BIOENG

Brazilian Journal of Biosystems Engineering (2026), 20 1389

REGULAR ARTICLE

Annotation-efficient weed mapping in sorghum fields using two-stage U-Net crop segmentation and HSV greenness analysis

Mohamed El Amine Bouhadjer [1]*; Sarah Mazari [1]; Miloud Chikr El Mezouar [1]

1 Department of Electronics, Djillali Liabes University, Sidi Bel Abbes, Algeria

Regular Section

Academic Editor: *Celso Antonio Goulart*

Statements and Declarations

Data availability

The code and derived datasets used in this study, "Annotation-Efficient Weed Mapping in Sorghum Fields Using Two-Stage U-Net Crop Segmentation and HSV Greenness Analysis," are publicly available on Zenodo (DOI: 10.5281/zenodo.17653873). The original sorghum UAV dataset published by Genze et al. (2022) is publicly accessible on Mendeley Data (DOI: 10.17632/4hh45vkp38.4).

Institutional Review Board Statement

Not applicable.

Conflicts of interest

The authors declare no conflict of interest.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or non-profit sectors.

Code/Software availability

The source code and software implementation developed for "Annotation-Efficient Weed Mapping in Sorghum Fields Using Two-Stage U-Net Crop Segmentation and HSV Greenness Analysis" are publicly available on Zenodo (DOI: 10.5281/zenodo.17653873).

Use of Generative AI

During the preparation of this work, the authors used ChatGPT by OpenAI and Claude by Anthropic for language editing, grammar correction, and writing verification purposes only. These tools were not used for generating, interpreting, or analyzing the scientific results or discussion of the study. The authors reviewed and edited all AI-generated output and take full responsibility for the content of this publication.

Author contribution (CRediT)

M.E.A.B.: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Writing – original draft, Visualization, Data curation; S.M.: Methodology, Validation, Formal analysis, Writing – review & editing, Supervision, Project administration; M.C.E.M.: Supervision, Project administration.

Abstract

Accurate weed mapping is essential for site-specific weed management in sorghum, but pixel-level weed annotation is labor-intensive and limits practical deployment. This work proposes an annotation-efficient two-stage approach using Unmanned Aerial Vehicle (UAV) RGB imagery that first segments crop pixels with U-Net and then detects weeds in non-crop regions using greenness-based analysis. Convolutional Neural Network (CNN)- and transformer-based encoders were evaluated for crop segmentation, and hue, saturation, value (HSV) filtering was compared with the Excess Green index (ExG) and the Color Index of Vegetation Extraction (CIVE) for weed detection. Among the evaluated configurations, the best-performing combination used U-Net with an EfficientNet-B5 encoder for crop segmentation and HSV-based filtering for weed detection. The approach was evaluated across sorghum growth stages (BBCH 15, 17, and 19) and produced practical weed-mapping outputs while substantially reducing annotation requirements, because only crop masks were needed and manual weed labeling was avoided. In total, 4,470 weed instances did not require manual annotation. When compared with a fully supervised end-to-end multi-class model, the proposed two-stage framework showed lower but practically useful segmentation performance while substantially reducing annotation requirements. The method also generated field-scale weed distribution maps and weed-cover estimates that can support targeted scouting and site-specific herbicide application. These results support the proposed approach as a practical and scalable alternative for precision weed management under real field conditions.

Keywords

Weed detection; Precision agriculture; Vegetation indices; Greenness analysis; Annotation-efficient learning

1. Introduction

Weeds pose a major challenge to crop production because they compete with crops for nutrients, water, sunlight, and space (Patel and Kumbhar, 2016). Weed control has traditionally relied on mechanical practices and chemical herbicides (Aktar et al., 2009). Although herbicides remain widely used because of their efficiency, environmental and health concerns have motivated the development of more sustainable weed management strategies (WHO, 1990; Hasan et al., 2021).

Recent advances in computer vision and deep learning have created new opportunities for weed detection and segmentation. Autonomous robots and UAVs equipped with imaging sensors and intelligent algorithms can support site-specific weed management, including targeted herbicide application and mechanical removal, thereby reducing herbicide use and improving crop productivity (Hasan et al., 2021; Ahmad et al., 2018). Semantic segmentation is particularly relevant in precision agriculture because it provides pixel-level delineation of plants, which is important for precision spraying and localized weed control (Li et al., 2023).

However, accurate weed detection within crop canopies remains difficult. In RGB imagery, weeds often share similar color, texture, and shape characteristics with crops, leading to pixel-level confusion (Wang et al., 2019). Field conditions such as illumination changes, shadows, motion blur, and occlusions further degrade segmentation reliability (Wang et al., 2019; Olsen et al., 2019). Weed detection is also affected by weed-species diversity and variation across growth stages, which limit model generalization (Wu et al., 2021). In addition, many deep segmentation methods require large, densely annotated datasets and computationally intensive training, while pixel-wise weed labeling is costly and time-consuming (Rai et al., 2023).

Previous studies have demonstrated strong performance of deep learning models such as U-Net and SegNet for weed and crop segmentation from UAV imagery (Asad and Bais, 2020;

*Corresponding author

E-mail address: mohamedelamine.bouhadjer@univ-sba.dz

https://doi.org/10.18011/bioeng.2026.v20.1389

Received: 10 March 2026 / Accepted: 19 May 2026 / Available online: 28 June 2026

Zou et al., 2021). Reviews have also highlighted both their effectiveness and their dependence on annotated data (Rai et al., 2023). Hybrid pipelines that combine deep learning and image processing have shown promise for reducing computational burden and simplifying downstream classification (Jin et al., 2021). Transformer-based segmentation models can achieve excellent weed-detection performance with fully annotated datasets (Garibaldi-Márquez et al., 2025), but they generally require dense pixel-wise labels and substantial computational resources. In parallel, greenness-based and vegetation-index methods remain attractive because they are interpretable and computationally efficient. Methods based on HSV and RGB-derived indices (e.g., ExG and CIVE) can provide effective vegetation discrimination under outdoor conditions, but their performance is sensitive to illumination, soil background, and crop–weed spectral similarity (Yang et al., 2015). Other studies have confirmed the utility of threshold-based indices in challenging environments (Štroner et al., 2023) and in UAV-based vegetation analysis (Chen et al., 2024; Macedo et al., 2025). Together, these findings support hybrid strategies that combine robust crop segmentation with low-cost greenness-based weed identification, while highlighting the need for annotation-efficient two-stage weed mapping approaches in sorghum UAV RGB imagery.

To address these limitations, this work proposes a two-stage weed detection framework for UAV RGB imagery of sorghum under real field conditions with natural weed infestation, evaluated across multiple BBCH growth stages (Genze et al., 2022). In the first stage, a U-Net-based crop segmentation model is trained using multiple CNN (ResNet, DenseNet, EfficientNet) and transformer (Swin Transformer, and Vision Transformer (ViT)) encoder backbones, and the best-performing backbone is selected to segment and remove crop pixels from the input images. In the second stage, greenness-based vegetation identification methods (HSV, ExG, and CIVE) are applied to the remaining vegetation to detect weeds. By combining deep learning-based crop segmentation with computationally efficient greenness analysis, the proposed framework aims to provide an annotation-efficient and scalable solution for weed mapping in precision agriculture. To assess the quality and robustness of the approach, we compare the proposed two-stage pipeline with a fully supervised end-to-end deep learning baseline for joint crop–weed segmentation. We also analyze the distribution of crop, weed, and background pixels across multiple crop growth stages (e.g., BBCH 15, 17, and 19) to evaluate stage-specific performance and temporal consistency of predictions.

2. Materials and methods

The methodology proposed in this study integrates deep learning and greenness identification techniques to achieve precise weed segmentation in crop fields. The approach is divided into two main phases: crop segmentation and weed detection.

2.1 Data Acquisition and Augmentation

This study used the sorghum UAV RGB dataset reported by Genze et al. (2022). Images were acquired over an experimental sorghum field using a DJI Mavic 2 Pro Drone equipped with a 20 MP Hasselblad camera (L1D-20c). The sorghum hybrid ‘Farmsugro 180’ was sown at 37.5 cm row spacing and a seeding density of approximately 25 seeds m^{-2} (≈250,000 seeds ha^{-1}).

During image acquisition (late spring, BBCH 17; BBCH refers to the Biologische Bundesanstalt, Bundessortenamt und CHemische Industrie phenological growth scale by Hess et al. (1997)), the field contained several weed species, including Chenopodium album L., Thlaspi arvense, Matricaria chamomilla, Veronica officinalis, and Onopordum acanthium. The UAV was flown at 5 m altitude, yielding a ground sampling distance of approximately 1 mm, which enabled clear visual separation of sorghum, weeds, and soil. At this altitude and with the 5472 × 3648 pixel sensor, each individual image covered a ground footprint of approximately 5.47 m × 3.65 m, corresponding to about 20 m^2 per frame.

A total of 60 RGB images (5472 × 3648 pixels) were captured with approximately 10% overlap between consecutive frames, resulting in motion blur. Considering the perframe footprint and the 10% overlap, the 60 images jointly cover an analyzed ground area on the order of 10^3 m^2 (≈ 0.1 ha) of the experimental sorghum plot, which was sufficient for BBCH 17 model development. From these, 19 images were selected for detailed analysis and manually annotated pixel wise into three classes (sorghum, weeds, and soil), with agronomic expert guidance to ensure annotation quality.

The number of annotated images reflects the high labor cost of dense pixel-wise annotation for crop/weed/soil classes. For crop segmentation model training in the present study, the annotations were converted into two classes: crop (sorghum) and background (weeds + soil). The 19 annotated BBCH 17 images were divided into non-overlapping 256 × 256 patches, yielding 6270 patches (3960 for training/validation and 2310 for testing). A 4-fold cross-validation protocol was applied to the training/validation subset using a 75%/25% split in each fold.

To assess performance across crop development stages, two additional UAV images from BBCH 15 and BBCH 19 (each 2560 × 2816 pixels) acquired from different parts of the same experimental field were included as supplementary test data. Each image was divided into 110 non-overlapping 256 × 256 patches and used to evaluate stage-specific robustness under earlier and later growth conditions.

Table 1 summarizes the main characteristics of the datasets used in this study, including the number of annotated patches and the number of sorghum and weed instances per different Sorghum growth stages (BBCH). Figure 1 shows an example UAV image and a corresponding 256 × 256 patch containing both sorghum and weeds.

Table 1. Summary statistics of the datasets used in this study, BBCH refers to the Biologische Bundesanstalt, Bundessortenamt und CHemische Industrie phenological growth scale (Hess et al., 1997).

Dataset Name	Sorghum BBCH	Annotated Patches	Sorghum Instances	Weed Instances
sorghum_17	17	6270	3056	3060
sorghum_15	15	110	115	429
sorghum_19	19	110	99	981

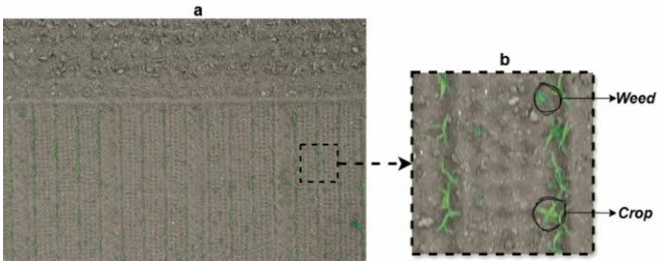

Figure 1. Example drone capture and related patch for the dataset sorghum_17. a UAV capture from one of the sorghum fields, b 256x256 pixel patch that contains both weeds and crops (sorghum).

To improve model robustness and generalization, data augmentation was applied to the training and validation patches. The augmentation strategy included horizontal and vertical flips, rotations (90°, 180°, and 270°), and random zoom-in/zoom-out operations of up to 20% to simulate variability in orientation, scale, and spatial arrangement commonly observed in UAV agricultural imagery. After augmentation, the training/validation dataset increased to 31,680 patches, including 23,760 training patches and 7,920 validation patches. This augmentation process improved dataset diversity and supported better generalization under varying field structures, lighting conditions, and crop–weed spatial configurations.

2.2 Crop Segmentation Using U-Net

To isolate crop regions before weed extraction, we adopted a U-Net-based segmentation framework and evaluated multiple encoder backbones, including CNN-based models (ResNet, DenseNet, and EfficientNet) and transformer-based models (Swin Transformer and Vision Transformer). The final model was selected using the lowest validation focal loss, while also considering the trade-off between segmentation performance and computational cost. The U-Net follows the classical encoder-decoder structure proposed by Ronneberger et al. (2015), where the encoder progressively extracts semantic features and the decoder reconstructs the segmentation map at the original resolution. Skip connections between corresponding encoder and decoder stages preserve fine spatial details and improve crop-boundary delineation in UAV RGB imagery (Ronneberger et al., 2015). Figure 2 shows the configuration used in this study.

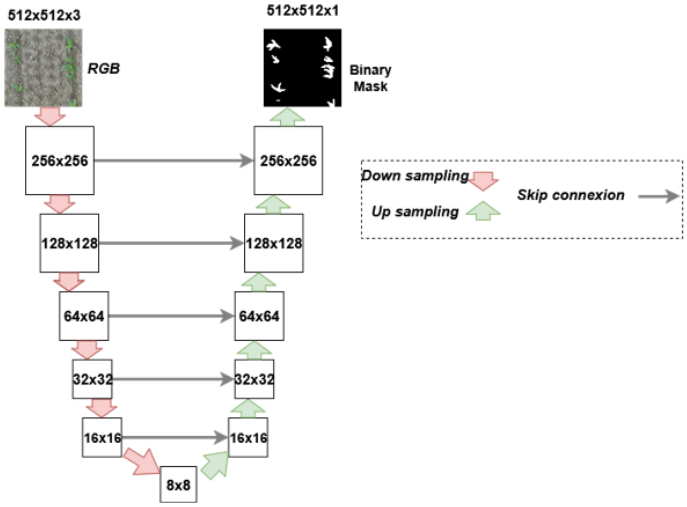

Figure 2. U-Net architecture with encoder–decoder structure and skip connections.

The evaluated CNN backbones provide complementary representational properties. ResNet variants use residual skip connections that facilitate optimization of deeper models (He et al., 2016), DenseNet variants promote feature reuse and gradient flow through dense connectivity (Huang et al., 2017), and EfficientNet variants improve the accuracy-efficiency trade-off through compound scaling of depth, width, and resolution (Tan and Le, 2019). Transformer-based encoders were also evaluated to improve context-aware segmentation. In contrast to convolutional models, transformers use self-attention to model long-range dependencies (Vaswani et al., 2017). Swin Transformer backbones use shifted-window attention and hierarchical representations that are well suited to encoder-decoder segmentation pipelines (Liu et al., 2021), whereas Vision Transformer (ViT) encoders process patch-based token representations with global self-attention (Dosovitskiy et al., 2021).

All backbone variants were trained under identical settings (input size 256 × 256, Adam optimizer), and validation focal loss was used as the primary model-selection criterion. Table 2 summarizes the channel progression and trainable parameters for U-Net models with CNN-based encoders. Table 3 summarizes the transformer-based encoder variants. In general, deeper or higher-capacity backbones can improve representation power but increase memory use and computational cost. The backbone comparison was therefore used to identify a model that provides strong crop-segmentation performance while remaining practical for the proposed two-stage pipeline.

Table 2. U-Net with different CNN backbones (ResNet, DenseNet, EfficientNet).

Backbone	Params (M)	Encoder Channels	Decoder Channels
ResNet18	12.91 M	64 → 64 → 128 → 256 → 512	256 → 128 → 64 → 64 → 32
ResNet34	23.01 M	64 → 64 → 128 → 256 → 512	256 → 128 → 64 → 64 → 32
ResNet50	50.07 M	64 → 256 → 512 → 1024 → 2048	1024 → 512 → 256 → 128 → 64
ResNet101	68.97 M	64 → 256 → 512 → 1024 → 2048	1024 → 512 → 256 → 128 → 64
ResNet152	84.67 M	64 → 256 → 512 → 1024 → 2048	1024 → 512 → 256 → 128 → 64
DenseNet121	13.65 M	64 → 128→ 256 → 512 →1024	512 → 256 → 128 → 64 → 64
DenseNet161	51.67 M	96 → 192 → 384 → 1056 → 2208	1056 → 384 → 192 → 96 → 64
DenseNet169	22.87 M	64 → 128 → 256 → 640 → 1664	640 → 256 → 128 → 64 → 64
DenseNet201	35.19 M	64 → 128 → 256 → 896 → 1920	896 → 256 → 128 → 64 → 64
EfficientNet-B0	14.52 M	96 → 184 → 360 → 752 → 1280	640 → 320 → 160 → 80 →64
EfficientNet-B1	17.02 M	96 → 184 → 360 → 752 → 1280	640 → 320 → 160 → 80 →64
EfficientNet-B2	20.50 M	104 → 200 → 400 → 824 → 1408	704 → 352 → 176 → 88 →64
EfficientNet-B3	25.59 M	120 → 224 → 432 → 904 → 1536	768 → 384 → 192 → 96 →64
EfficientNet-B4	38.10 M	136 → 256 → 504 → 1056 → 1792	896 → 448 → 224 → 112 →64
EfficientNet-B5	55.23 M	152 → 296 → 576 → 1200 → 2048	1024 → 512 → 256 → 128 →64
EfficientNet-B6	74.71 M	176 → 328 → 648 → 1352 → 2304	1152 → 576 → 288 → 144 →64
EfficientNet-B7	105.44 M	192 → 368 → 720 → 1504 → 2560	1280 → 640 → 320 → 160 →64

Table 3. U-Net with Transformers backbones (Swin Transformer, Vision Transformer).

Backbone	Params (M)	Encoder Channels	Decoder Channels
Swin-T	33 M	96 → 192 → 384 → 768	256 → 128 → 64 → 32 → 32
Swin-S	55 M	96 → 192 → 384 → 768	256 → 128 → 64 → 32 → 32
Swin-B	95 M	128 → 256 → 512 → 1024	256 → 128 → 64 → 32 → 32
ViT-B/16	92 M	64 → 128 → 768 → 768	256 → 128 → 64 → 32 → 32
ViT-B/32	94 M	64 → 128 → 768 → 768	256 → 128 → 64 → 32 → 32

2.3 Weed Detection Using Greenness Identification Techniques

After training and selecting the best-performing crop segmentation model based on validation focal loss, the U-Net with an EfficientNet-B5 encoder was applied to the test dataset to predict crop masks. This step reduced the likelihood of false weed detections and increased the robustness of subsequent greenness-based image processing. Following crop suppression, weed detection was performed using three greenness identification techniques: HSV color filtering, ExG, and CIVE. Each method was applied independently to the crop-free images in order to assess its effectiveness under identical conditions.

2.4 HSV Color Filtering

For HSV-based weed detection, the crop-suppressed RGB images were first converted into the HSV color space. This transformation decouples chromatic information from intensity, making the detection process more robust to illumination variations commonly encountered in UAV imagery. Weed candidates were then extracted by applying fixed thresholds on the hue, saturation, and value channels. The HSV interval was selected based on an analysis of the hue histogram of weed pixels, resulting in a lower bound of [35, 50, 40] and an upper bound of [88, 255, 255]. As illustrated in Figure 3, the dominant peak in the hue histogram between 35 and 88 corresponds to the green spectral range associated with weeds, while values outside this interval mainly represent soil and noise.

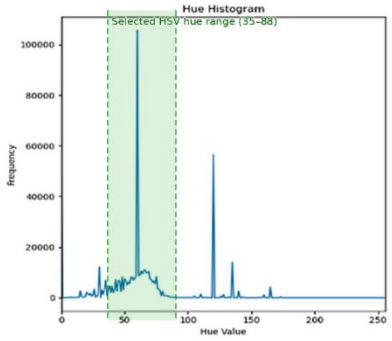

Figure 3. Hue Histogram.

After thresholding, small, connected components containing fewer than 50 pixels were removed from the binary mask. This post-processing step suppresses isolated noise regions and ensures that only spatially coherent weed patches are retained, leading to cleaner and more reliable weed segmentation results.

2.5 Excess Green Index (ExG)

As a second approach, weeds were detected using the Excess Green index, which enhances green vegetation in RGB images. The ExG index was computed as

$$ExG = 2 \times G - R - B \qquad (1)$$

where R, G, and B denote the red, green, and blue color channels, respectively (Woebbecke et al., 1995). The resulting ExG image was normalized to the range [0,255] to eliminate negative values and ensure consistent intensity scaling. To reduce pixel-level noise, a median filter with a 5×5 kernel was applied prior to binarization (Huang et al., 1979). Otsu's thresholding method was then used to convert the filtered ExG image into a binary mask separating vegetation from background (Otsu, 1979). Finally, connected components smaller than 50 pixels were removed to reduce spurious detections and improve mask cleanliness.

2.6 Color Index of Vegetation Extraction (CIVE)

The third method evaluated in this study was the Color Index of Vegetation Extraction (CIVE), which is designed to emphasize green vegetation while suppressing soil background. CIVE was computed using the formulation.

$$CIVE = 0.441 \times R - 0.811 \times G + 0.385 \times B + 18.78745 \qquad (2)$$

as proposed by Kataoka et al. (2003). Similar to the ExG-based pipeline, the CIVE image was normalized to the range [0,255], followed by noise reduction using a 5×5 median filter. Otsu's thresholding was then applied to generate a binary weed mask, and small connected components with fewer than 50 pixels were removed to further suppress noise. The complete processing pipeline, from crop segmentation to weed detection using HSV, ExG, and CIVE, is summarized in Figure 4.

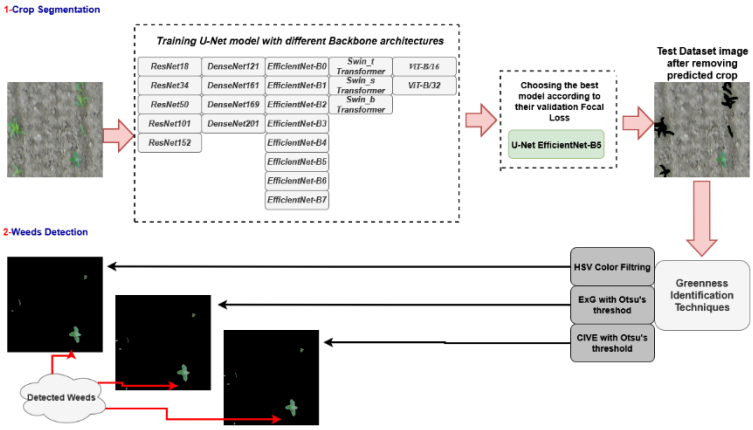

Figure 4. Diagram of the complete proposed weed detection process

2.7 Evaluation Metrics

In this work, we use accuracy, precision, recall, F1-score, Dice coefficient, IoU, and Focal Loss to evaluate segmentation performance, as summarized in Table 4.

Together, these metrics provide a complementary view of segmentation quality. Accuracy reflects the global rate of correct pixel classification, whereas precision and recall quantify, respectively, the reliability of positive predictions and the ability to retrieve all relevant pixels. The F1-score and Dice coefficient summarize this trade-off into a single overlap-oriented metric, while IoU offers a stricter assessment of region agreement between prediction and ground truth. Finally, Focal Loss is used during training to mitigate class

imbalance by emphasizing hard-to-classify pixels, which is critical in scenarios where vegetation covers only a small fraction of the image. For clarity on how metrics are reported in this study: Tables 6 and 7 correspond to binary segmentation (Crop–Background and Weed–Background). Metrics are computed on binarized masks, and in this setting the Dice coefficient is equivalent to the F1-score, so both values are reported identically. IoU is also reported as a stricter overlap metric. Table 9 corresponds to three-class segmentation (crop, weed, background), where results are reported as macro-averages across classes. In this case, precision, recall, F1-score, and IoU are computed per class and then averaged, and Dice is computed per class and then macro-averaged, therefore binary algebraic equivalences (e.g., Dice = F1) do not necessarily hold under macro-averaging.

Table 4. Evaluation metrics used for segmentation.

Metric	Formula	Interpretation/Note
Accuracy	$Accuracy = \frac{TP+TN}{TP+FP+TN+FN}$ (3)	Overall proportion of correctly classified pixels. Higher is better.
Precision	$Precision = \frac{TP}{TP+FP}$ (4)	Among pixels predicted as positive, fractions that are truly positive.
Recall	$Recall = \frac{TP}{TP+FN}$ (5)	Fraction of true positive pixels that are correctly detected.
F1-Score	$F1 = \frac{2\times Precision\times Recall}{Precision+Recall}$ (6)	Harmonic mean of precision and recall; robust to class imbalance.
Dice coefficient	$Dice = \frac{2TP}{2TP+FP+FN}$ (7)	Overlap between prediction and ground truth; for binary masks, equal to F1.
Intersection over Union (IoU)	$IoU = \frac{TP}{TP+FP+FN}$ (8)	Jaccard index; ratio of intersection to union of predicted and true regions.
Focal Loss	$FL(pt) = -\alpha(1-pt)\gamma\, log(pt)$ (9)	Training loss that down-weights easy pixels to focus on hard examples.

2.8 Experimental Setup

Experiments were conducted on a NVIDIA RTX 3060 GPU (12 GB VRAM). The implementation was developed in Python 3.9 using PyTorch 1.13.0, Torchvision, and CUDA 11.6 for accelerated computation. A 4-fold cross-validation strategy was used to improve robustness and reduce sensitivity to data partitioning. In each fold, the model was trained for up to 50 epochs with early stopping (patience = 10 epochs) based on validation focal loss. The initial learning rate was set to 0.001, the batch size was 8, and model parameters were optimized using the Adam optimizer. Figure 5 summarizes the model selection and evaluation workflow across the four folds. Because the learning stage of the proposed pipeline is limited to binary crop segmentation, the training setup is simpler than that of fully supervised multi-class crop–weed segmentation and is expected to reduce computational training cost.

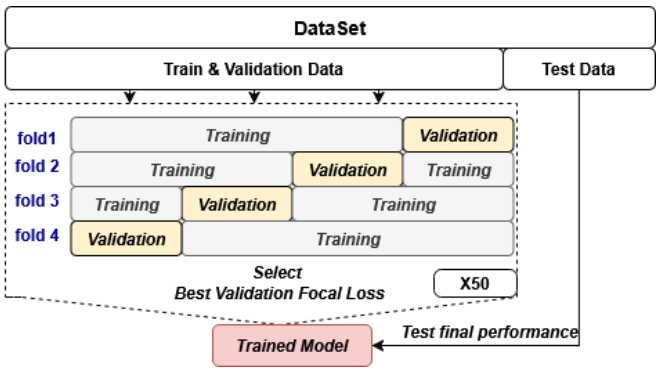

Figure 5. Overview of the model selection and evaluation process

For comparison with a fully supervised end-to-end baseline, a U-Net with an EfficientNet-B5 encoder was also trained from scratch to directly predict three classes (crop, weed, and background) using the same experimental settings (input size 256 × 256, Adam optimizer, initial learning rate 0.001, batch size 8, 4-fold cross-validation, maximum 50 epochs, and focal loss). For the proposed two-stage framework, a three-class prediction was reconstructed by combining the binary crop mask predicted by U-Net_EfficientNet-B5 with the binary weed mask produced by the HSV-based procedure, while remaining pixels were assigned to the background class. Both approaches were evaluated on the same test patches using macro-averaged metrics (Dice, F1-score, precision, recall, and IoU) across the three classes.

3. Results and discussion

3.1 Crop Segmentation Performance

To examine optimization behavior and generalization across encoder backbones, Figures 6 and 7 show the training and validation curves (focal loss and accuracy) for CNN-based and transformer-based U-Net variants, respectively.

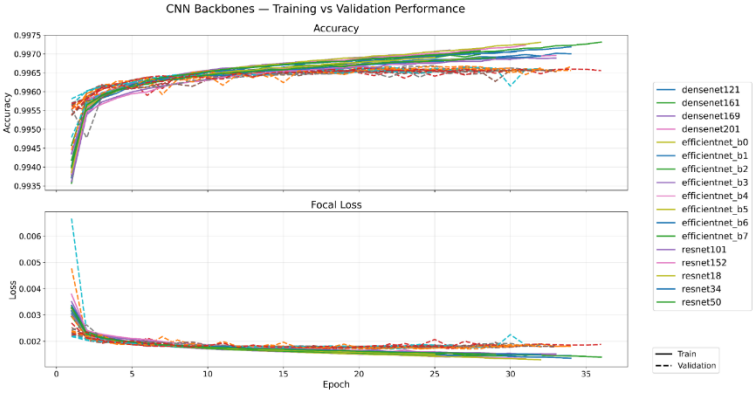

Figure 6. Training and validation curves of Focal Loss and Accuracy for U-Net with ResNet, DenseNet and EfficientNet backbones.

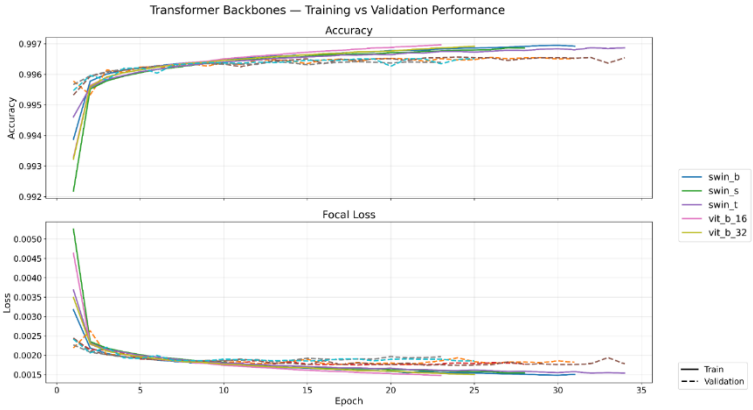

Figure 7. Training and validation curves of Focal Loss and Accuracy for U-Net with Swin and ViT transformer backbones.

In both model families, focal loss decreased steadily while accuracy increased during training, and training/validation curves remained close after the first epochs, indicating stable convergence without pronounced overfitting under the adopted training protocol. Across most backbone variants, peak performance was observed between epochs 19 and 22, where validation focal loss reached its minimum and validation accuracy reached its maximum. Table 5 summarizes the best-epoch performance of all evaluated U-Net backbones on the crop segmentation task.

Table 5. Experimental results in both Training and validation of the best epoch.

U-Net Backbone	Training Focal Loss	Validation Focal Loss
ResNet18	0.00166	0.001758
ResNet34	0.00160	0.001723
ResNet50	0.00164	0.001736
ResNet101	0.00161	0.001746
ResNet152	0.00165	0.001757
DenseNet121	0.00154	0.001749
DenseNet161	0.00162	0.001757
DenseNet169	0.00166	0.001748
DenseNet201	0.00167	0.001756
EfficientNet-B0	0.00161	0.001744
EfficientNet-B1	0.00160	0.001739
EfficientNet-B2	0.00159	0.001721
EfficientNet-B3	0.00158	0.001717
EfficientNet-B4	0.00151	0.001705
EfficientNet-B5	**0.00148**	**0.001688**
EfficientNet-B6	0.00157	0.001710
EfficientNet-B7	0.00157	0.001718
Swin-T	0.00160	0.001741
Swin-S	0.00165	0.001748
Swin-B	0.00155	0.001760
ViT-B/16	0.00166	0.001833
ViT-B/32	0.00167	0.001850

Overall, increasing model depth or capacity did not uniformly improve performance. Within the ResNet family, ResNet34 achieved the lowest validation focal loss (0.001723), indicating that moderate depth was sufficient for this task. DenseNet models showed stable but modest performance, with DenseNet121 performing best within that family (validation focal loss = 0.001749), while deeper variants did not provide clear gains.

EfficientNet backbones yielded the strongest overall results, with EfficientNet-B5 achieving the lowest training focal loss (0.00148) and the best validation focal loss (0.001688) among all evaluated encoders. This indicates that EfficientNet's compound scaling provides a favorable balance between representational capacity and efficiency for crop segmentation in UAV RGB patches (256 × 256). In the present setting, EfficientNet-B5 appears to provide a suitable capacity level: it is expressive enough to capture fine crop boundaries and local texture cues, while avoiding the additional complexity of larger variants that did not yield further validation gains. Transformer-based encoders were competitive but did not surpass the best CNN-based models under the present dataset size and resolution. Among transformer variants, Swin-T achieved the best validation focal loss (0.001741), whereas ViT-B/16 and ViT-B/32 produced the highest validation losses (0.001833 and 0.001850, respectively), suggesting that the available training conditions favored convolutional inductive biases for fine boundary localization.

Based on this comparative analysis, U-Net with an EfficientNet-B5 encoder was selected as the reference backbone for subsequent experiments. In addition to the lowest validation focal loss, it also achieved the highest validation accuracy (99.66%) and a strong Dice score (90.59%), indicating consistent superiority across optimization- and overlap-based metrics under the present training conditions. Table 6 reports test-set performance for binary crop–background segmentation using the selected U-Net_EfficientNet-B5 model (background = weeds + soil). The model achieved high overall performance (accuracy = 99.68%, Dice = 93.18%, IoU = 87.17%), confirming that crop regions can be reliably isolated before weed extraction. This high crop-segmentation quality is critical for the proposed two-stage pipeline because crop-suppression errors directly propagate to the subsequent greenness-based weed detection stage.

Table 6. Experimental average results (%) for binary Crop–Background segmentation on the testing set.

Model	Accuracy	Dice score	Precision	Recall	F1-score	IoU
U-Net EfficientNet-B5	99.68	93.18	97.69	89.07	93.18	87.17

Figure 8 displays 10 image patches from the test dataset, organized as follows: (a–d) contain both crops and weeds, (e–f) contain only crops, (g–h) contain only weeds, and (i, k) show only soil. For each patch, the figure presents the original image, the ground-truth crop mask (when crop pixels are present), the prediction produced by the trained U-Net–EfficientNet-B5 model, and the corresponding crop-suppressed image in which predicted crop pixels were removed. Because the ground-truth masks correspond to manually annotated crop pixels, no crop masks are shown for panels (g, h, i, k), which do not contain crops.

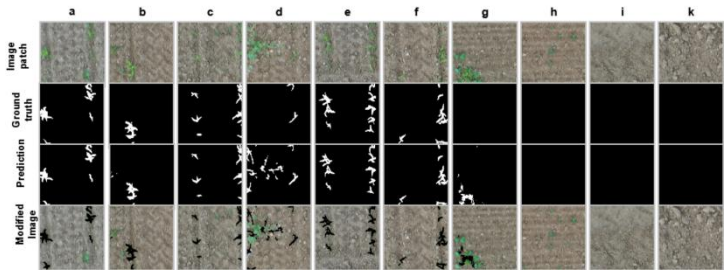

Figure 8. Qualitative results on the hold-out test-set of crop segmentation. The difference map shows test patches, ground truth of and prediction of segmented crop. a-d Examples of patches that contain both of crop and weed, e-f Examples of only crops, g-h examples of only weed, i-k examples of empty patches that have only soil.

Overall, the qualitative results confirm the strong performance of the crop segmentation model. Small errors may occur in regions where crops and weeds interact or partially overlap, and along with some leaf edges where boundaries are visually ambiguous, which can lead to slight crop-mask inaccuracies. Nevertheless, these errors remain limited, and the model still provides high-performance crop segmentation with reliable and accurate results for the subsequent crop-suppression and weed-detection stages.

3.2 Weed Segmentation Performance

After crop segmentation, predicted crop pixels were removed from the test images, yielding crop-suppressed images in which the remaining pixels corresponded primarily to weeds and soil, with a small number of residual crop pixels due to segmentation imperfections. Weed segmentation was then formulated as a binary weed–background task (weed vs. non-weed), and three greenness-based methods were evaluated under identical conditions: HSV color filtering, Excess Green (ExG), and CIVE.

In this study, all weed species present in the dataset were grouped into a single "weed" class for segmentation and evaluation. This formulation is aligned with the goal of early weed mapping for site-specific management, where reliable detection of weed-infested areas is often more immediately actionable than species-level discrimination.

Table 7 summarizes the performance of the three greenness-based methods on the BBCH 17 test set. Among the evaluated approaches, HSV color filtering achieved the strongest and most balanced performance (accuracy = 99.62%, precision = 82.21%, recall = 80.41%, F1-score = 81.30%, IoU = 68.66%). These results indicate that HSV-based filtering can detect most weed pixels while limiting false positives, making it the most suitable rule-based weed extraction step within the proposed two-stage framework.

Table 7. Experimental average results (%) for binary Weed–Background segmentation on the testing set.

	Accuracy	Dice score	Precision	Recall	F1-score	IoU
HSV	99.62	81.30	82.21	80.41	81.30	68.66
ExG	69.17	29.12	17.21	93.06	29.12	16.46
CIVE	70.48	32.11	19.62	92.81	32.11	18.71

By contrast, ExG and CIVE produced high recall (ExG: 93.06%; CIVE: 92.81%) but very low precision (ExG: 17.21%; CIVE: 19.62%), resulting in poor F1-score and IoU values. This pattern indicates systematic over-segmentation, where large numbers of background pixels are incorrectly labeled as weeds. Under the present field conditions, these methods were therefore too sensitive to background variability and non-target green responses to provide reliable weed maps as standalone binary weed detectors.

Figure 9 presents qualitative weed-detection results for the different greenness-identification techniques, including the crop-suppressed images, the weed ground-truth masks, and the predicted weed segmentations. Because the ground truth corresponds to weed masks, no masks are shown for panels (e, f, i, k), which do not contain weeds.

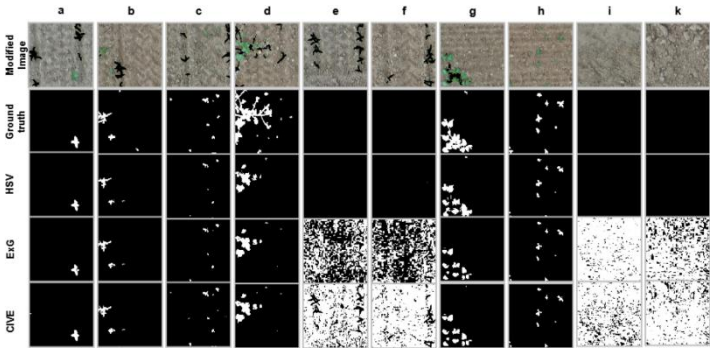

Figure 9. Qualitative results of weed detection. The difference map shows modified image same as figure.8, ground truth of and prediction of segmented weed by different processes.

As illustrated in Figure 9, HSV generally yields the most consistent weed segmentation across patch types and better separates weeds from soil and other background elements. Its main limitation appears at weed boundaries, particularly where brown or yellowish margins reduce apparent greenness, leading to partial under-segmentation along object edges.

Importantly, the HSV thresholds were deliberately chosen to be conservative. Expanding the threshold range to include weakly green or brownish edge pixels would increase soil–weed confusion, substantially raising false positives and degrading precision. This thresholding strategy therefore prioritizes robust weed localization over aggressive boundary inclusion. As a result, some boundary mismatches reduce overlap-based metrics (Dice/F1/IoU) even when the central weed regions are correctly detected. In contrast, ExG and CIVE combined with Otsu's thresholding generate acceptable results mainly in vegetation-dense patches (e.g., Figure 9a–c, h), but produce substantial noise in crop-only or soil-dominant patches (e.g., Figure 9 e, f, i, k), confirming their sensitivity to background variation and tendency toward over-segmentation.

The large performance gap observed between HSV and the two scalar vegetation indices (ExG and CIVE) can be explained by the different mathematical formulations of the three approaches and by their different sensitivity to illumination. ExG is defined as a linear combination of the three colour channels, $ExG = 2G - R - B$, in which the green channel receives a positive weight of +2 and the red and blue channels are penalized symmetrically. CIVE, on the other hand, uses the weighted combination $CIVE = 0.441R - 0.811G + 0.385B + 18.787$, in which the green channel receives a much stronger negative coefficient (so that vegetation pixels correspond to lower CIVE values) and the red and blue channels are treated asymmetrically. Because of these different algebraic forms, the two indices do not respond identically to yellowish-green weeds, sunlit crop foliage, or

reddish-brown soil pixels, which explains why CIVE achieved slightly higher precision than ExG (19.62% vs. 17.21%) in our experiments while both maintained comparably high recall (above 92%): the stronger weighting of the green channel in CIVE penalizes bright red soil more effectively than ExG. However, both indices share a fundamental limitation: they reduce the full three-dimensional colour information to a single scalar value that remains a linear function of the raw RGB intensities. Therefore, both ExG and CIVE inherit the overall brightness of each pixel: under the variable illumination conditions typical of UAV imagery (sunlit versus shaded regions, motion blur, partial cloud cover), background pixels with medium intensity in all three channels produce ambiguous index values that Otsu's global thresholding cannot reliably separate from true vegetation. This is precisely the source of the over-segmentation pattern observed in Table 7, where recall exceeds 92% for both indices but precision collapses below 20%.

The HSV colour space avoids this limitation by construction. Through the RGB → HSV transformation, chromatic information (Hue, Saturation) is explicitly decoupled from pixel brightness (Value). Sunlit sorghum foliage and shaded weed plants behave very differently with respect to the Value channel: sunlit crop leaves tend to saturate towards high V, while weeds partially occluded by shadows or affected by motion blur retain a lower V but a comparable Hue. By thresholding the three HSV channels jointly (H = [35, 88], S = [50, 255], V = [40, 255]), we simultaneously require (i) a green hue, (ii) a minimum color purity that excludes near grey soil, and (iii) a minimum brightness that excludes deep shadows and dark soil crumbs. None of these three criteria can be expressed by a single scalar index such as ExG or CIVE, and this is the structural reason why HSV filtering reached an IoU of 68.66% while the scalar indices remained below 20% IoU in the same experimental conditions. From a practical standpoint, this observation also has an important implication for the two-stage pipeline. Once the U-Net first stage removes crop pixels, the remaining scene is essentially composed of weeds, soil, and shadow regions. The task thus shifts from "green vs. non-green" (where scalar indices such as ExG and CIVE are acceptable) to "green vegetation vs. non-green residuals under variable illumination" a discrimination for which the multi-channel, brightness-aware nature of HSV becomes decisive.

Overall, Figure 9 visually supports the quantitative results reported in Table 7. The lower IoU observed for HSV-based weed segmentation is therefore not primarily due to incorrect weed localization, but rather to a deliberate trade-off between boundary completeness and background robustness required to limit soil–weed confusion in UAV imagery. From a practical standpoint, this result is important because it shows that a simple and interpretable HSV-based rule, when applied after accurate crop suppression, can provide useful weed-area maps while substantially reducing annotation burden. This supports a practical workflow for early weed scouting, weed-infested area delineation, and targeted intervention in UAV imagery without requiring dense weed labels for deep learning training. The implications of this trade-off for reconstructed three-class segmentation performance are examined in the subsequent comparison with the fully supervised end-to-end baseline.

3.3 Comparison with End-to-End Deep Learning Method

Table 8 reports the training and validation focal loss of the fully supervised end-to-end U-Net_EfficientNet-B5 baseline trained to jointly segment crop, weed, and background. Although focal loss values were slightly higher than in the binary crop-segmentation setting, the training curves indicated stable convergence of the multi-class model.

Table 8. Training and validation Focal Loss of the end-to-end multi-class U-Net_EfficientNet-B5 model.

U-Net Backbone	Training Focal Loss	Validation Focal Loss
EfficientNet-B5	0.0026	0.0031

Table 9 presents the quantitative comparison between the proposed two-stage framework and the end-to-end multi-class U-Net_EfficientNet-B5 baseline on the same hold-out test set, using macro-averaged metrics across the three classes (crop, weed, and background). As expected, the fully supervised end-to-end model achieved higher macro-averaged Dice, F1-score, recall, and IoU, reflecting the advantage of direct training with explicit weed annotations.

Table 9. Macro-averaged performance metrics (%) on the hold-out test set for the proposed two-stage method and the end-to-end multi-class baseline.

Model	Accuracy	Dice score	Precision	Recall	F1-score	IoU
Proposed two-stage method	99.44	87.05	81.59	72.54	75.95	65.39
End-to-end UNet Efficientnet-B5	99.72	93.99	94.72	86.33	90.12	82.89

Nevertheless, the proposed two-stage pipeline achieved useful performance (macro Dice = 87.05%, macro IoU = 65.39%) while relying only on crop masks and a simple HSV-based weed extraction step. The key contribution of the method is therefore not to outperform full supervision, but to provide an annotation-efficient alternative that reduces the need for dense weed labeling while preserving practical segmentation quality for weed mapping.

Confusion-matrix analysis (Figure 10) provides additional insight into this trade-off. In both approaches, crop pixels were segmented with high reliability, while the dominant error mode was confusion between weed and background. This confusion was stronger in the two-stage pipeline, especially at BBCH 15, where weak greenness and small weed size reduced the effectiveness of fixed HSV thresholding. At BBCH 19, however, both methods showed improved weed detection and reduced weed–background confusion as canopy development and weed greenness increased. Qualitative comparisons in Figure 11 support these trends and illustrate typical strengths and failure modes of both approaches across growth stages.

Overall, the two-stage framework and the end-to-end baseline produced visually consistent segmentations on many patches, in good agreement with the ground-truth masks. However, some failure cases highlight the specific limitations of the hybrid approach. For example, in patches a and d, small inaccuracies at crop boundaries in the first-stage crop segmentation led to thin edge regions being incorrectly labeled as weed after HSV-based detection, causing slight weed over-segmentation along crop contours. At BBCH 15 (patches g–h), the two-stage pipeline is further limited by weak greenness

contrast and the small size of both crop and weed plants, making early weeds harder to detect reliably. In contrast, at BBCH 19, and particularly in patch i, the two-stage pipeline performs comparatively better than the end-to-end model because the dense and highly green weed cover is more fully captured by the HSV-based greenness analysis, reducing weed under-segmentation. These observations indicate that the hybrid strategy is less effective under low-greenness early-growth conditions but can become advantageous in situations with strong weed presence and pronounced greenness, especially when the practical objective is weed-area mapping with reduced annotation cost.

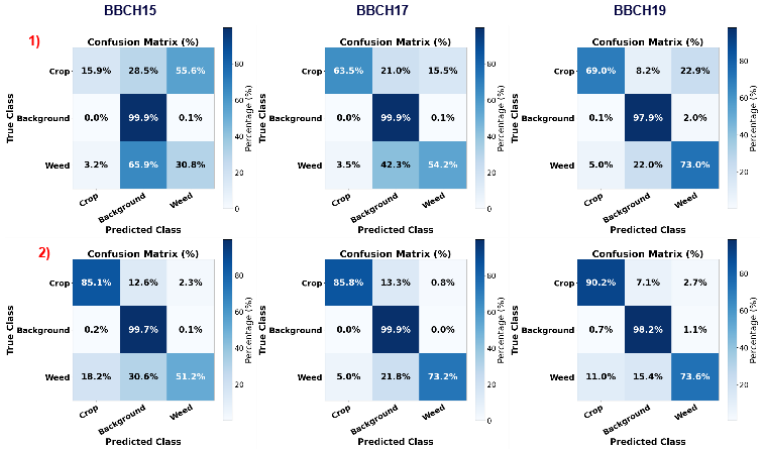

Figure 10. Normalized confusion matrices (%) showing pixel-wise classification results for the three classes (Background/soil = Background, sorghum canopy = Crop, and weeds = Weed) on the hold-out test set. Results are shown for sorghum at BBCH 15, BBCH 17, and BBCH 19 for both the end-to-end multi-class U-Net_EfficientNet-B5 baseline and the proposed two-stage framework.

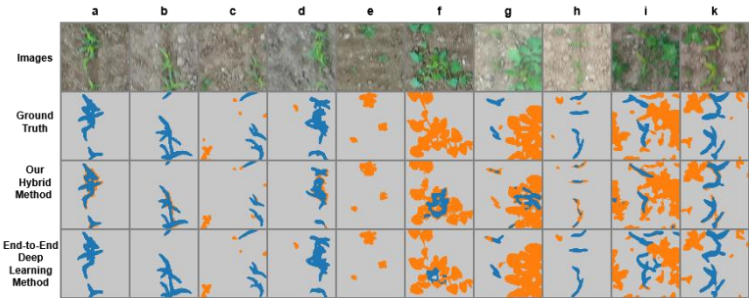

Figure 11. Qualitative segmentation results on the hold-out test set. Patch a–f show BBCH 17 examples (a–b: only crop, c–d: mixed crop and weeds, e–f: mainly weeds), patch g–h show BBCH 15, and patch i–k show BBCH 19. Background is shown in gray, crops in blue, and weeds in orange.

3.4 Comparison with the State of the Art

To position our two-stage framework within the broader literature, we compared it with recent UAV-based weed-segmentation studies that share our setup: RGB drone imagery, pixel-level segmentation, and U-Net variants. Table 10 reports the comparison using macro-averaged metrics across the three classes (crop, weed, background), so that all numbers refer to the same aggregation.

On the same sorghum dataset used here, Genze et al. (2022) trained a U-Net with a ResNet-34 encoder under full supervision and reported macro Dice = 99.69% and F1 = 89.37% on the hold-out test set. Their follow-up work (Genze et al., 2023) introduced DeBlurWeedSeg, a combined deblurring + segmentation model that improved the Sørensen–Dice coefficient by about 13.3% under motion-blurred conditions. Our own end-to-end baseline (U-Net with an EfficientNet-B5 encoder) reached macro Dice = 93.99%, F1 = 90.12%, and IoU = 82.89% on the same data, slightly above the F1 reported by Genze et al. (2022) (90.12% vs. 89.37%). Despite the differences in encoder and training protocol, our fully supervised baseline therefore matches the published state of the art on this dataset. The more interesting result is that our annotation-efficient two-stage pipeline reached macro F1 = 75.95% and IoU = 65.39% without using any pixel wise weed annotation. These numbers are lower than the fully supervised baselines, as one would expect, but they remain in the range needed for weed-area mapping and site-specific intervention and they are obtained without ever labelling the 4,470 weed instances present in the dataset.

We also found that our results on HSV vs. ExG vs. CIVE agree with what others have reported. Kawamura et al. (2021), in an independent UAV study on upland rice, compared the same three colour representations for crop/weed discrimination and concluded that HSV gave the best out-of-bag classification accuracy (0.904), ahead of RGB, CIE-Lab, ExG, and CIVE. We see the same pattern at the pixel-segmentation level: HSV reaches IoU = 68.66% in the binary weed–background setting, while ExG and CIVE drop below 20% because they over-segment the soil background. The fact that two independent studies converge on the same conclusion supports our choice of HSV for the weed-extraction step.

Table 10. Comparison of the proposed two-stage pipeline with recent UAV-based weed-segmentation studies. All values are macro-averaged across the three classes (crop, weed, background) on the hold-out test set, as reported by the original authors.

Study	Method	Annotation	Metrics (%)
Genze et al. (2022)	U-Net + ResNet-34	Full pixel-wise (3 classes)	Dice = 99.69; F1 = 89.37
Genze et al. (2023) – DeBlurWeedSeg	Deblur + U-Net	Full pixel-wise (3 classes)	+13.3% Dice vs. baseline
Kawamura et al. (2021)	SLIC + RF (HSV)	Full pixel-wise (3 classes)	OOB acc = 90.4
This work – end-to-end baseline	U-Net + EfficientNet-B5	Full pixel-wise (3 classes)	Dice = 93.99; F1 = 90.12; IoU = 82.89
This work – proposed two-stage	U-Net crop segm. + HSV weed	Crop masks only (no weed labels)	Dice = 87.05; F1 = 75.95; IoU = 65.39

The performance of the two-stage pipeline depends strongly on the sorghum growth stage. At BBCH 15, weeds are small, weakly pigmented, and often partly covered by soil residues, so a non-negligible fraction of weed pixels falls below the Value threshold used to filter out shadows. This is why weed–background confusion was strongest at BBCH 15 (Figure 10), and why the two-stage pipeline under-estimates vegetation cover at this stage. At BBCH 17 the weeds are fully developed, and their hue is firmly inside the green interval, so HSV thresholding works in its nominal regime. At BBCH 19, the weed canopy is dense and highly saturated, and the HSV filter actually captures weeds more completely than the end-to-end model Figure 11, patch i, shows a clear example where the two-stage prediction recovers weed clusters that the supervised model partially misses. This makes sense: HSV is a fixed rule, while the end-to-end network learns a probabilistic decision boundary that can over-suppress weeds

at growth stages under-represented in the training set (which here is dominated by BBCH 17).

The annotation effort saved by the two-stage approach can be quantified directly. The datasets used in this study contain 4,470 weed instances in total (3,060 at BBCH 17, 429 at BBCH 15, and 981 at BBCH 19; Table 1), and none of them needed pixel-wise weed labels only the sorghum crop masks were used to train the U-Net. Pixel-wise weed annotation of a single UAV image typically takes tens of minutes of expert time (Genze et al., 2022), so avoiding 4,470 such annotations is a real saving in human effort, and one that becomes more important as the system scales to larger plots or new sorghum varieties.

Finally, the two stages of our pipeline are not fully independent: errors in the first stage propagate into the second. Small inaccuracies in the U-Net crop mask at leaf boundaries (typically one- or two-pixel-thick under- or over-estimations, visible in Figure 11, patches a and d) survive the crop suppression step and are then re-labelled as weeds by the HSV rule, since the exposed edge pixels fall inside the green hue interval. This effect partly explains the gap between the two-stage macro IoU (65.39%) and the end-to-end baseline (82.89%), because the HSV filter has no way to tell a true weed pixel from a residual crop-boundary pixel. A simple post-processing step, for example, morphological erosion of the crop mask, or a confidence margin around crop contours could likely close part of this gap without bringing back the need for pixel-wise weed labels. We plan to test this in future work.

3.5 Measurement Analysis of Vegetation Cover

To complement the segmentation metrics, we performed a pixel-based analysis of vegetation cover using class areal fractions. In this measurement-oriented analysis, the measurands are the areal fractions of the three semantic classes (Background, Crop, and Weed) within the image domain. For each dataset configuration, class-wise pixel counts were computed from four sources: (i) the reference segmentation masks (ground truth), (ii) the end-to-end multi-class U-Net_EfficientNet-B5 predictions, (iii) the proposed two-stage framework predictions, and (iv) the HSV greenness indicator, which estimates total vegetation cover (Crop + Weed) by counting pixels within the fixed HSV interval [35, 50, 40]–[88, 255, 255]. To avoid trivial bias toward background-dominated scenes, only patches containing at least one crop or weed pixel in the reference segmentation were retained for this analysis. This filtering ensures that the estimated areal fractions provide a meaningful assessment of vegetation coverage and class-wise consistency across methods.

3.6 Overall Performance on the BBCH 17 Test Set

On the complete BBCH 17 hold-out test set (964 valid patches; 63,176,704 pixels), the reference masks indicate a strongly background-dominated scene (97.12% Background), with sparse vegetation (Crop = 1.97%, Weed = 0.90%; total vegetation = 2.87%). The end-to-end U-Net_EfficientNet-B5 baseline produced a similar global distribution (97.38% Background, 1.88% Crop, 0.73% Weed; total vegetation = 2.61%), and the two-stage framework also remained close to the reference (97.23% Background, 1.58% Crop, 1.19% Weed; total vegetation = 2.77%). The HSV greenness indicator measured 1,478,768 pixels as vegetation, corresponding to 2.34% of the image domain. Although this slightly underestimates the reference vegetation fraction (2.87%), it remains of the same order of magnitude and supports the use of fixed HSV thresholding as a reasonably unbiased indicator of total vegetation cover at BBCH 17. The close agreement among the reference masks, the end-to-end model, the two-stage framework, and the HSV-based indicator (Figure 12) therefore supports the measurement consistency of the proposed approach beyond macro-averaged classification metrics.

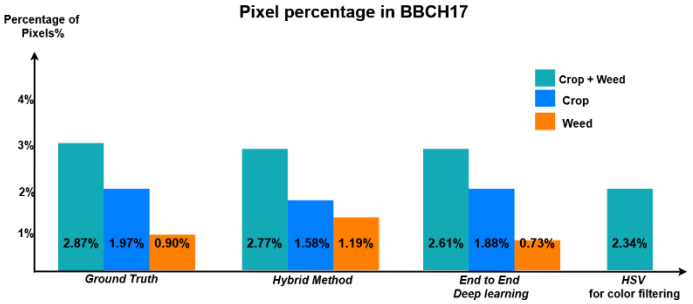

Figure 12. Pixel-wise areal fractions (%) of Crop and Weed on the BBCH 17 test set for the Ground Truth segmentation masks, the end-to-end multi-class U-Net_EfficientNet-B5 model, the proposed two-stage hybrid method, and the HSV greenness indicator (Crop + Weed inside the fixed HSV interval).

3.7 Measurement Analysis at Early Stage (BBCH 15)

At BBCH 15 (109 valid patches; 7,143,424 pixels), vegetation is less developed and the reference masks contain 96.03% Background, 1.64% Crop, and 2.33% Weed, corresponding to 3.97% total vegetation cover (Crop + Weed). The end-to-end model estimates 96.62% Background, 2.02% Crop, and 1.36% Weed (total vegetation = 3.38%), while the two-stage framework yields 97.81% Background, 0.35% Crop, and 1.84% Weed (total vegetation = 2.19%).

In this early-growth setting, the HSV greenness indicator detects only 154,003 pixels (2.16% of the image domain), which are substantially below the reference vegetation fraction (3.97%). This underestimation indicates that fixed HSV thresholding under-responds to early-stage vegetation with weak greenness and small spatial extent. The result is consistent with the lower weed-detection performance of the hybrid method at BBCH 15 and explains part of the increased weed–background confusion observed in Figure 10. From a measurement perspective, the HSV-based vegetation measurand is therefore less reliable at early growth stages unless thresholding is adapted to low-greenness conditions (Figure 13).

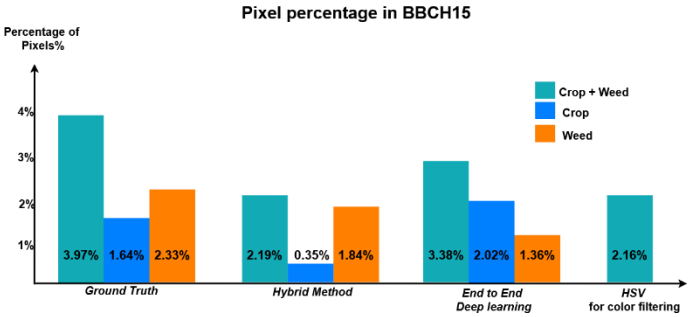

Figure 13. Pixel-wise areal fractions (%) of Crop and Weed on the BBCH 15 evaluation set for the reference masks (ground truth), the end-to-end multi-class U-Net_EfficientNet-B5 model, the proposed two-stage hybrid method, and the HSV greenness indicator. At this early growth stage, HSV underestimates the true vegetation fraction because both crop and weed exhibit limited greenness and small spatial extent.

3.8 Measurement Analysis at Late Stage (BBCH 19)

At BBCH 19 (110 valid patches; 7,208,960 pixels), weed pressure and canopy development are much higher. The reference masks contain 81.01% Background, 4.35% Crop, and 14.64% Weed, corresponding to 18.99% total vegetation cover (Crop + Weed). The end-to-end baseline produces 82.16% Background, 6.08% Crop, and 11.76% Weed (total vegetation = 17.84%), while the two-stage framework yields 82.86% Background, 3.81% Crop, and 13.33% Weed (total vegetation = 17.14%).

Under these conditions, the HSV greenness indicator detects 1,247,310 pixels (17.30% of the image domain), which is substantially closer to the reference vegetation fraction (18.99%) than in the BBCH 15 analysis. This improved agreement confirms that the fixed HSV interval becomes more reliable as canopy density and greenness contrast increase. Consistent with the confusion matrices and macro-averaged metrics (Figure 10; Table 9), the two-stage framework also achieves stronger weed detection at BBCH 19 than at BBCH 15. Overall, the results in Figure 14 show that the HSV-based vegetation measurand is stage-dependent and performs best under later growth conditions with dense, green canopies.

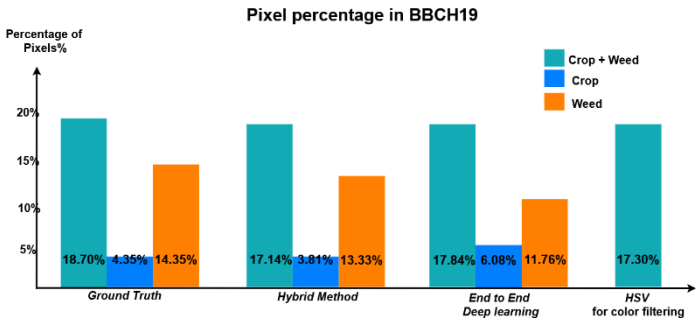

Figure 14. Pixel-wise areal fractions (%) of Crop and Weed on the BBCH 19 evaluation set for the reference masks, the end-to-end multi-class U-Net_EfficientNet-B5 model, the proposed two-stage hybrid method, and the HSV greenness indicator. At this late growth stage, the measured HSV vegetation percentage is close to the reference Crop+Weed fraction, leading to improved weed segmentation performance for the hybrid method.

4. Conclusions

This study proposed a two-stage framework for UAV-based weed mapping in sorghum fields that combines deep learning–based crop segmentation with HSV-based greenness analysis. By formulating the learning task as binary crop–background segmentation and extracting weeds in a second stage using a simple HSV rule, the method substantially reduces the need for labor-intensive pixel-wise weed annotation while simplifying model training and deployment. This design lowers computational and memory requirements, shortens training time, and facilitates implementation on standard computing platforms used in precision agriculture.

Among the evaluated backbones, EfficientNet-B5 was selected for crop segmentation based on the lowest validation Focal Loss (0.0016), and it also achieved the best segmentation quality (Accuracy: 99.66%, Dice: 90.59%), confirming the consistency between loss-based model selection and spatial performance. In the reconstructed three-class evaluation (background, crop, weed) on the BBCH 17 hold-out test set, the proposed two-stage framework achieved a macro-averaged Dice score of 87.05% and a macro-averaged IoU of 65.39%. Although the fully supervised end-to-end multi-class model provided higher overall segmentation performance, the proposed approach delivered practically useful weed maps while avoiding manual annotation of 4,470 weed instances across growth stages, representing a substantial reduction in expert labeling effort and cost.

The stage-wise analysis indicates that the framework is more reliable at mid to later growth stages (e.g., BBCH 17–19), when canopy density and vegetation greenness are sufficiently developed for HSV-based weed delineation. At earlier stages (particularly BBCH 15), performance decreases because greenness contrast is weaker, highlighting the importance of growth-stage awareness when applying greenness-based detection in field conditions.

This study also has important limitations. The framework was developed and evaluated on a limited number of annotated images from a single site and growing season, and the HSV weed-detection stage relied on a fixed threshold tuned on the main BBCH 17 dataset. Although patch-based training increased the number of training samples, it does not replace the diversity obtained from multi-site, multi-season acquisitions and broader weed communities. Therefore, the reported results should be interpreted as a controlled evaluation under the available field conditions rather than evidence of broad agronomic generalization.

Overall, the proposed framework provides a scalable and resource-efficient solution for site-specific weed management using UAV RGB imagery. It is suitable for applications such as targeted herbicide application, prioritization of manual scouting, and decision support for autonomous weeding systems. Future work will focus on improving robustness at early growth stages (especially BBCH 15), investigating adaptive or learned vegetation descriptors, validating the method on larger multi-site and multi-season datasets, and optimizing the framework for real-time or near real-time embedded deployment in UAV- or robot-assisted precision agriculture systems.

Acknowledgements

The authors would like to acknowledge the RCAM Laboratory, Djillali Liabes University (Sidi Bel Abbès, Algeria), for providing research facilities, technical assistance, and computational resources.

References

Ahmad, J., Muhammad, K., Ahmad, I., Ahmad, W., Smith, M. L., Smith, L. N., Smith, L. N., Jain, D. K., Wang, H., & Mehmood, I. (2018). Visual features based boosted classification of weeds for real-time selective herbicide sprayer systems. Computers in Industry, 98, 23–33. https://doi.org/10.1016/j.compind.2018.02.005

Aktar, M. W., Sengupta, D., & Chowdhury, A. (2009). Impact of pesticides use in agriculture: Their benefits and hazards. Interdisciplinary Toxicology, 2, 1–12. https://doi.org/10.2478/v10102-009-0001-7

Asad, M. H., & Bais, A. (2020). Weed detection in canola fields using maximum likelihood classification and deep convolutional neural network. Information Processing in Agriculture, 7, 535–545. https://doi.org/10.1016/j.inpa.2019.12.002

Chen, C., Yuan, X., Gan, S., Luo, W., Bi, R., Li, R., & Gao, S. (2024). A new vegetation index based on UAV for extracting plateau vegetation information. International Journal of Applied Earth Observation and Geoinformation, 128, 103668. https://doi.org/10.1016/j.jag.2024.103668

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. International Conference on Learning Representations. https://openreview.net/forum?id=YicbFdNTTy

Garibaldi-Márquez, F., Martínez-Barba, D. A., Montañez-Franco, L. E., Flores, G., & Valentín-Coronado, L. M. (2025). Enhancing site-specific weed detection using deep learning transformer architectures. Crop Protection, 190, 107075. https://doi.org/10.1016/j.cropro.2024.107075

Genze, N., Ajekwe, R., Güreli, Z., Haselbeck, F., Grieb, M., & Grimm, D. G. (2022). Deep learning-based early weed segmentation using motion blurred UAV images of sorghum fields. Computers and Electronics in Agriculture, 202, 107388. https://doi.org/10.1016/j.compag.2022.107388

Genze, N., Wirth, M., Schreiner, C., Ajekwe, R., Grieb, M., & Grimm, D. G. (2023). Improved weed segmentation in UAV imagery of sorghum fields with a combined deblurring segmentation model. Plant Methods, 19(1), 87. https://doi.org/10.1186/s13007-023-01060-8

Hasan, A. S. M. M., Sohel, F., Diepeveen, D., Laga, H., & Jones, M. G. K. (2021). A survey of deep learning techniques for weed detection from images. Computers and Electronics in Agriculture, 184, 106067. https://doi.org/10.1016/j.compag.2021.106067

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 770–778). https://doi.org/10.1109/CVPR.2016.90

Hess, M., Barralis, G., Bleiholder, H., Buhr, L., Eggers, T. H., Hack, H., & Stauss, R. (1997). Use of the extended BBCH scale – General for the descriptions of the growth stages of mono- and dicotyledonous weed species. Weed Research, 37(6), 433–441. https://doi.org/10.1046/j.1365-3180.1997.d01-70.x

Huang, G., Liu, Z., Van der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 2261–2269). https://doi.org/10.1109/CVPR.2017.243

Huang, T., Yang, G., & Tang, G. (1979). A fast two-dimensional median filtering algorithm. IEEE Transactions on Acoustics, Speech, and Signal Processing, 27(1), 13–18. https://doi.org/10.1109/TASSP.1979.1163188

Jin, X., Che, J., & Chen, Y. (2021). Weed identification using deep learning and image processing in vegetable plantation. IEEE Access, 9, 1666–1678. https://doi.org/10.1109/ACCESS.2021.3050296

Kataoka, T., Kaneko, T., Okamoto, H., & Hata, S. (2003). Crop growth estimation system using machine vision. In Proceedings of the 2003 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM 2003) (Vol. 2, pp. 1079–1083). https://doi.org/10.1109/AIM.2003.1225492

Kawamura, K., Asai, H., Yasuda, T., Soisouvanh, P., & Phongchanmixay, S. (2021). Discriminating crops/weeds in an upland rice field from UAV images with the SLIC-RF algorithm. Plant Production Science, 24(2), 198–215. https://doi.org/10.1080/1343943X.2020.1829490

Li X, Xu F, Liu F, Lyu X, Tong Y, Xu Z (2023). A synergistical attention model for semantic segmentation of remote sensing images. IEEE Transactions on Geoscience and Remote Sensing, 61, 1–16. https://doi.org/10.1109/TGRS.2023.3243954

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin Transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 10012–10022). https://doi.org/10.1109/ICCV48922.2021.00986

Macedo, F. L., Nóbrega, H., de Freitas, J. G. R., & Pinheiro de Carvalho, M. A. A. (2025). Assessment of vegetation indices derived from UAV imagery for weed detection in vineyards. Remote Sensing, 17(11), 1899. https://doi.org/10.3390/rs17111899

Olsen A, Konovalov DA, Philippa B, Ridd P, Wood JC, Johns J, Banks W, Girgenti B, Kenny O, Whinney J, Calvert B, Rahimi Azghadi M, White RD. (2019). DeepWeeds: A multiclass weed species image dataset for deep learning. Scientific Reports, 9, 3928. https://doi.org/10.1038/s41598-018-38343-3

Otsu, N. (1979). A threshold selection method from gray-level histograms. IEEE Transactions on Systems, Man, and Cybernetics, 9(1), 62–66. https://doi.org/10.1109/TSMC.1979.4310076

Patel, D. D., & Kumbhar, B. A. (2016). Weed and its management: A major threat to crop economy. Journal of Pharmaceutical Sciences and Bioscientific Research, 6, 453–758.

Rai, N., Zhang, Y., Ram, B. G., Schumacher, L., Yellavajjala, R. K., Bajwa, S., & Sun, X. (2023). Applications of deep learning in precision weed management: A review. Computers and Electronics in Agriculture, 206, 107698. https://doi.org/10.1016/j.compag.2023.107698

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In Proceedings of MICCAI (pp. 234–241). https://doi.org/10.1007/978-3-319-24574-4_28

Štroner, M., Urban, R., & Suk, T. (2023). Filtering green vegetation out from colored point clouds of rocky terrains based on various vegetation indices: Comparison of simple statistical methods, support vector machine, and neural network. Remote Sensing, 15(13), 3254. https://doi.org/10.3390/rs15133254

Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In Proceedings of the 36th International Conference on Machine Learning (ICML) (pp. 6105–6114). https://proceedings.mlr.press/v97/tan19a.html

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30. https://doi.org/10.48550/arXiv.1706.03762

Wang, A., Zhang, W., & Wei, X. (2019). A review on weed detection using ground-based machine vision and image processing techniques. Computers and Electronics in Agriculture, 158, 226–240. https://doi.org/10.1016/j.compag.2019.02.005

Woebbecke, D., Meyer, G., Von Bargen, K., & Mortensen, D. (1995). Shape features for identifying young weeds using image analysis. Transactions of the ASAE, 38, 271–281. https://doi.org/10.13031/2013.27839

World Health Organization. (1990). Public health impact of pesticides used in agriculture. Geneva, Switzerland: World Health Organization.

Wu, Z., Lin, M., Guo, L., & Wang, Y. (2021). Review of weed detection methods based on computer vision. Sensors, 21(11), 3647. https://doi.org/10.3390/s21113647

Yang, W., Wang, S., Zhao, X., Zhang, J., & Feng, J. (2015). Greenness identification based on HSV decision tree. Information Processing in Agriculture, 2(3–4), 149–160. https://doi.org/10.1016/j.inpa.2015.07.003

Zou, K., Chen, X., Zhang, F., Zhou, H., & Zhang, C. (2021). A field weed density evaluation method based on UAV imaging and modified U-Net. Remote Sensing, 13(2), 310. https://doi.org/10.3390/rs13020310