WeedSense: Multi-Task Learning for Weed Segmentation, Height Estimation, and Growth Stage Classification

Toqi Tahamid Sarker, Khaled R Ahmed, Taminul Islam, Cristiana Bernardi Rankrape, Karla Gage

Southern Illinois University Carbondale, USA

{toqitahamid.sarker, khaled.ahmed, taminul.islam, cris.rankrape, kgage}@siu.edu

Abstract

Weed management represents a critical challenge in agriculture, significantly impacting crop yields and requiring substantial resources for control. Effective weed monitoring and analysis strategies are crucial for implementing sustainable agricultural practices and site-specific management approaches. We introduce WeedSense, a novel multi-task learning architecture for comprehensive weed analysis that jointly performs semantic segmentation, height estimation, and growth stage classification. We present a unique dataset capturing 16 weed species over an 11-week growth cycle with pixel-level annotations, height measurements, and temporal labels. WeedSense leverages a dual-path encoder incorporating Universal Inverted Bottleneck blocks and a Multi-Task Bifurcated Decoder with transformer-based feature fusion to generate multi-scale features and enable simultaneous prediction across multiple tasks. WeedSense outperforms other state-of-the-art models on our comprehensive evaluation. On our multi-task dataset, WeedSense achieves mIoU of 89.78% for segmentation, 1.67cm MAE for height estimation, and 99.99% accuracy for growth stage classification while maintaining real-time inference at 160 FPS. Our multitask approach achieves 3× faster inference than sequential single-task execution and uses 32.4% fewer parameters. Please see our project page at weedsense.github.io.

1. Introduction

Weed management is a critical challenge in agriculture, with weeds causing global potential yield losses of 34% across major crops [33]. Precise identification and monitoring of weed growth are essential for effective and sustainable agricultural practices, yet traditional methods are often labor-intensive and lack the granularity needed for site-specific weed management (SSWM) [2, 8].

The timing of weed intervention is fundamentally linked to the Critical Period of Weed Control (CPWC) - the time-frame during which crops must remain weed-free to prevent significant yield losses [60]. Since weeds emerging with or shortly after crop emergence cause substantially greater yield losses [5, 45], precise early growth stage identification is critical. While traditional phenological scales like BBCH-scale [17, 26] are designed for crop development, weekly monitoring intervals align better with CPWC time-frames and agricultural scheduling practices. Current weed control methods face significant limitations: manual weeding is labor-intensive and impractical for large-scale operations [6, 9], while excessive herbicide use has led to environmental pollution and resistance [28, 43]. Precision herbicide application, enabled by accurate weed detection, offers potential for reducing chemical usage by 54% while maintaining efficacy [16], necessitating automated identification systems for real-time, spatially-precise detection.

Deep learning has advanced computer vision tasks, achieving state-of-the-art results in image classification [15, 42], object detection [11], and semantic segmentation [56, 59]. In agriculture, these techniques have been increasingly applied to crop/weed classification [21], disease detection [7, 12, 31], and yield prediction [23]. Computer vision, particularly deep learning, offers a promising avenue for automating weed monitoring, enabling rapid and accurate weed detection [13, 14], growth stage classification [1, 47], and plant height estimation [32].

For semantic segmentation, early approaches relied on hand-crafted features [53], while modern architectures like FCNs [29], U-Net [38], DeepLab [4], and BiSeNetV2 [56] have enabled more accurate pixel-wise classification. These methods have been adapted for weed segmentation [19, 50] but often lack the capability to analyze growth patterns over time, limiting their utility for comprehensive weed management systems.

Plant height estimation has evolved from Structure-from-Motion [52] and stereo vision [24] approaches that reconstruct 3D plant models [10, 20] to deep learning methods [58] that directly learn height-related features from images. Similarly, growth stage classification has progressed from manual inspection to CNNs [36] and recurrent net-
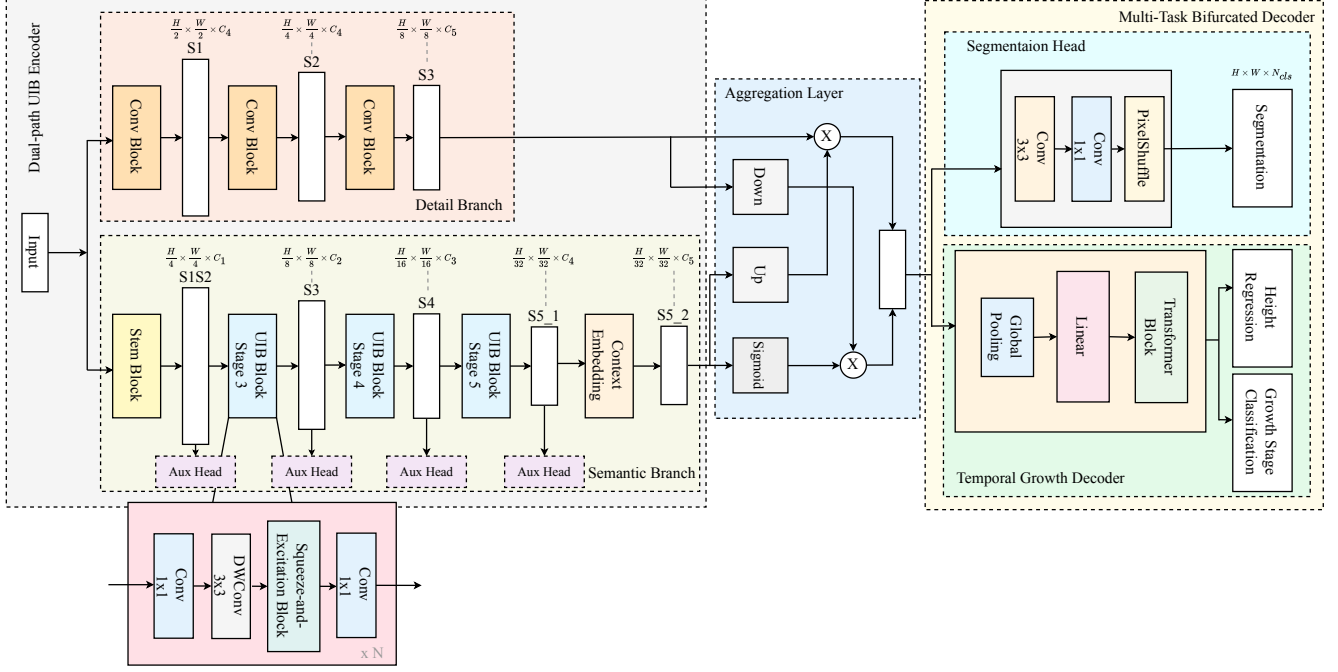
Figure 1. The proposed WeedSense framework consists of three main modules: A Dual-path UIB Encoder with parallel Detail and Semantic branches to extract multi-scale features; an Aggregation Layer that fuses these features through attention-guided operations; and a Multi-Task Bifurcated Decoder that simultaneously predicts semantic segmentation masks (17 classes including background), plant height (regression), and growth stage (11 classes). "UIB" indicates Universal Inverted Bottleneck blocks.

works like LSTMs [18] that can model temporal dependencies [51]. However, while deep learning approaches have been applied to plant height estimation in crops, their application specifically to weed height estimation represents a novel contribution, particularly when integrated with other analysis tasks.

Multi-task learning (MTL) aims to improve performance across multiple related tasks by leveraging shared representations [39]. While MTL has been applied to plant phenotyping [34], yield prediction [44], and disease detection [22], comprehensive weed analysis integrating segmentation, height regression, and growth stage classification remains largely unexplored.

To address this gap, we introduce WeedSense, which builds upon the efficient BiSeNetV2 [56] framework, incorporating Universal Inverted Bottleneck (UIB) blocks [35] in the Dual-path UIB Encoder (DUE) encoder for enhanced feature representation. A key innovation is our Multi-Task Bifurcated Decoder (MTBD) with a Temporal Growth Decoder (TGD) component, which leverages a transformer-based feature fusion mechanism to jointly learn height regression and growth stage classification from a shared feature representation. Our growth stage classification approach predicts growth stages using weekly development intervals. The TGD uses multi-head self-attention [49] to capture complex relationships between visual features and growth attributes. To support this research, we developed a

new, richly annotated dataset capturing the growth patterns of 16 distinct weed species over an 11-week period, from sprouting to flowering stage. Collected under controlled greenhouse conditions, it includes high-resolution video sequences with per-pixel semantic segmentation masks, precise weekly height measurements, and weekly growth labels. Our main contributions are:

- We introduce a new dataset of 16 weed species with annotations for semantic segmentation, height regression, and growth stage classification across their complete life cycle.
- We propose WeedSense, a multi-task learning architecture that jointly performs semantic segmentation, height estimation, and growth stage classification.
- We evaluate WeedSense's performance on our dataset and compare it with other state-of-the-art models.

The remainder of this paper is organized as follows: Section 2 introduces our dataset, Section 3 details the Weed-Sense architecture, Section 4 presents experimental results and ablation studies, and Section 5 concludes with a discussion of implications and future work.

2. Dataset

We present a novel multi-task temporal dataset of 16 weed species growth patterns for semantic segmentation, height regression, and growth stage classification. The dataset spans 11 weeks from sprouting through flowering, provid-
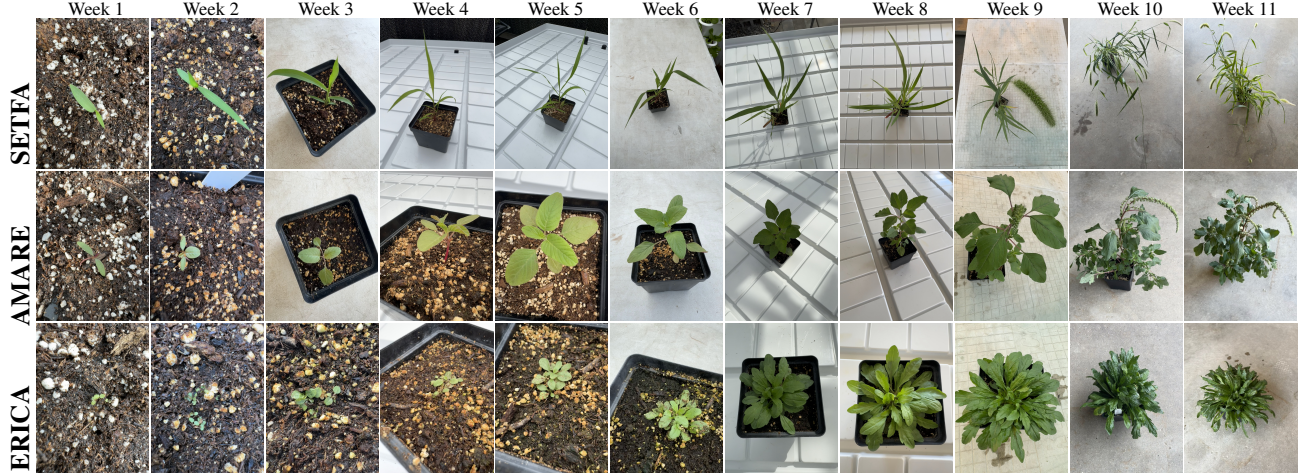
2

Figure 2. Growth progression over 11 weeks for three subjects (SETFA, AMARE, and ERICA), showing weekly development from week 1 to week 11. Best viewed on screen.

ing comprehensive coverage of the primary growth cycle for weed species commonly found in Midwestern cropping systems of the USA.

Species [25] (Scientific Name [3])	Image Count	Height (cm) Max	Std Dev	Growth (cm/week)
Fast-growing species (>10 cm/week)				
AMATU (Amaranthus tuberculatus (Moq.) Sauer.)	7,457	155.00	46.34	13.72
SORHA (Sorghum halepense (L.) Pers.)	5,540	121.00	45.57	14.06
SETFA (Setaria faberi Herrm.)	7,787	124.00	42.39	11.75
Medium-growing species (5-10 cm/week)				
SORVU (Sorghum bicolor (L.) Moench.)	4,787	100.00	38.21	9.84
PANDI (Panicum dichotomiflorum Michx.)	7,577	87.00	32.46	8.40
SETPU (Setaria pumila (Poir.) Roem.)	7,971	99.00	29.98	8.20
DIGSA (Digitaria sanguinalis (L.) Scop.)	9,211	77.00	27.84	7.53
ECHCG (Echinochloa crus-galli (L.) P. Beauv.)	8,562	80.00	26.35	7.38
SIDSP (Sida spinosa L.)	6,977	69.00	24.10	6.77
AMARE (Amaranthus retroflexus L.)	7,951	75.00	23.83	6.86
ABUTH (Abutilon theophrasti Medik.)	8,770	72.00	22.55	6.32
AMBEL (Ambrosia artemisiifolia L.)	8,630	71.00	22.07	6.19
Slow-growing species (<5 cm/week)				
AMAPA (Amaranthus palmeri S. Watson.)	9,080	62.00	19.82	5.66
CYPES (Cyperus esculentus L.)	8,131	56.00	18.26	5.42
CHEAL (Chenopodium album L.)	4,670	30.00	12.97	2.86
ERICA (Erigeron canadensis L.)	7,240	17.30	6.37	1.70

Table 1. Statistical summary of our weed species dataset (120,341 images) categorized by growth rates. The data shows significant variation in maximum height (17.3–155.0 cm) and growth patterns (1.70–14.06 cm/week), presenting diverse challenges for our multi-task learning approach.

Data Collection. We conducted this study during the spring and summer of 2024 at the SIU Horticulture Research Center greenhouse facility. The greenhouse was equipped with 1000W High Pressure Sodium grow lights maintaining optimal temperatures of 30–32°C. We used 32 square containers (10.7 cm × 10.7 cm × 9 cm) with two containers per species, where each container housed a single weed plant, creating replicate pairs for each species. Each container was filled with Pro-Mix® BX potting medium.

Weeds were maintained through regular watering based on soil moisture requirements and an all-purpose 20–20–20 nutrient solution applied every three days. Data acquisition utilized an iPhone 15 Pro Max positioned 1.5 feet above specimens, capturing 360-degree video documentation at 1440 × 1920 resolution and 30 FPS. We recorded 349 videos (15–30 seconds each) on a weekly basis throughout developmental stages, with the camera rotating around each plant to capture all viewing angles. Video recording was chosen over still images to ensure comprehensive angular coverage of each specimen in a single capture session. Minor protocol deviations occurred during collection: one plant each from SORVU and CHEAL duplicate pairs died after the third week, and SORHA weeds emerged only after the second week rather than the planned first week, while all remaining species successfully emerged within the first week. Despite these variations, consistent data collection procedures were maintained across all species throughout the complete 11-week monitoring period. We implemented systematic growth stage monitoring based on established phenological standards, with weekly imaging from week 1 (BBCH stage 11, first true leaf visible) through week 11 (BBCH stage 60, initial flower appearance). This temporal framework captured the complete vegetative growth cycle and transition to reproductive development across all species. Fig. 2 illustrates the temporal progression across three representative species (SETFA, AMARE, ERICA), demonstrating the diverse growth patterns and morphological changes captured in our dataset.

Data Preprocessing and Annotation. We applied two preprocessing steps: (1) temporal downsampling by extracting every 2nd frame from the 30 FPS videos to reduce redundancy, and (2) spatial downscaling to 720 × 960 pixels while preserving the original 3:4 aspect ratio. The resulting dataset contains 120,341 frames extracted from 349

videos, divided into training (80%), validation (10%), and test (10%) sets using frame-level splitting ensuring balanced species representation across splits.

For annotation, we employed the SAM2-Hiera-L [37] model to semi-automatically generate segmentation masks, with manual verification and correction. Plant components including stems, leaves, and flowers are labeled as a single foreground category. Each image is annotated with its corresponding growth week (1-11) based on capture date. We manually measured weed heights on a weekly basis, recording 325 measurements that revealed substantial variation across species, from 0.2 cm to 155 cm ($\Delta = 154.8$ cm). Table 1 presents key statistics for each species, showing significant inter- and intra-species variability. AMATU demonstrates the most aggressive growth, averaging 13.72 cm weekly increase and reaching 155 cm, while ERICA exhibits the slowest growth (1.70 cm/week) with a maximum height of just 17.3 cm. The height data reveals substantial intra-species variability, with standard deviations ranging from 6.37 cm to 46.34 cm, presenting challenging regression targets for our models. The resulting dataset combines high-resolution RGB images, segmentation masks, height measurements, and weekly growth stage labels, enabling comprehensive analysis of weed species growth patterns throughout their complete life cycle.

3. Method

This section introduces our Dual-path UIB Encoder with Multi-Task Bifurcated Decoder, an efficient framework for semantic segmentation and temporal growth analysis. As shown in Fig. 1, our model consists of two main encoder pathways that extract complementary information from input RGB images, and a bifurcated decoder that enables joint learning of multiple tasks. We denote feature maps as $C \times H \times W$ where C, H, and W represent spatial channel dimensions, height, and width respectively.

3.1. Dual-path UIB Encoder

Building upon the dual-path design of BiSeNetV2 [56], our encoder efficiently balances spatial detail preservation and semantic context extraction through specialized branches while incorporating UIB blocks for enhanced feature representation.

Detail Branch. The Detail Branch captures fine-grained spatial details crucial for accurate boundary delineation. Following a shallow-wide architecture inspired by VG-GNet [42], it consists of three sequential stages (S1, S2, S3) that generate feature maps with progressively reduced spatial resolutions ($H/2$, $H/4$, $H/8$) and expanded channel dimensions (64, 64, 128). Each stage applies 3×3 convolutions followed by batch normalization and ReLU activation.

Semantic Branch. The Semantic Branch follows a deep-narrow architecture with aggressive downsampling to ef-
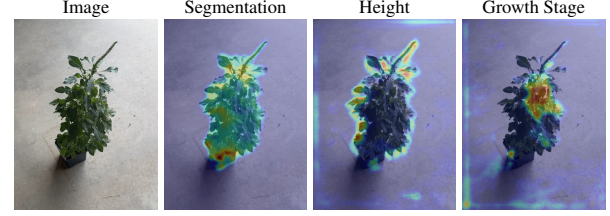

Figure 3. Visualization of attention activation maps from the aggregation layer for AMAPA at week 10. Each task displays distinct activation patterns: segmentation shows uniform boundary-focused activation, height estimation concentrates on plant extremities, and growth stage classification focuses on stem and mature leaf regions containing temporal growth indicators.

ficiently capture semantic context. It begins with a Stem Block for efficient initial feature extraction and downsampling, inspired by Inception [46] networks to balance computational efficiency with feature richness. The branch then employs hierarchical UIB blocks organized in three stages (S3, S4, S5_1) that progressively downsample features while increasing channel capacity, concluding with a context embedding block for global context enhancement.

Universal Inverted Bottleneck Blocks. As the core component of our Semantic Branch, we replace the original Gather-and-Expansion blocks [56] from BiSeNetV2 with UIB blocks from MobileNetV4 [35]. UIB blocks use configuration notation S[start]-M[mid]-E[end] representing kernel sizes for start, middle, and end depthwise convolutions, where kernel size 0 indicates the operation is skipped (identity mapping). Our S0-M3-E0 configuration employs only the middle 3×3 depthwise convolution:

$$
\begin{aligned}
F_{exp} &= \text{Conv}_{1 \times 1}(F_{in}) \cdot \text{ExpRatio} \\
F_{dw} &= \text{DWConv}_{3 \times 3}(F_{exp}) \\
F_{se} &= \text{SE}(F_{dw}) \\
F_{out} &= \text{LayerScale}(\text{Conv}_{1 \times 1}(F_{se})) + F_{in} \cdot \delta
\end{aligned}
\tag{1}
$$

where ExpRatio=6, SE enables adaptive channel recalibration with reduction ratio 0.25, LayerScale ensures stable training [48], and δ enables residual connections when dimensions match. The UIB blocks are organized hierarchically across three stages (S3, S4, S5_1) with progressive downsampling and channel expansion. To enhance global context modeling, we apply a Context Embedding Block after the final UIB stage that captures global statistical information through adaptive average pooling and residual connections.

Aggregation Layer. Having extracted complementary information through our dual-path design, we now need to effectively combine these heterogeneous features. The Aggregation Layer uses semantic features as intelligent guides to direct where detail features should focus, creating a unified representation that preserves spatial precision while incorporating semantic understanding. This aggregated representation then serves as input to our multi-task decoder.

Auxiliary Supervision Strategy. We implement four aux-

Model	Segmentation		Height Estimation							Growth Stage Classification	
	mIoU (%)↑	mF1 (%)↑	MAE (cm)↓	RMSE (cm)↓	R^2 ↑	Max Error (cm)↓	Within 1cm (%)↑	Within 2cm (%)↑	Within 5cm (%)↑	Accuracy (%)↑	F1 (%)↑
MTL-SegFormer	56.17	70.56	5.10	8.03	0.9289	61.07	25.21	40.15	65.16	97.96	98.22
MTL-UNet	51.51	66.3	2.15	3.03	0.9899	29.22	39.07	60.5	89.14	99.97	99.97
MTL-Poolformer	72.45	83.53	4.20	6.74	0.9499	52.52	29.08	45.29	72.41	98.70	98.89
MTL-BiSeNetV1	87.25	93.08	1.97	2.81	0.9913	**18.89**	41.22	63.90	91.51	99.99	99.99
MTL-BiSeNetV2	89.29	94.26	1.75	2.51	0.9930	24.39	**44.11**	68.13	94.30	**100.00**	**100.00**
MTL-SFNet	86.27	92.51	4.21	6.66	0.9511	69.24	26.36	44.53	72.33	98.45	98.74
WeedSense	**89.78**	**94.54**	**1.67**	**2.32**	**0.9941**	19.34	43.26	**70.37**	**95.49**	99.99	99.99

Table 2. Quantitative comparison of multi-task learning models for weed species analysis. Our WeedSense model achieves best or competitive performance across segmentation, height estimation, and growth stage classification metrics. Bold indicates best results.

iliary segmentation heads connected to different stages of the Semantic Branch (Stem Block, S3, S4, and S5_1), with stage-specific upsampling factors ($4\times$, $8\times$, $16\times$, and $32\times$ respectively). These heads facilitate effective gradient flow during training while being discarded during inference, thus incurring no additional computational cost.

3.2. Multi-Task Bifurcated Decoder

Our MTBD processes the aggregated features ($128 \times H/8 \times W/8$) from the encoder through parallel pathways for semantic segmentation and temporal growth prediction.

Semantic Segmentation Head. The segmentation head follows an encoder-decoder structure with progressive upsampling to recover full spatial resolution:

$$F_{mid} = \text{Dropout}(\text{Conv}_{3\times3}(F_{agg}))$$
$$M = \text{PixelShuffle}(\text{Conv}_{1\times1}(F_{mid})) \quad (2)$$

where F_{agg} represents the aggregated features and M is the predicted segmentation mask at resolution $N_{cls} \times H \times W$, with N_{cls} representing the number of classes (17 classes including background). PixelShuffle [41] provides parameter-free $8\times$ upsampling by reorganizing channel data into spatial dimensions.

Temporal Growth Decoder. The TGD predicts plant height and growth stage through global feature processing. Aggregated features undergo adaptive average pooling to produce a 128-dimensional representation, followed by linear projection to 512-dimensional embeddings. The projected features are processed by a transformer block with multi-head self-attention and layer normalization:

$$F_{attn} = \text{LayerNorm}(F_{proj} + \text{MHA}(F_{proj}, F_{proj}, F_{proj}))$$
$$F_{trans} = \text{LayerNorm}(F_{attn} + \text{FFN}(F_{attn})) \quad (3)$$

where MHA represents multi-head attention (8 heads) and FFN is a two-layer feed-forward network ($512 \rightarrow 2048 \rightarrow 512$ dimensions) with GELU activation. Finally, two parallel task-specific heads process the transformed features:

$$F_{task} = \text{ReLU}(\text{LayerNorm}(\text{Linear}_{512}(\text{Linear}_{1024}(F_{trans}))))$$
$$\hat{h} = \text{Linear}_1(F_{task}) \quad (4)$$
$$\hat{w} = \text{Linear}_{11}(F_{task})$$

where $\hat{h}$ represents plant height prediction (in cm) and $\hat{w}$ represents growth stage classification logits using weekly intervals (weeks 1-11). The architecture employs hard parameter sharing in the feature processing pipeline while maintaining separate head weights, facilitating knowledge transfer between related tasks [39].

Feature Activation Visualization. Figure 3 shows Grad-CAM [40] activation maps from our aggregation layer for AMAPA at week 10. The maps demonstrate task-specific feature extraction. Segmentation shows uniform activation across plant boundaries. This helps with accurate boundary delineation. Height estimation focuses on plant extremities and uppermost leaves. These regions correspond to vertical measurements. Growth stage classification targets the central stem and mature leaf regions. These areas contain visual features like stem thickness and leaf development that indicate developmental progression through weekly intervals. The distinct patterns confirm our multi-task architecture learns complementary representations. There is no interference between tasks. This enables high performance across all tasks while sharing network parameters.

4. Experiments and Discussion

4.1. Implementation Details

All experiments are conducted using PyTorch on an Intel Xeon Gold 6240 CPU @ 2.60GHz system with 124 GB RAM and an NVIDIA Tesla V100S GPU with 32GB memory. We evaluate our approach against six state-of-the-art architectures adapted for multi-task learning: Seg-Former [54], UNet [38], PoolFormer [57], BiSeNetV1 [55], BiSeNetV2 [56], and SFNet [27]. For fair comparison, each method uses its original encoder-decoder architecture for semantic segmentation while incorporating our proposed Temporal Growth Decoder (TGD) for height regression and growth stage classification tasks. The model is trained using a multi-task loss function that equally weights three components: pixel-level weighted cross-entropy for segmentation (including auxiliary supervision), Mean Squared Error for height regression, and cross-entropy for growth stage classification. All models are trained for 50 epochs from scratch without pre-trained weights. Training uses the Adam opti-
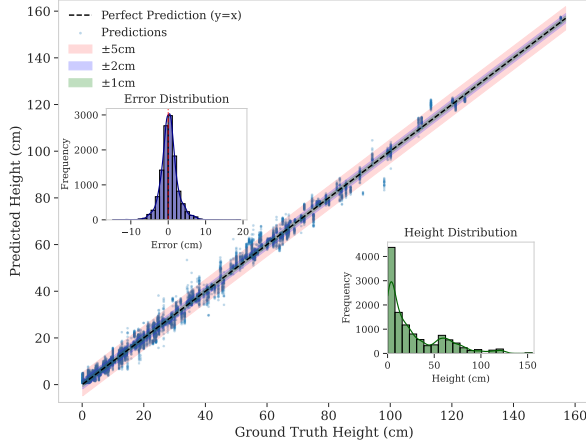
Figure 4. Height estimation performance of WeedSense across the full measurement range (0.2-155 cm). Predictions cluster tightly along the perfect prediction line (y=x) with tolerance bands showing ±1cm (green), ±2cm (blue), and ±5cm (red) accuracy zones. The error distribution (top left inset) is symmetric around zero with no systematic bias, while the height distribution (bottom right inset) shows the diverse range of plant sizes in our dataset.

mizer with initial learning rate $\eta_{\text{base}} = 2 \times 10^{-4}$, weight decay of 0.0001, and cosine annealing schedule [30]. A 1,500-iteration linear warmup period gradually increases the learning rate from $0.1\eta_{\text{base}}$ to η_{base}. Data augmentation includes random cropping (50%-200% of original size), horizontal flipping (50% probability), and ImageNet normalization. Random cropping is applied to all training images, while horizontal flipping is applied probabilistically to half the samples during each epoch. All images are resized to 512×512 pixels with batch size 8 during training and batch size 1 during evaluation.

4.2. Comparative Evaluation

We evaluate our proposed WeedSense against the multi-task learning variants we created: MTL-SegFormer, MTL-UNet, MTL-PoolFormer, MTL-BiSeNetV1, MTL-BiSeNetV2, and MTL-SFNet. Tables 2 to 4 summarize results across segmentation quality, height estimation, growth stage classification, and computational efficiency.

Segmentation Performance. As shown in Table 2, Weed-Sense achieves the best segmentation performance with 89.78% mIoU and 94.54% mF1 score. Among competitive methods, WeedSense surpasses MTL-BiSeNetV2 by 0.49 and 0.28 percentage points respectively, outperforms MTL-BiSeNetV1 by 2.53 percentage points, and exceeds MTL-SFNet by 3.51 percentage points in mIoU. WeedSense achieves significant gains over lower-performing methods: 38.27 percentage points better than MTL-UNet, 33.61 percentage points better than MTL-SegFormer, and 17.33 percentage points better than MTL-PoolFormer. The qualitative results in Fig. 5 further validate these findings, showing that WeedSense provides more accurate boundary delineation for AMATU, CYPES, ECHCG species with better

Model	Small↓ 0-20cm	Medium↓ 20-50cm	Large↓ 50-100cm	Very Large↓ >100cm
MTL-SegFormer	2.81 cm	7.50 cm	8.18 cm	11.29 cm
MTL-UNet	1.46 cm	3.03 cm	3.10 cm	3.01 cm
MTL-Poolformer	2.23 cm	6.11 cm	7.10 cm	8.89 cm
MTL-BiSeNetV1	1.35 cm	2.60 cm	2.96 cm	2.74 cm
MTL-BiSeNetV2	**1.20 cm**	2.53 cm	2.32 cm	3.12 cm
MTL-SFNet	2.12 cm	5.64 cm	7.65 cm	10.22 cm
WeedSense	**1.20 cm**	**2.28 cm**	**2.28 cm**	**2.60 cm**

Table 3. Height estimation error (MAE) across different plant size categories.

Model	Params (M)↓	GFLOPs↓	FPS↑
MTL-SegFormer	**8.19**	**7.94**	138
MTL-UNet	35.67	233.23	94
MTL-Poolformer	18.77	22.86	101
MTL-BiSeNetV1	17.31	13.30	**249**
MTL-BiSeNetV2	29.62	16.84	185
MTL-SFNet	18.62	30.77	151
WeedSense	30.50	16.73	160

Table 4. Computational efficiency comparison showing model size, computational complexity, and inference speed.

precision for fine-grained plant structures.

Height Estimation Performance. Height regression presents challenges due to the wide range of plant heights from 0.2 to 155 cm and visual occlusions during flowering stages. As detailed in Table 2, WeedSense achieves the best overall performance with 1.67 cm MAE, 2.32 cm RMSE, and 0.9941 R² value. Our approach achieves a maximum error of 19.34 cm versus the best competitor's 24.39 cm, representing a 20.7% reduction. For tolerance rates, Weed-Sense achieves 43.26% of predictions within 1 cm, 70.37% within 2 cm, and 95.49% within 5 cm. The 2 cm tolerance rate surpasses all competitors by 2.24-30.22 percentage points.

Table 3 reveals consistent performance across plant sizes. For small plants (0-20 cm), WeedSense matches the best performer at 1.20 cm error. For medium plants (20-50 cm), our approach achieves 2.28 cm error, representing a 9.9% improvement over the best competitor. For large (50-100 cm) and very large plants (>100 cm), WeedSense maintains 2.28 cm and 2.60 cm errors respectively, outperforming all other methods. Figure 4 visualizes predictions clustering tightly along the identity line with consistent accuracy across all height ranges.

Growth Stage Classification Performance. As shown in Table 2, WeedSense achieves 99.99% accuracy and F1 score on growth stage classification. The top-performing models achieve near-perfect results, with WeedSense ranking among the best performers alongside MTL-BiSeNetV2 and MTL-BiSeNetV1. Performance differences among top models are negligible for practical applications. The consistently high classification performance across most architectures indicates that growth stage classification is more
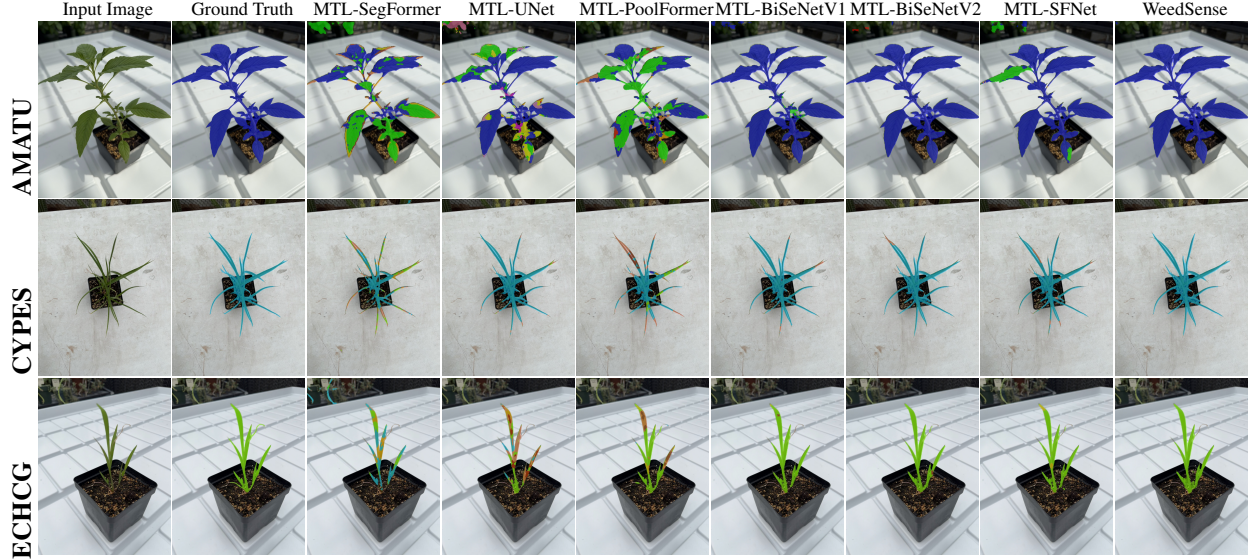
Figure 5. Qualitative comparison of different models, showing the input images, ground truth masks, and the segmentation outputs for AMATU, CYPES and ECHCG. Best viewed on screen.

tractable than segmentation or height estimation, likely due to distinctive visual patterns associated with different growth stages.

Real-Time Performance Analysis. As shown in Table 4, WeedSense achieves 160 FPS with 30.50M parameters and 16.73 GFLOPs, providing a balance between performance and efficiency. While MTL-BiSeNetV1 offers the highest frame rate at 249 FPS and MTL-SegFormer requires the fewest resources at 7.94 GFLOPs, WeedSense delivers competitive overall efficiency-performance trade-offs. WeedSense runs 1.7× faster than MTL-UNet while requiring 13.94× fewer computational resources and achieving 38.27 percentage points better segmentation mIoU. Compared to MTL-SFNet, our approach provides 1.84× computational efficiency improvement while maintaining 3.51 percentage points better segmentation performance.

4.3. Ablation Studies

To validate our architectural design choices and determine the optimal configuration for WeedSense, we conduct extensive ablation experiments. All experiments follow the training protocol described in Sec. 4.1.

Kernel Configuration. We first analyze the effect of kernel size configuration for UIB blocks, denoted as S[start]-M[mid]-E[end]. Table 5 shows the performance results across different configurations. The first thing to observe is that the S5-M3-E5 configuration achieves the highest segmentation mIoU at 89.53% but requires slightly more computational resources. The S0-M3-E0 configuration maintains competitive performance while offering the lowest computational cost. For height estimation, the S0-M3-E5 configuration achieves the lowest MAE at 1.64cm, while S1-M3-E0 provides competitive performance at 1.65cm.

Interestingly, all configurations achieve near-perfect growth stage classification accuracy. From these results, we select the S0-M3-E0 configuration as it provides the optimal efficiency-performance trade-off.

Squeeze-and-Excitation Module Effects. Building upon the S0-M3-E0 configuration, we analyze the impact of squeeze-and-excitation (SE) modules. As shown in Table 5, incorporating SE modules increase segmentation mIoU by 0.94 percentage points and reduces height estimation error by 0.07 cm. This performance gain comes at the cost of only 1.08M additional parameters with no change in computational complexity. From these results, we can conclude that adaptive channel recalibration improves feature quality with minimal computational overhead.

Channel Capacity. We now analyze the effect of channel capacity in the aggregation layer and SE modules on model performance. Table 5 shows performance across different configurations. The first thing to observe is that adding SE modules universally improves performance regardless of channel dimensions. When comparing configurations, the 256-channel with SE achieves better height estimation but similar segmentation performance compared to the 128-channel variant, yet requires 45.3% more computation. For segmentation, performance increases as channel capacity increases up to 128-channels, then plateaus with minimal gains at 256-channels. For height estimation, performance continues to improve with wider channels. The 64-channel variants show substantial performance degradation despite their efficiency advantages. From these results, we select the 128-channel with SE configuration as it offers the best balance between accuracy and efficiency for deployment.

Auxiliary Supervision. Table 5 shows the results with and without auxiliary heads. Removing auxiliary heads reduces

Config.	Seg.↑ mIoU (%)	Height↓ MAE (cm)	Week↑ Acc (%)	Params (M)↓	GFLOPs↓
UIB Kernel Configuration					
S0-M3-E0	88.84	1.74	100.00	29.42	16.73
S1-M3-E0	88.81	1.65	100.00	29.43	16.74
S0-M3-E1	89.43	1.67	100.00	29.43	16.74
S1-M3-E1	89.49	1.75	99.98	29.43	16.74
S5-M3-E0	89.41	1.67	99.99	29.44	16.75
S0-M3-E5	89.17	1.64	99.91	29.44	16.75
S5-M3-E5	89.53	1.70	99.96	29.46	16.77
Squeeze-and-excitation Module					
No SE	88.84	1.74	100.00	29.42	16.73
With SE	89.78	1.67	99.99	30.50	16.73
Channel Capacity Configuration					
C64-NoSE	86.50	2.28	99.87	28.45	13.70
C64-SE	87.31	2.00	99.99	29.53	13.70
C128-NoSE	88.84	1.74	100.00	29.42	16.73
C128-SE	89.78	1.67	99.99	30.50	16.73
C256-NoSE	89.37	1.52	100.00	32.13	24.31
C256-SE	89.70	1.52	100.00	33.21	24.31
Auxiliary Supervision Configuration					
No Aux	87.48	2.05	98.70	9.62	16.73
Aux S1-4	89.78	1.67	99.99	30.50	16.73
Model Size Configuration					
Small	77.17	2.66	99.69	21.50	3.56
Medium	89.78	1.67	99.99	30.50	16.73
Large	91.41	1.41	99.78	45.36	34.55
Task Configuration					
Single Task	93.67	1.53	99.55	-	-
Multi Task	89.78	1.67	99.99	-	-

Table 5. Comprehensive ablation analysis of WeedSense design components. Results demonstrate the progressive refinement from baseline S0-M3-E0 kernel configuration through the addition of squeeze-and-excitation modules, optimal channel capacity selection, auxiliary supervision, and model size scaling. The final comparison shows trade-offs between single-task and multi-task learning approaches.

Component	Small	Medium	Large
Detail Branch Channels	32-32-64	64-64-128	96-96-192
Semantic Branch Channels	8-16-32-64	16-32-64-128	24-48-96-192
UIB Blocks per Stage	1-1-2	2-2-4	3-3-6
Expansion Ratio	4	6	6
Transformer Embed Dim	256 (4 heads)	512 (8 heads)	768 (12 heads)
Task Head Dimensions	512→256	1024→512	1536→768

Table 6. Model size variant configurations. Branch channels correspond to stages S1-S2-S3 (Detail) and stem-S3-S4-S5 (Semantic). UIB blocks are distributed across S3-S4-S5 stages.

proach against dedicated single-task models using the same Medium-variant architecture. Table 5 shows that while single-task models achieve 3.89% better segmentation mIoU and 0.14 cm lower height MAE, our multitask model provides substantial computational advantages. Individual single-task models require 45.14M total parameters (segmentation: 30.50M, week: 7.32M, height: 7.32M) compared to our multitask model's 30.50M parameters, representing a 32.4% parameter reduction. For inference efficiency, sequential execution of single-task models achieves only 52.3 FPS, while parallel execution reaches 157.84 FPS but requires significantly more resources. Our unified approach achieves 160 FPS with a single model, outperforming even parallel single-task execution while using substantially fewer resources, demonstrating an optimal balance between performance and efficiency for practical deployment.

5. Conclusion

In this study, we proposed WeedSense, a novel multi-task learning architecture for comprehensive weed analysis that simultaneously performs semantic segmentation, height estimation, and growth stage classification using RGB imagery. We also introduced a novel dataset capturing 16 weed species over an 11-week growth cycle with pixel-level annotations, height measurements, and weekly growth stage labels to evaluate the model's performance across diverse growth patterns and species characteristics. Experimental evaluation demonstrated that WeedSense outperforms state-of-the-art models across key performance metrics including segmentation accuracy, height estimation precision, and growth stage classification using weekly intervals. Our unified multitask approach achieves computational efficiency gains, requiring 32.4% fewer parameters than separate single-task models while maintaining 160 FPS inference speed compared to 52.3 FPS for sequential single-task execution. Future research could explore the integration of WeedSense with morphology-based BBCH growth stage classification, field-based data collection to validate performance in real agricultural environments, and crop-weed discrimination tasks to enhance its adaptability in diverse agricultural environments while investigating its robustness across different lighting conditions and weather scenarios.

segmentation mIoU by 2.30 percentage points and increases height estimation error by 22.75%. Growth stage classification accuracy also drops from 99.99% to 98.70%. As described in Sec. 3.1, these auxiliary components are training-only and incur zero computational overhead during inference.

Model Size Scaling. We implement three model variants that systematically scale key architectural components to analyze capacity-performance trade-offs. Table 6 details the scaling strategy across both encoder branches and decoder components. As shown in Table 5, performance increases predictably with model size. The Large variant improves segmentation mIoU by 1.63 percentage points and reduces height MAE by 0.26 cm compared to Medium. However, these gains come at the cost of significantly increased computational requirements. In contrast, the Small variant offers substantially lower computational cost but with substantial accuracy degradation compared to Medium. From these results, we conclude that the Medium variant provides the optimal compromise for practical deployment.

Single vs. Multitask. We compare our multitask ap-

References

[1] Abeer M Almalky and Khaled R Ahmed. Deep learning for detecting and classifying the growth stages of consolida regalis weeds on fields. *Agronomy*, 13(3):934, 2023. 1

[2] TW Berge, AH Aastveit, and H Fykse. Evaluation of an algorithm for automatic detection of broad-leaved weeds in spring cereals. *Precision Agriculture*, 9:391–405, 2008. 1

[3] Thomas Borsch, Walter Berendsohn, Eduardo Dalcin, Maïté Delmas, Sebsebe Demissew, Alan Elliott, Peter Fritsch, Anne Fuchs, Dmitry Geltman, Adil Güner, et al. World flora online: Placing taxonomists at the heart of a definitive and comprehensive global resource on the world's plants. *Taxon*, 69(6):1311–1341, 2020. 3

[4] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 1

[5] DA Dew. An index of competition for estimating crop loss due to weeds. *Canadian Journal of Plant Science*, 52(6):921–927, 1972. 1

[6] William W Donald. Between-row mowing systems control summer annual weeds in no-till grain sorghum. *Weed Technology*, 21(2):511–517, 2007. 1

[7] Konstantinos P Ferentinos. Deep learning models for plant disease detection and diagnosis. *Computers and electronics in agriculture*, 145:311–318, 2018. 1

[8] C Fernández-Quintanilla, JM Peña, Dionisio Andújar, José Dorado, A Ribeiro, and F López-Granados. Is the current state of the art of weed monitoring suitable for site-specific weed management in arable crops? *Weed research*, 58(4):259–272, 2018. 1

[9] Frank Forcella. Rotary hoeing substitutes for two-thirds rate of soil-applied herbicide. *Weed Technology*, 14(2):298–303, 2000. 1

[10] Ryo Fujiwara, Tomohiro Kikawada, Hisashi Sato, and Yukio Akiyama. Comparison of remote sensing methods for plant heights in agricultural fields using unmanned aerial vehicle-based structure from motion. *Frontiers in Plant Science*, 13:886804, 2022. 1

[11] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 1

[12] Sandeep Goshika, Khalid Meksem, Khaled R Ahmed, and Naoufal Lakhssassi. Deep learning model for classifying and evaluating soybean leaf disease damage. *International Journal of Molecular Sciences*, 25(1):106, 2023. 1

[13] ASM Mahmudul Hasan, Ferdous Sohel, Dean Diepeveen, Hamid Laga, and Michael GK Jones. A survey of deep learning techniques for weed detection from images. *Computers and electronics in agriculture*, 184:106067, 2021. 1

[14] Sebastian Haug and Jörn Ostermann. A crop/weed field image dataset for the evaluation of computer vision based precision agriculture tasks. In *Computer Vision-ECCV 2014 Workshops: Zurich, Switzerland, September 6-7 and 12, 2014, Proceedings, Part IV 13*, pages 105–116. Springer, 2015. 1

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[16] T Heisel, S Christensen, and AM Walter. Whole-field experiments with site-specific weed management. 1999. 1

[17] Michael Hess, Gilbert Barralis, H Bleiholder, L Buhr, TH Eggers, H Hack, and R Stauss. 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, 1997. 1

[18] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 2

[19] Huasheng Huang, Yubin Lan, Jizhong Deng, Aqing Yang, Xiaoling Deng, Lei Zhang, and Sheng Wen. A semantic labeling approach for accurate weed mapping of high resolution uav imagery. *Sensors*, 18(7):2113, 2018. 1

[20] Sylvain Jay, Gilles Rabatel, Xavier Hadoux, Daniel Moura, and Nathalie Gorretta. In-field crop row phenotyping from 3d modeling performed using structure from motion. *Computers and Electronics in Agriculture*, 110:70–77, 2015. 1

[21] Andreas Kamilaris and Francesc X Prenafeta-Boldú. Deep learning in agriculture: A survey. *Computers and electronics in agriculture*, 147:70–90, 2018. 1

[22] Ali Seydi Keceli, Aydin Kaya, Cagatay Catal, and Bedir Tekinerdogan. Deep learning-based multi-task prediction system for plant disease and species detection. *Ecological Informatics*, 69:101679, 2022. 2

[23] Saeed Khaki and Lizhi Wang. Crop yield prediction using deep neural networks. *Frontiers in plant science*, 10:621, 2019. 1

[24] Wan-Soo Kim, Dae-Hyun Lee, Yong-Joo Kim, Taehyeong Kim, Won-Suk Lee, and Chang-Hyun Choi. Stereo-vision-based crop height estimation for agricultural robots. *Computers and Electronics in Agriculture*, 181:105937, 2021. 1

[25] J Kotleba. European and mediterranean plant protection organization (eppo). *Agrochemia (Slovak Republic)*, 34(4), 1994. 3

[26] Peter D Lancashire, Hermann Bleiholder, T van den Boom, P Langelüddeke, Reinhold Stauss, Elfriede Weber, and A Witzenberger. A uniform decimal code for growth stages of crops and weeds. *Annals of applied Biology*, 119(3):561–601, 1991. 1

[27] Junghyup Lee, Dohyung Kim, Jean Ponce, and Bumsub Ham. Sfnet: Learning object-aware semantic correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2278–2287, 2019. 5

[28] Fuhan Liu and Neil V O'Connell. Off-site movement of surface-applied simazine from a citrus orchard as affected by irrigation incorporation. *Weed Science*, 50(5):672–676, 2002. 1

[29] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 1

[30] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[31] Sharada P Mohanty, David P Hughes, and Marcel Salathé. Using deep learning for image-based plant disease detection. *Frontiers in plant science*, 7:215232, 2016. 1

[32] Rama Rao Nidamanuri. Deep learning-based prediction of plant height and crown area of vegetable crops using lidar point cloud. *Scientific Reports*, 14(1):14903, 2024. 1

[33] E-C Oerke. Crop losses to pests. *The Journal of agricultural science*, 144(1):31–43, 2006. 1

[34] Michael P Pound, Jonathan A Atkinson, Darren M Wells, Tony P Pridmore, and Andrew P French. Deep learning for multi-task plant phenotyping. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 2055–2063, 2017. 2

[35] Danfeng Qin, Chas Leichner, Manolis Delakis, Marco Fornoni, Shixin Luo, Fan Yang, Weijun Wang, Colby Banbury, Chengxi Ye, Berkin Akin, et al. Mobilenetv4: universal models for the mobile ecosystem. In *European Conference on Computer Vision*, pages 78–96. Springer, 2024. 2, 4

[36] Sanaz Rasti, Chris J Bleakley, Guénolé CM Silvestre, Nicholas M Holden, David Langton, and Gregory MP O'Hare. Crop growth stage estimation prior to canopy closure using deep learning algorithms. *Neural Computing and Applications*, 33(5):1733–1743, 2021. 1

[37] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 4

[38] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 1, 5

[39] Sebastian Ruder. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*, 2017. 2, 5

[40] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: visual explanations from deep networks via gradient-based localization. *International journal of computer vision*, 128:336–359, 2020. 5

[41] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883, 2016. 5

[42] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1, 4

[43] Niels Henrik Spliid and Benny Køppen. Occurrence of pesticides in danish shallow ground water. *Chemosphere*, 37(7):1307–1316, 1998. 1

[44] Zhuangzhuang Sun, Qing Li, Shichao Jin, Yunlin Song, Shan Xu, Xiao Wang, Jian Cai, Qin Zhou, Yan Ge, Ruinan Zhang,

et al. Simultaneous prediction of wheat yield and grain protein content using multitask deep learning from time-series proximal sensing. *Plant Phenomics*, 2022. 2

[45] CJ Swanton, S Weaver, P Cowan, R Van Acker, W Deen, and A Shrestha. Weed thresholds: theory and applicability. In *Expanding the context of weed management*, pages 9–29. CRC Press, 2020. 1

[46] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 4

[47] Nima Teimouri, Mads Dyrmann, Per Rydahl Nielsen, Solvejg Kopp Mathiassen, Gayle J Somerville, and Rasmus Nyholm Jørgensen. Weed growth stage estimator using deep convolutional neural networks. *Sensors*, 18(5):1580, 2018. 1

[48] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 32–42, 2021. 4

[49] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[50] Aichen Wang, Yifei Xu, Xinhua Wei, and Bingbo Cui. Semantic segmentation of crop and weed using an encoder-decoder network and image enhancement method under uncontrolled outdoor illumination. *Ieee Access*, 8:81724–81734, 2020. 1

[51] Chunying Wang, Weiting Pan, Xubin Song, Haixia Yu, Junke Zhu, Ping Liu, and Xiang Li. Predicting plant growth and development using time-series images. *Agronomy*, 12(9):2213, 2022. 2

[52] Matthew J Westoby, James Brasington, Niel F Glasser, Michael J Hambrey, and Jennifer M Reynolds. 'structure-from-motion'photogrammetry: A low-cost, effective tool for geoscience applications. *Geomorphology*, 179:300–314, 2012. 1

[53] DM Woebbecke, GE Meyer, K Von Bargen, and DA Mortensen. Shape features for identifying young weeds using image analysis. *Transactions of the ASAE*, 38(1):271–281, 1995. 1

[54] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021. 5

[55] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 325–341, 2018. 5

[56] Changqian Yu, Changxin Gao, Jingbo Wang, Gang Yu, Chunhua Shen, and Nong Sang. Bisenet v2: Bilateral network with guided aggregation for real-time semantic seg-

mentation. *International journal of computer vision*, 129:
3051–3068, 2021. 1, 2, 4, 5

[57] Weihao Yu, Mi Luo, Pan Zhou, Chenyang Si, Yichen Zhou,
Xinchao Wang, Jiashi Feng, and Shuicheng Yan. Metaformer
is actually what you need for vision. In *Proceedings of
the IEEE/CVF conference on computer vision and pattern
recognition*, pages 10819–10829, 2022. 5

[58] Genping Zhao, Weitao Cai, Zhuowei Wang, Heng Wu,
Yeping Peng, and Lianglun Cheng. Phenotypic parameters
estimation of plants using deep learning-based 3-d recon-
struction from single rgb image. *IEEE Geoscience and Re-
mote Sensing Letters*, 19:1–5, 2022. 1

[59] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang
Wang, and Jiaya Jia. Pyramid scene parsing network. In
*Proceedings of the IEEE conference on computer vision and
pattern recognition*, pages 2881–2890, 2017. 1

[60] Robert L Zimdahl. The concept and application of the critical
weed-free period. 1988. 1