Multi-modal Auto-labelling Pipeline for Sugar Beet Crop-Weed Classification

Andrew Zakhary[1], Ali Ehteshami-Bejnordi[1], Viktoria Sorokina[2], Dzmitry Yablonski[2], Stefan Henkler[1]

1Hamm-Lippstadt University of Applied Science, Lippstadt, Germany

2Geopard, Cologne, Germany

A paper from the Proceedings of the
17th International Conference on Precision Agriculture and 11th Brazilian Congress on Precision and Digital Agriculture
13-15 July 2026
Porto Alegre, Brazil

Abstract.

The deployment of robust machine learning models in precision agriculture is frequently bottlenecked by the scarcity of high-quality annotated data. Despite recent developments in ML-based crop–weed detection systems, the availability of large-scale, high-quality labelled datasets remains a major limitation. State-of-the-art models require extensive data that captures diverse growth stages, variable lighting and weather conditions, and a wide range of weed species. Creating such datasets manually from dense drone imagery is labour-intensive and costly, often requiring agricultural experts to distinguish subtle visual differences between crops and weeds. This process is prohibitively expensive and time-consuming, often creating a bottleneck for model development. As a result, adopting precision agriculture technologies can be particularly challenging for small-scale farmers and developers with limited access to expert annotation or public datasets.

To address these challenges, we propose a multi-modal auto-labelling pipeline for sugar beet crop–weed classification that is capable of producing high-quality labelled training data with minimal human labor. The proposed pipeline integrates two complementary architectures: a convolutional object detection model based on YOLO (You Only Look Once) and a multimodal representation learning model CLIP (Contrastive Language-Image Pre-Training), enabling the system to address both object detection and pixel-level segmentation tasks within a unified framework. The pipeline begins by training the YOLO model on a small seed set of manually labelled data, which is then used to generate pseudolabels by detecting and extracting individual plant instances from a much larger corpus of unlabeled drone imagery. This approach enables rapid dataset expansion, but the resulting predictions can contain a relatively high degree of uncertainty due to limited initial supervision.

To mitigate this uncertainty, the CLIP model is incorporated as a semantic verification stage within the pipeline. The extracted visual crops produced by the YOLO detector are fed into the CLIP model, which compares the image features against textual botanical and morphological descriptions of sugar beet crops and weed leaves. By leveraging both visual and semantic information, the CLIP model helps validate and refine pseudolabels without requiring extensive manual correction. We evaluate the proposed pipeline against a standard fully supervised YOLO model trained on a large-scale public dataset and also compare its performance to expert annotations. Experimental results demonstrate that the proposed approach achieves classification accuracy comparable to models trained on large public datasets, while requiring only a fraction of the time and effort associated with manual labelling. This reduction in annotation cost highlights the practicality of the framework and demonstrates its potential as a scalable, cost-effective solution for deploying smart farming systems in data- and resource-constrained environments.

Keywords.
Classification, Smart farming, precision weeding

Introduction

The success of modern machine learning systems is fundamentally dependent on the availability of large quantities of high-quality annotated data. Across computer vision applications, advances in model architectures have consistently outpaced the development of corresponding labelled datasets, creating a growing bottleneck for the deployment of robust and reliable learning systems. This challenge is particularly acute in agricultural imaging, where supervised models require extensive annotations collected across diverse environmental conditions, growth stages, and plant species. While deep learning has demonstrated considerable promise for automated crop monitoring and weed management, the cost and effort required to generate suitable training data remain a major barrier to widespread adoption (Zhao and Wang 2026).

The annotation problem is especially severe in crop–weed classification from unmanned aerial vehicle (UAV) imagery. High-resolution drone surveys capture thousands of individual plants across large agricultural fields, requiring annotations at a level of detail that often demands expert agronomic knowledge (Gomez-Zamanillo et al. 2025). Distinguishing crops from visually similar weed species can be difficult even for experienced annotators, and the dense vegetation present in agricultural imagery substantially increases the time required for manual labelling. As a result, creating large-scale datasets for crop–weed classification is both labour-intensive and expensive, limiting the development of machine learning solutions and restricting access to precision agriculture technologies for researchers and small-scale agricultural operators with limited resources.

Several public datasets have been introduced to support agricultural machine learning research, including benchmarks for sugar beet crop–weed discrimination. However, these datasets remain limited in scale, geographic diversity, and environmental variability, making it difficult to train models capable of generalising across real-world field conditions. At the same time, object detection architectures such as the YOLO (Redmon and Farhadi 2018) family have demonstrated increasingly strong performance in agricultural vision tasks, further highlighting the need for scalable methods capable of producing high-quality labelled data without relying on extensive manual annotation.

To address this challenge, researchers have increasingly explored semi-supervised learning and self-training approaches in which a model trained on a small labelled seed set generates pseudolabels for a much larger pool of unlabeled data (Khan et al. (2021); Nong et al. (2022); Kong et al. (2025)). By leveraging large quantities of unannotated imagery, these approaches offer a practical pathway toward reducing annotation costs. However, pseudolabel quality remains a critical limitation. Models trained on limited supervision frequently generate uncertain or incorrect predictions, and the accumulation of label noise can reduce rather than improve downstream model performance. Consequently, effective mechanisms for validating and refining automatically generated labels are essential if pseudolabelling is to become a reliable alternative to manual annotation.

Recent advances in multi-modal representation learning provide a promising solution to this verification problem. Models such as CLIP (Radford et al. 2021) learn joint representations of visual and textual information, enabling semantic reasoning about image content through natural language descriptions. Unlike conventional image classifiers, these models can evaluate visual observations against descriptive textual concepts, offering a potential means of assessing the reliability of automatically generated labels. Despite their success in a range of computer

vision tasks, the use of multi-modal models as semantic verification mechanisms within agricultural annotation pipelines remains largely unexplored.

In this work, we propose a multi-modal auto-labelling framework for sugar beet crop–weed classification that combines YOLO-based object detection with CLIP-based semantic verification to generate high-quality labelled training data while minimizing human annotation effort. By integrating object detection with semantic validation, the proposed framework seeks to reduce annotation cost while maintaining label quality. Experimental evaluation demonstrates that the resulting models achieve performance comparable to fully supervised approaches trained on large public datasets, while requiring significantly less manual labelling effort. This work highlights the potential of multi-modal auto-labelling as a scalable and cost-effective strategy for agricultural machine learning in data-constrained environments.

Related Work

Crop–Weed Detection from UAV Imagery

The use of UAVs for large-scale crop monitoring has produced a growing body of research on deep learning-based crop–weed detection. Early approaches relied on fully supervised convolutional architectures trained on dense, manually annotated imagery, with benchmark datasets such as the Sugar Beets dataset and the Lincoln Beet dataset providing standardised evaluation platforms (Salazar-Gomez et al. 2021). While these benchmarks enabled systematic comparison of detection methods, their limited scale and geographic scope constrain the ability of trained models to generalise across diverse field conditions. More recent work has applied transformer-based architectures and multi-scale feature extraction strategies to improve detection performance, but the underlying requirement for large quantities of labelled training data remains a fundamental bottleneck for practical deployment (Zhao and Wang 2026).

Several studies have investigated object detection models in UAV-based weed identification contexts. (Gallo et al. 2023) applied YOLOv7 to high-resolution drone imagery of chicory fields and evaluated performance against the Lincoln Beet benchmark, demonstrating competitive detection performance when domain-specific training data are available. Comparative evaluations of YOLO variants for weed detection tasks have confirmed that model performance is strongly sensitive to the availability and diversity of labelled training data, underscoring the importance of scalable annotation strategies (Sharma and Vardhan 2024).

Semi-Supervised and Self-Training Approaches in Agricultural Vision

To reduce dependence on large manually labelled datasets, researchers have increasingly explored semi-supervised and self-training strategies for agricultural image analysis. Khan et al. (2021) proposed a semi-supervised generative adversarial network (GAN) for crop–weed classification at early growth stages using UAV RGB imagery, demonstrating that competitive performance can be achieved with a small initial set of labelled examples by leveraging a generator to produce synthetic training data for the discriminator. Nong et al. (2022) presented SemiWeedNet, a semi-supervised segmentation framework that incorporates consistency regularisation and online hard example mining to exploit unlabelled data, achieving strong performance on publicly available crop–weed benchmarks while reducing annotation requirements. More recently, Kong et al. (2025) investigated semi-supervised learning for weed detection specifically in wheat, finding that self-training with confidence-based pseudo-label

filtering provided meaningful performance gains beyond the supervised baseline when labelled data were scarce.

A parallel line of work has applied unsupervised pseudo-labelling strategies that exploit structural cues present in agricultural imagery. RoWeeder (De Marinis et al. 2024) generates a pseudo-ground truth for weed segmentation by detecting crop rows and treating off-row vegetation as weeds, training a noise-tolerant deep learning model without any manual labels. While this approach eliminates the annotation burden entirely, it is limited to scenarios where reliable crop-row geometry can be detected and does not generalise to multi-class detection settings. In contrast, the pipeline proposed in the present work begins with a small manually labelled seed set and progressively expands the training corpus through iterative pseudolabelling, targeting a more general multi-class detection objective.

Pseudolabel Quality and Noise Mitigation

A central challenge in self-training pipelines is the accumulation of label noise introduced by incorrect pseudolabels generated during early training cycles, when the model has limited supervision. Confirmation bias—the tendency of a model to reinforce its own errors through repeated self-training—is a well-documented failure mode that can cause performance to stagnate or degrade when pseudolabel quality is low (Arazo et al. 2020). Various strategies have been proposed to mitigate this problem, including confidence-based filtering, consistency regularisation, and ensemble-based selection. At the image-to-label level, Wu et al. (2022) demonstrated that jointly considering instance-level uncertainty and dataset diversity in sample selection substantially improves the quality of the labelled subset used for retraining in object detection settings.

More recent work has explored the use of pre-trained vision–language models as external semantic validators for noisy label correction. Wei et al. (2024) introduced DeFT, a denoising fine-tuning framework that leverages the aligned textual and visual representations of CLIP-based models to identify and filter incorrectly labelled samples during fine-tuning. By learning class-specific positive and negative textual prompts, DeFT exploits the semantic knowledge encoded in pre-trained vision–language representations to discriminate clean from noisy labels without requiring additional manually labelled data. This work provides direct conceptual motivation for the semantic verification stage in our proposed pipeline, which similarly employs CLIP similarity scoring as an external quality filter on YOLO-generated pseudolabels.

CLIP and Vision–Language Models in Agricultural Contexts

CLIP learns a joint embedding space for images and natural language descriptions by training on large-scale paired image–text data, enabling zero-shot and few-shot transfer to downstream classification tasks. Its ability to evaluate visual content against textual descriptions without task-specific training makes it an attractive tool for domains where labelled data are scarce. Applications of CLIP in agricultural and plant science contexts have grown substantially. Campos-Mocholí et al. (2025) studied zero-shot and fine-tuned CLIP for plant disease classification, finding that CLIP's generalisation capabilities allowed it to identify disease symptoms in crop images with minimal task-specific supervision, particularly in settings where annotated examples of rare disease classes were unavailable.

Nawaz et al. (2024) introduced AgriCLIP, a domain-specialised vision–language model pre-trained on 600,000 agricultural image–text pairs, demonstrating that adapting CLIP to the agricultural domain significantly improves zero-shot classification accuracy across a range of crop and livestock recognition benchmarks. The strong performance of both general-purpose

CLIP and domain-adapted variants in agricultural vision tasks supports the hypothesis that image–text similarity provides a reliable semantic signal for class discrimination, even when specialised training data are limited. Despite this progress, the use of CLIP as an inline verification mechanism within an iterative detection pipeline—as proposed in the present work—has not been previously explored in the precision agriculture literature.

Annotation Efficiency and Active Learning

Reducing human annotation effort is a well-studied problem in the computer vision literature. Active learning approaches address this by selecting the most informative unlabelled samples for annotation, with the goal of achieving maximum model improvement per labelled example. In the object detection setting, bounding-box annotation is substantially more expensive than image-level labelling, making annotation efficiency particularly critical (Kao et al. 2019). Entropy-based and uncertainty-aware sample selection strategies have been shown to reduce the number of labelled examples required to achieve a target detection performance (Wu et al. 2022). The sequential retraining strategy employed in this work is related to active learning in that it iteratively expands the labelled pool through model-assisted annotation, shifting the human annotator from the role of creating labels from scratch to verifying and correcting model-generated predictions.

Taken together, the reviewed literature establishes the context and motivation for the proposed multi-modal auto-labelling pipeline. While semi-supervised detection, pseudolabelling, and vision–language models have each been studied independently in related domains, the integration of CLIP-based semantic verification into an iterative YOLO retraining pipeline for agricultural crop–weed classification represents a novel contribution. The following section describes the methodology in full detail.

Methodology

Dataset Acquisition and Benchmarking Baseline

The dataset utilized in this study was captured across commercial sugar beet fields using UAVs. High-resolution aerial video footage was systematically partitioned into discrete image frames. To optimize dataset variance and mitigate spatial redundancy, frame extraction was executed with minimal geometric overlap, preventing repetitive sampling of identical plant instances across successive frames. The dataset produced 750 images with approximately 50 instances per image.

The proposed pipeline follows an incremental sequential retraining strategy which is shown in Fig. 1. Initially, 50 manually labelled images are used to train a seed object detection model. Once trained, this model is applied to a previously unseen batch of images. Rather than

immediately incorporating these images into the training process, the model's predictions are evaluated and reviewed.

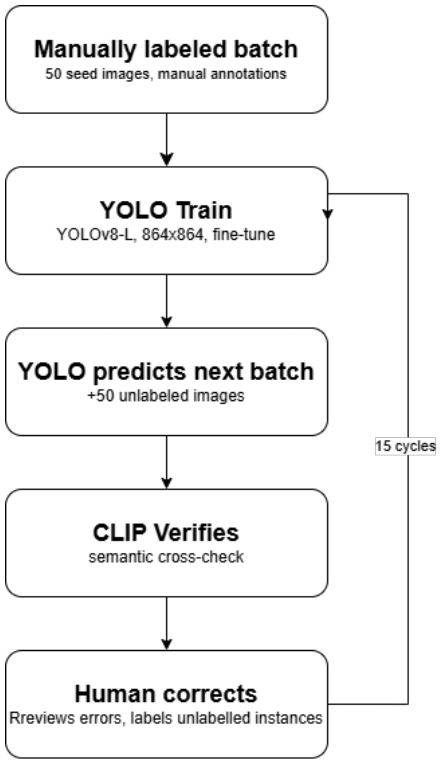

Fig. 1 Pipeline Flow

After review, the newly labelled images are added to the existing training dataset and the model is retrained using all available labelled data. This process is repeated iteratively, allowing the model to gradually improve as additional examples become available.

The primary objective of this strategy is to shift the role of the human annotator from creating annotations entirely manually to verifying and correcting model-generated predictions. As the model improves, the number of corrections required per image decreases, leading to substantial reductions in annotation time.

The pipeline was executed over 15 learning cycles. Each cycle introduced an additional batch of 50 images, resulting in a progressively larger training set ranging from 50 to 750 images.

YOLO Object Detection Model

Object detection within the pipeline is performed using the YOLO architecture. Unlike traditional two-stage detectors, which first generate candidate regions and then classify them, YOLO performs localisation and classification simultaneously in a single forward pass through the

network. This design provides an advantageous balance between detection accuracy and computational efficiency.

A pre-trained YOLOv8-Large model was selected as the detector backbone, as it has seen widespread adoption in recent agricultural studies including UAV-based weed detection (Pukrop et al., 2024), weed species differentiation in soybean crops (Sulzbach et al., 2025), and crop–weed classification in wheat fields (Liu et al., 2024). The Large variant was preferred over smaller variants for its greater representational capacity, enabling more discriminative feature extraction for the fine-grained visual differences between sugar beet and weed species in high-resolution UAV imagery. During the first cycle, the model was trained on the initial set of 50 manually labelled images. For subsequent cycles, the model was fine-tuned using all previously reviewed and accepted annotations.

Training was performed using an image resolution of 864 × 864 pixels and a batch size of four images. A lower learning rate was selected to preserve previously learned features while adapting the model to the specific appearance of sugar beet crops and weeds. Early stopping was used to prevent overfitting when validation performance ceased improving.

CLIP-Based Verification

Although YOLO is capable of generating object detections, predictions produced during the early learning cycles may contain substantial classification errors due to the limited amount of training data available. To reduce the impact of these errors, a secondary verification stage based on CLIP was incorporated into the pipeline. Fig. 2 shows the flow for the labels through the whole pipeline.

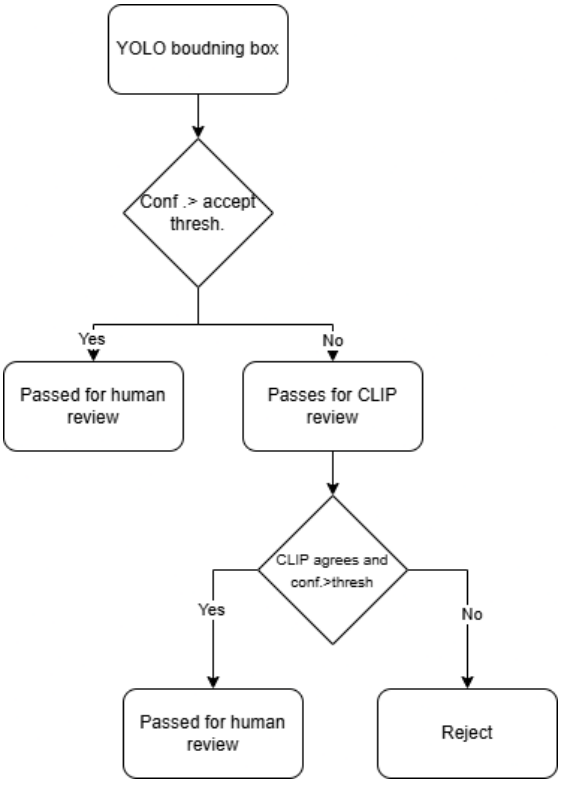

Fig. 2 Label decision flow

In this work, CLIP was used as a verification mechanism rather than a primary classifier. For every YOLO detection with a confidence score below a predefined threshold, the detected

image region was cropped and analysed by CLIP. The crop was compared against two botanical descriptions:

• “A sugar beet plant with ovate leaves and smooth margins.”

• “A grassy weed with jagged leaves and upright growth.”

CLIP computes similarity scores between the cropped image and both textual descriptions. The class with the highest similarity score is selected as the CLIP prediction. A consensus-based filtering strategy was then applied. If the class predicted by CLIP matched the class predicted by YOLO, the detection was retained. If the predictions disagreed, the detection was discarded. High-confidence YOLO detections were accepted directly without CLIP verification.

This mechanism acts as a semantic guardrail during the early stages of learning by removing uncertain detections that are likely to correspond to incorrectly classified plants or background regions. As a result, the quality of the generated annotations is improved before human review, reducing the number of false positives introduced into the training process.

Fully Supervised Model

In order to compare this approach with traditional methods, a standalone YOLO model was trained strictly on the public Lincoln Beet dataset (Salazar-Gomez 2021). In order to improve the model’s ability to generalize on our dataset, the model is finetuned on the same training/validation set for the 15th cycle of the pipeline consisting of 750 images. This transfer-learned model serves as a direct point of comparison to measure the accuracy, sample efficiency, and labor savings of our proposed multi-modal pipeline against traditional, large-scale fully supervised methods.

Evaluation Procedure

The performance and structural integrity of the generated annotations are systematically evaluated at the conclusion of each sequential learning cycle using standard statistical computer vision matrices. Spatial validation of predicted bounding boxes relative to ground-truth coordinates is governed by an Intersection over Union (IoU) threshold set at a standard $\boldsymbol{mAP_{50}}$, alongside an average across a tighter gradient of thresholds $\boldsymbol{mAP_{50-95}}$ (Redmon and Farhadi 2018) to evaluate localization density.

To map classification and localization accuracy for both sugar beet crops and weed variants, the system tracks individual and aggregate profiles for Precision $\boldsymbol{P}$, Recall $\boldsymbol{R}$, and the balanced F1-score $\boldsymbol{F1}$:

$$Precision\ (P) = \frac{TP}{TP + FP}, Recall\ (R) = \frac{TP}{TP + FN}, F1 = 2 \times \frac{P \times R}{P + R}$$

Where TP, FP, and FN represent True Positives, False Positives, and False Negatives, respectively. Additionally, an overall System Error Rate (ϵ) is monitored dynamically as a compound function of aggregate dataset noise introduced prior to human screening, calculated against total ground-truth instances (N_{gt})

$$\epsilon = \frac{FP + FN}{N_{gt}}$$

Multi-Modal Verification Assessment

To explicitly isolate and quantify the exact semantic contribution of the CLIP verification layer,

system metrics are captured via a synchronous comparative approach at every cycle:

- **Pre-Filter Baseline:** Performance metrics are logged immediately following raw YOLOv8 inference.
- **Post-Filter Evaluation:** Metrics are logged immediately following the CLIP consensus filtering protocol.

The mathematical delta (Δ) between these two synchronous measurements isolates the exact impact of multi-modal semantic verification on mitigating label uncertainty throughout the cumulative retraining sequence.

Finally, these curves are structured to be plotted directly against the fully supervised model which is transfer-learned onto our dataset —to compare domain-specific field readiness, sample efficiency, and human labor reduction.

Results

Sequential Training Performance

Fig. 3 Pipeline performance across 15 sequential training cycles (50–750 images). (a) Learning curves with transfer-learned fully supervised baseline. (b) Precision and recall trajectories with baseline precision and recall shown as dashed reference lines. (c) Human annotation effort measured as the number of required corrections per cycle (d) Class-level F1 scores per class with transfer-learned fully supervised baseline. (e) Per-cycle benefit of CLIP verification,

expressed as the absolute gain in precision and F1 score and the absolute reduction in error rate relative to YOLO-only predictions. (f) CLIP verification workload showing the number of low-confidence detections reviewed and rejected per cycle.

YOLO achieved a mean Average Precision (mAP@50) of 0.913 after training on 50 images, increasing to a maximum of 0.982 after 650 training images (Fig. 3(a)). Similarly, mAP@50–95 improved from 0.661 to 0.914 over the same period. Performance gains became progressively smaller beyond approximately 500 training images, suggesting that the detector had already learned most discriminative visual features required for reliable crop–weed detection.

Overall detection performance exhibited a similar trend. The baseline YOLO F1 score increased from 0.794 at 50 training images to a maximum of 0.866 at 700 training images, while the error rate decreased from 0.342 to 0.236. Precision improved steadily from 0.690 to 0.799, whereas recall remained consistently high throughout the experiment, typically exceeding 0.90 across all training cycles, as shown in Fig. 3(b). These results indicate that the primary benefit of additional training data was a reduction in false positive detections rather than a substantial increase in object discovery.

Class-Level Performance Analysis

Class-specific metrics revealed distinct learning behaviour for sugar beet and weed detection, as illustrated in Fig. 3(d). Beet detection performance was already strong during the initial training cycle, achieving an F1 score of 0.930 with only 50 training images. Across all training cycles, beet F1 remained relatively stable between approximately 0.90 and 0.94, demonstrating that sugar beet instances were readily distinguishable even when training data were limited.

In contrast, weed detection benefited substantially from increased training data. Weed F1 improved from 0.565 at 50 training images to a peak of 0.746 at 700 training images. This improvement was primarily driven by increased precision, which rose from 0.413 to 0.638 over the same period. Weed recall remained comparatively high throughout the experiment, generally ranging between 0.81 and 0.90. These findings suggest that the principal challenge was not locating weed instances but accurately distinguishing weeds from visually similar crop structures and background vegetation.

The differing learning trajectories indicate that sugar beet plants formed a relatively homogeneous visual class, whereas weed instances exhibited substantially greater intra-class variability. Consequently, additional labelled examples provided only marginal benefits for beet detection but continued to improve weed classification performance throughout most of the training process.

Effect of CLIP-Based Verification

The CLIP verification stage consistently improved label quality and overall detection performance across all training cycles. By validating YOLO detections using image–text similarity, CLIP reduced false positive predictions while preserving the majority of true positive detections.

During the initial training cycle, CLIP verification increased precision from 0.690 to 0.775 and improved overall F1 score from 0.794 to 0.826. The error rate decreased from 0.342 to 0.297, representing a relative reduction of approximately 13%.

These improvements remained consistent throughout the experiment, as summarised in Fig. 3(e). Across all fifteen training cycles, CLIP increased precision by approximately 0.03–0.10 points while improving F1 score by 0.01–0.05 points. Error rates were consistently reduced by

approximately 0.02–0.06 points relative to the baseline YOLO detector. Although CLIP verification typically resulted in a modest reduction in recall, the corresponding gains in precision produced a net improvement in overall F1 performance.

The largest benefits were observed during the early stages of training, when the detector was trained on relatively small datasets and generated a larger number of uncertain predictions. As training progressed and the baseline detector became increasingly reliable, the magnitude of improvement gradually diminished. Precision gains decreased from +0.085 in Cycle 1 to +0.038 in Cycle 15, while F1 improvements declined from +0.032 to +0.017 over the same period. This behaviour suggests that CLIP functions primarily as a corrective mechanism, providing the greatest benefit when detector uncertainty is highest.

CLIP–YOLO Agreement and Annotation Efficiency

The number of detections requiring CLIP review decreased substantially throughout the sequential training process, as shown in Fig. 3(f). The total number of reviewed detections fell from 1,169 in Cycle 1 to 479 in Cycle 15, indicating increasing detector confidence and stability as additional training data were incorporated.

Among reviewed detections, agreement between YOLO and CLIP remained consistently high. CLIP accepted between approximately 64% and 75% of reviewed detections across the training cycles, demonstrating substantial alignment between the visual representations learned by YOLO and the semantic representations encoded by CLIP.

More importantly, CLIP verification consistently reduced the amount of manual correction required, as shown in Fig. 3(c). Human correction counts decreased from 940 to 722 in Cycle 1 and from 662 to 571 in Cycle 15, corresponding to reductions of approximately 14–25% across the experimental runs. This reduction in annotation effort directly supports the primary objective of the proposed pipeline: generating high-quality labelled datasets while minimising human intervention.

Comparison with Fully Supervised Baseline

The baseline model achieved an overall F1 score of 0.797, a precision of 0.692, and a recall of 0.940, with a mAP@50 of 0.880 and an error rate of 0.337. At the class level, beet detection was strong, yielding an F1 score of 0.917 (precision 0.856, recall 0.987), while weed detection remained considerably weaker, with an F1 score of 0.644 (precision 0.495, recall 0.921). These results reflect the well-known tendency of models trained on public benchmarks to achieve high recall at the cost of elevated false positive rates when deployed on domain-specific imagery.

The proposed pipeline substantially outperformed the baseline across all primary metrics. At its peak (Cycle 14), the pipeline achieved an overall F1 score of 0.866 with YOLO alone and 0.880 with CLIP verification, compared to 0.797 for the baseline — an improvement of 0.069 and 0.083 points respectively (Fig. 3(a)). The pipeline error rate at Cycle 14 was 0.236 (0.214 post-CLIP), representing a relative reduction of approximately 30% compared to the baseline error rate of 0.337. Weed detection showed the largest gains: the pipeline weed F1 reached 0.746 (YOLO) and 0.769 (CLIP) at Cycle 14, compared to 0.644 for the baseline, corresponding to improvements of 0.102 and 0.125 points (Fig. 3(d)). Beet detection performance was broadly comparable between the two approaches, with the pipeline achieving a beet F1 of 0.922 against the baseline's 0.917.

These results demonstrate that the proposed multi-modal auto-labelling pipeline not only matches but exceeds the performance of a model trained on a large public dataset, while requiring only a small manually annotated seed set. The higher recall of the baseline model

reflects its training on a broader distribution of sugar beet imagery, yet this did not translate into superior overall performance due to substantially elevated false positive rates. The pipeline's higher precision and lower error rate indicate that iterative self-training with CLIP-based semantic verification produces more discriminative and reliable detections on the target domain, even with significantly fewer labelled examples.

Discussion

Performance in the Context of Expertly Annotated Benchmarks

The results achieved by the proposed pipeline can be contextualised against published detection performance on expertly annotated agricultural benchmarks. Gallo et al. (2023) reported competitive detection performance using YOLOv7 on the Lincoln Beet dataset under domain-specific training conditions, while Sharma and Vardhan (2024) demonstrated that YOLO-family models exhibit strong sensitivity to the availability and diversity of labelled training data. The pipeline proposed in this work achieves an overall F1 score of 0.880 and a mAP@50 of 0.982 at Cycle 14, results which are broadly comparable to or exceed those reported in equivalent fully supervised settings on similar crop–weed benchmarks. Crucially, these results are obtained from a seed set of only 50 manually labelled images, progressively expanded to 750 through iterative self-training—a fraction of the annotation effort required to construct the public benchmarks against which these models are typically evaluated.

This finding is further supported by the broader data-centric perspective on annotation quality in object detection. Hussain et al. (2026) systematically document that even widely used and carefully constructed benchmarks—including MS COCO, PASCAL VOC, and Open Images—contain systematic annotation errors such as missing labels, inaccurate bounding boxes, and misclassified instances. Their survey highlights that annotation noise is an inherent and under-addressed property of large-scale datasets, rather than a problem unique to automatically generated labels. In this context, the error rate of 0.214 achieved by the proposed pipeline after CLIP verification is noteworthy: it suggests that iterative pseudolabelling with semantic filtering can produce annotation quality competitive with, and in some cases exceeding, the reliability of manually curated datasets at scale.

Recall Trade-off and the Role of CLIP as a Conservative Verifier

A consistent observation across all training cycles is that the CLIP verification stage produces a modest but systematic reduction in recall relative to YOLO-only predictions. This is an inherent consequence of the consensus-based filtering strategy: CLIP, acting as a conservative secondary verifier, will always reject a proportion of detections where its semantic similarity score disagrees with the YOLO classification, including instances where YOLO was in fact correct. Detections that fall below the YOLO confidence threshold and are subsequently flagged for CLIP review represent the most uncertain predictions in the pipeline; by discarding disagreements rather than resolving them, the filter necessarily removes some true positives alongside the false positives it targets.

This trade-off is a deliberate design choice rather than a limitation. In the context of pseudolabel generation for iterative self-training, label cleanliness is more important than label completeness. Introducing a false positive into the training set causes the model to reinforce an incorrect prediction across subsequent cycles—a well-documented failure mode known as

confirmation bias (Arazo et al. 2020). A missed true positive, by contrast, simply results in one fewer training example, which is recoverable as the dataset expands.

This effect is most pronounced in the early training cycles, when the YOLO detector is least reliable and the proportion of uncertain low-confidence detections is highest. As training progresses and detector confidence improves, fewer detections are routed to CLIP for review, and the recall difference between YOLO-only and YOLO+CLIP narrows accordingly. The converging recall curves observed from Cycle 9 onwards suggest that the pipeline naturally transitions toward a regime in which CLIP verification provides diminishing returns in recall reduction, while continuing to offer meaningful precision gains. This behaviour is consistent with the interpretation of CLIP as a corrective rather than a generative mechanism: its value is greatest precisely when the primary detector is most uncertain, and its influence appropriately diminishes as model reliability increases.

Limitations

The present study has several limitations that should be considered when interpreting these results. The pipeline was evaluated exclusively on sugar beet imagery captured from a single geographic region under a limited range of environmental conditions. Generalisation to other crop types, weed species, or UAV imaging configurations has not been assessed, and the botanical CLIP prompts used for verification were designed specifically for the sugar beet–grassy weed dichotomy. The YOLO confidence threshold used to route detections to CLIP review was fixed across all cycles rather than adapted dynamically, which may not optimally balance the precision–recall trade-off at each stage of training. Finally, while the pipeline reduces human annotation effort, it does not eliminate it entirely: human review of model-generated labels remains necessary at each cycle to correct residual errors that both YOLO and CLIP fail to identify.

Acknowledgments

This paper is the result of preliminary work by Hamm-Lippstadt University of Applied Sciences, Germany, on the project AKI4KMU - Automated AI Framework for SMEs, supported by the Ministry of Economic Affairs, Industry, Climate Action and Energy of the State of North Rhine-Westphalia. Co-funded by the European Union. Grant number EFRE-20800498

References

Arazo, E., Ortego, D., Albert, P., O'Connor, N. E., & McGuinness, K. (2020). Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *2020 International Joint Conference on Neural Networks (IJCNN)* (pp. 1–8). IEEE. doi:10.1109/IJCNN48605.2020.9207304

Campos-Mocholí, M., Chacón-Albero, O., & Julian, V. (2025). CLIP in the field: A study of plant disease detection via zero-shot and fine-tuning. In *Lecture Notes in Computer Science*. Cham, Switzerland: Springer. doi:10.1007/978-3-032-05925-3_2

De Marinis, P., Vessio, G., & Castellano, G. (2024). RoWeeder: Unsupervised weed mapping through crop-row detection. In *European Conference on Computer Vision* (pp. 132–145). Cham, Switzerland: Springer Nature Switzerland.

Gallo, I., Rehman, A. U., Dehkordi, R. H., Landro, N., La Grassa, R., & Boschetti, M. (2023). Deep object detection of crop weeds: Performance of YOLOv7 on a real case dataset from UAV images. *Remote Sensing*, *15*(2), 539. doi:10.3390/rs15020539

Gómez-Zamanillo, L., Portilla, N., Picón, A., Egusquiza, I., Navarra-Mestre, R., Elola, A., &

Bereciartua-Perez, A. (2025). Reducing annotation effort in agricultural data: simple and fast unsupervised coreset selection with DINOv2 and K-means. *Frontiers in Plant Science*, *16*, 1546756. doi:10.3389/fpls.2025.1546756

Hussain, A., Ullah, K., Afaq, M. *et al.* Quality over quantity: a data-centric survey of annotation errors in object detection datasets. *Artif Intell Rev* 59, 107 (2026). https://doi.org/10.1007/s10462-026-11502-z

Kao, C.-C., Lee, T.-Y., Sen, P., & Liu, M.-Y. (2018). Localization-aware active learning for object detection. In *Asian Conference on Computer Vision* (pp. 506–522). Cham, Switzerland: Springer International Publishing.

Khan, S., Tufail, M., Khan, M. T., Khan, Z. A., Iqbal, J., & Alam, M. (2021). A novel semi-supervised framework for UAV based crop/weed classification. *PLOS ONE*, *16*(5), e0251008. doi:10.1371/journal.pone.0251008

Kong, X., Liu, T., Chen, X., Lian, P., Zhai, D., Li, A., & Yu, J. (2024). Exploring the semi-supervised learning for weed detection in wheat. *Crop Protection*, *184*, 106823.

Liu, Y., Zeng, F., Diao, H., Zhu, J., Ji, D., Liao, X. and Zhao, Z., 2024. YOLOv8 model for weed detection in wheat fields based on a visual converter and multi-scale feature fusion. *Sensors*, *24*(13), p.4379.

Nawaz, U., Muhammad, A., Gani, H., Naseer, M., Khan, F. S., Khan, S., & Anwer, R. (2025). AgriCLIP: Adapting CLIP for agriculture and livestock via domain-specialized cross-model alignment. In *Proceedings of the 31st International Conference on Computational Linguistics* (pp. 9630–9639).

Nong, C., Fan, X., & Wang, J. (2022). Semi-supervised learning for weed and crop segmentation using UAV imagery. *Frontiers in Plant Science*, *13*, 927368. doi:10.3389/fpls.2022.927368

Maren Pukrop, S.P., 2024. Weed detection with YOLOv8-seg in UAV-imagery. In *44. GIL-Jahrestagung, Biodiversität fördern durch digitale Landwirtschaft* (pp. 383-388). Gesellschaft für Informatik eV.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., et al. (2021). Learning transferable visual models from natural language supervision. *arXiv preprint*, arXiv:2103.00020.

Redmon, J., & Farhadi, A. (2018). YOLOv3: An incremental improvement. *arXiv preprint*, arXiv:1804.02767. doi:10.48550/arXiv.1804.02767

Salazar-Gomez, A., Darbyshire, M., Gao, J., Sklar, E. I., & Parsons, S. (2021). Towards practical object detection for weed spraying in precision agriculture. *arXiv preprint*, arXiv:2109.11048.

Sharma, S., & Vardhan, M. (2024). Advancing precision agriculture: Enhanced weed detection using the optimized YOLOv8T model. *Arabian Journal for Science and Engineering*, *50*, 7343–7360.

Sulzbach, E., Scheeren, I., Veras, M.S.T., Tosin, M.C., Kroth, W.A.E., Merotto Jr, A. and Markus, C., 2025. Deep learning model optimization methods and performance evaluation of YOLOv8 for enhanced weed detection in soybeans. Computers and Electronics in Agriculture, 232, p.110117.

Wei, T., Li, H.-T., Li, C.-S., Shi, J.-X., Li, Y.-F., & Zhang, M.-L. (2024). Vision-language models are strong noisy label detectors. *Advances in Neural Information Processing Systems*, *37*, 58154–58173.

Wu, J., Chen, J., & Huang, D. (2022). Entropy-based active learning for object detection with progressive diversity constraint. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 9397–9406).

Zhao, H., & Wang, Y. (2026). Deep learning–based approaches for weed detection in crops. *Frontiers in Plant Science*, *16*, 1746406. doi:10.3389/fpls.2025.1746406