Semantic Segmentation Using U-Net Deep Learning Network for Quince Phenotyping on RGB and HyperSpectral Images

K.Sudars[1], I.Namatēvs[1], A.Ņikuļins[1], R.Balašs[1], A.Peter[1], S.Strautiņa[2], E.Kaufmane[2], I.Kalniņa[2]
[1]Institute of Electronics and Computer Science,
Dzerbenes St. 14, LV-1006, Riga, Latvia
[2]Institute of Horticulture,
Graudu St. 1, LV-3701, Dobeles novads, Latvia
{kaspars.sudars, ivars.namatevs, arturs.nikulins, rihards.balass, astile.peter}@edi.lv;
{sarmite.strautina, edite.kaufmane, ieva.kalnina}@llu.lv

Abstract - Semantic segmentation based on the deep learning techniques can be used for the non-invasive phenotyping of quinces. In this paper we present a deep neural network for generating pixel wise mask from RGB and Hyperspectral images of the quinces using the U-Net architecture. The generated mask will be very useful for the experts involved in the phenotyping for the getting the dimension of the quinces. Also it can be used in the future for automatic plucking of quinces by the robot. This paper also compares the evaluation metrics of the model trained on both RGB and HSI data. We were able to achieve an accuracy of 93.33% and 70.225% for HSI and RGB data respectively. The developed segmentator is freely available in the GIT repository. The future works will include the model for detecting the ripeness of the quinces from the HSI data and also HSI images will be included in the dataset which will be helpful for the experts who are researching about the other fruits.

Index Terms—Deep learning, Deep neural networks, U-Net, Phenotyping.

I. INTRODUCTION

Deep learning techniques can help the agricultural sector by helping the farmers to estimate the yield, prevention of crop diseases, monitor the crops using drones, etc. Hyperspectral imaging (HSI) which contains more elaborate information about the spectrum of light for each pixel will help the non-invasive phenotyping of crops. In this paper, we will generate the masks from HIS of the Japanese quince (Chaenomeles japonica) using the deep learning architecture called U-Net. The architecture of the U-Net was modified to get better results for our purpose by adding a batch normalization layer and introducing the padding.

The predicted mask can be used for getting the dimension of the quinces and in the future maybe for the computer vision tasks. The images were captured using a hyperspectral camera called Specim IQ. The pre-porcessing of the data was quite important for the HSI model because the HSI dataset was quite complicated compared to the RGB dataset. The training has been done on both HSI data and RGB data. All the metrics obtained after training both datasets were compared to decide which approach is better. The paper also compares the predicted mask and ground truth pixel wise and shows the best and worst cases of the prediction according to their Jaccard indexes. We were able to acquire a mean accuracy of 93.33% with only 47 images in the training dataset and 12 in the validation dataset. Since the model and training was implemented in the PyTorch framework and the trained model is available freely in the Git, it will be useful for the researchers to implement or modify the parameters.

II. RELATED WORKS

Anand et Al [3] proposed a framework for IoT-assisted precision agriculture using the deep aerial semantic segmentation. Unmanned Aerial Vehicles (UAV) were used to acquire the images. DeepLabV3+ with ResNet (Output Channel = 128 and Output Stride = 8) was used for the feature extraction. Wang et Al [4] developed a semantic segmentation model based on the Encoder-Decoder network for crops and weeds. The images were pre-processed before feeding to the network. Transfer learning was also implemented due to less number of images for training and validation. The detection of multi-species fruit flowers with the help of semantic segmentation network [5]. In this paper, the model was fine tuned using the single set of apple flower images. The DEEPLAB and RGR refined module was combined to get better prediction and recall rates. Liyanage et Al [6] proposes a method to annotate RGB images and semantic segmentation for autonomous driving using the hyperspectral imaging. In the paper they compared the predicted mask with manually annotated ground data. Detection of plant responses to drought using the HSI and high throughput phenotyping platform [7]. In the paper they successfully detected the drought responses at early stages and revealed the recovery effects after re-watering period. Recognition and counting of spikes from images of wheat plants using deep learning frameworks [8]. In this context, T. Misra et Al developed Web-SpikeSegNet which will count the spikes in a non-destructive with the help of deep learning techniques and image analysis.

III. METHODOLOGY

A. U-Net

U-Net was developed in Germany by researchers at the University of Freiburg [9]. It is a fully convolutional neural network and was originally designed for processing biomedical images. When a raw image is fed into the network it will produce an output segmentation map and works decently with the touching objects of the same class. Since the Quinces will be very close to each other in the tree, the U-Net will be the perfect architecture for our purpose.

The U-Net consists of an Encoder and Decoder part, the encoder uses the convolutional and max pooling layers and it is further divided into many layers. Each layer consists of two 3x3 convolutional layers followed by a ReLu activation layer. The transition between the layer is done with the help of a max pooling layer. While the encoder is responsible for gaining knowledge about the image, the decoder is responsible for figuring out the location of pixels on the needed images. For localization the skip connections are deployed, where the feature map of the encoder is concatenated to the output of the transpose convolution of the same layer. But unlike the original paper, the size of output and input images will be the same. We modified the architecture by introducing the batch normalization layer between the convolutional and ReLu activation layer. The padding of one is also used in the convolutional layer to prevent further downsizing. The model was implemented with the help of PyTorch framework. The architecture of U-Net is shown in the Fig. 1.

B. Training and Validation Dataset

Hyper spectral images (HSI) were used for the training. The Quinces used for the dataset were grown locally in the Institute of Horticulture, Latvia. The HSI camera used for capturing the image was Specim IQ. The images captured using the camera had the size 512 X 512 and 204 number of channels. 47 images were used for the validation and 12 for the validation. Each images is the dataset has a wavelength band from 400nm to 1000nm. Since HSI is very rich in information, the memory occupied by the images was very high, so the batch size used for training was one. The image annotation was done with the help of VGG image annotator. An example of the preview of hyperspectral images represented using RGB format is shown in Fig. 2. The pre-processing is important before the training of the model. Since there will be two files associated with only one image and training is carried out in PyTorch framework. The PyTorch dataloader class will get confused during indexing. So in order to solve this problem the HSI data and its respective header file should be converted to the NumPy array and store this arrays as a database. The NumPy arrays along with its respective masks are fed into the network for the training and validation. The channel associated with the color can be extracted the HSI data by dividing the difference between the bandwidth of the Specim camera which is 600nm and wavelength of the particular color and the number of channels which is 204. The training was done on NVIDIA A100 GPU card.

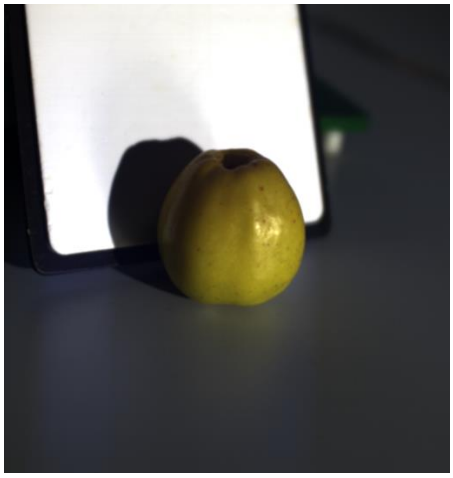

Fig. 2. Preview of image in the dataset.

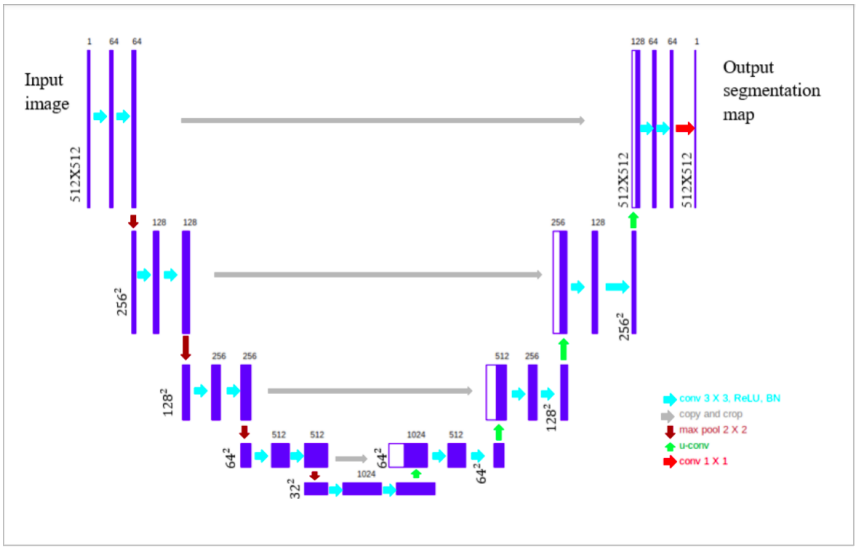

Fig. 1. U-net deep neural network architecture.

C. Evaluation

The model is evaluated by the metrics like Jaccard index, Accuracy, Precision, Recall and F1 score. Jaccard index: The intersection over union (IoU) metric also referred as Jaccard index is calculated by the equation:

$$IoU = \frac{AreaOfOverlap}{AreaOfUnion} \quad (1)$$

Accuracy is also an important metric to evaluate the performance of the network. The network is given by:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

TP - True Positive, TN - True Negative, FP - False Positive and FN - False Negative. Precision gives an idea about how good is the positive detection relative to the ground truth and it is defined by:

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

While recall gives the completeness of positive prediction compared to the ground truth:

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

F1 score is the harmonic mean between the precision and recall and it's given by the equation:

$$F1 = \frac{2 * TP}{2*TP+FP+FN} \quad (5)$$

The Loss is calculated for both the training and validation for saving the weights. Since it is only binary class segmentation which have only class, the BCE-Dice loss is calculated. This loss is a sum of binary cross entropy loss (BCE) and Dice loss.

IV. RESULTS

The number of epochs has been set to 50 epochs for the training of model on HSI data and 200 epochs for the RGB data. The weights will be saved based on the validation loss. If the validation loss of the current epoch is less than the validation loss of the previous epoch the weights, checkpoints will be saved. The training and validation losses of the models trained on both the data are shown in the Fig. 3 and Fig. 4 respectively. It is clear from the Fig. 3 and Fig. 4 that the validation loss is minimum at the 40th epoch for the model trained on the HSI data and 200th epoch for the RGB data. The RGB model was learning quite slow after the 100th epoch, but the loss was still decreasing. The Fig. 5 and Fig. 6 shows the ground truth and predicted mask generated by the RGB and HSI model respectively. The images for testing the model are fed to the network and the best and the worst cases generated by model trained on the HSI data according to the Jaccard index are shown in the Fig. 7. The blue and red portion represents the ground truth and predicted mask respectively. The only aspect that model trained on RGB data is leading is 12.83 frame per second (FPS) compared to 1.73.

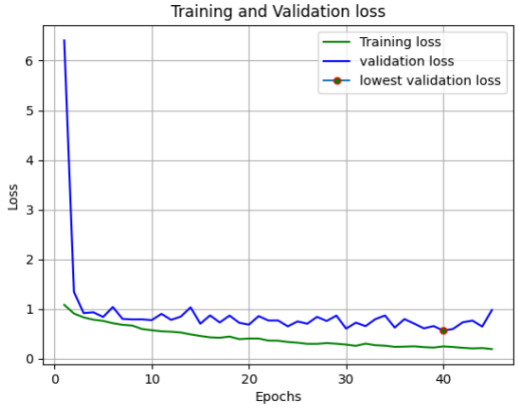

Fig. 3. Losses vs Epochs.

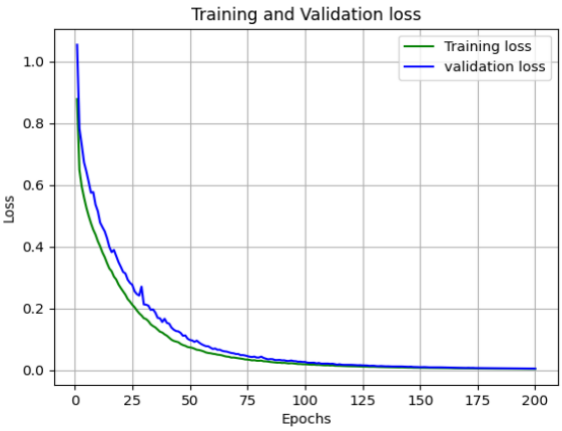

Fig. 4. The training vs validation loss of model trained on RGB data.

Fig. 5. The ground truth and predicted mask generated by model trained on RGB data.

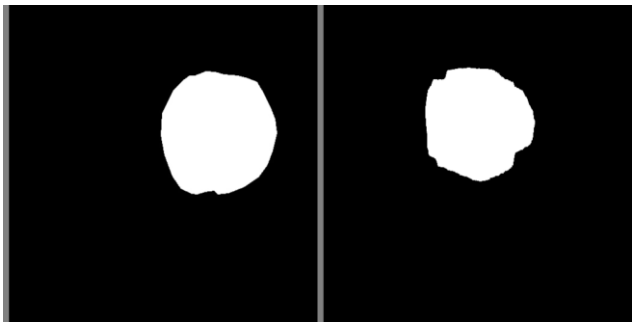

Fig. 6. The ground truth and predicted mask generated by model trained on HSI data.

The mean value of all the evaluation metrics obtained after training the model on both, the data are shown in the Table I and Table II respectively. Despite the same number of images on both, the dataset are same all the metrics observed after the training was considerable smaller for the model trained on the RGB images. These effects were visually observable in the predicted mask compared to the ground truth.

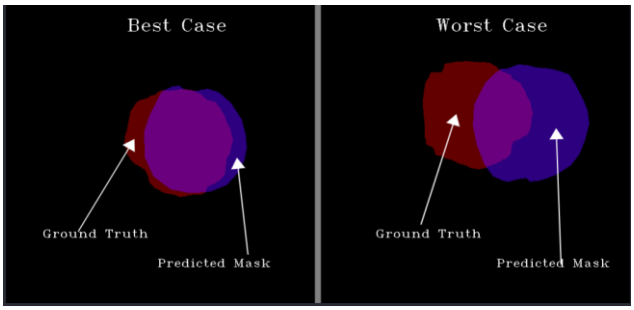

Fig. 7. Best and Worst Case.

TABLE I
EVALUATION METRICS OF MODEL TRAINED ON HSI DATA

Metric	Value
Jaccard index	0.0246
F1	0.0475
Recall	0.0686
Precision	0.0372
Accuracy	0.7025

TABLE II
EVALUATION METRICS OF MODEL TRAINED ON RGB DATA

Metric	Value
Jaccard index	0.5476
F1	0.6911
Recall	0.7550
Precision	0.6420
Accuracy	0.9330

V. CONCLUSIONS

This paper presented the generation of masks for Quinces using the U-Net architecture. Since the dataset was of Hyperspectral images which are rich in information, the model trained on the HSI data was better in any of the metrics as compared to model trained on the RGB data. FPS is the only metric were the model, trained on the RGB data, is leading. The weight obtained after the training for generating pixelwise mask was quite accurate in case of HSI data was quite accurate even though the number of the images in the dataset was very small. These generated masks will be quite helpful for the experts to get the dimension of the quinces for either research or estimation of the yield. The model can be improved by adding the HSI images to the already existing dataset and train the model on the enlarged data.

ACKNOWLEDGMENT

This work was funded by Latvian Council of Science project No. LZP-2020/1-0353 “Smart non-invasive phenotyping of raspberries and Japanese quinces using machine learning and hyperspectral and 3D imaging AKFEN”.

REFERENCES

[1] GIT internet repository for open quince segmentation: https://pubgit.edi.lv/kaspars.sudars/akfen-semantic-segmentation

[2] Quince HIS and RGB data used in experiments: https://makonis.edi.lv/s/nMo6nzELY3JmKWM

[3] T. Anand, S. Sinha, M. Mandal, V. Chamola and F. R. Yu, “AgriSegNet: Deep Aerial Semantic Segmentation Framework for IoT-Assisted Precision Agriculture,” in IEEE Sensors Journal, vol. 21, no. 16, pp. 17581-17590, 15 Aug.15, 2021, doi: 10.1109/JSEN.2021.3071290.

[4] A. Wang, Y. Xu, X. Wei and B. Cui, “Semantic Segmentation of Crop and Weed using an Encoder-Decoder Network and Image Enhancement Method under Uncontrolled Outdoor Illumination,” in IEEE Access, vol. 8, pp. 81724-81734, 2020, doi: 10.1109/ACCESS.2020.2991354.

[5] P. A. Dias, A. Tabb and H. Medeiros, “Multispecies Fruit Flower Detection Using a Refined Semantic Segmentation Network,” in IEEE Robotics and Automation Letters, vol. 3, no. 4, pp. 3003-3010, Oct. 2018, doi: 10.1109/LRA.2018.2849498.

[6] D. C. Liyanage, R. Hudjakov and M. Tamre, “Hyperspectral Imaging Methods Improve RGB Image Semantic Segmentation of Unstructured Terrains,” 2020 International Conference Mechatronic Systems and Materials (MSM), 2020, pp. 1-5, doi: 10.1109/MSM49833.2020.9201738.

[7] M. S. M. Asaari, S. Mertens, S. Dhondt, N. Wuyts and P. Scheunders, “Detection of Plant Responses to Drought using Close-Range Hyperspectral Imaging in a High-Throughput Phenotyping Platform,” 2018 9th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS), 2018, pp. 1-5, doi: 10.1109/WHIS- PERS.2018.8747228.

[8] T. Misra et al., "Web-SpikeSegNet: Deep Learning Framework for Recognition and Counting of Spikes From Visual Images of Wheat Plants," in IEEE Access, vol. 9, pp. 76235-76247, 2021, doi: 10.1109/ACCESS.2021.3080836.

[9] O. Ronneberger, P. Fischer, T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation" arxiv:1505.04597 Comment: conditionally accepted at MICCAI 2015.