. Conn, A., Pedmale, U.V., Chory, J., Navlakha, S.: High-resolution laser scanning reveals plant architectures that reflect universal network design principles. Cell systems **5**(1), 53–62 (2017)
9. Conn, A., Pedmale, U.V., Chory, J., Stevens, C.F., Navlakha, S.: A statistical description of plant shoot architecture. Current Biology **27**(14), 2078–2088 (2017)
10. Contributors, M.: Openmmlab pose estimation toolbox and benchmark. `https://github.com/open-mmlab/mmpose` (2020)
11. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Niessner, M.: ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2432–2443 (2017)
12. Dai, Z., Wang, G., Zhu, S., Yuan, W., Tan, P.: Cluster contrast for unsupervised person re-identification (2021), `https://arxiv.org/abs/2103.11568`
13. Di Cicco, M., Iocchi, L., Grisetti, G.: Non-parametric calibration for depth sensors. Robotics and Autonomous Systems **74**, 309–317 (2015)
14. Dutagaci, H., Rasti, P., Galopin, G., Rousseau, D.: Rose-x: An annotated data set for evaluation of 3d plant organ segmentation methods. Plant methods **16**(1), 1–14 (2020)
15. Edstedt, J., Athanasiadis, I., Wadenback, M., Felsberg, M.: DKM: Dense kernelized feature matching for geometry estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8922–8931 (2023)
16. Edstedt, J., Bökman, G., Wadenbäck, M., Felsberg, M.: DeDoDe: Detect, don't describe — describe, don't detect for local feature matching. In: Proceedings of the International Conference on 3D Vision. pp. 148–157 (2024)
17. Edstedt, J., Sun, Q., Bokman, G., Wadenback, M., Felsberg, M.: RoMa: Robust dense feature matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19790–19800 (2024)
18. Fong, W.K., Mohan, R., Hurtado, J.V., Zhou, L., Caesar, H., Beijbom, O., Valada, A.: Panoptic Nuscenes: A large-scale benchmark for LiDAR panoptic segmentation and tracking. IEEE Robotics and Automation Letters **7**(2), 3795–3802 (2022)

19. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? The KITTI vision benchmark suite. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3354–3361 (2012)
20. Geng, Z., Sun, K., Xiao, B., Zhang, Z., Wang, J.: Bottom-up human pose estimation via disentangled keypoint regression. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14676–14686 (2021)
21. He, S., Luo, H., Wang, P., Wang, F., Li, H., Jiang, W.: Transreid: Transformer-based object re-identification. In: Proceedings of the International Conference on Computer Vision. pp. 15013–15022 (October 2021)
22. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3.6M: Large scale datasets and predictive methods for 3D human sensing in matural environments. IEEE Transactions on Pattern Analysis and Machine Intelligence **36**(7), 1325–1339 (2014)
23. Jiang, T., Lu, P., Zhang, L., Ma, N., Han, R., Lyu, C., Li, Y., Chen, K.: RTMPose: Real-time multi-person pose estimation based on MMPose (2023), https://arxiv.org/abs/2303.07399
24. Li, Z., Snavely, N.: Megadepth: Learning single-view depth prediction from internet photos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2041–2050 (2018)
25. Li, Z., Chen, Z., Liu, X., Jiang, J.: Depthformer: Exploiting long-range correlation and local information for accurate monocular depth estimation (2022), https://arxiv.org/abs/2203.14211
26. Li, Z., Wang, X., Liu, X., Jiang, J.: Binsformer: Revisiting adaptive bins for monocular depth estimation (2022), https://arxiv.org/abs/2204.00987
27. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: Proceedings of the European Conference on Computer Vision. pp. 740–755 (2014)
28. LLC, U.: YOLOv5: Real-time object detection (2020), https://github.com/ultralytics/yolov5, version 7.0 [Accessed: 2025-07-24]
29. Maji, D., Nagori, S., Mathew, M., Poddar, D.: Yolo-pose: Enhancing yolo for multi-person pose estimation using object keypoint similarity loss. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2637–2646 (2022)
30. Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing. pp. 722–729 (2008)
31. Orbbec, I.: Gemini 335l (2025), https://www.orbbec.com/products/stereo-vision-camera/gemini-335l/ [Accessed: 2025-07-27]
32. Orbbec, I.: Gemini 336l (2025), https://www.orbbec.com/products/stereo-vision-camera/gemini-336l/ [Accessed: 2025-07-27]
33. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: Proceedings of the International Conference on Computer Vision. pp. 12159–12168 (2021)
34. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**(6), 1137–1149 (2017)
35. Schunck, D., Magistri, F., Rosu, R.A., Cornelißen, A., Chebrolu, N., Paulus, S., Léon, J., Behnke, S., Stachniss, C., Kuhlmann, H., Klingbeil, L.: Pheno4d: A spatio-temporal dataset of maize and tomato plant point clouds for phenotyping and advanced plant analysis. Plos one **16**(8), 1–18 (2021)

36. Shao, S., Pei, Z., Chen, W., Chen, P.C.Y., Li, Z.: Iebins: Iterative elastic bins for monocular depth estimation and completion. International Journal of Computer Vision **133**, 2463–2486 (2025)
37. Shen, P., Jing, X., Deng, W., Jia, H., Wu, T.: PlantGaussian: Exploring 3D Gaussian splatting for cross-time, cross-scene, and realistic 3D plant visualization and beyond. The Crop Journal **13**(2), 607–618 (2025)
38. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: Proceedings of the European Conference on Computer Vision. pp. 746–760 (2012)
39. Song, S., Lichtenberg, S.P., Xiao, J.: SUN RGB-D: A RGB-D scene understanding benchmark suite. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 567–576 (2015)
40. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5693–5703 (2019)
41. Sun, Y., Zhang, Z., Sun, K., Li, S., Yu, J., Miao, L., Zhang, Z., Li, Y., Zhao, H., Hu, Z., Xin, D., Chen, Q., Zhu, R.: Soybean-mvs: Annotated three-dimensional model dataset of whole growth period soybeans for 3d plant organ segmentation. Agriculture **13**(7) (2023)
42. Teichman, A., Miller, S., Thrun, S.: Unsupervised intrinsic calibration of depth sensors via slam. In: Robotics: Science and Systems. vol. 248 (2013)
43. Xiao, P., Shao, Z., Hao, S., Zhang, Z., Chai, X., Jiao, J., Li, Z., Wu, J., Sun, K., Jiang, K., Wang, Y., Yang, D.: PandaSet: Advanced sensor suite dataset for autonomous driving. In: Proceedings of the International Intelligent Transportation Systems Conference. pp. 3095–3101 (2021)
44. Xiao, S., Ye, Y., Fei, S., Chen, H., Zhang, B., Li, Q., Cai, Z., Che, Y., Wang, Q., Ghafoor, A., Bi, K., Shao, K., Wang, R., Guo, Y., Li, B., Zhang, R., Chen, Z., Ma, Y.: High-throughput calculation of organ-scale traits with reconstructed accurate 3D canopy structures using a UAV RGB camera with an advanced cross-circling oblique route. ISPRS Journal of Photogrammetry and Remote Sensing **201**, 104–122 (2023)
45. Xu, Y., Zhang, J., Zhang, Q., Tao, D.: ViTPose: Simple vision transformer baselines for human pose estimation. In: Proceedings of the Annual Conference on Neural Information Processing Systems. pp. 38571–38584 (2022)
46. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10371–10381 (2024)
47. Yeshwanth, C., Liu, Y.C., Nießner, M., Dai, A.: Scannet++: A high-fidelity dataset of 3d indoor scenes. In: Proceedings of the International Conference on Computer Vision. pp. 12–22 (2023)
48. Zhang, J., Tu, Z., Yang, J., Chen, Y., Yuan, J.: Mixste: Seq2seq mixed spatio-temporal encoder for 3d human pose estimation in video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13232–13242 (2022)
49. Zhao, Q., Zheng, C., Liu, M., Wang, P., Chen, C.: PoseFormerV2: Exploring frequency domain for efficient and robust 3D human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8877–8886 (2023)
50. Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., Liu, Y., Chen, J.: Detrs deat yolos on real-time object detection (2023), https://arxiv.org/abs/2304.08069

51. Zhou, K., Yang, Y., Cavallaro, A., Xiang, T.: Learning generalisable omni-scale representations for person re-identification. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(9), 5056–5069 (2021)
52. Zhu, J., Zhai, R., Ren, H., Xie, K., Du, A., He, X., Cui, C., Wang, Y., Ye, J., Wang, J., Jiang, X., Wang, Y., Huang, C., Yang, W.: Crops3D: A diverse 3D crop dataset for realistic perception and segmentation toward agricultural applications. Scientific Data **11**(1), 1438–1455 (2024)
53. Zhu, S., Liu, X.: Pmatch: Paired masked image modeling for dense geometric matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023)
54. Zhu, W., Ma, X., Liu, Z., Liu, L., Wu, W., Wang, Y.: MotionBERT: A unified perspective on learning human motion representations. In: Proceedings of the International Conference on Computer Vision. pp. 15085–15099 (2023)