Multimedia Tools and Applications
https://doi.org/10.1007/s11042-021-11770-7

Semantic segmentation for plant phenotyping using advanced deep learning pipelines

Pullalarevu Karthik[1] · Mansi Parashar[2] · S. Sofana Reka[1] · Kumar T. Rajamani[3] · Mattias P. Heinrich[3]

Received: 26 July 2021 / Revised: 23 September 2021 / Accepted: 25 November 2021

Abstract
Large strides have been made in the field of semantic segmentation which finds its application in extensive areas of research. However, these advancements have not been completely utilized in the field of plant phenotyping. Deriving quantitative plant phenotypes in a non-destructive manner from plant images is a key challenge that strongly relies on the precise segmentation of plant images. In this paper, we propose novel semantic segmentation pipelines for the task to improve the automated phenotyping process. In this work architectures such as U-Net, Attention-Net and Attention-Augmented Net are introduced that are trained on the Arabidopsis Thaliana plant dataset released under the CVPPP14 competition. Dice coefficient is used as the evaluation metric to compare performances of the proposed architectures, and also benchmark them against existing algorithms in literature. Results of semantic segmentation of Rosette plants shows the state-of-the-art results, with attention net achieving a 0.985 dice score that easily outperforms all the other deep learning and image processing techniques proposed earlier for plant segmentation in this domain. Results are exhibited with comparison analysis successfully with these advanced deep learning architectures and can be used as a base for plant phenotyping related applications.

Keywords Phenotyping · U-Net · Attention-Net · Attention augmented net · Semantic segmentation

1 Introduction

Keeping in mind the growing number and affluence of the population, there is a steep rise in the demand for agricultural produce [1–3] making it vital to take advanced steps for securing global food security. Related disquisition [10] has found that High Throughput Plant

✉ S. Sofana Reka
chocos.sofana@gmail.com

[1] School of Electronics Engineering, Vellore Institute of Technology, Chennai, India

[2] School of Computer Science and Engineering, Vellore Institute of Technology, Chennai, India

[3] Institute of Medical Informatics, University of Lübeck, Lübeck, Germany

Published online: 14 December 2021

Phenotyping holds the potential to alleviate the situation. Traditional methods of plant phenotyping consist of using destructive harvesting of plants for manual measurements and extraction of features making it difficult to extract temporal information from the crop. In this scenario, only one set of measurements can be obtained from each specimen leading to a bottleneck in phenotyping studies [8–13]. Computer Vision techniques for Image-based plant phenotyping [15, 26] were proposed to effectively break the phenotyping bottleneck. The images which are taken at regular periods along the duration of the experiment can be used to extract information without having to harvest the crop. Size, number of leaves, shape descriptors like area, perimeter, circularity, aspect ratio, and roundness can be derived from image analysis which can be used to predict plant mass [5, 6, 24] and hence the growth rate can be analyzed using computer vision. *Arabidopsis Thaliana* is extensively linked with plant phenotyping given its immense importance in mapping phenotypic traits to genotype in literature [24]. It is the first plant to have its genome sequenced and acts as a basis for studying plant traits for genetic, molecular, and cellular biology.

Arabidopsis Thaliana is a rosette plant [6] growing close to the ground allowing it to be treated as 2D objects and using RGB images to non-invasively extract essential phenotypic features [5]. Deriving the functional measurements requires semantic segmentation of the rosettes as the crucial base step for analysis [32]. Thereby, powerful image segmentation algorithms [16–22] are required to separate the background from the foreground in plant images, taking into consideration the complications arising due to overlapping leaves and variable scales at different stages of the life cycle. Given the importance of Arabidopsis Thaliana in plant phenotyping and taking into consideration the need for scientific innovation in computer-aided phenotyping and the complications which come with the task, it becomes important to come forward with novel, state of the art methods. In our work, we aim to develop decisive pipelines for auto segmentation of rosette plants to facilitate computer vision-based plant phenotyping. To review the related literature in this domain, *Pape* et al. proposed a Histogram based segmentation [25], using color and size-based histogram thresholding for segmentation. However, in presence of similar color noise, small and overlapping leaves, the image processing algorithm fails. To overcome the shortcomings of image processing techniques, Deep learning-based methodologies were proposed for the task. *Shubra Aich* et al. adopt the SegNet architecture [2] using an autoencoder framework without skip connections making the model vulnerable to vanishing gradients.

Similarly, *Sakurai* et al. [30] used transfer learning using ImageNet for the segmentation task [27–35] which makes use of domain adaptation to train weights that were trained on an ImageNet [29] dataset. Due to the difference in the network architecture between source and target domain, skip connections are not transferred which again makes them vulnerable to vanishing gradients as discussed in the previous work. With respect to this feature of extensive trainable parameters of CNN models [23] , *Atanbori* et al. [4] proposed a 'tiny' and 'very-tiny' net architecture. They reduced the number of model parameters by using separable convolutions and then combined it with Singular Value Decomposition (SVD) for diminishing the size of the weight matrices. A reduction in the number of parameters however led to a decrease in performance as compared to the uncompressed models.

1.1 Research gaps and motivation

In order to use plant images for phenotyping, it becomes imperative that the plant sample is segmented accurately even when the shape and color of leaves, the geometric and density eccentricity of the leaves differ. Therefore, image processing algorithms that contain parameters that are hand-tuned to specific data are at risk of being overly specified. While the previously proposed deep neural networks learn a representation of the data [37–41] without image parameters specified by hand, they work well when enough data is available for training. However, datasets of plant images for segmentation are not yet available on a large scale due to the considerable expense involved in collecting and annotating this type of data. Additionally, the proposed convolution methods take only local context into account, missing relative global context which can be leveraged to enhance the robustness of the segmentation.

To address these research gaps and to expand the relatively smaller amount of research done in the plant phenotyping domain for segmentation, in this work the authors contribute the following novel state-of-the-art architectures.

- The authors have introduced and adapted a U-Net inspired architecture, used generally for the task of medical segmentation. Further research shows that it has not yet been applied to the field of concern yet. We chose this network as it gives exemplary results with fewer training images [28] several, overcoming the drawback of lack of training images for plant phenotyping.
- The authors propose introducing attention to the architecture to allow capturing of pixel-based long-range interactions. This would allow global context to be taken into consideration along with local context. It will enable the model to focus on regions of interest hence enhancing the discriminative power for semantic segmentation.
- A hybrid of the self-attention and convolutional model is proposed for the plant phenotyping domain. This novel architecture presents a combination of attention and convolutional blocks to improve [7] semantic segmentation results.

Experimental results on our Arabidopsis Thaliana 2D RGB images dataset (http://www.plant-phenotyping.org/CVPPP2014-dataset) confirm that our model yields a much greater result contributing to improving the Background-Foreground segmentation of Rosettes and providing a boilerplate for automated and precise plant phenotyping.

The entire paper is done in three sections, Sect. 2 exhibits the proposed methodology involving the explanation of the proposed architecture for the task involved. Section 3 explains the experiments involved in achieving this task. Section 4 and Sect. 5 portrays the results and discussions with the conclusion.

2 Proposed methodology

2.1 Semantic segmentation

In this work, the main focus of the Computer Vision task is Semantic Image Segmentation. The task can be defined as the pixel-wise classification of the image into a corresponding class label, which in this case would be foreground and background pixels. For a given RGB input image (h x w×3) a segmentation map with each pixel representing the class label has to be outputted as (h x w×1). Here, the global context of "what" the pixel belongs to and the local context of "where" the pixel would be positioned [34] is information to be precisely extracted from the image for successful segmentation. For our proposed deep learning approach, we extracted the lower level features which are vital in distinguishing the classes a pixel belongs to by effectively down sampling the spatial resolution of the input. These feature representations are then up-sampled to generate a segmentation map of the same resolution as the input image, hence giving us pixel-wise classification at the same scale and pixel-position as the original image.

“Unpooling” is more or less the reverse operation of pooling. In this process, the spatial dimension is increased by repeating a pixel value in a neighborhood. For up sampling, instead of using the concept of unpooling, we use learned up sampling for a more precise segmentation map with regards to the original pixel positions. This process uses transposed convolutions where an inverse convolution operation of multiplying the weights of the filter by a single value from the low-resolution feature map is carried out and then these form the output feature map. This process exhibits a better resolution for this analysis. At the final analysis, the output image is obtained, which is of the same dimension as the input image with pixels classified into their corresponding labels.

2.2 Proposed architectures

For this analysis of the proposed approach involved, the following deep learning architectures were considered and further study was analyzed in these architectures and the results are also compared extensively with the methods used in the literature**.**

2.2.1 U-Net

Ronneberger et al. modify the FCN as proposed by *long* et al. [34] by introducing a modification where up sampling of the expansion path will also have a large number of feature channels to propagate context information which causes the expansive path to be symmetric of the contractive path. It consists of a contracting subsection followed immediately by an expansive subsection, hence giving it its unique shape as exhibited in Fig. 1, for which it is named. Features set from the contracting path are concatenated with the up sampling output. The contracting path encapsulates context and the expansion path captures the localization of the pixels.

For the plant phenotyping task, we propose a model comprising of 19 convolutional layers divided into a contracting path that follows a standard convolutional network. In this model, there are 3 × 3 convolutions that are repeated with each being followed by a RELU activation [17] process and a 2 × 2 max-pooling of stride 2. The down sampling path doubles the feature set. An expansive path consists of up sampling of the feature map accompanied by 2 × 2 convolution which results in halving the number of feature channels. In this model, as an important analysis, there is a concatenation with the feature map of the contracting path which corresponds to that particular convolution which is followed by two 3 × 3 convolutions each of which is followed by a RELU activation and dropout with varying dropout rates. The feature map of the contracting path is cropped before mapping to mitigate the loss of border pixels in every convolution. The final layer is a 1 × 1 convolution which maps each of the 64 component feature vectors into 2 classes, either plant pixel or background pixel.

2.2.2 Attention net

The attention maps are computed from queries and keys which primarily are weighted average operations. If F^{s}_{avg} and F^{s}_{max} are 2D tensors obtained after average pooling and max pooling respectively, then the output of the attention module would be a spatial attention

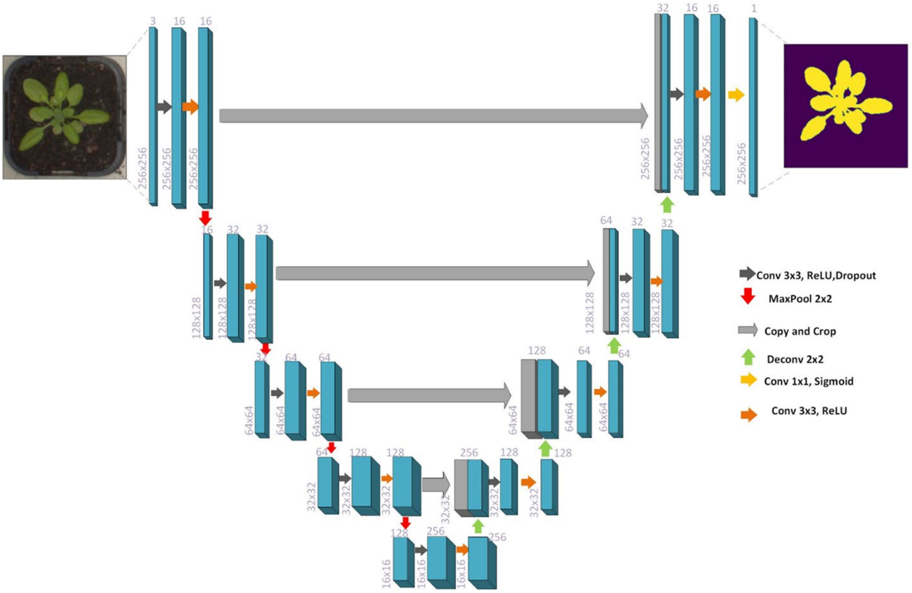

Fig. 1 U-Net architecture. Each blue box represents a multi-channel feature map and the number of channels are represented on top of them. Numbers on the left edge of each blue box represent the x–y size. Gray boxes signify the copied feature maps. Arrows represent various operations like convolution, MaxPooling

map $M_s(F)$ with encoded information which regions to accentuate and subdue. This map is computed as shown in the below equation

$$\mathbf{Ms(F)} = \boldsymbol{\sigma}\left(\mathbf{f}^{7\times7}\left[\mathbf{F^s}_{\mathbf{avg}};\mathbf{F^s}_{\mathbf{max}}\right]\right)$$

where,

σ is the sigmoid activation function
f^{x7} is the convolution operation of filter size 7×7

These enable dependency on the input signals and hence compute long-range interaction. The weights in the operation are produced dynamically with no dependence on their relative position.

In our model, the down sampling path consists of 3 layers where each layer has 2 convolutions of kernel size 3×3, with a stride of 1 and no padding followed by max-pooling. In the next layer, we introduce attention. Queries, keys, and values are computed by usual convolutions by passing a bias parameter and setting the kernel side to 1×1. The input channels are set equal to the output channels, in this case, 256. Figure 2 exhibits the Attention Net architecture for the model proposed.

2.2.3 Attention augmented net

In our Attention Augmented Net, convolutional feature maps and self-attention feature maps are combined; the convolutional feature maps are concatenated with feature maps being produced by self-attention. In other terms,

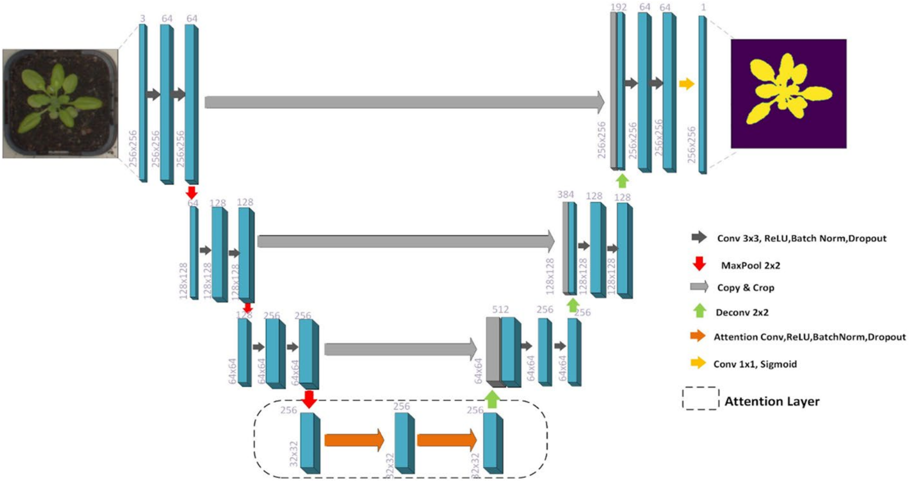

Fig. 2 Attention net architecture. Each blue box represents a multi-channel feature map and the number of channels are represented on top of them. Numbers on the left edge of each blue box represent the x–y size. Gray boxes are the copied feature maps. The orange arrow represents Attention mechanism followed by ReLU activation, batch normalization and dropout

$$\mathbf{AAConv(X) = Concat[Conv(X), MHA(X)]}$$

where MHA(X) stands for multi-headed attention feature maps.

Figure 3 shows the architecture of Attention augmented Net. The Attention Augmented net follows the same pattern as that of the Attention Net with few modifications [7]. Rather

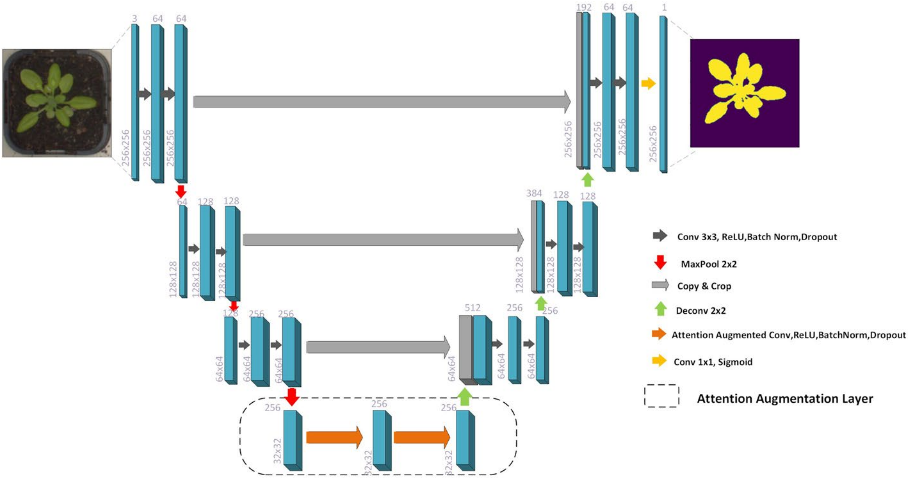

Fig. 3 Attention augmented net architecture. Each blue box represents a multi-channel feature map and a number of channels is represented on top of them. Numbers on the left edge of each blue box represent the x–y size. Gray boxes are the copied feature maps. The orange arrow represents augmented convolution followed by ReLU activation, batch normalization and dropout

than passing the output of the convolution layer into the attention layer to redefine the existing feature map, we simply concatenate the feature map produced by convolutions with the feature map produced by the attention layer to obtain the final output feature map. All the convolutions were of size 3×3. For the Max Pooling operations, kernel_size of 2×2 with a stride of 2 was used. In this analysis, sigmoid activation was used in the output layer to extract the output.

3 Datasets and experiments

In this section, we elucidate the dataset, training, and implementation of our models and analyses with a detailed perspective for the approach designed.

3.1 Dataset

In the work proposed, the dataset for the process is the CVPPP14 leaf segmentation challenge dataset (http://www.plant-phenotyping.org/CVPPP2014-dataset) which consists of 752 images. The images are RGB images of a top view of Rosette plants grown in various growth chambers. Some images are out of focus, and some have the presence of moss or water in the background which increases the complexity of the semantic segmentation task. For further processing of work, we use randomly sampled 601 images for training and 151 images for validation.

3.2 Training and implementation

The input and output image size was set as 256×256, where the input images were RGB and the output segmentation image was set to grayscale. In this analysis, the experimentation was done with the training batch size and validation batch size by varying them across architectures and found that the best results can be gleaned with a training batch size of 16 and a validation batch size of 10. At the first step, training was done on the U-Net model from the beginning without utilizing any pre-trained architecture's weights. For the segmentation task, a deeper network did not show any considerable improvement in results as in processing.

Thereby for this process, a 5-layer U-Net has been opted where the network was trained for 3,4,5 down-sampling and up-sampling layers. A dropout rate of 0.2 was set in order to prevent overfitting with a ReLU activation for convolutional layers. For the ideal result, the _normal weight initialization was used. The model was trained for 40 epochs with Adam Optimizer [36] and binary cross-entropy as the loss function.

For attention net, the effect of data augmentation on the results was analyzed at the outset. In this process rotated images in the dataset were randomly used with an implementation of random horizontal flipping and center cropping. On observation, the results did not show any significant improvement from the results without data augmentation. The convolutional layers are customized by using a dropout rate of 0.2 and a momentum rate of 0.1. To achieve the best performance for the model the services of PyTorch's LR finder were utilized. It has been found that this method is the most efficient way of finding the ideal learning rate [14] rather than manual hyperparameter tuning. In order to eliminate overfitting in this process batch normalization was performed. Based on the results obtained, learning rates of 0.002 were involved to train the images for 32 epochs with soft dice loss as the optimizer. Similarly, for the Attention Augmented net, a dropout rate of 0.4 and a momentum rate of 0.1 to the convolution layers were applied. The optimum learning rates were found to be 1e-3 and were utilized. For the process batch normalization was applied and was trained on the net for 51 epochs to get the best performance.

Training for U-Net was performed on a single Tesla-K80 machine using Tensorflow [19] as the deep learning framework. Additionally, both Attention-net and attention-augmented net were implemented using PyTorch on Azure Virtual Machine with 4-Tesla K80 GPUs.

Sigmoid activation was used for the output layer to get binary classification into the foreground or background pixel for all the 3 architectures to obtain the best performances.

4 Results and discussion

In this section, the results of the proposed advanced deep learning architectures are exhibited and studied, with comparative analysis with previous work published on the dataset.

Figure 4 portrays the sample images analysis with segmentation results. For this analysis proposed, the Dice coefficient is used as a statistical measure of the overlap between the two images, in this case, the predicted mask and the ground truth label. The Dice Score lies between 0 to 1, where a dice coefficient of 1 implies a perfect and complete overlap which is exhibited in the below function. Dice Coefficient can be expressed as -

$$\mathbf{Dice} = 2*|\mathbf{X}| \cap |\mathbf{Y}|/(|\mathbf{X}| + |\mathbf{Y}|)$$

where,

X Ground truth label
Y Predicted mask

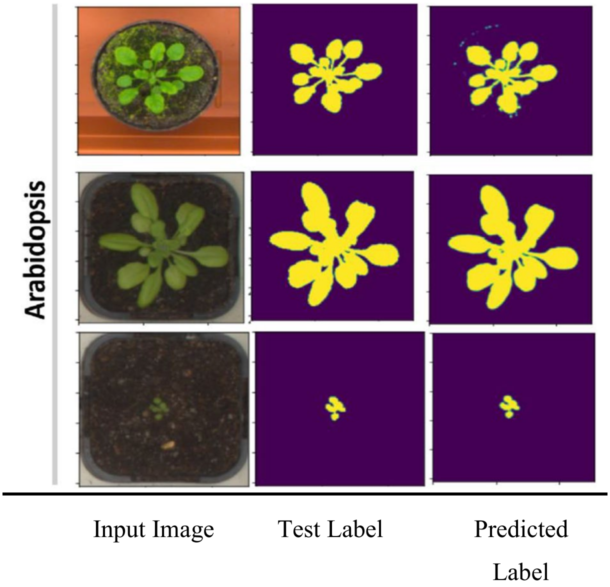

Fig. 4 Sample images with corresponding ground truths and segmentation results of U-Net architecture. Starting from the left, the first column is the input image, followed by the ground truth label and the last column is the output of attention net

Table 1 summarizes the performance of our model over the validation set.

From Table 1 we can observe that U-net with a score of 97.29 is outperformed by Attention-net which gives a score of 98.47. This is also slightly better than the performance of the attention-augmented net, at 98.35. Therefore, we can conclude that for the semantic segmentation in the plant images' domain, the attention net is able to outperform the hybrid attention-convolution net.

The training and validation Dice Score and Loss are recorded and depicted in Figs. 5, 6, 7 and 8 to provide an in-depth analysis of our proposed architectures' performance on the dataset.

Table 2 compares the performance of different architectures on the plant dataset. Dice Score is used as the statistical measure for comparison. The dice score, for *aiche* et

Table 1 Performance of the models proposed on the validation set. Dice coefficient, described in Sect. 4, which is also known as F-Score, is used as the performance metric of validation of the results

Architecture	Dice score (%)
U-Net	97.29
Attention net	**98.47**
Attention augmented net	98.35

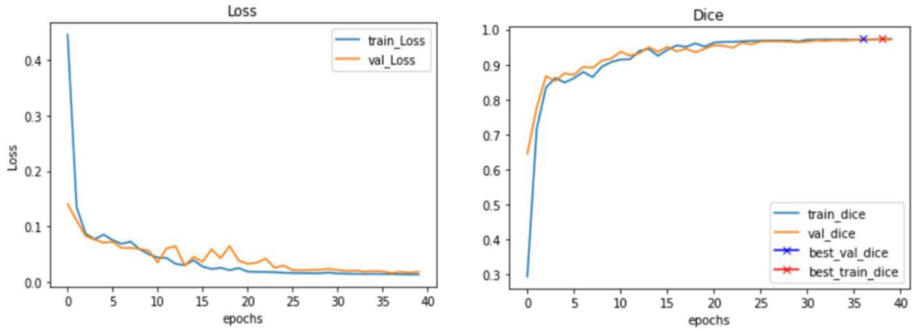

Fig. 5 Dice score and loss for U-Net

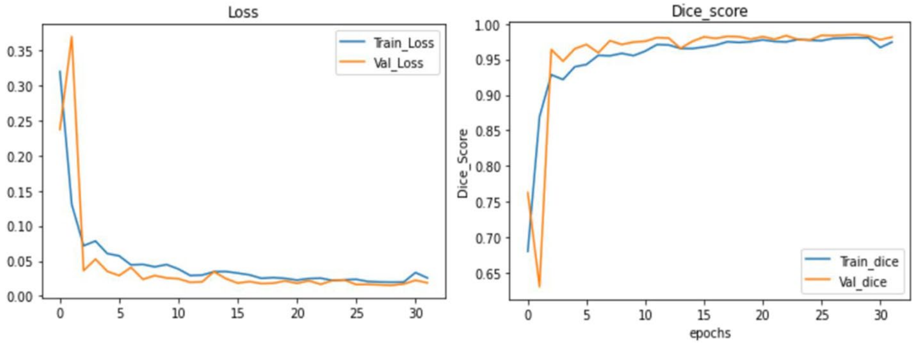

Fig. 6 Dice score and loss curves for attention net

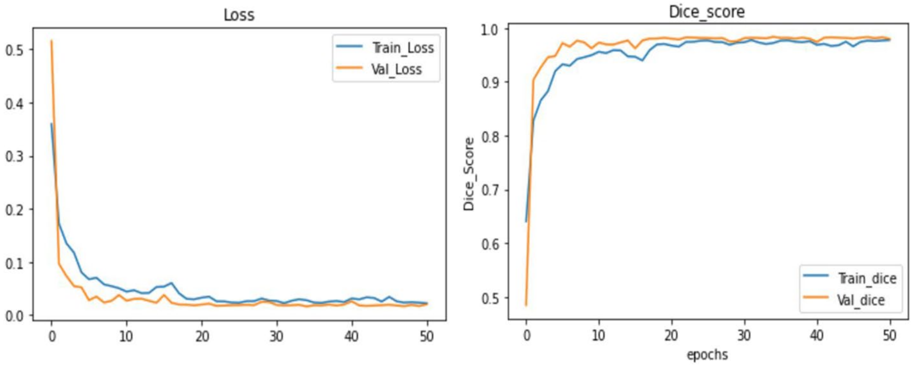

Fig. 7 Dice score and loss for attention augmented net

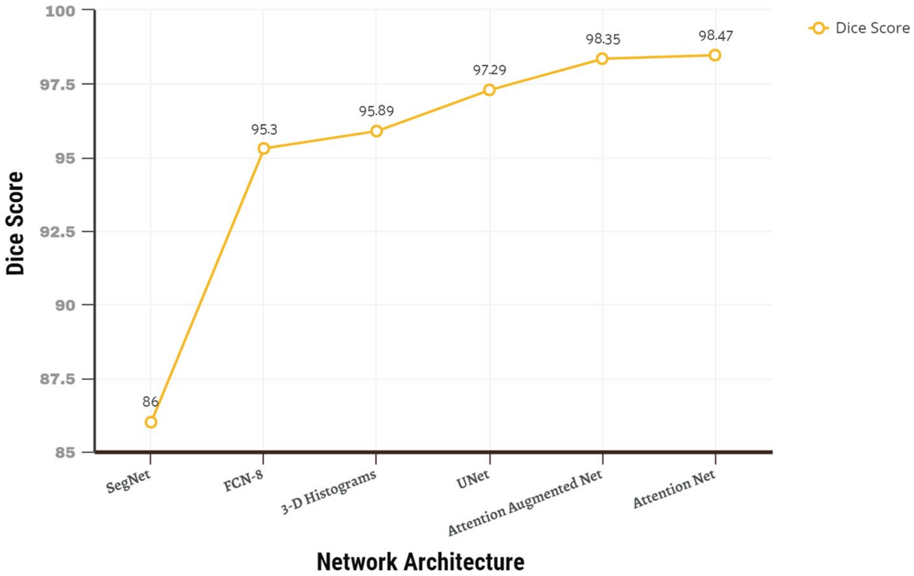

Fig. 8 Chart depicting network architecture comparison. Attention net shows the best performance among other architectures

Table 2 Architecture wise performance comparison. Highlighted scores indicate work done by the authors

Performance	Dice score (in %)
U-Net	97.29
Attention net	98.47
Attention augmented net	98.35
SegNet based [2]	86
FCN-8 based [30]	95.89
3D Histogram based [25]	95.3

al.'s "Deep Convolution and DeConv" method is calculated from their given Precision and Recall values.

$$\mathbf{F1Score} = 2((*\mathbf{Precision}*\mathbf{recall})/(\mathbf{precision} + \mathbf{recall}))$$

U-Net based pipeline, which gives the least score amongst our newly introduced pipelines for plant phenotyping, still gives a 1.46% increase in performance than the highest performing existing methodology as shown in Table 2.

Our best network ie. Attention Network gives a dice score of 98.47 which outperforms *Pape* et al. [25]'s method of Histogram-based segmentation by 3.32%, *sakurai* et al. [30]'s "Two-Step Transfer Learning" approach by 2.69% and SegNet based [2] results by 14.5%.

Hence, in this work, we introduced and built the 3 architectures from scratch and trained them on the plant phenotyping dataset released as a part of the CVPPP14 challenge. Upon testing the model, we compare the results obtained. We get state-of-the-art results, with U-net being able to perform segmentation with a loss of 2.71. Attention net and Attention Augmented net give further improved results as they procure a loss of 1.53 and 1.65 respectively. This clearly shows that the attention mechanism helps to improve the model performance by allowing it to focus on important features of the leaf class, taking into consideration the long-range interactions of the image pixels.

5 Conclusion and future work

In this work, we proposed the need of developing and using powerful deep learning methodologies to enhance the semantic segmentation of plant images which directly aids in high throughput phenotyping by allowing more precise features from plant images to be extracted non-destructively. We developed novel semantic segmentation pipelines for the task of foreground–background separation with the use of three deep learning architectures. According to our research, these architectures had never been utilized for the plant segmentation task before. Our evaluations indicate that we were successful in building deep learning-powered pipelines for the niche of semantic segmentation for plant phenotyping that is robust and precise in separating plant images from their background, even when the images are out of focus, have water or moss in the background. Our model outperforms the existing techniques and shows an unmatched accuracy. Furthermore, with a strong foundation for plant segmentation, the results can be utilized to further study and extract the various quantitative plant phenotypic features from the segmented images. Future work in this field presents the vast possibilities of completely automating the plant phenotyping pipeline where essential features like leaf count, leaf shape, density can be extracted from images and also predictions about the growth rate and yield can be made without physically harvesting the plant.

References

1. Aich S, van der Kamp W, Stavness I (2018) Semantic binary segmentation using convolutional networks without decoders. arXiv preprint arXiv:1805.00138
2. Aich S, Stavness I (2017) Leaf counting with deep convolutional and deconvolutional networks. In: Proceedings of the IEEE international conference on computer vision workshops. pp 2080–2089
3. Alexandratos N, Bruinsma J (2012) World agriculture towards 2030/2050: the 2012 revision

4. Atanbori J, French AP, Pridmore TP (2020) Towards infield, live plant phenotyping using a reduced-parameter CNN. Mach Vis Appl 31(1):2
5. Augustin M, Haxhimusa Y, Busch W, Kropatsch WG (2016) A framework for the extraction of quantitative traits from 2D images of mature Arabidopsis thaliana. Mach Vis Appl 27(5):647–661
6. Bell J, Dee HM (2016) Watching plants grow–a position paper on computer vision and Arabidopsis thaliana. IET Comput Vision 11(2):113–121
7. Bello I, Zoph B, Vaswani A, Shlens J, Le QV (2019) Attention augmented convolutional networks. In: Proceedings of the IEEE international conference on computer vision. pp 3286–3295
8. Chan W, Jaitly N, Le Q, Vinyals O (2016) Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp 4960–4964
9. Chorowski JK, Bahdanau D, Serdyuk D, Cho K, Bengio Y (2015) Attention-based models for speech recognition. In: Advances in neural information processing systems. pp 577–585
10. Danzi D, Briglia N, Petrozza A, Summerer S, Povero G, Stivaletta A, …, Janni M (2019) Can high throughput phenotyping help food security in the mediterranean area? Front Plant Sci 10:15
11. Das Choudhury S, Samal A, Awada T (2019) Leveraging image analysis for high-throughput plant phenotyping. Front Plant Sci 10:508
12. Dong X, Lei Y, Wang T, Thomas M, Tang L, Curran WJ, …, Yang X (2019) Automatic multiorgan segmentation in thorax CT images using U-net-GAN. Med Phys 46(5):2157-2168
13. Dong H, Yang G, Liu F, Mo Y, Guo Y (2017) Automatic brain tumor detection and segmentation using U-Net based fully convolutional networks. In: Annual conference on medical image understanding and analysis. Springer, Cham, pp 506–517
14. Dornbusch T, Lorrain S, Kuznetsov D, Fortier A, Liechti R, Xenarios I, Fankhauser C (2012) Measuring the diurnal pattern of leaf hyponasty and growth in Arabidopsis–a novel phenotyping approach using laser scanning. Funct Plant Biol 39(11):860–869
15. Furbank RT, Tester M (2011) Phenomics–technologies to relieve the phenotyping bottleneck. Trends Plant Sci 16(12):635–644
16. Giuffrida MV, Minervini M, Tsaftaris S (2015) Learning to count leaves in rosette plants. In: Tsaftaris SA, Scharr H, Pridmore T (eds) Proceedings of the Computer Vision Problems in Plant Phenotyping (CVPPP). BMVA Press, Swansea
17. Hahnloser RH, Sarpeshkar R, Mahowald MA, Douglas RJ, Seung HS (2000) Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. Nature 405(6789):947–951
18. Isensee F, Petersen J, Kohl SA, Jäger PF, Maier-Hein KH (2019) nnu-net: Breaking the spell on successful medical image segmentation. arXiv preprint arXiv:1904.08128, 1, 1–8
19. Kingma DP, Ba J (2014) Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980
20. Klose R, Penlington J, Ruckelshausen A (2009) Usability study of 3D time-of-flight cameras for automatic plant phenotyping. Bornimer Agrartechnische Berichte 69(93–105):12
21. Lozej J, Meden B, Struc V, Peer P (2018) End-to-end iris segmentation using U-Net. In: 2018 IEEE International Work Conference on Bioinspired Intelligence (IWOBI). IEEE, pp 1–6
22. Minervini M, Fischbach A, Scharr H, Tsaftaris SA (2016) Finely-grained annotated datasets for image-based plant phenotyping. Pattern Recogn Lett 81:80–89
23. Norman B, Pedoia V, Majumdar S (2018) Use of 2D U-Net convolutional neural networks for automated cartilage and meniscus segmentation of knee MR imaging data to determine relaxometry and morphometry. Radiology 288(1):177–185
24. O'Malley RC, Ecker JR (2010) Linking genotype to phenotype using the Arabidopsis unimutant collection. Plant J 61(6):928–940
25. Pape JM, Klukas C (2014) 3-D histogram-based segmentation and leaf detection for rosette plants. In: European conference on computer vision. Springer, Cham, pp 61–74
26. Pound MP, Atkinson JA, Townsend AJ, Wilson MH, Griffiths M, Jackson AS, …, Pridmore TP (2017) Deep machine learning provides state-of-the-art performance in image-based plant phenotyping. Gigascience 6(10):gix083
27. Ramachandran P, Parmar N, Vaswani A, Bello I, Levskaya A, Shlens J (2019) Stand-alone self-attention in vision models. arXiv preprint arXiv:1906.05909
28. Ronneberger O, Fischer P, Brox T (2015) U-net: Convolutional networks for biomedical image segmentation. In: International conference on medical image computing and computer-assisted intervention. Springer, Cham, pp 234–241
29. Russakovsky O, Jia D, Su H, Krause J, Sanjeev S, Ma S, Zhiheng H, Andrej K, Aditya K, Michael B, Alexander B, Li F-F (2015) ImageNet large scale visual recognition challenge. arXiv:409.0575v3
30. Sakurai S, Uchiyama H, Shimada A, Arita D, Taniguchi RI (2018) Two-step transfer learning for semantic plant segmentation. In: ICPRAM. pp 332–339

31. Santos TT, Koenigkan LV, Barbedo JGA, Rodrigues GC (2014) 3D plant modeling: localization, mapping and segmentation for plant phenotyping using a single hand-held camera. In: European conference on computer vision. Springer, Cham, pp. 247–263
32. Scharr H, Minervini M, French AP, Klukas C, Kramer DM, Liu X, …, Yin X (2016) Leaf segmentation in plant phenotyping: a collation study. Mach Vis Appl 27(4):585-606
33. Sevastopolsky A (2017) Optic disc and cup segmentation methods for glaucoma detection with modification of U-Net convolutional neural network. Pattern Recognit Image Anal 27(3):618–624
34. Shelhamer E, Long J, Darrell T (2017) Fully convolutional networks for semantic segmentation. IEEE Trans Pattern Anal Mach Intell 39(4):640–651
35. Smith LN (2017) Cyclical learning rates for training neural networks. In: 2017 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, pp 464–472
36. Wu Y, Liu L, Bae J, Chow KH, Iyengar A, Pu C, …, Zhang Q (2019) Demystifying learning rate policies for high accuracy training of deep neural networks. In: 2019 IEEE international conference on big data (Big Data). IEEE, pp 1971–1980
37. Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhudinov R, Bengio Y (2015) Show, attend and tell: Neural image caption generation with visual attention. In: International conference on machine learning. pp 2048–2057
38. Yang B, Wang L, Wong D, Chao LS, Tu Z (2019) Convolutional self-attention networks. arXiv preprint arXiv:1904.03107
39. Yu F, Koltun V, Funkhouser T (2017) Dilated residual networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp 472–480
40. Zambaldi V, Raposo D, Santoro A, Bapst V, Li Y, Babuschkin I, …, Shanahan M (2018) Deep reinforcement learning with relational inductive biases. In: International conference on learning representations
41. Zhou Z, Siddiquee MMR, Tajbakhsh N, Liang J (2018) Unet++: A nested u-net architecture for medical image segmentation. In: Deep learning in medical image analysis and multimodal learning for clinical decision support. Springer, Cham, pp 3–11

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.