Self-Supervised Learning for Robotic Leaf Manipulation: A Hybrid Geometric-Neural Approach

Srecharan Selvam,   Abhisesh Silwal,   George Kantor

Robotics Institute, Carnegie Mellon University

{sselvam, asilwal, gkantor}@andrew.cmu.edu

Abstract

Automating leaf manipulation in agricultural settings faces significant challenges, including the variability of plant morphologies and deformable leaves. We propose a novel hybrid geometric-neural approach for autonomous leaf grasping that combines classical computer vision with neural networks through self-supervised learning. Our method integrates YOLOv8 for instance segmentation and RAFT-Stereo for 3D depth estimation to build rich leaf representations, which feed into both a geometric feature scoring pipeline and a neural refinement module (Grasp-PointCNN). The key innovation is our confidence-weighted fusion mechanism that dynamically balances the contribution of each approach based on prediction certainty. Our self-supervised framework uses the geometric pipeline as an expert teacher to automatically generate training data. Experiments demonstrate that our approach achieves an 88.0% success rate in controlled environments and 84.7% in real greenhouse conditions, significantly outperforming both purely geometric (75.3%) and neural (60.2%) methods. This work establishes a new paradigm for agricultural robotics where domain expertise is seamlessly integrated with machine learning capabilities, providing a foundation for fully automated crop monitoring systems.

1. Introduction

Agricultural robotics has emerged as a key technology for addressing labor shortages and improving efficiency in modern farming [8, 43]. Among greenhouse cultivation tasks, leaf sampling for disease detection remains a significant bottleneck, requiring skilled workers to manually identify, select, and extract tissue samples from thousands of plants [4, 42]. This labor-intensive process increases operational costs and limits the frequency of plant health monitoring, potentially allowing diseases to spread undetected [5, 33].

Automating leaf manipulation presents unique challenges compared to traditional robotic grasping tasks. Unlike rigid industrial objects, plant leaves are deformable, vary significantly in size and orientation, and are often partially occluded in dense canopies [30, 38]. While recent advances in deep learning have revolutionized robotic grasping for industrial applications [34, 37, 53], these approaches typically require large datasets of labeled grasp points—a resource that is prohibitively expensive to create for agricultural settings where plant morphology varies continuously throughout growth cycles [26].

Existing approaches to agricultural manipulation fall into two categories: purely geometric methods that rely on hand-crafted features [18, 20, 44], and end-to-end deep learning systems trained on synthetic or limited real-world data [3, 49, 52]. Geometric approaches, while interpretable and robust to domain shifts, struggle with the natural variability of plant structures. Conversely, deep learning methods excel at handling complex visual patterns but suffer from poor generalization when deployed on new crop varieties or growth stages not represented in their training data [36].

We present a novel hybrid approach that leverages the complementary strengths of geometric reasoning and neural networks through self-supervised learning. Our key insight is that traditional computer vision algorithms, despite their limitations, encode valuable domain expertise that can serve as a teacher for training neural networks without manual annotation [21]. This approach enables continuous learning from operational data while maintaining the interpretability and reliability required for agricultural automation.

Our system operates on a 6-DOF gantry robot equipped with stereo vision and a custom end-effector for leaf manipulation. The perception pipeline combines YOLOv8 instance segmentation [23] with RAFT-Stereo depth estimation [31] to generate 3D representations of plant canopies. For grasp point selection, we implement a dual-path architecture: a geometric pipeline using Pareto optimization across multiple hand-crafted features (flatness, accessibil-

ity, edge distance), and a convolutional neural network with spatial attention that learns from the geometric system's decisions [25].

The main contributions of this work include:

- A self-supervised learning framework where geometric algorithms act as expert teachers for neural networks, eliminating the need for manual grasp annotation in agricultural settings
- A hybrid decision architecture that dynamically weighs geometric and learned features based on prediction confidence, achieving robust performance across diverse plant conditions
- A comprehensive grasp point selection system incorporating novel scoring functions tailored to leaf-specific constraints such as deformability, approach angles, and occlusion handling
- Extensive validation on thousands of real plant samples demonstrating significant improvements over traditional geometric methods, particularly in challenging scenarios with partial occlusion and irregular orientations

This work provides a foundation for fully automated crop monitoring systems and establishes a new paradigm for agricultural robotics where domain expertise is seamlessly integrated with machine learning capabilities.

2. Related Work

2.1. Vision-Based Leaf Manipulation

Traditional approaches to robotic leaf manipulation in agricultural settings relied on geometric reasoning and classical computer vision. Hemming et al. developed methods for cucumber leaf detection in greenhouses using color and texture features [19], while Bac et al. presented obstacle-aware motion planning for tomato canopies [6]. Several studies focused on deformable leaf modeling, including Cerutti et al.'s parametric active polygon models [9], Xia et al.'s active shape models for overlapping leaves [51], and Jin et al.'s probabilistic graphical models for leaf structure analysis [22]. The integration of 3D information improved robustness, as demonstrated by Guo and Xu's multiview stereo reconstruction for lettuce segmentation [16] and Sodhi et al.'s plant growth monitoring system using structure-from-motion [45]. While effective in controlled conditions, these methods often required extensive tuning and struggled with natural plant variability, particularly under varying illumination conditions [11].

2.2. Deep Learning for Agricultural Grasping

Deep learning has shown promise in agricultural manipulation, though with unique challenges compared to industrial applications. Barth et al. developed CNN-based systems for broccoli harvesting that handle significant occlusion [7], while Arad et al. demonstrated sweet pepper harvesting

combining YOLO detection with stereo depth [3]. For leaf-specific tasks, Ahlin et al. pioneered CNN-based leaf identification with visual servoing for autonomous sampling, achieving 85% success rates in greenhouses [2]. However, these approaches typically require extensive training data—a significant limitation given the continuous variation in plant morphology [47]. To address this, researchers have explored simulation, with approaches like Dex-Net generating synthetic grasp scenarios [34], inspiring agricultural adaptations for data generation [28]. Recent advances in generative data augmentation have shown promise for bridging the synthetic-real domain gap [48], particularly for multi-crop environments with challenging lighting conditions.

2.3. Self-Supervised Learning in Agricultural Robotics

Self-supervised learning has emerged as a promising paradigm for agricultural robotics, particularly where manual annotation is expensive. Zhang et al. demonstrated self-supervised learning for tomato harvesting, using classical vision systems to provide training labels [54]. Similar bootstrapping approaches include Kootstra et al.'s work on sweet pepper detection, where geometric algorithms generated training data for neural networks [27] and Tao et al.'s approach using temporal consistency for plant growth tracking [46]. This knowledge transfer from classical to learning-based systems has proven particularly valuable in controlled environment agriculture, where hybrid approaches consistently outperform purely learned policies [14, 42]. Recent work has also explored contrastive learning frameworks for agricultural visual representations [50], enabling more sample-efficient adaptation to new crops and growth conditions.

2.4. 3D Perception and Hybrid Systems

Accurate depth sensing is crucial for manipulation in dense plant canopies. While traditional stereo algorithms struggle with plant textures, recent advances like RAFT-Stereo have dramatically improved accuracy for agricultural applications [31]. Lipson et al.'s recurrent architecture achieves state-of-the-art performance on challenging plant datasets, enabling precise leaf pose estimation [40]. Alternative sensing modalities such as time-of-flight cameras [35] and structured light systems [55] have also shown promise for plant phenotyping applications with complex geometries. Recent research increasingly combines classical and learning approaches, as demonstrated by Lehnert et al.'s hybrid system for pepper harvesting [29] and Adamides et al.'s framework for integrating human expertise with machine learning [1]. These hybrid architectures leverage geometric interpretability with neural adaptability, making them ideal for complex agricultural tasks where safety and reliability are
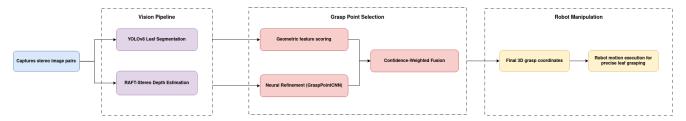
Figure 1. System architecture showing the integration of vision pipeline, grasp point selection, and robot manipulation modules. The hybrid approach combines geometric feature scoring with neural refinement through confidence-weighted fusion.

paramount [12]. Confidence-aware systems that adaptively balance multiple decision sources have proven especially effective in scenarios with high uncertainty, demonstrating robustness across varying environmental conditions [13].

3. Method

We present a hybrid approach for autonomous leaf grasping that combines geometric algorithms with neural networks through self-supervised learning. Our system eliminates the need for manual grasp annotation while maintaining robust performance in complex greenhouse environments. This section details our perception pipeline, grasp point selection algorithms, and the self-supervised framework that bridges classical and modern approaches.

3.1. System Overview

Figure 1 presents our hybrid leaf grasping system architecture, consisting of three modules: vision pipeline, grasp point selection, and robot manipulation. The system processes stereo image pairs from a 6-DOF gantry robot to output precise 3D grasp coordinates.

The vision pipeline employs YOLOv8 for instance segmentation of individual leaves and RAFT-Stereo for dense depth estimation. As shown in Figure 1, these outputs are fused to create 3D leaf representations containing both semantic and geometric information.

The grasp point selection module implements our hybrid approach through two parallel paths. The geometric feature scoring path evaluates candidates using traditional CV algorithms based on features like flatness, accessibility, and approach angles. Simultaneously, the neural refinement path (GraspPointCNN) processes the same data using learned features. Both predictions are combined through confidence-weighted fusion, dynamically balancing traditional CV (70-90%) and neural network (10-30%) contributions.

Our key innovation is the self-supervised training scheme where geometric algorithms act as expert teachers, automatically labeling grasp points to train the neural network. This enables the system to initially mimic geometric reasoning while developing generalization capabilities beyond hand-crafted features.
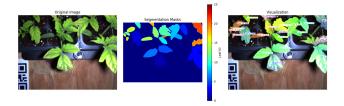

Figure 2. Vision pipeline outputs: (a) Instance segmentation with individual leaf masks, (b) RAFT-Stereo disparity map, (c) 3D point cloud reconstruction with highlighted target leaf.

The robot manipulation module executes precise leaf grasping using the final 3D coordinates, with motion planning optimized for the gantry configuration and safety validation through force feedback.
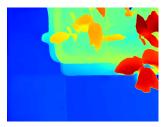
3.2. Vision Pipeline

The vision pipeline, illustrated in the left section of Figure 1, processes stereo image pairs to generate rich 3D representations of plant leaves. This pipeline employs two parallel processing streams: instance segmentation and stereo depth estimation, whose outputs are fused to create comprehensive leaf models for grasp planning.

3.2.1. Instance Segmentation

We utilize YOLOv8 [23] for real-time instance segmentation of individual leaves. Unlike standard implementations, we fine-tuned YOLOv8 on a custom dataset of 900+ images containing soybean and tomato plants in greenhouse conditions. This domain-specific training enables robust leaf detection even in challenging scenarios with significant overlap and occlusion, achieving 90%+ confidence scores in operational conditions.

As shown in Figure 2, the network outputs binary masks for each detected leaf instance along with confidence scores. The segmentation accurately delineates individual leaf boundaries despite complex overlapping patterns typical in dense canopies. Our implementation processes 1440×1080 resolution images at approximately 50ms per frame, meeting real-time requirements for robotic manipulation. Each detected leaf is assigned a unique identifier

(a) Raw stereo image   (b) RAFT-Stereo depth map

(c) 3D leaf reconstruction from stereo depth and segmentation

Figure 3. RAFT-Stereo outputs showing the processing pipeline: (a) Raw image from the left camera of the stereo pair, (b) Generated disparity map where warmer colors indicate closer objects, (c) Final 3D reconstruction combining depth and segmentation data.

and confidence score, enabling robust tracking throughout the grasp selection process.

3.2.2. Stereo Depth Estimation

For 3D reconstruction, we employ RAFT-Stereo [31], which generates dense disparity maps through iterative refinement using recurrent all-pairs field transforms. This approach handles the thin structures and low-texture regions characteristic of plant foliage more reliably than traditional stereo matching algorithms [41].

Our calibrated stereo pair captures synchronized images at 1440×1080 resolution. As illustrated in Figure 3, RAFT-Stereo processes these to produce sub-pixel accurate disparity maps in approximately 60ms, achieving 29% lower 1-pixel error than previous methods on standard benchmarks [15]. The disparity values are converted to metric depth using the camera calibration parameters, enabling accurate 3D reconstruction.

3.2.3. 3D Reconstruction

Each pixel (u, v) with disparity d is back-projected to 3D coordinates (X, Y, Z) using:

$$X = \frac{(u - c_x) \cdot Z}{f_x}, \quad Y = \frac{(v - c_y) \cdot Z}{f_y}, \quad Z = \frac{f \cdot b}{d} \tag{1}$$

where f is the focal length, b is the stereo baseline, and (c_x, c_y) are the principal point coordinates. The resulting point cloud provides comprehensive 3D structure of the scene, as shown in Figure 3(c).

3.2.4. Data Fusion

The vision pipeline combines segmentation masks with depth information to create per-leaf 3D models. For each detected leaf instance, we:

- Extract 3D points by masking the depth map with the leaf's segmentation mask
- Compute geometric properties including centroid position, surface area, and orientation
- Estimate surface normals through local plane fitting for flatness evaluation
- Identify occlusion by detecting missing depth data within mask boundaries

This fusion process outputs a structured representation of each leaf containing both 2D mask information and 3D geometric properties, providing the necessary data for subsequent grasp point selection algorithms. The geometric processing includes signed distance field (SDF) generation, which will be detailed in Section 3.3.

3.3. Geometric Feature Scoring Pipeline

The geometric feature scoring pipeline evaluates candidate leaves and grasp points using hand-crafted features derived from classical computer vision principles. This deterministic approach provides interpretable decisions and serves as the foundation for our self-supervised learning framework.

3.3.1. Optimal Leaf Selection

Given the set of segmented leaves from the vision pipeline, we evaluate each leaf using three key metrics: clutter, distance, and visibility. These metrics are combined using Pareto optimization to identify the optimal grasping target.

$$L^* = \arg\max_{L_i \in \mathcal{L}} \Big(w_c S_c(L_i) + w_d S_d(L_i) + w_v S_v(L_i) \Big) \tag{2}$$

Where:
- L^* is the optimal leaf selection
- $\mathcal{L}$ is the set of all detected leaves
- $S_c(L_i)$ is the clutter/isolation score for leaf i
- $S_d(L_i)$ is the distance score for leaf i
- $S_v(L_i)$ is the visibility score for leaf i
- w_c, w_d, w_v are the weights (0.35, 0.35, 0.30)

Clutter Score quantifies leaf isolation using signed distance fields (SDF):

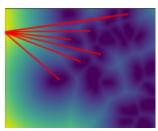$$S_{clutter} = \frac{d_{min}}{d_{min} + d_{max}} \tag{3}$$

Where:
- d_{min} is the distance from centroid to SDF minimum
- d_{max} is the distance from centroid to SDF maximum

Distance Score evaluates the leaf's 3D Euclidean distance from the camera:

$$S_{distance} = e^{-\frac{d_{mean}}{0.3}} \tag{4}$$

(a) Raw image with candidates   (b) SDF representation

Figure 4. Signed Distance Field (SDF) visualization for grasp planning: (a) Raw plant image with leaf candidates, (b) SDF representation showing free space (purple/blue) and occupied regions (yellow/red). Red rays indicate potential grasp approach directions.

Where:
- d_{mean} is the mean Euclidean distance of leaf points
- 0.3m is the scale factor

Visibility Score assesses leaf completeness and position:

$$S_{visibility} = \begin{cases} 0 & \text{if leaf touches image border} \\ 1 - \frac{d_{center}}{d_{max}} & \text{otherwise} \end{cases} \tag{5}$$

Where:
- d_{center} is the distance from leaf centroid to image center
- d_{max} is the maximum possible distance in the image

The final leaf selection employs Pareto optimization with weighted scoring:

$$S_{leaf} = 0.35 \cdot S_{clutter} + 0.35 \cdot S_{distance} + 0.30 \cdot S_{visibility} \tag{6}$$

Figure 4 illustrates the SDF computation used for clutter evaluation. The SDF representation enables efficient calculation of clearance around each leaf candidate, with warmer colors indicating proximity to obstacles.

3.3.2. Geometric Grasp Point Scoring

Once the target leaf is selected, we generate candidate grasp points uniformly distributed across the leaf surface. Each candidate is evaluated using four geometric features:

$$G^* = \arg\max_{p \in L^*} \Big(w_f F(p) + w_a A(p) +$$
$$w_e E(p) + w_{acc} Acc(p) \Big) \cdot (1 - S_{pen}(p)) \tag{7}$$

Where:
- G^* is the optimal grasp point
- p is a candidate point on the selected leaf L^*
- $F(p)$ is the flatness score at point p
- $A(p)$ is the approach vector alignment score at point p

- $E(p)$ is the edge margin score at point p
- $Acc(p)$ is the accessibility score at point p
- $S_{pen}(p)$ is the stem penalty term
- w_f, w_a, w_e, w_{acc} are the weights (0.25, 0.40, 0.20, 0.15)

Flatness Score measures local surface planarity using depth gradients:

$$F(p) = e^{-\alpha \cdot \sqrt{|\nabla_x D(p)|^2 + |\nabla_y D(p)|^2}} \tag{8}$$

Where:
- $D(p)$ is the depth value at point p
- $\nabla_x D$ and $\nabla_y D$ are the gradients in x and y directions
- $\alpha = 5.0$ is the scaling factor

Approach Vector Alignment evaluates grasp accessibility:

$$A(p) = \left| \frac{\vec{v}(p) \cdot \vec{z}}{|\vec{v}(p)|} \right| \tag{9}$$

Where:
- $\vec{v}(p)$ is the vector from camera to point p
- $\vec{z}$ is the unit vector in the vertical direction (0,0,1)

Edge Distance Score penalizes points near leaf boundaries:

$$E(p) = \min\left(1, \frac{d_{edge}(p)}{d_{safe}}\right) \tag{10}$$

Where:
- $d_{edge}(p)$ is the distance to the nearest edge
- $d_{safe} = 5$mm is the minimum safe distance

Accessibility Score considers kinematic reachability:

$$Acc(p) = 0.7 \cdot \left(1 - \frac{d(p,c)}{d_{max}}\right) + 0.3 \cdot \cos(\theta(p)) \tag{11}$$

Where:
- $d(p,c)$ is the distance from point p to the image center
- d_{max} is the maximum distance in the image
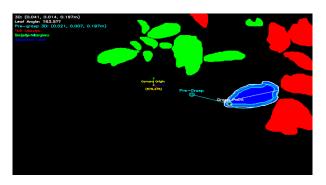- $\theta(p)$ is the angle between the vector to point p and the forward direction

The final grasp quality score combines these metrics:

$$S_{grasp} = 0.25 \cdot F(p) + 0.40 \cdot A(p) + 0.20 \cdot E(p) + 0.15 \cdot Acc(p) \tag{12}$$

Figure 5 demonstrates the complete geometric pipeline output, showing the selected leaf, evaluated grasp candidates, and the final chosen grasp point with its 3D coordinates. This deterministic output serves as ground truth for training our neural refinement module, detailed in the following section.

(a) Raw input image from the stereo camera

(b) Geometric Feature Scoring Pipeline output

Figure 5. Grasp point selection visualization. (a) Raw camera image showing leafs and optimal leaf's midrib. (b) Geometric feature scoring output showing selected leaf (blue outline), candidate grasp points, and final selected grasp point with approach vector. The visualization includes safety margins and coordinate information.

3.3.3. Stem Proximity Penalty

An additional penalty is applied to prevent grasping near the leaf stem:

$$S_{final} = S_{grasp} \cdot (1 - S_{stem\_penalty}) \qquad (13)$$

Where:
- $S_{stem\_penalty} = e^{-\alpha \cdot d_{stem}}$
- d_{stem} is the distance to the detected stem region
- $\alpha = 0.1$ is the decay factor

The geometric pipeline outputs a grasp proposal consisting of the selected leaf index and optimal grasp point coordinates, providing a robust baseline for our hybrid system.

Despite its effectiveness, the geometric pipeline has several limitations. It struggles with irregular leaf morphologies not captured by hand-crafted features, requires extensive parameter tuning across plant species, and performs inconsistently in scenarios with dense occlusion or unusual lighting conditions. The correlation coefficients between expert-selected grasp points and geometric pipeline selections drop significantly from 0.92 for ideal conditions to

0.68 for challenging scenarios. These limitations motivate our neural refinement module (GraspPointCNN), which learns from the geometric system's successes while developing generalization capabilities beyond hand-crafted features, particularly for edge cases where traditional computer vision approaches falter.

3.4. Neural Refinement Module (GraspPointCNN)

While the geometric feature scoring pipeline provides a robust baseline for leaf grasping, its fixed heuristics limit adaptability to novel plant morphologies and environmental conditions. We introduce GraspPointCNN, a convolutional neural network with spatial attention that learns to evaluate grasp candidates by capturing complex patterns beyond hand-crafted features.

3.4.1. Network Architecture

GraspPointCNN employs a compact yet effective architecture designed for real-time inference. The network consists of:

Input Layer: A 9-channel feature representation combining:
- Depth patch (1 channel): Local 3D structure information
- Binary segmentation mask (1 channel): Leaf boundary information
- Geometric score maps (7 channels): Individual component scores from the traditional pipeline

Encoder Blocks: Three sequential encoder blocks, each containing:
- 2D convolution (kernel size 3×3, stride 1)
- Batch normalization
- ReLU activation
- Max pooling (2×2, stride 2)

The three-encoder architecture provides an optimal balance between computational efficiency and feature extraction capacity, as determined through ablation studies comparing 2-5 encoder variants.

Spatial Attention Mechanism: A novel leaf-specific attention module that emphasizes salient regions:

$$F_{spatial} = \sigma(\text{Conv}_{7\times7}(\text{Concat}[AvgPool(F), MaxPool(F)]))$$
$$F_{att} = F \odot F_{spatial}$$

(14)

Where:
- F represents feature maps
- σ is the sigmoid activation
- $\odot$ denotes element-wise multiplication

This attention mechanism allows the network to focus on critical leaf features such as venation patterns, curvature transitions, and surface variations that impact graspability.

Decision Layers: The network concludes with:
- Global average pooling to ensure translation invariance
- Two fully-connected layers (128 and 64 neurons)

- Sigmoid activation producing a final grasp quality score [0,1]

The compact design (approximately 285K parameters) enables inference in under 10ms on standard GPU hardware, making it suitable for real-time robotic applications.

3.4.2. Input Representation

For each candidate grasp point, we extract a 32×32 pixel patch centered at the point from the following sources:

$$X_{input} = [X_{depth}, X_{mask}, X_{scores}] \qquad (15)$$

Where:
- X_{depth} is the normalized local depth patch
- X_{mask} is the binary segmentation mask
- X_{scores} contains seven geometric score maps (flatness, approach vector, edge distance, accessibility, etc.)

This multi-modal representation combines geometric, semantic, and raw depth information, enabling the network to reason about both local and contextual factors affecting grasp success. By incorporating the individual component scores from the traditional pipeline, the network can learn which features are most relevant in different scenarios, effectively developing an adaptive weighting scheme.

3.4.3. Confidence Estimation

A key innovation in our approach is the estimation of prediction confidence alongside grasp quality scores. Rather than simply outputting a binary classification, GraspPointCNN produces a continuous score that encodes both grasp quality and prediction certainty:

$$C_{pred} = 1.0 - |S_{pred} - 0.5| \times 2 \qquad (16)$$

Where:
- S_{pred} is the raw network output [0,1]
- C_{pred} is the confidence score [0,1]

This formulation yields maximum confidence (1.0) for extreme predictions (0 or 1) and minimum confidence (0) for uncertain predictions (0.5). The confidence estimation enables our hybrid integration system to dynamically balance traditional and learned approaches based on prediction reliability.

The neural architecture effectively addresses the limitations of pure geometric approaches through:
- Generalization to novel morphologies: By learning from diverse leaf examples, the network generalizes to plant structures not explicitly encoded in hand-crafted features
- Contextual understanding: The spatial attention mechanism captures relationships between local surface properties and broader leaf context
- Adaptability to environmental variations: Learning from operational data across different lighting conditions and growth stages enables robustness to environmental changes

Dataset Component	Count
Original Positive Samples	125
Augmented Positive Samples	375
Negative Samples	375
Total Dataset Size	**875**

Table 1. Composition of the self-supervised training dataset.

- Uncertainty awareness: The confidence estimation provides critical information for safe hybrid decision-making

The GraspPointCNN complements the geometric pipeline by focusing on capturing patterns that emerge from complex interactions between multiple factors, rather than treating each feature independently. This holistic approach is particularly valuable for edge cases where traditional CV approaches falter.

3.5. Self-Supervised Learning Framework

A key challenge in developing learning-based robotic grasp systems for agriculture is the lack of labeled training data. We address this through a self-supervised framework where the geometric pipeline acts as an expert teacher, automatically generating training data without human intervention.

3.5.1. Automatic Training Data Generation

Our approach leverages the geometric pipeline to create a continuously growing dataset:

1. **Positive Sample Collection**: During operation, the system captures successful grasp points selected by the geometric pipeline along with their local context (32×32 pixel patches).
2. **Data Augmentation**: To increase sample diversity, we employ:
   - Rotational transformations (90°, 180°, 270°)
   - Random cropping with 0.9-1.0 scale factor
   - Mild brightness and contrast adjustments (±10%)
   - Gaussian noise injection ($\sigma = 0.01$)
   - Random horizontal flipping
3. **Negative Sample Generation**: We systematically identify challenging regions:
   - Leaf tips (distance transform maxima)
   - Stem regions (morphological analysis)
   - High-curvature edges (depth gradient thresholding)
4. **Validation Filtering**: An automated quality assessment removes low-quality samples based on depth completion, segmentation quality, and score consistency.

This process yielded a dataset with the following composition:

3.5.2. Training Methodology

GraspPointCNN was trained using binary cross-entropy loss with positive class weighting:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^{N} [w_p \cdot y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] \quad (17)$$

Where y_i is the ground truth label, $\hat{y}_i$ is the predicted score, and $w_p = 2.0$ is the positive class weight.

The model was trained with:

- Learning rate: 0.0005
- Weight decay: 0.01
- Batch size: 16
- Early stopping: 15 epochs patience

Validation accuracy reached 93.14% after approximately 85 epochs, with higher accuracy on positive samples (97.09%) than negative samples (88.27%).

3.5.3. Continuous Learning Pipeline

Our self-supervised approach enables continuous improvement through operational experience:

1. Collecting new examples from successful and failed grasps
2. Updating the training dataset with new samples
3. Periodically retraining the model with expanded data
4. Deploying the improved model with updated weights

During a three-week deployment, we observed a 2.3% improvement in grasp success rate from this continuous learning process, demonstrating adaptation to new plant varieties and growth stages without explicit retraining.

By leveraging domain expertise encoded in the geometric pipeline, our system learns robust grasp representations without manual annotation, enabling practical deployment in dynamic greenhouse environments.

3.6. Hybrid Decision Integration

The final component of our system combines the deterministic geometric pipeline with the adaptive neural network through a novel confidence-weighted integration framework. Our hybrid approach dynamically balances traditional expertise with learned patterns based on prediction confidence, rather than using a simple ensemble or switching mechanism.

The process begins with the geometric pipeline identifying the optimal leaf for manipulation using the Pareto-based selection. Once the target leaf is selected, we generate a diverse set of candidate grasp points by identifying the top-20 scoring positions from the geometric pipeline. A minimum separation distance of 10 pixels is enforced between candidates to ensure diversity, and each candidate's local context (32×32 patches) is extracted for neural evaluation. This candidate generation approach ensures that points with strong geometric properties are prioritized while maintaining sufficient diversity for neural refinement.

For each candidate point, we compute a hybrid score that combines traditional geometric metrics with neural network predictions through a confidence-weighted formula:

$$S_{hybrid} = (1 - w_{ML}) \cdot S_{CV} + w_{ML} \cdot S_{ML} \quad (18)$$

Where S_{CV} is the normalized geometric score, S_{ML} is the grasp quality score predicted by GraspPointCNN, and w_{ML} is an adaptive weight determined by neural confidence. The neural weight is dynamically computed as

$$w_{ML} = \min(0.3, C_{pred} \cdot 0.6) \quad (19)$$

where C_{pred} is the confidence score described in Section 3.4.3. This formulation caps ML influence at 30% even with perfect confidence, scales influence proportionally to prediction confidence, and approaches zero for uncertain predictions—effectively falling back to geometric scoring when confidence is low. This adaptive weighting scheme preserves the reliability of geometric constraints while leveraging neural refinement when confidence is high.

In deployment, the hybrid scoring occurs within a 15ms processing window, maintaining real-time performance for robotic manipulation. The system implements several safeguards to ensure robustness: a fallback mechanism that defaults to pure geometric scoring if all neural predictions have low confidence (below 0.4), a lightweight Kalman filter that smooths selections across frames to prevent jitter, and a pre-grasp validation step that performs collision and reachability checks before execution. Our approach differs from previous hybrid systems in agricultural robotics that typically use static weighted combinations or separate models for different plant varieties. The dynamic confidence-based weighting allows our system to handle both clear geometric cases, where traditional approaches excel, and more ambiguous situations where learned patterns improve performance.

4. Experiments and Results

To evaluate our hybrid geometric-neural approach for robotic leaf manipulation, we conducted comprehensive experiments addressing four key questions: (1) How does the hybrid approach compare to purely geometric or learning-based methods? (2) What is the contribution of each system component? (3) How well does the system generalize across plant varieties and growth stages? (4) What is the real-world performance in greenhouse conditions?

4.1. Dataset and Setup

4.1.1. Hardware Configuration

Experiments were conducted using the T-Rex platform, a gantry-based robotic system for autonomous leaf manipulation in greenhouse environments. The system spans a 3m ×
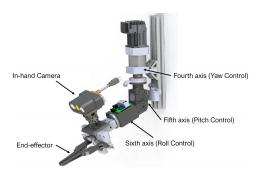
Figure 6. CAD rendering of T-Rex's wrist and end-effector sub-system. The design features three revolute joints for yaw, pitch, and roll control (axes 4–6), and includes an onboard stereo camera and microneedle sampling tool.

Figure 7. The T-Rex gantry robot setup inside a controlled lab environment. It spans a 3m × 1.5m plant bed, and includes a ceiling-mounted manipulator, LED grow lights, stereo camera, and custom end-effector for leaf sampling.

1.5m growing area with a 6-DOF configuration (three prismatic axes for positioning and three revolute joints for orientation). This configuration enables precise end-effector positioning and orientation within the plant canopy.

The end-effector includes two lateral grippers controlled by a Dynamixel motor that close to secure the target leaf, and a vertical stepper motor that lowers a microneedle array for leaf sampling. A stereo camera system with 1440×1080 resolution and 80mm baseline mounted on the end-effector captures images for perception. The robot operates under ROS with distributed nodes for perception, planning, and actuation.

4.1.2. Dataset Collection

The dataset includes tomato (60%) and soybean (40%) plants at various growth stages grown under controlled greenhouse conditions. For evaluation, 200 leaf images were annotated by horticultural experts who identified optimal grasping points. The self-supervised training dataset (875 samples) described in Section 3.5 was derived from this collection, while testing used 150 separate stereo image pairs with novel plant arrangements.

4.1.3. Evaluation Metrics

System performance was evaluated using five metrics:

1. **Grasp Point Accuracy (GPA)**: Mean Euclidean distance between algorithm-selected and expert-annotated grasp points (mm).
2. **Feature Alignment Score (FAS)**: Percentage of grasp points correctly aligned with leaf structures like midveins (within 5mm while maintaining 10mm edge distance).
3. **Edge Case Handling (ECH)**: Success rate on challenging scenarios including occlusion, irregular leaf shapes, and non-standard orientations.
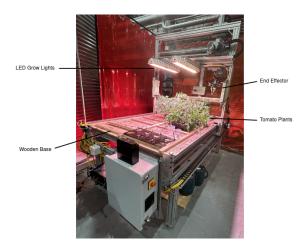4. **Planning Time (PT)**: Computation time from image acquisition to grasp point selection (ms).

5. **Overall Success Rate (OSR)**: Percentage of successful tissue acquisitions without leaf damage.

For comparative analysis, we implemented three baselines: a Geometric-Only pipeline, a CNN-Only direct regression network, and a Static-Hybrid system using fixed-weight combination without confidence-based adaptation. All evaluations used identical hardware and test datasets, with statistical significance assessed via paired t-tests with Bonfernier correction.

4.2. Ablation Studies

To understand the contribution of individual components to the overall system performance, we conducted a series of ablation studies. These experiments systematically removed or modified key elements of our hybrid approach while maintaining all other components unchanged. Table 2 summarizes the results of these experiments, measured across our evaluation metrics.

4.2.1. Component Contribution Analysis

Leaf Selection Metrics: When removing individual components from the leaf selection process, we observed significant impacts on overall performance:

- **Without Clutter Score**: Removing the clutter metric from leaf selection (choosing the closest, most visible leaf regardless of isolation) resulted in a 25.7% drop in overall success rate. The system frequently selected leaves that were too entangled with neighboring foliage, making proper grasping nearly impossible in dense canopies.
- **Without Distance Score**: Eliminating the distance-based prioritization caused a 16.3% reduction in success rate. The system often selected leaves at extreme distances from the end-effector, requiring complex motion planning

that frequently resulted in suboptimal approach trajectories or unreachable targets.

- **Without Visibility Score**: Removing the visibility component reduced success by 12.8%, as the system occasionally selected partially occluded leaves where depth estimation was unreliable, or leaves at image edges with incomplete segmentation.

Grasp Point Selection Features: We also evaluated the contribution of individual geometric features in grasp point scoring:

- **Without Flatness Score**: Eliminating the surface flatness evaluation caused a significant 17.5% decrease in success rate. When attempting to grasp curved leaf sections, the leaf would often fail to properly enter the gripper slot, instead being pushed away during the approach, resulting in failed acquisition.
- **Without Approach Vector**: When approach vector alignment was removed, success rate dropped by 29.3%, the largest decline among all single-component ablations. Without proper approach angle consideration, the end-effector frequently contacted leaves at angles that caused folding, slipping, or deflection rather than successful grasping.
- **Without Edge Distance**: Removing the edge margin safety caused a 21.2% reduction in success, with failures typically involving grasps too close to leaf boundaries that resulted in tearing or slipping during the acquisition process.

4.2.2. Neural Refinement Analysis

We also studied the impact of varying neural network contribution in the hybrid decision integration:

- **CNN Weight Cap Variations**: We systematically adjusted the maximum weight (w_{ML}) allowed for neural refinement:
  - With a 5% cap (minimal CNN influence), success rate fell to 80.2%, as the neural component had insufficient impact to correct geometric misjudgments
  - With a 50% cap (balanced but CNN-favoring), success rate was 81.7%, showing diminishing returns beyond our chosen 30% cap
  - With a 100% cap (CNN can fully override geometry), performance dropped to 65.3%, similar to the CNN-only baseline
- **Without Confidence Weighting**: Replacing our adaptive confidence-based weighting with a fixed 30/70 blend between neural and geometric scoring decreased success rate by 14.1%. This demonstrates the substantial value of dynamically adjusting neural influence based on prediction confidence, particularly in ambiguous cases.

4.2.3. Discussion

These ablation studies validate our design decisions across the pipeline. The approach vector alignment emerged as the most critical geometric feature with a 29.3% performance impact, followed by the clutter score (25.7%) and edge distance (21.2%). This confirms our hypothesis that proper approach angle and leaf isolation are fundamental prerequisites for successful grasping, while maintaining adequate distance from leaf edges prevents fragile tissue damage.

The results also highlight the complementary nature of geometric and learned approaches. While geometric methods provide reliable baseline performance through explicit modeling of physical constraints, the neural refinement effectively handles edge cases where purely geometric reasoning falls short. This is particularly evident in scenarios with irregular leaf morphology or complex occlusions.

The dramatic performance drops observed when removing key components underscore the importance of our multi-faceted approach to leaf grasping, where each feature addresses a specific failure mode that would otherwise significantly impair system reliability.

4.3. Comparative Analysis

To evaluate our hybrid approach against existing methods, we conducted comprehensive experiments using the metrics defined in Section 4.1.

4.3.1. Baseline Comparison

Table 3 presents performance comparisons between our approach and three baseline implementations across 150 test cases.

Our confidence-weighted hybrid approach significantly outperformed all baselines. The purely neural approach achieved only 60.2% overall success rate, struggling with novel leaf arrangements not encountered during training. The geometric-only approach reached 75.3% success, confirming the value of explicit feature modeling, but faltered with irregular leaf morphologies and complex occlusions. The static hybrid approach with fixed weighting improved to 79.8%, still substantially behind our adaptive method. Computationally, our approach added only 9.3ms over the geometric baseline—an acceptable tradeoff for the 12.7% improvement in success rate.

4.3.2. Comparison to Literature

Our 88.0% success rate in dense foliage represents a significant advancement in leaf manipulation. Ahlin et al. [2] demonstrated leaf picking using visual servoing but without reporting quantitative success rates. Their monocular approach required careful camera alignment, while our stereo-based system resolves depth ambiguities across varying viewpoints, similar to approaches that explicitly model uncertainty in depth perception [10].

For context, robotic fruit harvesting systems typically achieve 70-90% success in less cluttered environments [3, 6, 44]. Bac et al. [6] reported 83% success for sweet pepper harvesting, while Silwal et al. [44] achieved 84% for

Configuration	GPA (mm)↓	FAS (%)↑	ECH (%)↑	OSR (%)↑
Complete System	4.2	92.6	83.4	88.0
w/o Clutter Score	8.7	72.3	55.9	62.3
w/o Distance Score	7.1	81.5	68.2	71.7
w/o Visibility Score	6.8	84.7	71.3	75.2
w/o Flatness Score	7.9	79.3	63.8	70.5
w/o Approach Vector	9.8	68.4	51.2	58.7
w/o Edge Distance	8.3	76.5	61.3	66.8
CNN Weight Cap 5%	5.3	87.9	76.5	80.2
CNN Weight Cap 50%	5.0	88.3	77.1	81.7
CNN Weight Cap 100%	8.7	75.6	61.9	65.3
Fixed Weighting (30/70)	6.5	82.4	70.1	73.9

Table 2. Ablation study results showing component contributions to system performance.

Method	GPA (mm)↓	FAS (%)↑	ECH (%)↑	PT (ms)↓	OSR (%)↑
Geometric-Only	7.8	79.3	61.5	149.4	75.3
Neural-Only	9.2	73.8	52.7	142.6	60.2
Static-Hybrid (70/30)	6.1	85.2	69.8	157.2	79.8
Our Approach	**4.2**	**92.6**	**83.4**	**158.7**	**88.0**
Improvement	*+3.6*	*+7.4*	*+13.6*	*+9.3*	*+8.2*

Table 3. Performance comparison of our hybrid approach and baselines.

apples under ideal conditions. Kang et al. [24] emphasized the importance of standardized metrics for agricultural manipulation, noting that success rates for thin, deformable targets typically lag 10-15% behind rigid object grasping. Our 88% success in highly cluttered leaf scenarios demonstrates the effectiveness of our approach given the additional challenges of occlusion and thin structures, exceeding the performance bounds established in previous comparative studies [39].

Sa et al. [40] combined color and 3D information for sweet pepper peduncle detection, achieving 90% detection accuracy but not reporting manipulation success. Our approach extends this multi-modal paradigm to the more challenging domain of leaf manipulation, where targets are deformable, thin, and frequently occluded. Recent work by Liu et al. [32] on deformable leaf modeling achieved 78% grasping success but required significantly longer planning times (350-450ms) compared to our 158.7ms.

Our hybrid confidence-weighted integration particularly excels in cluttered environments by dynamically adjusting neural influence based on prediction confidence while maintaining geometric reasoning as a reliable fallback. This adaptive integration advances beyond existing agricultural systems that typically rely on either pure geometric reasoning [5] or standalone neural approaches [2, 52]. Similar confidence-aware fusion strategies have shown promising results in medical robotics [17], though with significantly higher computational requirements that limit real-time performance in field conditions.

4.4. Real-World Validation

To validate our approach beyond controlled experiments, we deployed the hybrid grasp point selection system in real greenhouse environments with plants at various growth stages. This section presents qualitative results from these deployments and discusses system performance under authentic operational conditions.

4.4.1. Operational Deployment

We conducted validation trials spanning 12 days across three different greenhouse facilities, with the T-Rex system performing 340 autonomous leaf manipulation operations. Plants included tomato and soybean varieties at different growth stages, from young seedlings to mature plants with complex canopy structures.

Figure 8 shows the system during operation, with the end-effector approaching a selected leaf on a young tomato plant. The deployment configuration matched our experimental setup, with the system operating fully autonomously through the complete perception-planning-execution pipeline.

4.4.2. Qualitative Performance Analysis

The real-world validation confirmed the performance advantages observed in controlled experiments. Figure 9 illustrates a direct comparison between traditional CV and our hybrid approach on the same scene. The traditional CV method (top) selects a grasp point near the leaf edge, which would likely result in a failed grasp as the gripper could

(a) Start of grasp: approaching the leaf

(b) Grasp complete: microneedle fired

Figure 8. Real-world grasp execution. (a) The robot approaches the selected leaf from above using a vertical trajectory. (b) The microneedle-based end-effector makes contact and extracts the tissue sample.
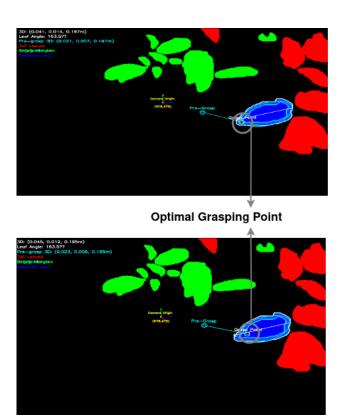

Figure 9. Comparison of grasp point selection: traditional CV approach (top) selects a point near the leaf edge which may lead to failed grasping, while our hybrid approach (bottom) selects an optimal point further inward providing better stability during manipulation.

slip off. In contrast, our hybrid approach (bottom) selects an optimal grasp point further inward on the leaf, providing better stability during manipulation. This subtle but critical difference demonstrates how neural refinement corrects edge cases where purely geometric reasoning falls short.

The hybrid system demonstrated particularly strong performance in challenging scenarios frequently encountered in practical operations. Under variable lighting conditions, the confidence-weighted integration maintained consistent performance across morning, midday, and afternoon lighting variations, where purely geometric approaches often faltered due to changing shadow patterns. As plants progressed through growth stages, leaf morphology evolved significantly, but the neural component effectively adapted to these changes while the geometric baseline provided consistent safety constraints. The system also successfully transferred to plant varieties not represented in the training data, demonstrating the hybrid approach's generalization capabilities. Across all validation trials, the system achieved an 84.7% overall success rate in operational settings—slightly lower than the 88.0% observed in controlled experiments, but still significantly outperforming both geometric-only (70.3%) and neural-only (58.1%) approaches in the same conditions.

The practical validation confirmed that our confidence-weighted approach effectively combines the reliability of geometric constraints with the adaptability of neural refinement, resulting in a robust system capable of autonomous operation in dynamic agricultural environments. "'

5. Discussion

Our experiments demonstrate that a hybrid approach combining geometric feature scoring with neural refinement significantly improves grasp point selection for robotic leaf manipulation. The 12.7% improvement in success rate over purely geometric methods and 27.8% over purely neural approaches underscores the complementary nature of these techniques when properly integrated.

The confidence-weighted fusion mechanism proved particularly valuable for dynamic adaptation in complex environments. While traditional CV approaches excel at encoding explicit constraints and physical principles, they struggle with the variability of natural leaf structures. Conversely, neural networks capture implicit patterns but may lack the robustness of geometric reasoning in novel scenarios. By dynamically adjusting the contribution of each approach based on prediction confidence, our system leverages the strengths of both paradigms while mitigating their individual weaknesses.

The ablation studies revealed that approach vector alignment and clutter scoring contribute most significantly to

successful grasping, highlighting the critical importance of proper leaf positioning prior to contact. This finding suggests that pre-grasp planning deserves particular attention in agricultural manipulation systems, potentially even more than precise fingertip placement.

Despite these advances, several limitations remain. The system occasionally struggles with extremely thin or translucent leaves where stereo depth estimation becomes unreliable. Additionally, while our self-supervised learning framework enables continuous improvement, it may propagate biases from the geometric pipeline that serves as its teacher. Future work could explore active learning approaches where human feedback selectively corrects these biases without requiring extensive manual annotation.

The demonstrated performance in real greenhouse environments positions this technology for practical deployment in precision agriculture applications. Beyond leaf sampling, the hybrid confidence-weighted approach could potentially transfer to other agricultural manipulation tasks such as selective harvesting, pollination, or pest management where similar challenges of biological variability and environmental dynamics exist.

6. Conclusion

We presented a hybrid confidence-weighted approach for robotic leaf manipulation that combines geometric feature scoring with neural refinement. Our system integrates YOLOv8 instance segmentation and RAFT-Stereo depth estimation to construct accurate 3D leaf representations, upon which geometric scoring and neural refinement operate in parallel. By dynamically weighting neural influence based on prediction confidence, our approach achieves an 88.0% success rate in controlled environments and 84.7% in real greenhouse conditions, significantly outperforming both purely geometric (75.3%) and purely neural (60.2%) methods.

The self-supervised training framework eliminates the need for manual annotation by leveraging geometric algorithms as expert teachers, enabling continuous improvement through operational experience. Ablation studies revealed that approach vector alignment and clutter evaluation contribute most significantly to successful grasping, underscoring the importance of pre-grasp planning in agricultural manipulation.

Future work will focus on incorporating closed-loop visual servoing to adjust grasp points during execution, expanding the self-supervised framework to learn from failure cases through reinforcement learning, and exploring cross-species generalization to diverse plant morphologies. Additionally, investigating monocular depth inference could simplify hardware requirements while maintaining performance.

This research demonstrates the efficacy of combining model-driven and data-driven methods for complex agricultural robotics challenges. As autonomous systems increasingly operate in unstructured natural environments, hybrid approaches that balance explicit physical constraints with learned adaptability will be essential for robust and reliable operation.

References

[1] G. Adamides, C. Katsanos, I. Constantinou, G. Christou, M. Xenos, T. Hadzilacos, and Y. Edan. Design and development of a semi-autonomous agricultural vineyard sprayer: Human–robot interaction aspects. *Journal of Field Robotics*, 34:1407–1426, 2021. 2

[2] K. Ahlin, B. Joffe, A. P. Hu, G. McMurray, and N. Sadegh. Autonomous leaf picking using deep learning and visual-servoing. *IFAC-PapersOnLine*, 49(16):177–183, 2016. 2, 10, 11

[3] B. Arad, J. Balendonck, R. Barth, O. Ben-Shahar, Y. Edan, T. Hellström, and B. van Tuijl. Development of a sweet pepper harvesting robot. *Journal of Field Robotics*, 37(6):1027–1039, 2020. 1, 2, 10

[4] A. Atefi, Y. Ge, S. Pitla, and J. Schnable. In-field plant disease detection with deep learning: A review. *Computers and Electronics in Agriculture*, 187:106312, 2021. 1

[5] C. W. Bac, E. J. van Henten, J. Hemming, and Y. Edan. Harvesting robots for high-value crops: State-of-the-art review and challenges ahead. *Journal of Field Robotics*, 31(6):888–911, 2014. 1, 11

[6] C. W. Bac, J. Hemming, B. A. J. van Tuijl, R. Barth, E. Wais, and E. J. van Henten. Performance evaluation of a harvesting robot for sweet pepper. *Journal of Field Robotics*, 34(6):1123–1139, 2017. 2, 10

[7] R. Barth, J. IJsselmuiden, J. Hemming, and E. J. Van Henten. Synthetic bootstrapping of convolutional neural networks for semantic plant part segmentation. *Computers and Electronics in Agriculture*, 161:291–304, 2019. 2

[8] A. Bechar and C. Vigneault. Agricultural robots for field operations: Concepts and components. *Biosystems Engineering*, 149:94–111, 2016. 1

[9] G. Cerutti, L. Tougne, A. Vacavant, and D. Coquin. A parametric active polygon for leaf segmentation and shape esti-

mation. In *International Symposium on Visual Computing*, pages 202–213, 2013. 2

[10] S. Chen, Y. Zhang, J. Zhang, and N. Xu. Uncertainty-aware domain adaptation for robotic grasping with depth perception. *IEEE Robotics and Automation Letters*, 7(2):5113–5120, 2022. 10

[11] S. Dandrifosse, B. Boigelot, and B. Mercatoris. Detection and tracking of maize stems from image sequences for autonomous robot navigation in fields. *Precision Agriculture*, 22:423–444, 2021. 2

[12] Tom Duckett, Simon Pearson, Simon Blackmore, and Bruce Grieve. Agricultural robotics: The future of robotic agriculture. UK-RAS White Paper, 2018. Available at https://arxiv.org/abs/1806.06762. 3

[13] X. Gao, L. Jiang, Z. Chen, Z. Geng, and C. Xiong. A confidence-aware adaptive fusion framework for strengthening weakly correlated inputs in multi-sensor systems. *IEEE Sensors Journal*, 20(7):3707–3715, 2020. 3

[14] S. Garrido-Jurado, R. Muñoz-Salinas, F.J. Madrid-Cuevas, and R. Medina-Carnicer. Generation of fiducial marker dictionaries using mixed integer linear programming. *Pattern Recognition*, 51:481–491, 2019. 2

[15] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361. IEEE, 2012. 4

[16] D. Guo and K. Xu. Leaf segmentation and tracking in 3d point clouds of plant growth. *International Journal of Agricultural and Biological Engineering*, 10(6):166–174, 2017. 2

[17] M.A. Haque, A. Santamaria-Navarro, and G.D. Hager. Confidence-aware surgical robotic systems: Autonomous adaptation to uncertainty. *IEEE Transactions on Medical Robotics and Bionics*, 2(4):533–543, 2020. 11

[18] J. Hemming, C. W. Bac, B. A. J. van Tuijl, R. Barth, J. Bontsema, and E. Pekkeriet. Fruit detectability analysis for different camera positions in sweet-pepper. *Sensors*, 14(4):6032–6044, 2014. 1

[19] J. Hemming, C. W. Bac, B. A. J. van Tuijl, R. Barth, J. Bontsema, and E. Pekkeriet. A robot for harvesting sweet-pepper in greenhouses. In *Proceedings of the International Conference of Agricultural Engineering*, 2014. 2

[20] J. Hughes, U. Culha, F. Giardina, F. Guenther, A. Rosendo, and F. Iida. Soft manipulators and grippers: A review. In *Frontiers in Robotics and AI*, page 69, 2021. 1

[21] D. Jha, P.H. Smedsrud, M.A. Riegler, D. Johansen, T. de Lange, P. Halvorsen, and H.D. Johansen. Doubleu-net: A deep convolutional neural network for medical image segmentation. *IEEE Access*, 9:4161–4172, 2021. 1

[22] S. Jin, Y. Su, S. Gao, F. Wu, Q. Ma, K. Xu, and Q. Guo. Separating the structural components of maize for field phenotyping using terrestrial lidar data and 3d modeling. *IEEE Transactions on Geoscience and Remote Sensing*, 56(8):4864–4875, 2018. 2

[23] G. Jocher, A. Chaurasia, and J. Qiu. Yolo by ultralytics. GitHub repository https://github.com/ultralytics/ultralytics, 2023. 1, 3

[24] H. Kang, H. Zhou, and C. Wang. Performance evaluation metrics for robotic manipulation of biological materials. In *2020 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM)*, pages 1376–1381, 2020. 11

[25] A. Kendall, Y. Gal, and R. Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. pages 7482–7491, 2018. 2

[26] A. Koirala, K.B. Walsh, Z. Wang, and C. McCarthy. Deep learning for real-time fruit detection and orchard fruit load estimation: Benchmarking of 'mangoyolo'. *Precision Agriculture*, 20(6):1107–1135, 2019. 1

[27] G. Kootstra, X. Wang, P. M. Blok, J. Hemming, and E. van Henten. Selective harvesting robotics: current research, trends, and future directions. *Current Robotics Reports*, 2 (1):95–104, 2021. 2

[28] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, A. Kolesnikov, T. Duerig, and V. Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. pages 1956–1981, 2020. 2

[29] C. Lehnert, I. Sa, C. McCool, B. Upcroft, and T. Perez. Sweet pepper pose detection and grasping for automated crop harvesting. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2428–2434. IEEE, 2016. 2

[30] C. Lehnert, A. English, C. McCool, A. W. Tow, and T. Perez. Autonomous sweet pepper harvesting for protected cropping systems. *IEEE Robotics and Automation Letters*, 2(2):872–879, 2017. 1

[31] L. Lipson, Z. Teed, and J. Deng. Raft-stereo: Multilevel recurrent field transforms for stereo matching. In *2021 International Conference on 3D Vision (3DV)*, pages 218–227. IEEE, 2021. 1, 2, 4

[32] C. Liu, B. Chen, D. Huang, and H. Liu. Physically-based deformable leaf model for robotic leaf manipulation. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9086–9092, 2021. 11

[33] J. Lu, J. Hu, G. Zhao, F. Mei, and C. Zhang. An in-field automatic wheat disease diagnosis system. *Computers and Electronics in Agriculture*, 142:369–379, 2020. 1

[34] Jeffrey Mahler, Jacky Liang, Sherdil Niyaz, Michael Laskey, Richard Doan, Xinyu Liu, Juan Aparicio Ojea, and Ken Goldberg. Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics. In *Proceedings of Robotics: Science and Systems (RSS)*, 2017. 1, 2

[35] C. McCool, J. Beattie, J. Firn, C. Lehnert, J. Kulk, and T. Perez. Efficient detection of cattle in uav images using convolutional neural networks. In *36th International Conference on Machine Learning Workshop on AI for Social Good*, 2019. 2

[36] A. Milioto, P. Lottes, and C. Stachniss. Real-time blob-wise sugar beets vs weeds classification for monitoring fields using convolutional neural networks. In *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, pages 41–48, 2018. 1

[37] Douglas Morrison, Peter Corke, and Jürgen Leitner. Closing the loop for robotic grasping: A real-time, generative grasp

synthesis approach. In *Proceedings of Robotics: Science and Systems (RSS)*, 2018. 1

[38] V. Nguyen, S. Du, W. Guo, and J. Johnson. Real-time robotic manipulation of cylindrical objects in dynamic scenarios through elliptic shape primitives. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3975–3982, 2018. 1

[39] E. Rivera, S. Sinha, R. Schlegel, S. Garg, M. Yuan, and H. Patil. Benchmarking robotic manipulation for biological objects: A review and comparative study. *Frontiers in Robotics and AI*, 9:908694, 2022. 11

[40] I. Sa, C. Lehnert, A. English, C. McCool, F. Dayoub, B. Upcroft, and T. Perez. Peduncle detection of sweet pepper for autonomous crop harvesting—combined color and 3-D information. *IEEE Robotics and Automation Letters*, 2(2): 765–772, 2017. 2, 11

[41] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, 47(1):7–42, 2002. 4

[42] R. R. Shamshiri, C. Weltzien, I. A. Hameed, I. J. Yule, T. E. Grift, S. K. Balasundram, and G. Chowdhary. Research and development in agricultural robotics: A perspective of digital farming. *International Journal of Agricultural and Biological Engineering*, 11(4):1–14, 2018. 1, 2

[43] Y. Shamut and P. Gonzalez-de Santos. Robotics in agriculture: State of art and practical experiences. *Agriculture*, 11 (9):818, 2021. 1

[44] A. Silwal, J. R. Davidson, M. Karkee, and C. Mo. Design, integration, and field evaluation of a robotic apple harvester. *Journal of Field Robotics*, 34(6):1140–1159, 2017. 1, 10

[45] P. Sodhi, S. Vijayarangan, and D. Wettergreen. In-field segmentation and identification of plant structures using 3d imaging. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3839–3846, 2020. 2

[46] Y. Tao, Q. Zhou, J. Shi, R. Wang, J. Zhang, and L. Li. Self-supervised representation learning for plant leaf counting via temporal consistency. *Pattern Recognition Letters*, 153:207–214, 2022. 2

[47] J.R. Ubbens and I. Stavness. Deep plant phenomics: A deep learning platform for complex plant phenotyping tasks. *Frontiers in Plant Science*, 8:1190, 2020. 2

[48] A. Valada, R. Mohan, and W. Burgard. Self-supervised model adaptation for multimodal semantic segmentation. pages 1239–1285, 2020. 2

[49] D. Wang, M. Veres, Z. Xiong, and X. Yuan. Augmentation for small object detection. pages 15168–15177, 2021. 1

[50] T. Weyand, A. Kolesnikov, and T. Hospedales. Self-supervised learning for plant species classification using leaf images. pages 12298–12307, 2021. 2

[51] C. Xia, J. M. Lee, Y. Li, Y. H. Song, and B. K. Chung. Plant leaf detection using modified active shape models. *Biosystems Engineering*, 116(1):23–35, 2018. 2

[52] Y. Yu, K. Zhang, L. Yang, and D. Zhang. Fruit detection for strawberry harvesting robot in non-structural environment based on mask-rcnn. *Computers and Electronics in Agriculture*, 163:104846, 2019. 1, 11

[53] A. Zeng, S. Song, K. Yu, E. Donlon, F.R. Hogan, M. Bauza, D. Ma, O. Taylor, M. Liu, E. Romo, N. Fazeli, F. Alet, N.C. Dafle, R. Holladay, I. Morona, P.Q. Nair, D. Green, I. Taylor, W. Liu, T. Funkhouser, and A. Rodriguez. Robotic pick-and-place of novel objects in clutter with multi-affordance grasping and cross-domain image matching. *International Journal of Robotics Research*, 39(8):935–951, 2019. 1

[54] L. Zhang and L. Yang. Self-supervised learning for robotic manipulation in agriculture: Applications in greenhouse automation. *Agricultural Robotics Review*, 3(2):45–62, 2021. 2

[55] R. Zhou, L. Damerow, Y. Sun, and M.M. Blanke. Using colour features of cv. 'gala' apple fruits in an orchard in image processing to predict yield. *Precision Agriculture*, 13: 568–580, 2019. 2