OBJECT SEGMENTATION

Devikalyan Das
Universität des Saarlandes
66123 Saarbrücken
deda00002@stud.uni-saarland.de

August 23, 2022

ABSTRACT

Image segmentation is about classification of image at the pixel level or in simpler terms, each pixel is assigned a label. This ensures a more deeper understanding of the images and hence has helped in solving many critical computer vision problems. With the advent of Deep Learning, Image Segmentation has been successfully applied in various fields such as medical, autonomous driver-less cars, satellite images with state-of-the-art performance compared to classical image processing based segmentation such as thresholding, edge detection or clustering. Here, in this report I analysed and discussed about some of the deep learning based image segmentation models that I tested over two benchmark datasets for Image Segmentation such as Pascal-VOC and Cityscapes. The experimental results and the performance evaluation of the segmentation models are made to get a better understanding about the topic.

1 Introduction

Image segmentation is the process of dividing an image into different segments according to the labels assigned to each of the pixel values present in the image. It has been employed to solve many Computer Vision problems such as determining the shape of the object, which pixel belongs to which object etc that cannot be solved by techniques like classification of entire image or using bounding boxes [1]. It has wide range of applications in automatic driving system to get understanding of the obstacles on the road, medical domain to get segmented body parts for performing diagnostic tests, satellite image analysis, agricultural engineering and other fields of computer vision. To achieve this earlier researcher used classical image processing techniques such thresholding using OTSU, edge detection etc [2]. The classical techniques though performed the desired task but visual results were not appealing. With the arrival of deep learning the task of segmentation saw substantial rise both in terms of performance and computing resources. This was not only limited to image segmentation, but also tasks like classification,detection and tracking etc. improved. Huge benchmark datasets are now available to test performance of the deeplearning models and provide state-of-the-art results. Models such as AlexNet [3], VGG [4], GoogleNet [5], Residual Net [6], DenseNet [7] etc. have been used for various segmentation and classification tasks. Also, UNet [8] was developed that proved to be breakthrough model for segmentation of medical images that is used quite frequently for any segmentation task across many domain because of its simple yet efficient architecture compared to above models. Majority of the other models are the modifications made on these models to get enhanced result. The major challenge on the path of better performance by these models is the availability of huge annotated datasets and huge computational resources. Here I present some of the Convolutional Neural Network based deep learning models such as U-Net [8], R2U-Net [9] and a proposed deep layer dilated model based on U-Net for different tasks asked in the use case of the NNTI Project. The U-Net model is trained and evaluated on Pascal VOC dataset and the other two on Cityscapes dataset. The two datasets are among the benchmark datasets and have enough classes(21,19) to test the efficiency of a model for segmentation task. The performance of the models have been evaluated based on various metrices usually used for segmentation task such as Accuracy, Dice or F1 [10], IOU [11], AUC [12], Sensitivity [13] and Specificity [13]. The novel contributions in this report for the project is the deep layer dilated model based on U-Net. The performance of this model is compared against R2U-Net. The paper is organized as follows: Section 2 discusses related work and project requirements. Section 3 is about the architectures of the proposed deep layer dilated based U-Net model. Section 4 is about Experimental Setup and Results explaining

everything about the project requirements, datasets, experiments, and results. The conclusion and future work are discussed in Section 5.

2 Related Works

Semantic segmentation has scaled great heights in terms of improvement in results with the use of deep convolutional neural networks (DCNNs) where DCNNs are used for classification of each pixel in the image. In the last few years many deeper network models have been proposed that have proved to enhance the segmentation tasks[16:5-sec paper]. But, the issue of vanishing gradients have hurt training of such deeper network models. To handle this, models came with various approaches like residual connections [17] and using Activation functions such as ReLU [18: 5,6-2 paper]. Various CNN models came such as FCN [19: 2-2 paper], DeepLab [20: 28- 2 paper], VGG [4], Segnet [21 : 5- 2 paper] etc. which improved semantic segmentation performance at various points of time but with enhancements there were also some drawbacks.
One of the popular approaches for semantic segmentation is U-Net [8]. The architecture for U-Net is shown in Fig.1. The network has two parts convolutional encoder and decoder units and they are symmetric to each other. In The encoding units, feature maps are down-sampled using 2 X 2 max-pooling and then the feature maps are up-sampled at the decoding units. The convolutional layers use ReLU as activation function in both the units. It uses Skip connections between the encoding and decoding units. The model is simple and easy to understand. The main advantage of this model is that it produces good performance even with small training samples for segmentation tasks. and provides both location and contextual information. Compared to patch based segmentation, it processes an entire image and produces segmentation masks thereby, preserves the full context of the input images [22: 12,14- 2 paper]. The drawbacks of this model is the lack of sufficient depth in the network which gives average performance against multiple classes. Many variants of the U-Net have come up that uses different connectivity pattern but base architecture is same.
Among variants of U-Net, Alom et al. proposed Recurrent convolutional neural network based on U-net(RU-Net) and Recurrent Residual CNN based on U-Net (R2U-Net) [23: 2 paper]. Both of these architectures use discrete time steps based recurrent convolutional layers. In RU-Net, recurrent convolutions are used before downsampling by maxpooling, before upsampling and before giving out segmented masks. In R2U-Net, the recurrent convolutional layers (RCLs) are stacked with residual connection in both the encoding and decoding units. R2U-Net, as claimed by the authors has addressed two issues of U-Net which are lack of depth and feature summation in Skip connections from one part of the network to other. The depth of the R2U-Net is achieved by the use of residual unit with RCLs which has also improved the feature accumulation from one part of the network to other part. The feature accumulation is happening inside the model compared to U-Net, where it happened outside. This ensured better convergence and extraction of low-level features. Also, in R2U-Net, instead of copy and crop unit as in U-Net, concatenation operation is performed which provided sophistication to the model with improved performance. However, the drawback is that the model parameters and size very large compared to U-Net. Also they don't take into consideration the resolution loss due to subsequent pooling layers which affects in the reconstruction of small objects in the images. The architecture of RU-Net is illustrated in Fig. 2. and the building blocks of the stacked convolutional units of RU-Net and R2U-Net compared with normal convolutional units and residual convolutional units is illustrated in Fig. 3
Alternatively, I have proposed a deep layer dilated convolutional neural network (DDU-Net) based on U-Net, which has addressed some of the issues of both the discussed networks. The detailed architecture of the model is provided in the next section.

3 DDU-Net Architecture

The DDU-Net is inspired from atrous convolution used in DeepLab [24] and U-Net. The atrous convolution helps to control the resolution at which feature maps are computed with-in the convolutional neural network models. It enlarges the receptive fields which will help to integrate larger context at multiple scales, thus generates large scale feature maps with rich spatial information without any significant increase of the number of parameters and amount of computation. In semantic segmentation, both pixel-level precision and multi-scale context information are important for better results. Hence, the dilated convolution play an important role in aggregating multi-scale contextual information without losing resolution. In my model, I have used U-Net as a baseline model with the dilated convolutions with dilation rate of 3 used at deeper downsampled and upsampled layers instead of normal convolutions. The reason to use dilated layers only at the deeper layers is that at the lower level small objects get vanished due to the down-sampling by max-pooling, hence to makes sure that all objects are present and contribute towards the generation segmentation masks. The architecture of my model is illustrated in Fig.4. The size of the input images fed into the model is kept at 256 x 256 and uses four consecutive down-sampling with corresponding up-sampling and skip connections. Instead of using up-sample layer for up-sampling, I have used transposed convolution [25] which ensures an optimal up-sampling

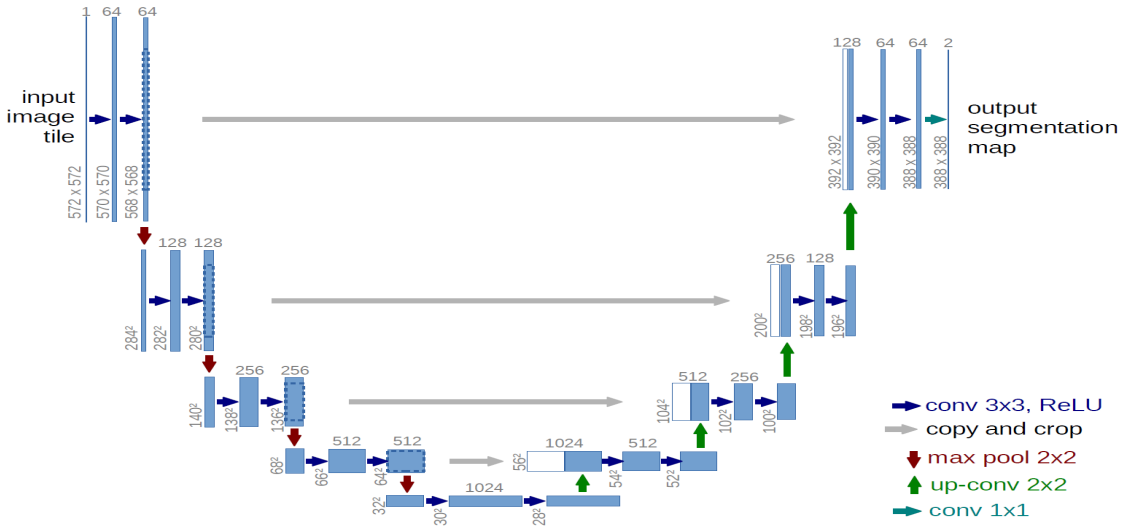


Figure 1: U-Net architecture consisted with convolutional encoding and decoding units that take image as input and produce the segmentation feature maps with respective pixel classes.

through learnable parameters without having to define any predefined interpolation method.

4 Experimental Setup and Results

4.1 Project Requirements and Dataset Summary

Based on the first project requirement, semantic segmentation was needed to be performed on Pascal VOC dataset. Hence, U-Net model was used for this task. The training and evaluation was done Google Colab with single Nvidia Tesla K80 GPU. The results are illustrated in the Colab Notebook.

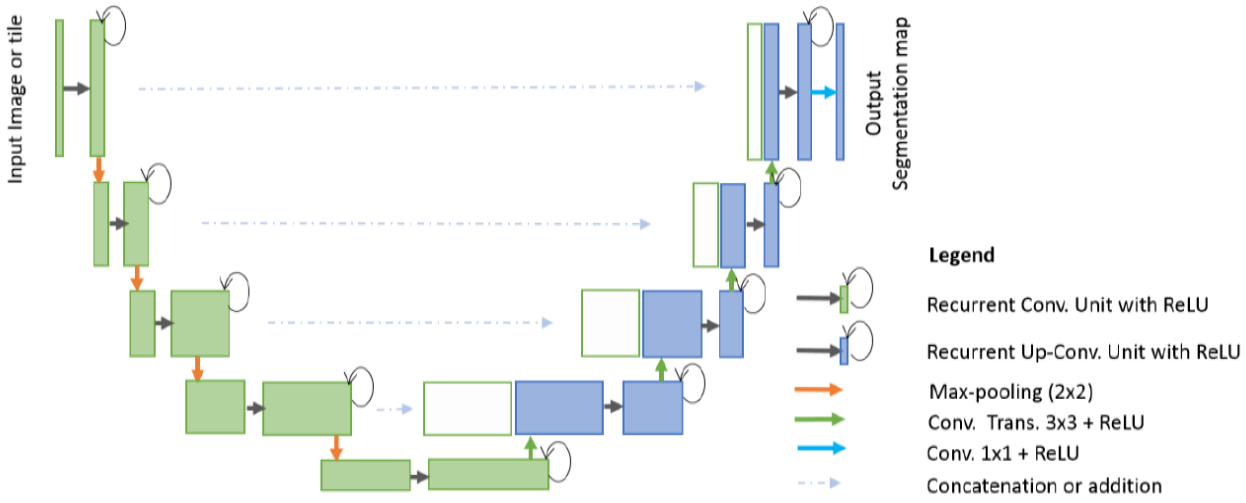


Figure 2: RU-Net architecture with convolutional encoding and decoding units using recurrent convolutional layers (RCL) based U-Net architecture. The residual units are used with RCL for R2U-Net architecture

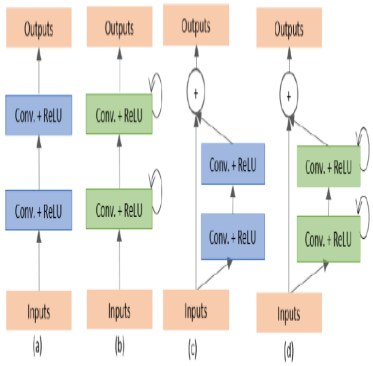


Figure 3: Different variant of convolutional and recurrent convolutional units (a) Forward convolutional units, (b) Recurrent convolutional block (c) Residual convolutional unit, and (d) Recurrent Residual convolutional units (RRCU).

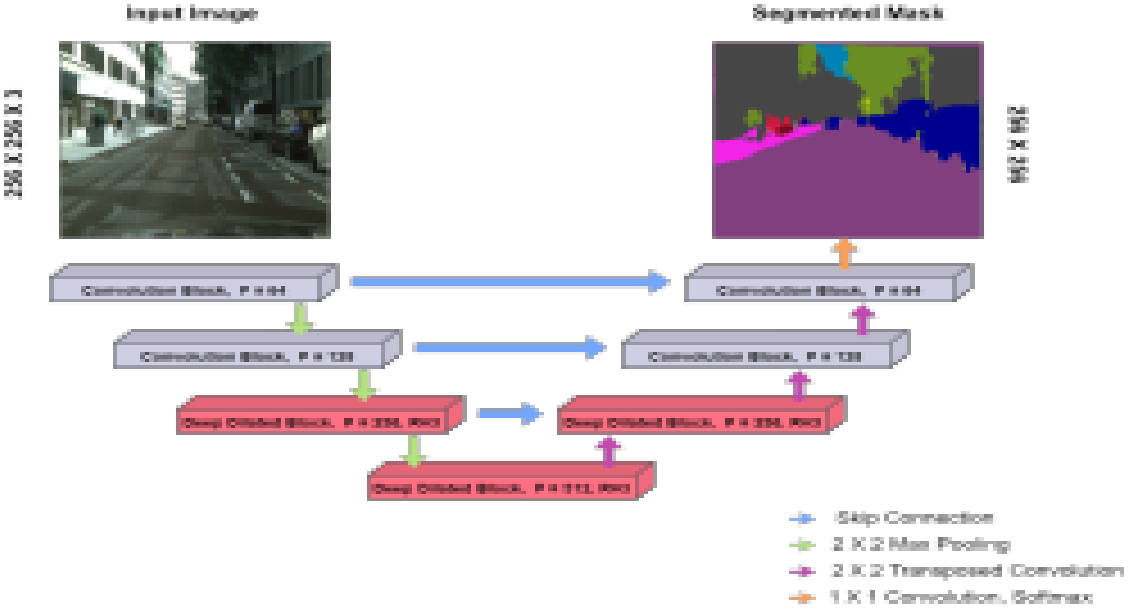


Figure 4: Architecture of the Proposed model DDU-Net

For the second and third requirement, the performance of the deep learning models i.e., R2u-Net on semantic segmentation and a novel model with improved performance compared to R2U-Net was required to be demonstrated, hence DDU-Net was proposed. The models needed to be trained and tested on Cityscapes dataset. The Cityscapes dataset contains 5000 annotated images with fine annotations with 30 classes for over 50 cities from several months (spring, summer, fall), varying weather conditions with large number of dynamic objects, Varying scene layout and Varying background. This report henceforth will discuss about results and implementation for this requirement.

4.2 Implementations

Most of the classes in Cityscapes Dataset out of 30 are void and only 19 classes are available. So, 19 classes were used for training and inference. For training and inference purpose, the dataset was resized to 256 X 256 with a batch size of 8 for the R2U-Net and proposed DDU-Net. The dataset has 2975 training and 500 validation images. The validation dataset was used for testing purpose. For, the implementation Pytorch framework was used on multiple GPU clusters (8 Nvidia Tesla V100 GPUs) depending on availability. The models were trained for 200 epochs and the weights were saved after every epochs. The evaluation was performed after each epoch. The inferencing was done on the validation set as for test set annotation is not available. For the training Cross-Entropy loss and ADAM optimizer was considered. The performance of Loss in R2U-Net and the proposed DDU-Net during training and validation is illustrated in Figure 5.

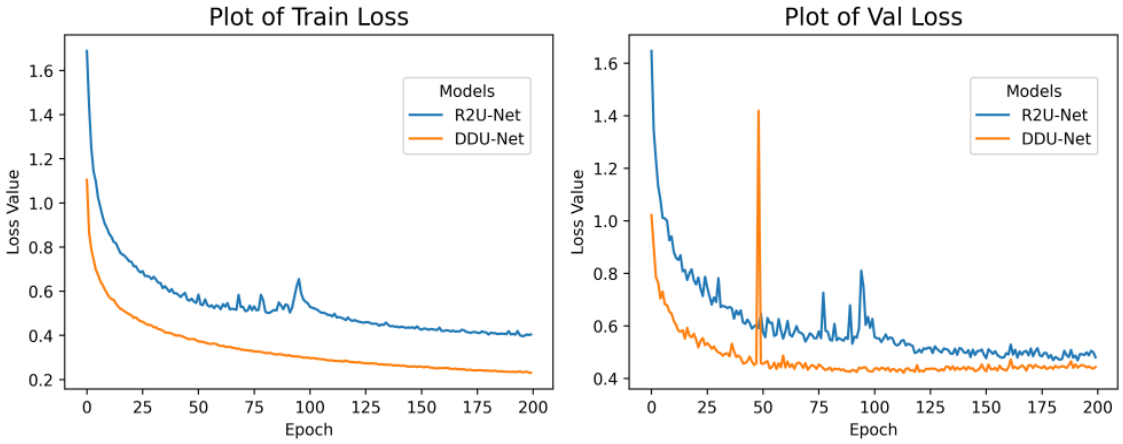


Figure 5: Loss Plot for training and validation

4.3 Quantitative Metrices

For quantitative analysis, various performnce metrices were considered such as accuracy (AC), sensitivity (SE), specificity (SP), F1-score, Dice coefficient (DC), and Jaccard similarity (JS). For this variables such as True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) are required.
The mathematical formulation for the used metrices are given as:

$$AC = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$SE = \frac{TP}{TP + FN} \tag{2}$$

$$SP = \frac{TN}{TN + FP} \tag{3}$$

$$DC = 2\,\frac{|GT \cap SR|}{|GT| + |SR|} \tag{4}$$

$$JS = \frac{|GT \cap SR|}{|GT \cup SR|} \tag{5}$$

For comparison of the metrices results only the micro level scores are considered.

4.4 Results

The training and validation accuracy on the Cityscapes dataset for the R2U-Net and the proposed DDU-Net are illustrated in Figure 6. The figures show that the proposed DDU-Net provide better performance during both training and validation phase when compared to R2U-Net. The curves for sensitivity, specificity, Dice and IOU are illustrated in Figure 7, Figure 8,Figure 9 and Figure 10 respectively. The various metrices plots for both training and validation demonstrates that the proposed DDU-Net is effective for segmentation tasks as compared to R2U-Net. The metrice values used for comparison of both the models are provided in table 1. From the table, the effectiveness of the proposed model can be inferred where it shows significant increment in the metrice values.
The qualitative comparison of the segmented mask for R2U-Net and the proposed DDU-Net along with ground truth labels are shown in Figure 11. From here it can be inferred that the segmented maps of the proposed model has more details about various classes present in the ground truth. Though the proposed model performed bad in some classes but overall it fared better compared to the R2U-Net.
Also a comparison for the numbers of parameters used in each model has been made. This has been depicted in Table 2. It can be clearly inferred that R2U-Net requires a large number of parameters for training. The proposed model requires very less compared parameters as compared to R2U-Net. This also indicates less computing power is needed to use the proposed model.

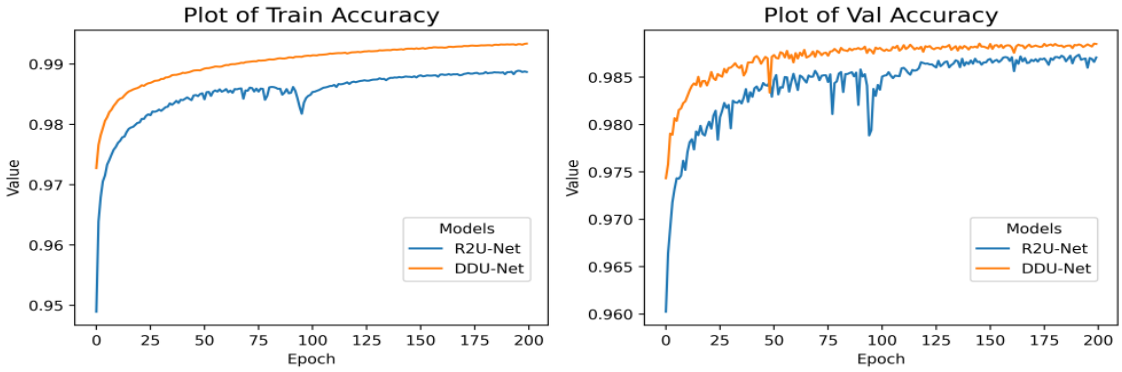


Figure 6: Accuracy Plot for training and validation against R2U-Net

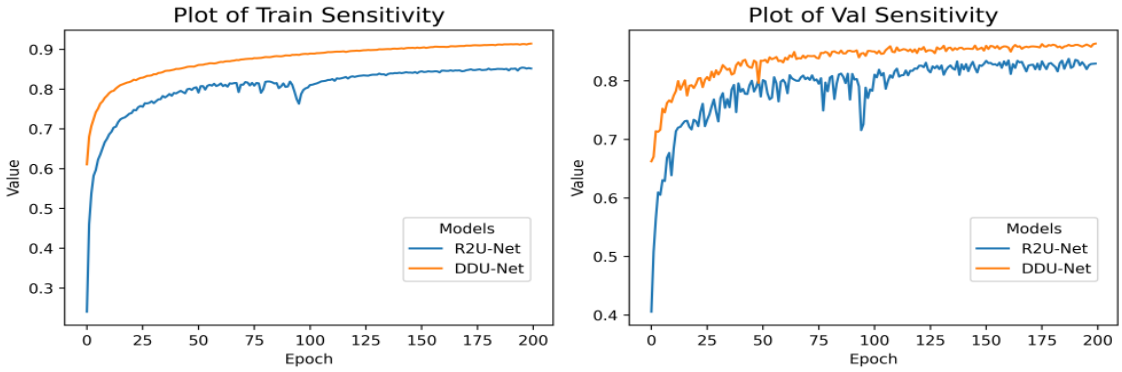


Figure 7: Sensitivity Plot for training and validation against R2U-Net

5 Conclusion and Future work

In this project, an extension of U-Net called DDU-Net was proposed for the purpose of semantic segmentation. The model was evaluated using Cityscapes dataset. The experimental results demonstrate that the proposed DDU-Net show better performance in segmentation tasks with the less number of network parameters when compared to R2U-Net. In addition, results show that the proposed model not only ensure better performance during the training but also in testing phase. Even though the proposed model fared well against its predecessors, there is still room for improvement. Semantic segmentation is still not achieved perfectly. Future works in the direction to have more annotated data for all classes in benchmark datasets are needed in order to overcome the class imbalances. In the direction of model

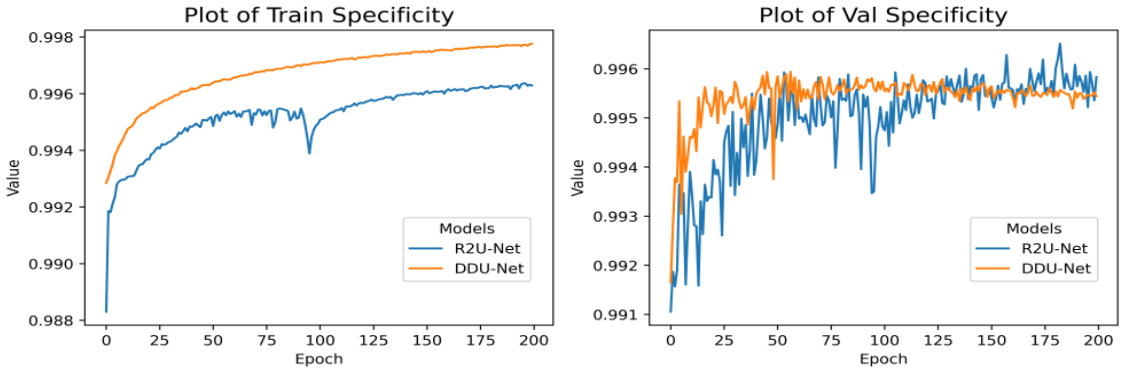


Figure 8: Specificity Plot for training and validation against R2U-Net

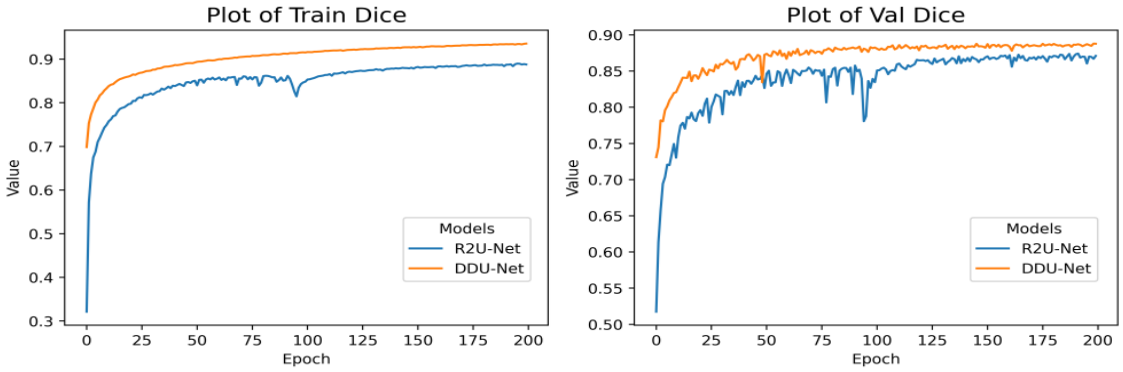


Figure 9: Dice Score Plot for training and validation against R2U-Net

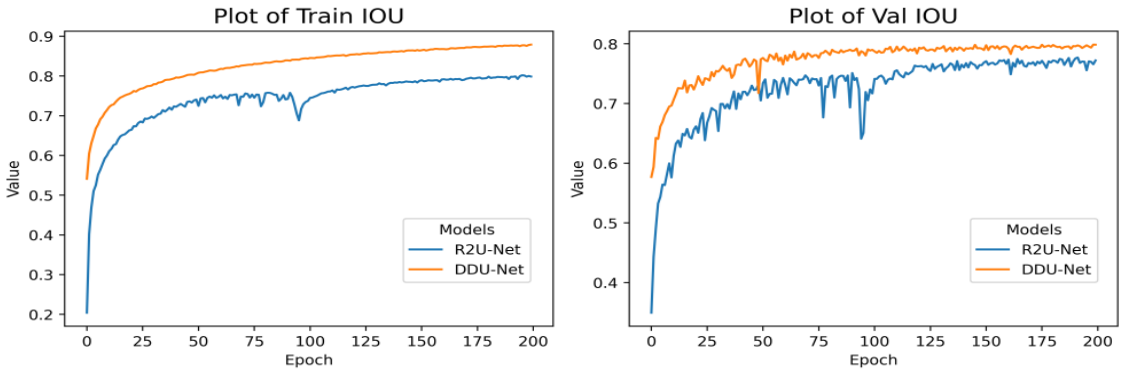


Figure 10: IOU Score Plot for training and validation against R2U-Net

architectures, more novel and deep models architecture can be designed in order to improve the performance of Semantic Segmentation task without increasing computational power.

References

[1] Ayoola Olafenwa. Image segmentation with 5 lines 0f code.

[2] Qiaowei Li, Shuangyuan Yang, and Senxing Zhu. Image segmentation and major approaches. In *2011 IEEE International Conference on Computer Science and Automation Engineering*, volume 2, pages 465–468, 2011.

[3] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[4] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

[5] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Computer Vision and Pattern Recognition (CVPR)*, 2015.

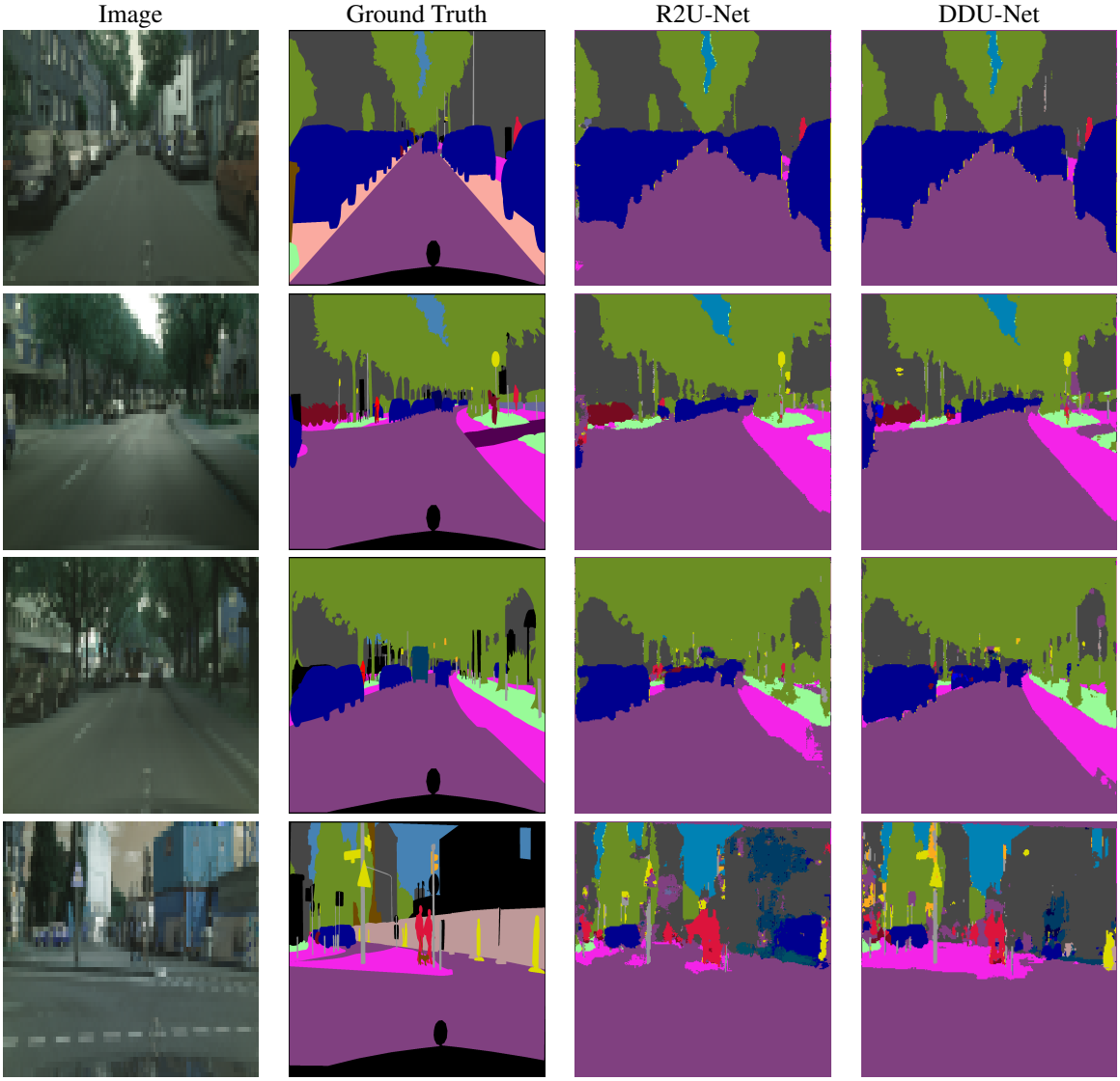
[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[7] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *CVPR*, pages 2261–2269. IEEE Computer Society, 2017.

[8] O. Ronneberger, P.Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 9351 of *LNCS*, pages 234–241. Springer, 2015. (available on arXiv:1505.04597 [cs.CV]).

Table 1: Experimental Results of Proposed approach DDU-Net against R2U-Net

Dataset	Type	Methods	Acc	SE	SP	Dc	IOU
Cityscapes	Train	R2U-Net	0.9887	0.7987	0.8878	0.8518	0.9963
		DDU-Net	**0.9934**	**0.8791**	**0.9356**	**0.9144**	**0.9978**
	Val	R2U-Net	0.9871	0.7722	0.871	0.8296	**0.9958**
		DDU-Net	**0.9885**	**0.7985**	**0.8878**	**0.8637**	0.9954



[9] Md. Zahangir Alom, Mahmudul Hasan, Chris Yakopcic, Tarek M. Taha, and Vijayan K. Asari. Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation. *CoRR*, abs/1802.06955, 2018.

[10] Francis C. Evans. Lee Raymond Dice (1887–1977). *Journal of Mammalogy*, 59(3):635–644, 08 1978.

[11] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union. June 2019.

[12] Andrew P. Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159, 1997.

Table 2: Parameters comparison of Proposed approach DDU-Net against R2U-Net

Methods	Parameters(in Million)
R2U-Net	80.94
DDU-Net	10.81

[13] Dirk Husmeier. Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks. *Bioinformatics*, 19(17):2271–2282, 11 2003.