Proceedings of the 5th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/15/20230800

Structural analysis of U-Net and its variants in the field of medical image segmentation

Yanjia Kan

School of Artificial Intelligence, Xi'an Jiaotong-liverpool University, Suzhou, 215123, China

Yanjia.Kan20@student.xjtlu.edu.cn

Abstract. Medical image segmentation can provide valuable information for doctors, it has important research value in the medical field. Meanwhile, U-Net, as the fundamental networks for such tasks, brings a substantial improvement in the segmentation performance of traditional medical images. With the increasingly widespread use of U-Net, researchers have designed various U-Net variants according to different task requirements. However, most of the current summaries of U-Net variants are divided according to the direction of network applications, and the structural relationship between the variant networks and U-Net is not elaborated. Therefore, this paper classifies U-Net variants according to their network framework by elaborating the principles of U-Net structure. According to the U-Net network structure, it is divided into three main categories: backbone improvement, module addition and cross-network fusion. Further, the characteristics, advantages and disadvantages of different categories of variants are introduced, and the directions of the variants for U-Net optimization are analyzed. Finally, the article summarizes the current development direction of U-Net variants and provides an outlook on the future directions that can continue to be optimized.

Keywords: U-Net, medical image segmentation, U-Net variants.

1. Introduction

In medical tasks, the main purpose of image segmentation is to segment regions with medical research value from images. Medical images are various, such as magnetic resonance images, ultrasound images, etc. The segmented regions generally have special features. This feature can assist in clinical diagnosis and treatment, as well as provide valid evidence for pathological studies. Therefore, a medical image with accurately segmented lesions can greatly improve the accuracy and effectiveness of disease treatment in the later stage for medical personnel. However, in practical situations, due to the limitations of multiple parties such as acquisition equipment, deterioration of the lesion, and structural peculiarities of the organ, the image segmentation is very difficult. Medical images are complex and some images lack obvious features, especially in the segmentation edges where discriminative linear features may be missing. In addition, medical images are often affected by noise and volume effects, such as uneven gray scale and artifacts. Therefore, traditional segmentation tasks often require mature doctors to complete. However, human judgment is often influenced by many factors, resulting in fluctuations in the accuracy of segmentation. As a result, image segmentation field introduced various segmentation algorithms and neural networks, which greatly reduces the consumption of human and material and the

DOI: 10.54254/2755-2721/15/20230800

losses caused by segmentation errors. As of today, neural networks have penetrated many levels of medical tasks with remarkable achievements. Meanwhile, with the continuous emergence of new technologies, the application of it also has more prospects.

Convolutional neural networks (CNN) have good feature extraction and generalization abilities, and image semantic segmentation is one of the important branches. The vigorous development of semantic segmentation is due to the FCN structure proposed by Jonathan Long and others [1]. Compared to CNN, FCN discards the structure used as classification output behind it and uses convolutional layer to replace its function. In addition, the network adopts a skip connection structure, which realize the conversion of network output from probability to image. Inspired by FCN, Ronneberger and others innovatively created U-Net based on FCN. As one of the variants of FCN, U-Net designs the network structure as a symmetrical network, and the skip connection part uses splicing operations instead of the pixel-by-pixel addition method of FCN, thereby greatly improving the network's feature extraction ability [2]. Therefore, with its simple and flexible characteristics and excellent segmentation ability, U-Net is preferred as the test standard for many segmentation tasks.

In view of this, this article takes U-Net as the core, introduces its network and typical network variants. Subsequently, based on its structure, the variants are divided into three categories (Figure 1). By explaining the U-Net variant structure, analyzing its optimization characteristics, and summarizing the optimization ideas of the U-Net network. Finally, the problems and challenges faced by U-Net are summarized, and the future development direction of the U-Net network is prospected.

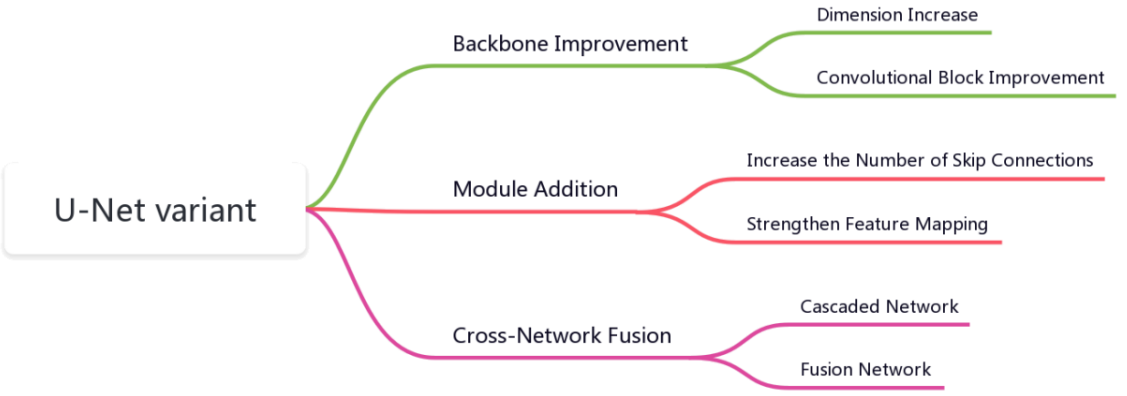

Figure 1. U-Net taxonomy.

2. 2D U-Net

U-Net is mainly adopting an encoder-decoder and skip connection structure to achieve fast and accurate end-to-end network training even with limited data. The encoder part is responsible for capturing contextual information while the decoder part is for mapping features to the required resolution. This includes continuous convolutional operations, bottleneck design, 4 downsampling, and 4 upsampling. In addition, to ensure that the network still retains low-level semantic features in deep structures, the network uses skip connections (Figure 2). This method combines the semantic features of the decoder at the same scale with the encoder’s deep features of to enrich the localization information in the mapping. Finally, the network also uses an overlapping tile strategy to solve the boundary information loss, uses data augmentation to solve the insufficient training data.

DOI: 10.54254/2755-2721/15/20230800

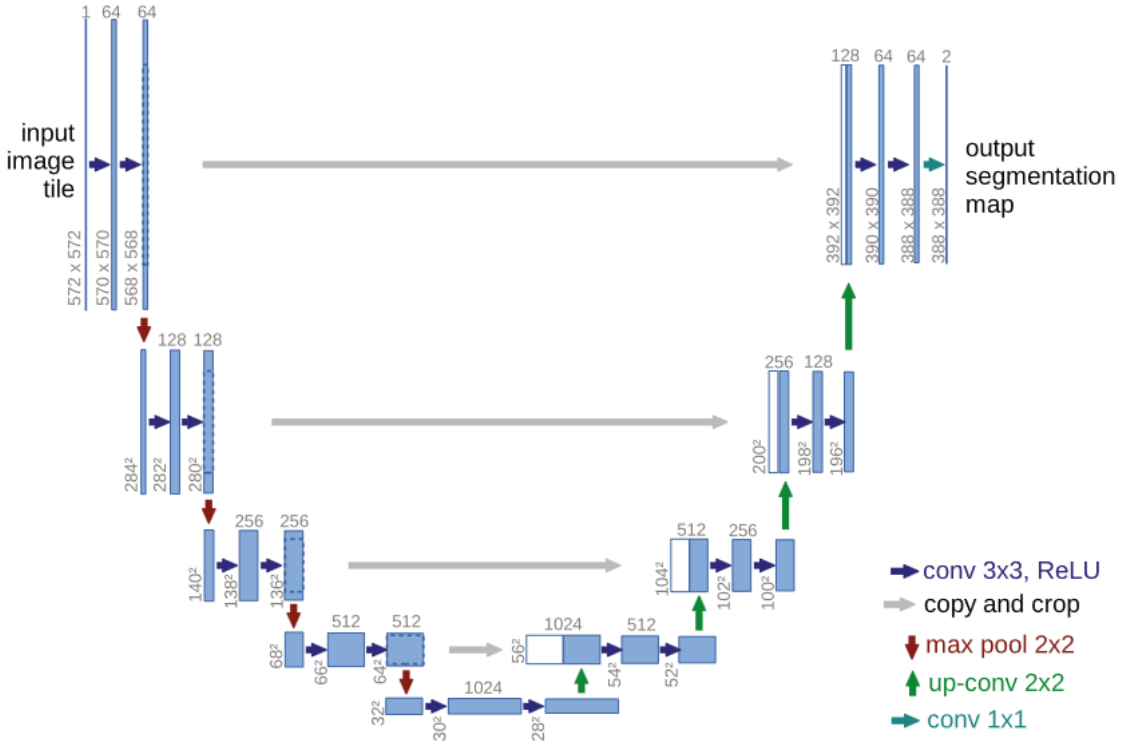

Figure 2. The architecture of U-Net [2].

U-Net, as an important network structure in the field of medical image semantic segmentation, is used to assist in several image analysis tasks. For example, magnetic resonance imaging analysis, computed tomography scan analysis, and ultrasound imaging analysis, etc. However, different images require different feature extraction methods. Therefore, in recent years, a lot of improved methods have been created to make U-Net have highly network suitability for specialized medical images. According to the unique network structure of U-Net, common improvement methods generally optimize the encoder-decoder structure and skip connection structure, resulting in various U-Net variants. These variants can be divided into three categories according to their optimized network positions: backbone improvement, module addition, and cross-network fusion.

3. Variant classification

3.1. Backbone improvements

In U-Net, the backbone network defines how the layers in the encoder are arranged, and its corresponding part is used to describe the architecture of the decoder. Therefore, backbone improvement mainly refers to the improvement of the network encoder-decoder part.

3.1.1. Dimension increase. U-Net is sometimes referred to as 2D U-Net because the network design is based on 2D images as input and output. However, medical images also include a lot of 3D image datasets. Traditional methods involve annotating data by slicing 2D images, which results in redundant data annotation between adjacent slices, tedious annotation process, huge calculation and other problems. Therefore, 3D U-Net, a 3D segmentation method which can converts 2D image operations to 3D image operations was proposed [3]. 3D U-Net retains the U-shaped network structure and makes adjustments to it. The upsampling and downsampling are set to three times, and the image channel remains the same in each deconvolution operation. Changing the channel operation is set in the first convolution after each

DOI: 10.54254/2755-2721/15/20230800

sampling. Finally, batch normalization is added to accelerate convergence. 3D U-Net learned from sparsely annotated 3D images and provided dense 3D segmentation results, which were validated in the African clawed frog kidney confocal microscopy dataset with higher accuracy than 2D U-Net. V-Net is also used for 3D image segmentation [4]. V-Net replaces the upsampling and downsampling parts in U-Net with 2×2×2 convolution kernels with a stride of 2. It also restores four upsampling and four downsampling operations and changes the number of feature maps during convolution (Figure 3). In addition, V-Net introduces residual networks and a new objective function, the dice coefficient, to achieve end-to-end training for prostate MRI images.

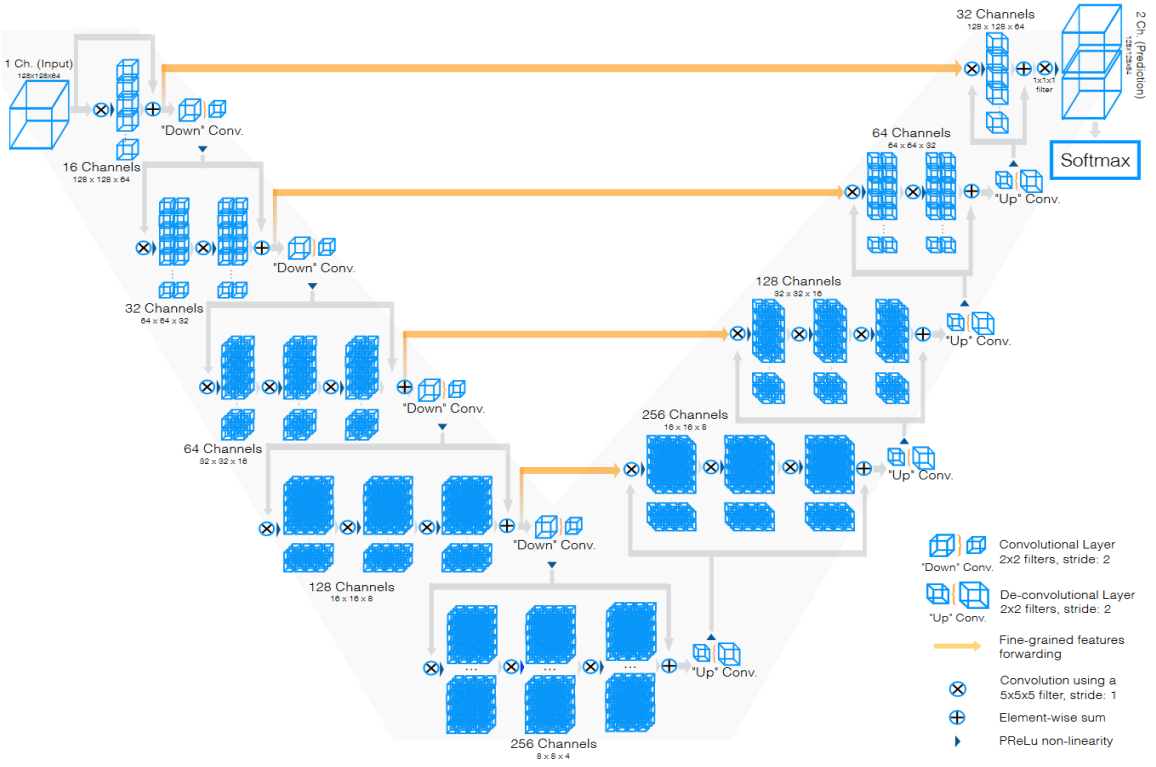

Figure 3. The architecture of V-Net [4].

3.1.2. Convolutional block improvement. The convolutional block of the U-Net network is also a structure that can be optimized. Since the introduction of ResNet, residual blocks have been widely used due to their excellent network performance enhancement capabilities. Since the residual block has a positive effect on avoiding spatial information loss and improving the accuracy of semantic segmentation networks. Therefore, the residual blocks are also used by V-net and ResU-Net to increase the extraction effect of features. Among them, ResU-Net, as a 2D network, combines with a weight mechanism based on the residual block and performs better than U-Net in solving retina segmentation problems (Figure 4) [5]. Ibtehaz further improved on the residual block and designed the MultiResUnet [6]. This network has MultiRes Block modules and Res Path modules. The MultiRes Block replaces the large convolution kernels, such as 5×5 and 7×7 convolution kernels in traditional residual blocks, with continuous 3×3 convolution kernels, and maintains image size by adding 1×1 convolution kernels. This reduces the calculation parameters while ensuring feature extraction. In addition, MultiResUnet also adds Res Path in the encoder-decoder. In each skip connection, Res Path introduces residual connections to connect feature graph to decoder through a convolution chain. This ensures that spatial information lost after each encoder is reversed through deconvolution can be transmitted to the decoder, reducing the semantic differences between corresponding levels. Compared with U-Net, MultiResUnet has significantly improved on challenging datasets, including endoscopic images, skin mirror images, etc.

DOI: 10.54254/2755-2721/15/20230800

Networks with residual blocks also include RU-Net and R2U-Net, designed by Alom and others [7]. The authors used recurrent convolutional layers (RCLs) to replace the original function and proposed four cell structures. Including the recurrent convolutional module used in RU-Net and the recurrent + residual module used in R2U-Net. This network achieves feature accumulation according to different strides, ensuring stronger feature representation. It also outperforms traditional U-Net in segmentation of multiple image datasets.

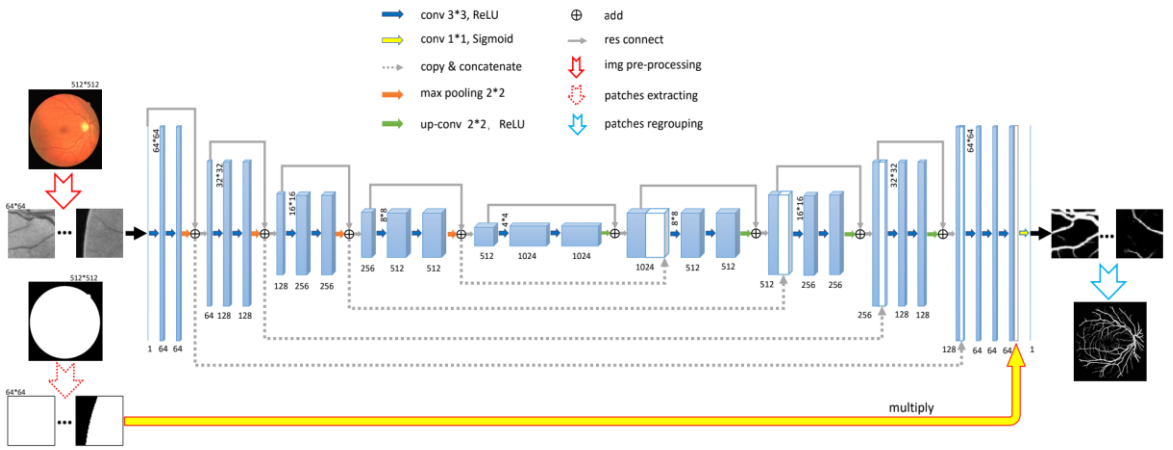

Figure 4. The architecture of ResU-Net [5].

The introduction of residual blocks greatly enhances the network's feature extraction capabilities, enriches the learnable feature quantity of the U-Net, and increases the depth and width of the network, improving its expression and accuracy. However, the large number of redundant features brought by residual blocks can cause the network to learn too many redundant features, increasing the complexity of the network, training burden, and time resource costs.

3.2. Module addition With the development of many functional modules, such as dense connection modules and attention modules, these modules have been widely used in multiple fields due to their outstanding specialty performance and excellent generalization ability. At the same time, skip connection modules play an important role in the U-Net network's semantic feature fusion. By adding and fusing modules, the U-Net network's feature learning ability can be further enhanced.

3.2.1. Increase the number of skip connections. U-Net combines shallow and deep features through skip connections, greatly improving the learning ability and the network's segmentation accuracy. Therefore, by increasing skip connections' number, the model can capture more semantic information and achieve better segmentation performance. U-Net++ is another U-Net architecture variant proposed by Zongwei et al., inspired by DenseNet, which enhances skip connections (Figure 5) [8]. Its structure reduces the semantic gap between corresponding layers. U-Net++ uses a dense skip connection network instead of traditional skip connections between U-Net layers. The network consists of multiple skip connection nodes and dense connections, and all feature mappings of the previous cell at the same level are received by each skip-connected cell after it. Therefore, a dense block can be thought for each layer. The semantic information loss in contraction path and expansion path is reduced by preserving feature mapping to a maximum extent through dense blocks.

DOI: 10.54254/2755-2721/15/20230800

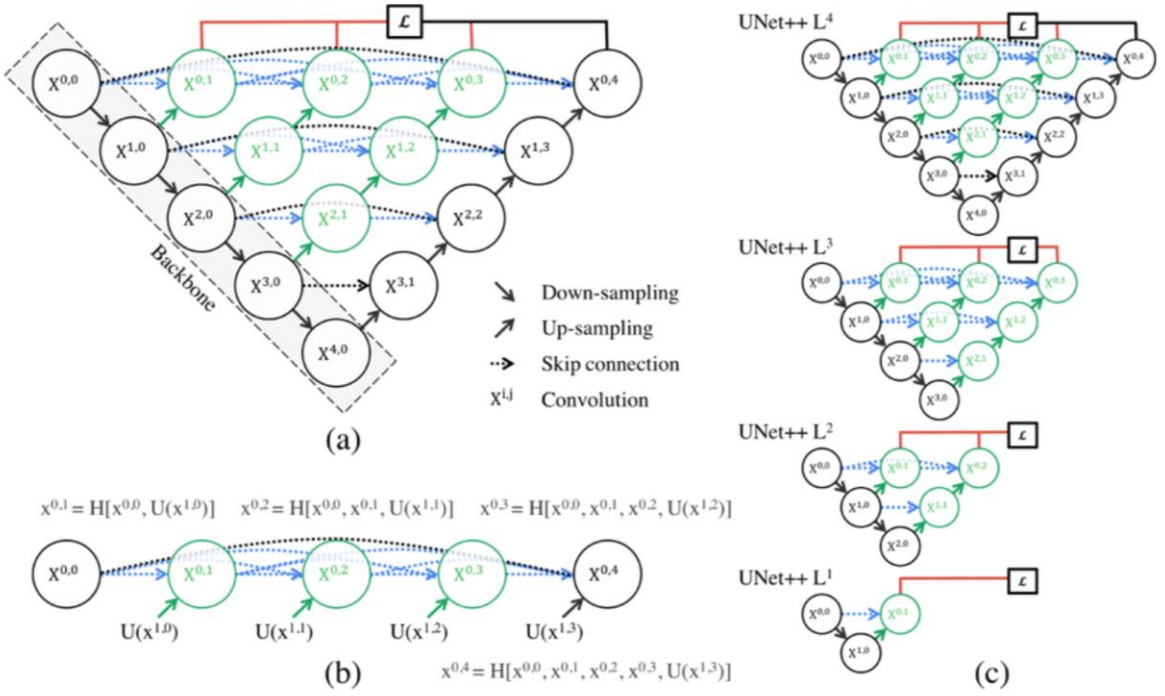

Figure 5. The architecture of U-Net++ [8].

A similar method is U-Net3+ [9]. It is based on U-Net++ and combines smaller features of the same scale retained the decoder's features with larger ones. At the same time, the fine-grained semantics and coarse-grained semantics on the complete scale are captured. The dense connection module improves the disadvantages of residual blocks to some extent. This optimization solution not only solves the limitation of feature fusion at skip connections in U-Net itself but also greatly reduces the network parameters by pruning while preserving the maximum features of the network, ensuring learning depth and network speed. In addition to the dense connection structure, Bio-Net introduces a reverse skip connection structure on the basis of the original forward skip connection [10]. The network establishes bidirectional connections to link the encoder and decoder, and recursively implements the feature mapping between the encoder and decoder. In addition, Bio-Net does not require additional training parameters but its network performance exceeds that of traditional U-Net.

3.2.2. Strengthen feature mapping. In optimizing the skip connection process of the network, variants introduce specific functions modules to enhance feature mapping in skip connections. The common additions are the function of attention module. Attention U-Net introduces an Attention Gate module [11]. This module multiplies the coarse-grained information extracted by the network and the attention coefficients and then fuses the output with the upsampled feature map (Figure 6). This enables the network to learn images of different sizes and shapes, calculate feature importance for each pixel, and suppress invalid information while highlighting prominent features of specified targets. A similar network with a similar function is BCDU-Net proposed by Azad et al [12]. The network also adopts the idea of dense connections to let the network learn more features but also references a new extension module, LSTM module. This module controls feature transmission through gate states and preserves feature data that is significant for segmentation. The network strengthens feature mapping and effective information preservation by combining feature maps with the nonlinear function in the bidirectional Convolutional LSTM module. The network has also demonstrated its excellent network performance in several segmentation tasks.

DOI: 10.54254/2755-2721/15/20230800

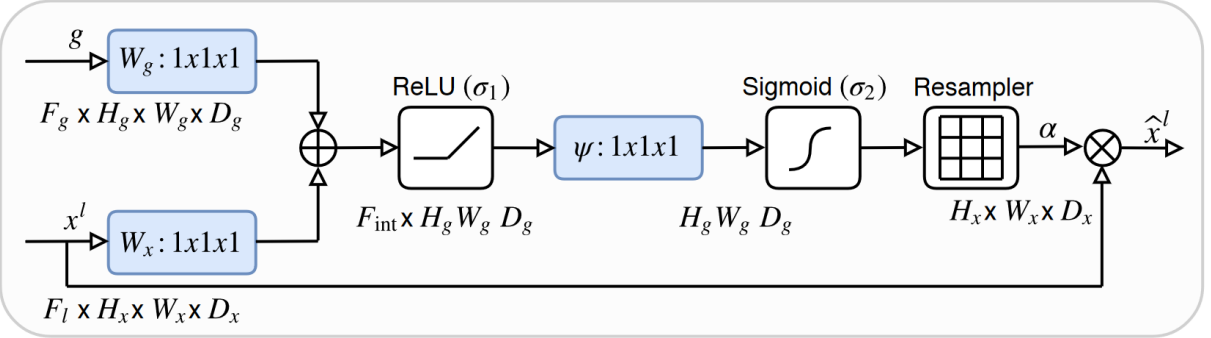

Figure 6. The architecture of Attention Gate [11].

Attention mechanism is also one of the methods to reduce network redundant features. This optimization approach simulates the process of human visual recognition of object features and focuses on enabling the network to make conditional choices to improve its accuracy. In addition, the parallel processing approach in the attention mechanism makes it not only has fewer parameters and better network performance, but also achieves a certain improvement in speed.

3.3. Cross-Network fusion

As one of the fundamental infrastructures of images semantic segmentation, the structure of U-Net has been used by researchers for feature processing of medical images. However, some U-Net variants introduce new network structures by applying other networks used in different fields to the U-Net.

3.3.1. Cascaded network. For tasks that require more processing or are more computationally demanding, a single U-Net network structure is often not enough. Therefore, DoubleU-Net was proposed (Figure 7) [13]. The module combines two U-Net structures, with the first one using VGG-19 as an encoder and both U-Nets using ASPP to capture contextual information. Due to VGG-19's lighter weight and compatibility with U-Net, the authors chose to incorporate VGG-19 into the network. Additionally, since deep networks can achieve more accurate segmentation, the authors added another U-Net to receive the results from the first network's feature output. The DoubleU-Net aims to solve challenging medical image segmentation problems, for which the authors used several datasets to test the network performance. A variety of medical imaging modalities such as colonoscopy, dermoscopy and microscopy are included and demonstrate the excellent network performance of DoubleU-Net. Similarly, the parallel U-Net was designed to better delineate the focal sites of ischemic stroke variant symptoms [14]. This network combines four 2D U-Nets and from CBV, CBF, MTT, and Tmax images extract valuable information about stroke lesion locations. Then, the U-Net probability map is used to determine the extent of the lesion at the pixel level to achieve accurate segmentation.

DOI: 10.54254/2755-2721/15/20230800

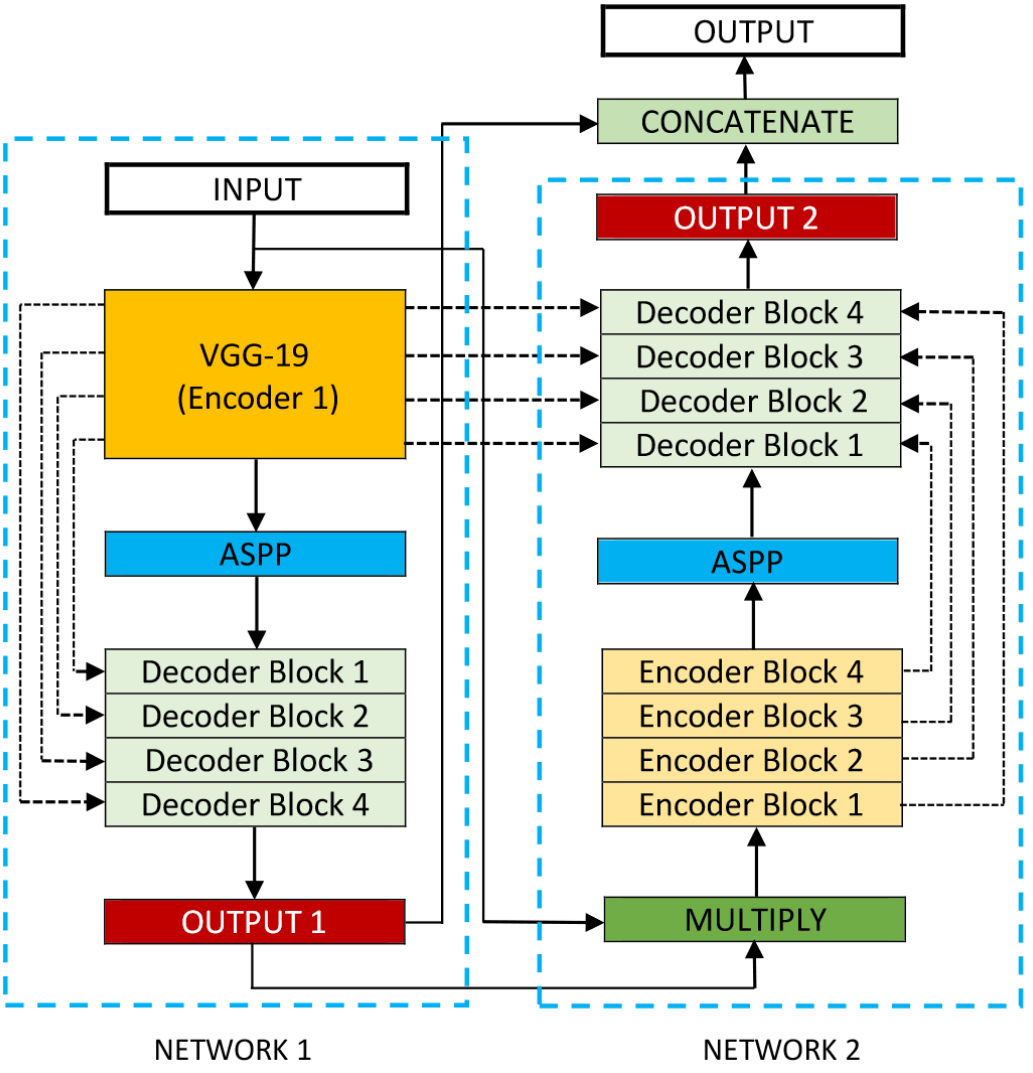

Figure 7. The architecture of DoubleU-Net [13].

3.3.2. Fusion network. In addition to using multiple parallel networks for image segmentation, specific networks with unique structures and functions can be decomposed and reconnected with U-Net's encoder-decoder parts to form a composite network composed of different network structures. TransUNet adopted the Transformer structure based on CNN features in the encoder part [15]. Since CNN can extract local details effectively and the Transformer structure can perceive global information well, Chen et al. combined them to design a fusion network. TransUNet introduces the Vision Transformer (ViT) and successfully applies it to full-size images (Figure 8). The network transforms images into sequences and then encodes global information, making effective use of shallow features to achieve high-precision image feature segmentation. Another example of network fusion is the Generative Adversarial U-Net [16]. The network addressed the issue of limited medical image labeling by introduced the Generative Adversarial Network (GAN) structure and designed a domain-impartial model that could be applied to various medical images. The network separates the U-Net generator parts and encoder, makes the overall network have the function of image generation by combining the generator and GAN. Moreover, the network optimizes the image generation process using conditional GANs and Wasserstein GANs. And it also achieves broad applicability to images with insufficient data.

DOI: 10.54254/2755-2721/15/20230800

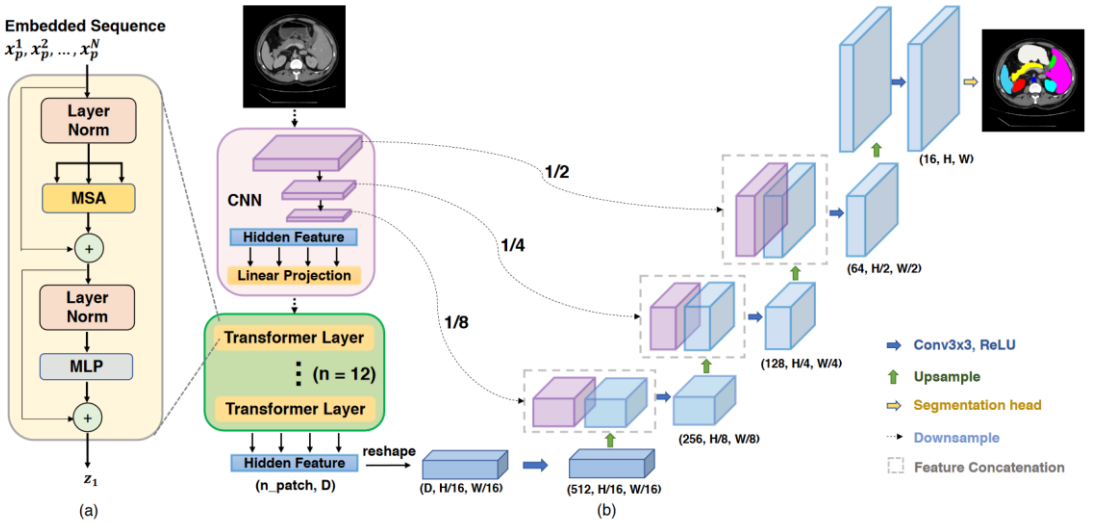

Figure 8. The architecture of TransUNet [15].

Cross-network fusion is a new optimization strategy proposed in recent years for U-Net networks. Technologies in the fields of CNN, RNN, and others are constantly innovating with the times. This optimization strategy is a practical approach to the network's flexibility and generalization. "Taking the strengths and weaknesses" into account not only solves the problems of the U-Net network itself but also extends its capabilities.

4. Conclusion

As an extremely important network in medical image segmentation, U-Net achieves its flexibility and generalization through its unique encoder-decoder structure and skip connections. Based on the demand for different medical image tasks, researchers have optimized the performance of U-Net networks in many directions, resulting in multiple U-Net variant networks. This article divides them into three categories based on the optimized structural parts: backbone improvement based on encoder-decoder structure, module addition based on skip connection structure, and cross-network fusion based on the entire network's functional structure. By discussing the network structures, the optimization strategies of U-Net variant networks are summarized.

It can be seen from the article that there is a demand for residual blocks, dense connections, and attention mechanisms in the future optimization and development of U-Net networks. These methods have an excellent effect on feature extraction and network learning capabilities. However, there is still room for development. While dense connections reduce the burden of network training to a certain extent compared to residual blocks, it still cannot achieve a balance between network training speed and performance. Therefore, how to solve the balance problem to achieve better optimization is a direction for development. Additionally, the effectiveness of the attention mechanism requires a large amount of data as a basis. At the same time, as the demand for segmentation networks in medical field increases, the attention features of some tasks will change due to factors such as time and space. Therefore, the controllability of attention mechanisms in terms of time and space is another area for optimization. In the end, multi-domain network fusion is a new direction for the development of U-Net variants. With the gradual complexity of image segmentation tasks in the medical field, the development of multifunctional composite networks will have extremely broad application prospects.

References

[1] Long, J., Shelhamer, E., & Darrell, T. Fully convolutional networks for semantic segmentation. 2015, *Computer Vision and Pattern Recognition*. 3431-3440.

DOI: 10.54254/2755-2721/15/20230800

[2] Ronneberger, O., Fischer, P., & Brox, T. U-net: Convolutional networks for biomedical image segmentation. 2015, *In Medical Image Computing and Computer-Assisted Intervention*, 234-241.

[3] Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., & Ronneberger, O. 3D U-Net: learning dense volumetric segmentation from sparse annotation. 2016, *In Medical Image Computing and Computer-Assisted Intervention* 424-432.

[4] Milletari, F., Navab, N., & Ahmadi, S. A. V-net: Fully convolutional neural networks for volumetric medical image segmentation. 2016, *In 3DV*. 565-571.

[5] Xiao, X., Lian, S., Luo, Z., & Li, S. Weighted res-unet for high-quality retina vessel segmentation. 2018, *In International Textile Machinery Exhibition* 327-331.

[6] Ibtehaz, N., & Rahman, M. S. MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation. 2020, *Neural networks*, **121**, 74-87.

[7] Alom, M. Z., Hasan, M., Yakopcic, C., Taha, T. M., & Asari, V. K. Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation. 2018, arXiv preprint arXiv:1802.06955.

[8] Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N., & Liang, J. Unet++: A nested u-net architecture for medical image segmentation. 2018, *In Deep Learning on Medical Image Analysis* 3-11.

[9] Huang, H., Lin, L., Tong, R., Hu, H., Zhang, Q., Iwamoto, Y., ... & Wu, J. Unet 3+: A full-scale connected unet for medical image segmentation. 2019 *In International Conference on Acoustics, Speech and Signal Processing* 1055-1059.

[10] Xiang, T., Zhang, C., Liu, D., Song, Y., Huang, H., & Cai, W. BiO-Net: learning recurrent bi-directional connections for encoder-decoder architecture. 2020 *In Medical Image Computing and Computer-Assisted Intervention*. 74-84.

[11] Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., & Rueckert, D. Attention u-net: Learning where to look for the pancreas. 2018, *arXiv preprint arXiv:1804.03999.*

[12] Azad, R., Asadi-Aghbolaghi, M., Fathy, M., & Escalera, S. Bi-directional ConvLSTM U-Net with densley connected convolutions. 2019, *International conference on computer vision workshops*.

[13] Jha, D., Riegler, M. A., Johansen, D., Halvorsen, P., & Johansen, H. D. Doubleu-net: A deep convolutional neural network for medical image segmentation. *In Conference Board of the Mathematical Sciences* 558-564.

[14] Soltanpour, M., Greiner, R., Boulanger, P., & Buck, B. Ischemic stroke lesion prediction in ct perfusion scans using multiple parallel u-nets following by a pixel-level classifier. 2019 *In Biological Information and Biomedical Engineering.* 957-963.

[15] Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., ... & Zhou, Y. Transunet: Transformers make strong encoders for medical image segmentation. 2019 *arXiv preprint arXiv:2102.04306*.

[16] Chen, X., Li, Y., Yao, L., Adeli, E., & Zhang, Y. Generative adversarial U-Net for domain-free medical image augmentation. 2021 *arXiv preprint arXiv:2101.04793*.