

Multiple U-Net-based CNNs in Ultrasonic Image Segmentation of hemangioma

Rongzhou Zhou 24520211154667 Yuehui Qiu 30920211154164
Guanyi Zheng 30920211154142 Zewei Tao 24520211154661
Xiaobo Liu 24520211154656

Abstract

At present, the clinical segmentation of hemangioma lesions is mainly drawn manually by experts. Affected by the level of clinical experience, the segmentation results will inevitably appear human errors. Using deep learning technology can help segment more accurately and efficiently, but the acquisition of labeled samples is difficult and costly. How to use few-shot datasets to train and get automatic and accurate segmentation of ultrasonic images has become one of the current research hotspots. In order to solve the problem, this paper uses four excellent U-Net based networks, embeds three of them into the nnU-Net (“No New-Net”) framework for training and inferring, and tries to use integrated inference and activity contour model for better final results. The best-performing model ranks ninth in the training match of CCF BDCI with a dice coefficient of 0.886.

KEY WORDS: Medical image segmentation
U-Net based networks Integrated inference

Introduction

Epidemiological statistics show that the incidence of infantile hemangioma is 10% to 12%, mainly seen in premature and female infants. Ultrasound examination is non-invasive and can provide clinical information such as location, shape and range of involvement of hemangioma, which can help guide doctors for further treatment. However, it takes a lot of time for doctors to manually outline the location and size of tumors. Some relevant studies have proposed such a high-precision auxiliary labeling algorithm, which can improve the work efficiency of clinicians by 40 % (Lin et al. 2019).

According to the different segmentation ideas, traditional medical image segmentation methods can be divided into regions (Wang et al. 2015; Ma, Yang, and Zhao 2020) and boundary-based methods (Chan and Vese 2001; Islam and Kashem 2021). However, traditional segmentation networks need to design their own features manually, which usually can only play a good effect on the corresponding data set, with poor generalization ability. With the progress of computing power and the development of deep learning, modern segmentation algorithms mostly focus on the study of deep learning.

At the same time, the training of a modern deep learning model usually depends on a large number of high-quality training samples, and the number and quality of training samples usually play a decisive role in the quality of a model. The privacy of medical imaging makes it extremely difficult to obtain a large number of samples. At the same time, the marking of medical images requires manual drawing by clinical experts, which requires a lot of manpower and inevitably has a certain subjective bias. It is a difficult task to obtain a high quality model when samples are scarce and label quality is not guaranteed.

Progress in segmentation accuracy is mainly promoted by convolutional neural network (CNN) (Simonyan and Zisserman 2015; He et al. 2015). However, due to the locality of convolution kernel, traditional segmentation models based on CNN (such as FCN (Shelhamer, Long, and Darrell 2016)) lack the modeling ability of long-term dependencies. In order to solve this problem, a large number of studies have adopted a variety of methods to model context relations. For example, the spatial pyramid based approach (Chen et al. 2018; Zhao et al. 2016; Gu et al. 2019) uses convolution kernels of different sizes to aggregate context information from different ranges in a single layer. The codec network (Ronneberger, Fischer, and Brox 2015; Ji et al. 2020) based on U-Net adopts the method of jump connection to segment coarse-grained deep features and fine-grained shallow features at the same scale, and obtains coarse-grained deep features and fine-grained shallow features at the same scale. Although these methods have achieved great success in density prediction, they are still limited by the inefficiency of non-local context modeling between arbitrary locations, which makes it difficult to further improve the accuracy of complex views.

Recently, Transformer (Vaswani et al. 2017) based on self-attention mechanism and global feature extraction has gained wide attention in various visual tasks (Dosovitskiy et al. 2021; Xie et al. 2021a; Zhu et al. 2021). For medical image segmentation, Chen et al. first proposed TransU-Net (Chen et al. 2021), which uses self-attention mechanism to calculate global context based on deep features extracted by CNN to ensure various range dependencies within a specific scale. This method can give full play to the advantages of both CNN and ViT, and better extract local and global feature representation information. However, the transformer

model usually performs well only on large data sets, and over-fitting can easily occur in training on few-shot sets.

In order to further improve the segmentation accuracy based on a very few-shot set, precise data preprocessing and multi-model integrated inference are used to achieve the desired results in this paper. Data preprocessing adopts the nnU-Net (Isensee et al. 2019) preprocessing mechanism, which adopts a series of heuristic rules to enhance data according to the data fingerprint of a given data set. Data enhancement methods include random rotation, random scaling, random elastic transformation, gamma correction, mirroring and other methods. Taking the most classical U-Net as an example, compared with the original method, the U-Net with this data enhancement method has improved 9.5 % in the dice coefficient of prediction, which is a huge improvement in accuracy. At the same time, due to the use of a 2D data set, the speed of inference is fast. In order to improve the accuracy of prediction and make full use of the advantages of each model, this paper adopts the method of integrated inference to achieve higher accuracy in the acceptable inference time. The integrated reasoning process in this paper is roughly as follows: Towards each pixel point on prediction mask of four trained models, which has only 0 (background) or 1 (foreground) two choices, we adopt the principle of the minority is subordinate to the majority, vote on every pixel point and finally get the integrated output , hoping to further improve the precision of prediction.

Related Work

With the great progress made by AlexNet(Krizhevsky, Sutskever, and Hinton 2012) in ImageNet competition, CNN gradually replaced traditional methods and became the mainstream in the field of computer vision. And some classic networks emerged, such as VGG(Simonyan and Zisserman 2015), ResNet(He et al. 2015) and DenseNet(Huang et al. 2017). Deep convolutional neural networks (CNNs) are driving the development of various computer vision tasks. Medical imaging also benefits from many of these aspects. A large number of segmentation algorithms have been proposed in recent years. Here are some outstanding examples: First introduced by U-Net (Ronneberger, Fischer, and Brox 2015), a variant of encoder-decoder style structures with skipped connections, including the introduction of residual connection (Milletari, Navab, and Ahmadi 2016), dense connection (Li et al. 2017), attention mechanism (Oktay et al. 2018), additional loss layer (Kayalibay, Jensen, and Patrick 2017), feature recalibration (Roy, NavAb, and Wachinger 2018; Qin et al. 2018). The specific modifications vary widely, but they all have one thing in common: they are all based on U-Net structures. But some of them do not perform as well as the well-designed classical U-Net. The result is a finely preprocessed and highly practical network called nnU-Net(Isensee et al. 2019). This network achieved SOTA effect in the decathlon segmentation competition and is widely used in 3D image segmentation.

Recently, Transformer, due to its success in the field of natural language processing, has also been used for medical image segmentation. Many scholars combine CNN with Transformer, and the key is to combine self-attention and

convolution in Transformer (Chen et al. 2021; Xie et al. 2021b; Zhou et al. 2021), but its basic architecture is still based on U-Net. In this way, CNN’s advantages of local feature extraction and Transformer’s advantages of global feature extraction can be better used to obtain better model accuracy.

In this paper, training will be carried out on U-Net and its derived networks and transformer. Meanwhile, the traditional active contour model (Chan and Vese 2001) will be used for post-processing of test results, combined with integrated inference in order to obtain higher accuracy.

Proposed Solution

This study aims to solve the problem of accurate segmentation of 2D ultrasound image of a hemangioma. In the field of medical image segmentation, U-Net integrates shallow features with deep features by virtue of its encoder-decoder structure, so as to achieve accurate segmentation of medical image boundaries and positioning of target regions. Many modules, mechanisms and learning methods from other fields are also used in medical image segmentation. In this study, several different networks based on U-Net structure are selected for experiments, which are as follows: The original U-Net, nnU-Net representing SOTA, CE-Net(Gu et al. 2019) using a pre-trained model as an encoder, and the recently popular nnFormer combining Transformer. Based on nnU-Net, the five models of five-fold cross-validation and the results of four networks are integrated. Considering that pixel-level voting may leads to some fragmented regions or rough boundaries, the traditional segmentation method – active contour model is used for simple post-processing of the integrated results. This study will follow the process in Fig.1

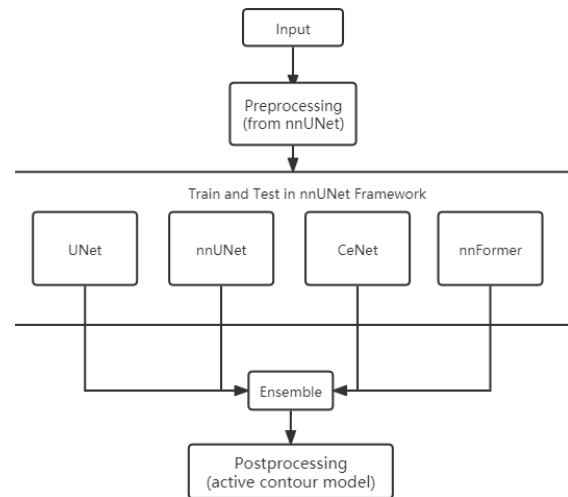


Figure 1: The overall flow figure of solution

U-Net

U-Net is a pioneering work in the field of medical image segmentation. After 2015, almost all medical image segmenta-

tion networks have been developed on this basis. Its network architecture is shown in Fig.2 When U-Net was proposed, its original text was designed for the segmentation of 2D data sets, which is similar to the data set used in this study. At the same time, U-Net is a light network, and the advantage of large network is stronger image expression ability, while the simple and small number of medical images do not have so much content to be expressed. Therefore, it is also found that in a few-shot data sets, large network has no significant advantage over lightweight U-Net.

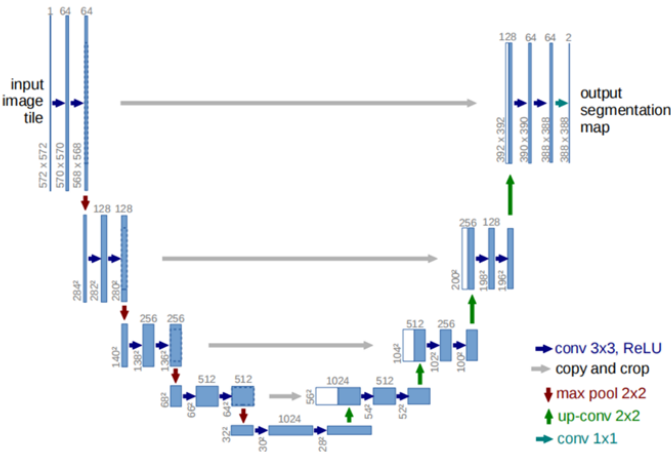


Figure 2: Network architecture diagram of U-Net

nnU-Net

nnU-Net is a medical image segmentation network proposed by Fabian Isensee et al in 2019. The main point is that despite the recent success of deep learning based segmentation methods, their applicability to specific image analysis problems of end users is generally limited. The task-specific design and configuration of methods requires a high level of expertise and experience, and small errors can lead to significant performance degradation. Thus, the authors make the assumption that the basic U-Net architecture is hard to beat if the proper processing can be designed. The authors propose nnU-Net, which enables successful 3D biomedical image segmentation for biomedical research applications, to perform data enhancement through delicate pre-processing to extract key representations. The nnU-Net exceeds the professionally designed model for a particular task in a large number of tasks, and its operational logic architecture is shown in Fig.3

This study is not based on the data set directly using nnU-Net's well-performed competitive network, mainly because nnU-Net has many three-dimensional image data of structure and process, and the modal task is somewhat different from ours. This is also mentioned in the official open source code. However, we can extend the processing of 3D images to 2D images by converting 2D images into pseudo-3D images that can be recognized by nnU-Net framework, so as to achieve the same SOTA effect. And the image obtained by

its unique pre-processing mechanism can be used in other networks to achieve the fusion between networks.

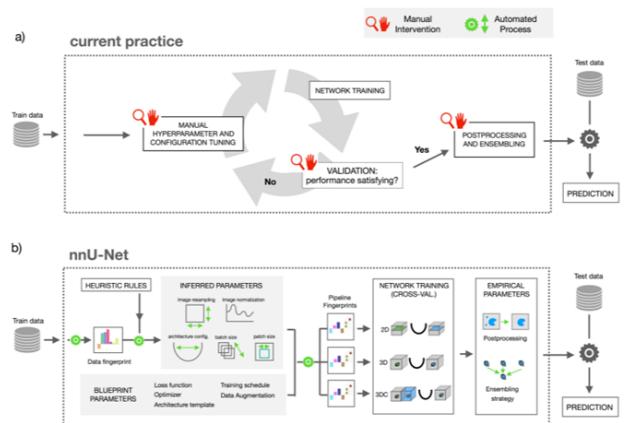


Figure 3: Network architecture diagram of nnU-Net

CE-Net

CE-Net(Gu et al. 2019) believed that continuous pooling and step-convolution operations would lead to the loss of some spatial information, and proposed a context encoder network to capture more high-level information and retain the spatial information used for 2D medical image segmentation. As shown in Fig.4, CE-Net mainly contains three main components: feature encoder module, context extractor module and feature decoder module. A pre-trained ResNet block is used as a fixed feature extractor. Context extractor module is composed of dense porous convolution (DAC) block and residual multicore pooling (RMP) block. CE-Net is applied to different 2D medical image segmentation tasks. The results show that the proposed method is superior to the original U-NET method and other advanced methods in optic disc segmentation, blood vessel detection, lung segmentation, cell contour segmentation and retinal optical coherence tomography layer segmentation. Since this study is also for the segmentation of 2D medical images, and this network has a good effect, CE-Net will be used in the subsequent experiments.

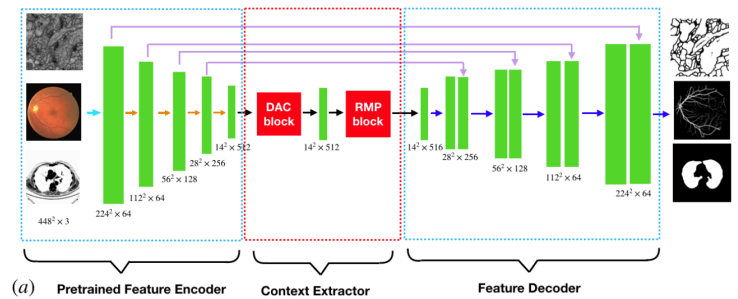


Figure 4: Structure figure of CE-Net

nnFormer

This model aims at inheriting nnU-Net preprocessing mechanism and introducing transformer architecture with good performance at present, because Transformer is expected to help atypical convolutional neural networks overcome their inherent shortcomings of spatial induction bias. However, most of the recently proposed Transformer-based segmentation methods only use Transformer as an auxiliary module to help encode the global context as a convolution representation, without studying how to best combine self-attention (the heart of Transformer) with convolution. To solve this problem, nnFormer, a powerful segmentation model, is proposed with an interlocking architecture based on the combination of self-attention and convolution experience. It has achieved good results on synapse multi-organ classification dataset, and its network architecture is shown in Fig.5

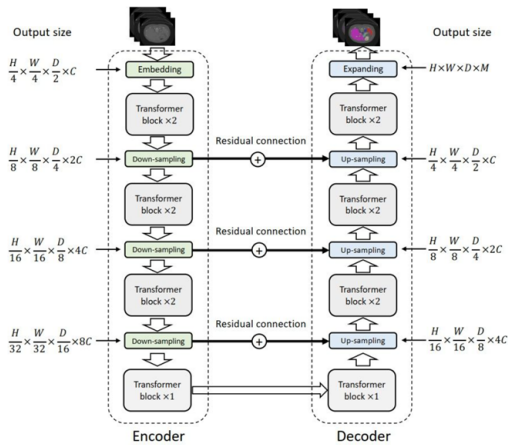


Figure 5: Network architecture diagram of nnformer

Experiment

The experiment of this study is based on nnU-Net, CE-Net, U-Net and nnFormer. These four U-Net based networks achieve excellent performance. Using nnU-Net’s preprocessing mechanism and training strategy, the four networks will output four prediction masks, which will be evaluated by the official evaluation machine respectively. The scores of the official evaluation machine are given different weights to carry out integrated inference. The flow chart of the overall framework is shown in Fig.1

Dataset

We used data sets from the hemangioma Ultrasound Image Segmentation Competition of CCF Big Data Computational Intelligence Competition. The dataset consists of 215 training samples and 107 test samples. Both training and test samples are 2D images in png format. The test samples are not labeled, and the test results were uploaded to the website for evaluation.

As for the preprocessing of data sets, nnU-Net has a self-contained data analysis framework and will generate data

fingerprints of corresponding data sets according to experience. In this experiment, nnU-Net’s data preprocessing and data enhancement methods will be used containing cropping, resampling, rotation, scaling, gaussian noise, gamma correction, mirroring and so on. Fig.6 shows the image changes before and after data pretreatment.

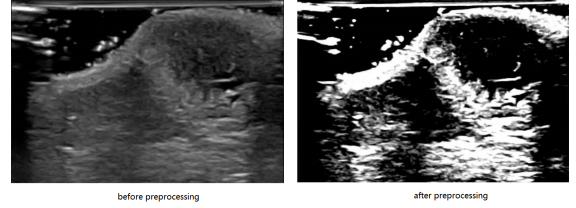


Figure 6: Comparison of image before and after nnU-Net pretreatment

Implementation Details

Our models is trained and tested by integrating the network structure into the nnU-Net framework, so the flow and parameter settings are similar to the default configuration of nnU-Net. SGD is used as the optimizer and momentum was 0.99. It is found in the experiment that the default learning rate of 1e-2 would cause the shock of training, so small initial learning rate like 1e-4 and poly learning rate strategy are adopted. The patch size is 512*320, the Batch size is 2, and the max training epoch is 500. Training for too many epochs will cause serious overfitting . As for the loss function, we use combination of dice loss and cross-entropy loss.

Experiment Results

The experiment is carried out based on U-Net, but as can be seen from Fig.7, the performance of the original U-Net is not good, the dice coefficient is only 0.74, and many images can not be recognized. Inspired by nnFormer, this study embed CE-Net and U-Net into nnU-Net architecture for training. At the same time, the internal verification of nnU-Net is carried out by five-fold cross-validation. Finally, an integration is carried out according to the internal verification of five different models. After comparison, it is found that the effect of integration is better than a single fold.

It can also be seen from Fig.7 that the network with the best performance is nnU-Net, whose dice coefficient reaches 0.886 and exceeds current transformer based nnformer and other U-Net based CNNs. The general reason can be attributed to the coupling between the network structure and its own processing flow design, which can achieve better results. However, the framework used by nnU-Net has significantly improved the U-Net model, reducing the fragmenting fragments. Meanwhile, the DICE coefficient has increased by 12.8 %, and the dice coefficient of embedded CE-Net has reached 0.882, indicating that the framework of nnU-Net is of certain portability. At the same time, it also shows that in image segmentation of few-shot sets, fine preprocessing may be more effective than adjusting network parameters or network architecture.

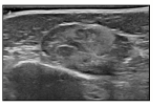






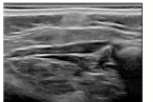






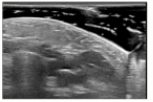






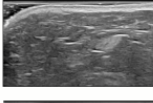








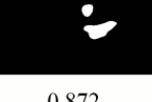



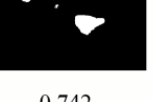
NO.	IMAGES	nnUNet	nnFormer	CeNet	UNet	Ensemble	Vanilla UNet
36							
85							
95							
99							
106							
DICE		0.886	0.872	0.882	0.837	0.883	0.742

Figure 7: Experimental results

However, the integrated inference based on the nnU-Net idea does not achieve the expected effect in this paper, which may be due to the following reasons: 1. All the networks are based on the U-Net architecture, which makes the features learned from the network are relatively similar and most of the segmentation inference results are relatively similar 2. Such high-precision models as nnU-Net already exist in these networks, and the effect of integration is not obvious. However, it ranked first in this competition with 0.9. The research results of this paper are close to the SOTA model, with only a gap of about 1%.

The traditional active contour model is used to post-process the integrated results, as shown in Fig 8. Due to the small changes, the segmented contours before and after post-processing are displayed in the same figure. The black line represents the contours before post-processing, while the bright line represents the contours after post-processing.

As shown in Fig.8, the active contour model can be expanded outward around the contour, and the extent of expansion is different for various samples. It tends to move towards the (bright) region with large gray value. This is because the minimum threshold is set in the algorithm, so the region with small gray value (dark) is not easy to be included. However, the gray values around the contour segmented by different samples vary, and the gray values around the contour of the same sample also vary. Therefore, local pixel information should be used to calculate the threshold, so as to produce better post-processing results.

Conclusion

In this paper, multiple convolutional neural networks based on U-Net architecture are used for segmentation prediction of ultrasound image of a hemangioma provided in the train-

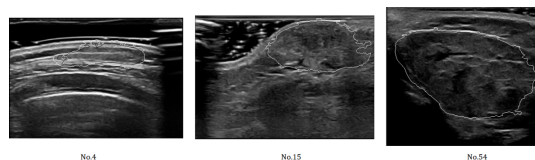


Figure 8: Results before and after active contour model

ing match of CCF BDCI, and the segmentation result has won the 9th place in this competition, achieving an effect close to SOTA. It is found that for some networks, embedding in nnU-Net framework can obviously improve the test accuracy. In addition, ensemble inference can not achieve better results than member model in all cases, and the rules of integration and the selection of member model need to be further explored. The post-processing using active contour model does not get expectant results, which is caused by the non-uniformity of gray level of medical images. Therefore, the local information of samples can be added into the calculation later.

After compositional experiments on a variety of models, we have a clearer understanding of the process of medical image segmentation and a better understanding of some new networks or structures. In the future, we will apply the experience gained from the experiment to our further scientific research.

References

- Chan, T. F., and Vese, L. A. 2001. Active contours without edges. *IEEE Transactions on Image Processing* 10(2):266–277.
- Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2018. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40:834–848.
- Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A. L.; and Zhou, Y. 2021. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv: Computer Vision and Pattern Recognition*.
- Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.
- Gu, Z.; Cheng, J.; Fu, H.; Zhou, K.; Hao, H.; Zhao, Y.; Zhang, T.; Gao, S.; and Liu, J. 2019. Ce-net: Context encoder network for 2d medical image segmentation. *IEEE Transactions on Medical Imaging* 38:2281–2292.
- He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Deep residual learning for image recognition. *arXiv: Computer Vision and Pattern Recognition*.
- Huang, G.; Liu, Z.; van der Maaten, L.; and Weinberger, K. Q. 2017. Densely connected convolutional networks. In *Computer Vision and Pattern Recognition*.
- Isensee, F.; Petersen, J.; Kohl, S. A. A.; Jäger, P. F.; and Maier-Hein, K. H. 2019. nnu-net: Breaking the spell on successful medical image segmentation. *CoRR* abs/1904.08128.
- Islam, M. M., and Kashem, M. A. 2021. Parametric active contour model-based tumor area segmentation from brain mri images using minimum initial points. *Iran Journal of Computer Science*.
- Ji, Y.; Zhang, R.; Li, Z.; Ren, J.; Zhang, S.; and Luo, P. 2020. Uxnet: Searching multi-level feature aggregation for 3d medical image segmentation. *arXiv: Computer Vision and Pattern Recognition*.
- Kayalibay, B.; Jensen, G.; and Patrick, V. 2017. Cnn-based segmentation of medical imaging data.
- Krizhevsky, A.; Sutskever, I.; and Hinton, G. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* 25(2).
- Li, X.; Hao, C.; Qi, X.; Qi, D.; Fu, C. W.; and Pheng-Ann, H. 2017. H-denseunet: Hybrid densely connected unet for liver and liver tumor segmentation from ct volumes. *IEEE Transactions on Medical Imaging* 1–1.
- Lin, L.; Dou, Q.; Jin, Y. M.; Zhou, G. Q.; Tang, Y. Q.; Chen, W. L.; Su, B. A.; Liu, F.; Tao, C. J.; and Jiang, N. 2019. Deep learning for automated contouring of primary tumor volumes by mri for nasopharyngeal carcinoma. *Radiology*.
- Ma, G.; Yang, J.; and Zhao, H. 2020. A coronary artery segmentation method based on region growing with variable sector search area. *Technology and health care: official journal of the European Society for Engineering and Medicine* 28(2):1–10.
- Milletari, F.; Navab, N.; and Ahmadi, S. A. 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)*.
- Oktay, O.; Schlemper, J.; Folgoc, L. L.; Lee, M.; Heinrich, M.; Misawa, K.; Mori, K.; McDonagh, S.; Hammerla, N. Y.; and Kainz, B. 2018. Attention u-net: Learning where to look for the pancreas.
- Qin, Y.; Kamnitsas, K.; Ancha, S.; Nanavati, J.; and Nori, A. 2018. Autofocus layer for semantic segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*.
- Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*.
- Roy, A. G.; NavAb, N.; and Wachinger, C. 2018. Concurrent spatial and channel squeeze excitation in fully convolutional networks. *Springer, Cham*.
- Shelhamer, E.; Long, J.; and Darrell, T. 2016. Fully convolutional networks for semantic segmentation.
- Simonyan, K., and Zisserman, A. 2015. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*.
- Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is all you need. In *Neural Information Processing Systems*.
- Wang, R.; Li, C.; Wang, J.; Wei, X.; Li, Y.; Zhu, Y.; and Zhang, S. 2015. Threshold segmentation algorithm for automatic extraction of cerebral vessels from brain magnetic resonance angiography images. *Journal of Neuroscience Methods* 241:30–36.
- Xie, E.; Wang, W.; Wang, W.; Sun, P.; Xu, H.; Liang, D.; and Luo, P. 2021a. Segmenting transparent object in the wild with transformer. *arXiv: Computer Vision and Pattern Recognition*.
- Xie, Y.; Zhang, J.; Shen, C.; and Xia, Y. 2021b. Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation.
- Zhao, H.; Shi, J.; Qi, X.; Wang, X.; and Jia, J. 2016. Pyramid scene parsing network. *arXiv: Computer Vision and Pattern Recognition*.
- Zhou, H. Y.; Guo, J.; Zhang, Y.; Yu, L.; and Yu, Y. 2021. nnformer: Interleaved transformer for volumetric segmentation.
- Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; and Dai, J. 2021. Deformable detr: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*.