# Mask-Guided Multi-Task Learning: Real-Time RGB Prediction of Plant Photosynthetic Efficiency at the Edge

Abdellah Lakhssassi[1] [0009-0004-8978-3917], Taminul Islam[1] [0000-0002-7803-4564], Cristiana Bernardi Rankrape[2] [0009-0006-6026-7876], Naoufal Lakhssassi[3] [0000-0002-8255-9419], Karla Gage[2] [0000-0002-6235-8804], and Khaled Ahmed[1] [0000-0002-3707-4316]

[1] School of Computing, Southern Illinois University, Carbondale, IL 62901, USA
{abdellah.lakhssassi,taminul.islam,khaled.ahmed}@siu.edu
[2] School of Agricultural Sciences, Southern Illinois University, Carbondale, IL 62901, USA
{cris.rankrape,kgage}@siu.edu
[3] Department of Biological Sciences, School of Science, Hampton University, Hampton, VA 23668, USA
naoufal.lakhssassi@hamptonu.edu

**Abstract.** Non-invasive prediction of the photosynthetic efficiency metric $F_v/F_m$ from standard RGB imagery would enable scalable, high-throughput crop stress monitoring. However, standard deep learning models struggle with this continuous regression task; retrofitted segmentation backbones degenerate by integrating irrelevant background noise. We propose PhotoSoyNet, a multi-task architecture for joint plant segmentation and per-image $F_v/F_m$ regression. Its core contribution is the Coupled Task Head, which aligns the heterogeneous dense and scalar tasks via two mechanisms: Plant-Weighted Spatial Pooling (PWSP) restricts the regression pool strictly to predicted plant tissue, while bidirectional cross-task channel gating enables symbiotic feature modulation. Evaluated on a novel multi-stress soybean dataset, state-of-the-art baselines up to 235M parameters fail to exceed $R^2 = 0.25$. In contrast, our edge-deployable variant, PhotoSoyNet-F (3.2M parameters), achieves $R^2 = 0.333$. Deployed on an NVIDIA Jetson Orin Nano, it operates at 47.8 FPS and 302 mJ per inference, running nearly 10× faster and 11.5× more efficiently than the SOTA segmentation baseline (SegFormer-B5), uniquely satisfying both real-time edge constraints and non-trivial physiological regression.

**Keywords:** Plant Phenotyping · Multi-Task Learning · Edge Computing · Photosynthetic Efficiency · Precision Agriculture.

## 1 Introduction

Non-invasive monitoring of plant health is central to precision agriculture and biodiversity conservation. The maximum quantum yield of photosystem II, $F_v/F_m$, is among the most reliable early-warning markers of abiotic stress, declining measurably before visible symptoms appear [15]. It is conventionally obtained with dedicated chlorophyll fluorometers requiring dark adaptation and leaf-level contact, limiting throughput to individual leaves. If $F_v/F_m$ could instead be predicted from RGB imagery, plant health screening could be performed with a commodity camera and edge device, opening practical avenues for real-time crop monitoring and in-situ ecological surveys.

*Plant phenotyping and stress detection.* Deep learning has accelerated non-invasive phenotyping, with CNNs routinely applied to disease detection, growth staging, and trait estimation, and lightweight variants enabling mobile deployment [24, 13, 21, 9]. Most plant-stress studies, however, frame the problem as *classification* rather than *continuous regression.* Deng et al. [5] reached 98.61% on soybean salt stress with ResNet50 fluorescence-feature fusion, and Jeon et al. [14] reached 96.9% with a multimodal 3D-CNN over RGB, depth, and 31 fluorescence channels; both require specialised fluorescence hardware and output categorical labels.

*Camera-based $F_v/F_m$ estimation.* A handful of studies have explored camera-based $F_v/F_m$ estimation, primarily through classical machine learning on hand-crafted vegetation indices. Wu et al. [25] reached $R^2 = 0.925$ on UAV imagery of spring wheat, but multispectral features carried most of the predictive power ($R^2 = 0.838$ for RGB alone). Zolin et al. [29] showed RGB indices such as Excess Green and VARI correlate with $F_v/F_m$ under drought, though the relationship is indirect and species-specific. Jiao and Hu [15] identify red and far-red as most predictive of $F_v/F_m$. Deep learning regression of $F_v/F_m$ directly from RGB remains an open problem.

*Efficient backbones and multi-task learning.* Lightweight architectures such as RepViT [23], MobileMamba [12], and UniForm [28] achieve strong accuracy at low latency on edge hardware including the Jetson family, but have been validated almost exclusively for classification, not continuous trait regression [9, 11]. For joint dense and scalar prediction, shared-encoder multi-task designs offer a natural fit: WeedSense [20] pairs 17-class segmentation with plant-height regression, MT-CYP-Net [16] couples crop-type segmentation with yield regression, and MTDLF [6] combines segmentation with disease-severity regression. Analogous designs exist in medical imaging [4, 18]. None address $F_v/F_m$.

*Gaps and contributions.* Several gaps remain. **(i)** RGB-to-$F_v/F_m$ prediction with deep learning is largely unexplored; existing approaches use classical models with hand-crafted indices or require multispectral sensors. **(ii)** Plant stress is overwhelmingly framed as classification rather than continuous regression. **(iii)** To our knowledge, no public dataset pairs multi-view RGB soybean imagery with ground-truth $F_v/F_m$ under controlled multi-stress conditions. **(iv)** Edge deployment of physiological-trait regression models remains underdeveloped. **(v)** No prior work, to our knowledge, combines plant segmentation and continuous $F_v/F_m$ regression in a single multi-task network.
We address these gaps with three contributions:

1. **Dataset.** A multi-view RGB video dataset of 16 soybean plants under water, light, and nutrient stress from V1 onward, comprising 6,427 frames from 68 videos paired with ground-truth chlorophyll fluorescence measurements (the OJIP transient, a four-phase fluorescence rise on dark-adapted leaves whose inflection points yield $F_v/F_m$ and other photosystem II parameters) collected every 3–5 days from an OS30p+ fluorometer.
2. **Architecture.** *PhotoSoyNet*, a multi-task family with an accuracy variant (PhotoSoyNet-L, 45.66M) and an edge variant (PhotoSoyNet-F, 3.21M), sharing a *Coupled Task Head* that adapts the task-alignment principle of TOOD [8] from object detection to segmentation–regression. Two task-specific mechanisms handle the dense/scalar asymmetry: *Plant-Weighted Spatial Pooling* (PWSP) restricts the regression pool to plant pixels, and *bidirectional cross-task channel gating* lets each branch modulate the other. An ablation isolates these as the source of regression quality.
3. **Edge deployment.** PhotoSoyNet-F predicts $F_v/F_m$ at 47.8 FPS and 302 mJ per inference on the Jetson Orin Nano, making it roughly 10× faster and 11.5× more energy-efficient than the SOTA segmentation baseline, SegFormer-B5, while also achieving substantially better regression quality.

## 2 Dataset and Experimental Setup

*Experimental design.* Sixteen soybean plants were grown in individual pots in a greenhouse at $\sim 33°$C and 50% RH. Plants were sown October 27, 2025 and monitored until December 12, 2025, when the experiment was terminated as the plants began vining and growing too tall, although they had not flowered. At V1 (first fully developed trifoliate leaf), plants were randomly assigned to four groups of four: **Control** (standard watering and fertilization), **Nutrient Stress** (no slow-release fertilizer), **Light Stress** (shading cloth), and **Water Stress** (reduced watering). Treatments and data collection began simultaneously at V1, with measurements every ∼4 days (range 3–6) yielding six sessions from November 4 to November 25, 2025. The treatments produced visibly different growth: Water Stress reached only V3 by experiment end while Control and Nutrient Stress reached V5; Light Stress grew taller but with weaker stems.

*Ground-truth acquisition.* Chlorophyll fluorescence was measured with an OS30p+ portable fluorometer [19], which records the OJIP transient on dark-adapted leaves. Each session produced JIP-test parameters including $F_o$, $F_m$, $F_v/F_m$, $V_j$, and PI. All available leaves of each plant were measured and the resulting $F_v/F_m$ values averaged to one per-plant ground-truth value per session. Healthy plants typically exhibit $F_v/F_m \in [0.75, 0.85]$ [15].

*Image acquisition and segmentation.* Immediately after fluorometer readings, RGB video was captured by orbiting a smartphone around each plant. We recorded 68 videos and extracted 6,427 frames (medians 90–99 per video, balanced across groups). Initial masks were generated with SAM 3 [3] and manually refined

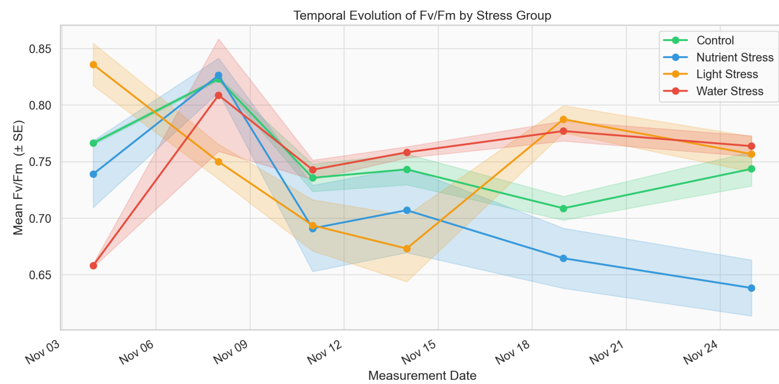


**Fig. 1.** Temporal $F_v/F_m$ per group with SE bands across the six measurement sessions. Only Nutrient Stress shows progressive PSII decline; Water Stress remains comparable to Control throughout, despite producing the most visible morphological effects.

around overlapping leaves and pot rims; SAM is used only at data-prep. Each frame inherits its plant's session-averaged $F_v/F_m$. Raw values ranged 0.60–0.87 and were min-max normalised to $[0, 1]$:

$$\hat{y} = \frac{F_v/F_m - 0.60}{0.87 - 0.60}. \tag{1}$$

Images were resized to 512×512 with a 4,503/991/933 train/val/test split at the video level: each of the 68 videos belongs to one partition.

*Ground-truth characterization.* The properties of the $F_v/F_m$ ground truth determine the difficulty ceiling for any vision-based predictor. Per-group statistics show clear differences in central tendency but substantial overlap: Control has a tight unimodal distribution (median 0.753, IQR 0.72–0.77), Nutrient Stress the lowest median (0.709) with the widest spread and a notably bimodal shape (some plants maintaining healthy PSII function, others falling to the dataset minimum 0.60), and Light and Water Stress overlap substantially with Control (medians 0.744 and 0.757). No pairwise group difference reaches $p < 0.05$ under Mann-Whitney U, and Kruskal-Wallis gives $H = 4.725$, $p = 0.193$. Figure 1 shows the temporal trajectories: all groups peak around November 8 (acclimation) and then diverge. Control declines gently by 6% (baseline drift); Nutrient Stress declines steeply by 22% from peak, reflecting progressive PSII repair-capacity depletion; Light Stress dips by 13% before partially recovering, consistent with photosynthetic acclimation; and Water Stress matches or slightly exceeds Control throughout, showing that three weeks of reduced watering was insufficient to produce measurable PSII damage despite visible morphology change.

This decoupling between visual appearance and photosynthetic efficiency is itself a fundamental challenge for RGB-based prediction. After normalisation, the target has mean 0.53 and standard deviation 0.22, so a mean-predictor yields

MAE $\approx 0.176$. The narrow original-scale range (0.27) means small absolute errors translate to modest $R^2$, and high intra-group variability sets a difficulty ceiling independent of model architecture.

## 3 Method

We propose *PhotoSoyNet*, a multi-task encoder–decoder family for joint plant/background segmentation and per-image $F_v/F_m$ regression. The two tasks have very different signal-to-noise profiles: segmentation is dense and structurally rich, whereas $F_v/F_m$ regression provides one scalar per image and is intrinsically noisy with respect to RGB appearance (Section 2). A naive multi-task network with a shared global pool tends to learn segmentation well but degenerate on regression. The central idea of PhotoSoyNet is therefore a *Coupled Task Head* that ties regression to segmentation by pooling regression features only over predicted-plant pixels, while letting the two branches modulate one another via cross-task channel gates. We instantiate the family at two operating points: PhotoSoyNet-L (45.66M, accuracy-oriented, also used as regression teacher) and PhotoSoyNet-F (3.21M, edge). The variants differ only in encoder and decoder; both use the same head and training recipe.

*Overall architecture.* Given $\mathbf{x} \in \mathbb{R}^{3\times H\times W}$, an encoder $\mathcal{E}$ produces a four-scale pyramid $\{\mathbf{F}_s\}_{s\in\{4,8,16,32\}}$, a decoder $\mathcal{D}$ fuses it into $\mathbf{F}_{\text{dec}} \in \mathbb{R}^{C\times H'\times W'}$, and the Coupled Task Head $\mathcal{H}$ produces both a segmentation logit map $\mathbf{S} \in \mathbb{R}^{2\times H\times W}$ and a scalar $\hat{y} \in [0, 1]$. Auxiliary $1\times1$ heads on each pyramid stage provide deep supervision during training. Figure 2 shows the edge variant.

*Backbones and Decoders.* **PhotoSoyNet-L** pairs RepMSCAN-L (a 34.5M-parameter SegNeXt-style encoder) with LightFPN+Hamburger, which fuses four scales at stride-4 and injects context via iterative matrix factorisation [20]. This four-way $H/4$ concatenation accounts for ∼half the model's FLOPs. **PhotoSoyNet-F** targets the edge, pairing UIB-DualPath [20] (a 1.9M-parameter dual-branch encoder) with FluoSegFastDecoder. To avoid heavy $H/4$ fusion, this decoder operates at stride-8 from a single primary feature, capturing context via a single-pass LKAContext module (effective RF $\sim19\times19$) with an optional $H/32$ semantic injection. This combination is an order of magnitude cheaper with no measurable segmentation loss at the 3.2M-parameter budget.

*Coupled Task Head.* The Coupled Task Head is the architectural contribution and the only component identical across both variants. We borrow the underlying philosophy from TOOD [8], which couples classification and localisation in object detection through task-interactive modules over a shared decoder feature. Our setting differs in two respects: the tasks are heterogeneous in granularity (dense per-pixel versus a single image-level scalar), and $F_v/F_m$ is biologically defined only over plant tissue. We address both with two mechanisms: PWSP and

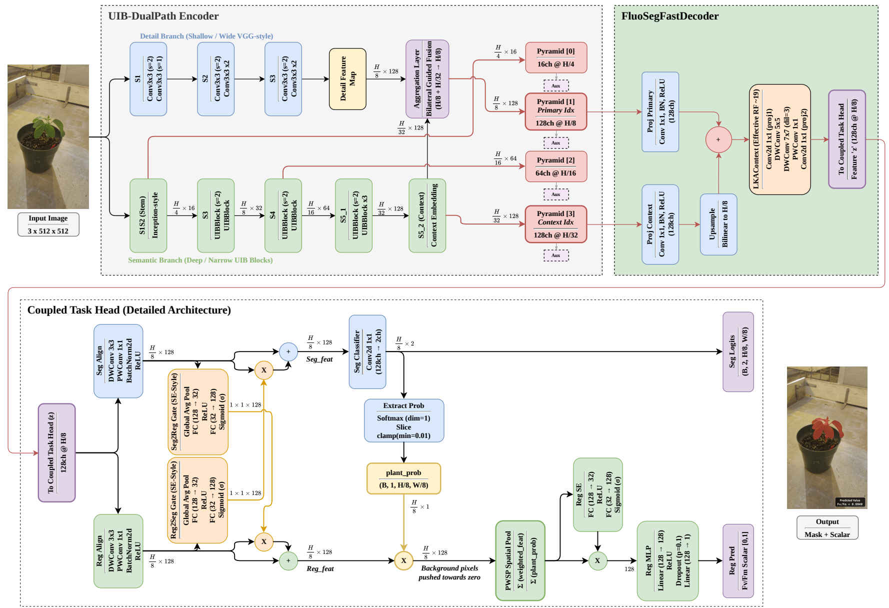

**Fig. 2.** PhotoSoyNet-F architecture. The UIB-DualPath encoder produces $\{\mathbf{F}_4, \mathbf{F}_8, \mathbf{F}_{16}, \mathbf{F}_{32}\}$. FluoSegFastDecoder operates on $\mathbf{F}_8$ with a single-pass LKAContext (effective RF $\sim$19) and an optional $\mathbf{F}_{32}$ injection. The Coupled Task Head couples segmentation and regression via PWSP and bidirectional cross-task gates (inset). PhotoSoyNet-L follows the same template with RepMSCAN-L as encoder and LightFPN+Hamburger as decoder; the head is identical.

bidirectional cross-task channel gating. The design follows two principles: the regression head must not integrate background pixels (since $F_v/F_m$ is plant-only), and the two branches should share information bidirectionally but in a controlled, channel-wise manner.

Two parallel depthwise separable convolutions (DW 3×3 followed by PW 1×1, BN, ReLU) project $\mathbf{F}_{\text{dec}}$ into a segmentation feature $\mathbf{F}_{\text{seg}}$ and a regression feature $\mathbf{F}_{\text{reg}}$, both of shape $C \times H' \times W'$. Each branch then produces a Squeeze-and-Excitation-style channel gate that modulates the *other* branch:

$$\mathbf{g}_{s\to r} = \sigma(\mathbf{W}_2\, \rho(\mathbf{W}_1\, \mathrm{GAP}(\mathbf{F}_{\text{seg}})))\,, \tag{2}$$

$$\mathbf{g}_{r\to s} = \sigma(\mathbf{W}_4\, \rho(\mathbf{W}_3\, \mathrm{GAP}(\mathbf{F}_{\text{reg}})))\,, \tag{3}$$

where $\rho$ is ReLU, $\sigma$ is sigmoid, and $\{\mathbf{W}_i\}$ are linear projections with reduction ratio 4. The features are then modulated via a residual gated connection:

$$\tilde{\mathbf{F}}_{\text{seg}} = \mathbf{F}_{\text{seg}} + (\mathbf{F}_{\text{seg}} \odot \mathbf{g}_{r\to s}), \qquad \tilde{\mathbf{F}}_{\text{reg}} = \mathbf{F}_{\text{reg}} + (\mathbf{F}_{\text{reg}} \odot \mathbf{g}_{s\to r}), \tag{4}$$

with $\odot$ channel-wise multiplication. This lets each branch up- or down-weight the other's channels using its own global context, without exchanging spatial information directly. $\tilde{\mathbf{F}}_{\text{seg}}$ is passed through a 1×1 classifier and bilinearly upsampled to give $\mathbf{S}$.

For regression, let $p(h,w) = \text{softmax}(\mathbf{S})_{\text{plant}}(h,w)$ be the plant probability at the head's working resolution, clamped to a minimum of 0.01 to ensure stable gradient flow early in training. PWSP averages the gated regression feature spatially, weighted by $p$:

$$\mathbf{f}_{\text{pwsp}} = \frac{\sum_{h,w} p(h,w) \cdot \tilde{\mathbf{F}}_{\text{reg}}(:,h,w)}{\sum_{h,w} p(h,w) + \varepsilon}, \tag{5}$$

with $\varepsilon = 10^{-6}$. The denominator normalises by total plant mass, making $\mathbf{f}_{\text{pwsp}}$ invariant to plant area. The pooled vector undergoes a final channel recalibration via an SE block, then passes through a small MLP (one hidden layer of 128 units, ReLU, dropout 0.1) to produce the scalar $\hat{y}$.

PWSP makes the regression prediction *differentiable through the segmentation probabilities*: gradient from the regression loss flows back through $p(h,w)$ to the segmentation classifier, encouraging masks that yield useful regression pools. Cross-task gating provides a second, channel-wise coupling. Together they implement the principle that segmentation quality should bound regression quality by construction. We isolate their contribution in Section 6 by replacing the head with a vanilla GAP regressor.

## 4 Experimental Setup

*Training protocol.* Both PhotoSoyNet variants are trained end-to-end on the 4,503-image training split for 100 epochs with AdamW [17] (lr $6\times10^{-5}$, weight decay 0.01, $\beta = (0.9, 0.999)$), 5-epoch linear warm-up, polynomial decay (power 1.0) to zero, batch 16 at 512×512, and mixed precision. An EMA [22] of student weights with decay 0.9998 is maintained throughout and used at evaluation. Augmentations are deliberately conservative, since aggressive perturbation risks destroying $F_v/F_m$-informative cues: random resize-and-crop (scale 0.5–2.0), horizontal flip ($p = 0.5$), photometric jitter (±20%), and ImageNet normalisation. Vertical flips and rotations are not used: greenhouse imagery has consistent gravitational orientation and rotation hurt in preliminary runs. The joint loss combines OHEM cross-entropy for segmentation (keep-ratio 0.25), smooth $\ell_1$ for regression ($\beta = 0.1$, weight 1.0), and auxiliary segmentation heads (weight 0.4). PhotoSoyNet-F additionally uses two-teacher distillation (segmentation weight 0.5, regression weight 0.3), whereas PhotoSoyNet-L is trained without distillation since it serves as the regression teacher. Validation runs every epoch and the best-mIoU checkpoint is used at test time.

*Baseline protocol.* To ensure fair comparison, *every baseline is trained from scratch on `fluosoy_v2` under the exact same schedule and data pipeline.* Because pure-segmentation baselines lack regression heads, we attach a generic *MLP+Attention* head (an SE-block, followed by a 512-to-256 MLP and a sigmoid scalar) to their deepest backbone feature. This represents the standard approach to retrofitting regression, pooling globally without plant-aware spatial restriction. We evaluate three families: **SegFormer** [27] (B0/B2/B5), **SegNeXt** [10] (T/S/L), and **UPerNet** [26] (Swin-B/Swin-L). **WeedSense** [20] is also included but retains its native multi-task design. All use official ImageNet weights, a 50-epoch mmsegmentation schedule, and cross-entropy/smooth $\ell_1$ losses.

*Metrics.* Models are evaluated on the 933-image test split. Segmentation: mIoU averaged over the two classes and plant IoU (pIoU), the latter being the harder and more biologically relevant of the two. Regression (in normalised $[0, 1]$ space): MAE and $R^2$. A mean-predictor yields MAE $\approx 0.176$ and $R^2 = 0$, so any $R^2 < 0$ indicates worse than trivial.

*Hardware and edge benchmarking.* Training and validation use a single NVIDIA A100 80GB with mixed-precision training (AMP). Edge benchmarking uses an NVIDIA Jetson Orin Nano 8GB at the `MAXN_SUPER` profile (PyTorch 2.8, CUDA 12.6) in FP32 at batch 1 and 512×512, the same conditions a deployed model would face on a single live frame. We deliberately avoid TensorRT, ONNX, or vendor-specific optimisation: these toolchains favour particular operator graphs and would make the comparison about engineering effort rather than model design. For each model we run 20 warm-up iterations followed by 100 timed forward passes; PhotoSoyNet variants are benchmarked in both *normal* and *deploy* (Conv–BN fused) mode, baselines in normal mode only. Power is read from the Jetson `VDD_IN` rail via `jetson-stats` synchronised with the forward passes; we report average power $\bar{P}$ and energy per inference $E_{\text{inf}} = \bar{P} \cdot \text{ms\_per\_image}$. A 60-second cool-down between models prevents thermal throttling. FLOPs are computed with `fvcore` [7] on 1×3×512×512 inputs, counting MACs as two FLOPs to match SegFormer/SegNeXt convention.

*Reproducibility.* All experiments use seed 42 and deterministic CuDNN; code, configs, all weights, and the Jetson harness will be released on acceptance.

## 5 Results

Table 1 reports both axes of comparison: segmentation/regression quality on `fluosoy_v2`, and Jetson edge inference. All numbers are produced under the protocols of Sections 4–4.

*Main comparison.* The headline observation is consistent across the baseline lineup: a stronger segmentation backbone improves segmentation but not $F_v/F_m$ regression. SegFormer-B5 (82M, pIoU 0.945) attains $R^2 = -0.018$, worse than the mean, and UPerNet-Swin-L (235M, pIoU 0.942) reaches $R^2 = -0.113$; across the eight non-WeedSense baselines the plant-IoU/$R^2$ correlation is mildly *negative*. PhotoSoyNet-L instead gives up ∼6.5 pp plant IoU (0.877 vs. 0.942) for MAE 0.108 and $R^2 = 0.446$, roughly twice the best baseline (SegNeXt-L,

**Table 1.** Main comparison on `fluosoy_v2`. Baselines use the unified MLP+Attention head (Section 4); WeedSense uses its native multi-task design. Edge metrics: NVIDIA Jetson Orin Nano (FP32, batch 1, 512×512); PhotoSoyNet variants in deploy mode. **Bold**: best per column among edge-deployable models ($\leq$ 15M); †: best across all evaluated models regardless of parameter budget.

| Method | Params | FLOPs | Segmentation | | Regression | | Edge | |
|---|---|---|---|---|---|---|---|---|
| | (M) | (G) | mIoU | pIoU | MAE | $R^2$ | FPS | mJ/inf. |
| SegFormer-B0 | 3.99 | 7.9 | **0.958** | **0.921** | 0.141 | 0.060 | 29.8 | 525 |
| SegFormer-B2 | 25.15 | 25.2 | 0.963 | 0.931 | 0.142 | 0.051 | 12.2 | 1358 |
| SegFormer-B5 | 82.40 | 74.6 | 0.970† | 0.945† | 0.159 | −0.018 | 4.9 | 3473 |
| SegNeXt-T | 4.50 | 6.3 | 0.951 | 0.908 | 0.143 | 0.216 | 18.5 | 723 |
| SegNeXt-S | 14.32 | 15.6 | 0.957 | 0.919 | 0.144 | 0.195 | 13.0 | 1051 |
| SegNeXt-L | 49.20 | 65.5 | 0.963 | 0.931 | 0.127 | 0.218 | 4.6 | 3184 |
| UPerNet-Swin-B | 121.95 | 298.2 | 0.966 | 0.937 | 0.153 | 0.024 | 2.5 | 7668 |
| UPerNet-Swin-L | 234.86 | 406.9 | 0.969 | 0.942 | 0.160 | −0.113 | 1.6 | 11684 |
| WeedSense | 12.72 | 12.7 | 0.950 | 0.907 | 0.149 | −0.116 | 43.2 | 350 |
| PhotoSoyNet-L | 45.66 | 313.2 | 0.934 | 0.877 | 0.108† | 0.446† | 2.6 | 6724 |
| PhotoSoyNet-F | 3.21 | 7.8 | 0.936 | 0.880 | **0.120** | **0.333** | **47.8** | **302** |

**Table 2.** PhotoSoyNet edge profiles on the Jetson Orin Nano (FP32, batch 1, 512×512, MAXN_SUPER). Power is averaged over 100 timed forward passes after a 20-iteration warm-up.

| Model | Mode | FPS | ms/img | $\bar{P}$ (W) | mJ/inf. |
|---|---|---|---|---|---|
| PhotoSoyNet-L | normal | 2.41 | 415.6 | 17.76 | 7380 |
| PhotoSoyNet-L | deploy | 2.64 | 378.3 | 17.77 | 6724 |
| PhotoSoyNet-F | normal | 38.61 | 25.9 | 12.96 | 336 |
| PhotoSoyNet-F | deploy | 47.75 | 20.9 | 14.41 | 302 |

0.218); PhotoSoyNet-F preserves most of this (MAE 0.120, $R^2 = 0.333$) at one-fourteenth the parameters. We attribute the baseline failure to global pooling: the MLP+Attention head integrates background pixels that carry no $F_v/F_m$ signal and only add variance. PWSP changes the pool's support to the plant region; cross-task gating adds channel-wise coupling. WeedSense, the only native multi-task baseline, still gets $R^2 = -0.116$: its temporal-fusion decoder does not transfer to single-image $F_v/F_m$, leaving PhotoSoyNet-F as the only model with both real-time throughput *and* non-trivial $R^2$.

*Edge deployment.* Table 2 reports normal- and deploy-mode profiles for both PhotoSoyNet variants, isolating the speedup from Conv–BN fusion at inference time.

PhotoSoyNet-F reaches 47.8 FPS at 302 mJ/inf., the most efficient profile in the table: versus SegFormer-B5 it is 9.7× faster and 11.5× more energy-efficient at 24% lower MAE; versus SegFormer-B0 (closest in size), 5× higher $R^2$ (0.333 vs. 0.060) at 1.6× the speed; versus WeedSense, comparable efficiency but $R^2$ moved from −0.116 to 0.333. Conv–BN fusion is mathematically lossless yet model-dependent: PhotoSoyNet-F gains +24% FPS and −10% energy (a raceto-idle effect from shorter, busier inference), while PhotoSoyNet-L gains only +10%/−9%, its runtime dominated by the LightFPN+Hamburger decoder. L itself sits with the heaviest baselines (2.6 FPS, 313 GFLOPs); it is the accuracy-first teacher, not an edge model. PhotoSoyNet-F preserves 90% of L's regression quality ($R^2 = 0.333$ vs. 0.446) at 14× compression and 18× throughput.

### 5.1 Qualitative Results

Figure 3 shows two representative test images: the case where PhotoSoyNet-F achieves its largest regression-margin advantage over every baseline (left), and the only case where it also out-segments every baseline (right). Heavier baselines (here SegNeXt-L at 49M and WeedSense as the only other multi-task model) produce strong plant masks but predict $F_v/F_m$ values clustered near a constant, far from the per-plant ground truth, mirroring their negative or near-zero $R^2$ on the full test set. PhotoSoyNet-F tracks the per-plant $F_v/F_m$ markedly better while keeping segmentation competitive. The visual examples illustrate the same trend as the quantitative results: strong segmentation alone does not guarantee accurate $F_v/F_m$ regression.

## 6 Ablation Study

The Method (Section 3) attributes PhotoSoyNet's joint-task quality to the Coupled Task Head, specifically PWSP and bidirectional cross-task channel gating, and the Results showed that standard backbones with a generic pooled regression head fail to recover non-trivial $F_v/F_m$ regression even at 235M parameters. The natural question is whether the head fixes this or whether other co-occurring choices (encoder, decoder, distillation, auxiliary supervision) explain the gap. We isolate the head's contribution with three complementary studies: a progression study tracing PhotoSoyNet-F's metrics across the design changes that built it (Table 3); a head ablation replacing the Coupled Task Head with a vanilla GAP regressor (Table 4); and a component-isolation study disabling PWSP or gating individually (Table 5). All runs use the protocol of Section 4.

*Component-wise progression.* Adding two-teacher distillation (a→b) lifts plant IoU by 1.2 pp and $R^2$ by 0.073 with no compute change, the expected behaviour of pairwise distillation when teachers exceed the student. The decisive jump is swapping LightFPN+Hamburger for FluoSegFastDecoder (b→c): plant IoU rises 6.3 pp, MAE improves marginally, $R^2$ slips slightly to 0.333, but parameters drop 5.74 → 3.21M and FPS rises 15.8 → 47.8, a 3× throughput gain at 44% of the parameter count. This is PhotoSoyNet-F. Two further changes branching off (c) were rejected. Two rejected branches off (c) confirm this: richer deep supervision with Lovász softmax [1] (c→d) and a 1.25× wider encoder (c→e) both nudge plant IoU up but leave regression flat or worse (Table 3).

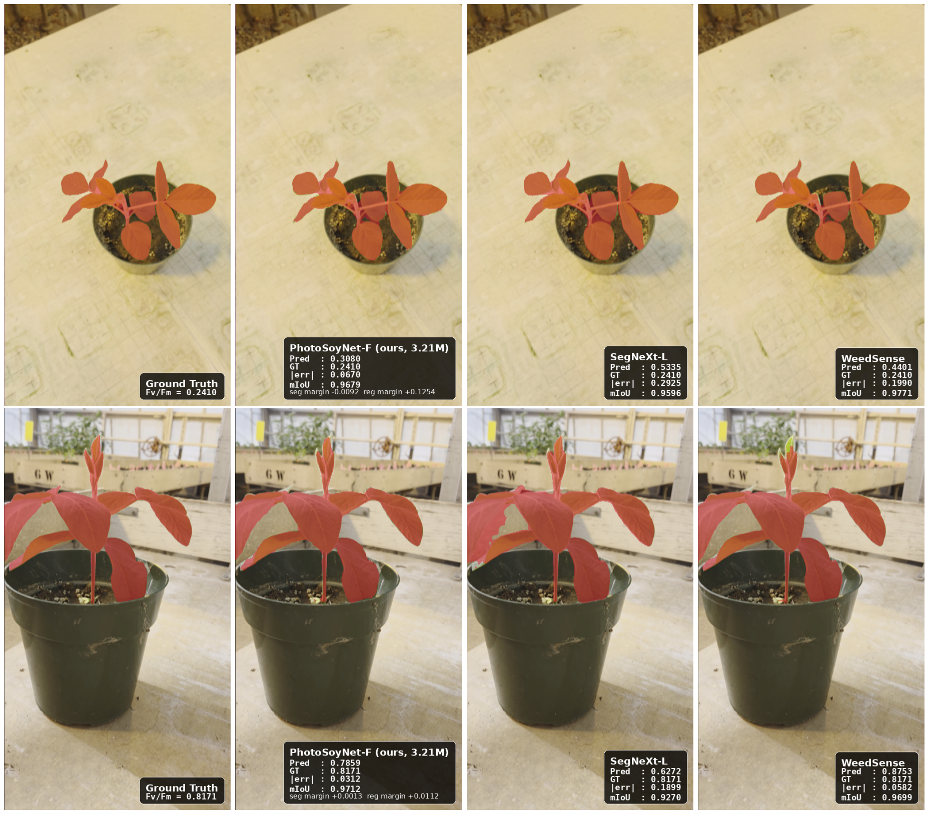


**Fig. 3.** Qualitative comparison on two representative test images. Columns: ground truth, PhotoSoyNet-F, SegNeXt-L, and WeedSense. Top row: case where PhotoSoyNet-F achieves its largest regression-margin advantage. Bottom row: case where PhotoSoyNet-F also out-segments the baselines. Masks are overlaid at 50% opacity; embedded cards report model name, predicted $F_v/F_m$, ground-truth $F_v/F_m$, absolute error, and mIoU.

*Head ablation.* We replace the Coupled Task Head with the simplest natural alternative: a GAP head that takes the same decoder feature, applies channel attention, and pools globally. Eq. 5 becomes $\mathbf{f}_{\text{gap}} = \frac{1}{H'W'}\sum_{h,w}\mathbf{F}_{\text{reg}}(:, h, w)$, and the two channel gates (Eq. 2, Eq. 3) are removed; encoder, decoder, distillation, optimiser, schedule, augmentations, and EMA are held fixed. Segmentation is unchanged (+0.2 pp plant IoU), while regression degrades on both metrics: MAE +8.3% and $R^2$ −31% relative. The GAP-head ablation at $R^2 = 0.229$ is comparable to SegNeXt-L (0.218 at 49M) and better than SegFormer-B5 (−0.018 at 82M), showing that the encoder, decoder, and distillation produce a regression-friendly representation, but one that hits a ceiling the standard pooled regressor cannot break through. The Coupled Task Head changes the pool's support from the entire image to the predicted plant region and lets the branches

**Table 3.** Component-wise progression on PhotoSoyNet-F. Rows (a)–(c) trace the build-up to the final model; rows (d) and (e) are branching experiments off row (c) that were rejected. **Bold**: row promoted to the final model. "Hvy. dec." is LightFPN+Hamburger; "Fast dec." is FluoSegFastDecoder. FPS is in deploy mode where applicable.

| Stage | Configuration | Change vs. baseline | Params (M) | pIoU | MAE | $R^2$ | FPS |
|---|---|---|---|---|---|---|---|
| (a) | UIB enc. + Hvy. dec., no KD | base | 5.74 | 0.805 | 0.129 | 0.285 | 15.8 |
| (b) | + KD (two teachers) | vs. (a): add seg+reg distillation | 5.74 | 0.817 | 0.122 | 0.358 | 15.8 |
| **(c)** | **+ FluoSegFastDecoder** | **vs. (b): swap decoder to LKA at $H/8$** | **3.21** | **0.880** | **0.120** | **0.333** | **47.8** |
| (d) | + Rich aux heads + Lovász | vs. (c): richer deep supervision | 3.56 | 0.882 | 0.125 | 0.238 | 42.1 |
| (e) | + Wider encoder (×1.25) | vs. (c): widen UIB-DualPath | 5.01 | 0.884 | 0.126 | 0.245 | 42.4 |

**Table 4.** Head ablation. PhotoSoyNet-F vs. the same network with the Coupled Task Head replaced by a GAP head with no cross-task gating. Both rows share encoder, decoder, distillation, and training protocol.

| Configuration | pIoU | MAE | $R^2$ | FPS |
|---|---|---|---|---|
| PhotoSoyNet-F (Coupled Task Head) | 0.880 | 0.120 | 0.333 | 47.8 |
| PhotoSoyNet-F-GAP (- PWSP, - gating) | 0.882 | 0.130 | 0.229 | 51.8 |
| $\Delta$ (ablation − full) | +0.002 | +0.010 | −0.104 | +4.0 |

modulate each other channel-wise, and that single change carries the model from the SegNeXt-L operating point to $R^2 = 0.333$.

*Component isolation.* To disentangle the contributions of PWSP and gating within the head, we train two further variants disabling each mechanism individually (Table 5). The result is sharper than expected: the two mechanisms are *synergistic, not additive*, and neither is sufficient on its own. Gating alone (no spatial restriction) achieves $R^2 = 0.214$, broadly matching the GAP-head baseline (0.229) and indicating that bidirectional channel modulation, applied to a globally pooled feature, offers limited benefit beyond what an SE-style attention already provides. PWSP alone (no gating) is more striking: it *collapses* regression to $R^2 = -0.460$, substantially worse than both the GAP baseline and a mean predictor. We attribute this to a noise-amplification effect: restricting the regression pool to the predicted-plant region narrows the feature support to a small, segmentation-noisy subset of channels, and without gating's channel-wise context the regressor has neither a stable spatial signal nor a way to recover global statistics. Cross-task gating is what makes mask-restricted pooling safe by reweighting the regression channels using global segmentation context. Segmentation is unchanged across all four variants (±0.5 pp pIoU), confirming the head's asymmetric role. The takeaway: PWSP encodes the correct biological prior (regression is plant-only) but is unstable in isolation; gating supplies the channel-level context that stabilises it; only together do they yield the full $R^2 = 0.333$.

**Table 5.** Component isolation within the Coupled Task Head. "PWSP only" retains the mask-weighted regression pool but removes both channel gates; "gating only" retains the gates but replaces PWSP with a global average over the full feature map. The GAP-head row from Table 4 is included for reference. All variants share encoder, decoder, distillation, and training protocol.

| Configuration | pIoU | MAE | $R^2$ |
|---|---|---|---|
| PhotoSoyNet-F (PWSP + gating) | 0.880 | 0.120 | 0.333 |
| PWSP only (no gating) | 0.873 | 0.153 | −0.460 |
| Gating only (no PWSP) | 0.877 | 0.127 | 0.214 |
| GAP head (neither, Table 4) | 0.882 | 0.130 | 0.229 |

# 7 Limitations and Conclusion

*Limitations and Conclusion.* We presented PhotoSoyNet, a multi-task family for joint plant segmentation and per-image $F_v/F_m$ regression from RGB, built around a Coupled Task Head that adapts TOOD's task-alignment principle [8] to a heterogeneous dense/scalar setting through PWSP and bidirectional cross-task gating. Ablating both mechanisms leaves segmentation nearly unchanged but worsens regression by +8.3% MAE and −31% $R^2$ relative, showing that the head is the main structural enabler. Across nine re-trained baselines (4–235M parameters), no baseline exceeds $R^2 = 0.25$, while PhotoSoyNet-L reaches $R^2 = 0.446$ and PhotoSoyNet-F reaches $R^2 = 0.333$ at one-fourteenth the size. On Jetson Orin Nano, PhotoSoyNet-F runs at 47.8 FPS and 302 mJ/inf., making it the only model that is both real-time and non-trivially predictive of $F_v/F_m$. By moving physiological-trait estimation from contact fluorometers onto commodity smartphones and edge devices, this work lowers the cost barrier for plant health screening and establishes RGB-based $F_v/F_m$ prediction as a viable triage tool (flagging stressed plants for targeted follow-up rather than replacing fluorometric measurement) with applications spanning precision agriculture, ecosystem monitoring, and in-situ ecological surveys. These conclusions are bounded by five limitations: per-plant averaged labels impose a likely noise floor (view-level appearance is not matched by view-level physiology); the split is video- not plant-level (though same-plant train/test videos are always $\geq 4$ days apart); only Nutrient Stress is clearly separable on $F_v/F_m$, so $R^2$ may partly reflect near-binary stress discrimination; the data covers one species, one greenhouse, and a V1–V5 window; and edge benchmarks use FP32 PyTorch without TensorRT/ONNX. Future work will expand species and growth stages, collect finer-grained physiological labels including per-leaf phenotyping, deploy and evaluate the system in open-field conditions, conduct per-group and per-plant temporal analyses to separate continuous $F_v/F_m$ tracking from stress-state discrimination, add an auxiliary stress-classification branch to the Coupled Task Head, and quantify deployment gains from engineering-toolchain optimisation.

**Acknowledgements** This work was supported by the SIUC Soybean Center 2025 grant program (Award #202196). 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 [2], which is supported by National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296.

## References

1. Berman, M., Triki, A.R., Blaschko, M.B.: The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4413–4421 (2018)
2. Boerner, T.J., Deems, S., Furlani, T.R., Knuth, S.L., Towns, J.: ACCESS: Advancing innovation: NSF's advanced cyberinfrastructure coordination ecosystem: Services & support. In: Practice and Experience in Advanced Research Computing (PEARC '23). pp. 173–176. ACM (2023)
3. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., Lei, J., Ma, T., Guo, B., Kalla, A., Marks, M., Greer, J., Wang, M., Sun, P., Rädle, R., Afouras, T., Mavroudi, E., Xu, K., Wu, T.H., Zhou, Y., Momeni, L., Hazra, R., Ding, S., Vaze, S., Porcher, F., Li, F., Li, S., Kamath, A., Cheng, H.K., Dollár, P., Ravi, N., Saenko, K., Zhang, P., Feichtenhofer, C.: Sam 3: Segment anything with concepts (2025), https://arxiv.org/abs/2511.16719
4. Cheng, J., Liu, J., Kuang, H., Wang, J.: A fully automated multimodal MRI-based multi-task learning for glioma segmentation and IDH genotyping. IEEE Transactions on Medical Imaging **41**(6), 1520–1532 (2022)
5. Deng, Y., Xin, N., Zhao, L., Shi, H., Deng, L., Han, Z., Wu, G.: Precision detection of salt stress in soybean seedlings based on deep learning and chlorophyll fluorescence imaging. Plants **13**(15), 2089 (2024)
6. Devi, A.G., Begum, S.S., Kocharla, S., Madhavi, P., Gorikapudi, S., Tirumalasetti, N.R.: Multi-task deep learning framework for segmentation and severity estimation of leaf diseases in multi-crop environments. International Journal of Innovative Technology and Interdisciplinary Sciences **9**(1), 210–237 (2026)
7. Facebook AI Research: fvcore: Collection of common code shared among different research projects in fair computer vision team. https://github.com/facebookresearch/fvcore, accessed: 2026-04-27
8. Feng, C., Zhong, Y., Gao, Y., Scott, M.R., Huang, W.: Tood: Task-aligned one-stage object detection. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3490–3499. IEEE Computer Society (2021)
9. Gunasekaran, H., Rajkumar, S., et al.: Lightweight deep learning for tomato disease detection: trends, challenges, and edge AI perspectives. Frontiers in Plant Science **16**, 1737208 (2026)
10. Guo, M.H., Lu, C.Z., Hou, Q., Liu, Z., Cheng, M.M., Hu, S.M.: Segnext: Rethinking convolutional attention design for semantic segmentation. Advances in neural information processing systems **35**, 1140–1156 (2022)
11. Han, S., Wang, H.: Application of deep learning technology in monitoring plant attribute changes. Sustainability **17**(17), 7602 (2025)
12. He, H., Zhang, J., Cai, Y., Chen, H., Hu, X., Gan, Z., Wang, Y., Wang, C., Wu, Y., Xie, L.: Mobilemamba: Lightweight multi-receptive visual mamba network. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 4497–4507 (2025)

13. Islam, S., Reza, M.N., Samsuzzaman, S.A., Cho, Y.J., Noh, D.H., Chung, S.O., Hong, S.J.: Machine vision and artificial intelligence for plant growth stress detection and monitoring: A review. Precision Agriculture **6**(1), 34 (2024)
14. Jeon, Y.J., Kim, H.S., Lee, T.S., Park, S.H., Yun, H., Jung, D.H.: Multimodal optical biosensing and 3D-CNN fusion for phenotyping physiological responses of basil under water deficit stress. Agronomy **16**(1), 55 (2025)
15. Jiao, Q., Hu, X.: Recent advances and emerging trends in chlorophyll fluorescence parameter Fv/Fm. Phyton **94**(9) (2025)
16. Liu, S., Wang, D., Guo, H., Han, C., Zeng, W.: Mt-cyp-net: Multi-task network for pixel-level crop yield prediction under very few samples. International Journal of Applied Earth Observation and Geoinformation **143**, 104748 (2025)
17. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019), https://openreview.net/forum?id=Bkg6RiCqY7
18. Marcus, A., Bentley, P., Rueckert, D.: Concurrent ischemic lesion age estimation and segmentation of CT brain using a transformer-based network. IEEE Transactions on Medical Imaging **42**(12), 3464–3473 (2023)
19. Opti-Sciences, Inc.: OS30p+ chlorophyll fluorometer (2025), https://opti-sciences.com/products/chlorophyll-fluorometers/os30p/, accessed: 2026-04-23
20. 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. pp. 7180–7190 (2025)
21. Song, X., Yan, L., Liu, S., Gao, T., Han, L., Jiang, X., Jin, H., Zhu, Y.: Agricultural image processing: Challenges, advances, and future trends. Applied Sciences **15**(16) (2025)
22. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: Advances in Neural Information Processing Systems. vol. 30, pp. 1195–1204 (2017)
23. 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)
24. Wang, T., Tong, R., Xu, T., Li, Y., Chen, Y.: Artificial intelligence in plant science: From image-based phenotyping to yield and trait prediction. Frontiers in Plant Science **16**, 1732979 (2026)
25. Wu, Q., Zhang, Y., Xie, M., Zhao, Z., Yang, L., Liu, J., Hou, D.: Estimation of Fv/Fm in spring wheat using UAV-based multispectral and RGB imagery with multiple machine learning methods. Agronomy **13**(4), 1003 (2023)
26. Xiao, T., Liu, Y., Zhou, B., Jiang, Y., Sun, J.: Unified perceptual parsing for scene understanding. In: Proceedings of the European conference on computer vision (ECCV). pp. 418–434 (2018)
27. 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)
28. Yeom, S.K., Kim, T.H.: Uniform: A reuse attention mechanism optimized for efficient vision transformers on edge devices. arXiv preprint arXiv:2412.02344 (2024)
29. Zolin, Y., Popova, A., Yudina, L., Grebneva, K., Abasheva, K., Sukhov, V., Sukhova, E.: RGB indices can be used to estimate NDVI, PRI, and Fv/Fm in wheat and pea plants under soil drought and salinization. Plants **14**(9), 1284 (2025)