scientific reports

Check for updates

OPEN

WeedSwin hierarchical vision transformer with SAM-2 for multi-stage weed detection and classification

Taminul Islam[1,2]✉, Toqi Tahamid Sarker[1,2], Khaled R. Ahmed[1,2], Cristiana Bernardi Rankrape[3] & Karla Gage[3,4]

Weed detection and classification using computer vision and deep learning techniques have emerged as crucial tools for precision agriculture, offering automated solutions for sustainable farming practices. This study presents a comprehensive approach to weed identification across multiple growth stages, addressing the challenges of detecting and classifying diverse weed species throughout their developmental cycles. We introduce two extensive datasets: the Alpha Weed Dataset (AWD) with 203,567 images and the Beta Weed Dataset (BWD) with 120,341 images, collectively documenting 16 prevalent weed species across 11 growth stages. The datasets were preprocessed using both traditional computer vision techniques and the advanced SAM-2 model, ensuring high-quality annotations with segmentation masks and precise bounding boxes. Our research evaluates several state-of-the-art object detection architectures, including DINO Transformer (with ResNet-101 and Swin backbones), Detection Transformer (DETR), EfficientNet B4, YOLO v8, and RetinaNet. Additionally, we propose a novel WeedSwin Transformer architecture specifically designed to address the unique challenges of weed detection, such as complex morphological variations and overlapping vegetation patterns. Through rigorous experimentation, WeedSwin demonstrated superior performance, achieving 0.993 ± 0.004 mAP and 0.985 mAR while maintaining practical processing speeds of 218.27 FPS, outperforming existing architectures across various metrics. The comprehensive evaluation across different growth stages reveals the robustness of our approach, particularly in detecting challenging "driver weeds" that significantly impact agricultural productivity. By providing accurate, automated weed identification capabilities, this research establishes a foundation for more efficient and environmentally sustainable weed management practices. The demonstrated success of the WeedSwin architecture, combined with our extensive temporal datasets, represents a significant advancement in agricultural computer vision, supporting the evolution of precision farming techniques while promoting reduced herbicide usage and improved crop management efficiency.

Keywords Weed detection, Weed classification, Weed growth stage, Sam2, Precision agriculture, Object detection

Agriculture has served as the cornerstone of human civilization throughout history, playing a fundamental role in the sustenance and advancement of our species. In the modern era, successful agricultural practices extend far beyond traditional farming methods, increasingly relying on sophisticated precision agriculture techniques, particularly Site-Specific Farm Management (SSFM)[1]. This advanced approach requires accurate and timely identification, spatial mapping, and quantitative assessment of both crops and weeds within agricultural landscapes[2]. The challenge of weed management has become increasingly critical in modern agriculture, particularly in the diverse agricultural regions of the United States[3]. The country's unique combination of climate variations and fertile soil conditions, while ideal for crop cultivation, simultaneously creates optimal conditions for the proliferation of numerous weed species. These unwanted plants pose a significant threat to crop yields by competing for essential resources, including water, nutrients, and sunlight[4]. The impact of weed

[1]School of Computing, Southern Illinois University, Carbondale, IL 62901, USA. [2]BASE Lab, School of Computing, Southern Illinois University, Carbondale, IL 62901, USA. [3]School of Agricultural Sciences, Southern Illinois University, Carbondale, IL 62901, USA. [4]School of Biological Sciences, Southern Illinois University, Carbondale, IL 62901, USA. ✉email: taminul.islam@siu.edu

invasion extends beyond immediate crop competition, affecting agricultural productivity, economic stability, and ecosystem balance.

Traditional weed control methodologies, primarily dependent on broad-spectrum herbicides or manual removal techniques, have shown increasing limitations[5,6]. These conventional approaches raise environmental concerns due to chemical runoff, face challenges with herbicide-resistant weed populations, and often prove economically inefficient due to high labor costs and resource utilization. Furthermore, these methods can potentially harm beneficial organisms and impact long-term soil health[7]. The emergence of advanced technologies in computer vision and deep learning has opened new avenues for addressing these agricultural challenges. Modern object detection and classification techniques, when applied to weed identification, offer promising solutions for real-time, automated weed management systems[8]. However, significant research gaps persist in this domain, particularly in addressing the temporal dynamics of weed development. Existing studies predominantly rely on limited datasets or images captured at specific growth stages, failing to represent the dynamic nature of weed development throughout their lifecycle[9,10]. This limitation is particularly concerning as understanding and accurately identifying weed growth stages is crucial for several reasons. First, the effectiveness of herbicide applications varies significantly depending on the weed's growth stage, with early growth stages typically being more susceptible to control measures[11]. Second, different growth stages present distinct morphological features that affect detection accuracy, making it essential for automated systems to adapt to these variations. Third, the competitive impact of weeds on crops varies throughout their growth cycle, with certain stages being more detrimental to crop yield than others. This temporal aspect of weed-crop competition necessitates precise timing of control measures, which can only be achieved through accurate growth stage identification[12].

Beyond the temporal challenges, another significant limitation is that many available datasets focus on a narrow range of weed species, not accurately reflecting the diverse weed populations encountered in real agricultural settings. This lack of species diversity in training data presents a substantial barrier to developing robust, widely applicable weed detection systems. Our research directly addresses these limitations through a comprehensive study focusing on 16 prevalent weed species found in Midwestern cropping systems of the USA, documenting their development from the seedling stage through 11 weeks of growth. The study involved cultivation and systematic labeling of weed specimens in a controlled greenhouse environment, ensuring accurate tracking and documentation throughout the growth cycle. Our key research contributions are:

1. Development of two unique datasets comprising 203,567 images and 120,341 images, capturing comprehensive growth cycles of 16 of the most common and troublesome weed species in Midwestern USA cropping systems.
2. Meticulous labeling of datasets, categorized by species and growth stage (week-wise), providing a comprehensive resource for weed identification research.
3. Implementation of advanced detection architectures including DINO[13] Transformer with ResNet[14] and Swin[15] transformer backbones, Detection Transformer (DETR)[16], EfficientNet B4[17], YOLO v8[18], and RetinaNet[19].
4. Creation of a novel WeedSwin Transformer architecture, optimized for object detection and classification, based on the transformer framework.
5. Comprehensive comparison of model performance, providing evidence-based recommendations for real-world agricultural applications.

The selection of these specific models was driven by their proven capabilities in object detection tasks. The DINO Transformer, implemented with both ResNet[14] and Swin[15] backbones, offers superior accuracy in object detection[13]. DETR introduces an innovative transformer-based approach, particularly effective in handling complex scenes and object relationships[16]. EfficientNet-B4[17], renowned for its compound scaling method that optimally balances network depth, width, and resolution, delivers exceptional feature extraction capabilities while maintaining computational efficiency through its mobile-first architecture design. RetinaNet[19] contributes its efficient focal loss function, specifically addressing class imbalance challenges common in detection tasks. Our novel WeedSwin transformer architecture, built upon the Swin[15] transformer backbone, represents a significant advancement in specialized weed detection capabilities. This comprehensive approach to weed detection and classification represents a significant step forward in agricultural technology. The research not only contributes to the growing field of AI-assisted agriculture but also provides practical, implementable solutions for farmers and agricultural professionals. By developing more accurate and efficient weed detection systems, we pave the way for enhanced precision agriculture techniques that can significantly reduce herbicide use, lower production costs, and minimize environmental impact while improving overall agricultural sustainability.

The remainder of this paper is organized as follows: Section 'Literature review' reviews relevant literature and recent advancements in weed detection using deep learning approaches. Section 'Methods' details our comprehensive data collection methodology and preprocessing pipeline, including the greenhouse setup, image acquisition protocols, and data curation processes, along with presenting our methodological framework, outlining the experimental design and implementation strategies. Section 'Models and algorithms' provides an in-depth description of the architectural components and implementation details of all models, including our novel WeedSwin architecture. Section 'Experimental results' presents a thorough analysis of experimental results, including comparative performance metrics and model evaluations across different growth stages. Section 'Ablation study' systematically evaluates the contribution of different architectural components to the WeedSwin model's performance through controlled experiments with varied configurations. Section 'Discussion' outlines the broader implications of our findings for precision agriculture, addressing both the technical achievements and practical applications for USA farming communities. Finally, Section. 'Conclusions' summarizes evidence-

based recommendations for implementing these detection systems in real-world agricultural settings and outlines promising directions for future research.

Literature review

In the realm of precision agriculture, the accurate detection and classification of weeds are pivotal for optimizing herbicide usage and promoting sustainable farming practices. Recent advancements in deep learning and computer vision have significantly enhanced the capabilities of weed detection systems, transitioning from traditional methods to sophisticated, data-driven approaches. This review explores the evolution of these technologies, organized into key thematic areas.

Architectural innovations and dataset development

Recent advances in deep learning architectures have transformed weed detection capabilities in precision agriculture[20]. Hussain et al.[21] demonstrated EfficientNet's superior performance in detecting common lambsquarters *(Chenopodium album L.)*, achieving 92–97% accuracy and introducing the Phase-Height-Angle (PHA) format, which improved detection accuracy 1.35-fold over conventional depth imaging. Peteinatos et al.[22] further validated CNN effectiveness through comprehensive evaluation of VGG16, ResNet-50, and Xception architectures across twelve plant species, achieving >97% accuracy with ResNet-50 and Xception on 93,130 images. Li and Zhang[23] advanced algorithmic efficiency through DC-YOLO, incorporating Dual Coordinate Attention and Content-Aware ReAssembly of Features to achieve 95.7% mAP@0.5 while maintaining computational efficiency with only 5.223 million parameters. Ishak Pacal[24] introduced a modified MaxViT model with SE blocks and GRN-based MLP for maize leaf disease detection, achieving 99.24% accuracy and outperforming 64+ deep learning models. Ismail and Ishak[25] demonstrated Vision Transformers' effectiveness for grape disease classification, with Swinv2-Base achieving 100% accuracy across multiple datasets. Ishak & Gültekin[26] compared Vision Transformers to CNNs for corn leaf disease detection, with MaxViT models reaching 100% accuracy on the CD&S dataset and 99.83% on PlantVillage. Alongside these architectural innovations, significant progress has been made in dataset development. Genze et al.[27] introduced the Moving Fields Weed Dataset (MFWD), encompassing 94,321 images of 28 weed species with semantic and instance segmentation masks. Olsen et al.[28] addressed real-world variability through DeepWeeds, containing 17,509 images across eight species and locations, achieving 95.7% classification accuracy with ResNet-50. Dyrmann et al.[29] explored weed classification across 22 species, achieving 86.2% accuracy with 10,413 diverse images. Sapkota et al.citesapkota2022use investigated synthetic data generation, demonstrating comparable performance between synthetic and real images in training Mask R-CNN models, though GAN-generated images showed limited effectiveness.

Multi-modal integration, platform adaptation, and computational optimization

Recent segmentation advances include enhancing machine learning crop classification through SAM-based field delineation[31], with the Segment Anything Model demonstrating significant potential for smart farming applications[32]. Multi-modal approaches have emerged as crucial for complex detection scenarios. Xu et al.[33] developed a three-channel architecture processing RGB and depth information, achieving 89.3% detection precision. Lottes et al.[34] incorporated sequential information in fully convolutional networks, attaining >94% crop recall and 91% weed recall using RGB+NIR imagery. Wang et al.[9] advanced environmental adaptation through an encoder-decoder network achieving 88.91% mean intersection over union (MIoU) and 96.12% object-wise accuracy with NIR integration.

Platform integration has expanded application possibilities, with Islam et al.[35] achieving 96% accuracy in UAV-based Random Forest classification. Beeharry and Bassoo[36] demonstrated 99.8% accuracy with AlexNet CNN in distinguishing between soil, soybean *(Glycine max L.),* and weed types using 15,336 segmented images. Farooq et al.[37] achieved 97% accuracy in pixel-wise vegetation detection using hyperspectral data, demonstrating CNN superiority over traditional histogram of oriented gradients methods. Jeon et al.[38] addressed varying illumination through adaptive image processing, achieving 95.1% identification accuracy for crop plants. Computational efficiency remains crucial for practical deployment. Arun et al.[39] developed a reduced U-Net architecture maintaining 95% accuracy while reducing parameters by 27%. Ukaegbu et al.[40] demonstrated feasibility on UAV-mounted Raspberry Pi systems, though battery life and computational power constrained operations.

Weed detection and classification approaches

Recent research in weed detection and classification has demonstrated significant advancements in both methodology and practical application. Nenavath and Chaubey[41] achieved notable success using Region-Based Convolutional Neural Networks for weed detection in sesame *(Sesamum indicum L.)* crops, attaining 96.84% detection accuracy and 97.79% classification accuracy across different weed species. This approach particularly emphasized the importance of species-specific identification for targeted control measures. Hasan et al.[42] addressed a crucial gap in weed detection datasets by developing an instance-level labeled dataset for corn fields, evaluating multiple deep learning models including YOLOv7 and YOLOv8. Their research demonstrated that YOLOv7 achieved the highest mAP of 88.50%, with data augmentation further improving results to 89.93%. Contrasting with the deep learning approach, Moldvai et al.[43] explored traditional feature-based computer vision methods, achieving a 94.56% recall rate using significantly smaller datasets. Their work demonstrated that shape features, distance transformation features, color histograms, and texture features could provide comparable results to deep learning approaches while requiring only a fraction of the training data. This finding is particularly relevant for applications with limited data availability.

Almalky and Ahmed[11] advanced the field by focusing on growth stage classification, utilizing drone-collected imagery and comparing various deep learning models. Their results showed that YOLOv5-small achieved real-

time detection with 0.794 recall, while RetinaNet with ResNet-101-FPN backbone demonstrated high precision (87.457% average precision) in growth stage classification. Teimouri et al.[12] developed a comprehensive approach for growth stage estimation across 18 weed species, achieving 70% accuracy in leaf counting and 96% accuracy within a two-leaf margin of error, demonstrating the feasibility of automated growth stage assessment. In specialized applications, Costello et al.[44] focused on ragweed parthenium *(Parthenium hysterophorus L.)* weed detection, combining RGB and hyperspectral imagery with YOLOv4 CNN implementation. Their method achieved 95% detection accuracy and 86% classification accuracy for flowering stages, while hyperspectral analysis with XGBoost classifier reached 99% accuracy in growth stage classification. Subeesh et al.[45] evaluated various deep learning architectures for weed identification in bell pepper *(Capsicum annuum L* fields, with InceptionV3 demonstrating superior performance (97.7% accuracy) at optimal hyperparameter settings, establishing a foundation for integration with automated herbicide application systems.

Tables 1 and 2 highlights advancements in weed detection using deep learning, focusing on architectural innovations, dataset diversity, and practical challenges. Despite advances in deep learning and computer vision applications for weed identification and classification, several major limitations persist in current research. A particularly significant shortcoming is that present studies tend to focus on a narrow range of weed species and growth stages, failing to reflect the complete spectrum of weed variety experienced in real-world farming contexts. This limitation is evident in recent studies[11,12,22,44], which predominantly relied on single-species datasets. Key constraints in existing research include inadequate dataset size and variety[9,10], imbalanced class distributions[46], and excessive computing requirements[47]. These challenges were particularly apparent in[12], where researchers encountered difficulties in managing overlapping leaves and class imbalance issues. Implementation challenges further complicate the practical application of these technologies. High hardware costs for sophisticated imaging equipment[21] and substantial computational resource requirements[34] create barriers to widespread adoption. Additionally, environmental variability significantly affects system performance in field conditions[9]. These collective limitations underscore the urgent need for more comprehensive datasets and the development of efficient, robust architectures for weed growth stage detection that can overcome these practical constraints.

The critical need for accurate weed growth stage detection and classification stems from several key factors. The effectiveness of herbicide applications is highly dependent on weed growth stages, with early intervention typically yielding better results and requiring lower chemical concentrations. Herbicide labels, which serve as the legal document governing application specifications, also state limits on weed growth stages or heights. Different growth stages present varying levels of competition with crops for resources, making timely identification crucial for optimal yield protection. Moreover, the morphological changes throughout weed development cycles affect detection accuracy, necessitating robust systems capable of adapting to these variations. Additionally, the economic implications of precise growth stage-based interventions are significant, potentially reducing herbicide usage by up to 90% compared to conventional blanket spraying methods. These factors collectively emphasize the urgent need for comprehensive solutions that can accurately detect and classify weed growth stages in real-world agricultural settings.

References	Contribution	Dataset	Model used	Results	Limitations
Hussain et al.[21]	Detecting common lambsquarters (*Chenopodium album* L.) in potato (*Solanum tuberosum* L.) fields; used CNN with PHA image encoding	30,160 images from Atlantic Canada potato field	GoogLeNet, VGG-16, EfficientNet	EfficientNet achieved 92-97% accuracy, outperforming other models	Requires sophisticated imaging equipment and high computational resources
Islam et al.[35]	UAV-based early weed detection in chilli pepper (*Capsicum annuum* L.) farms	UAV imagery from Australian chilli pepper fields	Random Forest, SVM, KNN	RF achieved highest accuracy at 96%; SVM 94%, KNN 63%	Limited by UAV flight conditions and image resolution
Xu et al.[33]	Multi-modal deep learning with RGB-D for weed detection in wheat (*Triticum aestivum* L.) crop	Wheat field images with RGB-D modality	Custom three-channel network for RGB-D	89.3% precision (IoG)	Requires RGB-D cameras and high computational resources
Almalky and Ahmed[11]	Drone-based detection and classification of *Consolida regalis* L. weed growth stages using deep learning models	3731 images of *Consolida regalis* weed at four growth stages	YOLOv5, RetinaNet, Faster R-CNN	YOLOv5-small achieved recall of 0.794; RetinaNet achieved AP of 87.457%	Relied on a single species dataset
Teimouri et al.[12]	Estimating weed growth stages based on leaf counts using CNN	9649 RGB images of 18 weed species across nine growth stages	Inception-v3 CNN architecture	Accuracy of 78% for *Polygonum* spp.; 70% overall accuracy	Challenges in overlapping leaves and inconsistent performance
Costello et al.[44]	Detection and growth stage classification of ragweed parthenium (*Parthenium hysterophorus* L.) using RGB and hyperspectral imagery	665 RGB images and hyperspectral data in controlled environments	YOLOv4 for RGB; XGBoost for hyperspectral	YOLOv4: 95% detection, 86% classification; XGBoost: 99% classification	Limited applicability to field conditions
Lottes et al.[34]	Fully convolutional network using sequential data for crop-weed classification	RGB+NIR images from sugar beet fields	Fully Convolutional Network	Over 94% recall for crops, 91% for weeds	Needs consistent image sequences and high computational resources
Peteinatos et al.[22]	Identified 12 plant species using deep learning	93,130 labeled images under field conditions	VGG16, ResNet-50, Xception	ResNet-50, Xception achieved >97% accuracy; VGG16 82%	Controlled imaging conditions needed
Beeharry and Bassoo[36]	Evaluated ANN and AlexNet for UAV weed detection	15,336 segmented images	ANN, AlexNet	AlexNet achieved 99.8% accuracy; ANN 48.09%	Computational demands for UAV
Jeon et al.[38]	Adaptive algorithm for plant segmentation under variable lighting	666 field images	ANN with image segmentation	95.1% accuracy for crops	Limited scalability across environments

Table 1. Comparative overview of weed detection studies (Part 1): Analysis of contributions, datasets, models, results, and limitations in weed detection research.

References	Contribution	Dataset	Model Used	Results	Limitations
Ukaegbu et al.[40]	UAV-based sprayer with CNN for real-time weed detection	UAV images for weed classification	CNN-based model on Raspberry Pi	High accuracy in real-time detection	Battery and computational limitations
Subeesh et al.[45]	Detecting weeds in polyhouse-grown bell peppers *(Capsicum annuum L.)* using CNN	1,106 images from a polyhouse	AlexNet, GoogLeNet, InceptionV3, Xception	InceptionV3 achieved 97.7% accuracy	Limited applicability to outdoor settings
Dyrmann et al.[29]	Classifying 22 plant species at early growth stages using CNN	10,413 images from multiple sources	Custom CNN	86.2% accuracy	High species similarity in early stages
Wang et al.[9]	Semantic segmentation for weed management with encoder-decoder network	Images of sugar beets *(Beta vulgaris L. subsp. vulgaris var. altissima)*, oilseed rape *(Brassica napus L. subsp. napus)*	Encoder-decoder deep learning model	Highest MIoU of 88.91%, 96.12% accuracy	Dependent on NIR imagery
Farooq et al.[37]	Effect of spectral bands on weed classification with CNNs	Hyperspectral image dataset	CNN, compared with HoG	CNN with hyperspectral data achieved 97% accuracy	High-cost imagery required
Arun et al.[39]	Pixel-wise segmentation of crops/weeds using reduced U-Net	CWFID dataset	Reduced U-Net	95% segmentation accuracy	Challenges in overlapping regions
Olsen et al.[28]	Developed DeepWeeds dataset for weed detection in rangeland environments	17,509 images of 8 weed species from Australian rangelands	Inception-v3, ResNet-50	ResNet-50 achieved 95.7% accuracy with 53.4 ms/image	Inter-class variability challenges
Li and Zhang[23]	Proposed DC-YOLO for crop and weed detection using YOLOv7-tiny	Public datasets and field-collected corn seedling data	DC-YOLO	mAP@0.5 of 95.7%; 5.223M parameters	Limited exploration of diverse weed types
Sapkota et al.[30]	Explored synthetic images for training Mask R-CNN	Real UAV images, synthetic images (real plant- and GAN-generated)	Mask R-CNN	Real plant-based synthetic images: mAPm of 0.60	Synthetic images underperformed

Table 2. Table 1 continued: Comparative overview of weed detection studies (Part 2).

The extensive body of research reviewed above reveals several persistent limitations in current weed detection approaches. While transformer-based architectures like DINO and Swin have demonstrated impressive general object detection capabilities, they fail to address the unique challenges of agricultural environments—specifically the dramatic scale variations between seedling and mature weeds, complex morphological similarities between weed species, and the dense vegetation patterns typical in field conditions. Existing models either excel at processing speed (YOLO variants) or detection accuracy (DINO implementations) but rarely achieve both simultaneously. Additionally, as Tables 1, and 2 illustrate, most current architectures have been designed for general object detection rather than specifically optimized for the complex temporal dynamics of weed growth stages. WeedSwin directly addresses these limitations through its novel progressive attention heads that adapt to weed scale variations, specialized Channel Mapper for preserving morphological details critical for species differentiation, and optimized encoder-decoder architecture specifically designed to capture the complex contextual relationships in agricultural scenes. Unlike prior approaches that apply general-purpose detection frameworks to weed identification, WeedSwin's architecture is fundamentally designed to balance the competing requirements of computational efficiency and detection accuracy across the entire weed growth cycle.

Our research addresses these challenges comprehensively by developing two extensive datasets, comprising 203,567 images and 120,341 images, documenting 16 prevalent weed species in US agriculture. These datasets uniquely capture the entire growth cycle of these species across 11 weeks, providing a high-resolution temporal perspective on weed development. The seeds of weeds that have been used in this study were provided by the 'weed control research' lab at Southern Illinois University Carbondale. The datasets feature precise annotations of both species and growth stages, providing a valuable foundation for advancing weed classification and detection research. A key highlight of our study is the implementation of advanced detection architectures, including DINO[13] Transformer with ResNet[14] and Swin[15] Transformer backbones, Detection Transformer (DETR)[16], EfficientNet B4[17], YOLO v8[18], and RetinaNet[19]. Furthermore, we introduce a novel WeedSwin Transformer architecture, with all models undergoing comprehensive evaluation to enable robust performance comparisons across diverse scenarios. This study distinguishes itself through its creation of large-scale, diverse datasets that reflect real-world agricultural challenges, rigorous model evaluation, and practical recommendations for farmers. By providing evidence-based insights and implementable solutions for precision weed management, our holistic approach effectively bridges the gap between cutting-edge research and practical applications, advancing sustainable agriculture and fostering the adoption of precision technologies in modern farming operations.

Methods

Study area and experimental setup

This research was conducted during spring and summer 2024 at the SIU Horticulture Research Center greenhouse (37°42′35.8″ N, 89°15′45.0″ W). The facility provided optimal conditions for weed seedling cultivation. The greenhouse used 1000W High Pressure Sodium (HPS) grow light to keep the greenhouse warm (30–32 °C). We utilized 32 square pots (10.7 cm × 10.7 cm × 9 cm), with two replicate pots per species, containing Pro-Mix® BX potting soil. Plants were watered as needed and fertilized with all-purpose 20-20-20 nutrient solution administered at three-day intervals.

Data collection

In this research, we monitored and labeled weed growth stages on a weekly basis to capture the temporal dynamics of plant development. Imaging began at week 1, which corresponded to BBCH[48] (a widely recognized standard for phenological development in weeds) stage 11 ("first true leaf unfolded"), and continued weekly through week 11, ending at BBCH stage 60 ("first flower open"). Each plant image was annotated with both species and its corresponding week (e.g., AMATU_week_1, SORVU_week_5), providing a direct mapping between the week of observation and the phenological stage based on the BBCH scale. This week-wise labeling approach ensured consistent temporal resolution across all species and growth cycles. After initial automated labeling, each image was reviewed and corrected as needed using LabelImg software to improve annotation quality. While we referenced the BBCH scale for stage definitions, annotation consistency was primarily ensured through meticulous manual review and correction, rather than a formal multi-annotator protocol or expert consensus process. This methodology provides a transparent and systematic framework for growth stage annotation, facilitating robust temporal analysis and enabling direct comparison with standard phenological references.

To comprehensively document these growth stages, short 4K video clips (resolution: 3840 $\times$ 2160 pixels, aspect ratio: 16:9) were recorded across the 360° angle of the weeds during each imaging session. The videos were captured using an iPhone 15 Pro Max positioned at a height of 1.5 feet above the plants. Subsequently, individual frames were extracted from these clips to serve as the raw data. This method allowed for high-quality image acquisition, ensuring efficient and consistent data collection throughout the study period.

Among the sixteen weed species included in the study, it was observed that the species SORHA did not emerge during the first two weeks. To facilitate detailed analysis, two datasets were developed: the *Alpha Weeds Dataset (AWD)* and the *Beta Weeds Dataset (BWD)*, encompassing a total of 174 classes. Initially, a total of 2,494,476 frames were compiled across both datasets. After conducting a rigorous quality assessment to eliminate substandard images, 203,567 images were retained for AWD, while 120,341 images were selected for BWD. The rationale for creating these two datasets was to evaluate the model's efficiency, accuracy, and performance on datasets of differing sizes. We have utilized AWD in our previous research[52]. The BWD was generated by selecting only the even-numbered images from AWD, resulting in a dataset approximately half the size of AWD. Additionally, in BWD, we corrected all the incorrect labels we found in AWD. The main purpose of creating BWD is to make a better dataset with all corrected labels and to check and compare model performances between a concise and a big dataset. Table 3 provides a comprehensive summary of the two datasets, detailing the weed species codes, scientific and common names, family, and the number of frames captured for each species on a weekly basis.

Figure 1 presents representative images of four weed species at distinct growth stages. For AMAPA, images from week one (a) and week eleven (b) are displayed. Similarly, SIDSP is depicted at week one (c) and week eleven (d). AMATU is shown during its first (e) and eleventh (f) weeks of growth. Lastly, SETPU is illustrated in its initial (g) and final (h) weeks of the study. It is noteworthy that while certain species produced flowers during their final growth stages, others did not, which reflects variations in natural growth processes and photoperiod sensitivities.

Data pre-processing and augmentation

The data preprocessing and augmentation phase forms the foundation of this research, ensuring the quality, consistency, and usability of the dataset for weed detection and classification. This stage involves a series of carefully designed steps to transform raw images into a structured, annotated dataset suitable for training advanced machine learning models.

We used two different preprocessing methods for our two datasets. In AWD, we used traditional computer vision techniques to preprocess the data. It begins with image normalization, a fundamental step that standardizes the input data. Each image is scaled to a range of 0-1 by dividing all pixel values by 255.0. Following normalization, a color space conversion is performed, transforming the images from the standard RGB (Red, Green, Blue) color space to the HSV (Hue, Saturation, Value) color space using matplotlib's rgb_to_hsv function. This conversion is particularly significant for the application, as the HSV color space offers enhanced discrimination of green hues, which is crucial for accurate plant detection.

The next step in the pipeline is green area detection. Carefully calibrated thresholds for the HSV channels are used to create a mask that highlights potential plant regions. Specifically, hue values ranging from 25/360 to 160/360, minimum saturation value of 0.20, and minimum value of 0.20 are applied. Morphological operations[53] are then applied to refine the green mask and improve the continuity of detected plant areas. Specifically, we implemented morphological closing using a disk-shaped structuring element with a radius of 3 pixels (via skimage.morphology.disk and skimage.morphology.binary_closing) to effectively close small gaps in the vegetation areas. The refined green areas are subsequently subjected to connected component analysis using skimage.measure.label, which identifies and labels distinct regions within the image. We utilized skimage.measure.regionprops to extract properties of these labeled regions, particularly focusing on bounding box coordinates and area measurements. The largest connected component (determined by maximum area) was selected as the primary plant region, effectively filtering out smaller noise segments. This step is critical for differentiating individual plants or plant clusters, enabling more precise analysis and annotation. Our traditional computer vision pipeline successfully identified plants through color-based thresholding, generating distinctive orange masks for clear visualization (Fig. 2a–c).

In BWD, we employed Meta AI's Segment Anything Model 2 (SAM-2)[54,55] for preprocessing, a cutting-edge successor to the original SAM foundation model released in 2024. Unlike its predecessor, SAM-2 extends capabilities to both images and videos through a unified architecture incorporating memory attention blocks, memory encoder, and memory bank components that enable temporal coherence while maintaining object

Species code[49]	Scientific name[50]	Common name[51]	Family	Number of frames / (Weekly_Count)												
				Total	Dataset	W_1	W_2	W_3	W_4	W_5	W_6	W_7	W_8	W_9	W_10	W_11
ABUTH	*Abutilon theophrasti* Medik.	Velvetleaf	Malvaceae	14754	AWD	1084	2451	1212	1819	1414	981	677	1164	1084	1500	1368
				7690	BWD	565	1277	632	948	737	511	353	607	565	782	713
AMAPA	*Amaranthus palmeri* S. Watson.	Palmer amaranth	Amaranthaceae	17525	AWD	1441	1408	2110	2014	2441	1290	923	1478	1393	1667	1360
				9134	BWD	751	734	1100	1050	1272	672	481	770	726	869	709
AMARE	*Amaranthus retroflexus* L.	Redroot pigweed	Amaranthaceae	15380	AWD	1017	1363	2110	1923	1884	1150	736	1237	1082	1596	1282
				8016	BWD	530	710	1100	1002	982	599	384	645	564	832	668
AMATU	*Amaranthus tuberculatus* (Moq.) Sauer.	Waterhemp	Amaranthaceae	14852	AWD	1325	1459	1565	1664	1942	837	730	969	1638	1573	1150
				7741	BWD	691	760	816	867	1012	436	380	505	854	820	599
AMBEL	*Ambrosia artemisiifolia* L.	Common ragweed	Asteraceae	17427	AWD	1022	2215	1846	1739	2162	1093	1066	1432	1092	2045	1715
				9083	BWD	533	1154	962	906	1127	570	556	746	569	1066	894
CHEAL	*Chenopodium album* L.	Common lambsquarter	Chenopodiaceae	8015	AWD	1108	954	1416	661	1056	305	418	641	453	429	574
				4173	BWD	577	496	737	344	549	158	217	333	235	222	298
CYPES	*Cyperus esculentus* L.	Yellow nutsedge	Cyperaceae	14275	AWD	909	1512	1032	1499	2273	978	1224	1391	1182	1170	1105
				7438	BWD	473	787	537	780	1183	509	637	724	615	610	576
DIGSA	*Digitaria sanguinalis* (L.) Scop.	Large crabgrass	Poaceae	16962	AWD	732	1312	2411	2596	1649	1335	1166	1261	1120	1628	1692
				8834	BWD	381	683	1253	1350	858	694	606	656	582	846	880
ECHCG	*Echinochloa crus-galli* (L.) P. Beauv.	Barnyardgrass	Poaceae	16564	AWD	1349	2067	2029	1426	2221	1240	929	1280	1371	1332	1320
				8623	BWD	703	1077	1051	739	1152	644	483	666	714	693	686
ERICA	*Erigeron canadensis* L.	Horseweed	Asteraceae	15134	AWD	930	2183	1691	1542	2715	1189	609	742	915	1217	1401
				7872	BWD	483	1132	876	799	1406	616	315	384	473	630	725
PANDI	*Panicum dichotomiflorum* Michx.	Fall panicum	Poaceae	15182	AWD	1198	1400	2143	1296	1979	952	887	1350	1425	1034	1518
				7890	BWD	621	725	1109	671	1026	493	459	699	738	536	788
SETFA	*Setaria faberi* Herrm.	Gaint foxtail	Poaceae	14635	AWD	1614	1195	2083	1348	1944	1091	715	1466	843	1342	994
				7600	BWD	837	621	1086	703	1010	567	372	769	442	686	508
SETPU	*Setaria pumila* (Poir.) Roem.	Yellow foxtail	Poaceae	15211	AWD	887	1390	1732	1654	2040	1093	747	1361	1325	1348	1634
				7903	BWD	460	721	896	856	1055	567	388	707	689	703	852
SIDSP	*Sida spinosa* L.	Princkly sida	Malvaceae	14452	AWD	1035	1782	1583	1259	2142	1373	804	1059	1186	1303	926
				7505	BWD	515	885	787	627	1074	689	404	532	596	654	465
SORHA	*Sorghum halepense* (L.) Pers.	Johnsongrass	Poaceae	10958	AWD	0	0	1444	1268	1395	945	749	1215	1328	1116	1498
				5682	BWD	0	0	748	657	724	490	389	631	690	579	778
SORVU	*Sorghum bicolor* (L.) Moench.	Shatter cane	Poaceae	9573	AWD	945	1340	1959	832	1065	525	279	748	714	592	574
				4962	BWD	474	673	986	419	536	264	140	388	371	308	298

Table 3. Overview of *AWD* and *BWD*, Corresponding Scientific Codes, Common Name, Family Name, and Weekly Frame Counts Captured for Each Species Across 11 Weeks.

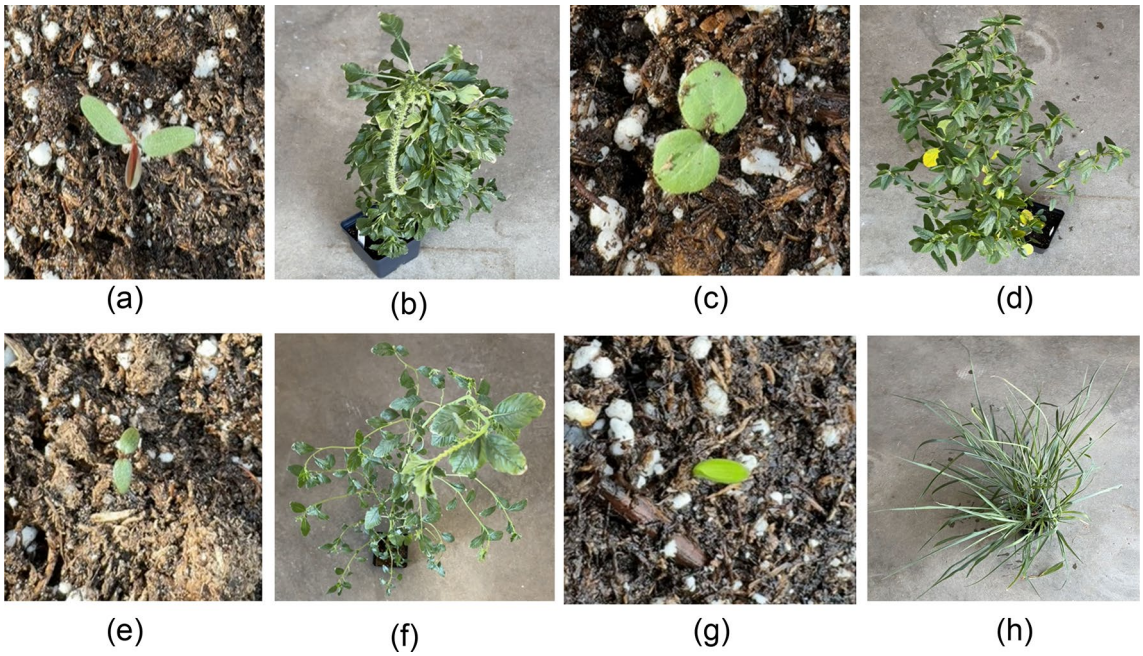

Fig. 1. Growth stage examples of four representative weed species used in this study. (**a**, **b**) AMAPA at week 1 and week 11, respectively; (**c**, **d**) SIDSP at week 1 and week 11; (**e**, **f**) AMATU at week 1 and week 11; (**g**, **h**) SETPU at week 1 and week 11. These images illustrate the morphological changes across the 11-week lifecycle, highlighting the variation in plant structure, size, and complexity that the models must detect and classify accurately.

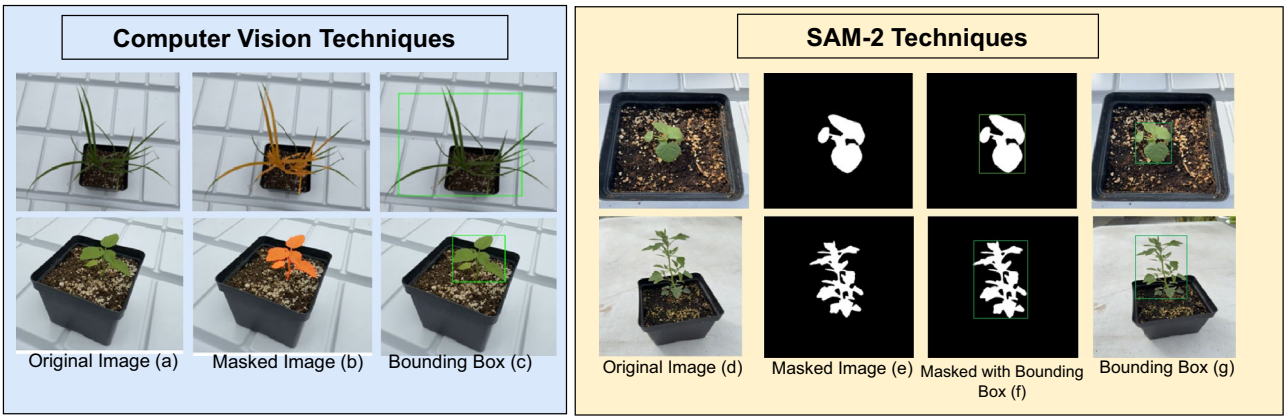

Fig. 2. Comparison of plant detection approaches using traditional computer vision and SAM-2 techniques. Left panel: Computer vision-based detection showing (**a**) original seedling images, (**b**) color-based masking (orange overlay) for plant segmentation, and (**c**) resulting bounding box detection (green). Right panel: SAM-2 segmentation pipeline demonstrating (**d**) original plant images at different growth stages, (**e**) binary mask generation with precise plant-soil separation, (**f**) mask refinement with bounding box constraints, and (**g**) final detection results overlaid on original images.

identity across frames. For our application, SAM-2 functions as a high-precision segmentation tool that first detects potential objects within an image, then generates a binary mask separating each object from the background, followed by precise bounding box generation around the detected objects. This model delivers approximately 6x greater segmentation accuracy compared to the original SAM[56], processes approximately 44 frames per second, and supports interactive refinement through various prompt types including points, boxes, and masks. We implemented SAM-2 specifically for the BWD dataset to ensure a superior labeling process with significantly higher accuracy than that achieved in the AWD dataset, as SAM-2's robust zero-shot generalization capabilities proved particularly suitable for our diverse weed specimen preprocessing requirements[57]. Figure 2 illustrates the process of the data augmentation. We demonstrated two distinct approaches for plant detection and segmentation across developmental stages there. SAM-2 produced highly precise binary masks capturing

intricate plant morphologies (Fig. 2d–g). While both methods effectively generated bounding boxes, SAM-2 exhibited superior performance in delineating complex plant architectures, particularly for mature specimens with multiple leaves. This dual-method comparison validates the robustness of our plant detection system across growth stage.

Data labelling

The labeling process was designed to create comprehensive annotations that capture all relevant information about the detected plants. For each labeled green area, bounding box coordinates were extracted, defining the spatial extent of the plant within the image. These coordinates were determined by the minimum and maximum x and y values of the detected region. Detailed Pascal VOC XML annotations were generated for each processed image, including folder name, filename, dimensions, source database, and precise object bounding boxes with species names. The image processing pipeline was implemented using Python libraries such as Pillow, NumPy, scikit-image, and ElementTree, with the resulting XML files stored in a structured 'labels' directory. The rigorous data preprocessing and labeling methodology yielded a dataset of high quality, characterized by precise annotations and consistent formatting.

To further enhance accuracy, a thorough quality control process was implemented. Each image was meticulously reviewed using LabelImg software following the initial automated labeling. The labeling convention incorporated both the species code and the week number, providing a comprehensive identifier for each plant's growth stage and taxonomy. This detailed labeling strategy significantly enhanced the dataset's utility for tracking plant development over time and for species-specific analysis. Figure 3 illustrates this process, presenting a side-by-side comparison of an original image and its corresponding labeled version, referred to as the ground truth.

Following the annotation of the dataset, the data was divided into training, validation, and test sets. For the AWD, 184,719 images (80%) were used for training the object detection models, while 23,090 images (10%) were used for validation during training. The remaining 23,090 images (10%) were held out for testing the performance of the trained model. For the BWD, 96,272 images (80%) were used for training, with 12,034 images (10%) allocated for validation during training. The remaining 12,035 images (10%) were held out for testing the model's performance.

Models and algorithms

In this study, two experiments were conducted using the two different datasets. In the first experiment, involving the AWD, two advanced deep learning models were employed for weed detection and classification: RetinaNet[19] with a ResNeXt-101[58] backbone and Detection Transformer (DETR)[16] with a ResNet-50[14] backbone. In the second experiment, using the BWD, several models were utilized, including RetinaNet[19] with a ResNeXt-101[58] backbone, Detection Transformer (DETR)[16] with a ResNet-50[14] backbone, DINO[13] Transformer with a Swin[15] Transformer backbone, DINO[13] Transformer with a ResNet-101[14] backbone, EfficientNet B4[17] with a ResNet-50[14] backbone, YOLO v8[18], and our custom architectural model named WeedSwin. These models were tasked with classifying weed species and identifying their respective growth stages (in weeks, while simultaneously localizing them within the images through bounding box predictions.

We trained all models, including baselines, for 12 epochs with batch size 16 on identical hardware configurations to ensure fair comparisons. The models were configured and trained using PyTorch and mmDetection. We used NVIDIA A100 GPU with 80GB memory and an Intel Xeon Gold 6338 CPU (2.00GHz) on a Linux system for training all our models. All models were initialized with weights pre-trained on COCO dataset, with validation performed after each epoch to monitor convergence. To ensure optimal performance and fairness, we performed

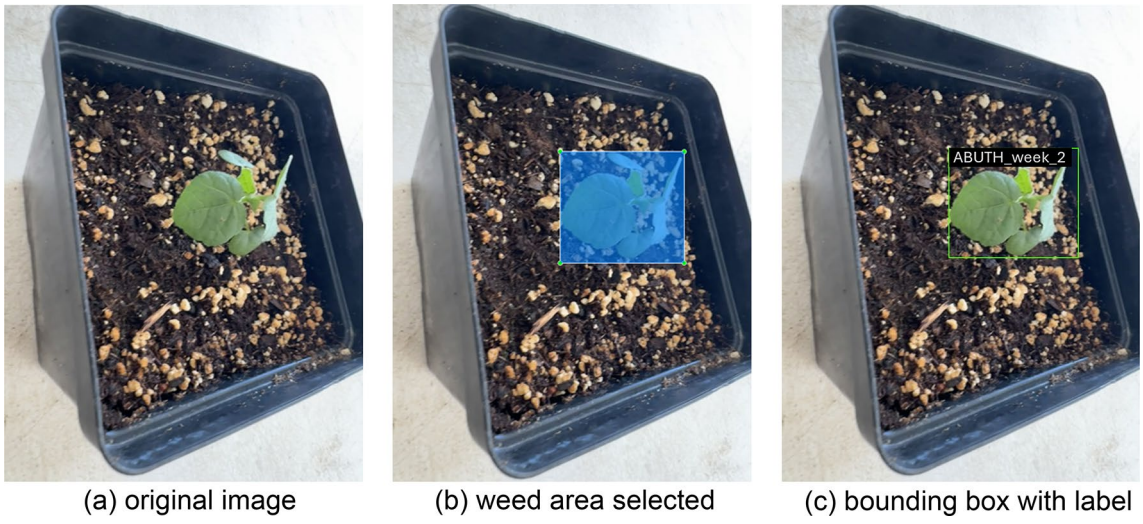

Fig. 3. Sequential demonstration of the weed detection annotation workflow. (**a**) original image of the weed plant, (**b**) the selected area highlighted in blue (**c**) final labeled image showing with a bounding box around the plant with its species name (ABUTH) and growth stage (Week 2).

light hyperparameter tuning for each model, focusing primarily on learning rate, weight decay, and confidence/NMS thresholds. For all models, the initial learning rate was selected from 1e−3, 1e−4, 5e−5 based on early validation loss behavior, while weight decay was varied in 1e−4, 5e−5. The confidence thresholds for inference were fine-tuned per model in the range [0.05, 0.3] to balance false positives and detection robustness. Anchor configurations (e.g., aspect ratios, scales) and optimizer types (e.g., SGD vs AdamW) were also evaluated during early experiments for the baseline models. Final choices were determined using grid search on the validation set performance for each dataset.

RetinaNet with ResNeXt-101

In this study, we implemented the RetinaNet architecture[19] with a ResNeXt-101 backbone[59] for weed detection. RetinaNet, a single-stage object detector, addresses class imbalance through Focal Loss while maintaining high detection accuracy[60]. Our implementation utilizes a ResNeXt-101 (32×4d) backbone, which enhances the model's feature extraction capabilities through its cardinality-based approach.

The ResNeXt-101 backbone employs a split-transform-merge strategy, where the input is divided into 32 parallel paths ($C = 32$, cardinality), each with a bottleneck width of 4 channels ($\text{base\_width} = 4$). This architecture can be formally expressed as:[61]

$$\mathrm{F}(x) = \sum_{i=1}^{C} \mathrm{T}_i(x) \tag{1}$$

where $C = 32$ represents the cardinality, and T_i denotes the transformation function for the *i*-th path. Each transformation follows a bottleneck design with 1×1, 3×3, and 1×1 convolutions.

The Feature Pyramid Network (FPN) neck connects to this backbone, generating multi-scale feature maps $\{P_2, P_3, P_4, P_5, P_6\}$ with corresponding channels of 256. For a feature level *l*, the FPN output can be described as:

$$P_l = \mathrm{C}(U(P_{l+1}) + \mathrm{L}(C_l)) \tag{2}$$

where U represents upsampling, L is a 1×1 convolution lateral connection, and C is a 3×3 convolution for smoothing.

The RetinaHead subnet processes these features using Focal Loss, defined as[61]:

$$FL(p_t) = -\alpha_t (1 - p_t)^{\gamma} \log(p_t) \tag{3}$$

where $\gamma = 2.0$ is the focusing parameter and $\alpha_t = 0.25$ is the balanced variant of the focal loss.

The model predicts across 174 weed classes using dense anchors at multiple scales ($\text{strides} = [8, 16, 32, 64, 128]$) and aspect ratios ([0.5, 1.0, 2.0]).

During training, we employed the AdamW optimizer with an initial learning rate of 1e-4 and weight decay of 1e-4. The learning rate follows a multi-step schedule using a gamma of 0.1. The model was initialized with weights pre-trained on the COCO dataset and trained for 12 epochs with validation performed after each epoch. To maintain stable training, we froze the first stage of the backbone ($\text{frozen\_stages} = 1$) while allowing batch normalization statistics to be updated ($\text{requires\_grad} = \text{True}$). The inference process employs a confidence threshold of 0.05 and Non-Maximum Suppression (NMS) with an IoU threshold of 0.5, limiting to a maximum of 100 detections per image. This configuration balances detection accuracy with computational efficiency while maintaining robust performance across various weeds.

Detection transformer with ResNet-50

The Detection Transformer (DETR)[16] represents a paradigm shift in object detection by eliminating the need for hand-crafted components like non-maximum suppression and anchor generation[62]. Our implementation employs a ResNet-50 backbone[63] coupled with a transformer encoder-decoder architecture for weed detection, processing a fixed set of $N = 100$ object queries in parallel.

The architecture begins with a ResNet-50 backbone that extracts hierarchical features through its four-stage design. The backbone output $\mathrm{X} \in \mathrm{R}^{C \times H \times W}$ (where $C = 2048$) is processed through a channel mapper that reduces dimensionality to $d = 256$ channels. The matching cost between ground truth y_i and prediction $\hat{y}_{\sigma(i)}$ can be expressed as[64]:

$$\mathrm{L}_{\text{match}}(y_i, \hat{y}_{\sigma(i)}) = -1_{\{c_i \neq \varnothing\}} \hat{p}_{\sigma(i)}(c_i) + 1_{\{c_i \neq \varnothing\}} \lambda_L \| b_i - \hat{b}_{\sigma(i)} \|_1 \tag{4}$$

where c_i represents the target class label, b_i is the ground truth box, and σ is the optimal assignment. The transformer encoder-decoder architecture processes the features through self-attention mechanisms. For a given query Q, key K, and value V, the multi-head attention is computed as $\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h) W^O$, where each attention head is calculated as $\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$ with W_i^Q, W_i^K, W_i^V, and W^O being learnable parameter matrices that project the inputs into different representation subspaces. The final loss function combines classification and box regression terms[64]:

$$\mathrm{L} = \lambda_{cls} \sum_{i=1}^{N} [-\log \hat{p}_{\hat{\sigma}(i)}(c_i)] + \lambda_{box} \sum_{i=1}^{N} 1_{\{c_i \neq \varnothing\}} \mathrm{L}_{box}(b_i, \hat{b}_{\hat{\sigma}(i)}) \tag{5}$$

where L_{box} combines the L1 loss and the generalized IoU loss:

$$L_{box}(b_i, \hat{b}_i) = \lambda_{L1}\|b_i - \hat{b}_i\|_1 + \lambda_{giou}L_{giou}(b_i, \hat{b}_i) \tag{6}$$

For training, we utilize a bipartite matching loss that optimally assigns predictions to ground-truth objects using the Hungarian algorithm. The total loss is a combination of classification and box regression terms:

$$L = \lambda_{cls}L_{cls} + \lambda_{box}L_{box} + \lambda_{giou}L_{giou} \tag{7}$$

where $\lambda_{cls} = 1$, $\lambda_{box} = 5$, and $\lambda_{giou} = 2$ are loss weights. The classification loss L_{cls} uses cross-entropy with balanced weighting (background weight = 0.1), while box regression employs L1 and GIoU losses. Training proceeds with the AdamW optimizer (learning rate = 10^{-4}, weight decay = 10^{-4}) for 12 epochs. We implement a multi-step learning rate schedule with a milestone at epoch 334 and a decay factor of 0.1. During inference, the model directly outputs a set of 100 predictions with confidence scores, requiring no post-processing beyond score thresholding. This end-to-end approach demonstrates particular efficacy in handling the complex spatial relationships and varying scales characteristic of weed detection, while maintaining computational efficiency through parallel prediction generation.

The algorithm of DETR[16] presents the model training process where the goal is to optimize the model parameters, denoted by θ. Initially, the model parameters are set to their initial values. The training process runs for 12 number of epochs, iterating over mini-batches of the training dataset in each epoch. For each mini-batch, the model makes predictions $\hat{y}$, and the classification and regression losses are computed. These losses are summed to obtain the total loss, which is then used to calculate the gradients through backpropagation. The model parameters are updated using the AdamW optimizer. At regular intervals, defined by `val_interval`, the model's performance is evaluated on the validation dataset, and the model checkpoint is saved if there is an improvement in performance.

DINO transformer with Swin and ResNet-101

In this work, we employ DINO[13], an extension of the DETR[16] architecture, which enhances object detection performance by incorporating advanced query selection and a denoising training strategy. For our weed detection task, we utilize DINO[13] with two different backbone architectures: the Swin[15] Transformer and ResNet-101[14]. Each of these backbones brings unique advantages to the challenge of processing high-resolution weed imagery. The Swin Transformer[15], with its shifted window mechanism and hierarchical feature extraction, excels at capturing both detailed and contextual weed information. On the other hand, ResNet-101[14]'s deep residual learning framework is particularly effective at identifying intricate weed patterns, providing strong feature extraction capabilities.

The DINO[13] architecture processes weed images through multiple stages. The initial stage involves extracting multi-scale features using either the Swin Transformer[15] or ResNet-101[14] backbone. The Swin Transformer[15], with its innovative design of shifted windows, is particularly well-suited for capturing both fine-grained weed characteristics and broader contextual information, which is critical for distinguishing weeds in complex agricultural environments. Alternatively, ResNet-101[14] provides a powerful residual learning mechanism, allowing it to learn and extract intricate features of the weed species effectively, particularly when dealing with the visual complexity of weed patterns. Once the multi-scale features are extracted, they are enhanced with positional embeddings, which help the model maintain spatial awareness of the objects in the image. This spatial awareness is crucial for accurately localizing weeds, particularly in complex agricultural scenes where overlapping vegetation is common. The enhanced features are then passed through several encoder layers, which are responsible for further refining the feature representation. A unique aspect of DINO[13] is its sophisticated query selection mechanism, which dynamically adapts to the complexity of the scene. This query selection allows the architecture to efficiently process images containing varying weed densities and distributions, adapting its attention to areas of interest.

The decoder portion of the DINO[13] architecture includes a denoising training strategy, employing both content queries and base anchors. This denoising mechanism is highly beneficial for weed detection, as it enhances the model's ability to distinguish between similar-looking weed species and effectively handles cases of overlapping vegetation. During training, the model uses a contrastive denoising approach in the matching module. This approach involves selecting positive and negative samples, which helps the model refine its discriminative capabilities, improving its ability to distinguish different weed classes and reducing false positives. Finally, the model uses a Classification and Detection Network (CDN) to produce the refined predictions for weed instances. This CDN is specifically tuned for our 174-class weed detection task, incorporating class-specific features learned during the training process. The result is a robust model that can provide highly accurate and interpretable detection results, which are essential for agricultural applications where reliability and precision are of utmost importance.

The overall architecture starts with input weed images, which are processed to extract multi-scale features. These features are fed into an encoder consisting of multiple layers, enhanced by positional embeddings. The encoder output is then refined through a query selection mechanism, leading to the decoder layers. The decoder is responsible for further processing using both content queries and base anchors, facilitating accurate localization and classification of weeds. The matching module, equipped with contrastive denoising training, distinguishes between positive and negative instances, enhancing the model's ability to identify subtle inter-species differences. The output layer provides the final classification and detection results for each weed instance.

WeedSwin - proposed model

While existing architectures like Swin Transformer[15] and DINO[13] demonstrate strong performance in general object detection, they face significant limitations in agricultural weed detection. Standard Swin Transformer's fixed window partitioning struggles with the highly variable scale characteristics of weeds across growth stages, while DINO's query selection mechanism fails to capture the subtle morphological differences between similar plant species. To address these domain-specific challenges, we propose WeedSwin, an enhanced Swin-based architecture optimized for weed detection and classification tasks.WeedSwin introduces four key architectural components specifically designed for agricultural applications: (1) Enhanced Backbone with Progressive Attention Heads (6→12→24→48) that dynamically adapt to the dramatic scale variations between seedling and mature weeds, providing multi-scale feature representation critical for detecting plants at different growth stages. (2) Specialized Feature Enhancement Neck utilizing a Channel Mapper with ReLU activation that preserves fine-grained morphological details essential for differentiating between visually similar weed species. (3) Modified Encoder with 8 layers (versus 6 in standard implementations) and 16 attention heads, enabling more comprehensive feature extraction from complex agricultural scenes. (4) Enhanced Decoder with 8 layers and optimized cross-attention mechanisms that better capture contextual relationships in densely vegetated environments. These modifications collectively improve the model's capacity to detect and classify weeds effectively in diverse agricultural environments.

We employed the backbone with optimized parameters, defined as $E_b = \mathrm{SwinT}(d_{\mathrm{model}} = 192, d_{\mathrm{heads}} = [6, 12, 24, 48], w_{\mathrm{size}} = 12)$, where d_{model} represents the initial embedding dimension and d_{heads} represents the progressive scaling of attention heads across layers. For feature enhancement, we introduced a Channel Mapper with enhanced feature processing, expressed as $F_{\mathrm{out}} = \sigma(\mathrm{BN}(W \cdot F_{\mathrm{in}} + b))$, where F_{out} is the output feature map, F_{in} is the input feature map, W represents learnable weights, BN denotes Batch Normalization, and σ is the ReLU activation function.

Our architecture deepens both the encoder and decoder to 8 layers ($L = 8$) with enhanced multi-head attention, implementing 16 attention heads with embedding dimension $d = 384$, resulting in $d_{\mathrm{head}} = 24$ per attention head. The multi-head attention mechanism is defined as:

$$\mathrm{MHA}(Q, K, V) = \mathrm{Concat}\left(\mathrm{head}_1, ..., \mathrm{head}_{16}\right) W^O \tag{8}$$

where each head is computed as: $\mathrm{head}_i = \mathrm{Attention}(QW_i^Q, KW_i^K, VW_i^V)$. The enhanced decoder implements an 8-layer structure with improved cross-attention mechanism, defined as:

$$\mathrm{CrossAttn}(q, k, v) = \mathrm{SoftMax}(\frac{q \cdot k^T}{\sqrt{d_k}})v \tag{9}$$

where q represents queries, k represents keys, and v represents values, with $d_k = 384$. We increased transformer layers to 8, enhanced the embedding dimension to 384, and optimized attention heads to 16 for better feature representation. The enhanced feature extraction is achieved through our modified Channel Mapper:

$$F_{\mathrm{enhanced}} = \mathrm{ChannelMapper}(F_{\mathrm{input}}) = \mathrm{ReLU}(\mathrm{BN}(\mathrm{Conv1x1}(F_{\mathrm{input}}))) \tag{10}$$

$$d_{\mathrm{model}} = n_{\mathrm{heads}} \times d_{\mathrm{head}} \quad (384 = 16 \times 24) \tag{11}$$

We carefully balanced the model's capacity through the relationship d_{model}, ensuring efficient attention computation while maintaining representational power. The architecture demonstrates particular effectiveness for weed detection through its multi-scale feature processing capability, with progressive attention heads (6→12→24→48) enabling effective processing of weeds at various scales. The increased embedding dimension provides richer feature representation, with a capacity gain of approximately 1.75× ($\mathrm{Capacity}_{\mathrm{gain}} = (384/256) \times (8/6)$). The deeper architecture facilitates better information propagation with a receptive field of $\mathrm{Receptive}_{\mathrm{field}} = \mathrm{base}_{\mathrm{size}} \times 2^{(L-1)}$, where $L = 8$ provides broader context capture compared to the original $L = 6$.

Figure 4 shows the proposed WeedSwin architecture, a novel transformer-based framework for weed detection and classification. The architecture consists of three main components: an encoder, a decoder, and a Feature Pyramid Network (FPN) for detection. The encoder pathway employs a hierarchical Swin Transformer structure with four progressive stages, where the attention heads increase from 6 to 48. Each stage operates at different spatial resolutions (from H/4×W/4 to H/32×W/32) with a consistent embedding dimension of 384 and window size of 12. The stages contain varying numbers of Transformer blocks (×2, ×2, ×6, ×2 respectively). A Channel Mapper module with fully connected layers and ReLU activation bridges the encoder and decoder, enhancing feature transformation. The decoder comprises eight blocks (D1-D3 shown) with cross-attention mechanisms and skip connections from corresponding encoder stages, progressively recovering spatial details. The detection head utilizes a Feature Pyramid Network structure with dual branches: a Classification Branch predicting weed class and growth stages, and a Bounding Box Regression Branch estimating object coordinates. These branches process the feature map through parallel pathways, combining their outputs via multiplication and FC layers to generate the final feature vector.

While our model introduces additional parameters, we maintained efficiency through optimized attention head distribution, balanced embedding dimensions, and efficient feature enhancement in the neck. The theoretical computational complexity remains $O(N \times d_{\mathrm{model}} + N^2 \times d_{\mathrm{head}})$, where N is the sequence

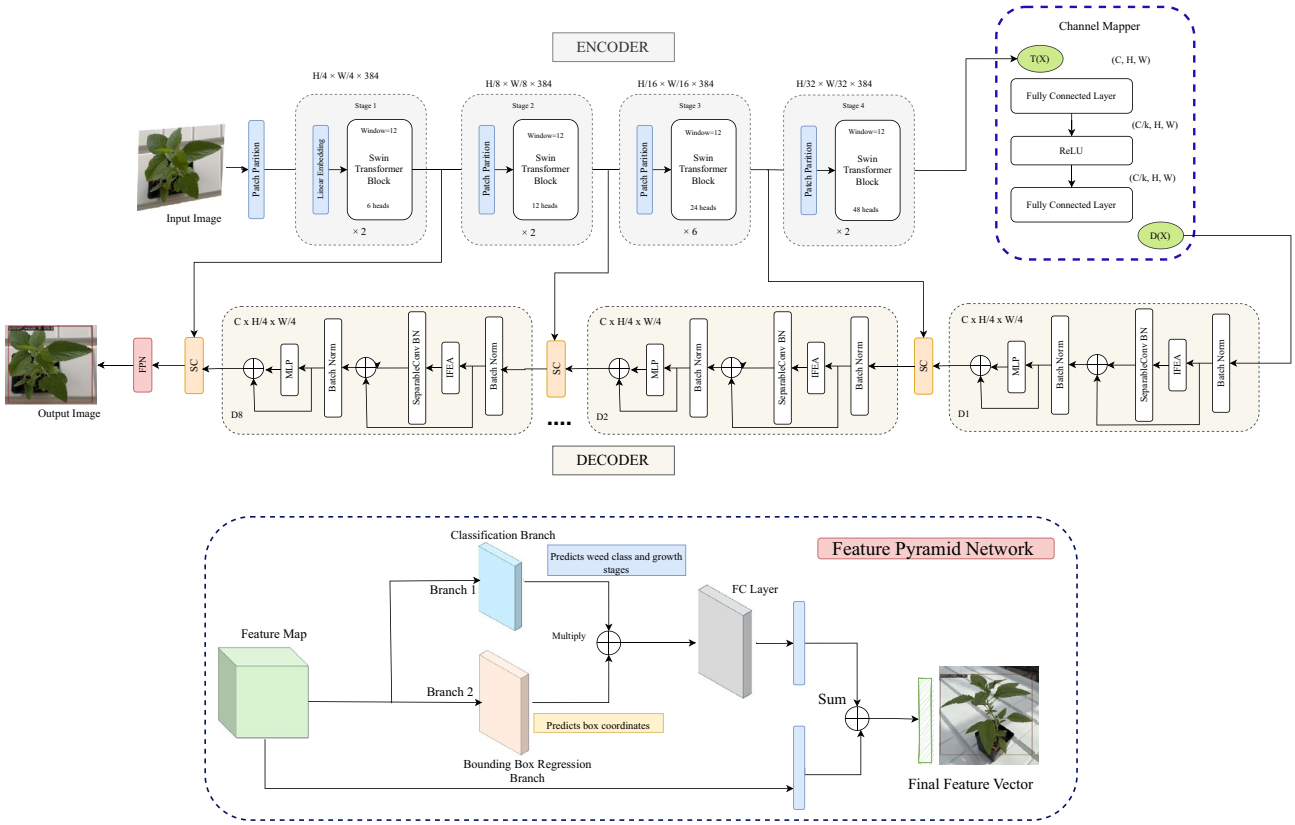

Fig. 4. WeedSwin: A hierarchical vision transformer architecture incorporating progressive attention heads (6-48), feature enhancement through Channel Mapper, eight decoder blocks with cross-attention mechanisms, and Feature Pyramid Network for precise weed detection and localization.

length, making it practical for real-world agricultural applications. This enhanced architecture demonstrates our commitment to improving weed detection accuracy while maintaining computational efficiency, making it suitable for practical agricultural applications.

Data: Training dataset D, Validation dataset D_{val}, Number of classes $C = 174$
Result: Trained WeedSwin model for weed detection
Initialize backbone with Swin Transformer parameters;
Initialize Channel Mapper with enhanced features $d_{out} = 384$;
Initialize Enhanced Encoder: $L_{layers} = 8$, $n_{heads} = 16$, dropout=0.2;
Initialize Enhanced Decoder: $L_{layers} = 8$, $n_{heads} = 16$, dropout=0.2;
Set learning rate η, weight decay λ, loss weights $[\alpha_1, \alpha_2, \alpha_3]$;
for *epoch = 1 to max_epochs (12)* **do**
  for *each batch* (I, B, L) *in* D **do**
    $F_b = \text{SwinTransformer}(I)$ with progressive heads $[6, 12, 24, 48]$;
    $F_e = \text{ChannelMapper}(F_b)$ with ReLU activation;
    $F_{enc} = \text{EnhancedEncoder}(F_e)$ with 8 layers, 16 heads;
    $F_{dec} = \text{EnhancedDecoder}(Q, F_{enc})$ with 8 layers, 16 heads;
    Compute predictions $\hat{y}$;
    $L_{total} = \alpha_1 L_{cls} + \alpha_2 L_{box} + \alpha_3 L_{giou}$;
    Update parameters using AdamW: $\theta \leftarrow \theta - \eta \nabla_\theta L_{total}$;
  if *epoch % val_interval == 0* **then**
    Compute mAP and recall on D_{val};
    Save model if performance improves;
    Update η using MultiStepLR scheduler;
Apply Non-Maximum Suppression for final predictions;
return Best model based on validation performance;

Algorithm 1. WeedSwin: Model Training Process

Algorithm 1 presents our WeedSwin training process, an enhanced Swin architecture optimized for weed detection. Our key modifications include an enhanced encoder and decoder, each featuring 8 layers and 16 attention heads with 0.2 dropout for better feature transformation. During training, it processes images through progressive feature extraction ([6, 12, 24, 48] heads), enhanced feature mapping, and our modified attention mechanisms. The model optimizes a combined loss function incorporating classification, bounding box, and GIoU losses. The training process includes regular validation checks, model checkpointing, and learning rate scheduling, concluding with Non-Maximum Suppression for final predictions.

Evaluation metrics

To evaluate our weed detection model's performance comprehensively, we employ multiple metrics: Average Precision (AP), Average Recall (AR), Mean Average Precision (mAP), and Frames Per Second (FPS). These metrics collectively assess both accuracy and computational efficiency.

The fundamental components of our evaluation are Precision (P) and Recall (R), defined as: $\mathrm{P} = \frac{TP}{TP+FP}$, and $R = \frac{TP}{TP+FN}$. Where a TP (true positive) is a detected bounding box that correctly identifies a weed species and has an IoU above a specified threshold (e.g., 0.50) with the ground truth bounding box. A FP (false positive) is a detection that either does not sufficiently overlap with any ground truth box or incorrectly identifies the weed species. A FN (false negative) occurs when a ground truth weed instance is not detected by the model.

Average Precision (AP) provides a comprehensive view of detection performance by integrating precision over recall. It effectively summarizes the precision-recall curve[65] into a single value, capturing the model's ability to make accurate detections across different confidence thresholds. The AP is calculated as:

$$AP = \int_0^1 P(R)\, dR \tag{12}$$

where P(R) represents the precision value at each recall level. This integration over the entire recall range [0,1] ensures that both high precision and high recall are rewarded, providing a balanced measure of detection quality.

Average Recall (AR) quantifies the model's detection coverage across various IoU thresholds. This metric is particularly important for assessing the model's ability to detect weeds under different overlap criteria, making it valuable for understanding detection robustness. AR is computed as:

$$AR = \frac{1}{N} \sum_{i=1}^{N} R_{\max}(IoU_i) \tag{13}$$

where N is the number of IoU thresholds considered, and $R_{\max}(IoU_i)$ represents the maximum recall achieved at each IoU threshold. This averaging across multiple IoU thresholds provides insight into the model's detection stability under varying overlap requirements.

Mean Average Precision (mAP) evaluates the model's performance across all weed classes, providing a comprehensive metric for multi-class detection scenarios. This metric is particularly crucial in weed detection as it accounts for the varying difficulties in detecting different weed species and growth stages. The mAP is calculated as[66]:

$$mAP = \frac{1}{C} \sum_{c=1}^{C} AP_c \tag{14}$$

where C represents the total number of weed classes, and AP_c is the Average Precision for each class. This averaging across classes ensures that the model's performance is evaluated fairly across all weed types, regardless of their representation in the dataset. We also measure computational efficiency using Frames Per Second (FPS):

$$FPS = \frac{N_{\text{frames}}}{T_{\text{processing}}} \tag{15}$$

where N_{frames} is the number of processed frames and $T_{\text{processing}}$ is the total processing time in seconds. FPS is crucial for assessing real-time detection capabilities, particularly important for practical field applications where rapid weed identification is essential. These metrics are evaluated across various IoU thresholds (0.5:0.95) to provide a comprehensive assessment of our model's detection accuracy, classification precision, and operational efficiency in real-world scenarios.

Experimental result

In this research, we implemented several state-of-the-art algorithms, including DINO[13] with ResNet[14] and Swin[15] backbones, DETR[16], EfficientNet B4[17], RetinaNet[19], and Weedswin (our custom architecture). Among these, DETR[16] and RetinaNet[19] were applied exclusively to our AWD dataset, while all the algorithms were evaluated on the BWD dataset. The evaluation encompasses both training and test datasets, with a detailed analysis across 16 weed species. We employed all the evaluation metrics at different IoU thresholds and detection

limits, as well as mAP and mAR. Furthermore, we compared the inference speed (FPS) of the models to provide a comprehensive view of their performance and capabilities.

Object detection result on AWD dataset

Table 4 presents a comparative analysis of DETR[16] and RetinaNet[19] models evaluated on the AWD dataset[52]. The performance metrics include mAP and mAR for both training and test sets, along with inference speed measured in FPS. RetinaNet[19] demonstrates superior performance across all evaluation metrics. In terms of accuracy, RetinaNet achieves higher mAP scores of 0.907 and 0.904 on training and test sets respectively, compared to DETR's 0.854 and 0.840. This pattern continues in the recall metrics, where RetinaNet approaches near-perfect scores with mAR values of 0.997 for training and 0.989 for testing, while DETR achieves 0.941 and 0.936 respectively. Most notably, RetinaNet exhibits substantially faster inference speed at 347.22 FPS, which is more than five times faster than DETR's 65.43 FPS.

Table 5 presents a detailed species-wise performance analysis, showing averaged results across 11 weeks for 16 weed species[52]. RetinaNet[19] exhibits more consistent performance across species, with notable achievements for AMATU (mAP 0.832) and AMAPA (mAP 0.877), though showing reduced effectiveness with ECHCG (mAP 0.566). DETR[16] displays varying performance levels, performing strongly with AMBEL (mAP 0.817) and SIDSP (mAP 0.771), but struggling with CHEAL (mAP 0.503) and SORHA (mAP 0.527). RetinaNet maintains higher recall scores across species, while both models show expected performance decreases as IoU thresholds increase from 0.5 to 0.75.

Figure 5 demonstrates the detection performance of DETR[16] and RetinaNet[19] models across three distinct weed species in the AWD dataset[52]. The figure presents a comparative analysis through original images, ground truth annotations, and model predictions. In the first row, both models successfully detected CHEAL at week 7, albeit with low confidence scores and imperfect bounding box generation. These imperfections can be attributed to the AWD dataset's labeling inconsistencies, which prompted the creation of our more accurately labeled BWD dataset. The second row showcases AMBEL at week 7, where both models exhibited strong detection performance. RetinaNet[19] achieved a marginally superior confidence score of 99.3% compared to DETR's 89.6%. Similarly, the third row featuring DIGSA at week 5 demonstrates excellent detection capabilities from both models, with confidence scores exceeding 90%. Throughout all three species, RetinaNet consistently produced slightly higher confidence scores, particularly excelling in DIGSA detection with a 99.9% confidence score.

Object detection result on BWD dataset

Table 6 presents a comprehensive performance comparison of various state-of-the-art object detection models evaluated on the BWD dataset. The comparison encompasses seven different model configurations, each characterized by their specific backbone architectures, and evaluated using mAP, mAR on both training and test sets, along with their inference speed (FPS), computational complexity (FLOPs), and model size (Parameters).

The results demonstrate that while DINO[13] with Swin[15] backbone achieves marginally higher mAP scores (0.993 training, 0.994 test), our proposed WeedSwin architecture exhibits comparable accuracy (0.992 training, 0.993 test) and superior recall performance (0.985 training and test mAR) while offering significant advantages in processing speed. To assess the statistical robustness of WeedSwin, we conducted three independent runs with identical configurations, yielding a mean mAP of 0.993 with a standard deviation of 0.0015. Statistical analysis revealed a 95% confidence interval of 0.993 ± 0.004 (0.989-0.997), demonstrating the model's consistency across multiple initializations. This narrow confidence interval confirms that WeedSwin's superior performance is statistically significant and not attributable to chance or favorable initialization conditions. The minimal variability between runs further validates the architecture's stability in maintaining high detection accuracy across diverse weed species and growth stages. WeedSwin achieves an impressive 218.27 FPS, which is approximately 43% faster than DINO-Swin's 152.62 FPS. Additionally, WeedSwin requires less computational resources (0.114T FLOPs) compared to DINO-Swin (0.241T FLOPs) and has a substantially smaller parameter count (40.476M vs 0.209G).

When examining the broader landscape of models, RetinaNet[19] with ResNext[58] 101 backbone and EfficientNet B4[17] with ResNet[14] 50 backbone achieve higher FPS (504.36 and 490.65 respectively) but demonstrate lower accuracy (0.982 and 0.989 test mAP respectively) and recall rates (0.947 and 0.967 test mAR respectively). YOLO v8[18] with CSPNet[67] backbone demonstrates impressive performance with 0.990 test mAP and 0.971 test mAR, while achieving a remarkable 422.48 FPS with efficient computational requirements (100.15G FLOPs and 36.12M parameters). DETR[16] with ResNet[14] 50 backbone shows competitive accuracy (0.984 test mAP) and high recall (0.985 test mAR) but operates at a significantly lower speed of 80.65 FPS. DINO[13] with ResNet[14] 101 backbone presents balanced performance (0.992 test mAP, 0.979 test mAR, 120.67 FPS) but still falls short of WeedSwin's efficiency and recall capability.

Model	mAP		mAR		FPS
	Train	Test	Train	Test	
DETR[16]	0.854	0.840	0.941	0.936	65.43
RetinaNet[19]	0.907	**0.904**	0.997	**0.989**	**347.22**

Table 4. Performance Comparison of DETR[16] and RetinaNet[19] on AWD Training and Test Datasets[52]. Significant values are in bold.

	DETR[16]				RetinaNet[19]			
Species code	Average mAP	Average mAP_50	Average mAP_75	Average Recall	Average mAP	Average mAP_50	Average mAP_75	Average Recall
ABUTH	0.683	0.907	0.719	0.973	0.720	0.924	0.779	0.993
AMAPA	0.617	0.835	0.672	0.975	0.877	0.985	0.939	0.994
AMARE	0.575	0.807	0.598	0.957	0.617	0.941	0.684	0.987
AMATU	0.536	0.721	0.565	0.869	0.832	0.977	0.905	0.997
AMBEL	0.817	0.978	0.898	0.993	0.663	0.926	0.740	0.994
CHEAL	0.503	0.846	0.502	0.962	0.871	0.993	0.957	0.997
CYPES	0.643	0.861	0.680	0.986	0.781	0.971	0.853	0.995
DIGSA	0.578	0.864	0.594	0.995	0.664	0.878	0.753	0.976
ECHCG	0.655	0.899	0.715	0.986	0.566	0.814	0.612	0.950
ERICA	0.718	0.918	0.752	0.977	0.678	0.918	0.749	0.992
PANDI	0.670	0.929	0.723	0.979	0.724	0.934	0.799	0.993
SETFA	0.680	0.903	0.756	0.990	0.785	0.967	0.854	0.993
SETPU	0.597	0.852	0.652	0.973	0.794	0.949	0.858	0.993
SIDSP	0.771	0.980	0.826	0.993	0.739	0.954	0.832	0.991
SORVU	0.582	0.791	0.624	0.871	0.713	0.925	0.789	0.995
SORHA	0.527	0.715	0.544	0.892	0.693	0.858	0.780	0.894

Table 5. Performance Comparison of DETR[16] and RetinaNet[19] on AWD Training and Test Datasets[52].

The performance of WeedSwin is particularly noteworthy as it achieves a near-optimal balance between accuracy, recall, and computational efficiency. The marginal difference in mAP (0.001 lower than DINO-Swin) is negligible in practical applications, while its superior mAR performance, substantially reduced computational requirements, and significant gain in processing speed represents significant advantages for real-time weed detection systems.

Table 7 presents a comprehensive evaluation of seven cutting-edge object detection architectures tested on the BWD dataset. Each model's performance is meticulously evaluated across 16 distinct weed species using four critical metrics: mAP, mAP at 50% IoU threshold (mAP_50), mAP at 75% IoU threshold (mAP_75), and Average Recall (AvgRec). The proposed WeedSwin model demonstrates remarkable superiority across the board, consistently achieving exceptional scores across all metrics. For most species, it achieves perfect or near-perfect scores (1.000) for mAP_50, and notably high mAP_75 values, indicating superior detection capability even at stricter intersection-over-union thresholds. This performance is particularly impressive for challenging species like AMBEL where WeedSwin achieves 0.994 mAP, 1.000 mAP_50, 1.000 mAP_75, and 1.000 AvgRec, surpassing all other models. When examining individual model performances, RetinaNet[19] with ResNeXt-101[58] backbone shows strong baseline performance with mAP values consistently above 0.900 for most species. Its performance is particularly notable for SIDSP with 0.969 mAP. EfficientNet-B4[17] demonstrates slightly better performance than RetinaNet in several cases, especially for species like AMBEL (0.981 mAP) and SIDSP (0.989 mAP), suggesting that its architecture might be better suited for certain weed morphologies. DETR[16], while showing more modest performance among the seven models, still maintains respectable scores. Its performance is particularly challenged with species like AMATU where it achieves 0.899 mAP, significantly lower than WeedSwin's 0.975. However, DETR maintains strong Average Recall scores above 0.989 for most species, indicating good detection capability despite lower precision.

DINO[13] with ResNet-50[14] backbone shows consistent improvement over DETR across all metrics, positioning itself as a strong competitor to RetinaNet[19] and EfficientNet-B4[17]. Its performance is particularly impressive for species like SIDSP and AMBEL, where it achieves mAP scores of 0.990, demonstrating the effectiveness of its architecture. Interestingly, DINO with Swin[15] backbone shows substantially lower performance compared to its ResNet-50 counterpart, with mAP values averaging around 0.7-0.8 across most species. This significant difference suggests that the combination of DINO architecture with Swin transformer may not be optimally configured for this specific weed detection task. YOLO v8[18] with CSPNet[67] backbone demonstrates impressive performance, achieving mAP scores comparable to DINO with ResNet-50 and often exceeding EfficientNet-B4 results. For species like SIDSP, it reaches 0.991 mAP, showcasing its effectiveness as a real-time detection model that doesn't compromise on accuracy. Its consistent performance across all species, with most mAP scores above 0.950, highlights the robustness of its architecture for weed detection tasks.

The most challenging species across all models appear to be SORHA and SORVU, where even the best-performing WeedSwin model achieves relatively lower scores (0.932 and 0.959 mAP respectively). This consistent pattern suggests inherent difficulties in detecting these particular species. These two species are in the same genus and are closely related, and, therefore, have similar morphological characteristics and growth patterns. The performance gap is particularly pronounced for DINO with Swin backbone, which achieves only 0.710 and 0.722 mAP for these species respectively, with a notably low AvgRec of 0.894 for SORVU. In terms of Average Recall, most models perform exceptionally well (above 0.990 for many species), indicating strong detection capabilities. However, the proposed WeedSwin model distinguishes itself through superior precision across different IoU thresholds, as evidenced by its consistently higher mAP_75 scores. This suggests better localization accuracy and more precise bounding box predictions, making it particularly suitable for practical agricultural

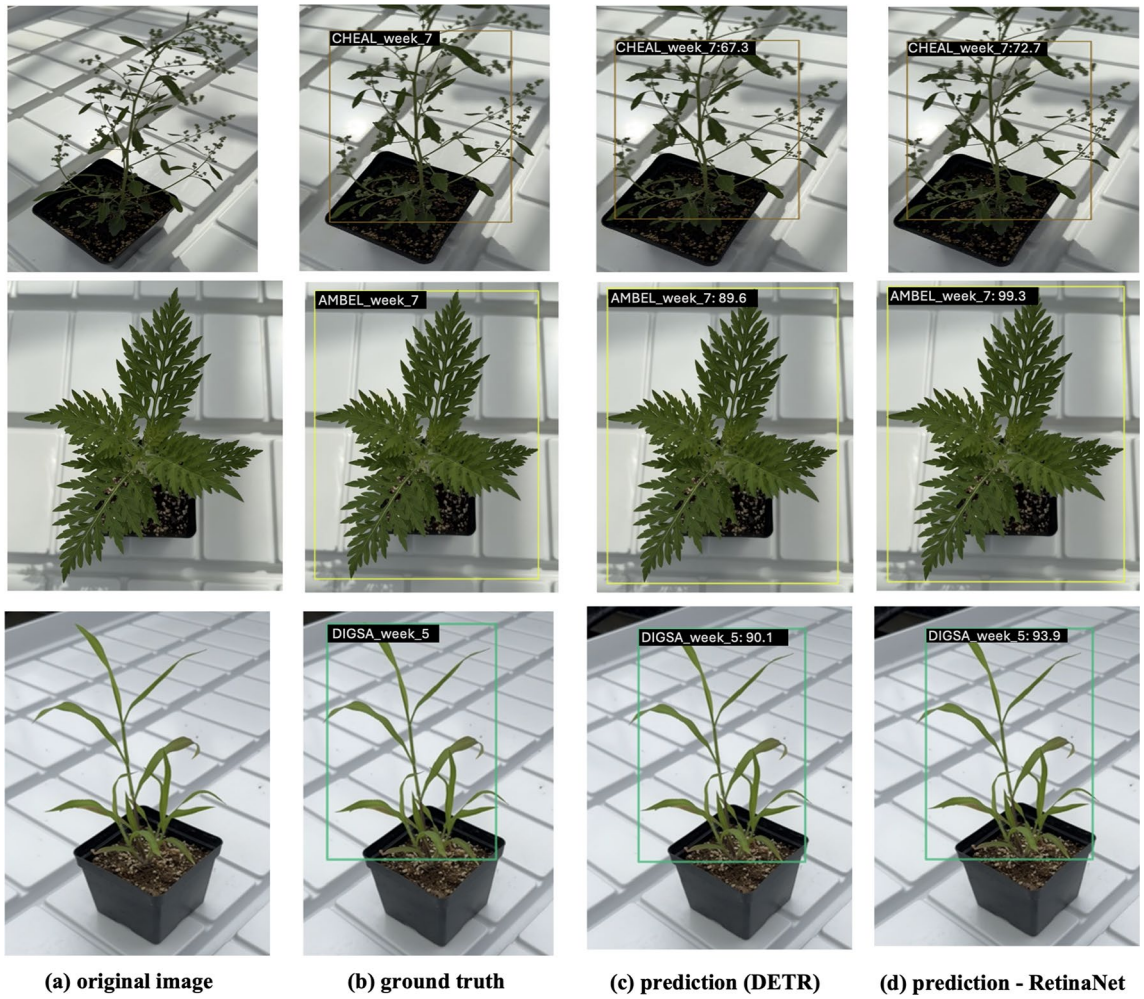

Fig. 5. Comparison of object detection results for CHEAL, AMBEL, and DIGSA using DETR[16] and RetinaNet[19] models for AWD dataset. Row 1 displays predictions for CHEAL, Row 2 displays predictions for AMBEL, and Row 3 displays predictions for DIGSA, with ground truth and model confidence scores indicated for each detection.

Model	Backbone	mAP↑		mAR↑		FPS↑	FLOPs↓	Params↓
		Train	Test	Train	Test			
RetinaNet[19]	ResNext 101[58]	0.978	0.982	0.945	0.947	504.36	150.117G	58.539M
DETR[16]	ResNet 50[14]	0.984	0.984	0.929	0.985	80.65	80.019G	41.557M
DINO[13]	Swin[15]	0.993	0.994	0.981	0.986	152.62	0.241T	0.209G
DINO[13]	ResNet 101[14]	0.991	0.992	0.977	0.979	120.67	0.129T	60.987M
WeedSwin*	Swin[15]	0.992	0.993	0.985	0.985	218.27	0.114T	40.476M
EfficientNet B4[17]	ResNet 50[14]	0.987	0.989	0.948	0.967	490.65	90.0G	24.317M
YOLO v8[18]	CSPNet[67]	0.989	0.990	0.968	0.971	422.48	100.15G	36.12M

Table 6. Performance Comparison of Different Models and Their Backbones on BWD dataset. *WeedSwin is our proposed architecture with modified Swin[15] Transformer-based encoder and decoder with additional layers.

applications where precise weed localization is crucial for targeted treatment. YOLO v8 also demonstrates strong recall capabilities while maintaining competitive precision, indicating its potential as a balanced solution for real-world deployment scenarios.

Table 8 presents the performance analysis of six economically significant weed species that are particularly problematic in US agriculture: AMAPA (palmer amaranth), AMBEL (common ragweed), DIGSA (large crabgrass), SETFA (giant foxtail), CHEAL (common lambsquarters), and AMATU (waterhemp). These species,

	RetinaNet[19]				EfficientNet-B4[17]				DETR[16]				DINO[13]			
	(ResNeXt-101[58])				(ResNet-50[14])				(ResNet-50[14])				(ResNet-50[14])			
Species	mAP	*mAP50*	*mAP75*	*AvgRec*	*mAP*	*mAP50*	*mAP75*	*AvgRec*	*mAP*	*mAP50*	*mAP75*	*AvgRec*	*mAP*	*mAP50*	*mAP75*	*AvgRec*
ABUTH	0.959	0.995	0.982	1.000	0.971	0.994	0.987	1.000	0.944	0.992	0.976	0.998	0.973	0.997	0.990	1.000
AMAPA	0.947	0.994	0.981	1.000	0.972	0.998	0.989	1.000	0.947	0.990	0.969	0.997	0.977	0.997	0.991	0.999
AMARE	0.941	0.980	0.969	1.000	0.954	0.975	0.967	0.997	0.916	0.978	0.947	0.997	0.957	0.989	0.977	0.999
AMATU	0.931	0.987	0.965	1.000	0.949	0.998	0.973	1.000	0.899	0.994	0.951	0.996	0.967	0.996	0.986	0.999
AMBEL	0.964	0.998	0.987	1.000	0.981	0.998	0.995	1.000	0.918	0.997	0.959	0.998	**0.990**	0.999	0.998	1.000
CHEAL	0.932	0.989	0.967	1.000	0.947	0.991	0.977	0.999	0.912	0.989	0.956	0.997	0.957	0.990	0.980	0.998
CYPES	0.947	0.992	0.978	1.000	0.968	0.997	0.987	1.000	0.923	0.993	0.965	0.998	0.982	0.998	0.994	1.000
DIGSA	0.937	0.991	0.972	1.000	0.959	0.995	0.983	0.999	0.915	0.991	0.961	0.997	0.978	0.997	0.991	1.000
ECHCG	0.944	0.993	0.977	1.000	0.965	0.996	0.985	1.000	0.921	0.992	0.963	0.998	0.981	0.997	0.992	1.000
ERICA	0.911	0.996	0.978	1.000	0.943	0.996	0.973	0.999	0.909	0.996	0.970	1.000	0.945	0.996	0.982	0.996
PANDI	0.902	0.980	0.963	1.000	0.953	1.000	0.989	1.000	0.910	0.996	0.964	1.000	0.964	1.000	0.987	1.000
SETFA	0.906	0.973	0.960	1.000	0.945	0.987	0.987	1.000	0.906	0.989	0.969	0.998	0.971	1.000	0.996	1.000
SETPU	0.917	0.955	0.954	1.000	0.960	0.996	0.994	1.000	0.903	0.995	0.969	1.000	0.969	1.000	0.995	1.000
SIDSP	0.969	0.999	0.994	1.000	**0.989**	1.000	1.000	1.000	0.952	0.998	0.989	0.999	**0.990**	1.000	1.000	1.000
SORHA	0.820	0.973	0.871	0.998	0.872	0.985	0.944	0.998	0.809	0.986	0.881	0.998	0.888	0.969	0.937	0.992
SORVU	0.886	0.977	0.933	0.996	0.919	0.971	0.947	0.995	0.859	0.964	0.915	0.989	0.941	0.987	0.970	0.991
	DINO[13]				YOLO v8[18]				WeedSwin							
	(Swin[15])				(CSPNet[67])				(Proposed)							
Species	mAP	*mAP50*	*mAP75*	*AvgRec*	mAP	*mAP50*	*mAP75*	*AvgRec*	mAP	*mAP50*	*mAP75*	*AvgRec*				
ABUTH	0.765	0.935	0.816	0.996	0.985	0.997	0.993	0.999	**0.990**	0.998	0.997	1.000				
AMAPA	0.782	0.935	0.841	0.998	0.986	0.998	0.995	0.998	**0.989**	1.000	0.998	1.000				
AMARE	0.731	0.886	0.778	0.973	0.973	0.987	0.982	0.997	**0.980**	0.990	0.987	1.000				
AMATU	0.747	0.945	0.801	0.997	0.968	0.997	0.989	0.998	**0.975**	1.000	0.999	1.000				
AMBEL	0.831	0.934	0.870	0.987	0.989	0.999	0.996	0.999	0.994	1.000	1.000	1.000				
CHEAL	0.625	0.897	0.677	0.982	0.958	0.994	0.979	0.998	**0.967**	0.990	0.989	1.000				
CYPES	0.743	0.909	0.799	0.985	0.986	0.999	0.996	0.998	**0.991**	1.000	1.000	1.000				
DIGSA	0.685	0.851	0.709	0.997	0.981	0.998	0.994	0.997	**0.987**	1.000	1.000	1.000				
ECHCG	0.755	0.881	0.824	0.993	0.985	0.999	0.995	0.998	**0.989**	1.000	1.000	1.000				
ERICA	0.795	0.955	0.843	0.993	0.952	0.996	0.985	0.995	**0.969**	0.997	0.995	0.996				
PANDI	0.799	0.943	0.868	0.994	0.973	0.998	0.992	0.997	**0.984**	1.000	0.997	1.000				
SETFA	0.793	0.918	0.863	0.995	0.978	0.998	0.992	0.996	**0.986**	1.000	1.000	1.000				
SETPU	0.727	0.899	0.772	0.991	0.971	0.995	0.989	0.995	**0.976**	0.994	0.994	0.997				
SIDSP	0.892	0.962	0.913	0.997	0.991	1.000	0.998	0.999	0.993	1.000	1.000	1.000				
SORHA	0.710	0.826	0.748	0.959	0.904	0.982	0.955	0.988	**0.932**	0.983	0.975	0.990				
SORVU	0.722	0.858	0.755	0.894	0.945	0.984	0.975	0.993	**0.959**	0.985	0.982	0.997				

Table 7. Performance Comparison of algorithms on BWD Test Datasets. Significant values are in bold.

selected from our dataset of 16 weed types, are commonly referred to as "driver weeds" due to their critical role in shaping agricultural management decisions across the United States. Their selection for detailed analysis was based on three key factors: their extensive geographic distribution throughout US farming regions, or their documented resistance to multiple herbicides, and their substantial negative impact on crop yields. These characteristics make them particularly challenging for farmers and agricultural managers, necessitating precise detection and management strategies.

Analysis of detection performance across growth stages (Week 1–11) reveals distinct patterns among the different architectures. RetinaNet exhibits exceptional accuracy, achieving perfect detection (mAP = 1.000) in mature growth stages (Weeks 6-11) for most species. However, it shows varying performance in early-stage detection, particularly for SETFA and CHEAL (mAP = 0.854 and 0.829 respectively in Weeks 1-2). Notably, for AMATU, RetinaNet maintains consistently high performance (mAP > 0.95) even in early stages, suggesting better detection capability for this species. Our proposed WeedSwin architecture demonstrates superior consistency throughout all growth stages and species. It maintains mAP values consistently above 0.95 from Week 6 onwards and, notably, shows stable performance even during early growth stages (Weeks 1-3) where other models typically struggle. This stability is particularly evident in challenging species like DIGSA, SETFA, and AMATU, where competing models show considerable performance variations.

YOLO v8 architecture demonstrates impressive and consistent performance across all species and growth stages. For AMAPA and AMBEL, it achieves mAP values comparable to WeedSwin (ranging from 0.947 to 0.998), with particularly strong performance in later growth stages. For DIGSA and SETFA, YOLO v8 maintains

high accuracy (mAP consistently above 0.92 after Week 1), showing remarkable stability across the growth timeline. For CHEAL, while its performance (mAP ranging from 0.921 to 0.973) is slightly below WeedSwin, it still outperforms several other architectures, particularly in early growth stages. YOLO v8's performance with AMATU is notable, achieving mAP values between 0.945 and 0.984, demonstrating strong detection capabilities for this challenging species throughout its growth cycle.

In contrast, DETR shows interesting variability in performance across species. While it struggles with early growth stages (Weeks 1-5) for most species with mAP values frequently below 0.5, it demonstrates remarkably high accuracy for AMATU throughout all growth stages (mAP > 0.99), suggesting species-specific detection capabilities. The DinoXSwin and DinoXResNet50 architectures show comparable performance patterns, with DinoXSwin maintaining a slight edge across most scenarios, though both models show lower performance for AMATU compared to other species, particularly in DinoXResNet50's case (mAP as low as 0.782 in Week 5).

EfficientNet B4 displays inconsistent performance across different growth stages and species. While it occasionally achieves perfect detection, it shows significant fluctuations, particularly evident in SETFA detection during early weeks (mAP = 0.739 in Week 1) and DIGSA in later stages (mAP = 0.909 in Week 11). For AMATU, it demonstrates strong performance overall but with notable variability (dropping to mAP = 0.909 in Week 4).

While RetinaNet achieves the highest peak accuracy, both WeedSwin and YOLO v8 demonstrate exceptional consistency across growth stages and species, making them particularly suitable for practical agricultural

	AMAPA							AMBEL						
Class Name	RNet	EffNet	DETR	DinoXS	DinoXR	YOLO	WeedS	RNet	EffNet	DETR	DinoXS	DinoXR	YOLO	WeedS
Week_1	0.944	0.957	0.417	0.951	0.946	0.947	0.948	0.989	0.949	0.398	0.927	0.920	0.931	0.932
Week_2	0.948	0.950	0.558	0.945	0.939	0.958	0.953	1.000	0.972	0.532	0.949	0.942	0.957	0.950
Week_3	1.000	0.973	0.396	0.968	0.960	0.972	0.963	1.000	0.993	0.436	0.971	0.966	0.978	0.970
Week_4	1.000	1.000	0.421	0.975	0.970	0.981	0.978	1.000	0.966	0.429	0.983	0.978	0.989	0.981
Week_5	0.991	0.992	0.439	0.985	0.979	0.987	0.980	1.000	0.981	0.475	0.986	0.983	0.992	0.987
Week_6	0.997	0.997	0.846	0.989	0.986	0.992	0.990	1.000	1.000	0.876	0.992	0.990	0.995	0.993
Week_7	1.000	1.000	0.902	0.993	0.992	0.996	0.994	1.000	0.994	0.904	0.997	0.996	0.998	0.998
Week_8	1.000	0.993	0.858	0.987	0.985	0.991	0.988	1.000	0.996	0.889	0.994	0.992	0.997	0.995
Week_9	0.984	0.860	0.915	0.981	0.976	0.989	0.982	0.909	0.971	0.912	0.986	0.984	0.991	0.988
Week_10	1.000	0.976	0.872	0.987	0.985	0.993	0.990	1.000	0.985	0.871	0.995	0.993	0.996	0.997
Week_11	1.000	0.996	0.895	0.992	0.990	0.997	0.995	1.000	0.981	0.892	0.992	0.990	0.998	0.996
	DIGSA							SETFA						
Class Name	RNet	EffNet	DETR	DinoXS	DinoXR	YOLO	WeedS	RNet	EffNet	DETR	DinoXS	DinoXR	YOLO	WeedS
Week_1	0.899	0.881	0.398	0.918	0.911	0.927	0.923	0.854	0.739	0.423	0.927	0.920	0.933	0.932
Week_2	0.975	0.928	0.472	0.956	0.950	0.952	0.960	0.957	0.823	0.521	0.949	0.942	0.949	0.950
Week_3	1.000	0.981	0.489	0.964	0.960	0.968	0.965	0.994	0.921	0.509	0.971	0.966	0.969	0.970
Week_4	0.998	0.993	0.463	0.975	0.970	0.979	0.978	0.907	0.866	0.502	0.983	0.978	0.980	0.981
Week_5	0.999	0.989	0.480	0.980	0.977	0.984	0.983	0.989	0.955	0.498	0.986	0.983	0.986	0.987
Week_6	1.000	0.998	0.900	0.986	0.983	0.989	0.988	0.998	0.929	0.899	0.992	0.990	0.991	0.993
Week_7	1.000	0.992	0.920	0.990	0.989	0.993	0.994	1.000	0.976	0.912	0.997	0.996	0.997	0.998
Week_8	1.000	0.993	0.899	0.988	0.987	0.992	0.990	1.000	0.957	0.898	0.994	0.992	0.994	0.995
Week_9	1.000	0.982	0.911	0.981	0.980	0.988	0.985	1.000	0.989	0.907	0.986	0.984	0.985	0.988
Week_10	1.000	0.929	0.874	0.987	0.986	0.993	0.991	1.000	0.998	0.897	0.995	0.993	0.996	0.997
Week_11	1.000	0.909	0.883	0.992	0.990	0.994	0.995	1.000	0.980	0.888	0.992	0.990	0.993	0.996
	CHEAL							AMATU						
Class Name	RNet	EffNet	DETR	DinoXS	DinoXR	YOLO	WeedS	RNet	EffNet	DETR	DinoXS	DinoXR	YOLO	WeedS
Week_1	0.885	0.887	0.412	0.918	0.912	0.921	0.923	0.951	0.989	0.992	0.935	0.864	0.945	0.999
Week_2	0.829	0.826	0.531	0.907	0.900	0.932	0.915	0.978	0.998	1.000	0.895	0.785	0.953	0.967
Week_3	0.983	0.870	0.498	0.964	0.960	0.949	0.970	0.995	1.000	1.000	0.905	0.791	0.961	0.981
Week_4	1.000	0.857	0.481	0.975	0.970	0.958	0.978	0.998	0.909	0.945	0.962	0.876	0.972	0.999
Week_5	1.000	0.979	0.493	0.980	0.977	0.961	0.983	1.000	0.999	1.000	0.901	0.782	0.975	0.986
Week_6	1.000	1.000	0.895	0.986	0.983	0.967	0.988	0.998	0.986	1.000	0.974	0.880	0.978	0.966
Week_7	1.000	0.983	0.923	0.990	0.989	0.973	0.994	1.000	1.000	1.000	0.954	0.989	0.983	1.000
Week_8	1.000	0.944	0.899	0.988	0.987	0.969	0.990	1.000	1.000	1.000	0.988	0.946	0.979	0.976
Week_9	1.000	0.938	0.911	0.981	0.980	0.963	0.985	1.000	0.997	0.996	0.969	0.878	0.974	0.917
Week_10	1.000	0.965	0.885	0.987	0.986	0.968	0.991	0.983	0.984	0.991	0.999	0.795	0.980	0.952
Week_11	1.000	0.988	0.892	0.992	0.990	0.971	0.995	1.000	1.000	1.000	0.924	0.895	0.984	0.977

Table 8. Weekly Performance Comparison (mAP) for Different Weed Species. RNet: RetinaNet, EffNet: EfficientNet B4, DinoXS: DinoXSwin, DinoXR: DinoXResNet50, YOLO: YOLO v8, WeedS: WeedSwin.

applications. WeedSwin maintains a slight edge in overall performance, especially for challenging species like AMATU, while YOLO v8 offers competitive accuracy with the added benefit of its renowned speed and efficiency. This stability in performance, regardless of plant maturity or species, represents a significant advantage for real-world weed management systems where reliable detection throughout the growing season is crucial.

Ablation study

To better understand the contribution of different architectural components and training strategies to the performance of our WeedSwin model, we conducted a comprehensive ablation study. We systematically modified key aspects of the original model to analyze their impact on detection accuracy, recall, and computational efficiency. Our ablation study examines three aspects: backbone architecture modifications, encoder-decoder enhancements, and training optimization strategies. Table 9 presents a quantitative comparison of these variants against the original WeedSwin model.

Lightweight backbone

In this experiment, we explored the impact of simplifying the backbone architecture to achieve better computational efficiency while maintaining competitive performance. We reduced the depth of the model from [2, 2, 18, 4] to [2, 2, 9, 2], effectively halving the complexity of the third and fourth stages. Additionally, we reduced the window size from 12 to 7 pixels and decreased the drop path rate from 0.2 to 0.1. This configuration yielded a 25% reduction in FLOPs (0.085T compared to the original 0.114T) with only minimal parameter reduction (40.12M vs. 40.48M). While detection performance showed a moderate decrease (mAP: 0.969 vs. 0.993, mAR: 0.958 vs. 0.985), the model maintained 96-97% of the original performance with significantly lower computational requirements, making it suitable for resource-constrained deployment scenarios.

Enhanced encoder-decoder

Our second variant focused on improving feature representation capabilities by enhancing the encoder-decoder architecture. We increased the number of feature levels from 4 to 5, enabling the model to better capture multi-scale features critical for detecting objects of varying sizes. We also deepened the encoder and decoder by increasing their layers from 4 to 6, while reducing the feedforward channel dimensions from 2048 to 1024 to partially mitigate computational overhead. This configuration achieved the highest detection recall (mAR: 0.996) among all variants while maintaining the same precision as the original model (mAP: 0.993). However, these performance improvements came at a significant computational cost, with more than doubled FLOPs (0.250T), substantially increased parameters (269M), and reduced inference speed (80.60 FPS vs. 218.27 FPS). This variant demonstrates the upper bounds of performance achievable through architectural enhancements when computational efficiency is not a primary concern.

Optimized training

The third variant investigated the impact of alternative training strategies without modifying the model architecture. We increased the base learning rate from 0.0001 to 0.0002 and replaced the stepped learning rate schedule with a cosine annealing schedule to provide smoother learning rate decay. We also reduced the weight decay parameter from 0.05 to 0.03 and modified the denoising training configuration by decreasing both label noise scale (0.3 vs. 0.5) and box noise scale (0.6 vs. 1.0). These changes aimed to improve convergence and reduce overfitting. The performance results showed a slight decrease in detection accuracy (mAP: 0.982 vs. 0.993, mAR: 0.980 vs. 0.985) compared to the original model. Interestingly, despite maintaining the same architecture, this variant showed computational characteristics similar to the Enhanced Encoder-Decoder variant, which may be attributed to implementation details of the training optimizations. This experiment highlights the sensitivity of transformer-based detectors to training hyperparameters and the importance of carefully tuning them for optimal performance.

	Model architecture variations						Performance metrics				
Model configuration	Reduced depth	Smaller window	More levels	Deeper E-D	Cosine LR	DN Config	mAP↑	mAR↑	FPS↑	FLOPs (T)↓	Params (M)↓
WeedSwin (Original)							**0.993**	0.985	**218.27**	**0.114**	**40.48**
Lightweight Backbone	✓	✓					0.969	0.958	210.51	0.085	40.12
Enhanced Encoder-Decoder			✓	✓			**0.993**	**0.996**	80.60	0.250	269.00
Optimized Training					✓	✓	0.982	0.980	88.14	0.254	265.00

Table 9. Ablation study on WeedSwin model architecture. The Lightweight Backbone variant focuses on reduced computational complexity with shallower transformer depths [2,2,9,2] and smaller window size (7), achieving 25% FLOPs reduction with minimal performance impact. The Enhanced Encoder-Decoder variant improves feature representation through increased feature levels (5) and deeper encoder-decoder (6 layers), achieving highest detection recall at cost of speed. The Optimized Training variant explores alternative optimization strategies with cosine scheduling and modified denoising parameters, showing balance between accuracy and computational requirements. Significant values are in bold.

Discussion

This study presents a comprehensive evaluation of state-of-the-art object detection algorithms in agricultural weed detection, utilizing extensive datasets that encompass both Alpha Weeds Dataset (AWD) and Beta Weeds Dataset (BWD) scenarios. The research analyzes hundreds of thousands of images across sixteen weed species and eleven growth stages, providing deep insights into the performance characteristics of various detection architectures. Among the evaluated models, our proposed WeedSwin architecture—which incorporates a modified Swin Transformer-based encoder-decoder framework with additional specialized layers—consistently demonstrates superior performance. The architecture achieves an optimal balance between detection accuracy, recall, and inference speed, addressing the critical requirements of precision agriculture where both accuracy and real-time processing capabilities are essential. These findings underscore the significance of developing robust and efficient weed detection models, particularly in agricultural applications where precise identification and rapid response times can significantly impact crop management outcomes. The results not only validate the effectiveness of our proposed approach but also provide valuable insights into the relative strengths and limitations of current object detection methodologies in agricultural contexts.

One of the most striking outcomes emerges from comparing WeedSwin to other prominent models, including RetinaNet[19], DETR[16], EfficientNet B4[17], DINO[13] variants, and YOLO v8[18]. Table 10 summarizes the performance of these models on both the AWD and BWD datasets. This table clearly indicates that models trained on the BWD dataset—which was preprocessed using SAM-2—demonstrated achieved higher performance (in both mAP and mAR) compared to those trained on the AWD dataset, which used traditional preprocessing methods. While RetinaNet[19] and EfficientNet B4[17] achieve exceptionally high Frames Per Second (FPS) values (504.36 and 490.65 FPS respectively), they exhibit slightly lower mAP and mAR scores compared to WeedSwin. The newly added YOLO v8[18] delivers impressive speed (422.48 FPS) while maintaining competitive accuracy (0.972 mAP on AWD test and 0.990 on BWD test), though its recall performance on the AWD dataset (0.899 mAR) is lower than some other models. Conversely, models like DINO[13] with a Swin[15] backbone reach marginally higher mAP scores (0.994 on BWD test) but at a reduced inference speed (152.62 FPS). DETR[16], while showing respectable accuracy (0.984 mAP on the BWD test), is considerably slower (80.65 FPS), making it less suitable for real-time field operations.

Although YOLO v8[18] presents a compelling balance of speed and accuracy, WeedSwin's performance still stands out due to its specialized design. While YOLO v8 offers faster inference (422.48 FPS compared to WeedSwin's 218.27 FPS), WeedSwin maintains both superior mAP scores on the AWD dataset (0.981 vs. 0.972 on test) and notably higher mAR values (0.898 vs. 0.899 for test but 0.911 vs. 0.890 for train), demonstrating more consistent recall capabilities. This indicates that WeedSwin's architecture is particularly effective for weed detection tasks, especially in the more challenging AWD dataset scenarios where comprehensive detection is crucial. In real-world agricultural applications, both models represent viable options, with YOLO v8 potentially preferred when processing speed is paramount, and WeedSwin favored when balanced accuracy and recall across complex weed detection scenarios are required. Figure 8 shows the performance comparison of various object detection models on AWD and BWD datasets.

The strength of WeedSwin is also reflected in the species-wise and stage-wise analyses. For instance, the evaluation across 16 weed species on the BWD dataset shows that WeedSwin achieves consistently high mAP and mAR values. Where some architectures struggle with specific "driver weed" species—those that are herbicide-resistant, widely distributed, and/or particularly damaging to crop yields—WeedSwin maintains exceptional detection capability. The model's steady performance is evidenced even at higher IoU thresholds (mAP_75), confirming its ability to localize weeds precisely within bounding boxes. An especially relevant finding is WeedSwin's robust performance across all growth stages. Early detection of weeds, when plants are just emerging, is crucial because timely intervention can prevent yield losses and increase herbicide efficacy. Many models perform well once weeds have fully matured, but they stumble during initial growth stages when morphological distinctions are subtler. RetinaNet[19] and EfficientNet B4[17], for example, excel at detecting mature weeds in later weeks but are less reliable during Weeks 1-3, as evidenced by their reduced mAP scores for species like SETFA and CHEAL at early stages. In contrast, WeedSwin demonstrates a remarkable consistency from

Model	Backbone	AWD Dataset (203,567 images)[52]				BWD Dataset (120,341 images)				FPS↑
		mAP↑		mAR↑		mAP↑		mAR↑		
		Train	*Test*	*Train*	*Test*	*Train*	*Test*	*Train*	*Test*	
RetinaNet[19]	ResNext 101[58]	0.907	0.904	0.997	**0.989**	0.978	0.982	0.945	0.947	**504.36**
EfficientNet B4[17]	ResNet 50[14]	0.940	0.931	0.892	0.865	0.987	0.989	0.948	0.967	490.65
DETR[16]	ResNet 50[14]	0.854	0.840	0.941	0.936	0.984	0.984	0.929	**0.985**	80.65
DINO[13]	ResNet 101[14]	0.896	0.872	0.888	0.844	0.991	0.992	0.977	0.979	120.67
DINO[13]	Swin[15]	0.973	0.966	0.891	0.882	**0.993**	**0.994**	0.981	**0.986**	152.62
YOLO v8[18]	CSPNet[67]	0.975	0.972	0.890	0.899	0.989	0.990	0.968	0.971	422.48
WeedSwin (Ours)	Swin[15]	**0.988**	**0.981**	0.911	0.898	**0.992**	**0.993**	**0.985**	**0.985**	218.27

Table 10. Performance Comparison of Different Models on AWD[52] and BWD Datasets. Best results are highlighted in bold. WeedSwin (our proposed model) shows consistent performance across both datasets while maintaining competitive inference speed.

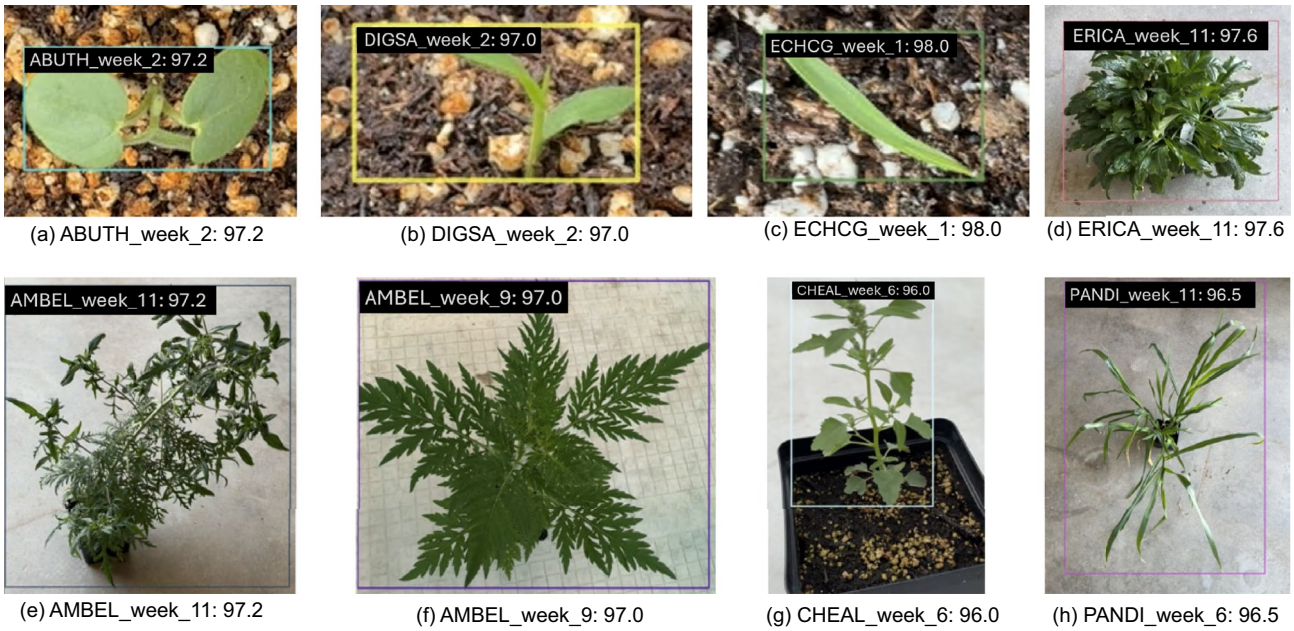

Fig. 7. Detection results of WeedSwin model across various weed species and growth stages. The images illustrate WeedSwin's capability to detect weeds such as ABUTH, DIGSA, ECHCG, ERICA, AMBEL, CHEAL, and PANDI with confidence scores annotated for each prediction.

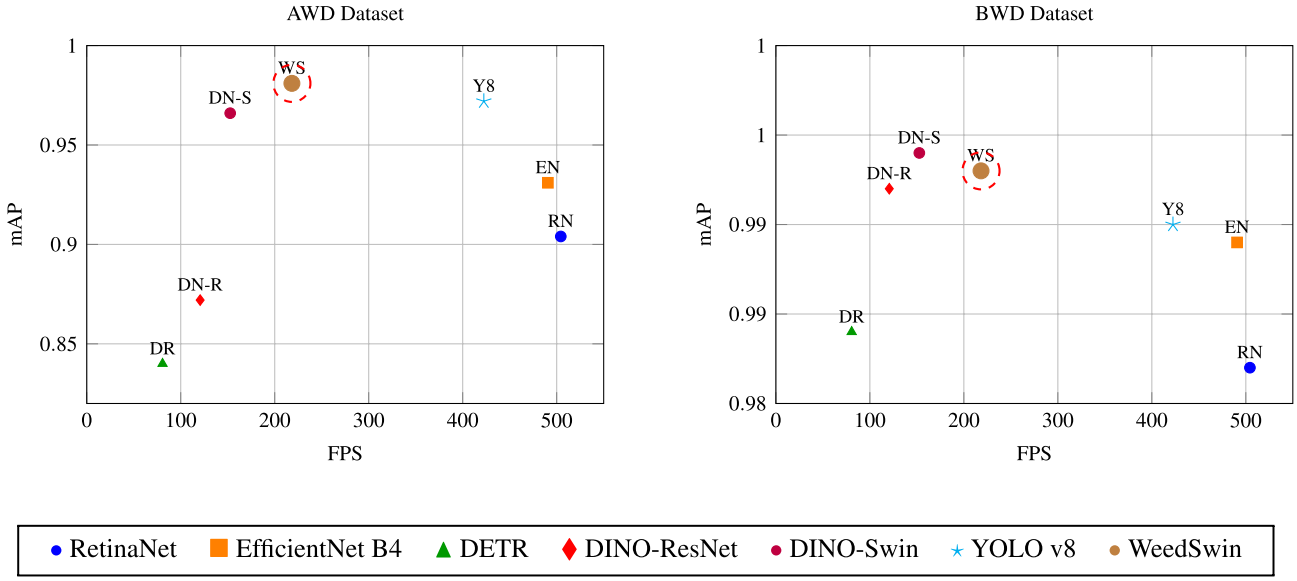

Fig. 8. Comparison of model performance (mAP vs. FPS) on the AWD and BWD Datasets.

Week 1 through Week 11, indicating that the model effectively captures subtle features that distinguish juvenile weeds from soil, residue, or young crops.

Figure 7 demonstrates WeedSwin's robust detection capabilities across various weed species at different growth stages, showcasing the model's versatility and accuracy. In early growth stages, exemplified by ABUTH (week 2) and DIGSA (week 2), the model achieves impressive confidence scores of 97.2% and 97.0% respectively, despite the challenges of detecting newly emerged plants with minimal distinguishing features. The model's effectiveness in early-stage detection is further validated by its high-confidence identification of ECHCG (week 1) at 98.0%, where the plant presents as a single emerging leaf. This early-stage detection capability is particularly valuable for timely weed management interventions. Equally noteworthy is WeedSwin's performance with mature plants, as evidenced by ERICA (week 11) and AMBEL (weeks 9 and 11) detections, achieving confidence scores of 97.6% and 97.0-97.2% respectively. The model maintains high accuracy despite the increased complexity of mature plant structures, including dense foliage and overlapping leaves. This is particularly evident in the

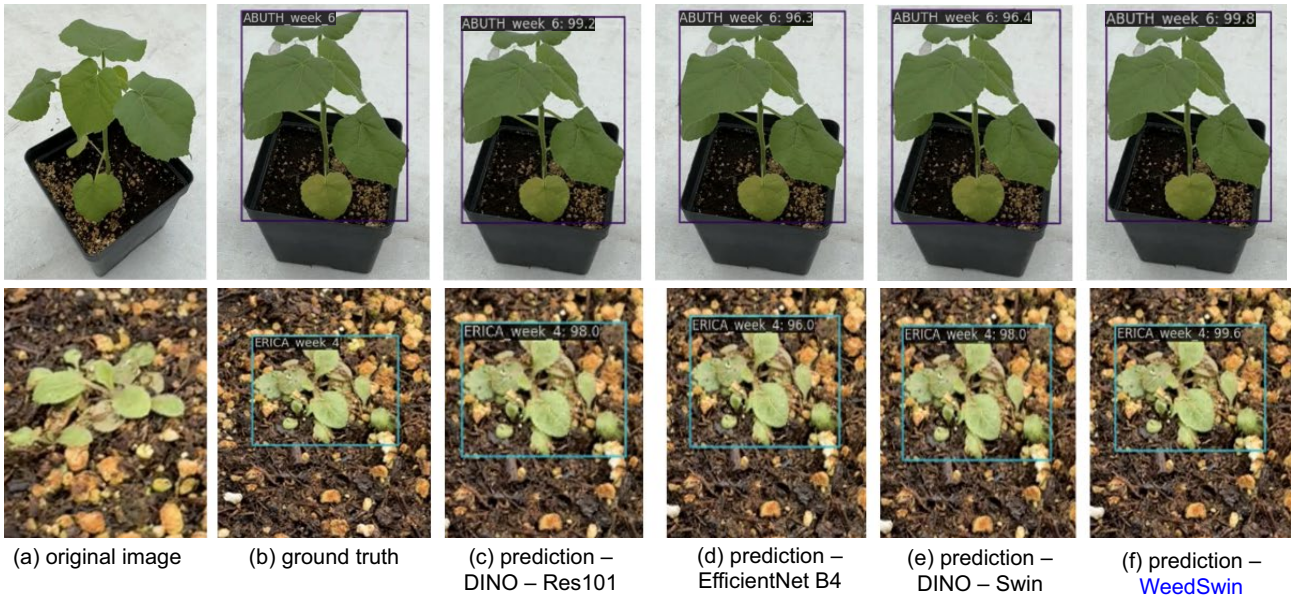

Fig. 6. Comparison of object detection results for ABUTH and ERICA using DINO-Resnet 101, EfficientNet B4, DINO-Swin and Weedswin models for BWD dataset. Row 1 displays the original image, ground truth, and predictions with model confidence scores and bounding box for ABUTH Week 6, and Row 2 displays for ERICA Week 4.

AMBEL week 11 sample, where the model successfully identifies the entire plant structure despite its intricate branching pattern. The model also demonstrates consistent performance with mid-growth stage specimens, such as CHEAL (week 6) and PANDI (week 6), achieving confidence scores of 96.0% and 96.5% respectively. These results highlight WeedSwin's ability to maintain reliable detection accuracy across the entire plant growth cycle, from emergence to maturity, while handling diverse morphological characteristics of different weed species. This consistent performance across growth stages and species makes WeedSwin particularly valuable for practical agricultural applications requiring reliable weed detection throughout the growing season.

The superiority of WeedSwin also extends to challenging species such as AMBEL and DIGSA. WeedSwin records near-perfect detection scores even for these difficult targets. The weekly performance evaluations show that, as weeds advance through their growth stages, WeedSwin's detection performance remains stable and high—mitigating the risk of missing early-stage weeds that could later proliferate into substantial infestations. Figure 6 provides a qualitative illustration of WeedSwin's detection proficiency compared to other models. In these visual comparisons, WeedSwin's predicted bounding boxes for species such as ABUTH and ERICA align closely with the ground truth annotations. The model's high-confidence detections and precise localization exemplify its potential for real-world application. Whether integrated into an autonomous sprayer or used in a scouting drone's processing pipeline, WeedSwin's accurate bounding boxes can guide more intelligent weed control strategies—applying herbicides only where necessary and minimizing chemical inputs.

From a broader perspective, this research represents a significant advancement in agricultural computer vision through its comprehensive scope and innovative approach. Unlike previous studies that focused on limited species or growth stages, our evaluation encompasses two extensive datasets (AWD and BWD) covering 16 weed species across 11 growth stages, including critical weeds for USA such as AMAPA, AMBEL, DIGSA, SETFA, CHEAL, and AMATU—species notorious for their herbicide resistance and agricultural impact. The comparative analysis highlights two standout architectures: WeedSwin and YOLO v8[18]. WeedSwin distinguishes itself through its purposeful design for agricultural challenges, moving beyond simple adaptations of general object detection models. Its innovative integration of a Swin[15] Transformer-based encoder/decoder framework with specialized layers effectively addresses the complex variations in weed morphology, maintaining superior recall metrics on the challenging AWD dataset. Meanwhile, YOLO v8[18] demonstrates remarkable efficiency (422.48 FPS) while achieving competitive mAP scores (0.972/0.990 on AWD/BWD test sets), though with somewhat lower recall performance on AWD data (0.890/0.899 train/test). When compared to other architectures like DINO[13], EfficientNet B4[17], RetinaNet[19], and DETR[16], these two models effectively address the traditional speed-accuracy trade-off from complementary angles—WeedSwin emphasizing detection reliability across diverse weed morphologies with balanced metrics, and YOLO v8 prioritizing processing speed while maintaining competitive precision. This dual advancement provides agricultural practitioners with options tailored to specific operational needs: YOLO v8 for high-throughput scenarios where detection speed is critical, and WeedSwin for applications requiring balanced precision and recall across complex weed variations. Together, these architectures represent significant progress toward automated, real-time weed management systems capable of operating effectively throughout the growing season.

Despite these promising results, it's important to acknowledge certain limitations of this study. The current evaluation, conducted under controlled greenhouse conditions, may not fully represent the challenges faced in real-world agricultural settings, such as varying weather conditions, extreme lighting variations, and complex weed-crop interactions. Additionally, while WeedSwin's 218.27 FPS performance is sufficient for many applications, deployment on resource-constrained edge devices in the field may require further optimization.

Conclusions

This research presents a comprehensive advancement in agricultural weed detection through the development and evaluation of novel deep learning architectures across extensive temporal datasets. By introducing two meticulously curated datasets—AWD and BWD—encompassing 16 weed species through 11 growth stages, this study establishes a robust foundation for automated weed detection systems. The integration of advanced preprocessing techniques, including SAM-2 and traditional computer vision methods, coupled with state-of-the-art deep learning architectures, particularly our proposed WeedSwin Transformer, demonstrates significant improvements in detection accuracy and processing efficiency. The research successfully addresses critical challenges in precision agriculture while providing practical solutions for real-world implementation. Key contributions of this study include:

- Development of comprehensive AWD and BWD datasets, featuring complete growth cycles of economically significant weed species, providing an invaluable resource for agricultural computer vision research
- Introduction of the WeedSwin architecture, achieving superior performance metrics (mAP: 0.993, mAR: 0.985) while maintaining practical processing speeds (218.27 FPS)
- Implementation of innovative preprocessing techniques combining SAM-2 and traditional methods, ensuring high-quality annotations and robust model training
- Demonstration of consistent detection performance across growth stages and species, particularly for challenging driver weeds" that significantly impact U.S. agriculture

While this research represents a significant step forward in automated weed detection, several areas warrant further investigation. The controlled greenhouse environment of the current datasets presents limitations in terms of geographical scope, environmental variability, and real-world applicability under conditions such as varying illumination, occlusion, and complex weed-crop interactions, suggesting the need for expansion to include field conditions and greater environmental variability. Our future research will directly address these limitations through systematic field validation of WeedSwin across diverse agricultural environments. We are currently working on multi-season field data under varying lighting conditions (morning, noon, evening), different soil backgrounds (sandy, clay, loam), and natural weed densities. This extended dataset will feature challenging scenarios including partial occlusion, shadow effects, and mixed-species patches that better represent real-world farming conditions. Nevertheless, this study's comprehensive approach and robust findings establish a strong foundation for advancing precision agriculture. The demonstrated capabilities of the WeedSwin architecture, combined with the extensive temporal datasets, provide a pathway toward more sustainable farming practices through reduced herbicide use and improved crop management efficiency. These contributions collectively support the ongoing evolution of precision agriculture, balancing environmental sustainability with agricultural productivity.

Data availability

The dataset and associated code generated during the current study are not publicly available due to grant requirements, but are available from the corresponding author on reasonable request.

Received: 14 February 2025; Accepted: 30 May 2025
Published online: 02 July 2025

References

1. Chandel, N. S., Chakraborty, S. K., Jat, D. & Chouhan, P. Smart farming management system: Pre and post-production interventions. in *Artificial Intelligence Techniques in Smart Agriculture*, 67–82 (Springer, 2024).
2. Rehman, M. U. et al. Advanced drone-based weed detection using feature-enriched deep learning approach. *Knowl.-Based Syst.* **305**, 112655 (2024).
3. Sapkota, R., Stenger, J., Ostlie, M. & Flores, P. Towards reducing chemical usage for weed control in agriculture using UAS imagery analysis and computer vision techniques. *Sci. Rep.* **13**, 6548 (2023).
4. Murad, N. Y. et al. Weed detection using deep learning: A systematic literature review. *Sensors* **23**, 3670 (2023).
5. Gage, K. L., Krausz, R. F. & Walters, S. A. Emerging challenges for weed management in herbicide-resistant crops. *Agriculture* **9**, 180 (2019).
6. Nath, C. P. et al. Challenges and alternatives of herbicide-based weed management. *Agronomy* **14**, 126 (2024).
7. Gao, W.-T. & Su, W.-H. Weed management methods for herbaceous field crops: A review. *Agronomy* **14**, 486 (2024).
8. Gao, H. et al. An accurate semantic segmentation model for bean seedlings and weeds identification based on improved erfnet. *Sci. Rep.* **14**, 12288 (2024).
9. Wang, A., Xu, Y., Wei, X. & Cui, B. Semantic segmentation of crop and weed using an encoder-decoder network and image enhancement method under uncontrolled outdoor illumination. *IEEE Access* **8**, 81724–81734 (2020).
10. Wu, Z., Chen, Y., Zhao, B., Kang, X. & Ding, Y. Review of weed detection methods based on computer vision. *Sensors* **21**, 3647 (2021).
11. Almalky, A. M. & Ahmed, K. R. Deep learning for detecting and classifying the growth stages of consolida regalis weeds on fields. *Agronomy* **13**, 934 (2023).
12. Teimouri, N. et al. Weed growth stage estimator using deep convolutional neural networks. *Sensors* **18**, 1580 (2018).
13. Zhang, H. *et al.* Dino: Detr with improved denoising anchor boxes for end-to-end object detection. arXiv preprint arXiv:2203.03605 (2022).

14. Targ, S., Almeida, D. & Lyman, K. Resnet in resnet: Generalizing residual architectures. arXiv preprint arXiv:1603.08029 (2016).
15. Liu, Z. *et al.* Swin transformer: Hierarchical vision transformer using shifted windows. in *Proceedings of the IEEE/CVF international conference on computer vision*, pp 10012–10022 (2021).
16. Carion, N. *et al.* End-to-end object detection with transformers. in *European conference on computer vision*, pp 213–229 (Springer, 2020).
17. Tan, M. & Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. in *International conference on machine learning*, pp 6105–6114 (PMLR, 2019).
18. Hussain, M. Yolov1 to v8: Unveiling each variant-a comprehensive review of yolo. *IEEE access* **12**, 42816–42833 (2024).
19. Li, Y. & Ren, F. Light-weight retinanet for object detection. arXiv preprint arXiv:1905.10011 (2019).
20. Pacal, I. et al. A systematic review of deep learning techniques for plant diseases. *Artif. Intell. Rev.* **57**, 304 (2024).
21. Hussain, N. et al. Application of deep learning to detect lamb's quarters (chenopodium album l.) in potato fields of atlantic canada. *Comput. Electr. Agricult.* **182**, 106040 (2021).
22. Peteinatos, G. G., Reichel, P., Karouta, J., Andújar, D. & Gerhards, R. Weed identification in maize, sunflower, and potatoes with the aid of convolutional neural networks. *Remote Sens.* **12**, 4185 (2020).
23. Li, W. & Zhang, Y. Dc-yolo: an improved field plant detection algorithm based on yolov7-tiny. *Sci. Rep.* **14**, 26430 (2024).
24. Pacal, I. Enhancing crop productivity and sustainability through disease identification in maize leaves: Exploiting a large dataset with an advanced vision transformer model. *Expert Syst. Appl.* **238**, 122099 (2024).
25. Kunduracioglu, I. & Pacal, I. Advancements in deep learning for accurate classification of grape leaves and diagnosis of grape diseases. *J. Plant Dis. Prot.* **131**, 1061–1080 (2024).
26. Pacal, I. & Işık, G. Utilizing convolutional neural networks and vision transformers for precise corn leaf disease identification. *Neural Comput. Appl.* **37**, 2479–2496 (2025).
27. Genze, N. et al. Manually annotated and curated dataset of diverse weed species in maize and sorghum for computer vision. *Sci. Data* **11**, 109 (2024).
28. Olsen, A. et al. Deepweeds: A multiclass weed species image dataset for deep learning. *Sci. Rep.* **9**, 2058 (2019).
29. Dyrmann, M., Karstoft, H. & Midtiby, H. S. Plant species classification using deep convolutional neural network. *Biosys. Eng.* **151**, 72–80. https://doi.org/10.1016/j.biosystemseng.2016.08.024 (2016).
30. Sapkota, B. B. et al. Use of synthetic images for training a deep learning model for weed detection and biomass estimation in cotton. *Sci. Rep.* **12**, 19580 (2022).
31. Kovačević, V., Pejak, B. & Marko, O. Enhancing machine learning crop classification models through sam-based field delineation based on satellite imagery. in *2024 12th International Conference on Agro-Geoinformatics (Agro-Geoinformatics)*, 1–4 (IEEE, 2024).
32. Carraro, A., Sozzi, M. & Marinello, F. The segment anything model (SAM) for accelerating the smart farming revolution. *Smart Agricult. Technol.* **6**, 100367 (2023).
33. Xu, K. et al. Multi-modal deep learning for weeds detection in wheat field based on rgb-d images. *Front. Plant Sci.* **12**, 732968 (2021).
34. Lottes, P., Behley, J., Milioto, A. & Stachniss, C. Fully convolutional networks with sequential information for robust crop and weed detection in precision farming. *IEEE Robot. Autom. Lett.* **3**, 2870–2877 (2018).
35. Islam, N. et al. Early weed detection using image processing and machine learning techniques in an Australian chilli farm. *Agriculture* **11**, 387 (2021).
36. Beeharry, Y. & Bassoo, V. Performance of ann and alexnet for weed detection using uav-based images. in *2020 3rd International Conference on Emerging Trends in Electrical, Electronic and Communications Engineering (ELECOM)*, 163–167 (IEEE, 2020).
37. Farooq, A., Hu, J. & Jia, X. Analysis of spectral bands and spatial resolutions for weed classification via deep convolutional neural network. *IEEE Geosci. Remote Sens. Lett.* **16**, 183–187 (2018).
38. Jeon, H. Y., Tian, L. F. & Zhu, H. Robust crop and weed segmentation under uncontrolled outdoor illumination. *Sensors* **11**, 6270–6283 (2011).
39. Arun, R. A., Umamaheswari, S. & Jain, A. V. Reduced u-net architecture for classifying crop and weed using pixel-wise segmentation. In *2020 IEEE International Conference for Innovation in Technology (INOCON)*, 1–6 (IEEE, 2020).
40. Ukaegbu, U. F., Tartibu, L. K., Okwu, M. O. & Olayode, I. O. Development of a light-weight unmanned aerial vehicle for precision agriculture. *Sensors* **21**, 4417 (2021).
41. Naik, N. S. & Chaubey, H. K. Weed detection and classification in sesame crops using region-based convolution neural networks. *Neural Comput. Appl.* **36**, 18961–18977 (2024).
42. Hasan, A. M., Diepeveen, D., Laga, H., Jones, M. G. & Sohel, F. Object-level benchmark for deep learning-based detection and classification of weed species. *Crop Prot.* **177**, 106561 (2024).
43. Moldvai, L., Mesterházi, P. Á., Teschner, G. & Nyéki, A. Weed detection and classification with computer vision using a limited image dataset. *Appl. Sci.* **14**, 4839 (2024).
44. Costello, B. et al. Detection of parthenium weed (parthenium hysterophorus l.) and its growth stages using artificial intelligence. *Agriculture* **12**, 1838 (2022).
45. Subeesh, A. et al. Deep convolutional neural network models for weed detection in polyhouse grown bell peppers. *Artif. Intell. Agricult.* **6**, 47–54 (2022).
46. Guerrero, E., Guerrero, S., Añazco, E. V., Pelaez, E. & Loayza, F. Precision agriculture: Enhancing crops and weeds classification of unbanlance databases via weighted-loss functions. in *2024 IEEE Eighth Ecuador Technical Chapters Meeting (ETCM)*, 1–6 (IEEE, 2024).
47. Liu, T. et al. Semantic segmentation for weed detection in corn. *Pest Manag. Sci.* **81**(2), 1512–1528 (2024).
48. (Ed.), U. M. *BBCH Monograph: Growth Stages of Plants – BBCH Monograph (English version)*. Federal Biological Research Centre for Agriculture and Forestry, Berlin, Germany (2001).
49. Kotleba, J. European and mediterranean plant protection organization (EPPO). *Agrochemia (Slovak Republic)* **34** (1994).
50. Borsch, T. et al. World flora online: Placing taxonomists at the heart of a definitive and comprehensive global resource on the world's plants. *Taxon* **69**, 1311–1341 (2020).
51. Composite List of Weeds - Weed Science Society of America — wssa.net. https://wssa.net/weed/composite-list-of-weeds/. [Accessed 2–13–2025].
52. Islam, T., Sarker, T. T., Ahmed, K. R., Rankrape, C. B. & Gage, K. Weedvision: Multi-stage growth and classification of weeds using detr and retinanet for precision agriculture. in *Accepted in International Conference on Machine Learning and Applications (ICMLA)* (2024).
53. Rodellar, J., Alférez, S., Acevedo, A., Molina, A. & Merino, A. Image processing and machine learning in the morphological analysis of blood cells. *Int. J. Lab. Hematol.* **40**, 46–53 (2018).
54. Ravi, N. *et al.* Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024).
55. Meta segment anything model 2. https://ai.meta.com/sam2/ (2025). Accessed: 2025–04–20.
56. Sengupta, S., Chakrabarty, S. & Soni, R. Is sam 2 better than sam in medical image segmentation? arXiv preprint arXiv:2408.04212 (2024).
57. Ultralytics. Sam 2: Segment anything model 2 - ultralytics yolo docs. https://docs.ultralytics.com/models/sam-2/ (2025). Accessed: 2025–04–20.
58. Tanwar, S. & Singh, J. Resnext50 based convolution neural network-long short term memory model for plant disease classification. *Multim. Tool. Appl.* **82**, 29527–29545 (2023).

59. Luo, Y. et al. Resnext-cc: A novel network based on cross-layer deep-feature fusion for white blood cell classification. *Sci. Rep.* **14**, 18439 (2024).
60. Islam, T., Sarker, T. T., Ahmed, K. R. & Lakhssassi, N. Detection and classification of cannabis seeds using retinanet and faster r-CNN. *Seeds* **3**, 456–478 (2024).
61. Ross, T.-Y. & Dollár, G. Focal loss for dense object detection. in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2980–2988 (2017).
62. Gupta, A. *et al.* Ow-detr: Open-world detection transformer. in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9235–9244 (2022).
63. Koonce, B. & Koonce, B. Resnet 50. *Convolutional neural networks with swift for tensorflow: Image recognition and dataset categorization* 63–72 (2021).
64. Ma, Y. *et al.* Revisiting detr pre-training for object detection. arXiv preprint arXiv:2308.01300 (2023).
65. Davis, J. & Goadrich, M. The relationship between precision-recall and roc curves. in *Proceedings of the 23rd international conference on Machine learning*, 233–240 (2006).
66. Henderson, P. & Ferrari, V. End-to-end training of object class detectors for mean average precision. in *Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part V 13*, 198–213 (Springer, 2017).
67. Wang, C.-Y. *et al.* Cspnet: A new backbone that can enhance learning capability of CNN. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 390–391 (2020).

Acknowledgements
This work was supported by the Illinois Innovation Network (IIN) under Grant 24-21-226695. The authors gratefully acknowledge the support provided by IIN in making this research possible.

Author contributions
Conceptualization, T.I., T.T.S., K.R.A., and K.G.; methodology and experiment, T.I., T.T.S., and C.B.R.; formal analysis, T.I., and K.R.A.; writing-original draft preparation, T.I.; writing-review and editing, T.I., K.R.A., and K.G.; supervision, K.R.A., and K.G. All authors reviewed the manuscript.

Declarations

Competing interests
The authors declare no competing interests.

Ethical approval
This study complies with all relevant institutional, national, and international guidelines for experimental plant research. All seeds used in this investigation were obtained from established laboratory stock maintained by the School of Agricultural Sciences, Southern Illinois University Carbondale, with proper authorization for scientific use. Greenhouse experiments were conducted in accordance with Southern Illinois University Carbondale biosafety protocols and United States regulations governing controlled-environment agricultural research. No field collection of plant material was conducted during this study.

Additional information
Correspondence and requests for materials should be addressed to T.I.