ITM Web of Conferences **54**, 01001 (2023) https://doi.org/10.1051/itmconf/20235401001
I3CS-2023

Efficient Pomegranate Segmentation with UNet: A Comparative Analysis of Backbone Architectures and Knowledge Distillation

Shubham Mane[1,*], *Prashant* Bartakke[1,**], and *Tulshidas* Bastewad[2,***]

[1]Department of Electronics and Telecommunication Engineering, COEP, Technological University, India-411005
[2]Department of Farm Implements and Machinery, Mahatma Phule Krishi Vidyapeeth, Rahuri-413722

Abstract. This work examines the segmentation of on-field images of pomegranate fruit using UNet model with different backbones. Precise and effective segmentation of pomegranate fruits on the field is essential for automating yield estimation, disease detection, and quality evaluation in the agricultural industry. The models have been trained and validated using actual images captured in a pomegranate field. The study assesses the performance of many backbones, including ResNet50, Inception ResNetV2, MobileNetv2, DenseNet121, EfficientNet, VGG16, and VGG19. The VGG19 backbone achieved the highest F1 score, 90.35 %, according to the data. In addition, we employed feature-based knowledge distillation to move the knowledge from the VGG19 backbone to the lighter MobileNetv2 backbone (45x smaller than VGG19 in number of parameters), which increased the F1 score of MobileNetv2 from 86.97% to 89.91%. Our findings show that the effectiveness of the UNet model for pomegranate fruit segmentation is greatly impacted by the selection of the backbone architecture, and that knowledge distillation can improve the accuracy of UNet models with lighter backbones without significantly increasing their computational complexity.

1 Introduction

The pomegranate is a widely cultivated fruit that has a high nutritional content and is valued for its abilities to improve health. Automated systems based on computer vision techniques have been developed for pomegranate yield prediction in response to the recent increase in demand for pomegranates and associated goods, disease diagnosis, and quality assessment. Segmentation by hand takes longer and is more prone to mistakes. The division of an image into different parts or segments based on their visual characteristics, such as colour, texture, or shape, is a process known as segmentation. It enables the extraction of relevant information and features from images, which is a crucial step in image analysis and computer vision. In the context of pomegranate fruit detection, segmentation can be utilised to separate the pomegranate fruit from other objects and the background [1]. In the agriculture

*e-mail: maness19.extc@coep.ac.in
**e-mail: ppb.extc@coep.ac.in
***e-mail: bastewadtb71@gmail.com

industry, precise segmentation of poemgranate fruit is an essential preprocessing step for various applications such as on-field disease detection, yield estimation, and quality assessment. The complex and diverse background can negatively impact the performance of these applications. Accurate segmentation is necessary to eliminate unwanted features that may confuse the model, ensuring reliable results in real-time scenarios by properly isolating the background from the image.

On-field pomegranate fruit images are difficult to segment because of a variety of issues, including their irregular shape, occlusions, variations in size and color, and lighting conditions [2]. To overcome these obstacles, it is necessary to create reliable computer vision algorithms that can account for these elements and segment pomegranate fruits accurately.

Deep learning models recognize and learn from data patterns using artificial neural networks. These models excel at image segmentation, which divides an image into regions based on its content. Deep learning models can accurately segment images by features. UNet models [3] are popular for image segmentation.The image segmentation architecture comprises of two modules namely encoder and decoder. UNet is widely used in medical industry for medical image segmentation. Its performance on images like pomegranate fruit has not been thoroughly investigated. Thus, the UNet model's pomegranate fruit segmentation performance must be evaluated.

To attain high accuracy, deep learning models need a lot of data, and gathering this data for the segmentation of pomegranate fruit might be challenging. Deep learning models are also expensive to compute, which may restrict their usefulness. Deploying such models on devices with limited resources can also be difficult. To overcome these difficulties, we suggest using knowledge distillation [4] to move information from a larger model to a smaller one, increasing the model's efficiency while preserving its accuracy.

We have made a number of contributions to the field of deep learning-based pomegranate fruit segmentation in this study. First, we developed a dataset of real-world images of pomegranate fruit and used data augmentation methods to boost its diversity. The effectiveness of UNet [3] models with various backbone architectures was assessed, and we showed how the choice of backbone architecture affected the model's performance. Third, we have demonstrated that feature-based knowledge distillation is capable of helping transfer knowledge from larger, more complex models to smaller, simpler models. This has allowed us to obtain nearly state-of-the-art performance with substantially lower computing costs. Last but not least, our research sheds light on the useful aspects for developing and testing deep learning models for image segmentation tasks. Overall, our work can contribute to the improvement of deep learning models for tasks like pomegranate fruit segmentation as well as additional agricultural image analysis tasks.

The remaining sections of this work are organized as follows: Work related to Section 2, Part 3 describes the study's methodology, including data preparation, model design, and performance assessment measures. Section 4 contains the experimental findings, followed by Section 5's discussion and conclusion. Section 6 is reserved for future work.

2 Related work

A semantic segmentation network's performance depends on its classification network. VGG-16, ResNet34, ResNet101, and Xception were compared as UNet, ResNet34, and AD-LinkNet backbones using the CVPR DeepGlobe road extraction dataset. VGG-16 extracted long and wide roads better than ResNet, but ResNet extracted small roads better. Xception, however, handled complex occlusion situations while retaining ResNet34's features [5]. Gómez-Flores et. al. have evaluated four CNN-based semantic segmentation models developed by the computer vision community: Fully Convolutional Network (FCN) with AlexNet,

Figure 1: Samples of captured on-field pomegranate fruit images

UNet, SegNet, and DeepLabV3+. This study compared eight popular CNN architectures using a larger BUS dataset than current methods. ResNet18 trains faster than SegNet and DeepLabV3+ networks. Early methods prioritized accuracy over computational efficiency [6]. Darknet19, MobileNet, and ShuffleNet backbone networks test LiteSeg architecture for multiple accuracy-computational cost trade-offs.At 161 fps and 640x360 resolution on Cityscapes, LiteSeg using MobileNetV2 as the backbone network achieves a mean intersection over union of 67.81 percent. [7].

To enhance the performance of compact networks, knowledge can be transferred from a complex model to a simple one using knowledge distillation [4]. It has been used to classify images by either transferring the intermediate feature maps [8, 9] or using the class probabilities generated by the bulky model as soft targets for training the small model [4, 10, 11]. Other applications exist as well, such as pedestrian re-identification [12], object detection [13], and others. Focusing on label variations between patches at the local level, the authors then distill the class probabilities for each pixel similar to pixel-wise distillation. and holistic distillation, which transmits holistic knowledge and records high-order data. We were inspired to develop our more straightforward methodology by independently developed applications for semantic segmentation [14, 15].

3 Methodology

The method for selecting a teacher backbone with higher segmentation accuracy and a student backbone that can be practically be deployed under real-world constraints, and then improving the accuracy of the student model, is described in next subsections: dataset preparation, model architecture, performance enhancement, and evaluation metrics.

3.1 Dataset Preparation

3.1.1 Collection

We used a smartphone to capture 3000 pictures of pomegranate fruits from a pomegranate field in Warkute Village, Baramati, in order to prepare the dataset. We made sure the pictures included a variety of viewpoints, lighting, and orientations as shown in Figure 1. Using the labelImg tool, we manually labelled each image to produce a ground truth mask of the pomegranate fruits.

3.1.2 Augmentation

We used many augmentation techniques on the original images to expand the dataset's size and diversity, as well as strengthen the model's robustness. These methods included flipping the image horizontally and vertically, random rotations, blurring, brightening, and random cropping.

3.1.3 Preprocessing

The RGB images were first preprocessed by being resized to 512x512x3 pixels and normalized to obtain the pixel values in the range of 0 to 1.

3.2 Model Architecture

3.2.1 Segmenatation model

The pomegranate fruit image segmentation problem is carried out using the UNet model, which comprises of an encoder and a decoder network as shown in Figure 2. Using convolutional layers, the encoder captures features from the input image, and the decoder upsamples the features to produce the segmentation mask. Skip connections are used by the model to increase segmentation precision. With the aid of this encoder-decoder architecture, a broad receptive field can be used to record contextual data while preserving the fine details of the segmented object. Another benefit of the UNet model is that it can accurately segment data with fewer training samples while maintaining spatial information.

3.2.2 Backbones

Seven widely used state-of-the-art classification models were used as backbones (encoder) to test the performance of the UNet model with various backbone architectures: DenseNet121 [16], EfficientNet [17], Inception ResNetV2 [18], MobileNetv2 [19], ResNet50 [20], VGG16 [21], and VGG19 [21]. Since different backbones have varied capacities for feature extraction from the input image, the backbone architecture of the UNet model serves as the feature extractor and is essential to the model's accuracy and efficacy.

3.3 Performance enhancement

3.3.1 Knowledge transfer to backbone

Here, we have employed feature-based knowledge [22] and offline distillation to transfer knowledge from a larger backbone UNet model (teacher) to a smaller, more efficient backbone (student). The strategy consisted of minimizing the distance between the intermediate features of the teacher and student models, with the feature maps of the three skip connections (1, 2 and 3) and the last layer as illustrated in figure 2 serving as specific objectives. This method is especially useful in situations where computational limitations exist, such as with mobile or embedded devices. The student model that resulted was more accurate than a model trained without knowledge distillation.

The loss function in feature knowledge distillation was usually the average of the mean squared error (MSE) of the feature maps of skip connections and last layer of the teacher and student models for a given input shown in equation 1.

$$L_{MSE} = \frac{1}{N_i} \sum_{i} \frac{1}{N_j} \sum_{j} \left\| f_j^{(t,i)} - f_j^{(s,i)} \right\|^2 \tag{1}$$

where N_i is the total number of images in the batch, N_j is the total number of feature maps in the intermediate layer, i is the index of the examples in the dataset, j is the index of the features in the intermediate layer, $f_j^{(t,i)}$ is the j-th feature map from the teacher model for the i-th example, $f_j^{(s,i)}$ is the j-th feature map from the student model for the i-th example, and $\|.\|^2$ denoted the squared $L2$ norm.

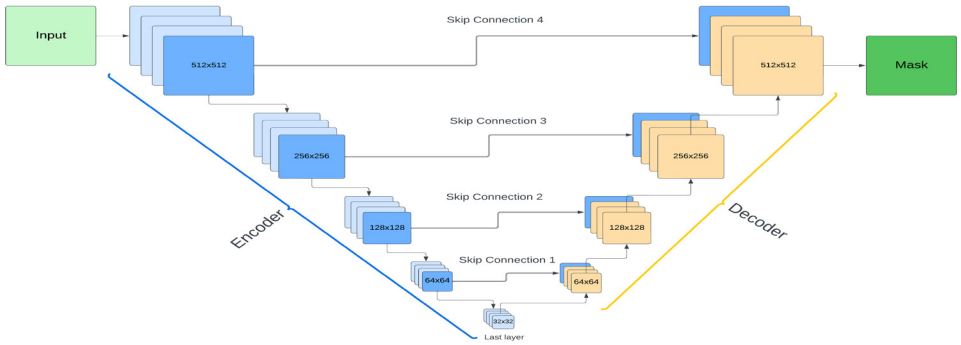

Figure 2: UNet Model

Loss of the skip connection layers and last layer have been averaged to obtain final loss. The purpose of using multiple losses in this way is to ensure that the model learns to reconstruct the output accurately and efficiently across all layers. By averaging the losses, we can get an overall sense of the model's performance, taking into account the various factors that contribute to its effectiveness. This approach can be useful for ensuring that the model is well-rounded and capable of producing high-quality results consistently, independent of the degree of complexity of the input data or the specific features being analysed. By minimising this loss, the student backbone was trained to mimic the behaviour of the teacher backbone and improve its performance on the task of interest.

3.3.2 Knowledge transfer to Decoder

We adhered to the knowledge transferring procedure for the decoder portion illustrated in Figure 2, which involves immediately copying the weights of the decoder parts from the UNet model with the teacher backbone to the UNet model with the student backbone. This procedure is utilised in all UNet models with various backbones.

In order to fine-tune the weights and optimise the total loss, we lastly trained the UNet model using a student backbone.

3.4 Evaluation metrics

Semantic segmentation models' efficacy, in which each pixel in an image is labelled with a class, could also be assessed using the F1 score. The F1 score was determined in this situation at the pixel level as opposed to the image level. Each pixel was handled as a binary classification problem in semantic segmentation in order to calculate the F1 score. The target class (for example, pomegranate fruit) was the positive class, while everything else was the negative class. The precision, recall, and F1 scores were then calculated individually for each class and averaged to provide the final result. For instance, if a semantic segmentation model was trained to distinguish between the classes of object and background, the F1 score (shown in Equation 2)for each class would be calculated.

$$F1score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{2}$$

4 Experimental results

On-field pomegranate fruit segmentation experiments were carried out utilizing the UNet model with various backbones and knowledge distillation. We made use of a dataset of real-world pomegranate fruit images. With a ratio of 80:10:10, the images were divided into training and validation sets, yielding 2400 training images, 300 validation images, and 300 testing images. By employing augmentation techniques, the training set was then enlarged with variance to 9600 images. DGX workstation with four Tesla-V100 NVIDIA GPUs was used for all the experiments. The Keras deep learning package with a TensorFlow backend in an Anaconda environment was used to train and test the UNet model.

Using a variety of backbone architectures, including DenseNet121, EfficientNet, Inception ResNetV2, MobileNetv2, ResNet50, VGG16, and VGG19, we trained and tested the UNet model. For all backbone architectures, the hyperparameters used to train the UNET model remained consistent. The learning rate was kept to 0.0001, the batch size to 8, and the epoch count to 200. Adam was employed as the optimizer, and the loss function was binary cross-entropy. Early stopping was also used to assess validation loss during UNET model training, with a patience of 10 epochs.

The performance of UNet models with various backbones is shown in Table 1. All models were trained and tested using the same set of hyperparameters, and the F1 score, precision, and recall were used to assess each model's performance. According to the data, the VGG19 backbone has the greatest F1 score (90.35%), followed by VGG16 (90.26%). The MobileNetv2 backbone had an F1 score of 86.97%, which was the lowest.

Table 1: Performance of UNet models with different backbones

Backbones	**F1 Score**	**Recall**	**Precision**
DenseNet121	90%	87.14%	93.55%
EfficientNetB0	88.45%	84.03%	93.96%
Inception_ResNetv2	88.46%	84.06%	94.07%
MobileNetv2	86.97%	82.87%	94.50%
ResNet50	89.11%	85.60%	93.47%
VGG16	90.26%	87.35%	93.80%
VGG19	**90.35%**	**88.23%**	93.04%

Since the MobileNetv2 backbone is faster than other backbones, We migrated knowledge from the VGG19 backbone to the MobileNetv2 backbone with the help of feature-based knowledge distillation. The weights from the decoder part of the UNet-VGG19 model were specifically copied to the decoder part of the UNet-MobileNetV2 model, and the distilled UNet-MobileNetV2 model was trained. Table 2 displays the performance outcomes for the distilled model.

Table 2: Performance of UNet models with different backbones

Model	**F1 Score**	**Recall**	**Precision**
UNet-MobileNetv2	86.97%	82.87%	94.50%
Distilled UNet-MobileNetv2	**89.91%**	**86.79%**	93.27%

5 Discussion and Conclusion

Our findings demonstrate the potential of data augmentation strategies for pomegranate fruit segmentation, both in terms of expanding dataset size and enhancing UNet model performance. We observed that the model's performance is significantly influenced by the backbone architecture selected, with VGG19 earning the maximum F1 score of 90.35% as shown in Table 1. As a result, it seems likely that this architecture's feature representation would be more effective for this purpose. It is crucial to keep in mind that these models demand greater computational power and might not be practical in all situations.

Our study, presented in Table 2, demonstrates that knowledge distillation can effectively transfer information from a more computationally expensive model to a smaller and more affordable one, resulting in expanding its potential to obtain superior performance. Specifically, we refined our UNet-MobileNetV2 model and achieved an F1 score of 89.91%, which is comparable to the UNet-VGG19 model but with significantly reduced computing cost. The reason why the UNet-MobileNetV2 model did not exceed the F1 score of the UNet-VGG19 model may be due to the fact that the MobileNetv2 backbone has a different and simpler architecture with 45 times fewer parameters compared to VGG19, which may limit its ability to capture complex features and patterns in images.

Overall, our research highlights the significance of selecting suitable backbone architectures and employing information distillation strategies to enhance the performance of deep learning models for image segmentation tasks. Our findings have practical applications in the agriculture industry and may be applied to other fruit segmentation tasks.

6 Future Work

The future research directions based on the findings include incorporating datasets from diverse locations, as the study used only one dataset from a specific location that may limit the generalizability of the findings to other pomegranate fields in different regions in different daylight conditions. Examining the effects of different loss functions may improve the model's accuracy for unbalanced datasets. This model has the potential to segment pomegranate fruit in real-time and can be deployed on mobile platforms, with significant implications for the agriculture sector.

Acknowledgement

We express our heartfelt gratitude to the farmers in Warkute Village, Baramati, who generously allowed us to gather on-field pomegranate fruit images from their farms, which were instrumental in the success of our research. We would also like to acknowledge the COE-SIP, the Center of Excellence in Signal and Image Processing at COEP Technological University, for providing all the resources necessary to train the models required for this project.

References

[1] P. Kantale, S. Thakare, *A review on pomegranate disease classification using machine learning and image segmentation techniques*, in *2020 4th International conference on intelligent computing and control systems (ICICCS)* (IEEE, 2020), pp. 455–460

[2] M. Fawakherji, A. Youssef, D. Bloisi, A. Pretto, D. Nardi, *Crop and weeds classification for precision agriculture using context-independent pixel-wise segmentation*, in *2019 Third IEEE International Conference on Robotic Computing (IRC)* (IEEE, 2019), pp. 146–152

[3] O. Ronneberger, P. Fischer, T. Brox, *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), pp. 234–241

[4] G. Hinton, O. Vinyals, J. Dean, arXiv preprint arXiv:1503.02531 (2015)

[5] R. Zhang, L. Du, Q. Xiao, J. Liu, *Comparison of backbones for semantic segmentation network*, in *Journal of Physics: Conference Series* (IOP Publishing, 2020), Vol. 1544, p. 012196

[6] W. Gómez-Flores, W.C. de Albuquerque Pereira, Computers in Biology and Medicine **126**, 104036 (2020)

[7] T. Emara, H.E. Abd El Munim, H.M. Abbas, *Liteseg: A novel lightweight convnet for semantic segmentation*, in *2019 Digital Image Computing: Techniques and Applications (DICTA)* (IEEE, 2019), pp. 1–7

[8] A. Romero, N. Ballas, S.E. Kahou, A. Chassang, C. Gatta, Y. Bengio, arXiv preprint arXiv:1412.6550 (2014)

[9] N. Komodakis, S. Zagoruyko, *Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer*, in *ICLR* (2017)

[10] J. Ba, R. Caruana, Advances in neural information processing systems **27** (2014)

[11] G. Urban, K.J. Geras, S.E. Kahou, O. Aslan, S. Wang, R. Caruana, A. Mohamed, M. Philipose, M. Richardson, arXiv preprint arXiv:1603.05691 (2016)

[12] Y. Chen, N. Wang, Z. Zhang, *Darkrank: Accelerating deep metric learning via cross sample similarities transfer*, in *Proceedings of the AAAI conference on artificial intelligence* (2018), Vol. 32

[13] Q. Li, S. Jin, J. Yan, *Mimicking very efficient network for object detection*, in *Proceedings of the ieee conference on computer vision and pattern recognition* (2017), pp. 6356–6364

[14] J. Xie, B. Shuai, J.F. Hu, J. Lin, W.S. Zheng, arXiv preprint arXiv:1810.08476 (2018)

[15] Y. Liu, K. Chen, C. Liu, Z. Qin, Z. Luo, J. Wang, *Structured knowledge distillation for semantic segmentation*, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), pp. 2604–2613

[16] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, *Densely connected convolutional networks*, in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2017), pp. 4700–4708

[17] M. Tan, Q. Le, *Efficientnet: Rethinking model scaling for convolutional neural networks*, in *International conference on machine learning* (PMLR, 2019), pp. 6105–6114

[18] C. Szegedy, S. Ioffe, V. Vanhoucke, A. Alemi, *Inception-v4, inception-resnet and the impact of residual connections on learning*, in *Proceedings of the AAAI conference on artificial intelligence* (2017), Vol. 31

[19] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.C. Chen, *Mobilenetv2: Inverted residuals and linear bottlenecks*, in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2018), pp. 4510–4520

[20] K. He, X. Zhang, S. Ren, J. Sun, *Deep residual learning for image recognition*, in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), pp. 770–778

[21] K. Simonyan, A. Zisserman, arXiv preprint arXiv:1409.1556 (2014)