Capítulo 7

Cognitive Computing Advancements: Improving Precision Crop Protection through UAV Imagery for Targeted Weed Monitoring

Publicación asociada a este capítulo Weed species classification with UAV imagery and standard CNN models: assessing the frontiers of training and inference phases

- Mesías-Ruiz GA, Peña JM, de Castro AI, Borra-Serrano I, Dorado J. Cognitive Computing Advancements: Improving Precision Crop Protection through UAV Imagery for Targeted Weed Monitoring. *Remote Sensing*. 2024; 16(16):3026. https://doi.org/10.3390/rs16163026

Abstract

Early detection of weeds is crucial to manage weeds effectively, support decision-making and prevent potential crop losses. This research presents an innovative approach to develop a specialized cognitive system for classifying and detecting early-stage weeds at the species level. The primary objective was to create an automated multiclass discrimination system using cognitive computing, regardless of the weed growth stage. Initially, the model was trained and tested on a dataset of 31,002 UAV images, including ten weed species manually identified by experts at the early phenological stages of maize (BBCH14) and tomato (BBCH501). The images were captured at 11 m above ground level. This resulted in a classification accuracy exceeding 99.1 % using the vision transformer Swin-T model. Subsequently, generative modeling was employed for data augmentation, resulting in new classification models based on the Swin-T architecture. These models were evaluated on an unbalanced dataset of 36,556 UAV images captured at later phenological stages (maize BBCH17 and tomato BBCH509), achieving a weighted average *F1-score* ranging from 94.8 % to 95.3 %. This performance highlights the system's adaptability to morphological variations and its robustness in diverse crop scenarios, suggesting that the system can be effectively implemented in real agricultural scenarios, significantly reducing the time and resources required for weed identification. The proposed data augmentation technique also proved to be effective in implementing the detection transformer architecture, significantly improving the generalization capability and enabling accurate detection of weeds at different growth stages. The research represents a significant advancement in weed monitoring across phenological stages, with potential applications in precision agriculture and sustainable crop management. Furthermore, the methodology showcases the versatility of the latest generation models for application in other knowledge domains, facilitating time-efficient model development. Future research could investigate the applicability of the model in different geographical regions and with different types of crops, as well as real-time implementation for continuous field monitoring.

7.1 Introduction

Weeds pose a significant challenge to crop protection by engaging in intense competition with crops for essential resources. This competition leads to a notable reduction in crop yields (Horvath et al., 2023), further complicating agricultural endeavors. Accurate monitoring of weed populations is crucial (Fernández-Quintanilla et al., 2020). Traditional methods of weed identification often rely on manual observation, which can be both time-consuming and error-prone, frequently constrained by human experience (Andújar et al., 2010). Given the similarities in physical traits among numerous weed species, the potential for erroneous identification and subsequent treatment decisions can lead to diminished control-measure effectiveness. Furthermore, owing to the morphological variations among weed species and their growing stages (Wang et al., 2019a), there is a critical demand for a system capable of accurately recognizing and distinguishing these changes.

In the third era of information technology, characterized by the adoption of cognitive computing (CC) systems, a synergistic interaction between humans and technology emerges, aiming at expanding our knowledge base, enhancing the efficient utilization of natural resources and optimizing production processes. CC is based on the idea that computers can learn and simulate human cognitive functions, such as perception, memory and reasoning, using advanced algorithms (Aghav-Palwe y Gunjal,

2021). Additionally, CC can accelerate the analysis of extensive datasets (Sreedevi et al., 2022), enabling the detection of patterns and trends that may not be obvious to human perception. Therefore, CC seeks to understand the relationships between different types of data and real-world occurrences, employing cutting-edge technologies such as artificial intelligence (AI), pattern recognition and machine learning (ML) (Dong et al., 2020).

An inherent challenge in CC is its adaptation to new environments and datasets without extensive training data, thereby minimizing risks associated with bias and overfitting. According to Lytras y Visvizi, 2021, computer vision possesses the potential to tackle environmental challenges and promote sustainable practices by efficiently processing vast amounts of data and emulating human cognitive capabilities. ML, a foundational component of cognitive systems (Remya et al., 2023), allows computers to learn autonomously without the need for explicit programming, discerning patterns through statistical methods. Deep learning (DL), inspired by the workings of the human brain, employs artificial neural networks to perform high-level tasks (LeCun et al., 2015). These neural networks, akin to the brain, enhance their performance by adjusting their connections based on patterns within the data.

Convolutional neural networks (CNNs) are widely used in DL models and provide a fundamental role in several research areas. These architectures are especially relevant in fields such as image processing (Zaidi et al., 2022), natural language processing and emotion recognition (Singh y Prasad, 2023), where they have proven to be instrumental in improving the accuracy and efficiency of automated systems. Using CNN models, data-driven decisions can be made and effective predictions can be generated from new data due to their learning capabilities. The advent of the transformer model (Vaswani et al., 2017) represents a significant innovation in natural language processing and ML, revolutionizing our approach to tasks such as machine translation, text generation and computer vision (Lin et al., 2022). This architecture works similarly to the human brain, using prior knowledge to make decisions in new situations, facilitated by its dynamic attention mechanism. The emergence of the vision transformer (ViT) model (Dosovitskiy et al., 2021) has yielded models comparable to CNNs in the domain of computer vision tasks (Liu et al., 2023; Yang et al., 2022).

Through the application of ML, CC systems can enhance their ability to recognize patterns, understand context and manage the complexities of food production systems by integrating data from multiple sources, thus supporting decision-making (Lonij y Fiot, 2016). CC is revolutionizing modern agriculture through the integration of soft computing techniques such as fuzzy logic, neural networks and fuzzy cognitive maps. These technologies enable more efficient and sustainable management of resources, optimizing crop yields, reducing environmental impact. In addition, these techniques demonstrate a high potential to improve decision-making in real time, adapting to changing conditions and the specific challenges of the agricultural environment (Huang et al., 2010; Mourhir et al., 2017; Munteanu et al., 2021).

Technological advancements in remote sensing have popularized the use of unmanned aerial vehicles (UAVs) in agriculture, particularly for crop protection (Maes y Steppe, 2019), enabling high-resolution data acquisition on crop conditions (Weiss et al., 2020). Nevertheless, the limited availability of data can pose challenges when training CC models (Rejeb et al., 2022). To address this constraint, data augmentation techniques manipulate existing images to generate new ones representing a broader spectrum of real-world diversity. This strategic approach mitigates overfitting and improves model performance and generalization ability by inferring new data (Mumuni y Mumuni,

2022). Among the array of data augmentation techniques, generative adversarial neural networks (GANs) stand out as promising tools for the generation of synthetic data.

The integration of CC in weed detection has significantly improved accuracy, efficiency and cost-effectiveness. Advancements in computer vision and DL models provide robust tools for precision agriculture, ultimately enhancing crop yield and sustainability. Therefore, building upon the progress in CC, with a particular focus on the application of DL models such as GANs and ViTs, the main objective of this research was to develop a robust and automated multiclass discrimination system capable of accurately identifying various weed species, irrespective of their growth stage. The specific objectives of this study are defined as follows:

1. Develop a comprehensive dataset including weed species in their early growth stages, intended for the training, validation and testing of the ViT model's classification performance on both maize and tomato crops.
2. Assess the performance of the previously tested ViT model on the dataset featuring the same weed species but in later growth stages. The hypothesis being that there could be a significant decrease in performance compared to the initial growth-stage findings.
3. Implement GAN-based data augmentation techniques to enhance the spatial resolution of the early growth-stage dataset. Then, create new classification models and evaluate the performance of weed species classification using ViT architecture on the subsequent growth-stage dataset.
4. Evaluate and quantify the impact of increased training data on the performance of an object detection model for weed detection by training three different models using incremental datasets.

7.2 Materials and Methods

This study employed the principles of CC to develop a methodology that enables the creation of a system capable of analyzing complex data, extracting knowledge and utilizing that knowledge for informed decision-making. Our approach involved training a ViT-based model using UAV imagery, enabling the identification of multiple weed species during their early growth stages. Furthermore, we assessed the model's capacity to transfer knowledge by applying it to a dataset depicting a subsequent growth stage (Figure 7.1).

7.2.1 Dataset Generation

The data collection process involved various geographical locations, each characterized by varying crop types and agricultural conditions. The selected sites included an experimental maize farm in Arganda del Rey (Madrid, Spain), affiliated with the Spanish National Research Council (CSIC). The maize planting area covered approximately 7400 m^2, and data were collected at two phenological growth stages (Figure 7.2): early crop growth BBCH14 (4 leaves unfolded) and subsequent crop growth BBCH17 (7 leaves unfolded) (Meier, 2018). In addition, we included two commercial tomato crops in Santa Amalia (Badajoz, Spain), with planting areas of approximately 12,000 m^2 and 14,000 m^2. Data collection for the tomato crops also involved an early growth stage BBCH501

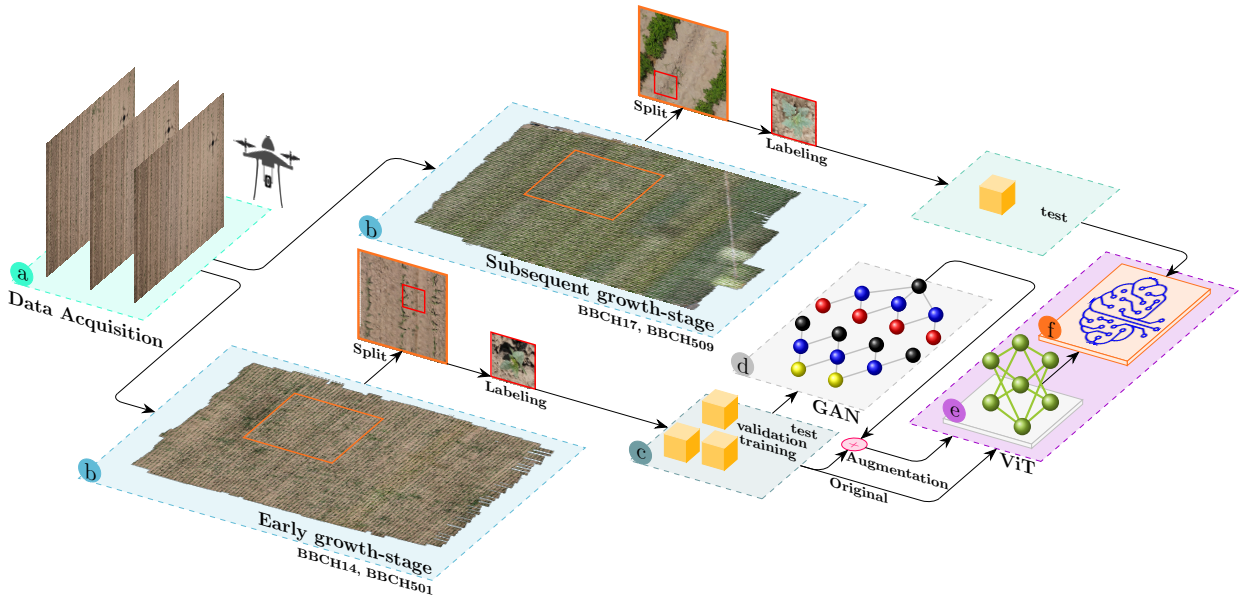
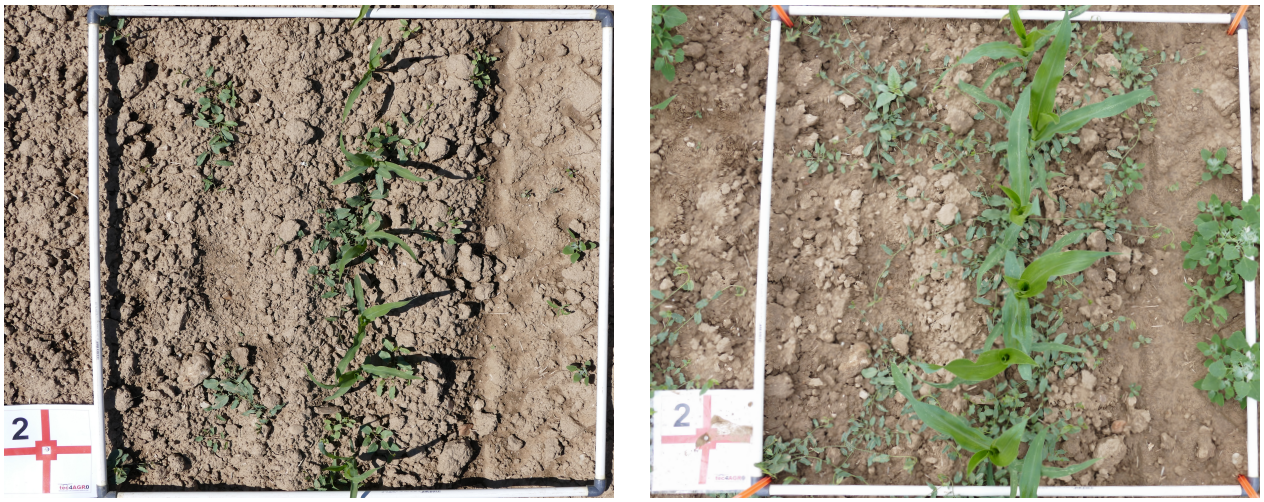


Figura 7.1: Flowchart depicting the research development process, which comprises the following steps: (a) UAV programmed flights at an altitude of 11 m above two crops; (b) orthomosaics building; (c) labeling and categorization of identified species by partitioning the orthomosaics + model building; (d) generation of synthetic images using GANs; (e) implementation of ViT classifiers using the dataset from the initial flights; and (f) assessment of the dataset related to the subsequent crop growth stage for comparative analysis.

(first flower bud visible) and subsequent growth stage BBCH509 (ninth flower bud visible) (Meier, 2018). Selection of these growth stages for this research was not arbitrary, but is optimal for applying effective weed control strategies. Natural weed infestations were observed in the study fields during the data collection process.

A Sony ILCE-6300L RGB visible light camera with an effective resolution of 24.2 megapixels was used, mounted on a UAV model Microdrones MD4-1000. Flying the UAV at an altitude of 11 m resulted in a ground sample distance (GSD) of 0.17 cm per pixel. To ensure comprehensive coverage, a 70 % overlap ratio was maintained both laterally and frontally during image capture. Each acquired image had dimensions of 6000 × 3376 pixels. Maize and tomato crop data were acquired in May 2020 and May 2021, respectively. The UAV images were taken at solar zenith (around 13:00 CET), with cloudless skies and constant lighting conditions. Collecting the images at this time is crucial to minimize the presence of plant shadows in the images, thus improving the quality of the image data and the reliability of the analysis results according to the field conditions.

A total of 568 and 565 images for the BBCH14 and BBCH17 growth stages in maize, respectively, and 895 and 950 images for the BBCH501 and BBCH509 growth stages in tomato, respectively, were used to create the orthomosaics of the entire fields used in the study. Subsequently, fragments of 1000 × 1000 pixels were extracted from these orthomosaics to facilitate the species identification and labeling task carried out by agronomy experts with rectangular boxes using the graphical tool LabelImg (Tzutalin, 2015). Only whole plants were labeled, which means that the plants divided by the image cuts were not considered. Through this process, the following species were identified:



(a) **(b)**

Figura 7.2: Images of the maize crop showing two sampling times: (a) early growth stage BBCH14 (4 leaves unfolded) and (b) subsequent growth stage BBCH17 (7 leaves unfolded). The images presented correspond to terrestrial images.

Atriplex patula, *Chenopodium album*, *Convolvulus arvensis*, *Cyperus rotundus*, *Datura ferox*, *Lolium rigidum*, *Portulaca oleracea*, *Salsola kali*, *Solanum nigrum* and *Sorghum halepense*. The number of labels per species and growth stage is shown in Table 7.1. As a starting point for our study, we allocated 100 labels per species for testing in the early growth stage. Subsequently, the remaining labels were divided into 80 % for training and 20 % for validation (Table 7.1). The dataset generated and used for this research can be found in (Mesías-Ruiz et al., 2024b).

Tabla 7.1: Distribution of labeled images for each weed species and crop in the training, validation and testing sets of the model developed during the early growth stage, along with the number of labeled images from the subsequent growth stage.

	Early growth stage				**Subsequent growth stage**
	Labels	**Train**	**Validation**	**Test**	**Labels**
Atriplex patula	1000	720	180	100	1459
Chenopodium album	1200	880	220	100	2175
Convolvulus arvensis	1200	880	220	100	1102
Cyperus rotundus	3090	2392	598	100	134
Datura ferox	683	466	117	100	589
Lolium rigidum	1000	720	180	100	80
Portulaca oleracea	1875	1420	355	100	177
Salsola kali	1200	880	220	100	1216
Solanum nigrum	1900	1440	360	100	2175
Sorghum halepense	1600	1200	300	100	103
Maize	12,364	9811	2453	100	24,614
Tomato	3890	3032	758	100	2732

7.2.2 Vision Transformer Neural Network for Weed and Crops Classification

ViTs represent an innovative extension of transformer models that apply the concept of attention to learn relationships between different parts of an image. ViTs achieve this by breaking down an image into uniformly sized patches and then converting each patch into a vector using an embedding layer. These patch vectors subsequently undergo a relationship learning process using transformers, which exploit attention mechanisms to understand the interconnections among these vectors. This modular structure allows for greater flexibility and efficiency compared to traditional convolutional architectures (Dosovitskiy et al., 2021).

The classification stage in our study used the Swin-T model (Liu et al., 2021), which was selected for its ability to attain high accuracy in image classification tasks while demanding less computational resources compared to other commonly employed ViT and CNN models. This model was specifically designed to maintain feature map resolutions akin to traditional convolutional networks such as VGG and ResNet. However, instead of conventional convolutional layers, Swin-T uses a hierarchical architecture of transformer blocks with sliding windows for processing image fragments and feature extraction (Liu et al., 2021). This approach enables multiscale modeling while preserving linear computational complexity concerning image size. The architecture of Swin-T consists of four stages (Figure 7.3e): linear embedding, Swin transformer block, patch partitioning and patch merging. Initially, the input image is divided into non-overlapping patches by patch partitioning. Subsequently, the patch merging process within the Swin transformer block combines these patches based on their adjacency in a 2×2 arrangement. Finally, the data stream is repeatedly subjected to patch merging and Swin transformer block operations to effectively process the information.

To explore the interpretability of the Swin-T model, the gradient-weighted class activation mapping (Grad-CAM) technique was used (Selvaraju et al., 2017). This technique facilitated an in-depth understanding of the mechanisms of self-attention and window shifting, and how these contribute to model decision-making. Grad-CAM not only improves interpretability by providing detailed, class-specific activation maps, but also maintains fidelity to the original model. This allows granular inspection of Swin-T model responses without the need to modify its architecture.

7.2.3 Model Training and Inference Details

The model training and inference include:

1. Preprocessing: The PyTorch library was used for label resizing to 224×224 pixels. Additionally, the library enabled various transformations, including random horizontal image flipping with a 50 % probability and normalization based on two sets of values: mean and standard deviation. These values, (0.485, 0.456, 0.406) and (0.229, 0.224, 0.225), correspond to the means and standard deviations of each color channel (red, green, blue), calculated from a reference dataset such as ImageNet (Figure 7.3b).

2. Custom sampler: Due to class imbalance within our dataset, we developed a custom sampler that relies on a calculation involving the number of samples per class in the training dataset and inverse class weights. These weights were assigned inversely proportional to the number of samples in each class, meaning that classes with fewer samples received higher weights, while those with more samples received lower weights. This sampler was used during training

to select batches of data. This approach facilitated the model in giving greater focus to underrepresented classes, mitigating any bias toward the majority classes in the trained model.

3. Hyperparameter optimization: In the cross-validation process, we utilized the AdamW algorithm (Loshchilov y Hutter, 2019) with a learning rate of $6 \times 10^{\breve{}4}$ and a weight decay factor of $1 \times 10^{\breve{}4}$. The chosen loss function was sparse categorical cross-entropy, with batches of size 32 and a total of 50 training epochs.

4. Performance metrics: To evaluate the performance of the classification model, several metrics were used. *Accuracy (ACC)* determines the overall accuracy by calculating the percentage of correctly predicted images in relation to the total number of images. However, it is crucial to acknowledge that accuracy is reliable only when the class distribution is balanced. *Precision (P)* represents the fraction of images correctly labeled as positive by the model. The *Recall (R)* metric measures the proportion of actual positive cases that the model correctly identifies. *F1-score* combines precision and recall into a single value, providing a comprehensive measure of classification performance. It proves particularly useful when managing class imbalances in the dataset.

$$ACC = \frac{\sum_{i=1}^{n} (T_{P_i} + T_{N_i})}{\sum_{i=1}^{n} (T_{P_i} + F_{P_i} + F_{N_i} + T_{N_i})} \tag{7.1}$$

$$P = \frac{\sum_{i=1}^{n} T_{P_i}}{\sum_{i=1}^{n} (T_{P_i} + F_{P_i})} \tag{7.2}$$

$$R = \frac{\sum_{i=1}^{n} T_{P_i}}{\sum_{i=1}^{n} (T_{P_i} + F_{N_i})} \tag{7.3}$$

$$\textit{F1-score} = 2 \cdot \frac{P \cdot R}{P + R} \tag{7.4}$$

7.2.4 Generative Adversarial Neural Network for Image Augmentation

Data augmentation is a technique that effectively enhances both the quantity and diversity of images within a training dataset (Mumuni y Mumuni, 2022). A recent advancement in this domain involves the use of GANs to augment datasets with a wide array of contrasting images (Olaniyi et al., 2022). In our work, we employed the pretrained generative facial prior GAN (GFPGAN) model, adapting it to align with the specific characteristics of our dataset. The GFPGAN is primarily a facial restoration model designed to achieve a balance between realism and fidelity in super-resolution images (Wang et al., 2021). The model consists of two core components (Figure 7.3c): a U-Net module for mitigating degradation and a StyleGAN2 module, which serves as a pretrained facial GAN used as a prior generator. These components are interconnected through a latent code mapping and multiple layers of specialized split-channel feature transforms. The model was trained using the Adam optimizer (Kingma y Ba, 2017) for both the generator and the discriminator with a learning rate of 2×10^{-3} for both modules, and the training process spanned a total of 800,000 iterations. Additionally, the model employs several loss functions to train the neural network and enhance the quality of facial restoration.

To prevent data leakage, data augmentation, as recommended by LeCun et al. (2015), was exclusively applied to the training images. Using the GFPGAN framework, images were generated with upscaling factors ×1 (GFPGAN×1), ×2 (GFPGAN×2) and ×3 (GFPGAN×3). These new datasets

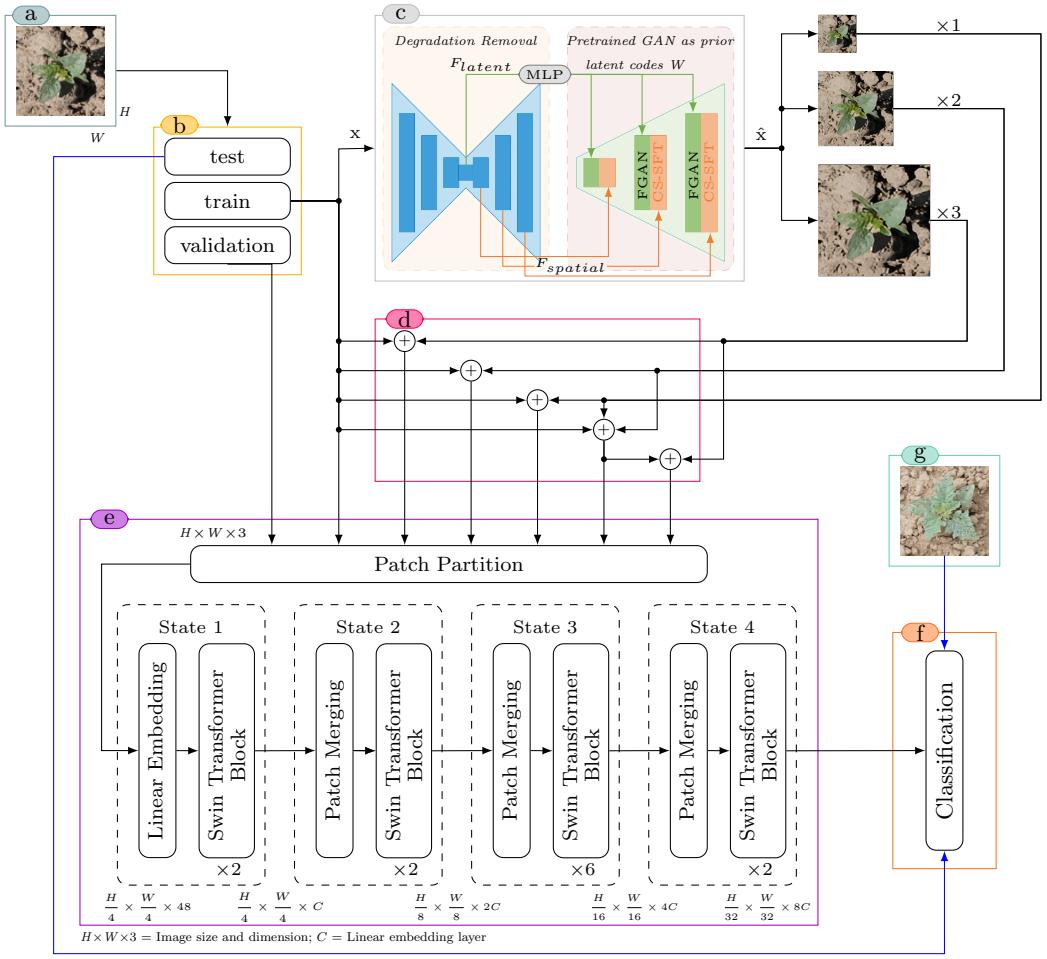


Figura 7.3: Flowchart of the CC system developed for weed monitoring at different phenological stages. (a) System input allows RGB images (12 classes) corresponding to BBCH14 and BBCH501 stages. (b) Preprocessing of the dataset. (c) GFPGAN framework. (d) Data augmentation. (e) Swin-T classification architecture. (f) Inference for BBCH17 and BBCH509 stages. Figure (c) adapted from (Wang et al., 2021). Figure (e) adapted from (Liu et al., 2021).

were used to augment the training data, resulting in the creation of various classification models (Figure 7.3d): Original + GFPGAN×1 (Ori + GFPGAN×1), Original + GFPGAN×2 (Ori + GFPGAN×2), Original + GFPGAN×3 (Ori + GFPGAN×3), Original + GFPGAN×1 + GFPGAN×2 (Ori + GFPGAN×1 + ×2), Original + GFPGAN×1 + GFPGAN×2 + GFPGAN×3 (Ori + GFPGAN×1 + ×2 + ×3).

Reference-complete quality metrics were employed to directly compare the target image (generated by GFPGAN) with the reference (original) image. The mean squared error (MSE) quantifies the root mean square difference between actual and ideal pixel values. While straightforward to compute, MSE may not accurately reflect human perception of quality. In contrast, the structural similarity index (SSIM) (Wang et al., 2004) integrates local image structure, luminance and contrast into a single local quality score. Here, structures refer to patterns of pixel intensities, particularly between adjacent pixels, normalized for luminance and contrast. Because the human visual system is very skilled at detecting structures, the SSIM metric is best suited for subjective quality assessment.

7.2.5 Vision Transformer Neural Network for Weed Detection

To validate our dataset and align our research with real-world conditions, we implemented the detection transformer (DETR) object detection model, based on ViTs, to identify and locate weed species in agricultural areas. Using DETR for this task offers significant advantages due to its ability to simplify the detection pipeline and enhance overall image reasoning. DETR eliminates the need for hand-designed components such as anchor generation and non-maximum suppression, which are common in other object detection models. Instead, it employs an encoder–decoder transformer architecture and global ensemble-based loss, guaranteeing unique predictions through bipartite allocation (Carion et al., 2020). This approach enables DETR to consider the relationships between objects and the global image context, producing the final set of predictions directly and in parallel. This feature is particularly beneficial for identifying multiple weed species in a single image, improving both the accuracy and efficiency of the detection process.

Three models were created using different combinations of data sets: Original, Ori + GFPGAN×1 and Ori + GFPGAN×1 + ×2. The PyTorch DL framework was used to train the DETR model. The hyperparameter settings included a learning rate of 1×10^{-4} and a weight decay of 1×10^{-4}. Batch sizes were set to 2 for the training data loader and 1 for the validation data loader. Training was conducted for a total of 20 epochs. In addition, the gradient trimming values were set to 0.1, gradient accumulation to 8 and a row record aggregation frequency of 5. This configuration was chosen to balance the stability and efficiency of the training, after testing several configurations.

To evaluate the performance of the DETR model during the training stage, several metrics were used. The intersection over union (IoU) metric evaluates the accuracy of model predictions by determining how well a predicted region overlaps with the actual object. IoU is calculated as the intersection area divided by the junction area between the ground truth box and the box predicted by the model. A high IoU value indicates a strong correspondence between predictions and actual annotations, essential for evaluating the spatial accuracy of the model. The mean average precision (mAP) assesses the model's accuracy and completeness in object detection. The mAP is calculated by averaging the accuracies at various retrieval levels, considering multiple IoU thresholds.

$$IoU = \frac{\textit{Area of Overlap}}{\textit{Area of Union}} \tag{7.5}$$

$$mAP = \frac{1}{|classes|}P \tag{7.6}$$

To comprehensively understand the model's performance in various real-world scenarios, several metrics were used. The mAP@[IoU = 0.50] and mAP@[IoU = 0.75] evaluate accuracy at the specific IoU thresholds 0.5 and 0.75, allowing a more detailed understanding of model performance at different levels of overlap. The mAP@[area = small/medium/large] metrics provide information about the model's effectiveness in detecting objects of different sizes. On the other hand, the recall metrics evaluate the model's ability to identify all relevant objects. Recall@[maxDets = 1/10/100] measures the recall considering only the top 1, 10 or 100 detections per image, respectively, which is useful to evaluate the performance at different levels of object density. Finally, Recall@[area = small/medium/large] complements the mAP metrics by evaluating the completeness of detections for objects of different sizes.

7.3 Results

7.3.1 Classification Inference Using the Original Datasets in Both Early and Subsequent Growth Stages

The results of inferring the Swin-T classification model on the dataset corresponding to the early growth stage are shown in Table 7.2. The evaluation highlights the outstanding performance of the multiclass classification in species identification, with all species achieving evaluation metrics above 97 %. This indicates the model's high accuracy in multiclass classification, showcasing its significant ability to effectively identify various species. Notably, certain species, such as *P. oleracea*, *S. nigrum*, *S. halepense*, maize and tomato, achieved 100 % accuracy and recall, highlighting the model's absolute reliability in identifying these particular species.

Tabla 7.2: Performance metrics for the Swin-T classification model applied to weed and crop species during the early growth stage (maize BBCH14 and tomato BBCH501).

Species	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	***F1-score* (%)**	**Support**
Atriplex patula	98.0	97.0	98.0	97.5	100
Chenopodium album	98.0	99.0	98.0	98.5	100
Convolvulus arvensis	99.0	100.0	99.0	99.5	100
Cyperus rotundus	99.0	100.0	99.0	99.5	100
Datura ferox	99.0	98.0	99.0	98.5	100
Lolium rigidum	98.0	100.0	98.0	99.0	100
Portulaca oleracea	100.0	99.0	100.0	99.5	100
Salsola kali	98.0	98.0	98.0	98.0	100
Solanum nigrum	100.0	99.0	100.0	99.5	100
Sorghum halepense	100.0	100.0	100.0	100.0	100
Maize	100.0	99.0	100.0	99.5	100
Tomato	100.0	100.0	100.0	100.0	100
Accuracy	99.1				
Macro average	99.1	99.1	99.1	99.1	
Weighted average	99.1	99.1	99.1	99.1	

The results of the Swin-T classification model's inference on the dataset corresponding to the subsequent growth stage are shown in Table 7.3. These inference results revealed varying performance in classifying different species. Notably, species such as maize and *S. halepense* exhibited a high level of precision and recall, with *F1-scores* close to 99 %. This suggests that the model excels in identifying these species. In contrast, *A. patula* had a low recall of 53.3 %, attributable to a high percentage of F_N between *A. patula* and *C. arvensis*, at 22 %. This suggested that the model had difficulty differentiating between these two species, leading to a higher omission rate of *A. patula* when it was actually present. The low accuracy of 55.6 % in *C. arvensis* was influenced by a significant percentage of F_N between *C. arvensis* and *C. album*, at 20 %. This level of F_N indicated that the model misclassified a considerable number of *C. album* instances as *C. arvensis*.

The low recall value of 52.2 % for *C. rotundus* was explained by the high rate of F_P with other species, specifically 16 % with maize and 14 % with *S. kali*. Confusion with *S. kali* was due to similar morphological characteristics. In the case of *L. rigidum*, its low accuracy of 37.1 % was due to a remarkable confusion with *S. kali* (53 %). This high F_P rate indicated that the model tended to misclassify instances of *S. kali* as *L. rigidum*. The confusion with *S. kali* was of particular concern and suggested considerable similarity in the data characteristics used by the model to differentiate these species.

Tabla 7.3: Performance metrics for the Swin-T classification model applied to weed and crop species during the subsequent growth stage (maize BBCH17 and tomato BBCH509).

Species	**Precision (%)**	**Recall (%)**	***F1-score* (%)**	**Support**
Atriplex patula	96.3	53.3	68.6	1459
Chenopodium album	91.9	82.0	86.7	2175
Convolvulus arvensis	55.6	97.3	70.7	1102
Cyperus rotundus	65.4	52.2	58.1	134
Datura ferox	88.7	91.7	90.2	589
Lolium rigidum	37.1	90.0	52.6	80
Portulaca oleracea	85.0	89.8	87.4	177
Salsola kali	98.5	88.3	93.2	1216
Solanum nigrum	98.2	100.0	99.1	2175
Sorghum halepense	67.8	100.0	80.8	103
Maize	99.0	98.0	98.5	24,614
Tomato	91.0	98.6	94.7	2732
Accuracy			94.7	36,556
Macro average	81.2	86.8	81.7	
Weighted average	95.9	94.7	94.8	

Figure 7.4 illustrates the original input images and their corresponding Grad-CAM activation maps. These maps highlight the areas within the input images that the model considers most relevant to its prediction. More intensely colored regions (usually red) indicate a greater contribution to the model's decision, while fainter regions (usually blue or uncolored) indicate less influence. Subimages (a), (d), (e), (j), (k) and (l) show the location of discriminative features that align with intuitive human features. In contrast, subimages (c), (f) and (g) show activations in the background rather than the main object, which may indicate model failure and bias.