

CWRepViT-Net: An Encoder-Decoder Deep Learning Framework with RepViT Blocks for Crop Weed Semantic Segmentation in Soybean Fields through their Life Journey

Masoomeh Gomroki¹, Dilshan I. Benaragma¹, Christopher J. Henry², Nasem Badreldin³ & Robert H. Gulden¹



¹Department of Plant Science, University of Manitoba, Winnipeg Manitoba Canada
²Department of Computer Science, University of Manitoba, Winnipeg Manitoba Canada
³Department of Soil Science, University of Manitoba, Winnipeg Manitoba Canada
 Contact: Rob.Gulden@umanitoba.ca



Introduction

With the rising global population, the demand for agricultural products is increasing, necessitating advanced technologies for efficient crop management. This study focuses on precision weed management using drone imagery and Deep Learning (DL) during early soybean growth stages. Drone images were collected at six intervals (21, 26, 33, 39, 45 and 52 days after seeding) to distinguish between five classes: soil, soybean, volunteer canola, broadleaf weeds, and grassy weeds. A deep learning model, Crop-Weed-RepViT-Net (CWRepViT-Net), based on an encoder-decoder architecture using RepViT and Modified UNet blocks, was developed for semantic segmentation. The five-step framework includes data preprocessing, training the network, segmentation, evaluation, and prediction. CWRepViT-Net achieved over 95% accuracy and a Kappa coefficient of 0.91, demonstrating high reliability in early-stage crop-weed differentiation.

The main objective of this study was to present a DL model for semantic segmentation of crops and weeds.

Key Findings

- The DL model overcomes the challenges of semantic segmentation at early crop and weed developmental stages including occlusion of leaves, and illumination effect and is robust against different environmental conditions and computationally efficient.
- The DL model benefits from the advantages of the Semi-Transfer Learning technique which improves the network training speed and enables it to overcome the limitation of GPU capacities.
- The network effectively segments into five classes including soybean as crop, volunteer canola, other broadleaf weeds, all grassy weeds and soil.

Results & Discussion

The CWRepViT-Net and other models were implemented using TensorFlow 2.10.1 and Python 3.9.21, and trained on a system with an NVIDIA RTX 4070 GPU, Intel i7-14700 CPU, and 64GB RAM. All networks were trained for 100 epochs with a batch size of 16 using the Adam optimizer and focal loss for multiclass semantic segmentation as follow (Gomroki et al., 2025; Liu et al., 2018):

$$L_{focal\ loss} = - \sum_{i=1}^c \alpha_i (1 - y_i)^{\gamma} t_i \log(y_i)$$

where c corresponds to the number of classes, t_i denotes the true probability distribution, y_i represents the probability distribution of prediction, γ is a hyperparameter of the loss function, which in this study equals to 2 and α_i is another hyperparameter representing the class weights.

Table 1 compares the performance of various encoder-decoder networks. CWRepViT-Net maintains efficient training speed despite having 13.63M parameters. The proposed CWRepViT-Net achieved the highest overall accuracy (95.87%) and Kappa coefficient (0.91), outperforming all other models in precision, F1-score, and IoU, demonstrating its effectiveness and high proficiency in segmenting crops and weeds in soybean fields.

Table 1. Quantitative evaluation results for the comparison of CWRepViT-Net with state-of-the-art models.

Method	Accuracy (%)	Kappa Coefficient (KC)	Precision (%)	F1-score (%)	Intersection over Union IoU (%)	Time of training (min sec)	Parameters (Million)	
							original form	encoder-decoder form
MobileNetV3	94.11	0.89	98.68	98.47	96.98	22min 45sec	5.5	6.5
MobileViT	92.65	0.87	97.72	98.05	96.17	34min 50sec	5.5	9.7
TinyNet	91.77	0.86	98.74	97.88	95.75	29min 35sec	6.2	7.2
TinyViT	92.20	0.86	98.61	98.01	96.01	32min 55sec	11.0	11.2
CWRepViT-Net (Proposed method)	95.87	0.91	98.95	98.76	97.35	33min 10sec	14.23	13.63

The dataset includes five classes soil, soybean, volunteer canola, other broadleaf weeds, and grassy weeds captured from the seedling to early reproductive stages of soybean growth. As shown in the confusion matrices, the proposed method achieved over 90% accuracy for most classes, though segmenting other broadleaf weeds remained the most challenging due to their visual similarity to soybean and volunteer canola, a difficulty also observed in the performance of the other networks.



Figure 2. Confusion Matrix of: (a) proposed method, (b) MobileNetV3, (c) MobileViT, (d) TinyNet and (e) TinyViT. In confusion matrices, canola, weed type I and weed type II stand for volunteer canola, other broadleaf weeds and all grassy weeds, respectively.

The CWRepViT-Net effectively segmented and distinguished visually similar species such as volunteer canola and broadleaf weeds (e.g., Redroot pigweed, Canada thistle, Lambs quarters) during early growth stages, outperforming other networks in challenging scenarios like poor illumination, overlapping leaves, and dense weed-crop mixtures. While other models struggled with accurate boundary preservation and class confusion, the proposed method consistently maintained precise segmentation, even under complex conditions illustrated in Figures 3(a-e).

Materials & Methods

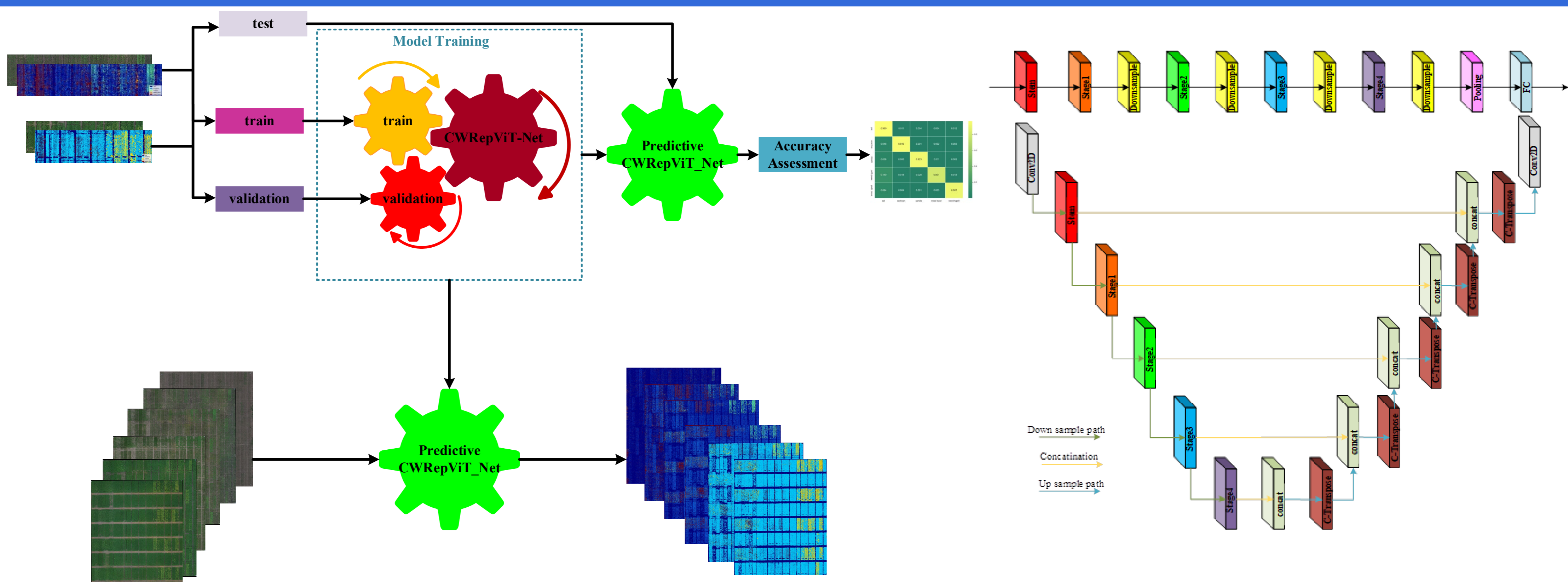


Figure 1. Left: Flowchart of proposed method for crop weed semantic segmentation. The rectangles indicate the respective test, training and validations sets. The yellow and red cogs indicate training and validation training process. They act like two powerful arms to generate the trained model which is shown by the maroon cog, and the green cog indicates the final model used for prediction, which is depicted via both an accuracy assessment (top right), and further prediction of field images (bottom). Right: Top: the structure of RepViT, which contains one stem and four stages (represented by different colours). Bottom: the structure of CWRepViT-Net using RepViT as an encoder path and MUNet as a decoder path. The encoder path is concatenated with the decoder path at five different resolutions (Stem, Stage1, Stage2, Stage3 and Stage4)

The proposed method involves three main steps: (1) pre-processing raster data, (2) training the CWRepViT-Net for crop and weed semantic segmentation, and (3) making predictions to assess the performance of the network and generate crop and weed maps during the soybean vegetative and reproductive stage. The workflow and data split 60% for training, 20% for validation, and 20% for testing are shown in Figure 1 (left). In the pre-processing stage, georeferenced orthophoto mosaics were generated, subdivided into 128x128 pixel sub-images, normalized, split into training, validation, and test sets, and augmented through rotations to enhance model generalization for deep learning input. The RepViT model (Revisiting Mobile CNN from a ViT Perspective) was proposed by Wang et al. (2024) for the first time. They revisited lightweight convolutional neural networks CNNs from the perspective of Vision Transformers (ViT) and introduced a revised model to leverage the

advantages of both architectures (Wang et al., 2024). In this study, a modification of UNet (MUNet) blocks are used as the decoder path as defined by Gomroki, Hasanlou, & Chanussot, 2023.

The CWRepViT-Net is a lightweight encoder-decoder deep learning network with about 15 million parameters. It uses pre-trained RepViT blocks in the encoder and MUNet blocks in the decoder to accurately segment soil, soybean, volunteer canola, broadleaf weeds, and grassy weeds from RGB drone images—without post-processing. By combining RepViT's hierarchical feature extraction with Conv2D-Transpose for spatial reconstruction, the model achieves high segmentation accuracy and fast training across different soybean growth stages. The network structure is illustrated in Figure 1 (right).

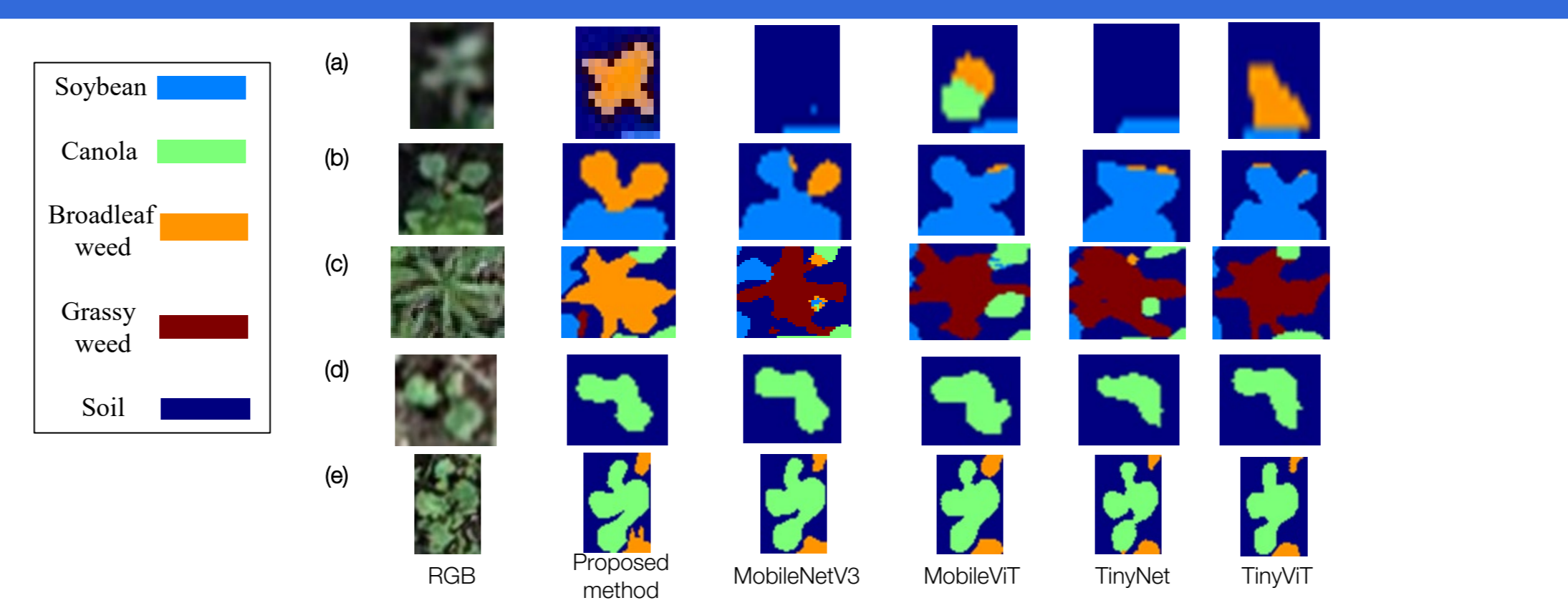
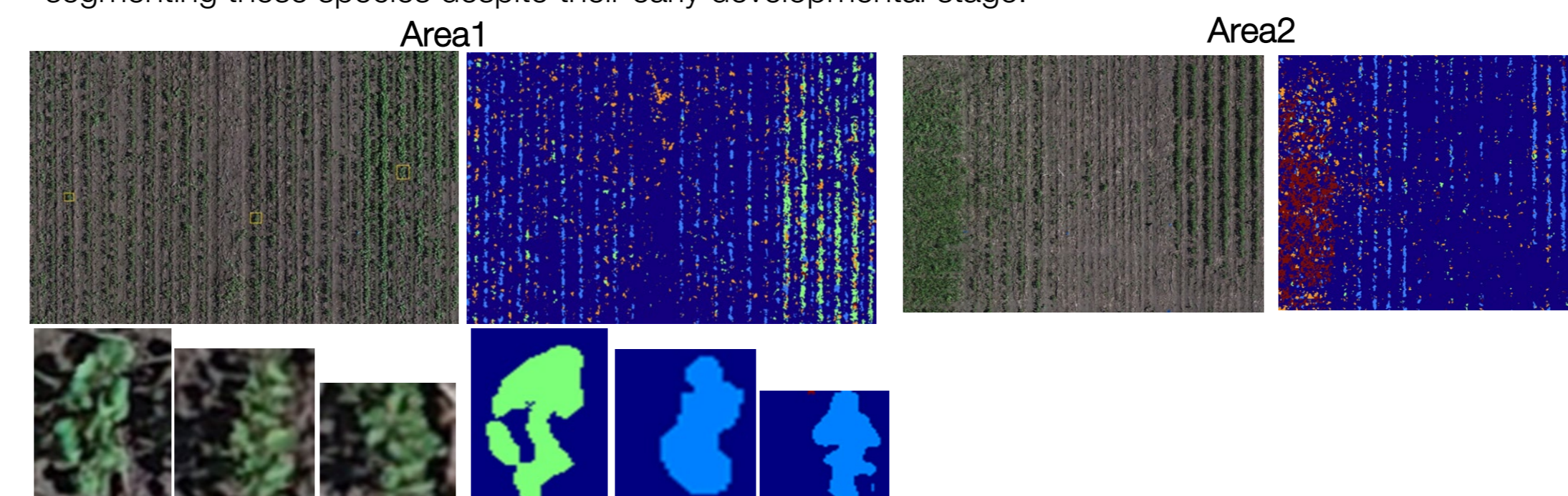
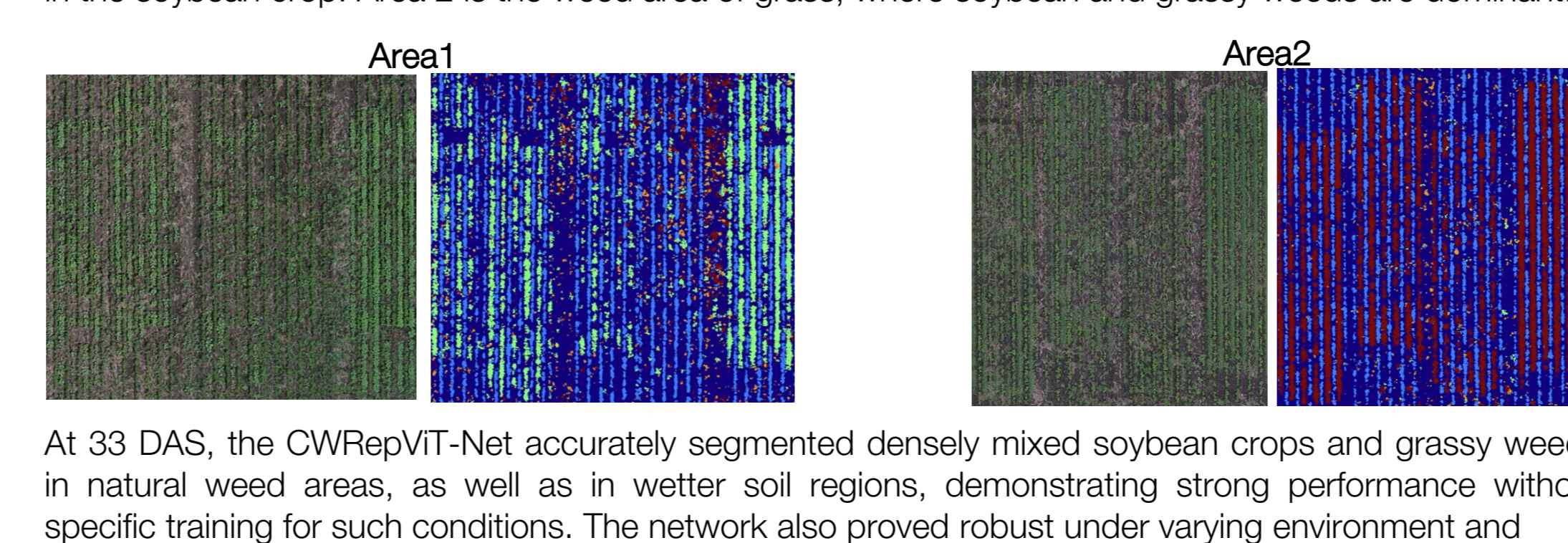


Figure 3. Results of broad leaf weeds (a-c) and volunteer canola (d-e) for all networks

At 21 DAS, when crops and weeds were still at the seedling stage, the CWRepViT-Net successfully segmented soybean, volunteer canola, and naturally occurring weeds in challenging conditions such as varying densities and shadow effects, as shown in zoomed-in views of Area1. In Area2, which contained a dense presence of grassy weeds, the network also demonstrated high accuracy in detecting and segmenting these species despite their early developmental stage.

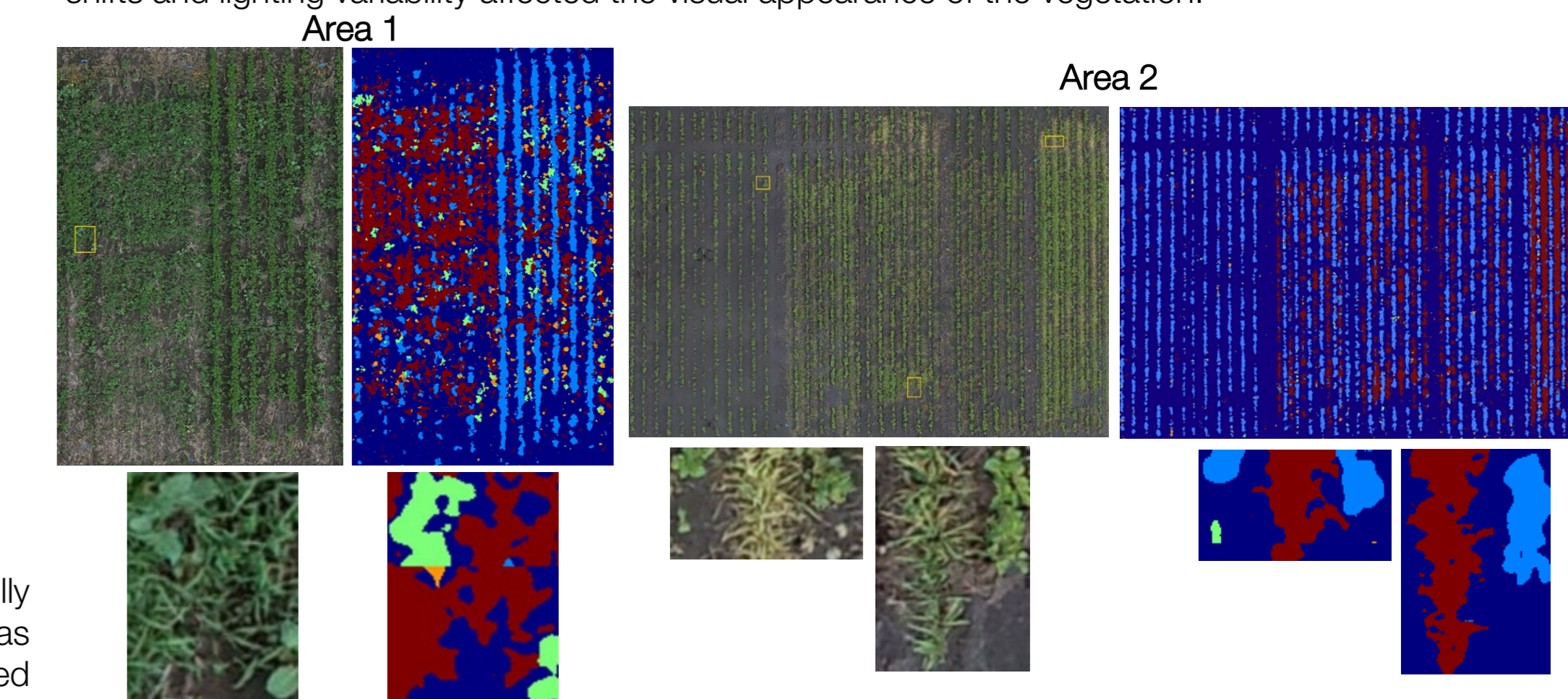


At 26 DAS, when both crops and weeds had grown larger, the CWRepViT-Net effectively segmented soybean, volunteer canola, and grassy weeds across different areas, accurately preserving leaf shapes and clearly distinguishing crop-weed boundaries even in complex and dense vegetation. Area1 belongs to the broadleaf weed portion of the experiment, where volunteer canola was the seeded broadleaf weed in the soybean crop. Area 2 is the weed area of grass, where soybean and grassy weeds are dominant.

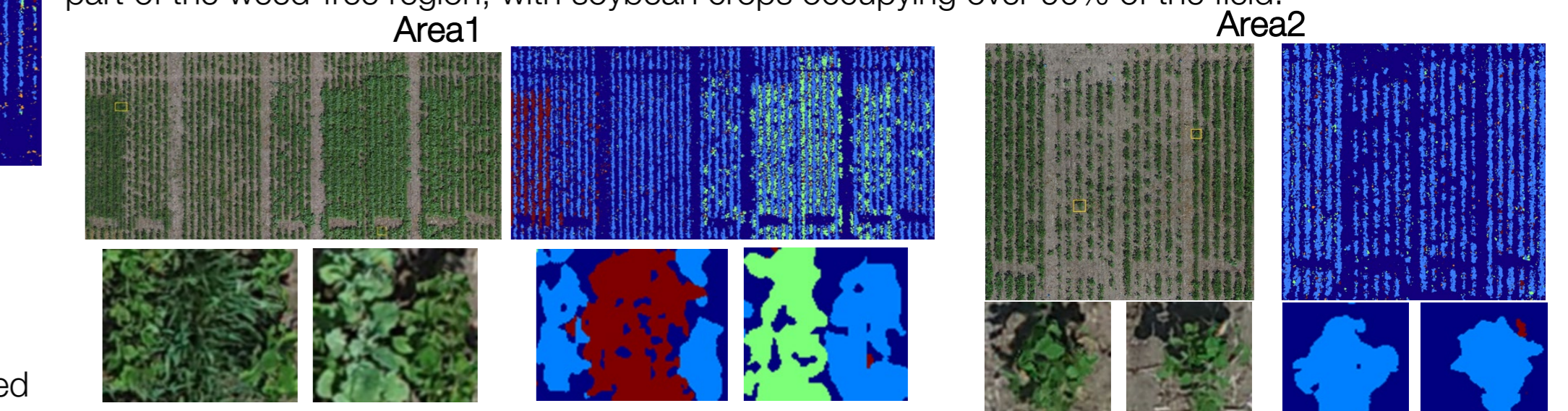


At 33 DAS, the CWRepViT-Net accurately segmented densely mixed soybean crops and grassy weeds in natural weed areas, as well as in wetter soil regions, demonstrating strong performance without specific training for such conditions. The network also proved robust under varying environment and

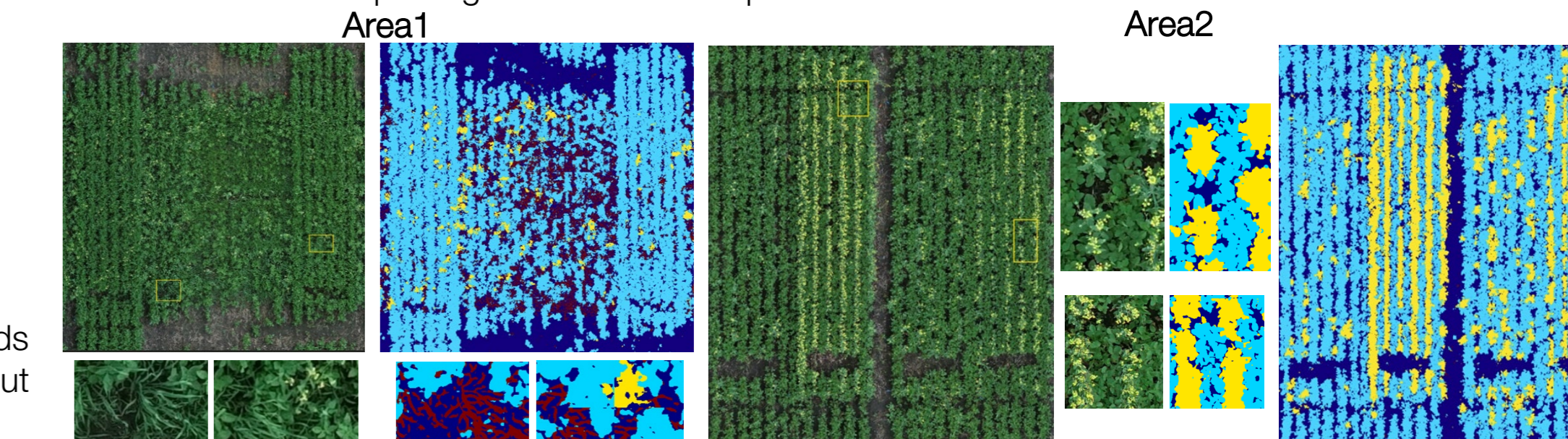
illumination conditions, successfully distinguishing soybean and grassy weeds even when color shifts and lighting variability affected the visual appearance of the vegetation.



At 39 DAS, despite increased crop and weed density, leaf overlap, and shadow or occlusion effects, the CWRepViT-Net accurately segmented soybean, volunteer canola, and weeds across both mixed and weed-free areas, effectively preserving class boundaries even under complex visual conditions. Area 1 is a part of the grassy weeds and broadleaf weeds patches. Area2 is a part of the weed-free region, with soybean crops occupying over 90% of the field.



At 45 DAS, when soybean reached peak vegetative growth and volunteer canola began to flower, the CWRepViT-Net effectively segmented high-density areas with overlapping soybean, grassy weeds, and sparsely scattered canola, accurately preserving class boundaries even under complex conditions. In both natural weed and broadleaf weed regions, including alternating row patterns and dense mixtures, the network successfully distinguished between soybean and volunteer canola despite significant leaf overlap.



References

- Wang et al. 2024. IEEE/CVF Conference. 15909-15920.
- Gomroki et al. 2023. RS MDPI. 15(5): 1232.
- Liu et al. 2018. IOP Conference. 428: 012043.
- Gomroki et al. 2025. IJRS. 1-24.

Funding

