A Coarse to Fine Framework for Multi-organ Segmentation in Head and Neck Images

Yan Pu
Graduate School of Information,
Production and System
Waseda University
Kitakyushu, Japan
7_yan9@toki.waseda.jp

Sei-ichiro Kamata
Graduate School of Information,
Production and System
Waseda University
Kitakyushu, Japan
kam@waseda.jp

Youjie Wang
Graduate School of Information,
Production and System
Waseda University
Kitakyushu, Japan
wang-youjie@fuji.waseda.jp

***Abstract*—Radiotherapy is widely used in the treatment of head and neck cancer. Due to the harmfulness of radiation, it is necessary to protect our healthy organs during the radiotherapy. Therefore, the accurate delineation of diseased region and surrounding healthy organs is the precondition for doctors to make the radiation plan. In real life, the delineation work is usually done manually. It is time-consuming and requires high professional skill. A fast and accurate organ segmentation method can greatly improve the efficiency of treatment. Most CT image datasets are 3D volumes and each volume can be divided into a series of 2D slice images. For multi-organ segmentation task, how to generate the stable organ features from CT images is still the plagued problem. For 2D framework, which processes the images slice by slice, the network cannot learn the correlation between continuous slices. It will lead to the loss of spatial information. For 3D framework, which processes the images volume by volume, the patch training is commonly used to against the massive increase of network parameters. The 3D patch will limit the maximum reception field of the network. For the organ, which is larger than the patch size, it is easy to lose global information. To solve these incompatible problems, we proposed a coarse to fine framework to take advantage of both 2D framework and 3D framework. The multi-view coarse network is designed to generate the organ probability maps and the coarse segmentation mask in 2D case. The organ volumes are extracted with the probability maps. These organ volumes are sent to the organ-based fine network to refine the mask of each organ in 3D case. Our proposed method is tested on the *Head and Neck Automatic Segmentation Challenge* datasets in 2015 and predict for 9 different organs. The result show that our framework performs the lowest error range for most organs and three of them achieve the top evaluation results in comparison with existing methods.**

***Contribution*—The main contribution of this paper is to propose a novel two-stage, Coarse to Fine, framework for multi-organ segmentation and verify its effectiveness in head and neck CT images.**

***Index Terms*—Organs segmentation, CT images, Head and neck**

I. INTRODUCTION

Cancer has long plagued humans with high death rates and widespread morbidity. According to the report [1] of the *World Health Organization*(WHO), more than 9 million people died of cancer in 2019. The head and neck contain a large number of tissues and organs, which are particularly prone to cancer in our bodies. In recent years, the high precision radiotherapy has become a common way for cancer treatment. Radiotherapy uses radiation to kill diseased tissues while minimizing the damage to healthy tissues. Therefore, the clinicians must accurately segment the *Brain Stem*, *Cross-optic Nerve*, *Parotid Gland* and other organs at risk(OARs) in advance. The multi-organs segmentation is a kind of instance segmentation problem. Need to find the edge profile of the targets. Our target is the above-mentioned OARs in head and neck part.

Due to the *Computed Tomography*(CT) images has spatial accuracy and high resolution and ability of 3D reconstruction. CT image can clearly show the structure of our body and provide information about the spatial density and the accurate position of structures. Therefore, it is the mainstream of clinical treatment to make radiotherapy plan based on CT images. In reality, the clinicians always manually segment the OARs layer by layer on the CT images, which is time-consuming and requires high professional skill. Manual segmentation may lead to the incorrect and neglectful situations. The segmentation result mainly depend on the subjective judgement of doctors. If an approach can accurately segment the target organs, it could save a lot of time and human cost. For this target, there are corresponding contests and challenges every year. For example, the head and neck auto segmentation challenge held in conjunction with the *Medical Image Computing and Computed-Assisted Intervention*(MICCAI) [2]. The research in this paper is based on the published data set on this challenge. The visualization of CT images and organ masks are shown in Figure 1.

In earlier years, deep learning algorithms have not shown the excellent performance in medical field. Automatic segmentation of head and neck OARs is mainly based on statistical shape modeling and atlas matching [3][4]. Karl et all [5] used the statistical appearance models and multi-atlas to achieve automatic segmentation. Haworth et all [6] based on the multi-atlas matching method, achieved the top position in the MICCAI 2015 head and neck organ segmentation challenge. The positions of the internal organs and the overall contour of the human body are relatively similar. Naturally, we can use the previously segmented data as a template. The test image and the template image are deformed and matched, and then the labels on the template are correspondingly labeled on the test

978-1-7281-9331-1/20/$31.00 ©2020 IEEE

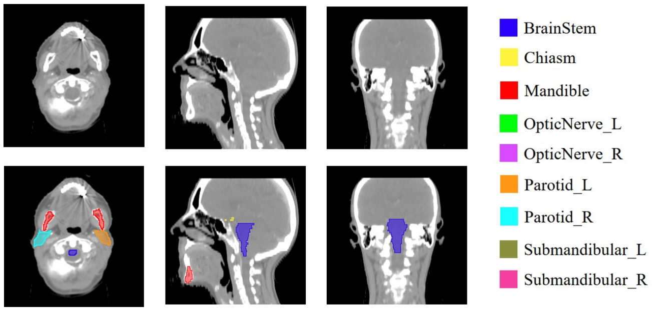

Fig. 1. The example of head and neck CT images in three perspectives. The first row is the original CT image. The OARs are shown on the bottom row and visualized with different color.

image. This method is more effective for smaller anisotropic organs, but the large differences in organ morphology may occur between patients. The similarity between test images and template images is low, which will decrease the accuracy of atlas-based methods. With the rapid development of deep learning methods, convolutional neural networks(CNN) begin to shows the excellent performance in medical imaging field. In 2015, fully convolutional networks(FCN) was first proposed by Jonathan[7], which became a milestone in the field of semantic segmentation. Under the excitation of FCN, a series of pixel-level segmentation algorithms were proposed. In the same year, Olaf et all[8] proposed a new end-to-end network structure named U-Net, which can achieve efficient segmentation results with very few data. Once proposed, it was widely applied to the medical field and derived a large number of improved algorithms based on U-Net. Bulat et all[9] first tried CNN method to solve the multi-organ segmentation problem, they designed a 2D structure and achieved state-of-the-art result at that time. For the head and neck segmentation, there are the following difficulties:

1) There is a huge imbalance between the background and the target organ pixels, which is shown in Figure 2. In each patient sample, the target organ pixels only occupy about 0.2% - 0.5%. Therefore, it is hard for network to learn and focus on the organ features in such an imbalanced data. This problem will lead to the low accuracy of organ segmentation, especially for small organs. Some small organs, like chiasm and optic nerves, only appear in 3-5 layers. Due to the different CT scanning technology, the space interval between layer and layer is different, which will also affect the performance of network. Therefore, the preprocessing of CT images need to design carefully.

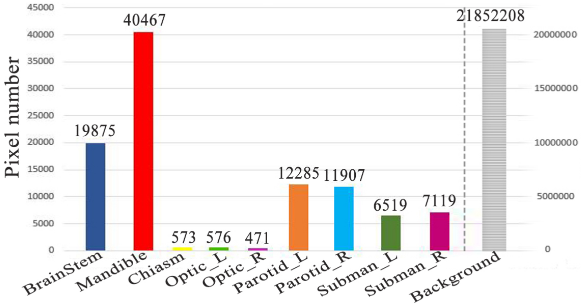

Fig. 2. The pixel distribution of one patient sample. The abscissas are 9 target OARs and the background. The ordinate is the number of pixel and the grid lines on both sides are different measurement standard.

2) It is difficult to use the spatial information in three-dimensional data. For 2D segmentation networks, the input is the two-dimensional slices, and the slices are discontinuous and unrelated. The network can not learn the relationship between layers. For 3D segmentation networks, the massive increase in the amount of parameters led to the use of batch training. The purpose is to reduce computational memory usage. The original volume needs to be cropped into a series of 3D patches for training. The 3D patch will limit the maximum reception field of the network. For the organ, which is bigger than patch size, it will loss global information during the learning. And the training of 3d networks requires more data, otherwise it is easy to cause overfitting.

II. RELATED WORK

A. *End-to-end approach*

The end-to-end approach mainly based on the encoder-decoder structure, which contains all the processing procedure. The network receives the input image, and the final output is the target we need, such as category labels or segmentation masks. U-Net[8] and 3D U-Net[10] are typical end-to-end methods. The completely symmetrical design keeps the size of the input and output same. Through skip-connection, high-resolution information and low-resolution information are effectively combined. In medical images, the organs boundary is blurred, and the gradient is complex. Requiring more high-resolution information for accurate segmentation. In human body, the internal structure is relatively fixed and the semantics are simple, so low-resolution information can be used for structural recognition. Therefore U-Net has excellent performance in medical images. FocusNet[11] integrates the main organ segmentation subnetwork, small organ locating subnetwork, and small organ segmentation subnetwork. On the premise of ensuring high segmentation accuracy of large organs, the network has improved the segmentation accuracy of small organs. At the same time, the DenseASPP[12] and SE-block[13] have been added to improve the U-Net structure.

B. *Two-stage approach*

The tweo-stage approach mainly contain more than one network. The task of segmentation is divided into several subtasks. Wang et all[14] proposed a two-stage method by using LocNet to locate the organs return the bounding box of target organ. Then crop the interest volume from original data. These interest volumes are used to train the SegNet, and then final return the segmentation mask. The two subnetwork is implemented base on 3D U-Net. Tong et al [15] also proposed a two-stage structure. First, the network is pre-trained by shape

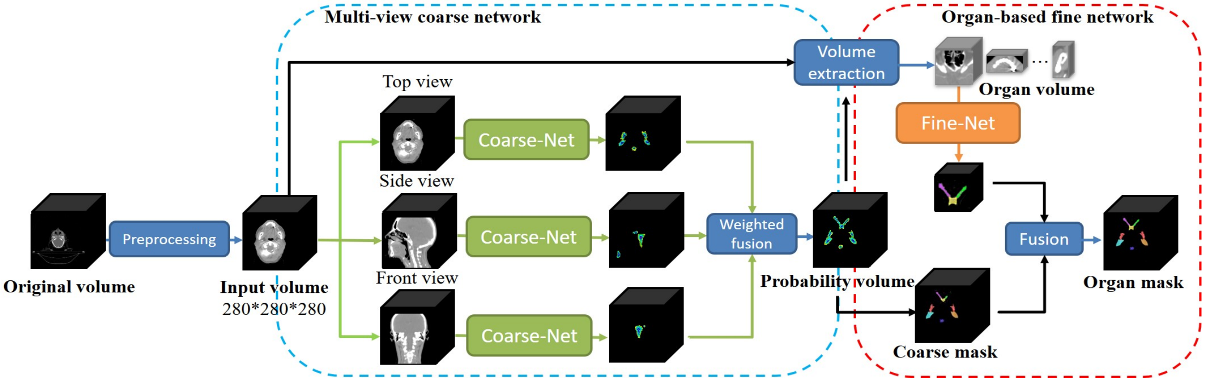

Fig. 3. Overall structure of our two-stage framework.

representation model(SRM), the purpose is to learn the shape information of the organs in advance. Then, segment with fully convolutional neural network(FCNN). Ren et all[16] use a multi-atlas method to segment coarsely and proposed an interleaved 3D-CNN to focus on three small organs. Compare with end-to-end methods, the two-stage approach is more targeted and flexible.

III. PROPOSED METHOD

Our proposed method is designed to a two-stage approach. First, the multi-view network is used to do coarse segmentation and generate organ distribution probability volume. With the probability information, we crop the proposal volumes for each organ. Then, the organ volumes were sent to fine segmentation network. The masks of all organ are fused with coarse mask to generate final result. The overall structure is given in Figure 3. The general architecture of coarse network and fine network is shown in Figure 4.

A. Multi-view coarse network

Coarse network. The coarse network is based on U-Net structure. We retain the skip connection in the original structure, in order to effectively utilize the feature information of different scales. The multiple pooling operations can lead to the loss of small organ feature. In this case, we only perform it with only three pooling operations. The usage of SE-block is to strengthen the proportion of organ features. SE-block is a novel unit to adaptively rejudge the channel-wise features. It improves the network by modeling the correlation between feature channels and enhancing the important features. Three different perspective networks are trained to generate different probability map separately.

Weighted fusion. In order to better integrate the output of the three perspective models, we use the weighted supervised fusion. The fused probability volume contains 10 channels, and the pixel value of different channels represents the probability that this pixel belongs to 10 categories. Then a softmax function can be used to achieve the target of segmentation. The probability map also reflects the distribution information of each organ. For different organs, find their corresponding feature channels. Then, through the pooling operation, you can get the position information of each organ region. Then combine the input volume to crop, you can get interest blocks containing different organs. These interest blocks are used as input to the fine network.

Loss function. For each coarse network we use a weighted cross-entropy as loss funtion, which is

$$WCE = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M}\left(\omega_j y_{ij} log\left(\hat{y_{ij}}\right)\right) \tag{1}$$

N is the pixel number and M is number of categories. ω_j is the weight of each category. y_{ij} represent the category label of pixel i. $\hat{y_{ij}}$ is the predict probability for category j. ω_j is the weight parameter, which is used to increases the contribution of small categories to the loss function. The weight parameters are calculated by the following formula.

$$\omega_j = \frac{N - \sum_i^N \hat{y_{ij}}}{\sum_i^N \hat{y_{ij}}} \tag{2}$$

B. Organ-based fine network

Fine network. For organ blocks of different sizes, we further use 3D segmentation networks to learn information between layers. This makes up for the loss of spatial information caused by the previous use of 2D networks. The fine network is a 3D version of the coarse network. The kernel of convolution and pooling that becomes 3 * 3 * 3 and 2 * 2 * 2.

Target volume extraction. With the 10 channels probability volume, we can get 10 binary mask volumes by sending each channel to the classifier. For each mask volume, we can generate a bounding block by evaluating the score value. The target bounding block need to have the smallest volume and the highest score. Then, we can crop the organ volume from the original CT image.

Loss function The 3D fine network is training with a generalized dice loss. It also include weight parameter, focus on solving imbalance problem between categories.

$$L_{gd} = 1 - \frac{1}{M}\frac{2\sum_{j=1}^{M}\omega_j\sum_{i=1}^{N} y_{ij}\hat{y_{ij}}}{\sum_{j=1}^{M}\omega_j\sum_{i=1}^{N}\left(y_{ij} + \hat{y_{ij}}\right)} \tag{3}$$

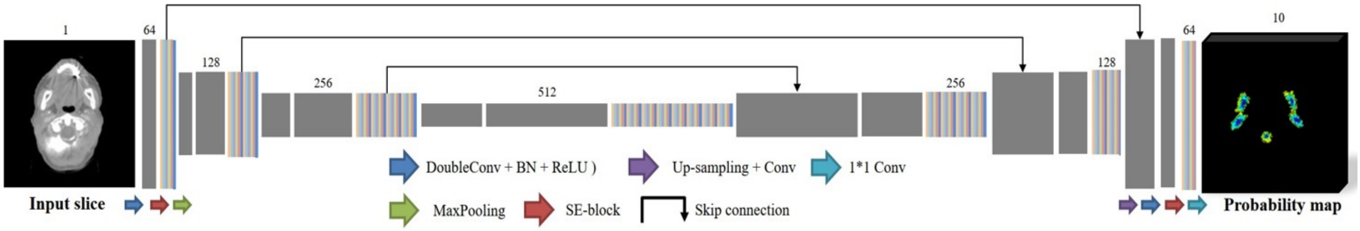

Fig. 4. The general segmentation structure of 2D coarse network and 3D fine network. We removed the final classify layer and the output is the probility map of organ features.

$$\omega_j = \frac{1}{\left(\sum_{i=1}^{N} y_{ij}\right)^2} \tag{4}$$

The formula is based on dice coefficient, which is designed to measure the similarity between the network output and the target.

IV. EXPERIMENTS

A. Data preprocessing

Window adjust and resample. Through statistics, it is found that the CT image values are varies from -1024 to 3000. Nowadays the initialization of network parameters generally uses Gaussian random initialization. Therefore, when the input data distribute close to Gaussian distribution, then the convergence speed of the network will be improve. Through experiments, it is found that using the window width of 1200 and window level of 200 to adjust original CT images can achieve better performance on segmentation. The are four examples of different values range, which are shown in Figure 5.

In order to eliminate the influence of the scale of the data collected by different CT scanning equipment, the data needs to be normalized, which is

$$I_{norm} = (I + 400) / (400 + 800) \tag{5}$$

Where, I_{n_out} means the normalized data. In order to make the input approximately follow the Gaussian distribution, the normalized data needs to be subtracted from the mean of the entire data set, which is

$$I_{mean} = \frac{1}{N} \sum_{j=1}^{N} I_{norm}^{j} \tag{6}$$

$$I = I_{norm} - I_{mean} \tag{7}$$

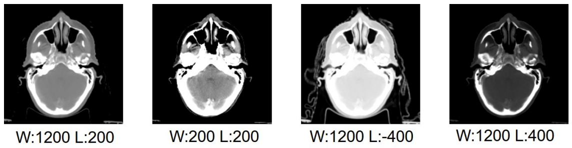

Fig. 5. The examples of different window width(ww) and window level(wl). We select the first set of parameters to adjust.

Where, I_{mean} means the average value of the entire data set and I is the results after data normalization. In addition, the data sets have different pixel pitches in the x, y, and z directions. In order to show good generalization capabilities of deep neural network, the images in the data set need to be resampled so that the pixel spacing of all images in the x and y directions is consistent. Data resampling is to scale the data by calculating a scaling factor, which is

$$F_{zoom} = S_{old} / S_{new} \tag{8}$$

Where, F_{zoom}, $S_{old}$and S_{new} means the resampling scaling factor, the old pixel pitch and the new pixel pitch. In the experiment, taking $S_{new} = 1$, that is, after resampling, the pixel pitch of the image in the x and y directions becomes 1 mm.

Crop. The shape of a CT volume is generally 512*512*slice, but after data resampling, the size of the CT image becomes inconsistent. For better network training, the data was cropped to the fix shape of 280*280*280. In this process, it is necessary to ensure that the entire head and neck image is included in the slices in all three directions.

B. Evaluate Criterion

In order to accurately evaluate the segmentation results, this article uses two evaluation indexes, Dice Similarity Coefficient and 95% Hausdorff Distance, to evaluate the segmentation results.

1) Dice Similarity Coefficient: Dice similarity coefficient (DSC) is a function that gathers similarity measures and is often used to judge the statistics of image segmentation quality. For a foreground class, DSC calculates the spatial overlap ratio of the area occupied by the class in the prediction result and Ground Truth, which is defined as

$$DSC = \frac{2\left|X \cap Y\right|}{\left|X\right| + \left|Y\right|} = \frac{2TP}{2TP + FP + FN} \tag{9}$$

Where, $|X|$ and $|Y|$ respectively represent the area occupied by the foreground class in the prediction result and Ground Truth.$|X \cap Y|$ represents the overlapping area of the two sets. DSC has a value range of $[0, 1]$, where 0 means that there is no overlap between the two sets, and 1 means that the two sets are completely overlapping.

2) 95% Hausdorff Distance: For two point sets $A = \{a_1, a_2, a_3...\}$ and $B = \{b_1, b_2, b_3...\}$ in Euclidean space, the Hausdorff Distance(HD), which is used to measure the distance between the two point sets of A and B, is defined as

$$H(A,B) = max\left(h(A,B), h(B,A)\right) \tag{10}$$

Where, $h(A,B) = \underset{a\in A}{max}\underset{b\in B}{min}\|a-b\|$, $h(B,A) = \underset{b\in B}{max}\underset{a\in A}{min}\|b-a\|$. $H(A,B)$ means the two-way HD between the point sets and $h(A,B)$ means the unidirectional HD from point set A to point set B. Similarly, $h(B,A)$ means the unidirectional HD from point set B to point set A. Distance is measured using Euclidean distance. For 3D point sets such as head and neck endangering organs, HD is calculating the maximum surface distance between the two regions. To obtain the Maximum Surface Distance(MSD), HD can be calculated with the organ segmentation results and the organ mask delineated by experts. The smaller MSD value,whcih means the two sets are closer, the better segmentation result. However, since HD is extremely susceptible to noise, if there is a discrete point that is far away from the data, the distance between this discrete point and the set becomes the farthest distance. Therefore, the distance between the two sets cannot be accurately measured. In the head and neck endangered organ segmentation, this kind of discrete mis-segmentation point does sometimes exist. In order to better evaluate the results, 95% Hausdorff distance (95HD) is introduced as

$$H_{95}(A,B) = max\left(h_{95}(A,B), h_{95}(B,A)\right) \tag{11}$$

Where, $h_{95}(A,B) = K_{95}\left(\underset{b\in B}{min}\|a-b\|\right)\forall a \in A$, $H_{95}(A,B)$ means the 95HD between two point sets, $h(A,B)$ means the distance from one point in set A to the set B. This distance should no less than the distance of remaining 95% of points in set A to the set B. $h(B,A)$ is the same meaning. K_{95} represents the top 5% value in the set. The smaller value of K_{95} means the higher similarity and better segmentation performance. The unit here is millimeters, and the pixel spacing of the CT sequence needs to be considered.

TABLE I
THE DSC SCORE OF EACH PERSPECTIVE IN MULTI-VIEW COARSE NETWORK.

Organs	Front view	Top view	Side view	Fusion
Brain Stem	81.2±2.6	84.3±1.8	80.4±1.7	**85.2±1.8**
Mandible	85.6±2.2	89.5±1.4	88.5±2.8	**90.1±2.0**
Chiasm	39.8±14.0	**43.1±15.2**	41.3±13.7	42.0±13.4
OpticNerve_L	62.4±3.8	**66.9±4.5**	61.8±4.1	65.4±4.4
OpticNerve_R	62.2±3.2	**66.6±5.1**	61.5±5.1	65.6±3.8
Parotid_L	76.0±2.1	79.2±2.4	76.4±2.6	**80.2±2.6**
Parotid_R	76.4±1.8	80.1±1.8	77.0±1.2	**81.0±2.5**
Submandibular_L	69.5±14.1	71.2±10.2	68.2±13.2	**72.2±9.3**
Submandibular_R	70.0±10.2	72.6±8.0	69.8±9.6	**74.0±8.6**

C. Implementation detials

All experiments are work on a workstation with GTX 1080Ti GPU under CUDA 10.0. The deep learning structure is built on *PyTorch*. Our two-stage method need to be trained separately. The 280*280*280 CT volume was cut to slices in three-view. All the slices is shuffled and composed to train set randomly. We also used translation and rotation to enhance the training set. The label images should do the same transformation as CT images. I start our training with 16 batch size and 0.0001 learning rate. We train each coarse network for 100 epochs. Every 10 epochs, we evaluate and save the best performance model. The three different perspective networks were trained in same way. Before fusion process, each probability map was send to a softmax classifier and evaluate with dice coefficient. The contrast result is shown in Table I.

After the extraction of organ volume. The 3D fine network start to train with these CT volumes. To offset the increase of parameters in 3D network training, we reduce batch size to 4. The trained network is used to predict 9 organs on the testing dataset.

TABLE II
THE PERFORMANCE OF DSC COMPARE WITH EXISITING METHODS(%)

Organs	U-Net[8]	TS 3D U-Net[14]	SRM+FCNN[15]	Ours
Brain Stem	82.2±1.7	87.5±2.2	**87.0±3.0**	88.0±1.4
Mandible	90.5±1.5	93.0±1.9	93.6±1.2	**94.0±1.8**
Chiasm	38.4±14.3	45.1±17.2	58.4±10.3	**61.2±9.8**
OpticNerve_L	64.5±7.2	73.7±7.6	65.3±5.8	**74.3±8.0**
OpticNerve_R	65.0±7.8	73.6±8.8	68.9±4.7	**76.8±7.2**
Parotid_L	77.2±6.5	86.4±2.6	83.9±2.9	**86.0±3.5**
Parotid_R	78.1±5.4	**84.8±7.0**	83.5±2.3	86.5±2.8
Submandibular_L	69.3±14.5	**75.8±14.7**	76.7±7.4	78.8±8.1
Submandibular_R	73.2±9.6	73.3±9.7	**81.1±6.5**	80.2±5.8

TABLE III
THE PERFORMANCE OF 95HD COMPARE WITH EXISITING METHODS(MM)

Organs	U-Net[8]	TS 3D U-Net[14]	SRM+FCNN[15]	Ours
Brain Stem	4.42±1.01	**2.01±0.33**	4.01±0.93	1.94±0.22
Mandible	1.90±0.35	1.26±0.50	1.50±.032	**1.18±0.45**
Chiasm	2.71±1.24	2.83±1.42	**2.17±1.04**	2.62±1.35
OpticNerve_L	2.90±2.20	2.53±2.34	2.52±1.04	**2.40±2.25**
OpticNerve_R	2.71±1.65	2.13±2.45	2.90±1.88	**2.01±2.35**
Parotid_L	4.16±1.61	**2.41±0.54**	3.97±2.15	2.35±0.44
Parotid_R	4.27±2.31	**2.93±1.48**	4.20±1.27	2.86±1.20
Submandibular_L	6.05±3.22	2.86±1.60	5.59±3.93	**3.05±2.12**
Submandibular_R	5.14±2.40	3.44±1.55	4.84±1.67	**3.12±1.41**

D. Result analysis

The multi-view results are shown in I, only chiasm and two optic nerves show best score in top view result instead of the fusion result, which means some organs feature are sensitive in some specific perspective. For most organs, the higher score proves the validity of model fusion.For comparison purpose, Table II and TableIII show the DSC and 95HD scores. The comparisons follow the same testing protocal. According to

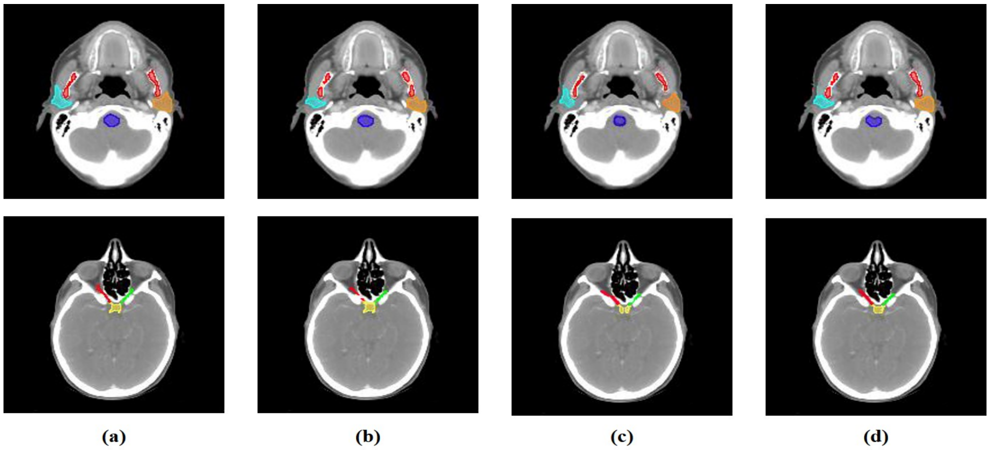

Fig. 6. The segmentation result of following methods:(a) Ground truth. (b) Two-stage 3D U-Net. (c) SRM+FCNN. (d) Our method. The different colors denote different target organs.

Table II and TableIII, our proposed method achieves the top performance on most organs. Especially for the mandible and two optic nerves, they get the best score on both DSC and 95HD. For DSC result, the scoring error of chiasm and brain stem become small, which means the predicition for these two organs are more stable. For 95HD result, there are also 5 organs perform better than other method.

V. Conclusion and Future Work

In this paper, we proposed a two-stage segmentation framework. Mainly focus on the head and neck CT images, we solve a part of the low performance problem on small organ by a fine segmentation network based on each organs. We also designed a weighted fusion process to combine the multi-view probability maps, which are used for coarse mask segmentation and organ volume extraction. For the imbalanced problem between organs and background, we design the specific data preprocessing and apply the SE-block to enhance the related feature channels. We also reduce the error range for most organs. The evaluation results prove the effectiveness of our design. The future work should further develop and verify the adaptivity of our framework. We merely focus and solve the specific problems in head and neck CT images. More contrast experiments need to be done on different medical datasets. We believe the hybird framework, which can fuse the advantages of 2D and 3D framework, is a novel research aspect for medical segmentation task.

References

[1] W. H. Organization, *World health statistics 2019: monitoring health for the SDGs, sustainable development goals*, World Health Organization, 2019.

[2] P. F. Raudaschl, P. Zaffino, G. C. Sharp, M. F. Spadea, A. Chen, B. M. Dawant, T. Albrecht, T. Gass, C. Langguth, M. Lüthi, *et al.*, "Evaluation of segmentation methods on head and neck ct: auto-segmentation challenge 2015," *Medical physics* **44**(5), pp. 2020–2036, 2017.

[3] X. Han, L. S. Hibbard, N. P. O'Connell, and V. Willcut, "Automatic segmentation of parotids in head and neck ct images using multi-atlas fusion," *Medical Image Analysis for the Clinic: A Grand Challenge* , pp. 297–304, 2010.

[4] J. E. Iglesias and M. R. Sabuncu, "Multi-atlas segmentation of biomedical images: a survey," *Medical image analysis* **24**(1), pp. 205–219, 2015.

[5] K. D. Fritscher, M. Peroni, P. Zaffino, M. F. Spadea, R. Schubert, and G. Sharp, "Automatic segmentation of head and neck ct images for radiotherapy treatment planning using multiple atlases, statistical appearance models, and geodesic active contours," *Medical physics* **41**(5), p. 051910, 2014.

[6] R. Mannion-Haworth, M. Bowes, A. Ashman, G. Guillard, A. Brett, and G. Vincent, "Fully automatic segmentation of head and neck organs using active appearance models," *MIDAS J* , 2015.

[7] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, 2015.

[8] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241, Springer, 2015.

[9] B. Ibragimov and L. Xing, "Segmentation of organs-at-risks in head and neck ct images using convolutional neural networks," *Medical physics* **44**(2), pp. 547–557, 2017.

[10] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: learning dense volumetric segmentation from sparse annotation," in *International conference on medical image computing and computer-assisted intervention*, pp. 424–432, Springer, 2016.

[11] Y. Gao, R. Huang, M. Chen, Z. Wang, J. Deng, Y. Chen, Y. Yang, J. Zhang, C. Tao, and H. Li, "Focusnet: Imbalanced large and small organ segmentation with an end-to-end deep neural network for head and neck ct images," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 829–838, Springer, 2019.

[12] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Yang, "Denseaspp for semantic segmentation in street scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3684–3692, 2018.

[13] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141, 2018.

[14] Y. Wang, L. Zhao, M. Wang, and Z. Song, "Organ at risk segmentation in head and neck ct images using a two-stage segmentation framework based on 3d u-net," *IEEE Access* **7**, pp. 144591–144602, 2019.

[15] N. Tong, S. Gou, S. Yang, D. Ruan, and K. Sheng, "Fully automatic multi-organ segmentation for head and neck cancer radiotherapy using shape representation model constrained fully convolutional neural networks," *Medical physics* **45**(10), pp. 4558–4567, 2018.

[16] X. Ren, L. Xiang, D. Nie, Y. Shao, H. Zhang, D. Shen, and Q. Wang, "Interleaved 3d-cnn s for joint segmentation of small-volume structures in head and neck ct images," *Medical physics* **45**(5), pp. 2063–2075, 2018.