Semantic Segmentation of Sugar Beet Fields with Pseudo-Attention Mechanisms

Miles Mena & Ryan Slocum

CSCI 5922: Neural Networks and Deep Learning
University of Colorado Boulder, CO, USA
April 29th, 2024

Abstract. Semantic segmentation of objects from RGB cameras enables autonomous systems to make decisions with more complete information. In agricultural robotics, semantic segmentation differentiates weeds, crops, and soil so that weeds can be autonomously removed, saving resources and labor. While existing semantic segmentation methods perform well enough to be implemented in commercial products, the methods themselves lack a singular approach towards robustness. Humans use a single model of vision, where as computer vision offers a wide array of preferment methods without a scientific understanding of the behaviors of each method. As such, we build upon the U-NET model to replicate attention with the aim of providing a scientific exploration of pseudo-attention in convolutional neural networks. We first add a convolutional block attention mechanism (CBAM) to U-NET. We also augment the data with a channel of YOLOv7 and a vegetation index. We present results on the PhenoBench Dataset, achieving a mean Intersection over Union (mIoU) score of 83.53 with U-NET CBAM.

1 Introduction

Agricultural robotics is a fast-growing application of deep learning in computer vision, with compelling outcomes. By being able to accurately detect the difference between weeds, crops, and other objects in an image, devices such as Verdant Robotic's Sharpshooter precisely spray or laser weeds, dramatically reducing herbicide usage [1]. In addition, the data gathered from these computer vision applications can assist farmers in better managing their crops and maximizing yields [2]. The semantic segmentation task is an important step in these computer vision pipelines, giving a pixel-by-pixel differentiation of weeds and crops in images taken from trailers, drones, and other robots used in agriculture.

Critical to this task is the proper and thorough segmentation of weeds. However, due to the relatively small size of most weeds in the sugar beet fields, these plants are easily missed by most current models. In the PhenoBench competition [3], in which this work enters a submission, the best models only achieve an Intersection over Union (IoU) score of less than 70% on weeds, which we believe could be significantly improved.

Vision Transformers (ViT) [4] successfully introduced attention to vision tasks using the attention module from [5]. In an attempt to replicate these attention modules, we propose two additions to the U-NET [6] model, with both aiming to increase the model's attentiveness towards weeds in the dataset. The first proposed addition highlights all plants in the image by feature engineering two new channels: vegetation indices from the existing RGB [7], and a YOLOv7 channel to provide bounding boxes for the locations of plants and weeds [8]. The second proposed addition integrates Convolutional Block Attention Modules (CBAM) within the layers of our U-NET model, to include more global information about the image. By manufacturing attentiveness with pseudo-attention mechanisms, we replicate vision attention first used in [4].

2 Related Work

2.1 Weed segmentation Current Methods

In [9] the authors develop a pipeline that combines ground and aerial images to segment crops and weeds. An encoder-decoder based neural network is trained on the ground imagery and augmented with cropped and resized aerial images. This outperforms models that are trained on only a single domain.

In [10], the authors improve upon the foundational U-NET model to segment seedling grass in the maize seedling stage. Instead of a sequence of double convolutions in the encoder stage, the authors down-sample a series of ResNeXt modules. Additionally, the decoder stage is composed of up-sampled deformable convolutions with a squeeze and excitation module attached after every convolution. These changes improve the models feature extraction in occluded targets and the sensitivity of the model to changes in seedling shape.

2.2 Attention Mechanisms in Weed Segmentation

In [11], the authors detect weeds with a modified Faster R-CNN that ingests RGB and Depth images on dual paths. The authors apply a squeeze and excitation (SE) module and a convolutional block attention module (CBAM). In their ablation study they showed that CBAM performed better than SE.

Knowing where to look is an important component of attention in vision tasks. In [12], YOLOv7 detects the location of weeds and then segments those patches with a series of CNN. [13] presents a lightweight attention mechanism based on CBAM for use in semantic segmentation that shows improvements in having the network identify objects of interest, especially weeds and crops.

2.3 U-NET with CBAM in Agricultural Segmentation

In [14], the authors present the use of an Efficient Channel Attention module in the semantic segmentation task for a pineapple field. The results show a better resiliency to changing field conditions, and a better global recognition of the scene. [15] utilizes the architecture we implement (shown in Fig 1), with CBAM modules after each convolutional layer, to improve the classification of winter wheat planting areas from satellite imagery. The results showed that the classification around the edges were improved using the attention module.

2.4 Novelty of the Proposed Approach

Our approach differentiates from these works by combining data augmentation with object detection models to replicate attention. Weed segmentation methods have explored the two methods separately, but not in conjunction. In particular, previous weed segmentation methods don't scientifically compare the two methods. Additionally, the dataset we train and test on has only recently become available, meaning that our proposed methods has the potential to generate top results in the associated competition.

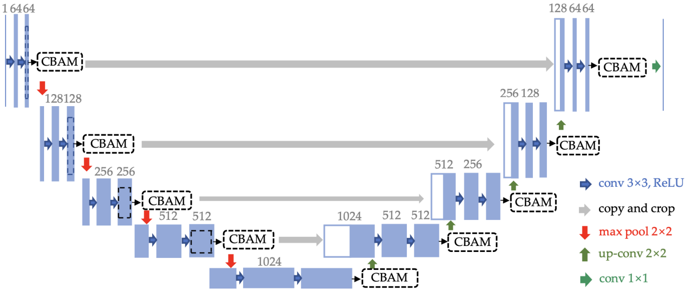

Fig. 1. U-NET/CBAM architecture presented in [15]

3 Methods

U-NET is an encoder-decoder neural network that was first introduced for biomedical image segmentation. As an encoder-decoder, U-NET learns how to compress an image's salient features, then decompress those features to predict classes for each pixel. Our additions to the U-NET architecture include the Convolutional Block Attention Module and data augmentation with YOLOv7 and the Color Index of Vegetation Extraction (CIVE).

We append CBAM to the architecture before a max pooling layer and before an up-convolution layer similar to [15], as shown in Fig. 1. CBAM consists of a channel attention and spatial attention mechanism. A feature map is passed into the channel attention mechanism to inform the model of the importance of the map's channels. Then the channel refined feature map is passed into the spatial attention mechanism to similarly extract salient areas in a single channel.

We leverage the contrast of color between plants, crops and weeds, and soil with the CIVE index. CIVE is a linear combination of RGB values developed in [16] and discussed in [17], that produces a grey scale image highlighting plants. An example of the our 5 channels are shown in Fig. 2

$$CIVE = 0.881G - 0.441R - 0.385B - 18.78745 \tag{1}$$

Alongside CIVE we augment the data with bounding box predictions of plants from YOLOv7. The Phenobench competition provides supplementary code with resources for the competition like dataloaders. Among these resources, they provide the YOLOv7 model that baselines plant and weed detection on their dataset for the plant detection task. They report the average precision (AP) and mean average precision (mAP) metrics found in Table 1 for their YOLOv7 object detection.

mAP	**mAP50**	**mAP75**	**Crops AP**	**Weeds AP**
60.48	82.47	62.30	83.06	37.91

Table 1. Phenobench YOLOv7 Performance

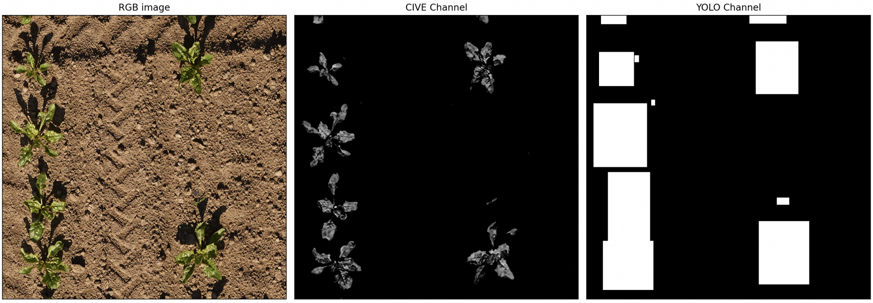

Fig. 2. The five channels provided to the U-NET CIVE+YOLOv7 model (R, G, B, CIVE, and YOLOv7)

These methods emulate the concept of attention, but they do not instantiate the natural language method of attention as introduced in [5]. We therefore call the methods we present ”pseudo-attention” to differentiate them from Vision-Transformer architectures like [4].

Our architectural parameters are listed Table 2. We will be running our training and experiments on a Windows 11 machine, with 12GB VRAM an AMD Ryzen 5 3600 6-core and a NVIDIA GeForce RTX 3060 GPU.

We report the individual Intersection-over-Union (IoU) for each class and the mean Intersection-over-Union (mIoU) across classes.

Parameter	**Value**
Epochs	100
Batch Size	2
Obejective Function	Weighted Cross Entropy: (1/88.45), (1/11.03), (1/.5)
Optimizer	Adam $\beta_1 = .9$, $\beta_2 = .999$
Framework	Pytorch 2.2.1
Learning Rate	.001
Seed	42

Table 2. Training Parameters of Experiment 1

4 Results

To evaluate the performance improvement of the two pseudo-attention mechanisms we've implemented, we use a control "vanilla" U-NET model with the same training parameters as the two test models. In Table 3 we report the mIOU for the test partition, as computed by the PhenoBench competition website.

Model	**mIOU Soil**	**mIOU Crops**	**mIOU Weeds**	**avg mIOU**
Vanilla U-NET	98.99	**93.01**	48.96	80.32
U-NET CIVE+YOLOv7	99.01	92.81	51.18	81.0
U-NET CBAM	**99.11**	92.99	**58.49**	**83.53**

Table 3. PhenoBench Semantic Segmentation per-model mIOU results

It's clear that both of the implemented pseudo-attention mechanisms made marginal gains over the control model. In the soil and crop categories, all three of the models performed at roughly the same level, with no noticeable differences in segmentation efficacy. This is unsurprising, as the mIOU scores for soil and crops is already fairly high in the control group.

However, there was a substantial increase in performance for the mIOU score for weeds. The vanilla U-NET scored just under 49 percent in this category, whereas the U-NET CIVE+YOLOv7 model scored just over 51 percent, and the U-NET CBAM model scored over 58 percent, representing an improvement of nearly 10 percentage points.

These pseudo-attention mechanisms improve upon the baseline by directing the model's focus on the relatively small weeds. However, the added computational cost of the CBAM modules may not be worth the slight gain in weed labeling accuracy. When labeling the test data using a 2021 Macbook Pro with a 16-core M1 Pro, the U-NET CBAM model took an average of 1.24 seconds per image, whereas the vanilla U-NET model took an average of 0.66 seconds per image. That is nearly double the time, for relatively minimal gain in labeling performance for weeds.

Visualizing each model's prediction on the same image, in Fig. 3, demonstrates the different behavior in each models predictions. We find that U-NET

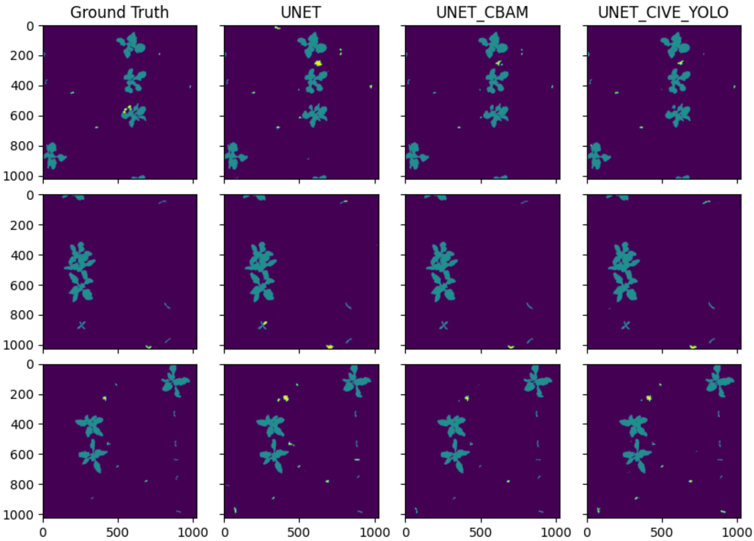

Fig. 3. Comparing the predictions of the each model on three examples from the validation set.

is imprecise in its prediction of weeds, and it is quick to over-predict an an area in the image as a weed. U-NET-CBAM outlines the edges of weeds better than U-NET, but is still susceptible to predicting a weed when there are no weeds present. Finally, U-NET-CIVE-YOLOv7 handles edges well, but can often confuse a small part of a crop with a weed.

4.1 Future research

This work leaves several research questions open for further investigation. Training for 100 epochs took over 72 hours, so we did not have the time or resources to conduct an extensive hyperparameter study. A through investigation of hyperparameters could greatly improve any model's performance in segmenting edges and small objects, particularly exploring the weights of the cross-entropy loss.

A second direction for this research is a complete ablation study of the data augmentation we've explored. There are several other vegetation indices with the possibility to improve the models. Also, PhenoBench's object detection model predicts weeds and plants, so differentiating between the two in our additional YOLOv7 channel could provide more relevant information to our base model. The provided YOLOv7 model does not detect objects perfectly, so improving the model or providing a prediction confidence could provide resilience against

YOLOv7's inaccuracies, while still integrating these initial guesses at weed and plant bounding boxes into the semantic segmentation task.

5 Conclusion

We have explored two pseudo-attention mechanisms for the semantic segmentation task for images of sugar beet fields. One mechanism involves the use of CBAM modules in a traditional U-NET architecture. The other mechanism involves the addition of feature-engineered YOLOv7 and CIVE channels to the RGB image. When testing with the Phenobench sugar beet dataset, we identified marginal improvement on weed detection when compared with our control U-NET model, though not enough to score highly in the associated competition. These methods show promise, but are likely not worth the added computational expense required for their implementation.

As with many works in deep learning and computer vision, it is important to consider ethical and societal impacts. Accurate weed detection can greatly reduce the amount of toxic herbicides released into an environment and produce more food with less inputs. However, problems such as contentions over the ownership of a farm's digital data, privacy concerns over autonomous UAV's, and labor saving technologies in high unemployment economies, are among a few that the scientific community has to discuss with farmers to ensure these technologies beget more benefits than burdens. [18]

References

1. Gerhards, R., Andujar Sanchez, D., Hamouz, P., Peteinatos, G.G., Christensen, S., Fernandez-Quintanilla, C.: Advances in site-specific weed management in agriculture—a review. Weed Research **62**(2) (2022) 123–133
2. Schunck, D., Magistri, F., Rosu, R., Corneließen, A., Chebrolu, N., Paulus, S., Léon, J., Behnke, S., Stachniss, C., Kuhlmann, H., Klingbeil, L.: Pheno4D: A spatio-temporal dataset of maize and tomato plant point clouds for phenotyping and advanced plant analysis . **16**(8) (2021)
3. Weyler, J., Magistri, F., Marks, E., Chong, Y.L., Sodano, M., Roggiolani, G., Chebrolu, N., Stachniss, C., Behley, J.: Phenobench–a large dataset and benchmarks for semantic image interpretation in the agricultural domain. arXiv preprint arXiv:2306.04557 (2023)
4. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
5. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
6. Ronneberger, O., Fischer, P., Brox, T.: 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, Springer (2015) 234–241
7. Milioto, A., Lottes, P., Stachniss, C.: Real-time semantic segmentation of crop and weed for precision agriculture robots leveraging background knowledge in cnns. In: 2018 IEEE international conference on robotics and automation (ICRA), IEEE (2018) 2229–2235
8. Redmon, J., Divvala, S.K., Girshick, R.B., Farhadi, A.: You only look once: Unified, real-time object detection. CoRR **abs/1506.02640** (2015)
9. Gao, J., Liao, W., Nuyttens, D., Lootens, P., Xue, W., Alexandersson, E., Pieters, J.: Cross-domain transfer learning for weed segmentation and mapping in precision farming using ground and uav images. Expert Systems with applications **246** (2024) 122980
10. Cui, J., Tan, F.: Improving u-net network for semantic segmentation of corns and weeds during corn seedling stage in field. Frontiers in Plant Science **15** (2024) 1344958
11. Xu, K., Yuen, P., Xie, Q., Zhu, Y., Cao, W., Ni, J.: Weedsnet: a dual attention network with rgb-d image for weed detection in natural wheat field. Precision Agriculture **25**(1) (2024) 460–485
12. Rai, N., Sun, X.: Weedvision: A single-stage deep learning architecture to perform weed detection and segmentation using drone-acquired images. Computers and Electronics in Agriculture **219** (2024) 108792
13. Bai, X., Xue, Y., Dai, H., Wang, L., Bai, X., Hu, X., Li, B.: Channel coordination attention for crop and weed segmentation neural networks. (2023)
14. Cai, Y., Zeng, F., Xiao, J., Ai, W., Kang, G., Lin, Y., Cai, Z., Shi, H., Zhong, S., Yue, X.: Attention-aided semantic segmentation network for weed identification in pineapple field. Computers and Electronics in Agriculture **210** (2023) 107881
15. Zhao, J., Wang, J., Qian, H., Zhan, Y., Lei, Y.: Extraction of winter-wheat planting areas using a combination of u-net and cbam. Agronomy **12**(12) (2022) 2965

16. TOSAKA, N., HATA, S.i., OKAMOTO, H., TAKAI, M.: Automatic thinning mechanism of sugar beets (part 2) recognition of sugar beets by image color information. Journal of the Japanese Society of Agricultural Machinery **60**(2) (1998) 75–82
17. Kataoka, T., Kaneko, T., Okamoto, H., Hata, S.: Crop growth estimation system using machine vision. In: Proceedings 2003 IEEE/ASME international conference on advanced intelligent mechatronics (AIM 2003). Volume 2., IEEE (2003) b1079–b1083
18. Rose, D.C., Wheeler, R., Winter, M., Lobley, M., Chivers, C.A.: Agriculture 4.0: Making it work for people, production, and the planet. Land use policy **100** (2021) 104933