Wheat Plant Counting Using UAV Images Based on Semi-supervised Semantic Segmentation

Hamza Mukhtar
Alkhawarizmi Institue of Computer Science
University of Engineering and Technology
Lahore, Pakistan
hamza.mukhtar@kics.edu.pk

Muhammad Zeeshan Khan
Alkhawarizmi Institue of Computer Science
University of Engineering and Technology
Lahore, Pakistan
zeeshan.khan@kics.edu.pk

Muhammad Usman Ghani Khan
Alkhawarizmi Institue of Computer Science
University of Engineering and Technology
Lahore, Pakistan
usman.ghani@kics.edu.pk

Tanzila Saba
AIDA Lab
CCIS Prince Sultan University
Riyadh, Saudi Arabia
tanzilasaba@yahoo.com

Rabia Latif
AIDA Lab
CCIS Prince Sultan University
Riyadh, Saudi Arabia
rlatif@psu.edu.sa

2021 1st International Conference on Artificial Intelligence and Data Analytics (CAIDA) | 978-1-6654-1511-8/20/$31.00 ©2021 IEEE | DOI: 10.1109/CAIDA51941.2021.9425252

***Abstract*— Plant counting in major grain crops like wheat through aerial images still poses a challenge due to the very high infield density of plants and occlusion. Annotation of aerial images for counting through perfect detection or segmentation is extremely difficult due to a large number of extremely small plant instances. In this paper, we present a semi-supervised method based on cross-consistency for the semantic segmentation of field images and an inception-based regression network for plant counting. Through loosely semantic segmentation, tiny plant clusters are extracted from the RGB image and fed to a regression network to get the count. Cross-consistency under the cluster assumption is a powerful semi-supervised training technique to leverage the unlabeled images. In this work, it is observed that regions with lower density are more detectable within hidden representations as compared to inputs. Supervised training of an encoder in a shared fashion and the main decoder is carried out on the RGB images and the corresponding mask. Consistency between the prediction of main and auxiliary decoders is imposed to leverage the unlabeled images. Induction of inception in the regression network benefits in extracting the multi-scale features which are very important because of quite tiny plant instances as compared to the whole image. The proposed plant counting framework achieves very high performance having a standard deviation of 0.94 and a mean of 0.87 of absolute difference in the count given the semi-supervised nature. Our network has performed reasonably well as compared to supervised detection and segmentation-based counting framework. Moreover, labeling for detection or segmentation is a quite tedious task, so our network has the leverage to train the model with few labeled and large numbers of unlabeled images which also provides the advantage to train the system for other crops like rice and maize with few labeled images.**

Keywords*— *Plant counting, crop estimation, semi-supervised learning, semantic segmentation, consistency training, encoder.

I. Introduction

For grain crops like wheat, the overall productivity of the crop hugely depends on the emergence (plant density in the field) [1]. Optimum wheat plant density is crucial because a vigorous plant density is required for the growth of the crop which largely depends on sunlight, fertilizer, nutrients, and moisture, eventually for achieving the higher yield. Conventional phenotypic estimation based on the manual measurements is labor-intensive and also error prune at the large-scale wheat yield estimation. Furthermore, traditional phenotypic traits are estimated on the sub-sampling approach

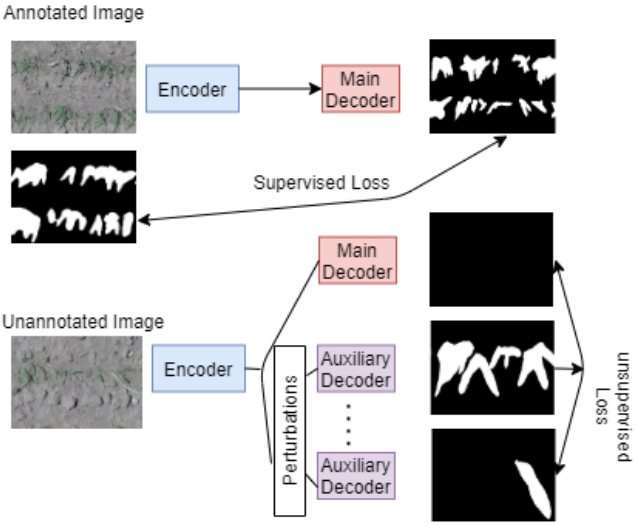

Fig. 1. Proposed consistency training enforced on segmentation model

where estimation is made on specified small fields. To make an efficient, and accurate estimation at a large scale, Computer Vision-based phenotypic measurements are becoming popular due to the performance of deep learning-based techniques, such as object classification, detection, localization, generation [20], and segmentation [21]. Plant count is associated with the well-research problem of leaves counting from plant images, but much challenging because of high occlusion and a large number of plants in the images captured through unmanned aerial vehicles (UAV).

Accurate plant counting is very important for crop management and agricultural experimentation. For the field management purpose, plant counting is used to estimate the seed's growth ratio and yield, whereas, for the latter, it is used to analyze the crop yield at a large scale, food security, procurement, and management for the next crop season. Different plant and different object counting approaches exist in the literature. Recent technological advancements make it more feasible to build non-destructive and autonomous approaches based on remote sensing. Utilization of UAVs in different fields offering an efficient solution for remote data acquisition on the crop fields. Usage of UAVs for agricultural purposes is also increasing for data acquisition along with the different deep learning and Computer vision techniques to

extract relevant information like plant counting from acquired images [2].

Galaxy of researchers has worked on object counting through various state-of-the-art deep learning techniques including counting through the regression [3], segmentation [4], and detection [5]. Although counting through regression is simple and accurate due to explicit learning from images, it requires a huge amount of annotated dataset and works poorly when the number of object instances is large [6]. On the other hand, counting through instances segmentation and detection requires a great amount of highly precise annotation for training which is very tedious to get, and becomes even more difficult for wheat plant counting where images have a huge number of instances per image. Mentioned approaches are not extendable in the standalone form on other object types without training on a large number of images of the new object type.

To build the accurate and extendable crop plant counting mechanism, we have proposed a completely data-driven approach for wheat plant counting from high-resolution UAV images based on semi-supervised semantic plant segmentation to extract the tiny plant cluster from RGB images, and then direct count from extracted small patches.

We have introduced the cross-consistency training [12] concept in our semantic segmentation framework to do training in a semi-supervised fashion to leverage the unannotated images under cluster assumption. Consistency training enforces invariance of the network's prediction over the small perturbations used on the inputs. As a benefit, our segmentation model will be adoptive to variation in the inputs. The robustness of the consistency training is directly proportional to the distribution of data, i.e., the cluster assumption, where the classes need to separate by regions with lower density. Semantic segmentation does not follow the presence of regions with low density distinguishing the classes in the input, rather in the output of the encoder. Based on this analysis, we have enforced consistency over the various perturbations used on the encoder's output. To train the segmentation network on labeled images, a shared and the main decoder are considered, while to leverage the unlabeled data, multiple auxiliary decoders have been used, whose inputs are the perturbed forms of the shared encoder's output. The consistency is enforced between the predictions of those the auxiliary and the main decoder's prediction (Fig. 1.). This enforcement of consistency enhances the representations of the shared encoder through extra training signals extracted from unlabeled images. The injected auxiliary decoders have a very tiny number of parameters than the encoder. During testing, only the main decoder is used which further reduces the overhead of auxiliary decoders. Our proposed approach is efficient and extendable to use pixel-level and weak labels of slightly different crops like rice in a semi-supervised manner.

To the best of our knowledge, this is the first attempt to count the plant from aerial images in a semi-supervised manner. Our contributions are three-fold:

- We have proposed a semi-supervised training through consistency training for which enforce the invariance of predictions into the encoder by injecting various perturbations.
- Plant counting through direction regression on cluster assumption.
- Our design is extendable to different crops like rice and maize by reducing the requirement of a massive annotated image dataset.

In section 2 of this article, various proposed techniques related to plant counting, semi-supervised learning, and semi-supervised semantic segmentation. Section 3 explains the methodology of our consistency training-based semi-supervised binary semantic segmentation model, and direction regression, while section 4 explains the dataset, training environment, provides the evaluation of segmentation, and also compares performance with other approaches. In the end, section 5, concludes the article while giving a future path.

II. Related work

Precision agriculture is receding dividends from UAV images and deep learning techniques which are not only highly accurate but have the robustness to work in the variational environment. In recent times, the Convolutional neural network has played a great part in counting the plant and their leaves through the regression [7, 8]. Inspired by their work, we have adopted a similar approach for wheat plant counting in a semi-supervised manner. Mask-RCNN is also used for corn plant counting [24].

Although object counting through object detection is a very robust and widely used technique in various domains, counting by plant detection is rarely found in the literature. Banana plants are counted using CNN to detect the plant from annotated aerial images of banana fields captured from 40 to 60 meters altitude, and it achieved fairly good results. Instead of detection or segmentation of plants, count through direction is also tried using Inception-v3 network [11]. Various studies have tried to estimate the plant density in wheat [7, 9] and maize [10] from aerial RGM images through semantic image segmentation. And then regression. These approaches follow a supervised learning mechanism where RGB images with corresponding semantic masks are used for the training of such methods. An image augmentation technique based on the randomized minimal region super-pixels is used to train the data-hungry semantic segmentation network [7]. Mask annotation of aerial images of crops like wheat for detection is an extremely difficult and time-consuming task which is also not flexible enough to extend to different crops.

To leverage the availability of large amounts of unannotated data, various methods, such as entropy minimization, pseudo labeling, and graph-based methods, have proposed to introduce semi-supervised learning in deep learning models. In our work, we have focused on consistency training methods [12] only that assume invariance predictions to a large extent when a realistic form of perturbation is enforced on the unannotated data samples. Models of low-density regions are favored with decision boundaries that give consistent predictions for similar input data samples. Π-Model, under dropout effect and various types of data augmentation, applies a consistency on two perturbed forms of the input samples. Virtual Adversarial Training (VAT) estimates the perturbations that will be a major contributor to variation in predictions. Our proposed method gets the semi-supervised effects by applying the consistency of prediction over various perturbations, that are enforced on the output of the encoder rather than the input inputs, between the main decoder and all the auxiliary decoders.

Many approaches use a small number of pixel-level, region-level, or image-level annotations combined with a big number of very loose annotations [18]. Primary localization

Fig. 2. Work of proposed counting method

maps through class activation mapping (CAM) [19] are produced for image-level loose supervision. By combining with pix-level labels, produced maps are refined and used for the training of a segmentation network.

III. METHODOLOGY

This section illustrates the proposed semi-supervised loose binary semantic segmentation-based counting approach where tiny patches extracted from the RGB image by segmentation based on cluster assumption and then fed for counting the number of wheat plants through direct regression. Although segmentation and counting modules are trained separately, however, both modules are not completely independent since the counting module uses the tiny patches for training and testing extracted through semantic segmentation based on the cluster assumption. Fig. 1 Shows the workflow of our plant counting mechanism.

A. Segmentation

Regression counting on the whole image at once is not only computationally expensive in very high-resolution images (~ 2000 x 6000). Moreover, direct regression requires a large number of images, while we not only have a small number of very high resolution but also difficult to count the plants from them for regression training.

1) Cluster Assumption

We have started by analyzing the cluster assumption for the consistency training approach by calculating the localized variation between the value of every pixel and its neighboring pixel to estimating the local smoothness. Mean Euclidean distance at a very spatial location with 16 intermediate neighboring pixels. Following [13], the mean distance of a patch centered at a particular spatial position from its neighboring locations is measured to form a realistic receptive field. Then the feature map is unsampled to the size equal to input, mean distance between the neighboring activations (feature vector of 1024-dimensions). On the input level, regions with lower density are not completely aligned with the class boundaries, that's why cluster assumption is not fully observed. On the other hand, this assumption is fully enforced at the encoder's output, having a higher mean distance of class boundaries as compared to corresponding lower density regions. This shows that perturbations are enforced better on the encoder's output than the input.

2) Consistency Training

Our objective is to leverage the unannotated images to train the semantic segmentation network that can work efficiently on the images taken from similar data distribution as the training images. Our semantic segmentation network (Fig. 3) has a main decoder D and a shared encoder E which together build the network There are also a set of N auxiliary decoders. Our segmentation model is trained on the tiny annotated image dataset in a supervised manner, while the auxiliary decoder network is trained on the unannotated image dataset by applying consistency on the predictions between the main and the auxiliary decoders. The main encoder receives uncorrupted representation, and perturbed versions of the encoder's output are fed to every auxiliary decoder. This mechanism of the main encoder and auxiliary decoders enhances the representation learning of the encoder E leverages the unannotated images. This improvement in the encoder propagates and enhances the learning of the semantic segmentation model.

Extra training signals are extracted by applying the consistency on annotated dataset between the main and the auxiliary decoder network. Formally, our supervised loss L_s is trained with Cross-Entropy for annotated training images x_i^l, and its pixel-level label y_i between the main and auxiliary decoder's output.

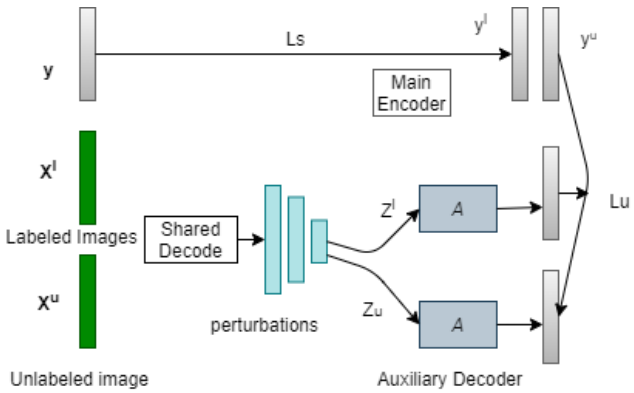

Fig. 3. For one training iteration, a labeled image and label y together with an unlabeled image are sampled to feed to the encoder and the main decoder, which produces two outputs, and supervised loss is calculated and then apply different perturbation to feature map Z to produce auxiliary outputs. Finally, unsupervised loss, between the outputs of the auxiliary and the main decoder is calculated.

$$Ls = \frac{1}{|Dl|} \sum_{x,y=0}^{l} H(y_i, f(x_i{}^l))) \quad (1)$$

With H, for annotated images xi^u, intermediate input's representations are measured from the shared encoder. From R perturbation functions, N perturbed versions of the intermediate representation are produced which are fed to the corresponding auxiliary decoder while assuming that one perturbation function p can be associated with more than one auxiliary decoders. Here training objective is to minimize the unsupervised loss Lu that calculates the level of inconsistency

$$Lu = \left(\frac{1}{|D_u|}\right)\frac{1}{N}\sum_{x^u=0}\sum_{n=0}^{N} D(d(p_i), d^n(p_i))\,) \quad (2)$$

Here **D** is the Mean Square Error (MSE) as a distance measure between the output of the main decoder and that of auxiliary outputs where *SoftMax* function's output is used over the dimensions of the channel. The aggregated loss *L* for consistency based semi-supervised learning model is measured

$$L = L_s + w_u L_u\,) \quad (3)$$

Here w_u is the weight associated with the unsupervised loss. The supervised loss is measured through the main encoder's output and the ground truth annotations, while unsupervised loss is measured through MSE between the output of the auxiliary and the main decoder. At each iteration, the exactly same number of images are drawn from an annotated and unannotated dataset. Finally, the aggregated loss is backpropagated to train the semantic segmentation network along with the auxiliary decoder's network.

3) Perturbation functions

Three perturbation functions, random, feature, and prediction-based, are applied to the hidden representation to establish consistency.

a) Random perturbations: Such perturbations are based on spatial dropout. [14].

b) Feature-based functions: Such consists of noise addition or dropping the encoder's feature map activations. To introduce the noise uniformly, a noise tensor of the equal size as a feature map is injected into the encoder's output through the multiplication of the noise tensor with the feature map. A fraction of activation ranging from 60% to 80% is sampled and normalized after summing on the channel dimension to produce the mask to get the perturbed form of the feature map. This process masks the 20% to 40% regions of the feature map which are most active.

c) Prediction-based function: These perturbation functions are based on the main and auxiliary decoder's output. Masking-based perturbations (Con-Mask, Obj-Mask, and G-Cutout) are used along with adversarial (I-VAT [15]). Two perturbed forms of feature map through masking of the detected object (Obj-mask), and the masking based on the context (Con-Mask). [16]. For the auxiliary decoder, adversarial perturbations are built that change the prediction by injecting the noise.

B. Regression

Based on the plant cluster, we have extracted the small patches from the RGB image, and a CNN network [17] is trained to count the number of plants from each tiny patch, and then aggregate the count from all patches of one image. To inject higher laten concepts, we have replaced the convolutional layers with the inception units.

IV. Experiment and Evaluation

A. Segmentation architecture

1) Encoder: Our encoder is based on Inception-ResNet [22] pre-trained on the ImageNet dataset given by [23]. We have replaced the last two strided convolutional of encoder with the dilated convolutions.

2) Decoders: To reduce the contain the number of parameters and take efficiency in regard, only 1 x 1 convolutions are used. After the 1 x 1 convolution, three sub-pixel convolutions with ReLU activation function to upsample the output to make it the same size as of input

B. Dataset

The dataset for the plant count is taken from [7] which has very high-resolution (~ 2000 x 6000) images of 24 wheat plots. We have created 120 non-overlapping images having size 420 x 640 from high-resolution original images which are annotated as well. These images are used for the training of our network along with the1500 unannotated images in a semi-supervised fashion. Fig. 4 shows some sample images and corresponding mask images.

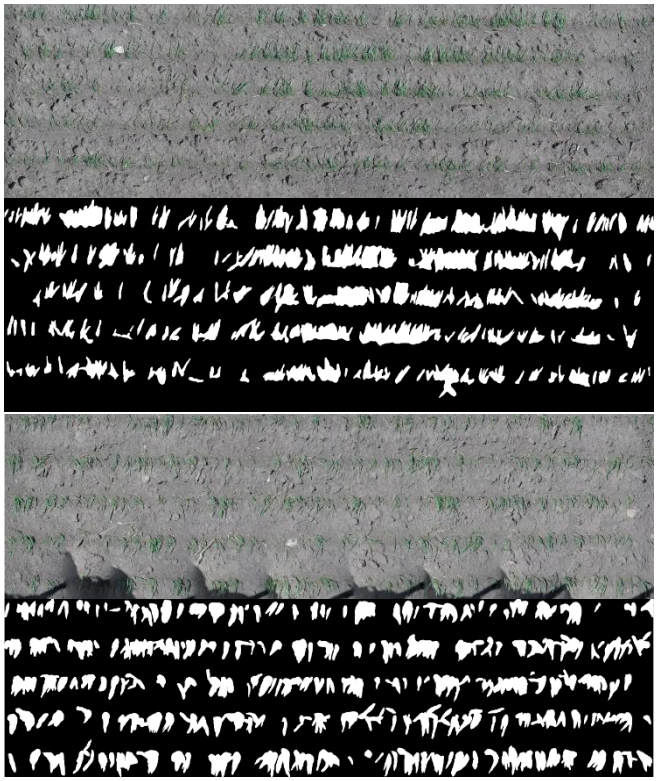

Fig. 4. Samples images and corresponding masks.

C. Training setting

We have trained the segmentation network for 250 epochs with SGD-momentum as an optimizer having a learning rate of 0.02 with momentum of 0.6 and weight decay 0.0002. Our counting network is trained on 10000 images extracted from 100 images along with augmented forms.

D. Result and Comparison

Our plant counting model is evaluated in two phases. First, we have evaluated the performance of semi-supervised segmentation for producing the binary segmentation of RGB images. Then we have measured the performance of the

counting network on Mean Absolute deviation and Std absolute difference (SDAD). Precision, recall, and accuracy are used as the performance measure of the segmentation network. For the evaluation of the semantic segmentation network, our validation dataset has 120 images and corresponding annotation.

TABLE I. PERFORMANCE EVALUATION OF SEGMENTATION NETWORK

Measure	%
Precision	88.83
Recall	85.86
Accuracy	95.02

Our segmentation network has achieved (~88) precision and (~86) recall, both values are low to some extent because annotations are not perfectly annotated and unannotated images also played their part in these lower values of recall and precision. However, our objective is to get an accurate count, not the precise binary segmentations.

TABLE II. PERFORMANCE COMPARISON OF COUNTING NETWORK

Model	MAD	SDAD	Semi-supervised
Leaf counting [17]	1.62	2.30	No
Wheat plant counting [7]	1.05	1.40	NO
Our Counting network	0.87	1.14	Yes

1120 small labeled images having count labels ranging from 0 to 19 are used for the validation of the counting module. Tab. 2 is showing a performance comparison of our plant counting module with few similar approaches. We have obtained MAD of 0.87 and SDAD of 1.14 which shows that our model is more accurate while working on a semi-supervised mechanism.

V. CONCLUSION

In this article, we have tried to use the abundance of unlabeled images for the semi-supervised binary semantic segmentation task that further is used to extract the small plant cluster regions to perform plant counting. Consistency training is a quite efficient and flexible method for semi-supervised training of the segmentation network. Although we have not evaluation on different grain crops, we expect that our method is extendable to other crops with fine-tuning of the segmentation and counting network on fewer labeled examples.

ACKNOWLEDGMENT

Authors pay their thanks to Intelligent Criminology Lab, National Center for Artificial Intelligence, Al-Khawarizimi Institute of Computer Science, UET Lahore for providing research platform, technical support, and financial assistance.

REFERENCES

[1] Ribera J, Chen Y, Boomsma C, "Delp EJ. Counting plants using deep learning,". In 2017 *IEEE global conference on signal and information processing (GlobalSIP)* 2017 Nov 14 (pp. 1344-1348). IEEE.

[2] Valente J, Sari B, Kooistra L, Kramer H, Mücher S, "Automated crop plant counting from very high-resolution aerial imagery," *Precision Agriculture*. 2020 Dec;21:1366-84.

[3] Onoro-Rubio D, López-Sastre RJ, "Towards perspective-free object counting with deep learning," In *European conference on computer vision* 2016 Oct 8 (pp. 615-629). Springer, Cham.

[4] Cholakkal H, Sun G, Khan FS, Shao L, "Object counting and instance segmentation with image-level supervision," In Proceedings of the *IEEE/CVF Conference on Computer Vision and Pattern Recognition* 2019 (pp. 12397-12405).

[5] Wang Y, Hou J, Hou X, Chau LP, "A Self-Training Approach for Point-Supervised Object Detection and Counting in Crowds," *IEEE Transactions on Image Processing*. 2021 Feb 4;30:2876-87.

[6] Chattopadhyay P, Vedantam R, Selvaraju RR, Batra D, Parikh D, "Counting everyday objects in everyday scenes," In Proceedings of the *IEEE Conference on Computer Vision and Pattern Recognition* 2017 (pp. 1135-1144).

[7] Aich S, Josuttes A, Ovsyannikov I, Strueby K, Ahmed I, Duddu HS, Pozniak C, Shirtliffe S, Stavness I, "Deepwheat: Estimating phenotypic traits from crop images with deep learning," In 2018 *IEEE Winter Conference on Applications of Computer Vision (WACV)* 2018 Mar 12 (pp. 323-332). IEEE.

[8] Aich S, Stavness I, "Leaf counting with deep convolutional and deconvolutional networks," In Proceedings of the *IEEE International Conference on Computer Vision Workshops* 2017 (pp. 2080-2089).

[9] Liu S, Baret F, Andrieu B, Burger P, Hemmerle M, "Estimation of wheat plant density at early stages using high resolution imagery," *Frontiers in Plant Science*. 2017 May 16;8:739.

[10] Gnädinger F, Schmidhalter U, "Digital counts of maize plants by unmanned aerial vehicles (UAVs)," *Remote sensing*. 2017 Jun;9(6):544.

[11] Ribera J, Chen Y, Boomsma C, Delp EJ, "Counting plants using deep learning," In 2017 *IEEE global conference on signal and information processing (GlobalSIP)* 2017 Nov 14 (pp. 1344-1348). IEEE.

[12] Ouali Y, Hudelot C, Tami M, "Semi-Supervised Semantic Segmentation with Cross-Consistency Training," *Supplementary Material*.

[13] French G, Aila T, Laine S, Mackiewicz M, "Finlayson G. Semi-supervised semantic segmentation needs strong, high-dimensional perturbations," *arXiv preprint arXiv*:1906.01916. 2019 Jun 5.

[14] Tompson J, Goroshin R, Jain A, LeCun Y, Bregler C, "Efficient object localization using convolutional networks," In Proceedings of the *IEEE conference on computer vision and pattern recognition* 2015 (pp. 648-656).

[15] Miyato T, Maeda SI, Koyama M, Ishii S, "Virtual adversarial training: a regularization method for supervised and semi-supervised learning," *IEEE transactions on pattern analysis and machine intelligence*. 2018 Jul 23;41(8):1979-93.

[16] Oliva A, Torralba A, "The role of context in object recognition," *Trends in cognitive sciences*. 2007 Dec 1;11(12):520-7.

[17] S. Aich and I. Stavness, "Leaf counting with deepconvolutional and deconvolutional networks," *CoRR*,abs/1708.07570, 2017.1,2,4,7

[18] Wei Y, Xiao H, Shi H, Jie Z, Feng J, Huang TS, "Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation," In Proceedings of the *IEEE Conference on Computer Vision and Pattern Recognition* 2018 (pp. 7268-7277).

[19] Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A, "Learning deep features for discriminative localization," In Proceedings of the *IEEE conference on computer vision and pattern recognition* 2016 (pp. 2921-2929).

[20] Khan MZ, Jabeen S, Khan MU, Saba T, Rehmat A, Rehman A, Tariq U, "A realistic image generation of face from text description using the fully trained generative adversarial networks," *IEEE Access*. 2020 Aug 10.

[21] Khan MZ, Hassan MA, Hassan SU, Khan MU, "Semantic Analysis of News Based on the Deep Convolution Neural Network" In2018 *14th International Conference on Emerging Technologies (ICET)* 2018 Nov 21 (pp. 1-6). IEEE.

[22] Szegedy C, Ioffe S, Vanhoucke V, Alemi A, "Inception-v4, inception-resnet and the impact of residual connections on learning," InProceedings of the *AAAI Conference on Artificial Intelligence* 2017 Feb 12 (Vol. 31, No. 1).

[23] Ansheng You, Xiangtai Li, Zhen Zhu, and Yunhai Tong, "Torchcv: A pytorch-based framework for deep learning in computer vision," https://github.com/donnyyou/ torchcv, 2019.

[24] Machefer M, Lemarchand F, Bonnefond V, Hitchins A, Sidiropoulos P, "Mask R-CNN Refitting Strategy for Plant Counting and Sizing in UAV Imagery," *Remote Sensing*. 2020 Jan;12(18):3015.