# WeedRepFormer: Reparameterizable Vision Transformers for Real-Time Waterhemp Segmentation and Gender Classification


Toqi Tahamid Sarker[0000−0003−2482−8059], Taminul Islam[0000−0002−7803−4564], Abdellah Lakhssassi[0009−0004−8978−3917], Cristiana Bernardi Rankrape[0009−0006−6026−7876], Kaitlin E. Creager[0009−0009−7115−2726], Karla Gage[0000−0002−6235−8804], and Khaled R. Ahmed[0000−0002−3707−4316]

Southern Illinois University Carbondale, Carbondale, IL, USA
`{toqitahamid.sarker, taminul.islam, abdellah.lakhssassi, cris.rankrape, kaitlin.creager, kgage, khaled.ahmed}@siu.edu`



**Abstract.** We present WeedRepFormer, a lightweight multi-task Vision Transformer designed for simultaneous waterhemp segmentation and gender classification. Existing agricultural models often struggle to balance the fine-grained feature extraction required for biological attribute classification with the efficiency needed for real-time deployment. To address this, WeedRepFormer integrates structural reparameterization across a Vision Transformer backbone, a RepLR-ASPP decoder, and a reparameterizable classification head, decoupling training-time capacity from inference-time latency. We also introduce a comprehensive waterhemp dataset containing 10,264 annotated frames from 23 plants. On this benchmark, WeedRepFormer achieves 92.18% mIoU for segmentation and 81.91% mAcc for gender classification using only 3.59M parameters and 3.80 GFLOPs. At 108.95 FPS, our model outperforms the state-of-the-art iFormer-T by 4.40% in classification accuracy while maintaining competitive segmentation performance and significantly reducing parameter count by 1.9×. Critically, WeedRepFormer attains the highest female classification accuracy of 76.27%, surpassing all baselines on the agriculturally consequential class that drives herbicide resistance spread. Materials are available at: https://github.com/toqitahamid/weedrepformer.



**Keywords:** Waterhemp · Semantic segmentation · Structural reparameterization · Multi-task learning.


## 1 Introduction

Waterhemp (*Amaranthus tuberculatus* (Moq.) Sauer) has emerged as one of the most troublesome weed species in North American agriculture, particularly threatening corn and soybean production across the Midwestern United States [28]. This dioecious species exhibits remarkable adaptability, with separate male and female plants requiring cross-pollination, generating extensive genetic diversity that facilitates rapid evolution of herbicide resistance [17,6]. Currently, waterhemp populations have developed confirmed resistance to seven herbicide sites of action (SOAs), with an eighth resistance to glufosinate (SOA Group 10) currently being confirmed [11,21]. This eighth resistance is particularly concerning as glufosinate represents the last remaining postemergence herbicide option in soybean production [21]. The severity of this problem is compounded by the species' exceptional fecundity, with female plants producing between 300,000 to over 2 million seeds per plant [10], contributing to rapid population expansion and yield losses up to 74% in corn [29] and 43-73% in soybean [9] under heavy infestation.

The dioecious nature of waterhemp presents both a challenge and an opportunity for precision weed management. Unlike monoecious weeds, waterhemp requires both male and female plants for reproduction, with only female plants producing seeds. Male plants produce wind-borne pollen, leading to extensive gene flow that carries herbicide resistance genes up to 800 m from source plants [15]. Recent studies have shown that herbicide exposure can alter population sex ratios [25], with PPO-inhibitor treatments shifting male-to-female proportions in ways that could affect seedbank composition and resistance evolution. These dynamics suggest that targeted removal of female plants before seed set could significantly reduce population growth and slow herbicide resistance evolution. However, manual identification of waterhemp gender in field conditions is labor-intensive and impractical at scale, while visual differentiation between male and female plants requires expertise, as morphological differences are subtle and vary with growth stage. Automated gender classification through computer vision could enable selective herbicide application or mechanical removal strategies, potentially reducing chemical inputs while improving long-term management efficacy.

Recent advances in deep learning have demonstrated remarkable success in agricultural computer vision tasks, with semantic segmentation methods achieving state-of-the-art performance in crop-weed discrimination and identifying weed plant species [27,14]. Multi-task learning frameworks that jointly optimize related tasks through shared representations have shown improved efficiency and generalization compared to single-task approaches [24]. For waterhemp management, simultaneous segmentation and gender classification are naturally related tasks that benefit from shared feature learning, as both require understanding of plant morphology and spatial structure.

However, practical deployment of deep learning models for agricultural robotics faces significant constraints. Real-time operation requires efficient processing on resource-constrained edge devices, such as those mounted on autonomous tractors or unmanned aerial vehicles. While Vision Transformers have demonstrated superior performance in semantic segmentation due to their global receptive fields [34], their computational cost has limited adoption in resource-constrained agricultural scenarios. Standard models like DeepLabV3+ [2] and U-Net [23] achieve high accuracy but require substantial computational resources. Recent lightweight architectures using depthwise separable convolutions, neural architecture search, or knowledge distillation often sacrifice accuracy for speed or require complex training procedures.

Structural reparameterization offers a promising alternative, enabling networks to train with multiple parallel branches for increased capacity while deploying as efficient single-path models through algebraic fusion [4,31]. This approach has achieved state-of-the-art efficiency in image classification and detection [30], and has been successfully applied to agricultural CNN-based architectures for disease detection, weed identification, and multi-task learning [19,8]. However, to the best of our knowledge, structural reparameterization has not yet been applied to Vision Transformers for multi-task dense prediction combining segmentation with fine-grained biological attribute classification in agriculture.

We propose WeedRepFormer, a Vision Transformer architecture that achieves an optimal balance of accuracy and efficiency for simultaneous waterhemp segmentation and gender classification via systematic structural reparameterization. In summary, our contributions are as follows:

- We propose a fully reparameterizable multi-task Vision Transformer that systematically applies structural reparameterization across backbone, segmentation head, and classification head.
- We introduce a waterhemp dataset with 10,264 annotated frames from 23 plants, including pixel-level segmentation masks and plant-level gender labels, and establish baseline results for this task.

## 2 Related Work

**Weed Detection and Segmentation.** Deep learning has revolutionized weed detection in precision agriculture. Early work adapted general semantic segmentation architectures such as U-Net [23] for agricultural applications, with successful deployment for real-time crop-weed classification [16]. Recent work has explored lightweight architectures for edge devices [33], but these methods primarily address binary crop-weed segmentation without species-specific biological attribute classification.

**Vision Transformers for Segmentation.** Vision Transformers [5] capture long-range dependencies through self-attention but incur high computational costs. Hierarchical variants address this limitation, with SegFormer [34] combining efficient Transformer encoders with lightweight MLP decoders for scalable dense prediction. Such architectures have shown success in general computer vision and are increasingly adopted for crop disease detection and weed species classification, though their application to fine-grained multi-task agricultural problems remains limited.

**Multi-Task Learning.** Multi-task learning (MTL) improves efficiency by sharing representations across related tasks [24]. In agriculture, MTLSegFormer [7] performs multi-task semantic segmentation with attention-based feature sharing between tasks, while WeedSense [27] jointly performs semantic segmentation, height estimation, and growth stage classification. These methods demonstrate the benefits of shared feature learning for related agricultural tasks.

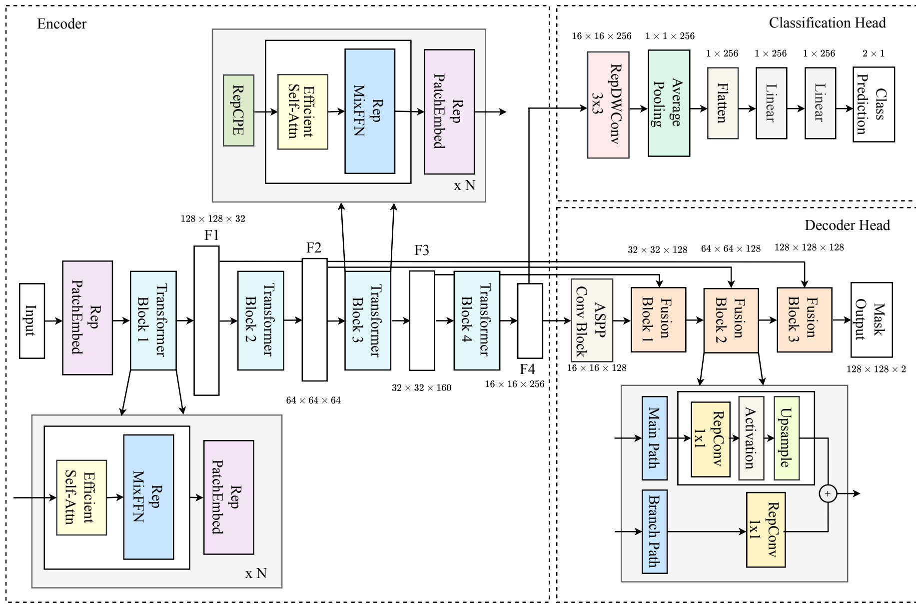
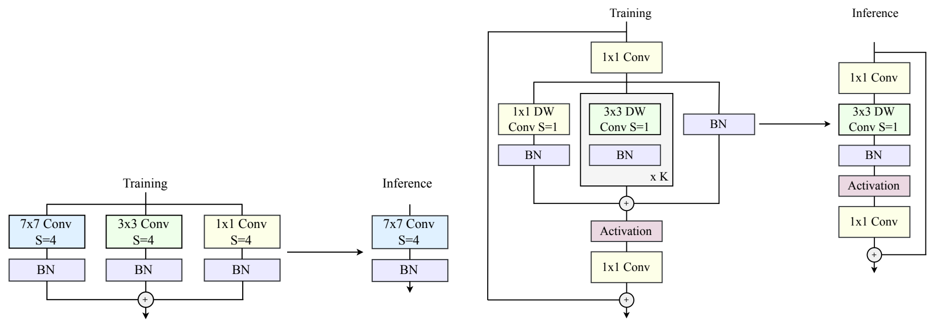



**Fig. 1.** Overview of our multi-task architecture. (a) Network consists of four-stage hierarchical Vision Transformer backbone with reparameterizable components. (b) Reparameterizable patch embedding with multi-branch convolutions. (c) RepLR-ASPP decoder with reparameterizable convolutions. (d) Classification head with optional SE attention.

**Structural Reparameterization.** Structural reparameterization enables multi-branch training with single-path inference through algebraic fusion. RepVGG [4] pioneered this approach for CNNs, MobileOne [31] achieved millisecond-scale mobile latency, and FastViT [30] extended the technique to Vision Transformers achieving significant speedups over hybrid architectures. In agriculture, reparameterization has been successfully applied to CNN-based architectures for disease detection [19] and weed detection [8]. However, existing agricultural applications primarily use CNN-based backbones such as YOLO and RepVGG variants. We extend this paradigm by introducing reparameterizable Vision Transformers for joint segmentation and gender classification, combining the global receptive fields of transformers with the efficiency benefits of structural reparameterization.

## 3 Method

We propose WeedRepFormer, a fully reparameterizable architecture for real-time waterhemp segmentation and gender classification (Fig. 1). During training, every block uses multiple parallel branches. At inference, the branches fuse into a single path with no extra cost [4,31]. This gives the classification head extra capacity at training time without slowing inference. The architecture has three parts: a




**(a)** RepPatchEmbed **(b)** RepMixFFN

**Fig. 2.** Reparameterizable components with train-time overparameterization. (a) RepPatchEmbed uses three parallel branches that fuse at inference. (b) RepMixFFN employs $K$ parallel depthwise convolutions with identity connection.

Vision Transformer backbone with RepPatchEmbed, RepCPE, and RepMixFFN; a RepLR-ASPP decoder; and a reparameterizable classification head.

### 3.1 Structural Reparameterization

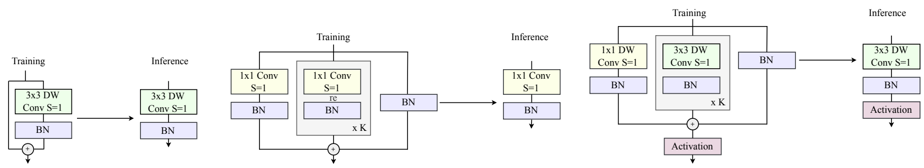
Each reparameterizable block has $K+2$ parallel branches at training time. There are $K$ depthwise convolutions of kernel size $k$ with weights $\{\mathbf{W}_m\}_{m=1}^{K}$, one $1\times1$ depthwise branch with weight $\mathbf{W}_{1\times1}$, and an identity branch represented as a centered Dirac kernel $\mathbf{W}_{\mathrm{id}}$. Each branch is followed by BatchNorm. At inference, we fold the BN parameters into each branch and sum the kernels into a single $k\times k$ depthwise convolution:

$$\mathbf{W}_{\mathrm{deploy}} = \gamma \odot \Big( \sum_{m=1}^{K} \widetilde{\mathbf{W}}_m + \mathrm{pad}_{k\times k}(\widetilde{\mathbf{W}}_{1\times1}) + \widetilde{\mathbf{W}}_{\mathrm{id}} \Big), \tag{1}$$

Here $\widetilde{\mathbf{W}}$ is the BN-folded weight [4]. The operator $\mathrm{pad}_{k\times k}(\cdot)$ zero-pads a $1\times1$ kernel to size $k\times k$. The vector $\gamma \in \mathbb{R}^C$ is the layer scale (initialized to $10^{-5}$), which we absorb into the fused weights. Biases are summed in the same way. The result is a single convolution at inference with no overhead compared to the post-fusion topology.

### 3.2 Vision Transformer Backbone

We extend the hierarchical Mix Transformer (MiT) backbone from SegFormer [34]. We replace its standard components with reparameterizable ones. The backbone has four stages. Stage $i$ produces a feature map $\mathbf{F}_i$ at resolution $H/2^{i+1} \times W/2^{i+1}$, with channel dimensions $[32, 64, 160, 256]$ across the four stages. We use efficient multi-head self-attention with spatial reduction ratios $r \in \{8, 4, 2, 1\}$ to keep




**(a)** RepCPE **(b)** RepConv1x1 **(c)** RepDWConv3x3

**Fig. 3.** Reparameterizable convolution modules used in positional encoding, decoder, and classifier. (a) RepCPE combines depthwise convolution with identity for positional encoding. (b) RepConv1x1 with $K$ parallel $1{\times}1$ branches for efficient channel mixing. (c) RepDWConv3x3 with $K$ parallel $3{\times}3$ depthwise branches for spatial feature refinement.

complexity low at high resolutions. We inherit the MiT-B0 MLP ratio of 4 and apply stochastic depth with rate 0.1.

**RepPatchEmbed.** RepPatchEmbed reparameterizes the patch embedding layer, following MobileOne [31] (Fig. 2a). At training time, it uses three parallel branches: a large-kernel convolution of size $p_i{\times}p_i$, a $3{\times}3$ convolution, and a $1{\times}1$ convolution. We set $p_i \in \{7, 3, 3, 3\}$ for stages 1–4, validated in Sect. 5.2. The multi-scale branches help the model learn diverse low-level features. At inference, the branches fuse into a single $p_i \times p_i$ convolution via Eq. 1.

**RepCPE.** Conditional positional encodings (CPE) [3] generate position information dynamically from local context, unlike absolute encodings. We apply reparameterizable CPE [30] only in the deeper stages 3 and 4 (Fig. 3a). The reason is that positional context helps semantic-level discrimination more than low-level texture; the ablation in Sect. 5.2 confirms this. RepCPE adds a $3 \times 3$ depthwise convolution to its input: $\mathrm{RepCPE}(\mathbf{F}) = \mathbf{F} + \mathrm{DWConv}_{3\times3}(\mathbf{F})$. At inference, the residual and the convolution fuse into a single $3 \times 3$ depthwise convolution via Eq. 1.

**RepMixFFN.** RepMixFFN enhances the standard MixFFN with a multi-branch depthwise convolution (Fig. 2b). It applies a $1{\times}1$ channel expansion $\mathbf{W}_{\mathrm{fc}_1}$ with MLP ratio 4, then a reparameterizable depthwise block, then a GELU activation, then a $1{\times}1$ projection $\mathbf{W}_{\mathrm{fc}_2}$. The whole module is wrapped in a DropPath residual:

$$\mathrm{RepMixFFN}(\mathbf{F}) = \mathbf{F} + \mathrm{DropPath}\Big(\mathbf{W}_{\mathrm{fc}_2}\,\sigma\big(\mathrm{RepConv}_{3\times3}(\mathbf{W}_{\mathrm{fc}_1}\mathbf{F})\big)\Big), \quad (2)$$

Here $\sigma$ is GELU and $\mathrm{RepConv}_{3\times3}$ is the $K{+}2$-branch block defined by Eq. 1. We set $K{=}2$, chosen by ablation (Sect. 5.2). Layer scale initialized to $10^{-5}$ stabilizes training.

### 3.3 Segmentation Decoder and Classification Head

**RepLR-ASPP Decoder.** For multi-scale feature aggregation, we adapt Lite R-ASPP [12]. Lite R-ASPP is a lightweight version of ASPP [1] that uses global


**Fig. 4.** Example image-mask pairs from the waterhemp dataset. Left pair shows a female specimen, right pair shows a male specimen. Binary segmentation masks delineate complete plant architecture against background.

pooling in place of dilated convolutions. We replace every convolution in the decoder with a RepConv1x1 block (Fig. 3b). Each RepConv1x1 has $K$ parallel $1{\times}1$ branches with layer scale, which fuse at deployment. The decoder runs in three steps. First, we apply global context attention to $\mathbf{F}_4$ via global average pooling and sigmoid gating. Next, we upsample the features stage by stage. At each stage, we fuse the upsampled features with the lateral features from $\{\mathbf{F}_1, \mathbf{F}_2, \mathbf{F}_3\}$ using channel concatenation followed by RepConv1x1. The final output is the segmentation prediction.

**RepClsHead.** RepClsHead is the gender classification head in our joint segmentation and classification framework. We adapt the reparameterizable refinement block from FastViT [30] (Fig. 3c). The block uses $K$ parallel $3 \times 3$ depthwise convolutions, a $1 \times 1$ depthwise branch, and an identity branch, all with layer scale. At inference, the block fuses to a single $3 \times 3$ depthwise convolution via Eq. 1. RepClsHead differs from FastViT's classifier in two ways. First, we replace FastViT's single linear classifier with a two-layer MLP (hidden dim 256, ReLU, dropout 0.5). The two-layer MLP lets the head re-encode features that were trained mostly for segmentation into features useful for gender classification. Second, the head is trained jointly with the segmentation loss (Sect. 3.3). Our ablation (Sect. 5.2) shows this design adds +4.19 pp absolute mAcc over a plain GAP+MLP head at the same FLOPs.

**Multi-Task Loss.** We train the network on both tasks at once with the multi-task loss

$$\mathcal{L} = \mathcal{L}_{\text{seg}} + \lambda \mathcal{L}_{\text{cls}}, \qquad \lambda = 0.5.$$

Here $\mathcal{L}_{\text{seg}}$ is pixel-wise cross-entropy over background and plant classes, and $\mathcal{L}_{\text{cls}}$ is binary cross-entropy for gender classification.

## 4 Dataset

We introduce a novel waterhemp plant dataset designed for joint semantic segmentation and gender classification tasks. The dataset comprises 23 individual

**Table 1.** Waterhemp dataset statistics. Percentages show distribution within 10,264 annotated frames.

| Gender Type | Frames | Percentage | Plants | Train (70.5%) | Val (15.5%) | Test (13.9%) |
|---|---|---|---|---|---|---|
| Male | 4,158 | 40.5% | 10 | 2,692 | 780 | 686 |
| Female | 6,106 | 59.5% | 13 | 4,544 | 816 | 746 |
| **Total** | **10,264** | **100.0%** | **23** | **7,236** | **1,596** | **1,432** |

mature waterhemp plants, specifically 13 females and 10 males identified with the assistance of an agricultural expert, which were collected from a field site and transplanted into pots for imaging. All plants were at the flowering stage, exhibiting fully developed morphological characteristics essential for gender differentiation. The complete dataset statistics are presented in Table 1. Representative samples from the dataset are shown in Fig. 4.

**Plant Collection.** Waterhemp plants were collected from a soybean field at SIUC University Farms on October 14, 2024 and transplanted into 6-inch diameter plastic pots for imaging. The field had been fertilized with 200 lbs per acre of potash in spring 2024. Soybeans (Pioneer P37A18E) were planted on June 13, 2024 at 180,000 seeds per acre with 15-inch row spacing. The field was tilled at planting to remove weed competition. A single early postemergence herbicide application was made on August 12, 2024, consisting of glufosinate (Liberty, 32 oz/acre), glyphosate (Roundup Powermax 3, 1 qt/acre), citric acid (2.4 oz/acre), and ammonium sulfate (2.4 lb/acre). Soybeans were harvested on October 21, 2024 with a yield of approximately 25 bu/acre.

**Data Collection and Preprocessing.** We captured one video per plant using an iPhone 15 Pro Max, performing a complete 360-degree horizontal rotation to ensure comprehensive morphological coverage. To balance dataset size with temporal diversity, we sample every third frame, effectively reducing redundancy while preserving motion variation. Frames are resized to $720 \times 960$ pixels, maintaining sufficient spatial resolution for fine-grained structural analysis while halving storage requirements. The final dataset comprises 10,264 frames across 23 plants.

**Annotation Protocol.** We employ SAM2.1 Hiera Large [22] for efficient semi-automatic mask generation. For each video sequence, we manually annotate positive and negative prompt points on the initial frame using an interactive bounding box interface. SAM2.1 then propagates these sparse annotations across all subsequent frames via its video predictor, ensuring temporal consistency. The resulting binary masks delineate the complete plant structure (stems, leaves, flowers) as foreground, with all other regions classified as background. Gender labels are assigned at the plant level based on botanical verification during the flowering stage.

**Dataset Split Strategy.** To prevent data leakage, we partition the dataset at the plant level rather than the image level. Specifically, all frames extracted from a single plant are assigned exclusively to one split—training, validation, or test. This ensures that no visual information from an individual plant appears

**Table 2.** Comparison with state-of-the-art methods on waterhemp multi-task learning. All models trained with identical protocols. Best results in **bold**, second best underlined.

| Method | Segmentation | | Classification | | Efficiency | | |
|---|---|---|---|---|---|---|---|
| | mIoU (%)↑ | mFscore (%)↑ | mAcc (%)↑ | mF1 (%)↑ | FPS↑ | GFLOPs↓ | Params (M)↓ |
| *Conv-based Models* | | | | | | | |
| BiSeNet [36] | 92.27 | 95.87 | 67.25 | 66.87 | <u>233.45</u> | 14.821 | 13.455 |
| BiSeNet V2 [35] | 92.30 | 95.89 | 58.17 | 57.96 | 159.61 | 12.286 | 14.820 |
| DDRNet [18] | 92.81 | 96.18 | 45.74 | 44.94 | 151.58 | 4.560 | 5.766 |
| Fast-SCNN [20] | 88.27 | 93.51 | 59.57 | 59.41 | 225.70 | **0.927** | **1.488** |
| ICNet [37] | 91.58 | 95.48 | 56.08 | 56.02 | 134.73 | 15.426 | 47.859 |
| MobileNetV2 [26] | 91.50 | 95.43 | 70.88 | 69.34 | 179.47 | 39.261 | 9.793 |
| MobileNetV3 [12] | 92.28 | 95.88 | 68.99 | 67.92 | 121.55 | 8.692 | 3.529 |
| MobileOne-S0 [31] | 73.02 | 82.76 | 60.61 | 59.23 | **296.97** | 8.213 | 28.915 |
| RepViT-M0.9 [32] | **94.31** | **97.01** | 69.20 | 68.86 | 82.23 | 25.404 | 8.954 |
| *Transformer-based Models* | | | | | | | |
| SegFormer-B0 [34] | <u>94.10</u> | <u>96.90</u> | 74.72 | 75.34 | 115.27 | 7.885 | 3.782 |
| iFormer-T [38] | 94.05 | 96.87 | <u>77.51</u> | <u>77.08</u> | 99.05 | 24.267 | 6.804 |
| FastViT-T8 [30] | 93.33 | 96.47 | 55.24 | 54.41 | 189.71 | 23.806 | 7.382 |
| **WeedRepFormer (Ours)** | 92.18 | 95.82 | **81.91** | **81.90** | 108.95 | <u>3.801</u> | <u>3.592</u> |

across multiple splits, thereby providing a more realistic evaluation of model generalization to unseen specimens. Unlike image-level splitting, which can inadvertently leak visual patterns across splits when consecutive frames share similar appearances, our plant-level strategy guarantees complete independence between training and evaluation data.

The plant-level split allocation is as follows: 15 plants for training (9 female, 6 male), 4 plants for validation (2 female, 2 male), and 4 plants for testing (2 female, 2 male). Table 1 summarizes the resulting frame-level distribution across splits. This yields a split ratio of approximately 70.5% for training, 15.5% for validation, and 13.9% for testing. The dataset exhibits a natural gender imbalance, with female plants contributing 59.5% of total frames. This distribution reflects inherent biological differences: female waterhemp plants typically develop denser foliage and more complex branching structures compared to their male counterparts, resulting in longer video sequences and more extracted frames per plant.

## 5 Experiments

**Implementation Details.** All models are trained for 80K iterations using AdamW optimizer with learning rate $6 \times 10^{-5}$, weight decay 0.01, and momentum $(0.9, 0.999)$. The learning rate follows a polynomial decay (power 1.0) after a 1,500-iteration linear warmup with start factor $10^{-6}$. We use batch size 8 on an NVIDIA A100 GPU at $512 \times 512$ resolution with standard augmentations including random horizontal flip, resize, and photometric distortion. The multi-task loss uses $\lambda = 0.5$ as described in Sect. 3.3. For fair comparison, all comparison methods utilize a standardized MLP classification head comprising global pooling followed by two linear layers with Xavier initialization, except Fast-SCNN which loads Cityscapes weights [20]. Our WeedRepFormer employs the proposed RepClsHead and initializes from SegFormer-B0 ImageNet weights [34], while reparameterizable modules inherit weights into their primary branch matching the source kernel size,

**Table 3.** Class-wise performance breakdown.

| Method | Segmentation | | | | Classification | | | |
|---|---|---|---|---|---|---|---|---|
| | IoU (%)↑ | | F-score (%)↑ | | Acc (%)↑ | | F1 (%)↑ | |
| | BG | Weed | BG | Weed | Male | Female | Male | Female |
| *Conv-based Models* | | | | | | | | |
| BiSeNet | 98.48 | 86.06 | 99.23 | 92.51 | 81.34 | 54.29 | 70.41 | 63.33 |
| BiSeNet V2 | 98.51 | 86.10 | 99.25 | 92.53 | 68.08 | 49.06 | 60.93 | 55.00 |
| DDRNet | 98.63 | 86.99 | 99.31 | 93.05 | 60.35 | 32.31 | 51.59 | 38.28 |
| Fast-SCNN | 97.59 | 78.95 | 98.78 | 88.24 | 55.69 | 63.14 | 56.89 | 61.93 |
| ICNet | 98.34 | 84.82 | 99.16 | 91.79 | 62.10 | 50.54 | 57.53 | 54.52 |
| MobileNetV2 | 98.34 | 84.66 | 99.16 | 91.69 | **97.38** | 46.51 | 76.21 | 62.47 |
| MobileNetV3 | 98.51 | 86.05 | 99.25 | 92.50 | 91.11 | 48.66 | 73.79 | 62.05 |
| MobileOne-S0 | 93.56 | 52.49 | 96.67 | 68.84 | 82.51 | 40.48 | 66.75 | 51.71 |
| RepViT-M0.9 | **98.93** | **89.69** | **99.46** | **94.56** | 83.24 | 56.30 | 72.14 | 65.57 |
| *Transformer-based Models* | | | | | | | | |
| SegFormer-B0 | 98.88 | 89.31 | 99.44 | 94.36 | 90.09 | 60.59 | 77.35 | 71.41 |
| iFormer-T | 98.87 | 89.24 | 99.43 | 94.31 | 95.19 | 61.26 | 80.22 | 73.95 |
| FastViT-T8 | 98.73 | 87.93 | 99.36 | 93.58 | 71.72 | 40.08 | 60.55 | 48.26 |
| **WeedRepFormer** | 98.45 | 85.91 | 99.22 | 92.42 | 88.05 | **76.27** | **82.34** | **81.46** |

with additional branches using Kaiming initialization. The backbone learning rate is scaled by 0.1× while the classification head uses 1.0× the base rate. We report mean IoU (mIoU) and mean F-score (mFscore) for segmentation, mean accuracy (mAcc) and mean F1 (mF1) for classification, and inference FPS.

### 5.1 Comparison with State-of-the-Art Methods

We compare our WeedRepFormer against state-of-the-art efficient segmentation and multi-task models on the waterhemp dataset. Table 2 presents comprehensive results across segmentation, classification, and efficiency metrics.

**Classification Performance.** WeedRepFormer achieves 81.91% mAcc and 81.90% mF1, outperforming all methods on gender classification. It improves over the second-best iFormer-T by 4.40% mAcc and 4.82% mF1. Convolution-based models struggle on this task despite competitive segmentation. DDRNet, for example, reaches only 45.74% mAcc. These results suggest that structural reparameterization captures the discriminative features needed for fine-grained classification.

**Segmentation Performance.** WeedRepFormer yields 92.18% mIoU and 95.82% mFscore. The top segmenters RepViT-M0.9 and SegFormer-B0 gain only 1.9–2.1 mIoU points over our model. These gains come at a steep cost. RepViT-M0.9 needs 6.7× our FLOPs and loses 12.71 mAcc points on classification, while SegFormer-B0 needs 2.1× our FLOPs and loses 7.19 mAcc points. Among efficient models under 5 GFLOPs, WeedRepFormer matches DDRNet's segmentation while delivering 36.17 mAcc points more on classification.

**Efficiency Analysis.** With only 3.59M parameters and 3.80 GFLOPs, WeedRepFormer runs at 108.95 FPS while maintaining high multi-task accuracy. It reduces computational cost by 51.7% and parameter count by 5.0% over the widely used SegFormer-B0 with negligible impact on throughput. Fast-SCNN is

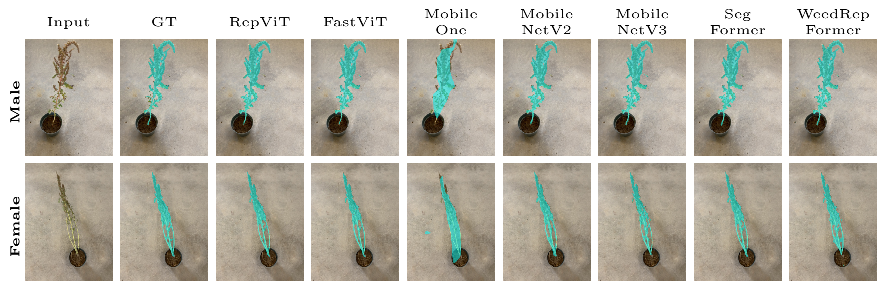



**Fig. 5.** Qualitative comparison on male and female waterhemp. Best viewed on screen.

lighter but suffers significant accuracy degradation, while higher-capacity models like iFormer-T and RepViT-M0.9 incur 6.4× and 6.7× more computation. Our reparameterization further yields a 1.54× inference speedup, collapsing from 4.63 GFLOPs during training to 3.80 GFLOPs at deployment. These results position WeedRepFormer for resource-constrained agricultural robotics.

**Class-wise Performance Analysis.** Table 3 reveals a pattern hidden by aggregate metrics. Most baselines exhibit severe male-female accuracy disparities. MobileNetV2 attains the highest male accuracy (97.38%) yet collapses to 46.51% on females, a gap of 50.87 points. iFormer-T, SegFormer-B0, and DDRNet show gaps of 33.93, 29.50, and 28.04 points. In contrast, WeedRepFormer achieves the best female accuracy (76.27%) and F1-score (81.46%) with the smallest gap of just 11.78 points. This balance is agriculturally consequential. Female waterhemp plants produce the seeds that drive herbicide resistance spread, so a male-biased detector is operationally useless regardless of its mean accuracy.

**Qualitative Results.** Fig. 5 illustrates segmentation results on representative male and female Waterhemp samples. MobileOne produces degraded segmentation with coarse boundaries and missed regions, whereas WeedRepFormer matches the visual quality of higher-capacity models like RepViT-M0.9 and SegFormer-B0, confirming that structural reparameterization preserves segmentation quality.

### 5.2 Ablation Study

We conduct a systematic ablation to validate architectural decisions. All experiments utilize the MixVisionTransformer backbone and RepLR-ASPP decoder unless otherwise specified. Results are summarized in Table 4.

**Reparameterizable Backbone Components.** We first evaluate the impact of including RepMixFFN, RepPatchEmbed, and RepCPE in the backbone. As shown in Table 4, RepCPE alone achieves the highest segmentation but suboptimal classification. Combining all three components achieves the best multi-task balance (92.15% mIoU, 77.09% mAcc), improving classification by +9.0% mAcc over RepCPE alone with only a minor segmentation trade-off.

**Classification Head Design.** We replace the standard MLP head with our proposed RepClsHead. Even with a single branch ($K = 1$), accuracy improves

**Table 4.** Comprehensive ablation study of architectural components. Gray rows indicate selected configurations.

| Configuration | mIoU | mAcc |
|---|---|---|
| *Reparameterizable Backbone Components* | | |
| RepMixFFN | 92.76 | 62.85 |
| RepPatchEmbed | 91.93 | 61.24 |
| RepCPE | 93.74 | 68.09 |
| RepCPE + RepMixFFN | 92.72 | 52.72 |
| RepCPE + RepPatchEmbed | 91.66 | 66.27 |
| RepMixFFN + RepPatchEmbed | 91.89 | 73.95 |
| All Components | 92.15 | 77.09 |
| *Classification Head Design* | | |
| $K$=1, Standard MLP | 92.15 | 77.09 |
| $K$=1, RepClsHead | 92.32 | 79.19 |
| $K$=2, Standard MLP | 92.25 | 77.72 |
| $K$=2, RepClsHead | 92.18 | **81.91** |
| *Parallel Branch Count ($K$)* | | |
| $K$=1, RepClsHead | 92.32 | 79.19 |
| $K$=2, RepClsHead | 92.18 | 81.91 |
| $K$=3, RepClsHead | 92.25 | 75.56 |
| $K$=4, RepClsHead | 92.17 | 78.84 |

| Configuration | mIoU | mAcc |
|---|---|---|
| *RepPatchEmbed Factorization* | | |
| Standard Conv, $K$=1, Standard MLP | 92.15 | 77.09 |
| Standard Conv, $K$=1, RepClsHead | 92.32 | 79.19 |
| Standard Conv, $K$=2, RepClsHead | 92.18 | 81.91 |
| DW-PW Conv, $K$=1, Standard MLP | 90.14 | 59.01 |
| DW-PW Conv, $K$=1, RepClsHead | 90.30 | 59.01 |
| DW-PW Conv, $K$=2, RepClsHead | 88.58 | 56.22 |
| *CPE Placement (Stages 1–4)* | | |
| (T, T, T, T) | 91.98 | 69.90 |
| (T, T, F, F) | 92.03 | 65.78 |
| (F, F, T, T) | 92.18 | 81.91 |
| (T, F, T, F) | 92.06 | 68.09 |
| (F, T, F, T) | 92.35 | 73.46 |
| *RepPatchEmbed Kernel Size* | | |
| (7, 3, 3, 3) | 92.18 | 81.91 |
| (7, 7, 7, 7) | 92.08 | 81.08 |
| *SE in Classification Head* | | |
| Without SE | 92.18 | 81.91 |
| With SE | 92.12 | 76.19 |

by +2.1%. With $K = 2$ branches, RepClsHead achieves 81.91% mAcc versus 77.72% for the simple baseline. This shows that the classification head benefits from the increased capacity of multi-branch training, which is then compressed for inference.

**Branch Count ($K$) & Efficiency.** We investigate the training-time width of the reparameterizable blocks ($K \in \{1, 2, 3, 4\}$), observing a performance peak at $K = 2$. Increasing to $K = 3$ or $K = 4$ leads to diminishing returns and overfitting, degrading accuracy to 75–78%. The $K = 2$ configuration incurs computational cost only during training. It collapses to the same inference cost as the $K = 1$ baseline while providing a +2.7% accuracy boost.

**Patch Embedding Factorization.** We evaluate depthwise–pointwise factorization (DW-PW Conv) against standard multi-branch convolution (Standard Conv) within RepPatchEmbed. While the DW-PW variant offers a theoretical reduction in GFLOPs (3.80 $\rightarrow$ 3.50), it causes a catastrophic drop in classification accuracy (> 20% decline). This suggests that the dense feature interaction of regular convolutions is critical for weed gender classification.

**Conditional Positional Encoding Placement.** We analyze the optimal insertion points for Conditional Positional Encodings (CPE) across the four backbone stages. Applying RepCPE in early stages hurts classification, whereas restricting it to the deeper layers (Stages 3 & 4) yields the best balance. Early stages encode local texture, while deeper stages capture high-level semantics that benefit most from conditional spatial context.

**RepPatchEmbed Kernel Size Configuration.** We compare a uniform large-kernel strategy [7, 7, 7, 7] against a progressive reduction strategy [7, 3, 3, 3] in the embedding modules. The uniform configuration uses 1.6× more parameters without improving performance, whereas the progressive strategy achieves higher

accuracy at a smaller size. This confirms that a large receptive field is crucial for early tokenization, while later stages benefit from smaller, more efficient kernels.
**Squeeze-and-Excitation Attention.** Finally, we add Squeeze-and-Excitation (SE) attention [13] to the RepClsHead. Contrary to expectations, SE degrades classification accuracy by 5.72%, so we exclude it from the final architecture.

## 6 Conclusion

This paper introduces WeedRepFormer, which systematically integrates structural reparameterization throughout a Vision Transformer architecture for simultaneous waterhemp segmentation and gender classification. Unlike prior agricultural work that primarily uses CNN-based reparameterization, we apply it to hierarchical Vision Transformers across backbone, decoder, and task-specific heads for joint dense prediction and biological attribute classification. WeedRepFormer outperforms all compared methods in classification accuracy while maintaining competitive segmentation with significantly reduced computational cost. Notably, our model achieves the best female classification performance with the smallest male-female accuracy gap, addressing a critical performance disparity. From an agricultural perspective, reliable female plant detection enables targeted weed management since female waterhemp are the primary seed producers responsible for population spread and herbicide resistance.

**Acknowledgments.** This work used the Delta system at the National Center for Supercomputing Applications (award OAC 2005572) through allocation CIS260273 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by U.S. National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296.

**Disclosure of Interests.** The authors have no competing interests to declare.

## References

1. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. In: arXiv preprint arXiv:1706.05587 (2017)
2. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European conference on computer vision (ECCV). pp. 801–818 (2018)
3. Chu, X., Tian, Z., Zhang, B., Wang, X., Shen, C.: Conditional positional encodings for vision transformers. In: International Conference on Learning Representations (ICLR) (2023)
4. Ding, X., Zhang, X., Ma, N., Han, J., Ding, G., Sun, J.: Repvgg: Making vgg-style convnets great again. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13733–13742 (2021)
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021)

6. Gage, K.L., Krausz, R.F., Walters, S.A.: Emerging challenges for weed management in herbicide-resistant crops. Agriculture **9**(8), 180 (2019). https://doi.org/10.3390/agriculture9080180
7. Goncalves, D.N., Junior, J.M., Zamboni, P., Pistori, H., Li, J., Nogueira, K., Goncalves, W.N.: Mtlsegformer: Multi-task learning with transformers for semantic segmentation in precision agriculture. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6290–6298 (2023)
8. Guo, A., Jia, Z., Ge, B., Chen, W., Song, S., He, C., Zhou, G., Wang, J., Lv, X.: Rlcfe-net: A reparameterization large convolutional kernel feature extraction network for weed detection in multiple scenarios. Expert Systems with Applications **274**, 126941 (2025)
9. Hager, A.G., Wax, L.M., Stoller, E.W., Bollero, G.A.: Common waterhemp (amaranthus rudis) interference in soybean. Weed science **50**(5), 607–610 (2002)
10. Hartzler, R.G., Battles, B.A., Nordby, D.: Effect of common waterhemp (amaranthus rudis) emergence date on growth and fecundity in soybean. Weed Science **52**(2), 242–245 (2004)
11. Heap, I.: The international herbicide-resistant weed database. Online (2025), accessed: Sept. 30, 2025. [Online]. Available: http://www.weedscience.org
12. Howard, A., Sandler, M., Chu, G., Chen, L.C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., et al.: Searching for mobilenetv3. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1314–1324 (2019)
13. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: CVPR. pp. 7132–7141 (2018)
14. Islam, T., Sarker, T.T., Ahmed, K.R., Rankrape, C.B., Gage, K.: Weedswin hierarchical vision transformer with sam-2 for multi-stage weed detection and classification. Scientific Reports **15**(1), 23274 (2025)
15. Liu, J., Davis, A.S., Tranel, P.J.: Pollen biology and dispersal dynamics in waterhemp (amaranthus tuberculatus). Weed Science **60**(3), 416–422 (2012)
16. Lottes, P., Behley, J., Milioto, A., Stachniss, C.: Fully convolutional networks with sequential information for robust crop and weed detection in precision farming. IEEE Robotics and Automation Letters **3**(4), 2870–2877 (2018)
17. Ma, R., Skelton, J.J., Riechers, D.E.: Measuring rates of herbicide metabolism in dicot weeds with an excised leaf assay. Journal of Visualized Experiments: Jove (103), 53236 (2015)
18. Pan, H., Hong, Y., Sun, W., Jia, Y.: Deep dual-resolution networks for real-time and accurate semantic segmentation of traffic scenes. IEEE Transactions on Intelligent Transportation Systems **24**(3), 3448–3460 (2022)
19. Pei, G., Qian, X., Zhou, B., Liu, Z., Wu, W.: Research on agricultural disease recognition methods based on very large kernel convolutional network-replknet. Scientific Reports **15**(1), 16843 (2025)
20. Poudel, R.P., Liwicki, S., Cipolla, R.: Fast-scnn: Fast semantic segmentation network. In: Proceedings of the British Machine Vision Conference (BMVC) (2019)
21. Rankrape, C.B., Lago, E., Miller, E.J., Werle, I., Tranel, P.J., Gage, K.L.: Evaluating glufosinate resistance in a waterhemp (amaranthus tuberculatus) population from southern illinois. In: North Central Weed Science Society. Kansas City, MO (December 2024)
22. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. In: International Conference on Learning Representations (ICLR) (2025)

23. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
24. Ruder, S.: An overview of multi-task learning in deep neural networks. arXiv preprint arXiv:1706.05098 (2017)
25. Rumpa, M.M., Babaei, S., Krausz, R.F., Gibson, D.J., Miller, E.J., Gage, K.L.: Does exposure to ppo-inhibiting herbicides alter the male-to-female sex ratio of palmer amaranth? Weed Technology **39**, e58 (2025)
26. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4510–4520 (2018)
27. Sarker, T.T., Ahmed, K.R., Islam, T., Rankrape, C.B., Gage, K.: Weedsense: Multi-task learning for weed segmentation, height estimation, and growth stage classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops. pp. 7180–7190 (October 2025)
28. Steckel, L.E.: The dioecious amaranthus spp.: here to stay. Weed Technology **21**(2), 567–570 (2007)
29. Steckel, L.E., Sprague, C.L.: Common waterhemp (amaranthus rudis) interference in corn. Weed Science **52**(3), 359–364 (2004)
30. Vasu, P.K.A., Gabriel, J., Zhu, J., Tuzel, O., Ranjan, A.: Fastvit: A fast hybrid vision transformer using structural reparameterization. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5785–5795 (2023)
31. Vasu, P.K.A., Gabriel, J., Zhu, J., Tuzel, O., Ranjan, A.: Mobileone: An improved one millisecond mobile backbone. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7907–7917 (2023)
32. Wang, A., Chen, H., Lin, Z., Han, J., Ding, G.: Repvit: Revisiting mobile cnn from vit perspective. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15909–15920 (2024)
33. Wang, J., Qi, Z., Wang, Y., Liu, Y.: A lightweight weed detection model for cotton fields based on an improved yolov8n. Scientific Reports **15**(1), 457 (2025)
34. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. Advances in neural information processing systems **34**, 12077–12090 (2021)
35. Yu, C., Gao, C., Wang, J., Yu, G., Shen, C., Sang, N.: Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation. International journal of computer vision **129**(11), 3051–3068 (2021)
36. Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., Sang, N.: Bisenet: Bilateral segmentation network for real-time semantic segmentation. In: Proceedings of the European conference on computer vision (ECCV). pp. 325–341 (2018)
37. Zhao, H., Qi, X., Shen, X., Shi, J., Jia, J.: Icnet for real-time semantic segmentation on high-resolution images. In: Proceedings of the European conference on computer vision (ECCV). pp. 405–420 (2018)
38. Zheng, C.: iformer: Integrating convnet and transformer for mobile application. In: International Conference on Learning Representations (ICLR) (2025)