ICPA 17th INTERNATIONAL CONFERENCE ON PRECISION AGRICULTURE
ConBAP CONGRESSO BRASILEIRO DE AGRICULTURA DE PRECISÃO E DIGITAL 2026
13-16 July 2026 | Porto Alegre, Brazil

Enhancing Weed Detection in Corn Crops Through Attention-based Models and Curated Datasets

Thiago Mantovani Martins[a], Everton Castelão Tetila[b], Jayme Garcia Arnal Barbedo[b], Liang Zhao[a], Joaquim Cezar Felipe[a]

[a] Faculty of Philosophy, Sciences and Letters at Ribeirão Preto (FFCLRP), University of São Paulo, Ribeirão Preto, 14040-901, SP, Brazil.

[b] Embrapa Digital Agriculture, Av. Dr. André Tosello, 209, Campinas, 13083-886, SP, Brazil.

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

Abstract.

Weed infestation is one of the leading causes of global agricultural productivity losses, directly impacting production costs, environmental sustainability, and food security. In precision agriculture, automated weed detection from aerial imagery enables site-specific herbicide application, reducing chemical overuse and environmental impact. Deep learning-based computer vision techniques have been widely adopted for this purpose, with Convolutional Neural Networks (CNNs) historically dominating object detection tasks. The publicly available WEED6C dataset was developed in a previous study to detect six plant species in Unmanned Aerial Vehicle (UAV)-acquired images of maize fields and was originally evaluated using convolutional architectures, including Faster R-CNN, YOLOv3, SABL, and FoveaBox. Recently, attention-based deep learning models, particularly Vision Transformers (ViT), have gained relevance for image-based tasks due to their adaptability and performance. This study extends the WEED6C benchmark by evaluating state-of-the-art attention-based object detection architectures for weed detection. In addition, a systematic annotation review was conducted, resulting in a refined version of the dataset (WEED6C V2), in which the total number of labeled objects increased from 9,967 to 15,289, mainly through the addition of previously unlabeled instances, together with the removal of inconsistent annotations and the correction of class labels. RT-DETR and Cascade R-CNN with a Swin Transformer backbone were evaluated, with Faster R-CNN as a convolutional baseline, under a common multilabel-stratified 5-fold cross-validation protocol; global metrics were reported as the macro-average of the per-class results. On the revised dataset, both attention-based architectures outperformed the Faster R-CNN baseline, which achieved an F1-score of 79.60%, an accuracy of 66.91%, and an mAP@0.50 of 76.85%. Cascade R-CNN with Swin Transformer obtained the highest F1-score (82.59%) and accuracy (71.11%), while RT-DETR achieved the highest recall (85.51%) and the highest mean Average Precision (81.13% at mAP@0.50 and 41.04% at mAP@[0.50:0.95]). These results indicate that the best-performing

The authors are solely responsible for the content of this paper, which is not a refereed publication. Citation of this work should state that it is from the Proceedings of the 17th International Conference on Precision Agriculture and 11th Brazilian Congress on Precision and Digital Agriculture. EXAMPLE: Last Name, A. B. & Coauthor, C. D. (2026). Title of paper. In Proceedings of the 17th International Conference on Precision Agriculture and 11th Brazilian Congress on Precision and Digital Agriculture (unpaginated, online). Monticello, IL: International Society of Precision Agriculture and Brazilian Congress on Precision and Digital Agriculture.

architecture depends on the evaluation criterion, and that Transformer-based architectures constitute competitive and promising alternatives for UAV-based weed detection. The findings highlight both the potential of attention mechanisms in agricultural environments characterized by occlusion and illumination variability, and the relevance of systematic dataset curation in manually annotated agricultural imagery. Ongoing work includes evaluation under multiple input resolutions and the integration of slicing-based inference strategies to improve small-object detection. The proposed approach contributes to advancing precision spraying systems by promoting more efficient herbicide use and supporting sustainable crop management practices.

***Keywords**.*

Weed detection, Precision agriculture, Attention-based models, UAV-Imagery, Deep learning

Introduction

Weed infestation is among the most significant biotic constraints on global crop production, competing with cultivated plants for water, nutrients, light, and space, and causing substantial yield and economic losses when left uncontrolled (Toscano et al. 2024; Gharde et al. 2018). In maize in particular, early-season weed competition can severely compromise stand establishment and final productivity (Gharde et al. 2018). Conventional weed control relies largely on the uniform application of herbicides across entire fields, a practice that increases production costs, contributes to the development of herbicide-resistant weed populations, and raises environmental and food-safety concerns (Toscano et al. 2024). Within this context, precision agriculture has emerged as a management paradigm in which spatial and temporal variability within the field is explicitly considered, so that inputs are applied only where and when they are required. Site-specific weed management, in which herbicides are directed to the actual location of weeds, is one of the most promising applications of this paradigm, as it can reduce chemical use while maintaining or improving control effectiveness.

The practical implementation of site-specific weed management depends on the ability to detect and localize weeds accurately across large areas. Unmanned Aerial Vehicles (UAVs) equipped with imaging sensors have become a central tool for this purpose, as they enable the rapid, low-cost, and non-destructive acquisition of high-resolution imagery over agricultural fields (Toscano et al. 2024). Compared with satellite and manned-aircraft platforms, UAVs offer greater operational flexibility, higher spatial resolution, and the capacity to acquire data on demand, which is particularly relevant for capturing small targets such as individual weed plants at early growth stages (Zhang et al. 2025). RGB cameras remain the most widely adopted sensors on these platforms, owing to their low cost, high spatial resolution, and the maturity of the computer-vision methods developed for three-channel imagery, although multispectral sensors have also been increasingly explored to exploit spectral differences between crops and weeds (Zhang et al. 2025). The large volume of imagery generated by UAV surveys, however, makes manual interpretation impractical, motivating the development of automated analysis methods.

Deep learning has become the dominant approach for the automated interpretation of agricultural imagery, and object detection, the task of simultaneously localizing objects with bounding boxes and assigning them a class, is particularly well suited to weed detection, since it identifies and counts individual plants or weed clusters directly. Convolutional Neural Networks (CNNs) have historically dominated this task, and detectors such as Faster R-CNN, together with single-stage architectures of the YOLO family, established the methodological standard for weed detection in UAV imagery (Shah and Tembhurne 2023). More recently, attention-based architectures have gained prominence in computer vision (Jamil et al. 2023). The Vision Transformer (ViT) demonstrated that the self-attention mechanism, originally introduced for natural-language processing, could be applied directly to image patches, enabling the modeling of long-range spatial dependencies that convolutional operations capture less naturally (Dosovitskiy et al. 2021; Shah and Tembhurne 2023). In agricultural settings, Vision Transformers have already shown promise for the classification of weeds and crops in high-resolution UAV imagery (Reedha et al. 2021). Despite their growing adoption in general-purpose detection benchmarks, the comparative performance of these attention-based architectures against well-established convolutional baselines remains underexplored in the specific context of UAV-based weed detection.

The WEED6C dataset was introduced in a previous study to support weed detection in UAV-acquired maize fields, comprising images annotated for six plant species (Tetila et al. 2025). In its original release, the benchmark was evaluated with convolutional detectors, including Faster R-CNN, YOLOv3, SABL, and FoveaBox (Tetila et al. 2025), leaving the behavior of attention-based architectures on this benchmark untested. Furthermore, the quality and completeness of manual annotations are known to exert a strong influence on the performance and the fairness of comparisons between detection models, a factor that is frequently overlooked in agricultural datasets assembled under field conditions.

This study extends the WEED6C benchmark by evaluating state-of-the-art attention-based object detection architectures for weed detection in UAV imagery and by contrasting them with a representative convolutional baseline. Two attention-based detectors were assessed, RT-DETR and Cascade R-CNN with a Swin Transformer backbone, while Faster R-CNN was adopted as a convolutional baseline representative of the state of the art in purely convolutional detection. As part of this work, a systematic review of the original annotations was conducted, resulting in a refined version of the dataset, referred to as WEED6C V2, intended to improve annotation quality and dataset representativeness. All models were trained and evaluated under an identical multilabel-stratified five-fold cross-validation protocol, ensuring a fair and consistent comparison. The main contributions of this work are therefore threefold: (i) the systematic revision and refinement of the WEED6C annotations; (ii) the first evaluation of attention-based detectors on this benchmark under a controlled cross-validation protocol; and (iii) a comparative analysis of attention-based and convolutional architectures for UAV-based weed detection in maize fields.

Materials and Methods

Dataset

The publicly available WEED6C dataset was used in this study (Tetila et al. 2025). The imagery was acquired with a UAV flying at an altitude of approximately 10 m over six maize cultivation areas located in the municipalities of Caarapó and Dourados, in the state of Mato Grosso do Sul, Brazil (Tetila et al. 2025). Maize was the main crop in all areas, and any plant other than maize was treated as undesirable, so that the six annotated plant species comprise both weed species and volunteer plants considered as weeds. The dataset comprises 1,391 images, originally acquired at a resolution of 5,472 × 3,648 pixels, and contains annotations for six plant species: *Cardamine bonariensis* (Bittercress), *Digitaria insularis* (Sourgrass), *Cocos nucifera* (Coconut), *Dysphania ambrosioides* (Mastruz), *Glycine max* (Soybean), and *Sorghum halepense* (Johnsongrass). Representative examples of each class are shown in Fig 1.

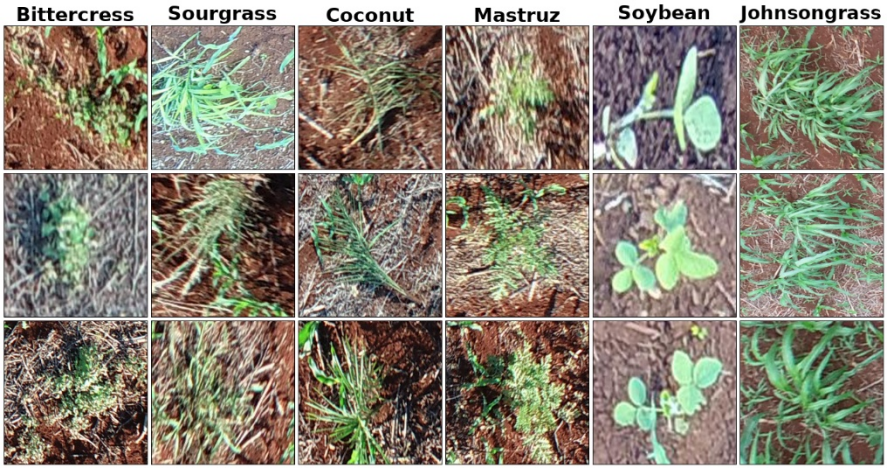

Fig 1. Representative cropped instances of the six annotated classes in WEED6C V2, with three randomly selected examples per class

The original annotations were provided as plain-text files containing, for each object, the class identifier and the bounding-box coordinates (xmin, ymin, xmax, ymax) expressed in absolute pixel values. For compatibility with the detection frameworks adopted in this work, the annotations were converted to the COCO format (Lin et al. 2014), which is widely used in object-detection tasks. An example of a field image containing multiple co-occurring species is shown in Fig 2.

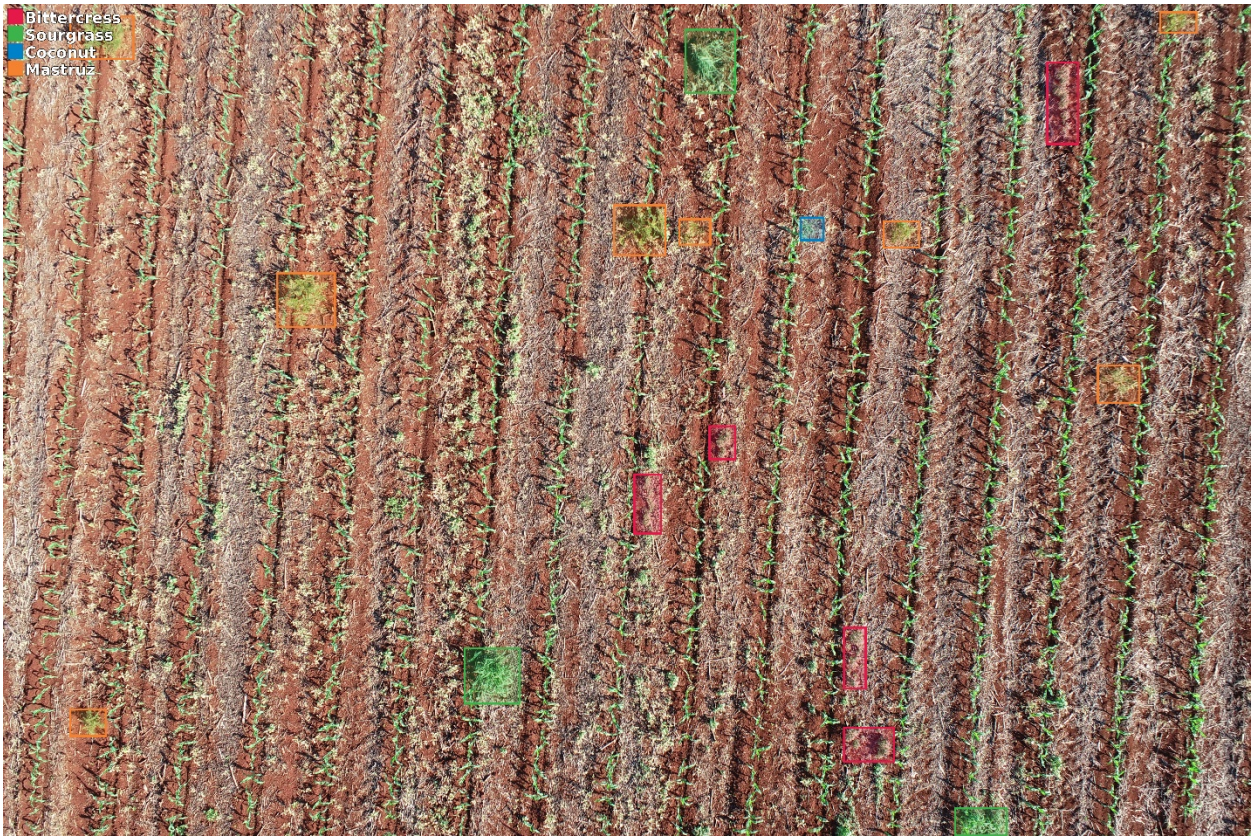

Fig 2. Example of a single UAV image from WEED6C V2 containing multiple co-occurring species, with bounding boxes color-coded by class

Annotation Review and WEED6C V2

Annotation quality is a critical factor in supervised object detection, as missing, imprecise, or mislabeled instances directly affect both model training and the fairness of comparisons between architectures. Accordingly, a systematic review of the original WEED6C annotations was conducted using the Computer Vision Annotation Tool (CVAT). Each image was inspected, and annotations were added, corrected, or removed according to consistent criteria, yielding a refined version of the dataset referred to as WEED6C V2.

The revision increased the total number of labeled objects from 9,967 to 15,289, a net increase of 5,322 objects. Previously unlabeled instances were added, inconsistent annotations were removed, imprecise bounding boxes were geometrically corrected, and a small number of class labels were reassigned. The per-class distribution before and after the revision is presented in Table 1. The original dataset exhibited a pronounced class imbalance, with soybean and Mastruz as the most frequent classes and Johnsongrass and coconut as the least represented. The revision increased the number of objects in every class and attenuated, although it did not eliminate, this imbalance, which motivated the adoption of stratified cross-validation in the experimental protocol. The refinement was intended to improve annotation completeness and consistency, thereby providing a more reliable basis for model training and evaluation.

Table 1. Summary of the annotation review from WEED6C to WEED6C V2

Class	Species	V1	V2
Bittercress	*Cardamine bonariensis*	1,541	3,807
Sourgrass	*Digitaria insularis*	1,454	1,737
Coconut	*Cocos nucifera*	719	1,061
Mastruz	*Dysphania ambrosioides*	2,598	3,658
Soybean	*Glycine max*	3,064	3,765
Johnsongrass	*Sorghum halepense*	591	1,261
Total	-	9,967	15,289

Object Detection Architectures

Three object detection architectures were evaluated in this study, selected to contrast attention-based designs with a representative convolutional baseline.

Faster R-CNN (Ren et al. 2017) was adopted as the convolutional baseline. It is a two-stage detector in which a Region Proposal Network generates candidate regions that are subsequently classified and refined by a detection head. As one of the most widely adopted convolutional detectors, it constitutes a strong and well-established reference for purely convolutional detection. A ResNet-50 backbone with a Feature Pyramid Network was used, initialized with weights pretrained on the COCO dataset.

RT-DETR (Zhao et al. 2024) is a real-time, fully attention-based detector derived from the DETR formulation (Carion et al. 2020), which reformulates object detection as a direct set-prediction problem and removes the need for hand-designed components such as anchor generation and non-maximum suppression. RT-DETR employs an efficient hybrid encoder that decouples intra-scale interaction from cross-scale fusion, together with a Transformer decoder, retaining the end-to-end set-prediction paradigm while substantially improving inference efficiency. The large variant (RT-DETR-L) was used, initialized with weights pretrained on the COCO dataset.

Cascade R-CNN with a Swin Transformer backbone combines the multi-stage Cascade R-CNN detection framework (Cai and Vasconcelos 2018) with the Swin Transformer (Liu et al. 2021) as the feature-extraction backbone. The Swin Transformer applies self-attention within shifted local windows, producing hierarchical feature representations suitable for object detection, while the cascaded detection heads progressively refine predictions under increasing intersection-over-union thresholds. The tiny variant of the backbone (Swin-T) was used, initialized with weights pretrained on the ImageNet-1K dataset.

Experimental Setup

All models were initialized with weights pretrained on large-scale datasets and subsequently fine-tuned on WEED6C V2 through transfer learning. The Faster R-CNN and RT-DETR backbones were initialized with weights pretrained on the COCO dataset, while the Swin Transformer backbone of the Cascade R-CNN was initialized with weights pretrained on ImageNet-1K. For all models, the input images were resized to a maximum size of 1,333 × 800 pixels, a resolution adopted to enable direct comparison with results reported in previous work on the original dataset.

Each architecture was trained using the default configuration and data-augmentation pipeline provided by its respective framework, with only the input resolution standardized across models. All frameworks were built on the PyTorch library (Paszke et al. 2019): Faster R-CNN was implemented in Detectron2 (Wu et al. 2019), RT-DETR in the Ultralytics framework (Jocher and Qiu 2023), and Cascade R-CNN with the Swin Transformer backbone in MMDetection (Chen et al. 2019). Consequently, the number of training epochs, batch size, and optimizer followed each framework's conventions rather than a single shared configuration, as summarized in Table 2. For Faster R-CNN, the default Detectron2 augmentation was applied, consisting of multi-scale resizing of the shorter image side and random horizontal flipping. For Cascade R-CNN with Swin Transformer, the pipeline comprised resizing with aspect-ratio preservation, random horizontal flipping with a probability of 0.5, channel normalization, and padding to a multiple of 32, with mixed-precision training enabled. For RT-DETR, the default Ultralytics augmentation pipeline was used. Under these conditions, the shared protocol concerns the dataset partitioning and the evaluation procedure, described in the following subsection, rather than identical training hyperparameters.

The experiments were conducted on a single workstation equipped with two NVIDIA RTX A4000 GPUs, each with 16 GB of memory (NVIDIA Corporation, Santa Clara, CA, USA), running CUDA 12.4. All models were trained using both GPUs.

Table 2. Training configuration for each evaluated model

Model	Framework	Backbone	Pretraining	Epochs	Effective batch size	Optimizer
Faster R-CNN	Detectron2	ResNet-50-FPN	COCO	30	2	SGD
RT-DETR	Ultralytics	RT-DETR-L	COCO	30	4	AdamW
Cascade R-CNN + Swin	MMDetection	Swin-T	ImageNet-1K	20	1	AdamW

Evaluation Protocol and Metrics

A multilabel-stratified five-fold cross-validation protocol was adopted to ensure a fair and consistent comparison among architectures. The dataset was partitioned into five folds while preserving the distribution of classes across partitions. In each iteration, four folds were used for training and the remaining fold for testing, so that every image was assigned to the test set exactly once over the complete cross-validation cycle.

For each test image, the detections produced by the model were first filtered with a confidence-score threshold of 0.5, and a class-agnostic non-maximum suppression step was applied to remove redundant boxes. Each remaining detection was then matched to the ground-truth annotations, a detection being counted as a true positive (TP) when its intersection over union with a ground-truth box of the same class was at least 0.3; unmatched ground-truth boxes were counted as false negatives (FN) and unmatched detections as false positives (FP). The intersection-over-union threshold of 0.3 follows the original WEED6C study (Tetila et al. 2025), which adopted a lower-than-usual value because the targets are small bounding boxes, for which a stricter threshold would penalize correct detections that are only slightly misaligned; this choice also ensures comparability with the previously reported results. The TP, FP, and FN counts were aggregated by class across the five folds, and per-class precision, recall, F1-score, and accuracy were computed from the aggregated counts, as defined in Equations (1) to (4):

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{1}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \tag{2}$$

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision+Recall}} \tag{3}$$

$$\text{Accuracy} = \frac{\text{TP}}{\text{TP+FP+FN}} \tag{4}$$

The global performance reported for each model was obtained as the macro-average of the per-class metrics, that is, the unweighted arithmetic mean of the six per-class values of each metric. Macro-averaging assigns equal importance to every class regardless of its frequency, which is particularly appropriate for WEED6C V2 given the pronounced class imbalance, as it prevents the most frequent classes from dominating the reported results. As a complementary measure of detection quality, the mean Average Precision (mAP) was also computed following the standard COCO protocol, at an intersection-over-union threshold of 0.50 (mAP@0.50) and averaged over thresholds from 0.50 to 0.95 (mAP@[0.50:0.95]).

Results and Discussion

Comparison Among Architectures

Table 3 presents the performance of the three evaluated architectures on the WEED6C V2 dataset at an input resolution of 1,333 × 800 pixels. Precision, recall, F1-score, and accuracy correspond to the macro-average across the six classes, computed from the true-positive, false-

positive, and false-negative counts aggregated over the five cross-validation folds. The mean Average Precision is reported following the standard COCO protocol.

Table 3. Global performance of the evaluated models on WEED6C V2 (input resolution 1,333 × 800 pixels). Precision, recall, F1-score, and accuracy are macro-averaged across classes

Model	Precision	Recall	F1-score	Accuracy	mAP@0.50	mAP@[0.50:0.95]
Faster R-CNN	78.50	80.88	79.60	66.91	76.85	34.40
RT-DETR	77.52	**85.51**	81.29	69.18	**81.13**	**41.04**
Cascade R-CNN + Swin	**82.63**	82.65	**82.59**	**71.11**	78.29	36.62

Both attention-based architectures outperformed the convolutional baseline across every metric considered. Under the F1-score and accuracy, which were computed at the operating point defined by a confidence threshold of 0.5 and an intersection-over-union threshold of 0.3, the Cascade R-CNN with a Swin Transformer backbone achieved the best results, reaching an F1-score of 82.59% and an accuracy of 71.11%, exceeding the Faster R-CNN baseline by 2.99 and 4.20 percentage points, respectively. This model also attained the highest precision (82.63%) and the most balanced precision-recall trade-off, which is consistent with the progressive refinement performed by the cascaded detection heads.

RT-DETR achieved the highest recall (85.51%) and the second-best F1-score (81.29%), and it led under the mean Average Precision, reaching 41.04% for mAP@[0.50:0.95] and 81.13% for mAP@0.50, against 36.62% and 78.29% for the Cascade R-CNN with Swin Transformer. This advantage was more pronounced at stricter localization thresholds, indicating accurate bounding-box regression across the confidence range. These results show that the relative ranking of the two attention-based detectors depends on the evaluation criterion: the Swin-based model is superior at the fixed operating point used for the F1-score, whereas RT-DETR is superior when performance is integrated over the full range of confidence scores.

Per-Class Analysis

Table 4 reports the per-class F1-score for the three models. The ranking of class difficulty was consistent across architectures: Bittercress and Johnsongrass were the hardest classes for every model, while Sourgrass, Soybean, and Mastruz were the easiest. The difficulty of Bittercress and Johnsongrass is consistent with their being among the less represented or more visually ambiguous classes, and it was the main factor limiting the macro-averaged performance of all models.

Table 4. Per-class F1-score (%) for the evaluated models on WEED6C V2

Model	Bittercress	Sourgrass	Coconut	Mastruz	Soybean	Johnsongrass
Faster R-CNN	69.23	87.56	83.60	86.81	83.97	66.44
RT-DETR	70.18	89.44	84.82	85.84	86.68	70.76
Cascade R-CNN + Swin	**70.32**	**90.75**	**86.24**	**87.08**	**88.54**	**72.64**

Comparison with Previous Work

The original WEED6C study reported, for Faster R-CNN, a precision of 74.04%, a recall of 78.76%, and an F1-score of 76.30%, together with an mAP of 0.37 and an mAP@0.50 of 0.71 (Tetila et al. 2025). In the present work, evaluated on the revised WEED6C V2 dataset under a five-fold cross-validation protocol, Faster R-CNN reached an F1-score of 79.60% and an mAP@0.50 of 76.85%. These figures are reported here only as a reference point, and they are not directly comparable, since the two studies differ in several respects, including the version of the dataset, the data-partitioning and evaluation protocols, the training configuration, and the computational environment. The original study established the WEED6C benchmark and the convolutional baselines on which the present work builds.

Within the present study, the annotation review described in the Materials and Methods section was conducted to increase the completeness and internal consistency of the labels for the specific

purpose of training and comparing the architectures evaluated here, and the resulting WEED6C V2 was used for all experiments. More generally, the performance of supervised detectors is known to be sensitive to the number and quality of the available annotations, which makes systematic dataset curation a relevant component of experimental work with manually annotated agricultural imagery, where occlusion, illumination variability, and the small size of the targets make annotation a demanding task.

Discussion

The results obtained in this first stage support three main observations. First, attention-based detection architectures constituted competitive and, in this experiment, superior alternatives to a strong convolutional baseline for UAV-based weed detection, with consistent gains over Faster R-CNN under every metric considered. This is consistent with the capacity of self-attention to model long-range spatial dependencies, which is advantageous in scenes characterized by visually similar plant species and cluttered backgrounds.

Second, the comparison between the two attention-based detectors did not yield a single dominant model, but rather a trade-off governed by the evaluation criterion and, ultimately, by the intended deployment. The Cascade R-CNN with a Swin Transformer backbone offered the best balance between precision and recall at a fixed operating point, which is desirable when detections are consumed directly at a fixed confidence threshold. RT-DETR, in turn, offered higher recall and superior mean Average Precision, especially under stricter localization, which is advantageous when minimizing missed weeds is a priority or when detections are subsequently filtered or ranked; a comparable behavior of RT-DETR, favoring recall in multiclass weed detection in maize, has also been reported by García-Navarrete et al. (2025). For site-specific weed management, where a false negative may be more costly than a spurious detection, the higher recall of RT-DETR is a relevant property.

Third, the per-class and area-based results indicate that the main limitation common to all models lies in small and visually ambiguous instances. The mean Average Precision restricted to small objects was substantially lower than for medium and large objects across all architectures, which is consistent with the difficulty of detecting individual weed plants at early growth stages in high-resolution aerial imagery. This finding, together with the comparatively low F1-scores observed for bittercress and Johnsongrass, indicates that resolution and small-object handling, rather than the detector family alone, are dominant factors limiting performance. The role of annotation completeness and consistency, ensured here through the curation of WEED6C V2, is an additional factor that supports reliable training and evaluation under these conditions.

These findings should be interpreted in light of the limitations of the present study. The evaluation was restricted to RGB imagery resized to a single input resolution, and the training configurations followed each framework's defaults rather than a fully harmonized setting, so the shared protocol concerns the data partitioning and the evaluation procedure rather than identical hyperparameters. The detection of small and densely distributed targets at the original acquisition resolution, as well as the exploitation of complementary spectral information, were not addressed at this stage and are left for future work.

Conclusion

This study extended the WEED6C benchmark by evaluating attention-based object detection architectures for weed detection in UAV-acquired maize imagery and by contrasting them with a representative convolutional baseline. As part of the work, a systematic review of the annotations was conducted, producing a refined version of the dataset (WEED6C V2) in which the number of labeled objects increased from 9,967 to 15,289, with the aim of improving annotation

completeness and consistency for the training and evaluation of the models considered. All architectures were trained and evaluated under an identical multilabel-stratified five-fold cross-validation protocol, ensuring a fair and consistent comparison.

On the WEED6C V2 dataset, both attention-based detectors outperformed the Faster R-CNN baseline across all the metrics considered. The Cascade R-CNN with a Swin Transformer backbone obtained the highest F1-score (82.59%) and accuracy (71.11%), while RT-DETR achieved the highest recall (85.51%) and the highest mean Average Precision (81.13% at mAP@0.50 and 41.04% at mAP@[0.50:0.95]). The best-performing architecture therefore depended on the evaluation criterion, with the Swin-based model favored at a fixed operating point and RT-DETR favored when performance was integrated over the full range of confidence scores. Taken together, these results indicate that attention-based architectures constitute competitive and promising alternatives for UAV-based weed detection.

The analysis also indicated that small and visually ambiguous instances were the main factor limiting performance across all architectures, as reflected in the lower average precision obtained for small objects and in the comparatively low scores for the least represented classes. Future work includes the evaluation of the architectures under multiple input resolutions and the integration of slicing-based inference strategies to improve the detection of small and densely distributed targets at the original acquisition resolution. The incorporation of complementary spectral information beyond the RGB domain is also envisaged, with the aim of further improving discrimination between visually similar plant species. By advancing accurate and efficient weed detection from UAV imagery, the proposed approach contributes to the development of precision spraying systems that promote more rational herbicide use and support sustainable crop management practices.

References

Cai, Z., and Vasconcelos, N. (2018). Cascade R-CNN: delving into high quality object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 6154–6162).

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. (2020). End-to-end object detection with transformers. In *Computer Vision – ECCV 2020* (Lecture Notes in Computer Science, Vol. 12346, pp. 213–229). Cham: Springer.

Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., et al. (2019). MMDetection: open MMLab detection toolbox and benchmark. *arXiv preprint arXiv*:1906.07155.

CVAT.ai Corporation (2023). Computer Vision Annotation Tool (CVAT). https://github.com/cvat-ai/cvat. Last accessed 13 June 2026.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2021). An image is worth 16×16 words: transformers for image recognition at scale. In *International Conference on Learning Representations*.

García-Navarrete, O. L., Camacho-Tamayo, J. H., Bregón, A. B., Martín-García, J., and Navas-Gracia, L. M. (2025). Performance analysis of real-time detection transformer and you only look once models for weed detection in maize cultivation. *Agronomy*, 15(4), 796.

Gharde, Y., Singh, P. K., Dubey, R. P., and Gupta, P. K. (2018). Assessment of yield and economic losses in agriculture due to weeds in India. *Crop Protection*, 107, 12–18.

Jamil, S., Piran, M. J., and Kwon, O.-J. (2023). A comprehensive survey of transformers for computer vision. *Drones*, 7(5), 287.

Jocher, G., and Qiu, J. (2023). Ultralytics YOLO (Version 8) [Computer software]. https://github.com/ultralytics/ultralytics. Last accessed 13 June 2026.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., et al. (2014). Microsoft COCO: common objects in context. In *Computer Vision – ECCV 2014* (Lecture Notes in Computer Science, Vol. 8693, pp. 740–755). Cham: Springer.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., et al. (2021). Swin Transformer: hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 10012–10022).

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al. (2019). PyTorch: an imperative style, high-performance deep learning library. *In Advances in Neural Information Processing Systems* (Vol. 32, pp. 8024–8035).

Reedha, R., Dericquebourg, E., Canals, R., and Hafiane, A. (2021). Vision transformers for weeds and crops classification of high resolution UAV images. *arXiv preprint* arXiv:2109.02716.

Ren, S., He, K., Girshick, R., and Sun, J. (2017). Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6), 1137–1149.

Shah, S., and Tembhurne, J. (2023). Object detection using convolutional neural networks and transformer-based models: a review. *Journal of Electrical Systems and Information Technology*, 10, 54.

Tetila, E. C., Wirti, G., Higa, G. T. H., da Costa, A. B., Amorim, W. P., Pistori, H., et al. (2025). Deep learning models for detection and recognition of weed species in corn crop. *Crop Protection*, 195, 107237.

Toscano, F., Fiorentino, C., Capece, N., Erra, U., Travascia, D., Scopa, A., et al. (2024). Unmanned aerial vehicle for precision agriculture: a review. *IEEE Access*, 12, 69188–69205.

Wu, Y., Kirillov, A., Massa, F., Lo, W.-Y., and Girshick, R. (2019). Detectron2. https://github.com/facebookresearch/detectron2. Last accessed 13 June 2026.

Zhang, S., Wang, X., Lin, H., Dong, Y., and Qiang, Z. (2025). A review of the application of UAV multispectral remote sensing technology in precision agriculture. *Smart Agricultural Technology*, 12, 101406.

Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., et al. (2024). DETRs beat YOLOs on real-time object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 16965–16974).