Ground-Truth-Free Relative Yield Estimation of Cacao Pods Using YOLOv8-Based Morphological and Self-Supervised Visual Features

Francis Wedemeyer Dayagro
College of Computer Studies
Cebu Institute of Technology - University
Cebu, Philippines
franciswedemeyer.dayagro@cit.edu

Felicity Orate
College of Computer Studies
Cebu Institute of Technology - University
Cebu, Philippines
felicity.orate@cit.edu

Zedric Marc Tabinas
College of Computer Studies
Cebu Institute of Technology - University
Cebu, Philippines
zedricmarc.tabinas@cit.edu

Moriel Edgar Deandre Bien
College of Computer Studies
Cebu Institute of Technology - University
Cebu, Philippines
morieledgardeandre.bien@cit.edu

***Abstract*—**

Cacao pod yield estimation is a critical component of farm planning, harvest optimization, and supply chain forecasting, yet conventional approaches rely heavily on manual measurements, destructive sampling, or historical averages that are labor-intensive and often inaccurate. This study proposes a ground-truth-free framework for relative yield estimation of cacao pods using YOLOv8-based object detection combined with morphological analysis and self-supervised visual feature learning. The system detects individual cacao pods from field images and extracts geometric and visual attributes—such as pod size, shape, surface texture, and spatial distribution—without requiring explicit yield labels. Self-supervised representation learning is employed to capture latent structural cues related to pod maturity and biomass, enabling relative yield comparison across trees and plots. The methodology involves automated pod detection, feature embedding generation, clustering and ranking for yield estimation, and validation through consistency analysis and expert-informed field observations rather than direct weight measurements. The proposed approach offers a scalable, low-cost, and practical decision-support tool for cacao yield monitoring, particularly in data-scarce agricultural environments

Keywords—Artificial Intelligence, Computer Vision, Self-Supervised Learning, Object Detection, Cacao Pods, Yield Estimation

I. INTRODUCTION

Cacao is a crucial agricultural crop that sustains smallholder farmers, yet reliable yield estimation remains a major challenge in many cacao-producing regions, particularly in the Philippines. Conventional yield assessment methods rely on manual pod counting, destructive sampling, or post-harvest measurements, which are labor-intensive, error-prone, and impractical for timely decision-making. Variations in pod size, maturity, and field conditions—driven by climate, soil health, and farm management—further reduce estimation accuracy. Recent advances in computer vision and deep learning, especially real-time object detection models such as YOLOv8, offer promising non-destructive alternatives for yield monitoring. By combining pod detection with morphological analysis and self-supervised visual feature learning, yield-related patterns can be inferred without requiring explicit ground-truth labels. However, challenges related to visual variability, occlusion, and scalability remain. This study proposes a ground-truth-free framework for relative cacao pod yield estimation to support practical, low-cost, and scalable yield monitoring in data-scarce farming environments.

II. REVIEW OF RELATED LITERATURE

A. Computer Vision Applications in Agriculture

B. Fruit and Pod-Level Morphological Analysis Using Images

C. Yield Estimation Methods and Ground-Truth Dependence

D. Self-Supervised Learning for Low-Annotation Agricultural Vision Tasks

E. Relative Ranking and Decision-Oriented AI Systems in Agriculture

F. Limitations of Conventional Cacao Yield Assessment Methods

G. Conclusion

III. RESEARCH DESIGN AND METHODOLOGY

A. Research Design

The research adopts a multi-stage experimental research design that integrates computer vision, self-supervised representation learning, and heuristic-based analysis to estimate relative cacao pod yield from field images. The proposed framework is designed to operate without ground-truth yield labels during model training, addressing the practical constraints of data-scarce agricultural environments.

The methodology consists of four sequential stages: (1) cacao pod detection using a YOLOv8 object detection model, (2) extraction of morphological features from detected pods, (3) extraction of self-supervised visual embeddings using SimCLR, and (4) fusion of these features followed by heuristic-based ranking to produce relative yield estimates. The ranking output provides an ordering of pods by inferred yield potential rather than an absolute yield measurement.

The study is *ground-truth-free with respect to yield estimation*. No pod weight, bean count, or fullness labels are used during detection training, feature learning, or ranking. Human expert input is introduced only at the evaluation stage as an external reference to assess whether the system's relative rankings are agriculturally reasonable.

B. Research Questions and Hypotheses

Research Question 1: How accurately can YOLOv8 detect cacao pods in the images?
Hypothesis (H1): YOLOv8 will achieve high precision and recall in pod detection, reflecting the state-of-the-art real-time object detection capability of modern YOLO models.

Research Question 2: Do morphological features (pod area, aspect ratio, etc.) correlate with pod fullness?
Hypothesis (H2): Larger pod area and certain aspect ratios will be indicative of fuller pods, so these simple features will positively correlate with the expert's fullness ranking.

Research Question 3: Can self-supervised (SimCLR) embeddings capture visual cues of pod fullness without labels?
Hypothesis (H3): SimCLR will produce embeddings where visually similar pods (in terms of fullness) lie closer together in the learned feature space, even though no explicit fullness labels are used during training.

Research Question 4: Does combining morphological and SimCLR features improve ranking accuracy?
Hypothesis (H4): The fused feature set will yield higher agreement with the expert ranking than using either feature set alone, since each modality provides complementary information.

C. Research Variables

Independent Variables: The inputs to the models are the image-derived features of each pod. After detection, each pod is represented by two independent feature vectors: (a) a morphological vector (e.g. pod area, bounding-box width, height, aspect ratio) and (b) a visual embedding vector (e.g. 128-D SimCLR embedding).

Dependent Variables: The primary dependent variable is the model-generated relative yield ranking of pods within an image. Secondary dependent variables include object detection metrics (precision, recall, and mAP) used to evaluate the reliability of the detection stage.

Control Variables: Data splits, preprocessing procedures, model architectures, and training hyperparameters are held constant across experiments, except where intentionally modified for ablation analysis. Environmental factors such as lighting, occlusion, and background variation are not explicitly controlled but are assumed to reflect realistic field conditions.

D. Data Collection and Labeling

The study utilizes the publicly available Cacao Tupi Dataset hosted on Roboflow, which contains approximately 2,900 field images of cacao pods captured under natural conditions. Each image includes bounding-box annotations for the class "Cacao-Pod," which are used exclusively to train and evaluate the object detection component.

No yield-related annotations (e.g., pod weight, bean count, or fullness score) are provided or created. As a result, the dataset supports the study's objective of developing a yield estimation framework that does not rely on ground-truth yield labels.

E. Preprocessing and Feature Extraction

YOLOv8 Detection: All images are resized to a uniform resolution (e.g., 640×640) and normalized prior to training. Data augmentation techniques, including random horizontal flips and random cropping, are applied during training to improve robustness. The trained YOLOv8 model outputs bounding boxes and confidence scores for detected pods. Early stopping is applied based on validation mAP to prevent overfitting.

Morphological Feature Computation: For each detected pod, geometric features are computed directly from the bounding box, including width, height, area, and aspect ratio. These features are normalized to ensure comparability across images. Bounding-box-based morphology is selected over segmentation-based descriptors to reduce annotation requirements and computational complexity, aligning with the study's emphasis on efficiency and scalability.

SimCLR Embedding Extraction: Each detected pod is cropped and resized (e.g., 224×224) before being processed by a SimCLR model with a ResNet-50 encoder. SimCLR is chosen for its simplicity, strong empirical performance, and ability to learn meaningful representations without labels or momentum encoders. During training, two augmented views of each pod image are generated using random crops, flips, color jitter, grayscale conversion, and Gaussian blur. The model is trained using the NT-Xent contrastive loss. After training, the projection head is discarded, and the encoder produces a 128-dimensional embedding for each pod.

Feature Fusion: Morphological features and SimCLR embeddings are concatenated into a single fused feature vector. Z-score normalization is applied to balance the contribution of low-dimensional geometric features and high-dimensional visual embeddings.

F. Model Selection and Training

YOLOv8 Training: Cacao pod detection is performed using the YOLOv8 model, initialized with COCO-pretrained weights for improved convergence and generalization. The model is trained using a learning rate of approximately 0.001 and a batch size of 16–32 for up to 100 epochs, with early stopping based on validation loss to prevent overfitting. Training optimizes bounding-box regression and objectness classification losses. Detection performance is evaluated using precision, recall, and mean Average Precision (mAP) on a validation set to ensure reliable pod localization.

SimCLR Training: Detected pods are cropped and used to train a self-supervised SimCLR model with a ResNet-50 encoder and a 128-dimensional projection head. The model is trained without labels using a large batch size (e.g., 128) and standard SimCLR augmentations, including random cropping, flipping, color jittering, and blurring. Training runs for up to 100 epochs until the contrastive loss converges. After training, the projection head is removed, and the encoder generates fixed-length visual embeddings for downstream analysis.

Ranking Procedure: To maintain a ground-truth-free framework, no supervised ranking model is used. Instead, a heuristic scoring approach computes the L2 norm of each pod's fused morphological and visual feature vector as a relative yield score. Pods are ranked in descending order of this score, producing a relative yield ordering without reliance on yield labels.

G. Evaluation Metrics

Object Detection Performance: Detection reliability is evaluated using precision, recall, and mAP at an Intersection-over-Union (IoU) threshold of 0.5. A high-performing detector is a prerequisite for downstream feature extraction and ranking.

Internal Representation Quality: The quality of SimCLR embeddings is assessed using unsupervised analyses, including dimensionality reduction (t-SNE) and clustering metrics such as silhouette score and intra-cluster versus inter-cluster similarity. These analyses evaluate whether the learned embeddings exhibit coherent structure without relying on yield labels.

Relative Ranking Evaluation: Model-generated rankings are compared against an expert-informed reference ranking using rank correlation metrics (Kendall's Tau and Spearman's rho) and pairwise agreement analysis. The expert ranking serves as an external plausibility check rather than a training signal or absolute ground truth. The system's validity is primarily supported by internal consistency, ablation performance, and superiority over heuristic and random baselines.

Statistical Analysis: Differences between system variants (morphological-only, embedding-only, and fused features) are analyzed using paired statistical tests. Mean performance values and standard deviations are reported to ensure that observed differences are not attributable to random variation.

H. Comparative Analysis

Feature Ablation: We evaluate three system variants: (1) ranking by morphological features only, (2) by SimCLR embeddings only, and (3) by the fused features. Comparing their ranking accuracies on the test set reveals the contribution of each feature type. We expect the fused system to outperform the single-modality systems, demonstrating that combining shape and visual cues improves yield estimation.

Heuristic Baselines: As simple baselines, pods are ranked by trivial heuristics (e.g. bounding-box area alone) and also randomly. The model should significantly outperform these baselines (random ≈50% pairwise agreement). Demonstrating this confirms that the learned features provide genuine insight beyond naive methods.

Failure Mode Analysis: We qualitatively inspect cases where the model's ranking disagrees with the expert. For instance, partially occluded pods or unusual lighting conditions may lead to errors. Documenting these failure cases provides insight into the method's limitations. Such qualitative analysis is reported to ensure transparency and to avoid overstating the system's capabilities.

I. Ethical Considerations

The study uses publicly available plant imagery and involves no human subjects, minimizing privacy concerns. Ethical considerations focus on transparency, responsible deployment, and avoidance of misuse. The system is explicitly designed to provide relative yield estimates, not absolute production forecasts, and this limitation is clearly stated. Model generalizability is discussed, and the need for retraining or adaptation when applied to new regions or cacao varieties is acknowledged. Efficient model architectures are used to reduce computational and environmental cost, aligning with sustainable AI principles.

IV. RESULTS AND DISCUSSION

V. CONCLUSION

REFERENCES