# U-Net School for Medical Image Segmentation

Quan Liu, Tianyao Shi, Yiyang Chen
Department of Computer Science - Shanghai Jiao Tong University
517021910675, 517021910623, 517021910585

***Abstract*—Medical image segmentation has always been a difficult task in computer vision, due to lack of data, hish demand on image details and huge size of input images. In this project, we adopt 3 different networks from the U-Net school to solve this problem: U-Net, U-Net++ and U-Net+++. Several data augmentation methods including rotation, shifting and clipping are used, to boost the performance of model. We conduct extensive experiment to select the optimal binarization threshold and other hyperparameters, and finally compare the pros and cons of the 3 different structures.**

***Index Terms*—U-Net, Medical Image Segmentation, Data Augmentation**

## I. Main Ideas

Image classification challenge has been one of the most-discussed problems in the past decades. However, better and better network structures have been found, whose accuracy on natural image classification has risen to an extent of general satisfactory among almost all datasets, except for medical segmentation tasks. Medical image segmentation is still challenging because of its importance as a nature and its commitment for finding tiny structures which could affect diagnosis results.

Some Challenges have been asking for better methods for medical image segmentation, like ISBI. During the search for domain-specific networks, we noted that there are three networks in a series all belonging to the U-Net school, which, seen in a row, can uncover the progress the field very well. These methods are: U-Net, UNet++ and UNet+++.

We did experiments[1] on these methods with the provided ISBI dataset, compared the procedure of leveraging all these methods and listed the performances of all. Although all three methods are all implemented on different machines with different calculation capacities, the experiments could still uncover some interesting traits among all these methods. The results of all methods show that UNet+++ performs best, with an accuracy of 0.923, with U-Net following it at 0.919, and UNet++ ranking the worst at 0.917. We also did more comparison in the final section.

## II. Methods

### A. U-Net

#### a.1) Motivation

Here we explain why we choose U-Net as our first and baseline method to experiment with. Published in [1], it has been cited over 10000 times and widely used as a benchmark in medical image segmentation. Based on U-Net, many vairations are designed to pursue better performance, as we will see in later parts. Nevertheless, U-Net itself was a big breakthrough at its publish time, and it is very beneficial to study with its network structure.

#### a.2) Detailed Description

Image segmentation has long been a major task in computer vision, where the input is an image and the expected output is the mask of image—where all pixels belonging to the same object are labeled out (See Figure 1).

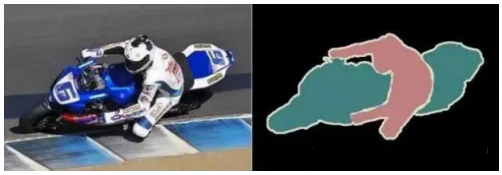

Fig. 1. An illustration of the goal of iamge segmentation.

The first try to combine deep learning methods with this task is Fully Convolutional Networks by [2], which became a milestone. We know that in traditional CNN, the feature dimension is decreased by convolution and pooling operations, and the reception region change from local to global gradually as the image propagate through the network. This makes it difficult to do segmentation, as we have to restore the size of image. FCN uses reversed convolution and upsampling operation to restore image size, and conduct pixel-level classification.

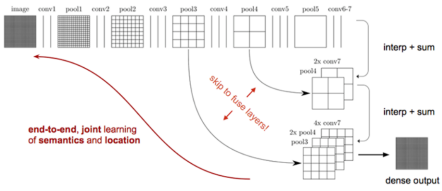


Fig. 2. Structure of Fully Convolutional Network, where it fuses pool4, pool3 and feature map to concate features.

However, FCN is not good at image details (see Figure 3). The result is often blurred and smooth, thus not suitable for medical image segmentation, where we especailly care about edges and details in the image, in case the doctors make wrong diagnosis. So there came U-Net, which imporve based on FCN.

Figure 4 shows the structure of U-Net. It uses many feature channels to allow features containing more information on the texture of original image to propagate between high-resolution layers. And it is specaily designed for medical tasks, where it is very difficult to separate the connected same-category cells. The authors proposed weighted loss function, where it gives the ground truth of connected cells more consideration. U-Net

[1]The code of this project can be found in https://github.com/cyy0915/MLCode

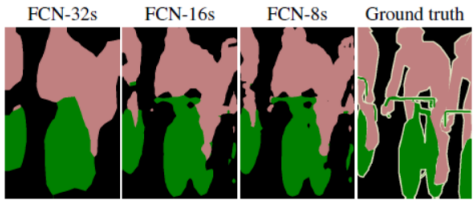


Fig. 3. Some results of FCN compared with the ground truth. Details of the cycling man isn't very good.

can perform well on small datasets like ISBI challenge, and does not require demanding GPU memory. Of course, data augmentation needs to be done before training.

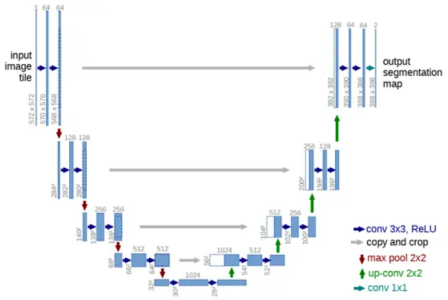


Fig. 4. Sructure of U-Net. It looks like the character 'U'.

## B. UNet++

### b.1) Motivation

It's clear that methods before UNet++[3] can solve general image segmentation problem with decent accuracy, especially when segmenting natural images. However, medical image segmentation require more accuracy in the detail of the segmentation, for example, tumors with ragged edges and more blood vessels on the edge are more likely to be malignant. This puts a higher target for neural networks dedicated to solving medical image segmentation problem, calling for higher-accuracy methods. Also, it's hard for U-Net users to prune U-Net architecture to gain optimal inference time.
As a result, two propositions have been made:

- Better networks are needed to solve medical image segmentation problems;
- Networks need to be more friendly to production situations where recognition speed is vital.

In order to meet these requirements, UNet++ was designed above U-Net[1] with skip connections replaced with dense convolution blocks. We will discuss detailed designs in the next section.

### b.2) Detailed Description

The biggest modification from U-Net into UNet++ was its skip connections. Where were direct skip connections in U-Net was replaced with an array of dense convolutional layers, reforming the network shape as shown in Figure.5.

The basic idea of UNet++ was that, with more dense convolution layers in between, the encoder activation domain would become closer to the decoder network activation domain, thus making finding exact segmentation a easier job. From another perspective, the network structure of UNet++ can

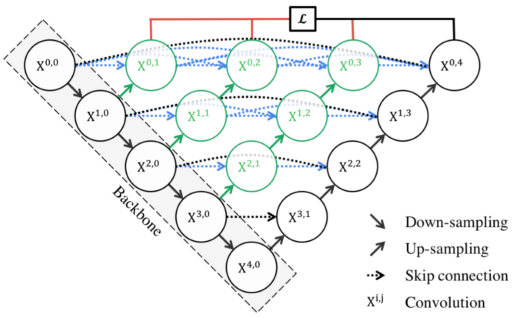


Fig. 5. Structure of Unet++, where black nodes represent basic U-Net, green nodes are dense convolution layers, blue and green arrows are dense connections and up-sampling, and red connections represent deep supervision.

be separated into several independent networks, as shown in Figure6, there was 4 of them, each having more layers than the last, inheriting the last one's encoder feature map, and adding another layer beyond that. Applying loss at all 4 output positions for the full-fledged network leverages the popular deep supervision method, not only helping the network to converge better, but also enabling users to prune the network during inference time as users can only use the first several layers of output to determine the result. Consequently, UNet++ with its special dense convolution connections could deal with the two requirements that we mentioned at the same time, thus being more strong and efficiency-aware than bare U-Net.

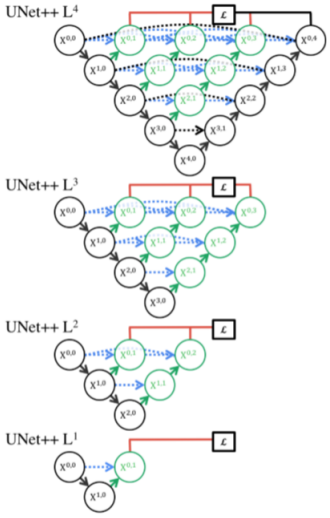


Fig. 6. Pruning structure of UNet++.

## C. UNet+++

### c.1) Motivation

After UNet and UNet++, we also try UNet+++[4], which was developed and published in 2020 and maybe the latest UNet-like deep learning network. In the following sections, we first briefly introduce UNet+++ according to the paper, and then introduce our training methods and results.

### c.2) Network description

In image segmentation, Combining multi-scale features is one of important factors for accurate segmentation. In recent

deep learning network like UNet and UNet++, feature maps in different scale explore distinctive information. Lowlevel detailed feature maps capture rich spatial information, while high-level semantic feature maps embody position information. Nevertheless, these exquisite signals may be gradually diluted when progressively down- and up-sampling. However, by implementing full-scale skip connections, UNet+++ can make full use of the multi-scale features, incorporating low-level details with high-level semantics from feature maps in different scales. Figure./reffig:unetpppGraph shows simplified overviews of UNet+++. The image is from [4].

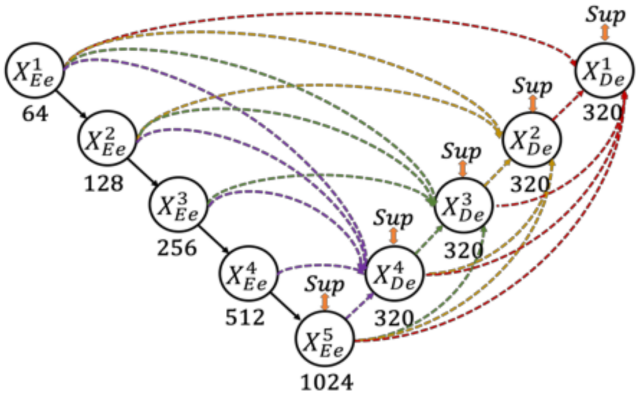


Fig. 7. UNet+++ network structure

## III. Algorithm

### A. U-Net

There are mainly 3 tricks to train a U-Net, namely overlap-tile strategy, data augmentation and weighted loss.

#### *a.1) Overlap-tile strategy*

This strategy allows the seamless segmentation of arbitrarily large images. See Figure 8 To predict the pixels in the border region of the image, the missing context is extrapolated by mirroring the input image. This tiling strategy is important to apply the network to large images, since otherwise the resolution would be limited by the GPU memory.

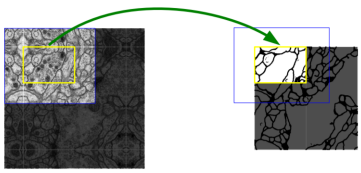

Fig. 8. Overlap-tile strategy for seamless segmentation of arbitrary large images. Prediction of the segmentation in the yellow area, requires image data within the blue area as input. Missing input data is extrapolated by mirroring

#### *a.2) Data augmentation*

As for our tasks there is very little training data available, we use excessive data augmentation by applying elastic deformations to the available training images. This allows the network to learn invariance to such deformations, without the need to see these transformations in the annotated image corpus. This is particularly important in biomedical segmentation, since deformation used to be the most common variation in tissue and realistic deformations can be simulated efficiently.

In case of microscopical images we primarily need shift and rotation invariance as well as robustness to deformations and gray value variations. Especially random elastic deformations of the training samples seem to be the key concept to train a segmentation network with very few annotated images. We generate smooth deformations using random displacement vectors on a coarse 3 by 3 grid. The displacements are sampled from a Gaussian distribution with 10 pixels standard deviation. Per-pixel displacements are then computed using bicubic interpolation. Drop-out layers at the end of the contracting path perform further implicit data augmentation.

#### *a.3) Weighted loss*

We pre-compute the weight map for each ground truth segmentation to compensate the different frequency of pixels from a certain class in the training data set, and to force the network to learn the small separation borders that we introduce between touching cells.

The separation border is computed using morphological operations. The weight map is then computed as

$$w(\mathbf{x}) = w_c(\mathbf{x}) + w_0 \cdot \exp\left(-\frac{(d_1(\mathbf{x}) - d_2(\mathbf{x}))^2}{2\sigma^2}\right)$$

where $w_c : \Omega \rightarrow R$ is the weight map to balance the class frequencies, $d_1 : \Omega \rightarrow R$ denotes the distance to the border of the nearest cell and $d_2 : \Omega \rightarrow R$ the distance to the border of the second nearest cell.

### B. UNet++

There are several unique algorithms and tricks used in UNet++, including deep supervision and its unique loss function. We will discuss all these below in detail.

#### *b.1) Deep supervision*

We have mentioned deep supervision in the last part about the description of UNet++, here we will discuss it in detail. Let's reflect on Figure.5, where all the outputs of the four branches connect to a single loss function. It's worth noticing that, in normal deep supervision procedure, losses at supervision layers decay with time, while in this setup, no loss would decay through time. This was because of the proposition that UNet++ was designed to support pruning during inference time, which can be done through only taking the output at layer $X^{0,t}, t \in \{1, 2, 3, 4\}$. As a result, the deep supervision in UNet++ is not in its standard form, whose effects will be discussed in the experiment section.

#### *b.2) Loss Function of UNet++*

In the paper[3], the loss function at arbitrary output point has a uniform formula, which consists of a cross-entropy term and a Dice-coefficient term:

$$L(Y, \hat{Y}) = -\frac{1}{N}\sum_{b=1}^{N}\left(\frac{1}{2} * Y_b * log\hat{Y}_b + \frac{2 * Y_b * \hat{Y}_b}{Y_b + \hat{Y}_b}\right) \quad (1)$$

Where $Y_b$ and $\hat{Y}_b$ denote the flattened probability vector output for the $b^{th}$ image.

Totalling all these things together, we can see that UNet++ has 4 streams of constant gradient flow feeding back into the network during training time.

### C. UNet+++

We use Pytorch to train UNet+++, so we use Pytorch's loss function and optimizer.

In details, because the result picture only have two class 0,1 per pixel, we use loss function BCEWithLogitsLoss, it combines a Sigmoid layer and the BCELoss. BCELoss is a criterion that measures the Binary Cross Entropy between the target and the output, the formula is:

$$\mathbf{l}(\mathbf{x}, \mathbf{y}) = \{l_1, \cdots, l_N\}^T,$$
$$l_n = -w_n[y_n \cdot \log x_n + (1 - y_n) \cdot \log(1 - x_n)]$$

Where N is the batch size. BCEWithLogitsLoss is more numerically stable than using a plain Sigmoid followed by a BCELoss as, by combining the operations into one layer, we take advantage of the log-sum-exp trick for numerical stability.

We use Adam algorithm as optimizer. Actually we test Adam, RMSprop, SGD and we find that Adam performs best.

## IV. Experimental Settings

### A. U-Net

Here we briefly descirbe the environment and tools we use to train a U-Net ourservles. We use TensorFlow with Keras API, and the training is done on a NVIDIA GTX 1060 (6GB). We use standard Adam optimizer with learning rate 1e-4, and train the net for 10 epochs for 5 hours, each epoch containing 2000 steps. The data augmentation is done via Keras ImageGenerator API, with rotation, shift and horizontal flip.

### B. UNet++

We will discuss the settings of UNet++ throughout our experiments in this section. The network structure was the same as depicted in Figure.5. Due to the limited capacity of our hardware, the training time batch size was very limited, which is actually only 1 picture per batch. We understand the effect that small batch size hinds convergence, but we had no choice. The optimizer was SGD, with a learning rate of 1e-3, momentum of 0.9 and weight decay of 1e-4. We were only able to train the network for 100 epochs due to the limited memory.

### C. UNet+++

We use the given dataset to train and test the network, the learning rate is the default value(0.001) of torch.optim.Adam. Because the network is complex, we can only set batch size to 1, otherwise the GPU memory is not enough.

## V. Result

### A. U-Net

To make sure we are on the right way, we just train for 5 epochs before we determine the optimal training and testing settings. We run into a few problems when we start. At first we find the original input size of U-Net is 256 square while the training data size is 512 square. Though the training is super fast and the validation accuracy seems pretty good (as the validation set is also from resized input), when we output the predicted image, use the provided code to test accuracy, the result can be as low as 60 percent. This is beacause that the provided code compare two images pixel-wise, and different dimension difinitely causes big problem.

We try different methods to fix it: (1)resize the predicted image to 512 square; (2) resize the label to 256 square; (3) change the network structure to allow input=512, and retrain the model. TableI shows the result.

TABLE I
3 WAYS TO FIX THE SIZE ISSUE

| Method | resize predtion | resize label | change network |
|---|---|---|---|
| Accuracy | 0.6754 | 0.7545 | 0.6996 |

Seems working, but still not so good. And here comes new question: why does resizing label perform better? Theoretically, resizing leads to information loss. To answers these questions, we need to take a closer look at the data. The labels depict the edges of cell stuctures using pure black, and on any other parts of the images, the pixels are pure white. White make up most of the image, so even if you always predict Positive (i.e. no edge, white on image), you get around 60 accuracy. So after shrinking label, it is very possible that the area coverage of white pixels increase, such that the performance is even better than the retrained model.

But what about the 70 percent accuarcy? The validation information during training process says the accuracy is as high as 97 percent? The answer lies in the following two images, on the left is the predicted image and on the right the label, as is shown in Figure 9.

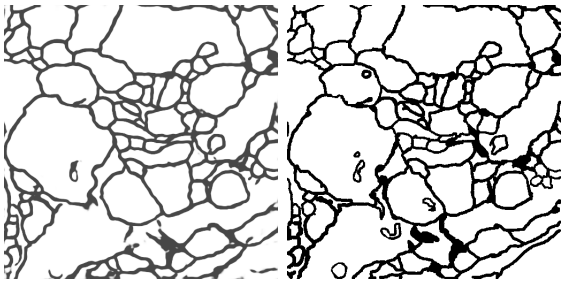

Fig. 9. Please look carefully at the two images. they are pretty the same, but wait, isn't the black in the left image not so black?

Hey! The network is not 100 percent sure about the prediction, so the output is possibility, which after converting to 0-255 uint8 value, is not 0 (pure black) nor 255 (pure white), but something in between. And the evaluation code requires that each pixel value to be the same as the label, to be count as a TP or TN. So if we want to use this code, we must conduct binarization on predictions before we compare pixels.

Then we conduct experiment to select the optimal threshold of binarization, whose result in this case, is 87. See Figure 10 for details.

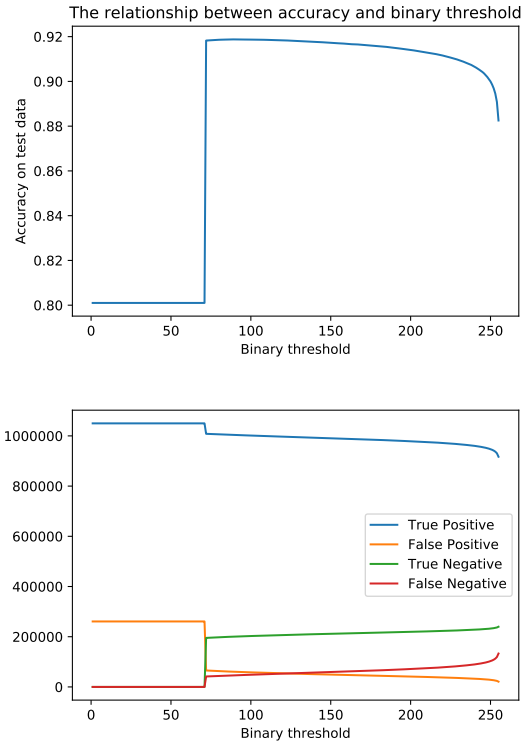


Fig. 10. Selecting optimal binarization threshold based on retrained model.

Then using this threshold, we trained for more 5 epochs and Figure 11 shows the performance and training epochs. The best accuracy is 0.9189. It seems not improving much while the epochs increase, this may be that there is too many steps (2000) in an epoch, and the work is pretty much done in the first epoch.

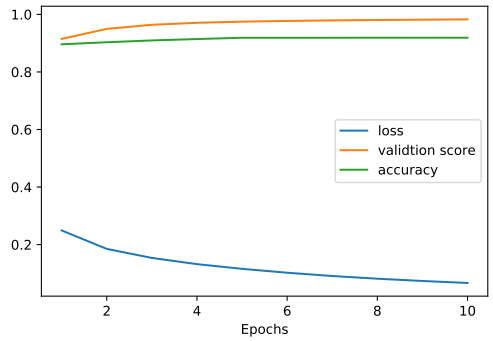


Fig. 11. Training history of U-Net.

### B. UNet++

We trained UNet++ with the designated parameters, along with deep supervision and without supervision, which results in the accuracies in Table.II.

As shown in Table.II, deep supervised network performance increases with output layer, while still performing worse than

TABLE II
UNET++ RESULTS

| | Out 1 | Out 2 | Out 3 | Out 4 |
|---|---|---|---|---|
| with deep supervision | 0.878 | 0.886 | 0.891 | 0.894 |
| without deep supervision | / | / | / | 0.917 |

the network without deep supervision, where in the UNet++ paper[3], the gap was not that big, and deep-supervised network perform generally better than normal ones. We suppose that was caused by our non-ideal training parameter setting. We tend to think that deep-supervised networks actually converge slower in this setting because of two reasons:

- Non-decaying deep supervision force the network to figure out a good approximation right at the beginning of the network, which may over-exploit the potential of the first layer, leaving less work to do for the following layers, and making the network less competent as a whole. We argue that this setup actually slows down the training process, which will be shown in the next experiment.
- Small training epochs lead to early exit and non-optimal performance during training.

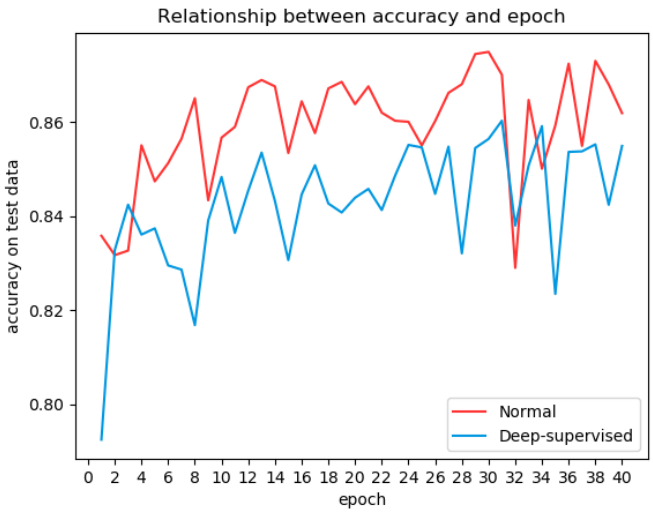


Fig. 12. UNet++ training accuracy-epoch curve.

As shown in Figure.12, training accuracy of normal UNet++ are generally higher than deep-supervised UNet++, which is consistent with our former proposition. We can observe that the deep-supervised network, trained upon our poor laptop GPU with few epochs, could reach no better result than normal UNet++. We can conclude that deep supervision really slows down the training process. Although it may perform better in the end, it may take a much longer training time, which is still a problem for those who want to save the training calculation time.

### C. UNet+++

We test the relationship between epoch and the accuracy of test data, in order to find the best epoch value. In the test, the value of epoch range from 1 to 40. In each iteration, we test the current model on test dataset, and record the accuracy. The result is given Figure.13.

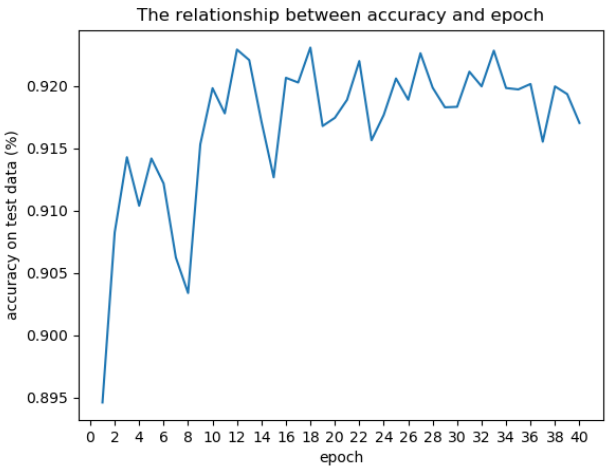


Fig. 13. UNet+++ training accuracy-epoch curve

The curve shows that as the epoch value increase, the accuracy first increase rapidly and than decrease slowly, which means that when epoch is too large, it becomes overfitting. The best accuracy is 0.923, with the epoch value 18

## VI. Conclusion

In the project, we study the deep learning network UNet, UNet++ and UNet+++, and implement all three network on the given dataset. The experiments on the three network show lots of information. First, all experiments get relatively high accuracy on test data, which means that UNet and its variants work well in image segmentation. Second, these network can converge quickly. In the training process, Although the given training dataset is not big(25 images), the accuracy reaches highest when iterate for about 20 times, a relatively small value. Third, the modification for Unet in other two networks improve the network performance, especially UNet+++. The best accuracy of UNet and UNet++ is approximately 0.918, and that of UNet+++ can be 0.923, a impovement of 0.5%.

In conclusion, we learn a lot from the project. We get a better understand of image segmentation and deep learning. However, Because of time limit and device limit, our experiments lack further optimization on some parameter, so the accuracy in some experiments above can be higher. We may refine it in the futrue.

## Referências

[1] Olaf Ronneberger, Philipp Fischer e Thomas Brox. “U-Net: Convolutional Networks for Biomedical Image Segmentation”. Em: Cham: Springer International Publishing, 2015, pp. 234–241. ISBN: 978-3-319-24574-4.

[2] Jonathan Long, Evan Shelhamer e Trevor Darrell. “Fully convolutional networks for semantic segmentation”. Em: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. IEEE Computer Society, 2015, pp. 3431–3440. DOI: 10.1109/CVPR.2015.7298965. URL: https://doi.org/10.1109/CVPR.2015.7298965.

[3] Zhou. Zongwei e Rahman. Siddiquee. Md. Mahfuzur. “UNet++: A Nested U-Net Architecture for Medical Image Segmentation”. Em: Cham: Springer International Publishing, 2018, pp. 3–11. ISBN: 978-3-030-00889-5.

[4] Huimin Huang, Lanfen Lin, Ruofeng Tong, Hongjie Hu, Qiaowei Zhang, Yutaro Iwamoto, Xianhua Han, Yen-Wei Chen e Jian Wu. “UNet 3+: A Full-Scale Connected UNet for Medical Image Segmentation”. Em: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2020, pp. 1055–1059.