Neurocomputing xxx (xxxx) xxx

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Cascaded SE-ResUnet for segmentation of thoracic organs at risk

Zheng Cao [a,b], Bohan Yu [a,b], Biwen Lei [a,b], Haochao Ying [b,c,*], Xiao Zhang [d], Danny Z. Chen [e], Jian Wu [b,c,*]

[a] College of Computer Science and Technology, Zhejiang University, China
[b] Real Doctor AI Research Centre, Zhejiang University, China
[c] School of Public Health, Zhejiang University, China
[d] School of Computer Science and Technology, Shandong University, China
[e] Department of Computer Science and Engineering, University of Notre Dame, United States

ARTICLE INFO

Article history:
Received 1 December 2019
Revised 25 May 2020
Accepted 21 August 2020
Available online xxxx

Keywords:
Organs at risk
SE-ResUnet
Segmentation

ABSTRACT

Computed Tomography (CT) has been widely used in the planning of radiation therapy, which is one of the most effective clinical lung cancer treatment options. Accurate segmentation of organs at risk (OARs) in thoracic CT images is a key step for radiotherapy planning to prevent healthy organs from getting over irradiation. However, known automatic image segmentation methods can hardly yield desired OAR delineation results, while manual delineation tends to take long time and tedious effort. In this paper, we propose a novel deep learning network, called cascaded SE-ResUnet, for automatic segmentation of thoracic organs including left lung, right lung, heart, esophagus, trachea, and spinal cord. Specifically, we first use a coarse segmentation network to identify the regions of interest (ROIs), and then a fine segmentation network is applied to achieve refined segmentation results, organ by organ. Finally, different configured models are ensembled to obtain the final segmentation results. In the StructSeg 2019 Challenge, we showed the capability of our new framework and won the 1st place at the test phase. Our code is available open-source at https://github.com/zjuybh/StructSeg2019.

1. Introduction

Radiation therapy is a common option for treating lung cancer, which can efficiently destroy cancer cells with exterior radiation beams. Developing an effective treatment plan is vital for radiation therapy, which can determine the appropriate distribution of radiation dose in target tumors and nearby organs, called organs-at-risk (OARs). On one hand, tumor cells should receive enough radiation dose to be completely killed. On the other hand, OARs usually contain much sensitive organs with respect to radiation, which must be protected during radiation therapy to avoid extra damage to the patient. For example, the patient runs a high risk of blindness if the optical nerves or chiasma receive too much radiation [1].

Hence, accurate segmentation of OARs is of great significance in radiotherapy treatment planning. In the past, radiologists need to manually delineate the target OARs, which is a time-consuming process and suffers large subjective differences. In addition, due to large location and contour variability, small organ sizes, and low contrast in CT scans, some organs such as esophagus and trachea are very difficult to delineate. It will be highly meaningful to reduce the treatment planning time and overall cost of radiation therapy if a computer-aided automatic delineation system of OARs is developed.

As a common disease, lung cancer has a high mortality, for which radiation treatment is a proper choice. In thoracic CT scans, there are several OARs such as the heart, esophagus, trachea, spinal cord, and left/right lungs, which need to be delineated during treatment planning. Thoracic CT scans contain both larger OARs like the heart and smaller OARs like esophagus which is hardly distinguishable, as shown in Fig. 1.

In this paper, to attain accurate segmentation of each organ-at-risk in thoracic CT scans, we propose a new deep learning network called Cascaded SE-ResUnet. First, we design a basic segmentation network SE-ResUnet, inspired by the structure of U-Net [2]. Specifically, residual paths [3] are applied to encourage smoother gradient flow, while squeeze-and-excitation blocks [4] are employed to serve as an attention mechanism for feature weighting to obtain better feature representation. Moreover, our overall framework uses six SE-ResUnet networks to perform coarse localization and refined segmentation of OARs, and thus it is called cascaded.

* Corresponding authors at: School of Public Health, Zhejiang University, China. (Haochao Ying and Jian Wu)

E-mail addresses: z.cao@zju.edu.cn (Z. Cao), 3140103928@zju.edu.cn (B. Yu), 21821214@zju.edu.cn (B. Lei), haochaoying@zju.edu.cn (H. Ying), xiaozhang@sdu.edu.cn (X. Zhang), dchen@nd.edu (D.Z. Chen), wujian2000@zju.edu.cn (J. Wu).

https://doi.org/10.1016/j.neucom.2020.08.086
0925-2312/

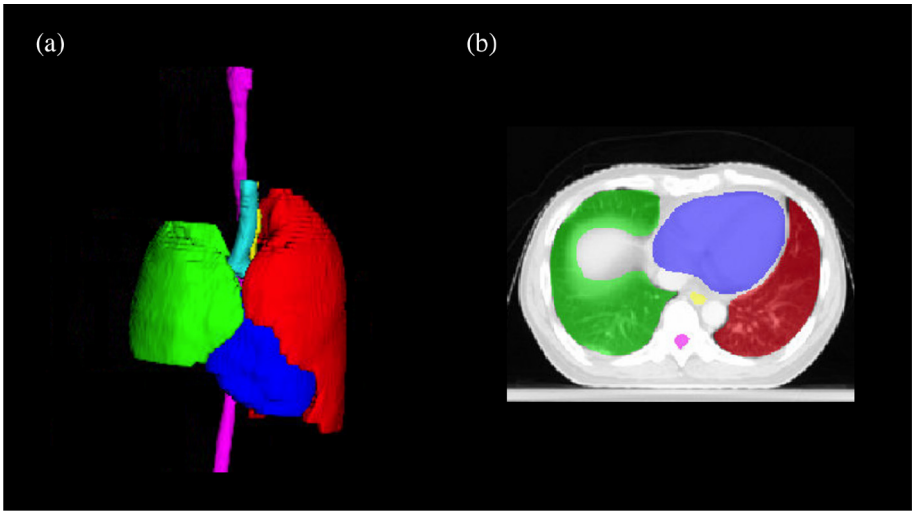

Fig. 1. A thoracic CT image example with annotation: (a) A 3D visualization; (b) a 2D slice (green: left lung; red: right lung; purple: heart; yellow: esophagus; blue: trachea; pink: spinal cord).

Finally, to ensure the coarse-to-fine structure to attain best possible performance, different window widths and window levels of CT images are selected for each SE-ResUnet network to outline the target organs.

In summary, the main contributions of this work are three folds. (1) We propose a novel deep architecture, Cascaded SE-ResUnet (C-SE-ResUnet), for segmenting OARs in thoracic CT images, which adopts a coarse-to-fine strategy along with a novel SE-ResUnet block to attain accurate boundary segmentation of each organ. (2) We propose a set novel approaches such as WCE/Dice loss and adjustive CT window to enhance the robustness for the segmentation of small size organs. (3) The proposed C-SE-ResUnet achieves an average Dice Similarity Coefficient (DSC) of 91.56%, and it shows effective results on the StructSeg 2019 Challenge dataset (ranked 1/574).

The rest of this paper is organized as follows. Section 2 briefly reviews the related work. Our proposed Cascaded SE-ResUnet method is described in Section 3. Section 4 reports the experimental results. Finally, Section 5 concludes the work.

2. Related work

Many methods have been proposed for computer-assisted CT image segmentation (see [5–7] for comprehensive reviews of such methods). In thoracic CT Organs-at-risk segmentation, traditionally there are three common categories of methods for automatic image segmentation: thresholding-based methods, region-based methods, and learning-based methods. Thresholding-based methods are easy to implement which need to only set a threshold (e.g., for the gray level of images), but this is also quite rough due to variations of image characteristics [8]. Region-based methods can perform well in noisy images but may be computationally expensive and inaccurate [9]. Learning-based methods include diverse statistical approaches such as clustering, regression, and neural networks, which can be applied and have good interpretability [10]. But, due to low contrast among organs at risk in CT images, these traditional methods often cannot attain satisfactory performance.

Since deep convolutional neural networks (CNNs) achieved a great deal of successes (e.g., ImageNet), deep learning has been widely applied to computer vision tasks such as image classification and object detection [11]. Also, more and more deep learning based methods were developed for image segmentation. In particular, fully connected networks (FCNs) can perform end-to-end segmentation and be effective in diverse imaging applications (e.g., semantic segmentation [12,13], video object detection [14,15]). However, due to its fully connected structure, FCN uses plenty of parameters, thus incurring difficulties in model training. SegNet [16] was presented with an encoder-decoder architecture for accelerating the training process. Based on FCNs and SegNet, an improved network, U-Net [2], employed an encoder-decoder architecture and used skip connections between the up-sampling and down-sampling layers to combine high-resolution features with the up-sampled output. Some variants of U-Net have also been proposed to enhance performance, such as 3D U-Net [17], V-Net [18], UNet++ [19], and attention U-Net [20].

Besides the known general image segmentation frameworks mentioned above, some dedicated deep learning models have also been developed. Specifically, for segmenting OARs in thoracic CT images, multiple deep learning based models have been devised. Skourt et al. used the U-Net architecture directly to segment lung [21]. Hamidian et al. proposed a 3D FCN network to detect lung nodules [22]. Moeskops et al. utilized different image modalities to train a multi-task segmentation model [23]. Trullo et al. introduced the structure of a conditional random field module as RNN into FCN [24]. Furthermore, a new segmentation model with two collaborative structures was developed [25]. Negahdar et al. applied volumetric fully CNN to 3D lung segmentation [26]. Jin et al. trained a generative adversarial network (GAN) to simulate lung delineation and used the generated images to enhance the performance of a progressive holistically nested network (P-HNN) [27].Moreover, deep learning methods have been highly successful other medical image segmentation tasks, such as segmentation of cells [28], head and neck (HaN) [29], liver [30], brain [31], and optic disk [32].

3. Materials and methods

3.1. Overview of the methodology

We propose the cascaded SE-ResUnet architecture to delineate 6 organs-at-risk (OARs) in chest CT scans. This architecture contains two parts: (1) a new deep learning network based on U-Net, called SE-ResUnet, and (2) a coarse-to-fine OAR segmentation framework, which consists of 6 SE-ResUnet nets, called cascaded SE-ResUnet. In this work, we aim to improve the performance of

the general deep learning segmentation networks, and achieve high quality segmentation results of each type of OARs in chest CT images. An overview of our proposed approach is shown in Fig. 2.

3.2. SE-ResUnet model for organ localization and segmentation

We use a 2D U-Net [2] as backbone, which has been widely used in medical image segmentation. Our proposed SE-ResUnet is a classic encoder-decoder deep learning network, as shown in Fig. 3. Its encoder part is used to extract high-level semantic information by convolution blocks and max-pooling operations. Skip connections transfer fine-grained local information before each max-pooling to the decoder part. The decoder fuses such information for accurate localization and segmentation. In the encoder part, residual path and SE-block are used to obtain better feature representation. Specifically, four continuous down blocks make up the encoder part. Each down block contains a 2×2 max pooling, two 3×3 convolution kernels, ReLU function, batch normalization, residual path, and SE-block. The number of output channels of each down block is twice of the fed input channels. The decoder architecture is similar to the inverse procedure of the encoder, except that bilinear interpolation is used to upscale the feature maps instead of downscale convolution layers in the encoder. At the end, the combined feature maps containing both high-level and low-level information are fed to the softmax layer to obtain the segmentation result.

3.3. Cascaded SE-ResUnet architecture

Due to the large size of a single 3D image (e.g., $512 \times 512 \times 150$) in our dataset and the relatively small sizes of some organs (e.g., esophagus and trachea), passing all the slices of a 3D images together to the model to segment the organs is computationally expensive and inaccurate. In this work, we adopt a coarse-to-fine strategy to resolve this issue. Specifically, two SE-ResUnets are trained separately: In the coarse phase, all the slices along the z-axis of a 3D image are fed to an SE-ResUnet to roughly localize all the organs at once, slice by slice. Then, using the coarse localization results, minimum bounding cubes of each organ's prediction are cropped. Finally, the other SE-ResUnet is trained using the slices of the cropped volume, volume of interest (VOI), to segment the organs one by one. During inference, the localization and segmentation processes are similar as in training. The difference is that the slices for the segmentation results of a single organ are stacked to 3D and the 3D results of all the organs are merged and padded to the input size to produce the final segmentation. In this Challenge, 6 organs need to be segmented: left lung, right lung, heart, esophagus, trachea, and spinal cord. We simply segment both lungs in a single SE-ResUnet model. Thus, our framework contains 6 SE-ResUnets, one for coarse localization, and other five for fine segmentation of each organ.

3.4. Loss functions

For medical image segmentation, voxel-wise cross entropy loss, Dice loss, and focal loss are commonly used loss functions [18,21,26]. Inspired by that, we design a WCE/Dice loss in C-SE-ResUnet. In the coarse localization phase, considering the large image size and contrast variability among different organs, the class weighted cross entropy (WCE) loss is employed, as defined in Eq. (1), with *class* $\in$ {Left lung, Right lung, Heart, Esophagus, Trachea, Spinal Cord}. Considering the voxel value distribution of each

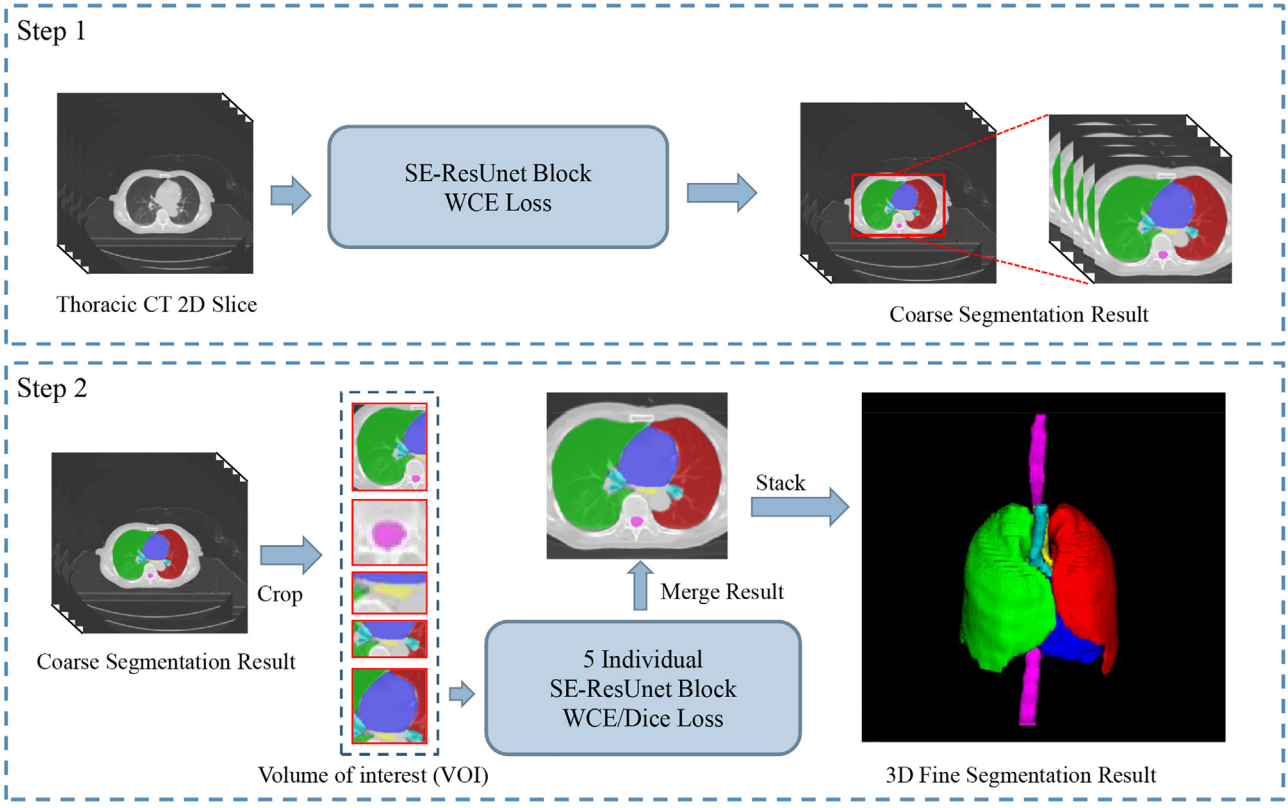

Fig. 2. Our cascaded SE-ResUnet framework for multi-organ segmentation.

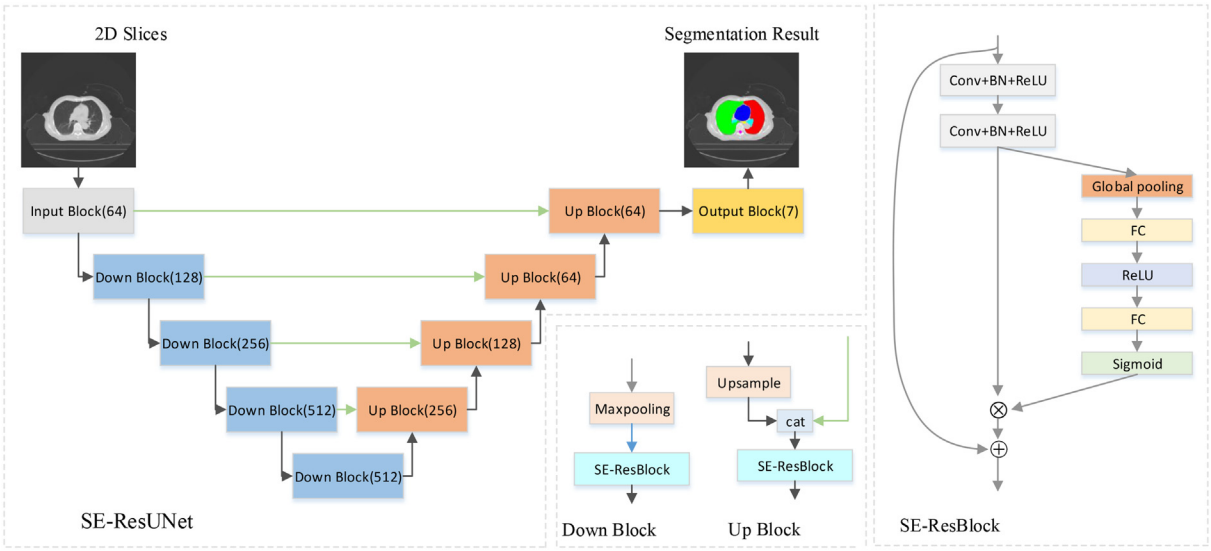

Fig. 3. An overview of the SE-ResUnet network.

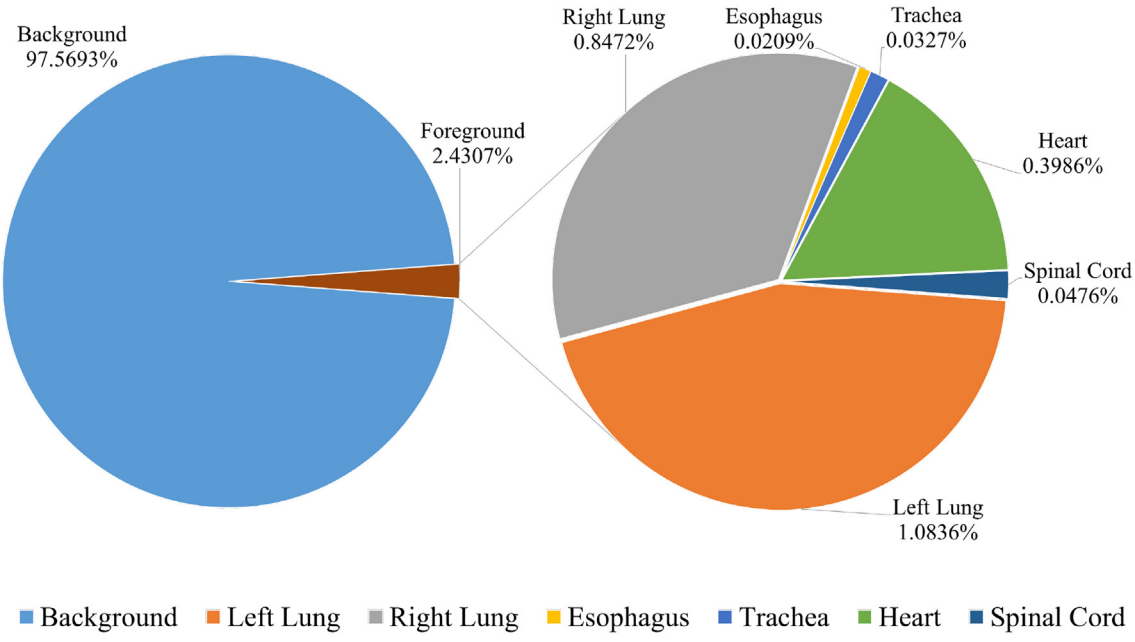

Fig. 4. The distribution of the average voxel numbers of the six organs.

Table 1
The weight value in the WCE loss for each organ.

Class	Background	Lungs	Heart	Esophagus	Trachea	Spinal Cord
Weight	1	1	1	15	5	10

organ in Fig. 4, the weight value for each class in the WCE loss is shown in Table 1.

$$\mathscr{I}_w = loss(x, class) = -weight[class] \log \frac{exp(x[class])}{\Sigma_j exp(x[j])} \quad (1)$$

Next, in the fine segmentation phase, we use different loss functions: the WCE loss for both lungs and trachea, and the Dice loss for heart, esophagus, and spinal cord. The Dice loss can accommodate minor details. For fine segmentation, since the lungs (left and right in a single model) and trachea tend to have more than one connected area in a slice, the Dice loss may be suboptimal for them. The Dice loss is defined in Eq. (2):

$$\mathscr{I}_D = loss(X, Y) = 1 - \frac{2|X \cap Y|}{|X| + |Y|} \quad (2)$$

where X and Y represent the probability map of the softmax layer output and the annotated 3D mask, respectively. Thus, Our WCE/Dice loss function can be defined in Eq. (3).

Table 2
CT window level and window width configuration.

Organs	Coarse	Left lung	Right lung	Heart	Esophagus	Trachea	Spinal Cord
Level/Hu	−500	−500	−500	30	85	−440	0
Width/Hu	1800	1800	1800	350	324	1180	600

$$\mathscr{I} = \begin{cases} \mathscr{I}_W & \text{if in the coarse segmentation phase} \\ \mathscr{I}_W & \text{if class} \in \{\text{Left Lung, Right Lung, Trachea}\} \\ \mathscr{I}_D & \text{if class} \in \{\text{Heart, Esophagus, Spinal cord}\} \end{cases} \quad (3)$$

3.5. Dataset

As one part of the StructSeg 2019 Challenge, a dataset of totally 50 whole-volume chest CT images from lung cancer patients was collected. 6 OARs of each 3D chest CT image were annotated and checked by experienced oncologists, along with their importance weights. The importance of left lung, right lung, heart, and spinal cord is all of 100% weight, while the importance of esophagus and trachea is of 70% and 80% weights, respectively. The in-plane resolution is around $1.2\ mm \times 1.2\ mm$, with the same slice thickness of $5\ mm$. The slice counts of the images range from 80 to 127, and every axial slice has a voxel dimension of 512×512. To train our deep learning model, the training dataset is split into 5 folds and the results are based on fold-5 (i.e., images 1–40 for training and images 41–50 for validation). Different organs in the CT scans have diverse distributions, as shown in Fig. 4. In spite of 97.5% background voxels, the space of two lungs takes almost 80% of the foreground. The heart accounts for 16.4% in the foreground while the other organs together are only for less than 4%.

3.6. Performance evaluation

The performance of our proposed architecture is evaluated using two evaluation metrics, DSC and 95%HD. DSC stands for Dice Similarity Coefficient, as shown in Eq. (4), where TP, FP, and FN are for true positives, false positives, and false negatives, respectively. It measures the similarity between the predicted and annotated labels.

$$DSC = \frac{2TP}{2TP + FN + FP} \quad (4)$$

95%HD represents the maximum Hausdorff distance from the annotated boundary to the prediction boundary, as shown in Eq. (5), where X and Y represent a predicted 3D mask and an annotated 3D mask, respectively. It measures the difference between the predicted and annotated labels.

$$\begin{aligned} d_H(X,Y) &= max\{d_{XY}, d_{YX}\} \\ &= \max\{\max_{x\in X}\min_{y\in Y}(x,y), \max_{y\in Y}\min_{x\in X}(y,x)\} \end{aligned} \quad (5)$$

4. Experiments and results

4.1. Pre-processing and post-processing

4.1.1. Adjustment of CT window level and window width

In CT scans, the Hounsfield scale is used in quantitative description of radiodensity, which is also known as the CT number, and the unit is called Hounsfield unit (HU). The dataset provides CT images ranged from −3000 HU to 3000 HU. However, normally the soft tissues in human body vary between 20 HU and 50 HU. To address the organs to be delineated, specific window level and window width of CT images are manually selected as shown in Table 2. In the coarse localization phase, the window level and window width are set as those of lungs due to their high proportion in the images. Comparison examples of such adjustment are shown in Fig. 5, from which it is obvious that each organ is under-

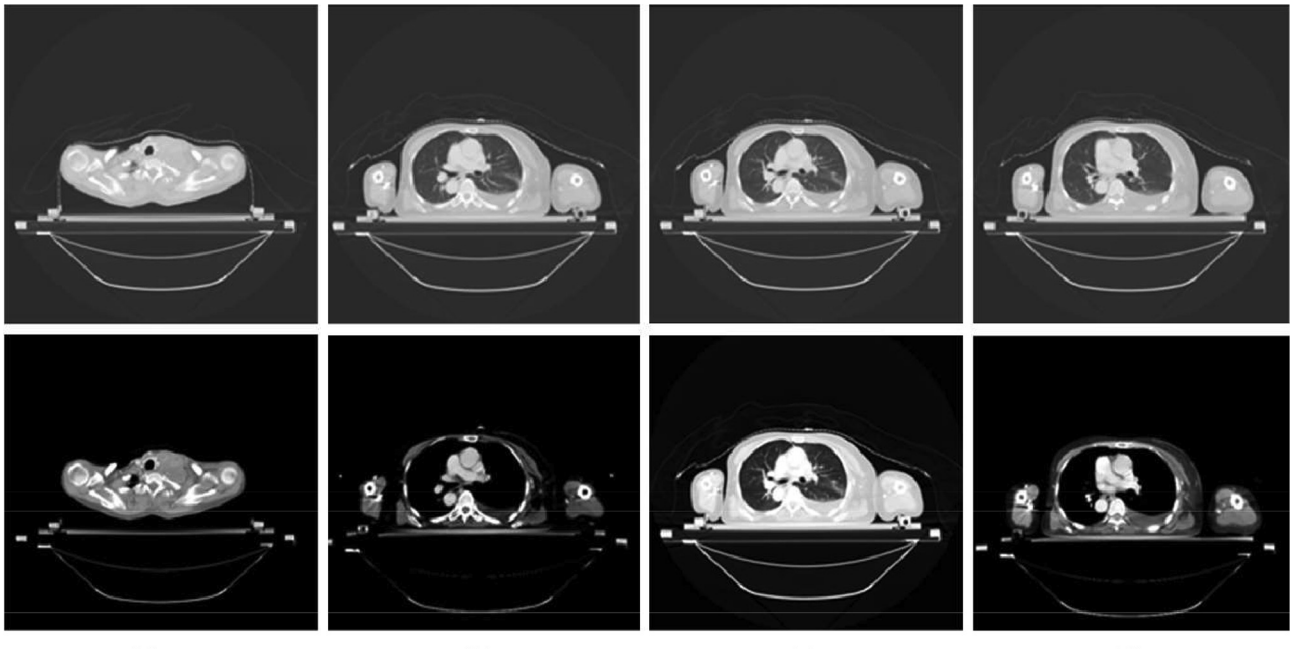

Fig. 5. Comparison of CT window level and window width adjustment: (a) heart, (b) trachea, (c) esophagus, and (d) spinal cord.

Table 3
Comparison of OARs segmentation results with different methods.

Algorithm	DSC (%)	95%HD (mm)	Inference time (s)
U-Net [2]	88.08	3.71	1.92
DeeplabV3+ [33]	88.66	3.68	2.43
Res-Unet	89.09	3.45	1.93
SE-ResUnet	89.31	3.24	1.95
C-SE-ResUnet	**91.56**	**2.54**	4.58

lined in the fine segmentation phase. After that, Z-score normalization is employed to normalize the image intensities. The intensity values are linearly normalized into the range $[-1, 1]$.

4.1.2. Data augmentation

We should mention that for this problem, specific data augmentation is very useful to improve the segmentation results (e.g., local distortion for esophagus). To reduce overfitting, we apply data augmentation such as random crop, shearing, zoom, and rotation, adding Gaussian noise, contrast adjustment, and local distortion.

The original prediction results have some noise in both the coarse and fine phases. In order to remove such noise, after coarse localization, we use the largest connected component algorithm to maintain only the main areas of the predicted organs of interest. Then after fine segmentation, we remove those connected components with a volume less than 200 voxels.

4.2. Training process and details

Our coarse localization SE-ResUnet is trained using 200 epochs on all the 512×512 size slices of 40 patients in the training data. We use SGD as the optimizer with the initial learning rate of 1e-3, which decays at 80, 120, and 150 epochs. Using the cropped 3D volumes from the localization results for each organ, 5 fine segmentation SE-ResUnets are trained only using the slices of these volumes. In addition, these slices are resized to 256×256 for lungs and 128×128 for the other organs. In the fine segmentation phase, SE-ResUnets are also trained using 200 epochs with SGD and an initial learning rate of $1e{-}3$, which decays at 80, 120, and 150 epochs.

Our experiments are performed on a workstation platform with Intel(R) Xeon(R) CPU E5-2630 v4 @ 2.20 GHz, 256 GB RAM and 8x NVIDIA Titan Xp GPU with 12 GB GPU memory. The code is implemented with PyTorch 0.4.1 in Ubuntu 18.04.

4.3. Results

Comparative Experiments. The results obtained on the StructSeg 2019 dataset are reported. In each test case, 2D slices from 3D CT images are used as input. For test result evaluation, joint 2D segmentation results of each case are considered. DeepLab is a widely-used pixel-wise segmentation tool [33], which also uses an encoder-decoder structure. Here, we use U-Net and DeepLabV3+ as baseline models. Res-Unet is implemented for the ablation experiments as well. Both the coarse SE-ResUnet and cascaded SE-ResUnet are tested. All the deep learning models are tested on the same validation set, using 2D image slices as input, and the results are shown in Table 3. The inference time in Table 3 is the average run time of model inference on the test set with a single GPU (Nvidia 1080Ti).

Table 4 shows the Dice coefficients of our coarse localization model and cascaded SE-ResUnet segmentation model for each organ. One can see that the coarse model performs well in localizing all the organs, and cascaded SE-ResUnet improves the overall performance with a large margin, especially on the esophagus and trachea segmentation results.

Some examples of segmentation results are shown in Figs. 6–9. In all these figures, (a) is for the delineation labels, and (b), (c), (d), and (e) are for the results of U-Net, ResUnet, coarse SE-ResUnet, and cascaded SE-ResUnet, respectively. Note that all the results shown here are from a single patient case. Fig. 6 shows the spinal cord delineation results at the same position; our method yields a closest shape to the ground truth while the other models just delineate some part of it. Likewise, an example of the heart segmentation is shown in Fig. 7; only C-SE-ResUnet attains the bulge curve on the bottom-left boundary. One may notice that U-Net, ResUnet, and coarse SE-ResUnet produce some false positive segmentation of trachea in Fig. 8(b)–(d), while cascaded SE-ResUnet in Fig. 8(e) does not (in the red boxes). Also, our method avoids the false negative delineation in Fig. 9. Overall, our proposed cascaded SE-ResUnet yields the best performance.

Rectangle box plot of DSC for each test case is shown in Fig. 10. Likewise, cascaded SE-ResUnet shows superior robustness, especially on the segmentation results of small size organs such as esophagus, spinal cord, and trachea.

Finally, we submitted our code package by docker online, and our team was listed as "zju_realdoctor", ranked the 1st place out of 574 participating teams with 0.9066 in DSC and 2.6642 in 95% HD.

Ablation Experiments. Apart from the above comparison with the state-of-the-art methods, we also conduct extensive ablation

Table 4
Comparison of DSC values of each OAR in the validation set.

DSC(%) Algorithm / Organ	U-Net	DeepLabV3+	ResUnet	SE-ResUnet	C-SE-ResUnet
Left Lung	96.39	96.38	96.62	96.70	**97.01**
Right Lung	95.94	95.96	96.03	96.28	**96.63**
Heart	93.44	93.19	93.70	93.97	**94.49**
Trachea	73.34	70.88	71.36	77.02	**80.69**
Esophagus	83.10	83.56	84.02	84.04	**84.98**
Spinal Cord	87.36	86.91	89.00	89.18	**91.00**
Mean	88.08	88.66	89.09	89.31	**91.56**

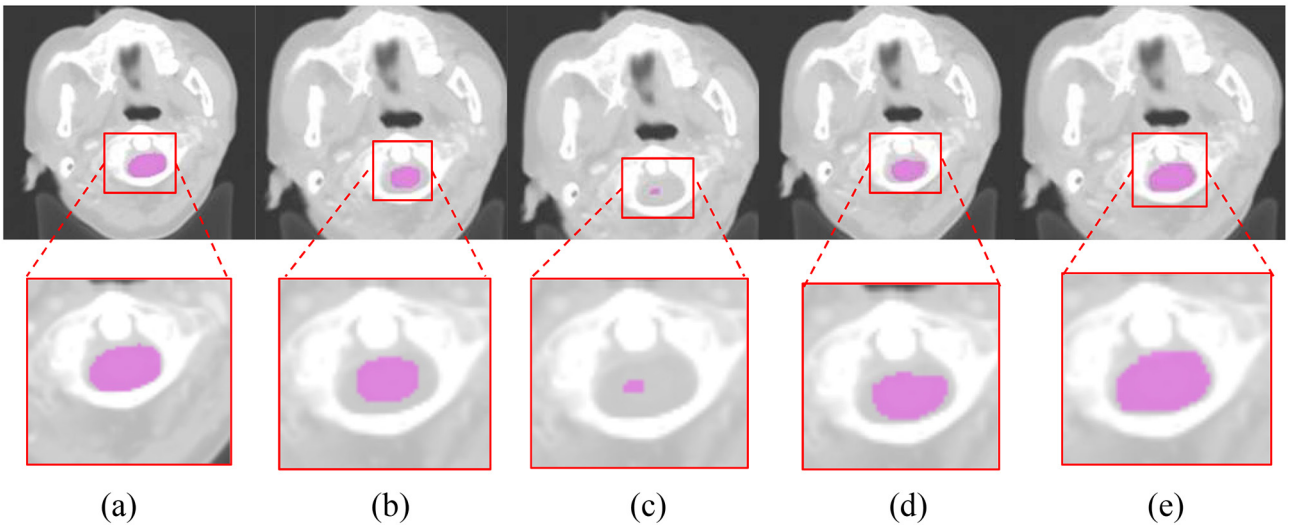

Fig. 6. Comparison of segmentation results: Spinal cord delineation. (a) Ground truth, (b) U-Net, (c) Res-Unet, (d) SE-ResUnet, and (e) C-SE-ResUnet.

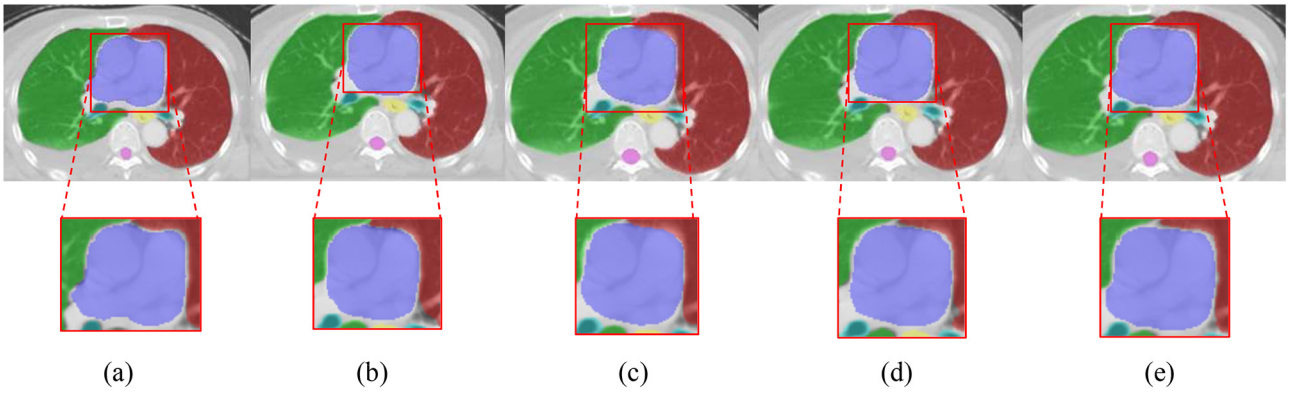

Fig. 7. Comparison of segmentation results: Heart delineation. (a) Ground truth, (b) U-Net, (c) Res-Unet, (d) SE-ResUnet, and (e) C-SE-ResUnet.

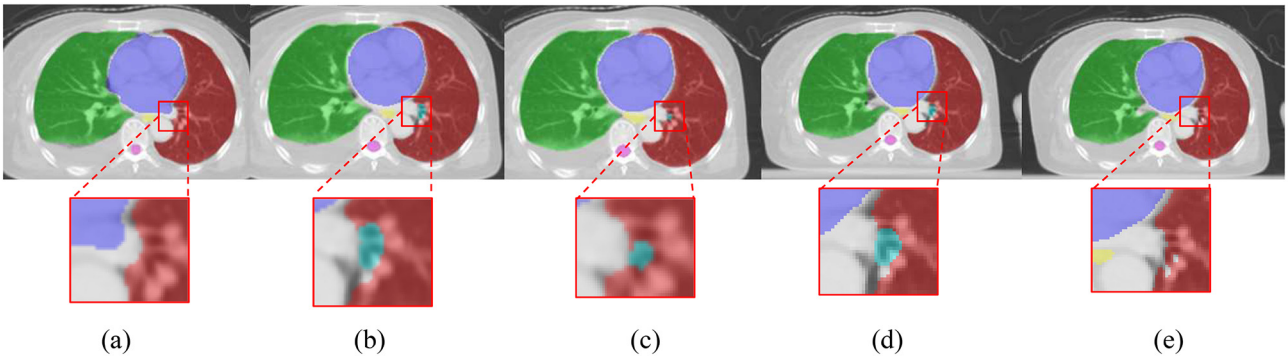

Fig. 8. Comparison of segmentation results: False positive trachea delineation. (a) Ground truth, (b) U-Net, (c) Res-Unet, (d) SE-ResUnet, and (e) C-SE-ResUnet.

experiments to validate the effectiveness of our method. Table 5 reports the values of DSC and 95%HD for different experiments. The baseline is U-Net [2], which is a classic encoder-decoder deep learning network for medical image segmentation. According to the ablation experimental results in Table 5, the implementation of residual path and squeeze-and-excitation blocks in U-Net are effective for improving the model performance. Although SE-ResUnet shows limited improvement on DSC over the baseline, it helps reduce about 12.7% on 95%HD. Besides, the coarse-to-fine method can significantly improve the model performance. In cascaded SE-ResUnet, the adjustment of CT window width and window level in different organs and the choice of the

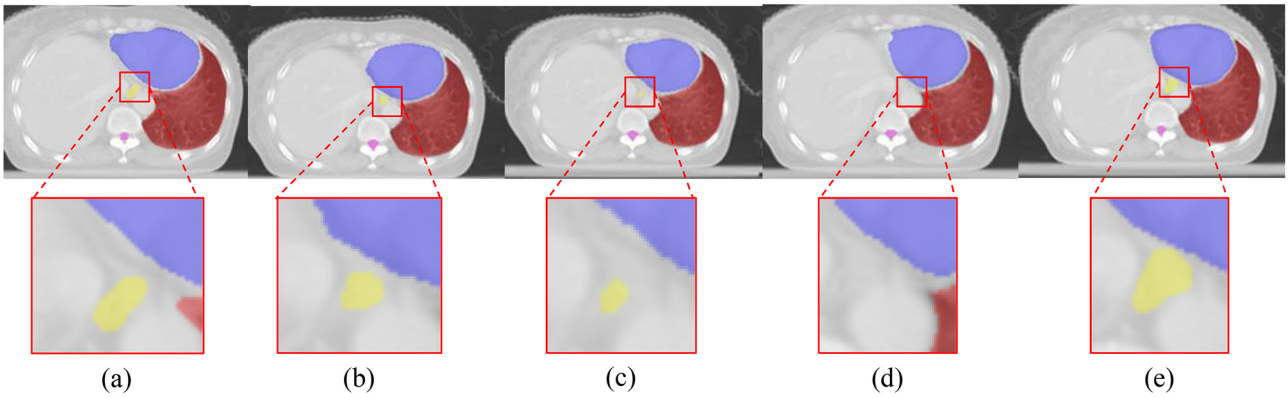

Fig. 9. Comparison of segmentation results: False negative esophagus delineation. (a) Ground truth, (b) U-Net, (c) Res-Unet, (d) SE-ResUnet, and (e) C-SE-ResUnet.

WCE/Dice loss are also meaningful to achieving better segmentation results.

A 3D visualization of segmentation results is shown in Fig. 11. The predicted mask of our proposed method shows smoother surface and less isolated regions (false positives) than the other methods.

4.4. Discussion

Choices of WCE/Dice losses. In the coarse segmentation phase, only a part of slices contains the foreground, as shown in Fig. 4. Due to the existence of negative slices, the dice loss is not very stable during the training process, which also has a worse performance than WCE loss (DSC 89.12% versus 89.31% in Table 5). We would rather choose WCE loss due to the training stability when the results are similar. However, in the fine segmentation phase, dice loss can delineate small organs better (DSC 91.21% versus 90.54% in Table 5). Therefore, we choose WCE/Dice losses in the proposed architecture to take advantage of both these loss functions.

Effect of the cascaded network. The coarse-to-fine strategy is beneficial to achieving more accurate segmentation results, especially for those tasks with unbalanced target sizes. As shown in Table 3, cascaded SE-ResUnet yields 2.25% DSC improvement and 21.6% 95%HD improvement over a single SE-ResUnet. However, the cascaded network still has some issues. On one hand, the cascaded network is harder to train due to more parameters of the model. On the other hand, the complexity of the cascaded network leads to a longer inference time (4.58s versus 1.95s in Table 3).

Comparison of 2.5D/3D segmentation methods. Our proposed cascaded SE-ResUnet is based on individual 2D slices, without considering the relationship among adjacent slices, and it leads to the lack of exploring inter-slice information. For example, the esophagus and trachea are long and occupy a number of consecutive slices. 2D segmentation methods may yield false delineation results as shown in Figs. 8 and 9. Possible better solutions may utilize 3D information in the networks, such as 2.5D anisotropic networks [34].

5. Conclusions

In this paper, we proposed a new cascaded SE-ResUnet framework for accurate segmentation of six organs at risk in 3D thoracic CT images. In particular, we developed a state-of-the-art segmentation network SE-ResUnet, implementing squeeze-and-excitation blocks and shortcut residual blocks into the common encoder-decoder U-Net structure. We devised a coarse-to-fine strategy for this segmentation problem. After performing coarse localization, five SE-ResUnet networks are trained to segment organs with different characteristics. The coarse-to-fine strategy improves segmentation accuracy by a large margin. We conducted experiments on the dataset of the StructSeg 2019 Challenge, and the validation and test studies showed the capability of our new approach for this segmentation task. Comparison and ablation experiments also suggested that our new cascaded architecture yields very good performance, especially in segmenting organs of small volumes.

CRediT authorship contribution statement

Zheng Cao: Conceptualization, Methodology, Writing - original draft. **Bohan Yu:** Methodology, Software. **Biwen Lei:** Visualization, Validation. **Haochao Ying:** Supervision, Writing - review & editing. **Xiao Zhang:** Validation. **Danny Z. Chen:** Supervision, Writing - review & editing. **Jian Wu:** Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was partially supported by the National Research and Development Program of China under grant No. 2019YFB1404802, No. 2019YFC0118802, and No. 2018AAA0102102, the National Natural Science Foundation of China under grant No. 61672453, the Zhejiang University Education Foundation under grants No. K18-511120-004, No. K17-511120-017, and No. K17-518051-02, the Zhejiang public welfare technology research project under grant No. LGF20F020013, the Medical and Health Research Project of Zhejiang Province of China (No. 2019KY667), the Wenzhou Bureau of Science and Technology of China (No. Y2020082), and the Key Laboratory of Medical Neurobiology of Zhejiang Province. D. Z. Chen's research was supported in part by NSF Grant CCF-1617735.

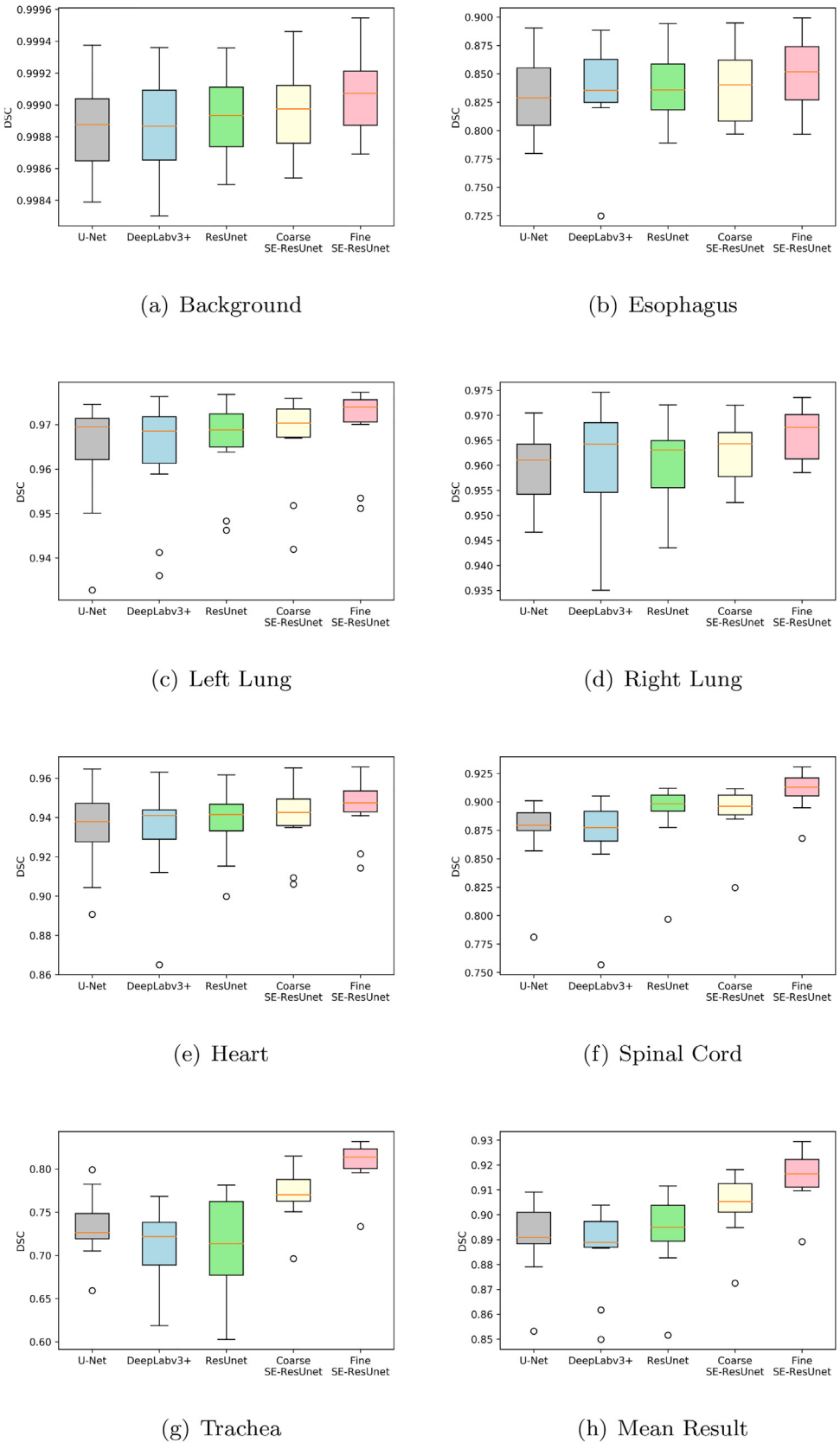

Fig. 10. Rectangle box plot of DSC values for comparison experiments. Each case result in the test set is recorded. (a) Background, (b) Esophagus, (c) Left Lung, (d) Right Lung, (e) Heart, (f) Spinal Cord, (g) Trachea, and (h) the mean result.

Table 5
Ablation study results.

Algorithm	Loss Function	CT Window	DSC (%)	95%HD (mm)
U-Net [2]	WCE	Unified	88.08	3.71
Res-Unet	WCE	Unified	89.09	3.45
SE-Unet	WCE	Unified	89.22	3.41
SE-Res-Unet	Dice	Unified	89.12	3.42
SE-Res-Unet	WCE	Unified	89.31	3.24
C-SE-ResUnet	WCE	Unified	90.22	2.79
C-SE-ResUnet	WCE	Adjustive	90.54	2.71
C-SE-ResUnet	Dice	Adjustive	91.21	2.60
C-SE-ResUnet	**WCE/Dice**	**Adjustive**	**91.56**	**2.54**

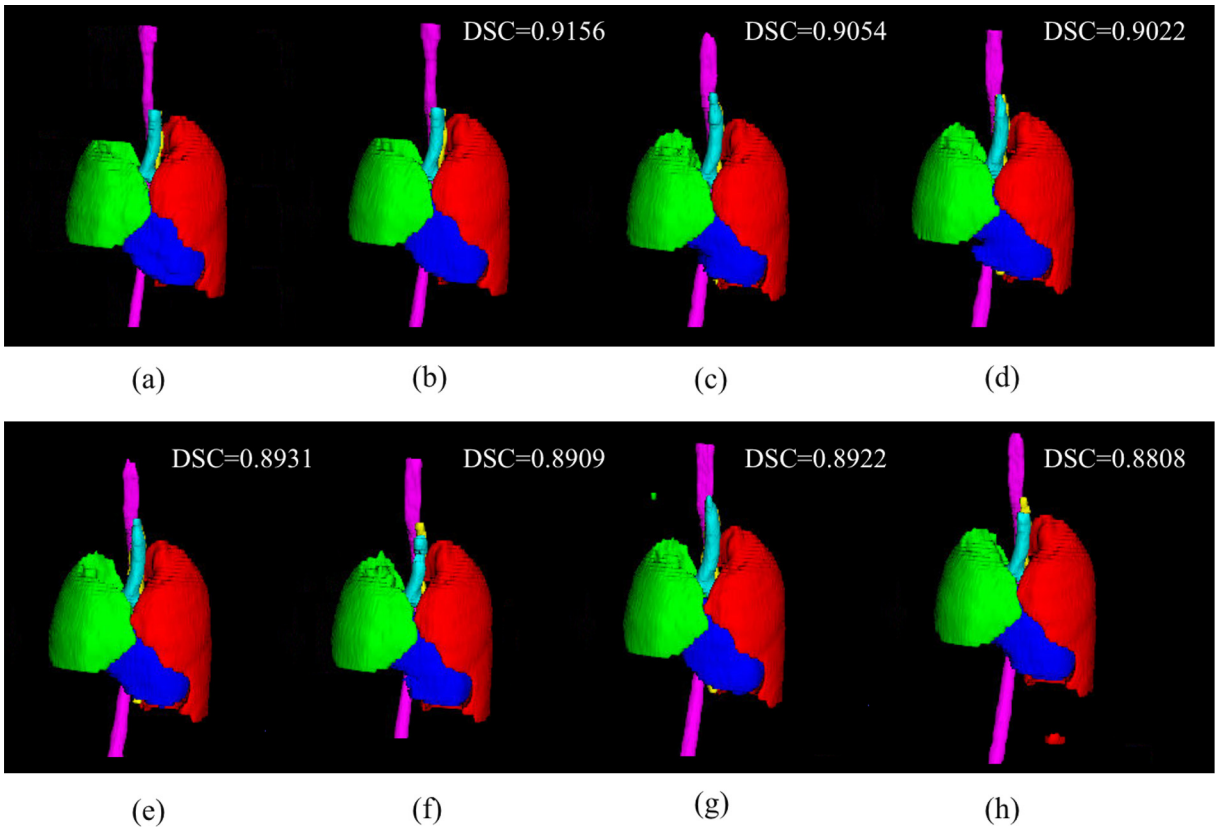

Fig. 11. 3D visualization of ablation experimental results: (a) ground truth, (b) C-SE-ResUnet with adjustive CT window and WCE/Dice loss, (c) C-SE-ResUnet with adjustive CT window and cross-entropy loss, (d) C-SE-ResUnet with unified CT window and cross-entropy loss, (e) SE-ResUnet, (f) SE-Unet, (g) Res-Unet, and (h) U-Net.

References

[1] J.A. Purdy, Dose to normal tissues outside the radiation therapy patients treated volume: a review of different radiation therapy techniques, Health Physics 95 (5) (2008) 666–676.
[2] O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2015, pp. 234–241.
[3] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[4] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141.
[5] B. Foster, U. Bagci, A. Mansoor, Z. Xu, D.J. Mollura, A review on segmentation of positron emission tomography images, Computers in Biology and Medicine 50 (2014) 76–96.
[6] N. Alsufyani, C. Flores-Mir, P. Major, Three-dimensional segmentation of the upper airway using cone beam CT: A systematic review, Dentomaxillofacial Radiology 41 (4) (2012) 276–284.
[7] S. Luo, X. Li, J. Li, Review on the methods of automatic liver segmentation from abdominal images, Journal of Computer and Communications 2 (02) (2014) 1.
[8] E.M. Van Rikxoort, W. Baggerman, B. Van Ginneken, Automatic segmentation of the airway tree from thoracic ct scans using a multi-threshold approach, in: Proc. of Second International Workshop on Pulmonary Image Analysis, 2009, pp. 341–349..
[9] J. Dehmeshki, H. Amin, M. Valdivieso, X. Ye, Segmentation of pulmonary nodules in thoracic ct scans: a region growing approach, IEEE Transactions on Medical Imaging 27 (4) (2008) 467–480.
[10] J. Ma, L. Lu, Hierarchical segmentation and identification of thoracic vertebra using learning-based edge detection and coarse-to-fine deformable model, Computer Vision and Image Understanding 117 (9) (2013) 1072–1083.
[11] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems, 2012, pp. 1097–1105..
[12] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440.
[13] Y. Li, H. Qi, J. Dai, X. Ji, Y. Wei, Fully convolutional instance-aware semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2359–2367.
[14] W. Wang, J. Shen, L. Shao, Video salient object detection via fully convolutional networks, IEEE Transactions on Image Processing 27 (1) (2017) 38–49.
[15] S. Caelles, K.-K. Maninis, J. Pont-Tuset, L. Leal-Taixé, D. Cremers, L. Van Gool, One-shot video object segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 221–230..

[16] V. Badrinarayanan, A. Kendall, R. Cipolla, SegNet: A deep convolutional encoder-decoder architecture for image segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence 39 (12) (2017) 2481–2495.
[17] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, O. Ronneberger, 3D U-Net: Learning dense volumetric segmentation from sparse annotation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2016, pp. 424–432..
[18] F. Milletari, N. Navab, S.-A. Ahmadi, V-Net: Fully convolutional neural networks for volumetric medical image segmentation, in: 2016 Fourth International Conference on 3D Vision (3DV), IEEE, 2016, pp. 565–571.
[19] Z. Zhou, M.M.R. Siddiquee, N. Tajbakhsh, J. Liang, UNet++: A nested U-Net architecture for medical image segmentation, in: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, Springer, 2018, pp. 3–11.
[20] O. Oktay, J. Schlemper, L.L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N.Y. Hammerla, B. Kainz, et al., Attention U-Net: Learning where to look for the pancreas, arXiv preprint arXiv:1804.03999..
[21] B.A. Skourt, A. El Hassani, A. Majda, Lung CT image segmentation using deep neural networks, Procedia Computer Science 127 (2018) 109–113.
[22] S. Hamidian, B. Sahiner, N. Petrick, A. Pezeshk, 3D convolutional neural network for automatic detection of lung nodules in chest CT, in: Medical Imaging 2017: Computer-Aided Diagnosis, Vol. 10134, International Society for Optics and Photonics, 2017, p. 1013409..
[23] P. Moeskops, J.M. Wolterink, B.H. van der Velden, K.G. Gilhuijs, T. Leiner, M.A. Viergever, I. Išgum, Deep learning for multi-task medical image segmentation in multiple modalities, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2016, pp. 478–486..
[24] R. Trullo, C. Petitjean, S. Ruan, B. Dubray, D. Nie, D. Shen, Segmentation of organs at risk in thoracic CT images using a sharpmask architecture and conditional random fields, in: 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017), IEEE, 2017, pp. 1003–1006..
[25] R. Trullo, C. Petitjean, D. Nie, D. Shen, S. Ruan, Joint segmentation of multiple thoracic organs in CT images with two collaborative deep architectures, in: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, Springer, 2017, pp. 21–29.
[26] M. Negahdar, D. Beymer, T. Syeda-Mahmood, Automated volumetric lung segmentation of thoracic CT images using fully convolutional neural network, in: Medical Imaging 2018: Computer-Aided Diagnosis, Vol. 10575, International Society for Optics and Photonics, 2018, p. 105751J..
[27] D. Jin, Z. Xu, Y. Tang, A.P. Harrison, D.J. Mollura, CT-realistic lung nodule simulation from 3D conditional generative adversarial networks for robust lung segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2018, pp. 732–740..
[28] T. Falk, D. Mai, R. Bensch, Ö. Çiçek, A. Abdulkadir, Y. Marrakchi, A. Böhm, J. Deubner, Z. Jäckel, K. Seiwald, et al., U-Net: Deep learning for cell counting, detection, and morphometry, Nature Methods 16 (1) (2019) 67..
[29] W. Zhu, Y. Huang, L. Zeng, X. Chen, Y. Liu, Z. Qian, N. Du, W. Fan, X. Xie, AnatomyNet: Deep learning for fast and fully automated whole-volume segmentation of head and neck anatomy, Medical Physics 46 (2) (2019) 576–589.
[30] X. Han, Automatic liver lesion segmentation using a deep convolutional neural network method, arXiv preprint arXiv:1704.07239..
[31] R. Mehta, J. Sivaswamy, M-Net: A convolutional neural network for deep brain structure segmentation, in: 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017), IEEE, 2017, pp. 437–440..
[32] A. Sevastopolsky, Optic disc and cup segmentation methods for glaucoma detection with modification of U-Net convolutional neural network, Pattern Recognition and Image Analysis 27 (3) (2017) 618–624.
[33] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.L. Yuille, DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs, IEEE Transactions on Pattern Analysis and Machine Intelligence 40 (4) (2017) 834–848.
[34] G. Wang, J. Shapey, W. Li, R. Dorent, A. Demitriadis, S. Bisdas, I. Paddick, R. Bradford, S. Zhang, S. Ourselin, et al., Automatic segmentation of vestibular schwannoma from t2-weighted mri by deep spatial attention with hardness-weighted loss, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2019, pp. 264–272..

Zheng Cao received the dual B.S. degree in applied chemistry from University of Reading and Nanjing University of Information Science and Technology in 2019. He is currently a Ph.D. candidate in College of Computer Science, Zhejiang University. His research interests include medical imaging, computer-aided diagnosis and machine learning.

Bohan Yu received the B.S. degree in biomedical science from Zhejiang University of Technology in 2018. He is currently working toward the M.A. degree in the College of Computer Science, Zhejiang University. His research interests include bio-image processing and medical artificial intelligence.

Biwen Lei received the B.S. degree in electronic business from Wuhan University in 2018. He is currently working toward the M.S. degree in the College of Computer Science, Zhejiang University. His research interests include computer vision and medical imaging.

Haochao Ying is currently an assistant professor in the School of Public Health, Zhejiang University. He received the Ph.D. degree in the College of Computer Science from Zhejiang University in 2019, and the B.S. degree in computer science and technology from Zhejiang University of Technology in 2014. His research interests include data mining for healthcare and personalized recommender system. He has authored some papers at prestigious international conferences and journals, such as World Wide Web Journal, IJCAI, CVPR, WSDM and PAKDD.

Xiao Zhang received the Ph.D. degree in the department of computer science and technology, Nanjing University in 2019. He is currently an assistant professor in the school of computer science and technology, Shandong University. His research interests include data mining, emotion recognition, and machine learning.

Danny Z. Chen received the B.S. degrees in Computer Science and in Mathematics from the University of San Francisco, California, USA in 1985, and the M.S. and Ph. D. degrees in Computer Science from Purdue University, West Lafayette, Indiana, USA in 1988 and 1992, respectively. He has been on the faculty of the Department of Computer Science and Engineering, the University of Notre Dame, Indiana, USA since 1992, and is currently a Professor. Dr. Chen's main research interests include computational biomedicine, biomedical imaging, computational geometry, algorithms and data structures, machine learning, data mining, and VLSI. He has published over 130 journal papers and 220 peer-reviewed conference papers in these areas, and holds 6 US patents for technology development in computer science and engineering and biomedical applications. He received the CAREER Award of the US National Science Foundation (NSF) in 1996 and the 2017 PNAS Cozzarelli Prize of the US National Academy of Sciences. He is a Fellow of IEEE and a Distinguished Scientist of ACM.

Jian Wu received the Ph.D. degree in Computer Science and Technology from Zhejiang University in 1998. He is an IEEE member, CFF member, CCF TCSC member, CCF TCAPP member and member of the "151 Talent Project of Zhejiang Province". Prof. Jian Wu is recently the director of Real Doctor AI Research Centre of Zhejiang University and Vice-president of National Research Institute of Big Data of Health and Medical Sciences of Zhejiang University. His research interests include Medical Artificial Intelligence, Service Computing and Data Mining.