International Journal of Internet, Broadcasting and Communication Vol.15 No.1 269-274 (2023)
http://dx.doi.org/10.7236/IJIBC.2023.15.1.269

IJIBC 23-1-32

Ensemble UNet 3+ for Medical Image Segmentation

JongJin Park

Professor, Dept. of Computer Engineering, Chungwoon Univ., Incheon, Korea)
jjpark@chungwoon.ac.kr

Abstract

In this paper, we proposed a new UNet 3+ model for medical image segmentation. The proposed ensemble(E) UNet 3+ model consists of UNet 3+s of varying depths into one unified architecture. UNet 3+s of varying depths have same encoder, but have their own decoders. They can bridge semantic gap between encoder and decoder nodes of UNet 3+. Deep supervision was used for learning on a total of 8 nodes of the E-UNet 3+ to improve performance. The proposed E-UNet 3+ model shows better segmentation results than those of the UNet 3+. As a result of the simulation, the E-UNet 3+ model using deep supervision was the best with loss function values of 0.8904 and 0.8562 for training and validation data. For the test data, the UNet 3+ model using deep supervision was the best with a value of 0.7406. Qualitative comparison of the simulation results shows the results of the proposed model are better than those of existing UNet 3+.

Keywords: *Ensemble UNet 3+, Medical image segmentation, Deep learning, UNet 3+, Deep supervision*

1. Introduction

Automatic organ segmentation in medical images is a critical step in many clinical applications. The state-of-the-art models for medical image segmentation are variants of U-Net and fully convolutional networks (FCN). They are networks of encoder-decoder structure which enables taking input of arbitrary size and produce correspondingly-sized output with efficient inference and learning. Their success is largely attributed to their skip connections, which combine deep, semantic, coarse-grained feature maps from the decoder sub-network with shallow, low-level, fine-grained feature maps from the encoder sub-network, and have proven to be effective in recovering fine-grained details of the target objects[1,2]. UNet, which is build upon FCN, proposed so called encoder-decoder architecture which consists of a contracting path to capture context and a symmetric expanding path to enable precise localization[3]. It supplements a usual contracting network by successive layers, where pooling operators are replaced by upsampling operators. Hence, these layers increase the resolution of the output. In order to localize, high resolution features from the contracting path are combined with the upsampled output. A successive convolution layer can then learn to assemble a more precise output based on this information.

Combining multi-scale features is one of important factors for accurate segmentation. UNet++ was

Manuscript Received: February. 2, 2023 / Revised: February. 5, 2023 / Accepted: February. 7, 2023
Corresponding Author: jjpark@chungwoon.ac.kr
Tel: ***-****-****
Professor, Department of Computer Engineering, Chungwoon University, Incheon, Korea

developed as a modified UNet by designing an architecture with nested and dense skip connections[4]. It is essentially a deeply-supervised encoder-decoder network where the encoder and decoder sub-networks are connected through a series of nested, dense skip pathways. The re-designed skip pathways aim at reducing the semantic gap between the feature maps of the encoder and decoder sub-networks. However, it does not explore sufficient information from full scales and there is still a large room for improvement. UNet 3+, which takes advantage of full-scale skip connections and deep supervisions. The full-scale skip connections incorporate low-level details with high-level semantics from feature maps in different scales, while the deep supervision learns hierarchical representations from the full-scale aggregated feature maps[5]. UNet 3+ has better accuracy especially benefiting for organs that appear at varying scales and has fewer network parameters to improve the computation efficiency. It uses hybrid loss function and a classification-guided module to enhance the organ boundary and reduce the over-segmentation in a non-organ image. UNet 3+ contributed in four ways: (i) making full use of the multi-scale features by introducing full-scale skip connections, which incorporate low-level details with high-level semantics from feature maps in full scales, but with fewer parameters; (ii) developing a deep supervision to learn hierarchical representations from the full-scale aggregated feature maps, which optimizes a hybrid loss function to enhance the organ boundary; (iii) proposing a classification-guided module to reduce over-segmentation on none-organ image by jointly training with an image-level classification; (iv) conducting extensive experiments on liver and spleen datasets, where UNet 3+ yields consistent improvements over a number of baselines[5].

In this paper, we proposed Ensemble(E) UNet 3+ for medical image segmentation. It consists of U-Net 3+s of varying depths into one unified architecture. E-UNet 3+ combines the multi-scale features as well as utilizing a full-scale deep supervision like UNet 3+ and bridge semantic gap between encoder and decoder nodes of UNet 3+. Proposed E-UNet 3+ is applied to teeth segmentation and compared with UNet 3+ in performance.

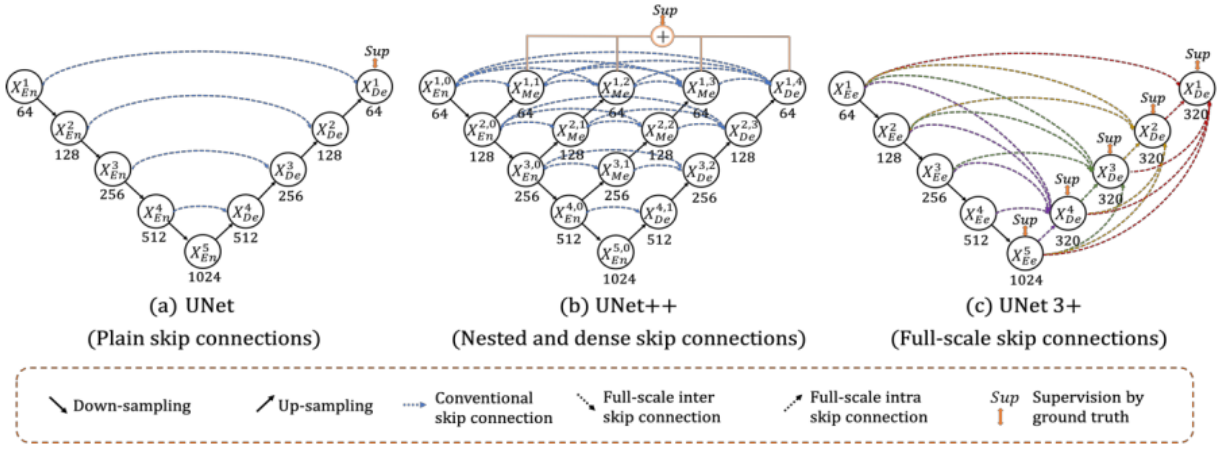


Figure 1. Comparison of UNets

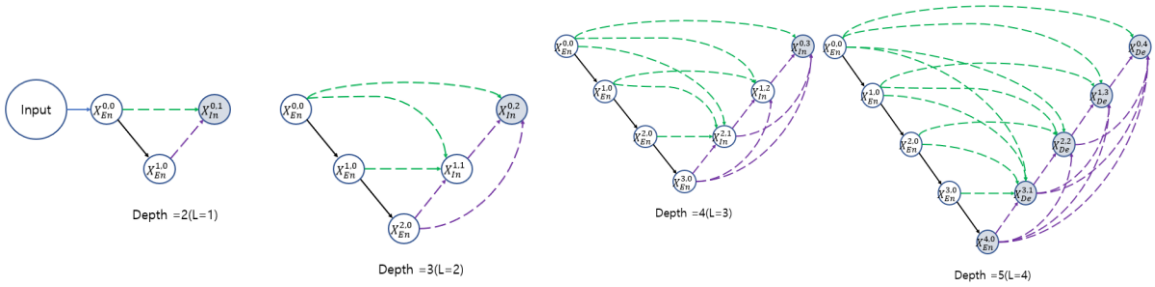
2. UNet 3+ explained

Fig. 1 gives simplified overviews of UNet, UNet++ and UNet 3+. UNet 3+ combines the multi-scale features by re-designing skip connections as well as utilizing a full-scale deep supervision, which provides fewer parameters but yields a more accurate position-aware and boundary-enhanced segmentation map[5]. The full-scale skip connections convert the inter-connection between the encoder and decoder as well as intra-

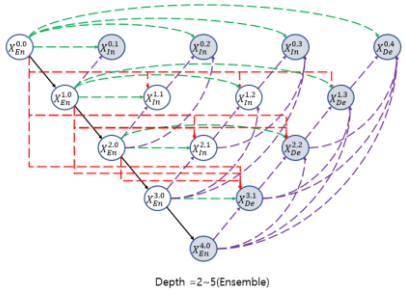
connection between the decoder sub-networks. To remedy the defect in UNet and UNet++, each decoder layer in UNet 3+ incorporates both smaller- and same-scale feature maps from encoder and larger-scale feature maps from decoder, which capturing fine-grained details and coarse-grained semantics in full scales. It is worth mentioning that UNet 3+ is more efficient with fewer parameters[5]. In order to learn hierarchical representations from the full-scale aggregated feature maps, the full-scale deep supervision is further adopted in the UNet 3+. To realize deep supervision, the last layer of each decoder stage is fed into a plain 3 × 3 convolution layer followed by a bilinear up-sampling and a sigmoid function. To further enhance the boundary of organs, we propose a multi-scale structural similarity index (MS-SSIM) loss function to assign higher weights to the fuzzy boundary[6]. To tackle false-positives in a non-organ images, classification-guided module (CGM) is added, which is designed for predicting the input image whether has organ or not.

3. Ensemble UNet 3+

UNet 3+ achieved better performances than UNet and UNet++ in semantic segmentation of the medical images. We think that there is still a room for improvement in the network. Therefor Ensemble UNet 3+ is proposed, which consists of U-Net 3+s of varying depths into one unified architecture. All UNet 3+s (partially) share the same encoder, but have their own decoders. Fig. 2 shows evolution process and structure of E-UNet 3+. All the UNet 3+s of varying depths in Figure 2(a) is ensembled into as E-UNet 3+ as seen in Figure 2(b).



(a) Evolution process of UNet 3+



(b) Structure of E-UNet 3+

Figure 2. Evolution and structure of E-UNet 3+

Each node in the Figure represents a convolution block, downward arrows indicate down-sampling, upward arrows indicate up-sampling, and green and red dot arrows indicate skip connections. Gray filled nodes mean that they are possibly included in DSV(Deep Supervision) to improve performance. They can bridge semantic gap between encoder and decoder nodes of UNet 3+. As a result, each node in the UNet 3+ decoders, from a horizontal perspective, combines multiscale features from its all preceding nodes at the same resolution, and from a vertical perspective, integrates multiscale features across different resolutions from its preceding node.

The E-UNet 3+6 benefits from knowledge sharing, because all UNet 3+s within the ensemble partially share the same encoder.

4. Simulation and Results

Simulation is carried out for teeth segmentation with the proposed E-UNet 3+ and the results are compared with those of UNet 3+ in performance. In this paper, to carry out simulation with the proposed model CBCT data of the teeth of 3 patients are used for training, evaluation and test. The tooth dataset used for learning is CBCT data of two patients (Patients 1 and 2), and one CBCT consists of 280 slice images. The CBCT data of the remaining patient (patient 3) was used as a test for the evaluation of the learned model. Among a total of 560 slice images as data for learning, 448 slices were randomly used for training and the remaining 112 slices were used for validation. 150 slice images of patient 3 were used to test the performance of the proposed model on CBCT data that were not used for learning. Parameters for learning of the models are like these: batch_size is 4, number of epochs, 50, depth of UNet 3+s, 5, initial number of features, 32, and image size is 400 x 400. The initial number of feature maps is 32, and the number of feature maps doubles as the depth of the next level increases. The loss function, which is an evaluation index used to learn the E-UNet3+ model, is a mixture of Focal loss and IoU (Intersection over Union) Loss, which is widely used as an evaluation index in the field of object detection.

The results of simulation for train, validation and test dataset by proposed UNet 3+ model are shown in the table 1. In order to show the performance of the proposed model is better, it was compared with the results of UNet 3+s. As a result of the simulation, the performance of the proposed model with deep supervision was the best for the training, and validation datasets. For both UNet 3+ and E-UNet 3+ models with deep supervision (DSV) performed better than those without deep supervision. The number of parameters to be learned is almost same between the models. That of E-Unet 3+ DSV is bigger than others. For the test dataset the performance of the UNet 3+ DSV is best, but only 0.006 difference with that of E-UNet 3+ DSV.

Table 1. Results of simulation

Model	Train	Validation	Test	Parameters (millions)
UNet 3+	0.8399	0.7973	0.6877	7.6
UNet 3+ DSV	0.8848	0.8415	**0.7406**	7.6
E-UNet 3+	0.8469	0.8203	0.6935	7.6
E-UNet 3+ DSV	**0.8904**	**0.8562**	0.7346	8.0

Figure 3 shows the qualitative comparisons of the teeth image segmentation results of patients 1 and 2 (for learning datasets) by each UNet 3+ model. Figure 4 shows the qualitative comparison of the results of patient 3 (test dataset) teeth image segmentation by each model. In the pictures of each slice, the picture in the first column is the original image, and the picture in the second column is the ground truth. The third column is the result of UNet3+ DSV, and the fourth and fifth columns are the results of the proposed E-UNet (without DSV, with DSV), respectively. The proposed E-UNet 3+ model shows excellent results by extracting the location and boundary of each teeth from CBCT images consisting of multiple teeth. The area marked in green represents TP(True Positive), red represents FP(False Positive), and yellow represents FN(False Negative). As shown in the figures, it is seen that the result by the proposed E-UNet 3+ model is better in the qualitative comparison of the segmentation result image.

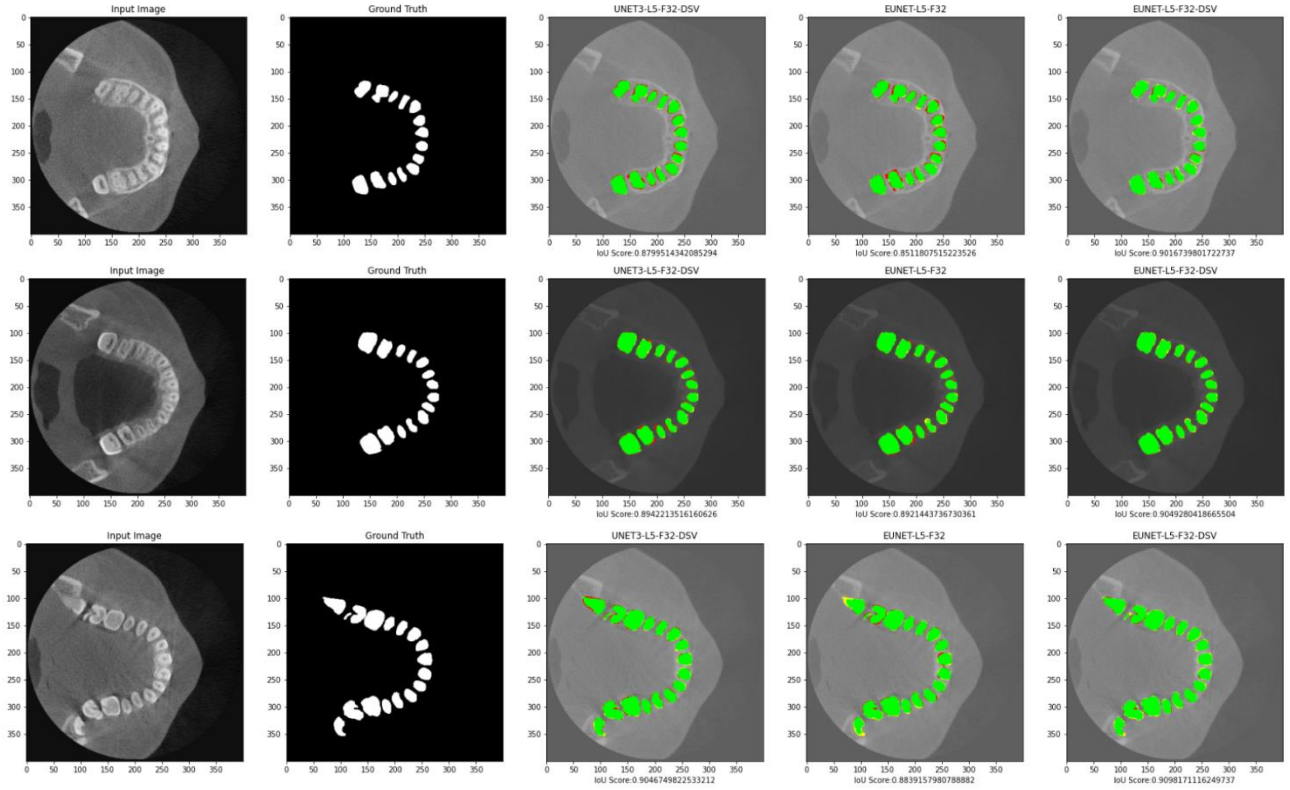


Figure 3. Results of teeth segmentation of patient 1 & 2 by UNet 3+ Models(Training data)

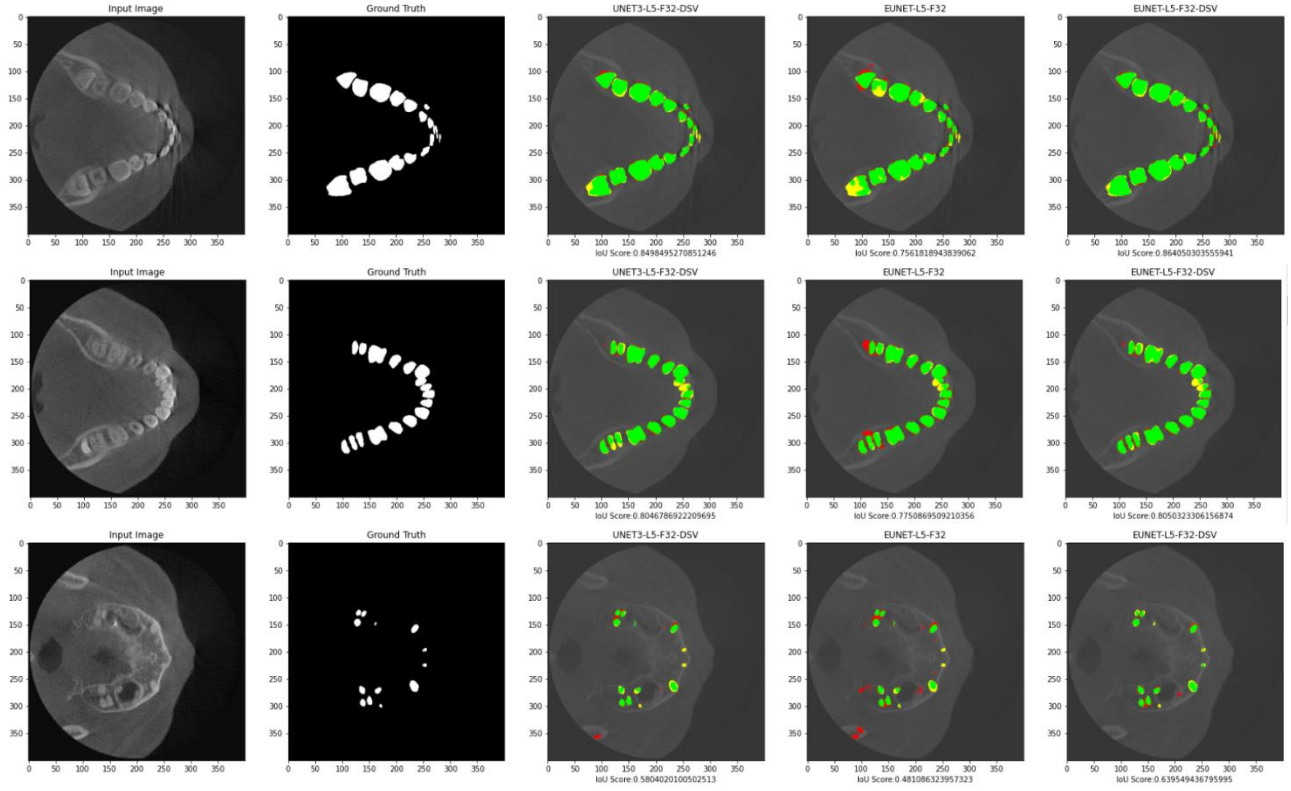


Figure 4. Results of teeth segmentation of patient 3 by UNet 3+ Models(Test data)

5. Conclusions

In this paper, we proposed a new UNet 3+ model for medical image segmentation. The proposed Ensemble(E) UNet 3+ model uses both inter-connection and intra-connection proposed in the existing UNet 3+ model and obtains U-Net 3+s of varying depths to be one unified architecture. UNet 3+s of varying depths have same encoder, but have their own decoders. They can bridge semantic gap between encoder and decoder nodes of UNet 3+. For performance improvement, deep supervision was used for learning on a total of 8 nodes of the E-UNet 3+. The proposed E-UNet 3+ model shows better segmentation results than those of the UNet 3+. As a result of the simulation, the E-UNet 3+ model using deep supervision was the best with loss function values of 0.8904 and 0.8562 for training and validation data. For the test data, the UNet 3+ model using deep supervision was the best with a value of 0.7406. Through qualitative comparison of the simulation results, the excellent result of the proposed model is confirmed.

References

[1] Z. W. Zhou, M. R. Siddiquee, N. Tajbakhsh, J. Liang, "UNet++: Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation", IEEE Trans Med Imaging, 2020 June; 39(6): 1856–1867. DOI:https://doi.org/10.1109/TMI.2019.2959609

[2] J. Long, E. Shelhamer and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," The IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431-3440, 2015. DOI:https://doi.org/10.48550/arXiv.1411.4038

[3] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," International Conference on Medical image computing and computer-assisted intervention, pp. 234–241. Springer, 2015. DOI-https://doi.org/10.48550/arXiv.1505.04597

[4] Z.W. Zhou, M.M.R. Siddiquee, N. Tajbakhsh and J.M. Liang, "UNet++: A Nested U-Net Architecture for Medical Image Segmentation," Deep Learning in Medical Image Anylysis and Multimodal Learning for Clinical Decision Support, pp: 3-11, 2018. DOI-https://doi.org/10.48550/arXiv.1807.10165

[5] Huimin Huang, et al., "Unet 3+: A full-scale connected unet for medical image segmentation", 2020 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP), pp. 1055–1059. IEEE, 2020. DOI- https://doi.org/10.48550/arXiv.2004.08790

[6] Z. Wang, E.P. Simoncelli and A.C. Bovik, "Multiscale structural similarity for image quality assessment," The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003. DOI- https://doi.org/10.1109/ACSSC.2003.1292216

[7] O.O et al., "Attention u-net: Learning where to look for the pancreas," Medical Imaging with Deep Learning, 2018. DOI- https://doi.org/10.48550/arXiv.1804.03999

[8] T.-Y. Lin, P. Goyal, R. Girshick, K.M. He and P. Dollar. "Focal loss for dense object detection," The IEEE international conference on computer vision, pp. 2980-2988, 2017. DOI- https://doi.org/10.48550/arXiv.1708.02002