A Comparative Analysis on Architectures of UNet Family for Image Segmentation

Yueru Gong

Khoury College of Computer Sciences, Northeastern University, Boston, MA, 02120, United State

gong.yuer@northeastern.edu

Abstract. Image segmentation is an important visual task that involves categorizing each pixel within an image, assigning it a specific label that defines its content. The advancement of deep learning has revolutionized this field, with Convolutional Neural Networks (CNNs) marking a significant improvement in accuracy and efficiency. These models learn from the raw image data to the segmentation labels in an end-to-end manner, enabling them to produce precise segmentation outcomes with minimal human intervention. Among the spectrum of segmentation models, the UNet architecture has emerged as a standout performer, renowned for its robust performance and adaptability. It has not only set a high benchmark but also inspired a plethora of enhancements and adaptations to cater to various applications. This paper presents a systematic analysis of UNet, hoping to share the UNet model with its extensions. So that beginners can have a preliminary understanding of this famous family of deep-learning network models. In this paper, the author first introduces the basic UNet model. Then introduce eight classic UNet variants. This paper will guide the reader to understand the compilation and improvement principles of the UNet family and provide ideas and a foundation for the reader's subsequent research on UNet.

Keywords: Image segmentation, deep learning, UNet.

1. Introduction

The UNet model and UNet family are undoubtedly one of the most successful convolutional neural networks (CNNs) for image segmentation tasks, especially for medical image segmentation. The subsequent years of development have also confirmed it as an all-rounder in semantic segmentation tasks. UNet's lightweight network architecture also makes it easier to improve the model based on specific task characteristics, enabling faster development.

With the increasing usage and application of the UNet model, researchers have made many changes to the basic UNet model to improve the performance of the UNet network in general or in some specialized areas. This paper selects the eight most classical models from all the improved UNet networks and categorizes and describes these eight classical UNet models.

Ronneberger et al presented the UNet model. The basic structure of the model is shown in Fig. 1, which presents a completely symmetric "U" shape. The UNet structure can be divided into three parts: downsampling, upsampling, and skip connection.

According to Ronneberger et al on the structure of the model, UNet has nine layers. During the first four layers of the encode (left side in the model image), each layer has two layers of convolution plus one layer of max-pooling. Then, starting from the fifth layer after the max-pooling of the fourth layer (right side in the model image), each layer contains two layers of convolution and one layer of max-pooling through the parallel layers. Starting from the fifth layer (right side in the model image) after the max-pooling of the fourth layer, each layer has two layers of convolution, plus one up-convolutional layer concatenated with the current layer by copying and cropping the parallel layers [1].

After multiple downsampling of the low-resolution information, UNet can provide the segmentation target in the whole image contextual semantic information, which can be interpreted as features that reflect the relationship between the target and its environment. After concatenating operation from the encoder directly to the high-resolution information on the same height as the decoder, UNet can provide more detailed features for segmentation, such as gradient and so on.

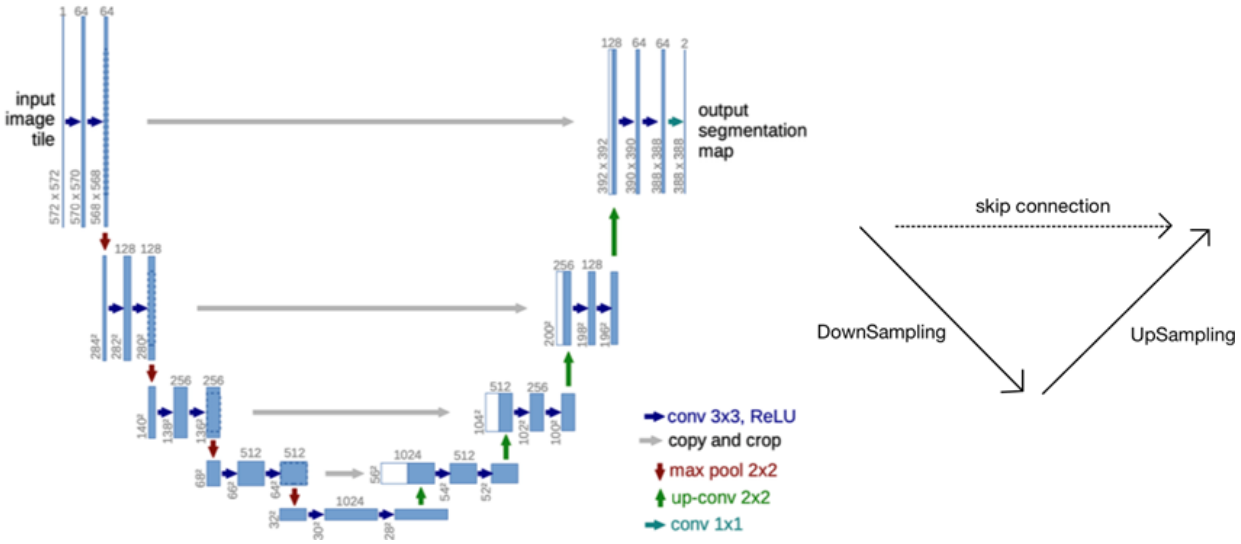

Figure 1. Architecture of UNet [1]

2. UNet Family

With the continuous development of the image segmentation field and the increasing requirements on model performance, researchers are trying to improve UNet in different ways. Different degrees of improvement of UNet can not only improve the performance of semantic segmentation but also enable the model to be optimized in specific domains.

Many variants of UNet are being proposed every year. This paper will present eight classical UNet variants. Then categorize these eight UNet variants according to their improvement methods. The current improvements are categorized into the following four categories: improving convolutional operations, enhancing skip connection, using a novel attention mechanism, and extending spatial dimensions. This classification method visually demonstrates the differences between the variants in terms of specific improvements, allowing researchers and engineers to select and optimize models in a more targeted manner.

2.1. Improvement for Convolution

2.1.1 Res-UNet

Residual Network (ResNet) consists of many residual blocks. The correctness rate hits saturation when the number of layers in the network increases, and if the number of layers is extended by adding a short-circuiting mechanism, the correctness rate drops. This degradation problem is solved by residual blocks. ResNet's key concept is to retain network layer complexity by doubling the number of feature maps when the feature map size is cut in half. Unlike regular networks, ResNet incorporates a short-circuiting mechanism every two levels to provide residual learning [2].

Res-UNet is an image segmentation network structure that combines ResNet and UNet. It improves the performance of the network by integrating the residual block (ResBlock) into UNet. Xiao et al. claim that Res-UNet is inspired by residual connections and incorporates a weighted attention mechanism and a skip connection into the model, replacing each UNet sub-module with a residual connection form. [3] The structure is shown in Fig. 2, where the gray solid line indicates the residual connection added in each module.

In the encoding and decoding process, this model replaces the standard convolutional block in UNet with ResBlock. To aid in the network's improved reconstruction of the details, high-resolution features are sent straight to the decoder through jump connections between the encoder and decoder.

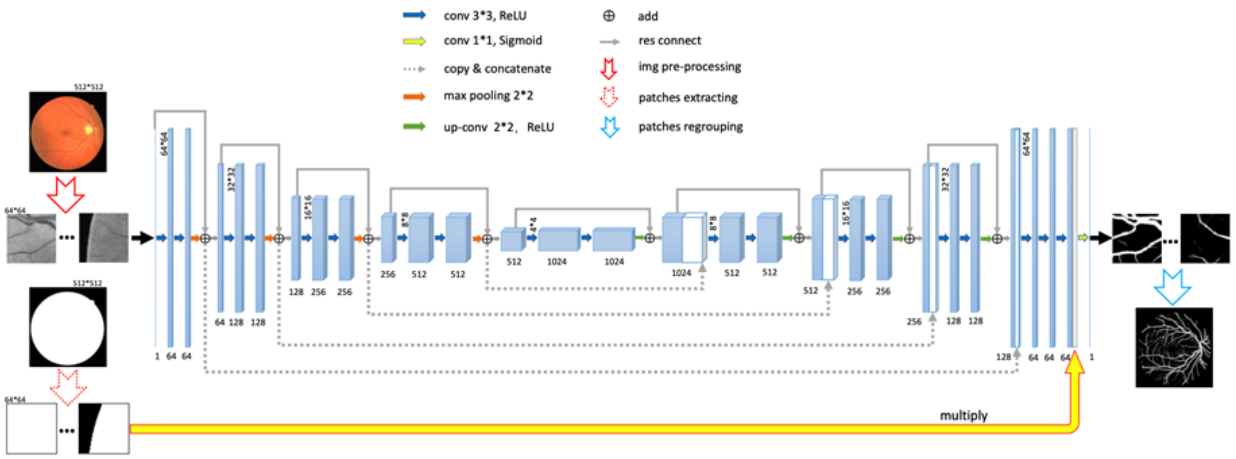

Figure 2. Architecture of Res-UNet [2]

2.1.2 Recurrent Residual CNN-based UNet (R2UNet)

Alom et al. state that this new model mostly replaces the original UNet sub-modules by integrating the ResNet and Recurrent Neural Network (RNN) structures into an encoder-decoder structure. In order to replace the original UNet sub-module, this new model primarily merges the architectures of RNN and ResNet into the encoder-decoder structure. This network is proposed for medical image segmentation [4].

According to Fig. 3, this model combined recurrent convolutional units with each layer. Also, Alom et al designed several different sub-modules as shown in Fig. 3. From left to right, the first one is the conventional method used in UNet. Then the second one is based on the first one which loops a convolutional layer containing an activation function. The third one uses a residual join approach. The fourth one is the cyclic residual convolution module combining (b) and (c) proposed in the paper.

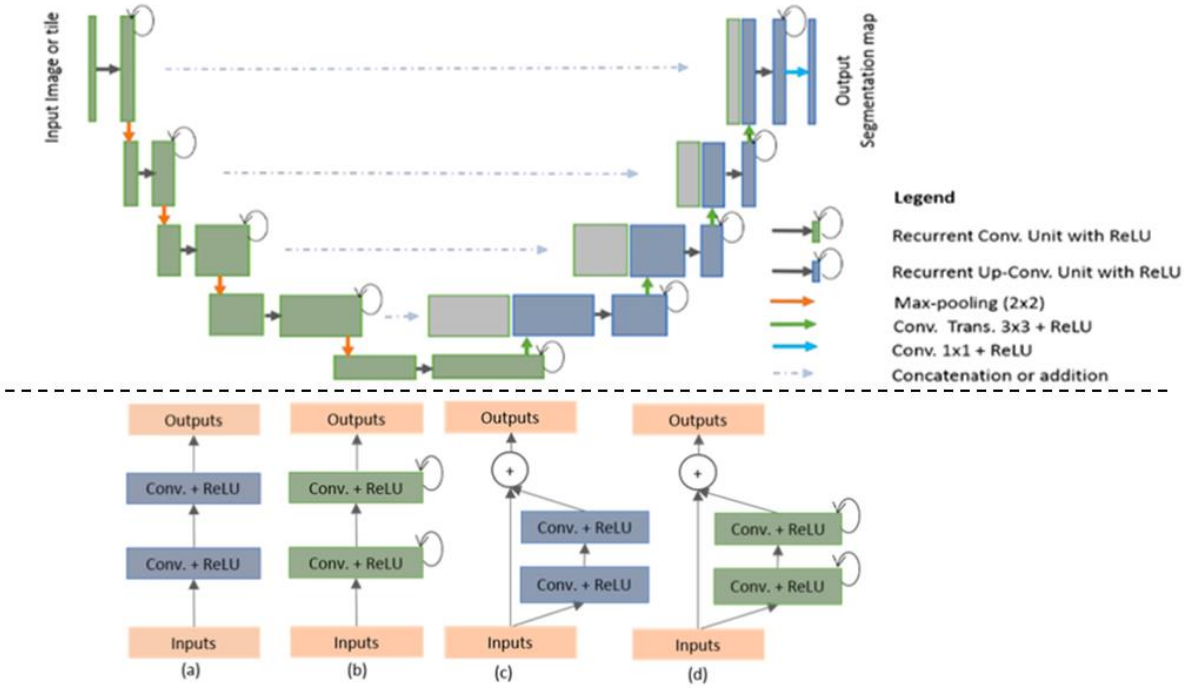

Figure 3. Architecture of R2UNet [4]

2.1.3 MultiResUNet

According to Ibtehaz et al, this paper proposes to combine the Multi-Res block with UNet. Developing the proposed MultiRes block undergoes the following steps: According to Fig. 4, firstly, three convolutional filters, 3*3, 5*5, and 7*7, are arranged in parallel, which are connected in series to generate the feature maps. Then, as shown in (b), the 5*5, and 7*7 filters are decomposed into two

3*3 filters. Then, (c) the results of the three 3x3 convolutional filters in this module are stitched together as a combined feature map, and then added with the results of the input feature maps obtained from the 1x1 convolutional filters, to obtain the MultiRes block [5].

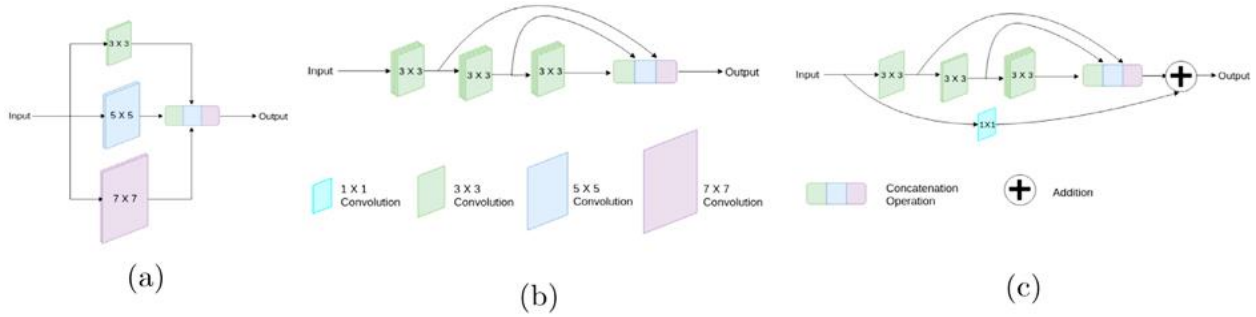

Figure 4. Architecture of R2UNet [5]

To lessen the semantic differences between the features of the encoder and the decoder, the network also suggests a Res Path, which involves passing the encoder characteristics through a number of convolutional layers. Before the encoder features are spliced with the corresponding features in the decoder, Ibtehaz et al. suggest doing a few more convolutional processes [5].

Res Path adapts and enhances the representativeness of the feature map layer by layer through multi-layer residual learning, which helps the network to better capture features at different scales. This makes the information conveyed richer to avoid important details from being omitted.

2.2. Improvement for Skip Connection

2.2.1 UNet++

To avoid the different layers affecting the accuracy of the model, UNet++ connects all the UNets of each layer, as shown in Fig. 5. The short connections are used to train the model, and then the long connections are used to get more information. According to Zhou et al, the advantage of this structure is that no matter which depth features are valid, all of them are used, and the network learns the importance of different depth features by itself. Moreover, UNet++ shares a single feature extractor. Therefore, different levels of features are reduced by different decoder paths. This encoder is still flexible enough to be replaced by a variety of different [6].

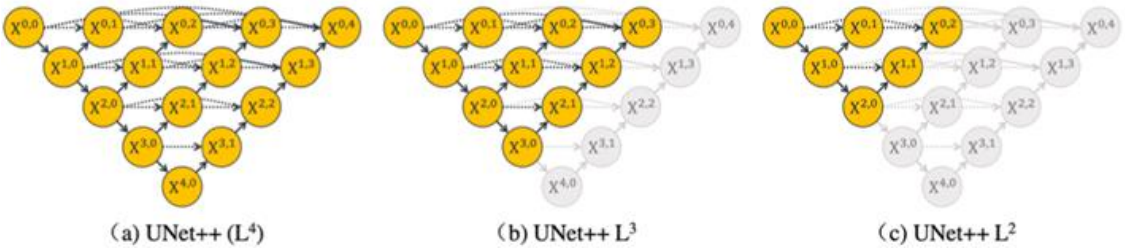

Figure 5. Architecture of UNet++ [6]

In addition, UNet++ can capture different levels of features and integrate them into different levels of features, or different sizes of receptive fields, by means of feature stacking. In this way, after the features with large receptive fields recognize the large object, the features with small receptive fields can help when the edge information of the large object and the small object itself are lost by the deep network's downsampling and upsampling over and over again.

Moreover, UNet++ uses deep supervision. The specific implementation is to add a 1x1 convolutional kernel behind X0,1, X0,2, X0,3, X0,4 in the graph, which is equivalent to supervising the output of UNet at each level, or each branch. This solves the problem of not being able to train that structure. The “pruning” operation is done. The amount of pruning is determined by the results of the subnetwork in the validation set. When L2 yields similar results to L4, pruning can be used so that L2 replaces L4. In this case, the speed and memory can be increased by an objective amount.

2.2.2 UNet3+

According to Huang et al, UNet3+ argues that UNet++ does not fully utilize multi-scale information despite the use of dense skip connections, thus UNet3+ proposes full-scale skip connections, full-scale deep supervisions, and classification-guided modules. Its architecture is demonstrated in Fig. 6 [7].

Full-scale skip connections combine high-level semantics with low-level semantics from feature maps at different scales. That is, each decoder layer combines the encoder with small-scale and same-scale feature maps and the decoder large-scale feature maps. This step captures both coarse- and fine-grained semantics at full-scale.

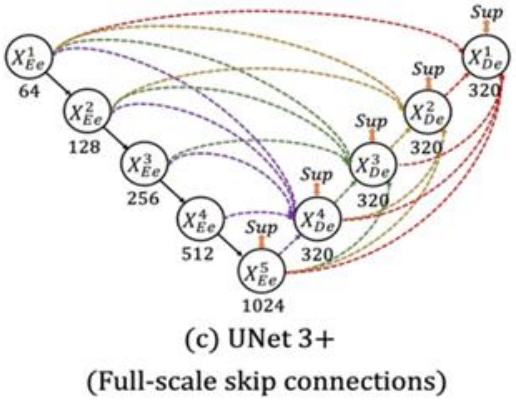

Figure 6. Architecture of UNet3+ [7]

Conversely, multi-scale aggregated feature maps teach hierarchical representations to full-scale deep supervisions. To put it another way, auxiliary loss functions, such as hybrid loss for segmentation in the three-level hierarchy of pixel-level, pixel-level, and Intersection over Union (IoU) loss, are introduced at various intermediate levels of the network to generate full-scale deep supervisions. direct supervision at the following levels of the hierarchy: the pixel, patch, and map levels.

The classification-guided module acts as an extra classification task, which utilizes the global classification information to guide the local pixel-level segmentation, helping the model to better understand and differentiate between different classes of regions.

2.3. Improvement Using Attention Mechanisms

2.3.1 Attention UNet

Attention is divided into Hard Attention and Soft Attention. Hard Attention focuses on a few key locations by selecting them, while Soft Attention weights each pixel of an image to focus on different regions. Instead of Hard Attention, Soft Attention can be used in UNet to learn the weights of attention by calculating the gradient of the neural network and using forward propagation and backward feedback, effectively suppressing activations in irrelevant regions and reducing the redundant part of the skip.

According to Oktay et al, Attention UNet adds the Attention Gate model to the skip connection process in the original UNet structure, as shown in Fig. 7. Training with this model can focus the attention on the useful places and extract only the useful things. According to Fig. 7, the Attention Gate model is combined with ReLU and Sigmoid by 1x1x1 convolution respectively to generate a weight map/alpha. It is then corrected by multiplying it with the features in the encoder [8].

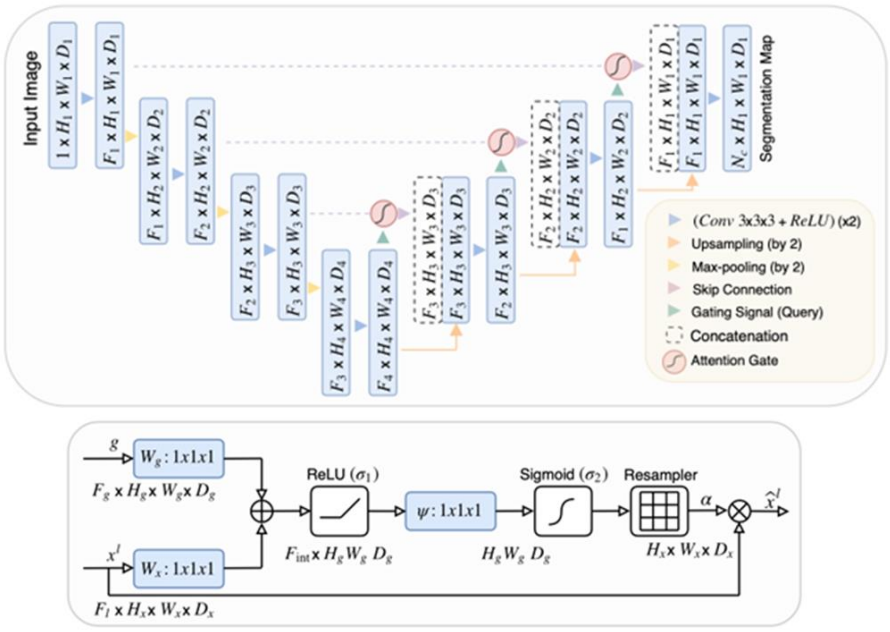

Figure 7. Architecture of Attention UNet [8]

2.3.2 TransUNet

Considering the shortcomings of CNN's convolutional kernel in modeling long-range dependency information, it is a good idea to use Transformer to make up for its shortcomings. However, if simply using the Transformer, the result is not good, probably because of its shortcomings in processing local information. Therefore, according to Chen et al, it proposes to first expand the feature map into pixel-level sequences, which are used as inputs to the Transformer. Then, in the Decoder stage, to compensate for the Transformer's lack of local information processing, the feature map extracted in the CNN encoder stage is combined with the Transformer's hidden feature in different stages (layers), so that both global and local details are available in the decoding process. Compared to utilizing a pure Transformer as an encoder, this hybrid CNN-Transformer encoder performs better [9].

2.4. Extension on Spatial Dimensions

3D UNet was released shortly after UNet, and is a simple extension of UNet for 3D image segmentation. The network uses 3D volumes as input, and operates with 3D convolutions, 3D max pooling, and 3D up-convolutional layers. In this way volume segmentation is achieved [10]. This network employs batch normalization following every convolutional layer, utilizing only three downsampling operations in contrast to UNet's use of dropout.

These modifications lead to the following advantages of 3D UNet: 1. the ability to learn from sparse annotations, which reduces the need for dense annotations and saves the cost of data annotation. 2. the ability to extract useful information from sparse annotations and generalize it to the entire three-dimensional space, generating a complete segmentation result. 3. the ability to provide detailed information about the three-dimensional structure, which supports downstream medical applications. 4. the capacity to get superior generalization outcomes with minimal data training. outcomes of generalization based on sparse training data.

3. Conclusion

In this paper, the author conducts a systematic analysis of UNet. First, the basic structure of the UNet model is elaborated. Then this work categorizes eight classical UNet models and introduces the improvement strategies and methods for each model. These eight UNet network architectures contain the mainstream improvement strategies of the UNet model, providing a foundation for readers' future innovative work. The UNet network performs almost perfectly in solving the medical image

segmentation problem and has been widely utilized in other fields such as computer vision and remote sensing image segmentation. The UNet network has also inspired the design of many subsequent image segmentation networks. However, UNet still requires a large amount of training data and computer resources. Moreover, UNet is still limited in its ability to capture boundary details, handle images with multi-scale targets, distinguish between foreground and background, generalize, and process time-series data. These shortcomings and deficiencies still require further research and innovation to improve.

References

[1] Ronneberger Olaf, Philipp Fischer, Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. Medical image computing and computer-assisted intervention, 2015: 234-241.

[2] He Kaiming, Zhang Xiangyu, Ren Shaoqing, Sun Jian. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[3] Xiao Xiao, Lian Shen, Luo Zhiming, Li Shaozi. Weighted res-unet for high-quality retina vessel segmentation. International conference on information technology in medicine and education, 2018: 327-331.

[4] Alom Md Zahangir, Hasan Mahmudul, Yakopcic Chris, et al. Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation. arXiv preprint, 2018: 1802.06955.

[5] Ibtehaz Nabil, M. Sohel Rahman. MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation. Neural networks, 2020, 121: 74-87.

[6] Zhou Zongwei, Siddiquee Md Mahfuzur Rahman, Tajbakhsh Nima, Liang Jianming. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. IEEE transactions on medical imaging, 2019 39 (6): 1856-1867.

[7] Huang Huimin, Lin Lanfen, Tong Ruofeng, et al. Unet 3+: A full-scale connected unet for medical image segmentation. IEEE international conference on acoustics, speech and signal processing, 2020: 1055-1059.

[8] Oktay Ozan, Schlemper Jo, Folgoc Loic Le, et al. Attention u-net: Learning where to look for the pancreas. arXiv preprint, 2018: 1804.03999.

[9] Chen Jieneng, Lu Yongyi, Yu Qihang, et al. Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint, 2021: 2102.04306.

[10] Çiçek Özgün, Abdulkadir Ahmed, Lienkamp Soeren, et al. 3D U-Net: learning dense volumetric segmentation from sparse annotation. Medical Image Computing and Computer-Assisted Intervention, 2016: 424-432.