Journal of Radiation Research, 2020, pp. 1–10
doi: 10.1093/jrr/rraa094

A slice classification model-facilitated 3D encoder–decoder network for segmenting organs at risk in head and neck cancer

Shuming Zhang[1], Hao Wang[1], Suqing Tian[1], Xuyang Zhang[1,4], Jiaqi Li[1,5], Runhong Lei[1], Mingze Gao[2], Chunlei Liu[2], Li Yang[2], Xinfang Bi[2], Linlin Zhu[2], Senhua Zhu[2], Ting Xu[3] and Ruijie Yang[1,*]

[1]Department of Radiation Oncology, Peking University Third Hospital, Beijing, China
[2]Beijing Linking Medical Technology Co., Ltd, Beijing, China
[3]Institute of Science and Technology Development, Beijing University of Posts and Telecommunications, Beijing, China
[4]Cancer Center, Beijing Luhe Hospital, Capital Medical University, Beijing, China
[5]Department of Emergency, Beijing Children's Hospital, Capital Medical University, Beijing, China
*Corresponding author. Department of Radiation Oncology, Peking University Third Hospital, 49 North Garden Road, Haidian District, Beijing, 100191, P R China. Tel: +86-010-82264926; Fax: 86-01-62017700; Email: ruijyang@yahoo.com
(Received 5 March 2020; revised 30 May 2020; editorial decision 6 September 2020)

ABSTRACT

For deep learning networks used to segment organs at risk (OARs) in head and neck (H&N) cancers, the class-imbalance problem between small volume OARs and whole computed tomography (CT) images results in delineation with serious false-positives on irrelevant slices and unnecessary time-consuming calculations. To alleviate this problem, a slice classification model-facilitated 3D encoder–decoder network was developed and validated. In the developed two-step segmentation model, a slice classification model was firstly utilized to classify CT slices into six categories in the craniocaudal direction. Then the target categories for different OARs were pushed to the different 3D encoder–decoder segmentation networks, respectively. All the patients were divided into training ($n = 120$), validation ($n = 30$) and testing ($n = 20$) datasets. The average accuracy of the slice classification model was 95.99%. The Dice similarity coefficient and 95% Hausdorff distance, respectively, for each OAR were as follows: right eye (0.88 ± 0.03 and 1.57 ± 0.92 mm), left eye (0.89 ± 0.03 and 1.35 ± 0.43 mm), right optic nerve (0.72 ± 0.09 and 1.79 ± 1.01 mm), left optic nerve (0.73 ± 0.09 and 1.60 ± 0.71 mm), brainstem (0.87 ± 0.04 and 2.28 ± 0.99 mm), right temporal lobe (0.81 ± 0.12 and 3.28 ± 2.27 mm), left temporal lobe (0.82 ± 0.09 and 3.73 ± 2.08 mm), right temporomandibular joint (0.70 ± 0.13 and 1.79 ± 0.79 mm), left temporomandibular joint (0.70 ± 0.16 and 1.98 ± 1.48 mm), mandible (0.89 ± 0.02 and 1.66 ± 0.51 mm), right parotid (0.77 ± 0.07 and 7.30 ± 4.19 mm) and left parotid (0.71 ± 0.12 and 8.41 ± 4.84 mm). The total segmentation time was 40.13 s. The 3D encoder–decoder network facilitated by the slice classification model demonstrated superior performance in accuracy and efficiency in segmenting OARs in H&N CT images. This may significantly reduce the workload for radiation oncologists.

Keywords: automatic segmentation; deep learning; head and neck; organs at risk; radiotherapy

INTRODUCTION

During the rapid development of radiotherapy technology in the last decades, high precision radiotherapy techniques such as intensity-modulated radiation therapy (IMRT) and volumetric modulated arc therapy (VMAT) have been widely used in head and neck (H&N) cancer [1]. Accurate organs at risk (OARs) segmentation is a prerequisite to precisely delivering dose to the tumor while sparing OARs in radiotherapy [2]. However, OARs segmentation, which is usually performed manually by radiation oncologists, is tedious and time-consuming [3]. Moreover, the accuracy of manual segmentation is highly dependent on the knowledge and experience of the radiation oncologists [4], and there exists remarkable inter-observer variation for manual segmentation [5]. So, it is necessary and pressing to develop an auto-segmentation model for OARs delineation.

Auto-segmentation models based on deep learning have developed rapidly in recent years [6, 7]. Ibragimov and Xing [8] used convolutional neural networks (CNN) for the segmentation of OARs in H&N CT images. They used the Dice similarity coefficient (DSC) to evaluate the segmentation result. They found that DSC was superior for the OARs with a large volume (e.g. mandible: 0.895) while it was inferior for the OARs with a small volume (e.g. left optic nerve: 0.639 and right optic nerve: 0.645). To improve the segmentation accuracy of small volume OARs, Tong *et al.* [9] proposed a shape representation model to learn the shape representation of OARs, which could constrain the results of fully convolutional neural networks (FCNN). However, though their results demonstrated that the shape representation model improved the segmentation accuracy for all OARs, the accuracy of optic nerves were still dissatisfactory (left optic nerve: 0.653 and right optic nerve: 0.689). Liang *et al.* [10] added a detection model to detect the structures in the axial plane of the CT image and constrain the segmentation region. The results were improved by adding a detection model, but the DSC was still inferior for small volume OARs (e.g. optic nerves: 0.689).

Some other studies were also devoted to improving the segmentation results of small volume structures by modifying the network loss function [11] or augmenting training data [12]. All these endeavors stated that it was difficult to gain a matched segmentation accuracy for both small and large volume structures simultaneously. This class-imbalance problem of OARs segmentation in H&N becomes more serious when the task is to segment small OARs from all the CT images [11]. In this case, most of the CT slices uncovering a specific OAR are redundant for the OARs segmentation model and may result in a time-consuming calculation. Cascaded networks consisting of a coarse and fine network were also developed for small OARs segmentation [13–15]. The coarse network was used to detect and/or coarsely segment the OARs, which reduced the redundant region, and then the fine network finely segmented the OARs based on the results of the coarse network. However, these methods still did not solve the false-positives problems on irrelevant slices [14].

In this study, a slice classification model was proposed to classify CT slices into six categories in the craniocaudal direction. Then the slices in the corresponding categories were pushed to a refined 3D segmentation network for target OARs segmentation. The redundant CT slices that were useless for target OARs segmentation were excluded. This method provides a way to solve the class-imbalance problem and false-positives on irrelevant slices. This two-step segmentation model is expected to improve the segmentation accuracy of OARs in H&N and reduce segmentation time.

MATERIALS AND METHODS

Data preparation

This study was approved by our Institutional Review Board (approval no. LM2018118). All the image data were de-identified by anonymization and analyzed retrospectively.

Image data acquisition

Twelve OARs were included in this study [brainstem, right/left eye, right/left optic nerve, right/left temporal lobe, right/left parotid, right/left temporomandibular joint (TMJ) and mandible]. A total of 170 patients with all the above OARs that were intact and not invaded by the primary tumor or metastatic lymph nodes were included in this study (brain metastases: 46, nasopharyngeal carcinoma: 40, lymphoma: 18, laryngeal carcinoma: 15, hypopharyngeal carcinoma: 9, tongue cancer: 4, maxillary sinus carcinoma: 4, other H&N cancer: 34). All patients were immobilized with a thermoplastic mask in the supine position and scanned from cranium to clavicle on a CT simulator (Brilliance Big Bore, Philips Medical Systems). Both enhanced and unenhanced CT images were scanned. The CT images were axially reconstructed, and the scan matrix was 512 × 512. The resolution of all axial plane images was between 0.69 and 1.34 mm, and the slice thickness varied between 1.5 and 3.0 mm. Magnetic Resonance images were also acquired to assist manual segmentation.

Image labelling

OARs delineation

All the OARs were delineated on the CT images for each patient by experienced radiation oncologists according to the recommended guidelines [16]. These OARs delineations, called ground truth, were used for segmentation model training.

Slice classification

The CT slices of each patient were classified into six categories (Fig. 1) in the *z*-direction (craniocaudal direction) by experienced radiation oncologists with the following boundary definitions according to the position and anatomy features of the landmark structures on images. The boundary between categories I and II was the first slice of the skull; the boundary between categories II and III was the first slice of the eyes; the boundary between categories III and IV was the last slice of the eyes; the boundary between categories IV and V was the last slice of the cerebellum; and the boundary between categories V and VI was the last slice of the mandible.

Preprocessing

The resolution of axial plane images was resampled to 1 × 1 mm using bilinear interpolation. To improve the generalization ability of the segmentation network, both random rotation within 20 degrees and random axial translation within 20 pixels were used to augment the diversity of training image data.

Experimental setup

To evaluate the effect of the slice classification model, we developed two auto-segmentation models: the segmentation model with slice classification model (two-step segmentation model) and the segmentation model without slice classification model (segmentation-only model). The deep learning networks were constructed and implemented with Keras [17] using a TensorFlow [18] backend. All computations were implemented on a computer with an Intel® Core™ i7–7700 CPU, hard disk of 4 TB, RAM of 64 GB and a Nvidia GTX 1080 GPU.

Slice classification model

The slice classification network mainly consisted of depth-wise separable convolution [19] and max-pooling layers (Fig. 2). Global average pooling was used to generate the classification tensor and reduce the

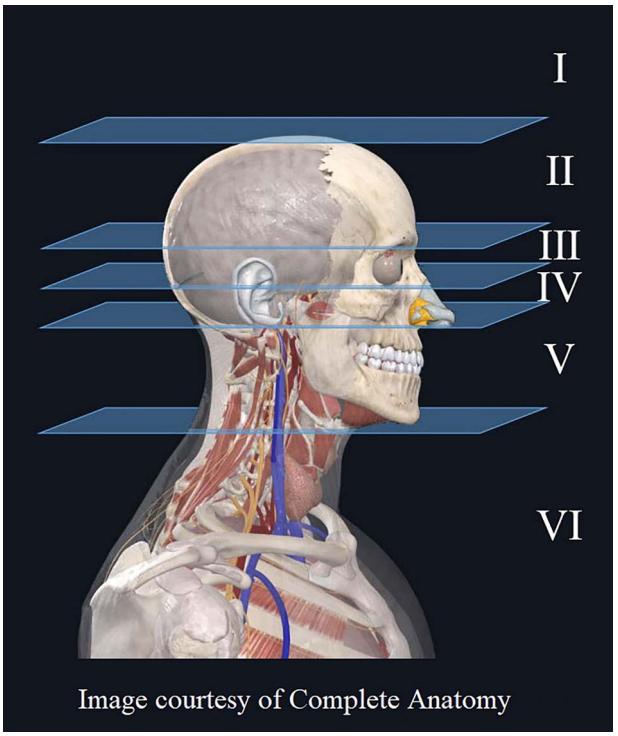

Fig. 1. Illustration of the categories in the slice classification model. The boundary between categories I and II was the first slice of the skull; the boundary between categories II and III was the first slice of the eyes; the boundary between categories III and IV was the last slice of the eyes; the boundary between categories IV and V was the last slice of the cerebellum; and the boundary between categories V and VI was the last slice of the mandible.

network parameters compared with the usage of a fully connected layer. The output layer would classify the input CT slices into six categories. The categorical cross-entropy loss function [20] and Adam optimizer [21] were used for network training, with the initial learning rate being 0.001. The categorical cross-entropy loss function is defined as

$$L = -\frac{1}{n} * \sum_{i=1}^{n} \sum_{j=1}^{c} \left(y_{i,j} * \log \left(\hat{y}_{i,j} \right) \right) \quad (1)$$

$\hat{y}_{i,j}$ where represents the predicted probability of the ith sample (total n samples) belonging to the jth class (total c classes), $y_{i,j}$ while represents the corresponding ground truth probability.

For the slice classification model, postprocessing was applied to correct for discontinuity of the predicted categories in the craniocaudal direction. For example, if the predicted categories for the continuous CT slices were in the order, e.g. I-I-I-II-I-I-I, the predicted II in the middle could be modified to I, making the final prediction I-I-I-I-I-I-I, based on the anatomical consistency.

Segmentation model

The segmentation network (Fig. 3A) was refined based on the encoder–decoder architecture of a 3D encoder–decoder. We first constructed a 3D encoder–decoder network, which consisted of an encoder path (7 down-sampling dilated convolution stacks) and a decoder path (7 up-sampling dilated convolution stacks). Four convolution modules with different dilation rates (rates = 1, 2, 3, 4) were used in the dilated convolution stacks to extract context features at different scales (Fig. 3B). In the segmentation model, the network accepted a CT volume of size 16 × 256 × 256 as the input. The three dimensions of the CT volume represented the z, x and y-axes, respectively. The CT slices of a patient were divided into several volumes sequentially with an overlay of 3 slices in the z-axis. For the case in which the number of slices was <16 in the last volume, the absent slices would be recruited from the previous volume, and they were concatenated at the top of the last volume.

In the 3D encoder path, with a 1 × 2 × 2 max-pooling size, the axial plane resolution was reduced to 1/8 of the original size after three max-pooling operations while the z-direction plane resolution remained unchanged. On the other side, the 3D de-convolution module was used in the decoder path to recover the image resolution to the original size correspondingly. In addition to the concatenation between the lower-level feature and the higher-level feature in the network, extra concatenations were set between two dilated convolution stacks to merge the front and rear feature maps to increase the feature extraction richness. Finally, all convolutions were followed by an activation layer with a scaled exponential linear unit (SELU) function [22]. Dice loss function [23] and Adam optimizer were used for network training, with the initial learning rate being 0.001. The Dice loss function is defined as:

$$L_{DSC} = 1 - DSC \quad (2)$$

where DSC is defined in equation (3) as:

$$\mathrm{DSC} = \frac{2 \mid \mathrm{V_A} \cap \mathrm{V_B} \mid}{\mid \mathrm{V_A} \mid + \mid \mathrm{V_B} \mid} \quad (3)$$

where $\mathrm{V_A}$ is the 3D region representing ground truth of an OAR; $\mathrm{V_B}$ is the 3D region representing auto-segmentation of the same OAR; and $\mathrm{V_A} \cap \mathrm{V_B}$ is the volume that ground truth and auto-segmentation have in common.

Network training and testing

The dataset contained CT images for 170 patients. For both the OARs delineation task and slice classification task, the labeled datasets were split into three parts: 120 patients were used to train the model parameters; 30 patients were used as held-out validation data; and 20 patients were used for testing.

Network training. During network training, the OARs delineation labels were used to train the segmentation model, and the slice classification labels were used to train the slice classification model. Each OAR had its corresponding segmentation model.

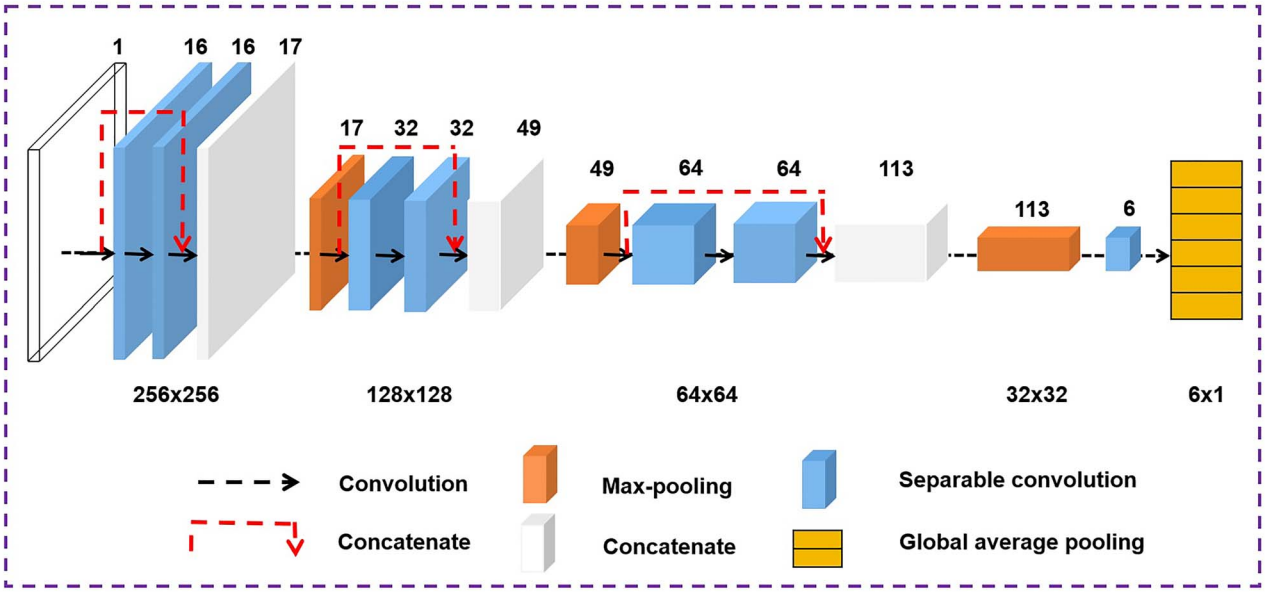

Fig. 2. The architecture of the slice classification model. It mainly consisted of separable convolution, max-pooling and global average pooling modules. The concatenations between two separable convolutions were used to merge the front and rear feature maps to extract more features.

Particularly, the early-stop mechanism was used to monitor the network iteration process, and it stopped the network training when the network performance on the validation dataset was not continuing to improve. The model parameters giving the best network performance on the validation dataset were saved and tested on the testing dataset.

Network testing. For the segmentation-only model, the original preprocessed CT images of each patient were used as the input of the segmentation model. In contrast, for the two-step segmentation model, the segmentation network was the same as the one used in the segmentation-only model, but the CT images feed for segmentation were additionally processed by the slice classification model. In detail, the original preprocessed CT images of each patient were firstly passed through the slice classification model, and then the slices containing target OARs recognized by the slice classification model were pushed into the segmentation network. Additionally, considering that the accuracy of the slice classification model may not achieve 100%, 5 extra slices [13] at the top and bottom boundary were appended to the original target categories CT slices to guarantee that the target OARs were completely covered in the slices. For example, during the testing phase, the slices in category III and 5 extra slices at the top and bottom boundary were pushed to the segmentation network for eye segmentation.

Postprocessing

OARs probability maps produced by the segmentation network were binarized at the threshold of 0.5 to get the OARs masks. Then morphological processing, including erosion and dilation (circular shaped kernel with a radius of 2 pixels and one iteration) was sequentially applied on the masks to smooth the OARs edges. The OARs contours were finally gained by carrying out edge detection on the smoothed OARs masks.

Performance evaluation

The accuracy of the slice classification model was evaluated using the percentage of correct classification of all the slices in each category and the number of slices categorized incorrectly for each category at the boundary. Also, the number of CT slices excluded/extracted by the slice classification model for each OAR was counted.

DSC [24] and 95% Hausdorff distance (95HD) [25] were used to evaluate segmentation accuracy. DSC [equation (3)] is a metric to measure the volumetric overlap between the ground truth and auto-segmentation. Its value was between 0 and 1 (0 = no overlap, 1 = complete overlap). HD is the maximum distance from a point in A to the nearest point in B, and it is defined as

$$\mathrm{HD}(A,B) = \max\left(h(A,B), h(B,A)\right) \tag{4}$$

$$h(A,B) = \max_{a\in A}\min_{b\in B}\|a-b\| \tag{5}$$

where points a and b belong to A and B respectively, and $||a - b||$ is the Euclidean distance. Smaller value usually means high segmentation accuracy. However, outliers have a great influence on HD and result in questionable values [26]. Therefore, 95HD was used in this study to avoid this problem. The 95HD was defined as the 95th percentile of the distance between points on the boundary of A and B [27].

The differences in segmentation performance between the two-step segmentation model and the segmentation-only model were evaluated by paired Student's t-test or non-parametric Wilcoxon signed-rank test according to the test of normality (Shapiro–Wilk

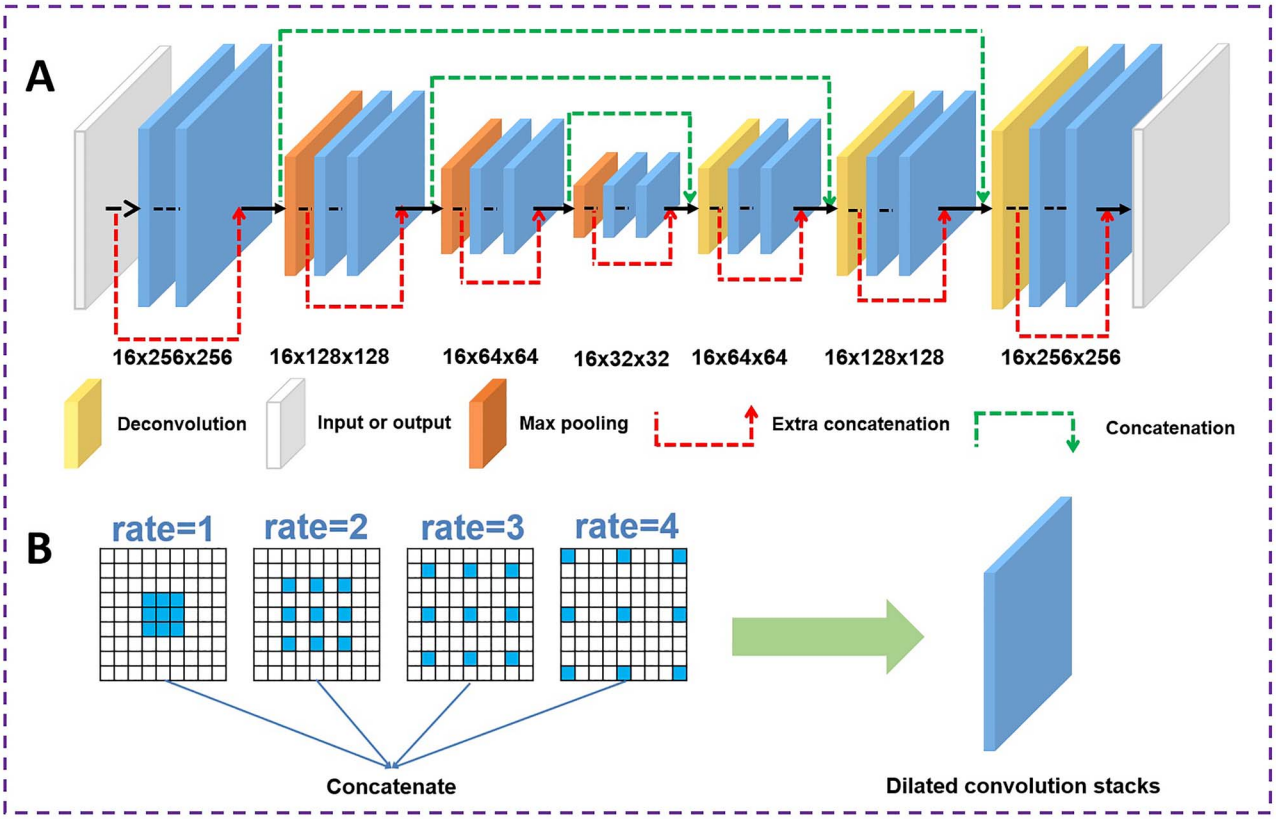

Fig. 3. The architecture of the refined 3D encoder–decoder network. (A) The refined 3D encoder–decoder network was constructed with a down-sampling path and an up-sampling path in which the dilated convolution stacks were used to extract image features. (B) The dilated convolution stack consisted of four convolution modules with different dilation rates (rate = 1, 2, 3, 4).

test). Values are presented as mean with standard deviation (SD). The statistical analyses were carried out with the Statistical Package for Social Science software (SPSS, version 24.0, IBM) and the statistical significance level was set at $\alpha = 0.05$.

RESULTS

Classification results

The percentage of correct classification of all the slices in each category is shown in Table 1. The slice classification model showed the highest accuracy on category I (99.12%) and the lowest accuracy on category IV (89.78%). The average accuracy was 95.99%.

The number of slices categorized incorrectly for each category at the boundary is shown in Fig. 4. The positive value $(+n)$ indicated that there were n redundant slices in this category at the boundary, and the negative value $(-m)$ indicated that there were m missed slices in this category at the boundary. The maximum number of slices categorized incorrectly at the boundary was 3, which appeared in two patients. Fewer than 2 slices were categorized incorrectly in 90% of the tested patients.

The average total number of CT slices for one patient was 127. The average numbers of CT slices for categories II, III, IV and V were 20, 10, 14 and 19, respectively. The average number of CT slices that was pushed/not pushed to the segmentation model of two-step segmentation model for the testing data is shown in Fig. 5. The white area (first number) indicates the number of CT slices that was not pushed to the segmentation model of the two-step segmentation model. The gray area (second number) indicates the number of CT slices that was pushed to the segmentation model of the two-step segmentation model. For example, the number of CT slices for the eye that were pushed to the segmentation model of two-step segmentation

Table 1. Percentage of correct classification of all the slices in each category

Category	Accuracy
I	99.12%
II	97.26%
III	92.53%
IV	89.78%
V	98.21%
VI	99.01%

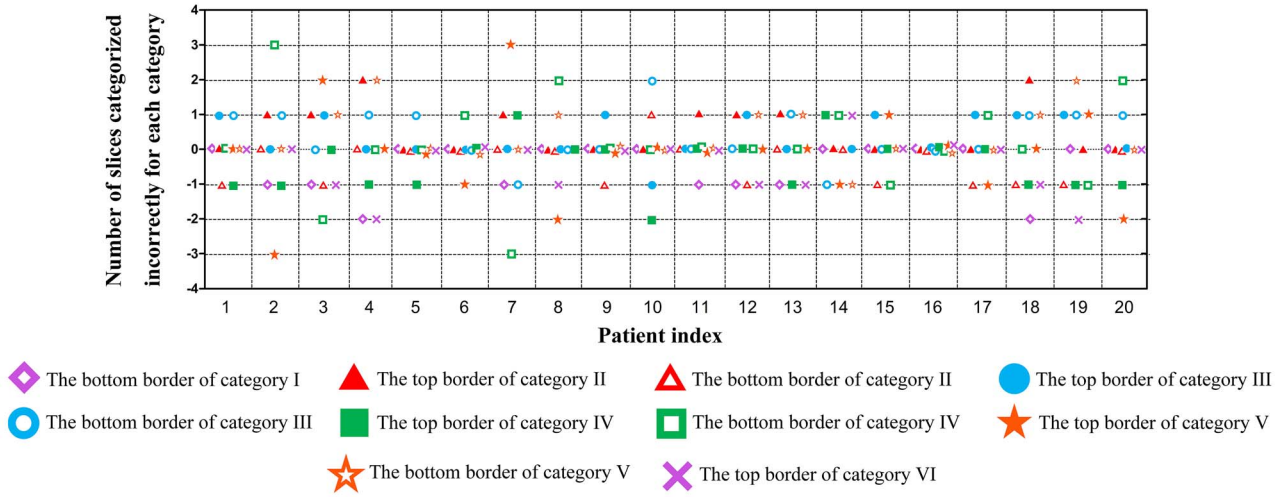

Fig. 4. The number of slices categorized incorrectly for each category at the boundary for each testing patient. The positive value $(+n)$ indicated that there were n redundant slices in this category at the boundary, and the negative value $(-m)$ indicated that there were m missed slices in this category at the boundary.

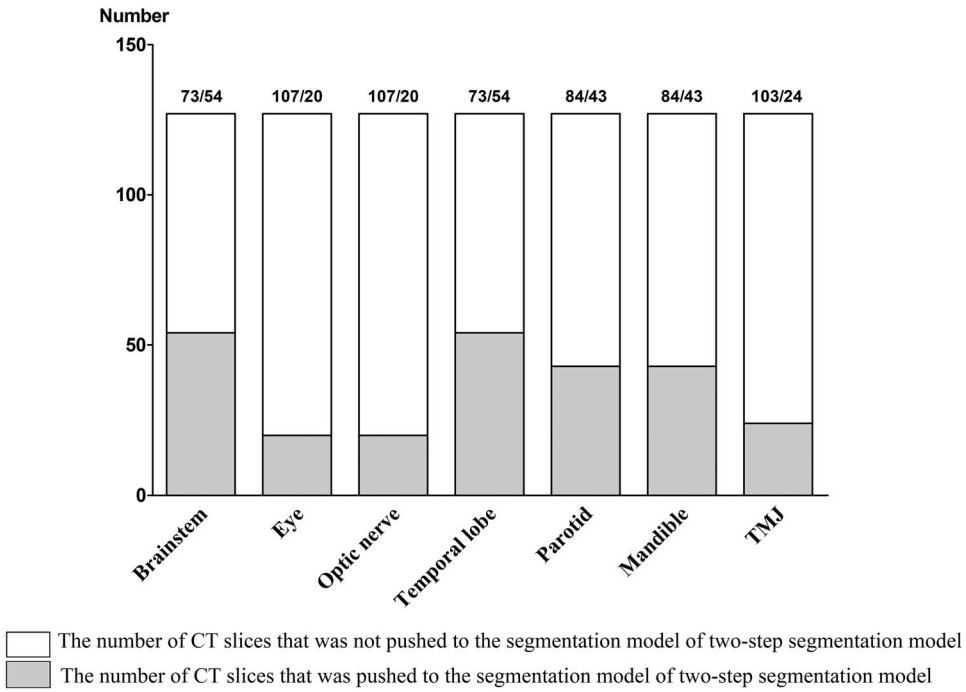

Fig. 5. The average number of CT slices that was pushed/not pushed to the segmentation model of two-step segmentation model for the testing data. The white area (first number) indicated the number of CT slices that was not pushed to the segmentation model of two-step segmentation model. The gray area (second number) indicated the number of CT slices that was pushed to the segmentation model of two-step segmentation model.

model was 20, which included 10 slices in category III and 5 extra slices at the top and bottom boundary, respectively. The remaining 107 CT slices were not pushed to the segmentation model. This means that 84.25% (107/127) of CT slices were excluded from the segmentation using the slice classification model. The excluded CT slices proportion for brainstem, optic nerve, temporal lobe, parotid, mandible and TMJ was 57.48 (73/127), 84.25 (107/127), 57.48 (73/127), 66.14 (84/127), 66.14 (84/127) and 81.10% (103/127) respectively. The average excluded CT slices proportion for one OAR was 70.98%.

Table 2. Results of DSC for the 3D encoder–decoder network with or without the slice classification model; values are given as mean ± standard deviation

	DSC		
OARs	With	Without	*P*
Right eye	0.88 ± 0.03	0.87 ± 0.04	0.111
Left eye	0.89 ± 0.03	0.84 ± 0.11	0.022∗
Right optic nerve	0.72 ± 0.09	0.69 ± 0.09	0.027∗
Left optic nerve	0.73 ± 0.09	0.70 ± 0.10	0.013∗
Brainstem	0.87 ± 0.04	0.80 ± 0.08	<0.001∗
Right temporal lobe	0.81 ± 0.12	0.69 ± 0.13	0.004∗
Left temporal lobe	0.82 ± 0.09	0.67 ± 0.14	<0.001∗
Right TMJ	0.70 ± 0.13	0.68 ± 0.13	0.355
Left TMJ	0.70 ± 0.16	0.66 ± 0.13	0.155
Mandible	0.89 ± 0.02	0.88 ± 0.03	0.071
Right parotid	0.77 ± 0.07	0.76 ± 0.07	0.167
Left parotid	0.71 ± 0.12	0.68 ± 0.20	0.811

∗ $P < 0.05$ was considered significant.

Table 3. Results of 95HD (mm) for the 3D encoder–decoder network with or without the slice classification model; values are given as mean ± standard deviation

	95HD		
OARs	With	Without	*P*
Right eye	1.57 ± 0.92	2.25 ± 1.17	0.023∗
Left eye	1.35 ± 0.43	2.57 ± 1.68	0.003∗
Right optic nerve	1.79 ± 1.01	1.90 ± 1.11	0.063
Left optic nerve	1.60 ± 0.71	1.76 ± 0.75	0.043∗
Brainstem	2.28 ± 0.99	5.50 ± 3.12	<0.001∗
Right temporal lobe	3.28 ± 2.27	14.49 ± 10.23	<0.001∗
Left temporal lobe	3.73 ± 2.08	16.59 ± 8.72	<0.001∗
Right TMJ	1.79 ± 0.79	2.52 ± 1.06	0.003∗
Left TMJ	1.98 ± 1.48	3.24 ± 2.22	0.001∗
Mandible	1.66 ± 0.51	3.73 ± 2.40	0.001∗
Right parotid	7.30 ± 4.19	12.32 ± 6.47	0.001∗
Left parotid	8.41 ± 4.84	13.73 ± 6.98	0.001∗

∗$P < 0.05$ was considered significant.

Segmentation accuracy

The results of DSC and 95HD for the two-step segmentation model and the segmentation-only model are shown in Tables 2 and 3, respectively. For the two-step segmentation model, the DSC and 95HD for each OAR were as follows: right eye (0.88 ± 0.03 and 1.57 ± 0.92 mm), left eye (0.89 ± 0.03 and 1.35 ± 0.43 mm), right optic nerve (0.72 ± 0.09 and 1.79 ± 1.01 mm), left optic nerve (0.73 ± 0.09 and 1.60 ± 0.71 mm), brainstem (0.87 ± 0.04 and 2.28 ± 0.99 mm), right temporal lobe (0.81 ± 0.12 and 3.28 ± 2.27 mm), left temporal lobe (0.82 ± 0.09 and 3.73 ± 2.08 mm), right TMJ (0.70 ± 0.13 and 1.79 ± 0.79 mm), left TMJ (0.70 ± 0.16 and 1.98 ± 1.48 mm), mandible (0.89 ± 0.02 and 1.66 ± 0.51 mm), right parotid (0.77 ± 0.07 and 7.30 ± 4.19 mm) and left parotid (0.71 ± 0.12 and 8.41 ± 4.84 mm). For the segmentation-only model, the DSC and 95HD for each OAR were as follows: right eye (0.87 ± 0.04 and 2.25 ± 1.17 mm), left eye (0.84 ± 0.11 and 2.57 ± 1.68 mm), right optic nerve (0.69 ± 0.09 and 1.90 ± 1.11 mm), left optic nerve (0.70 ± 0.10 and 1.76 ± 0.75 mm), brainstem (0.80 ± 0.08 and 5.50 ± 3.12 mm), right temporal lobe (0.69 ± 0.13 and 14.49 ± 10.23 mm), left temporal lobe (0.67 ± 0.14 and 16.59 ± 8.72 mm), right TMJ (0.68 ± 0.13 and 2.52 ± 1.06 mm), left TMJ (0.66 ± 0.13 and 3.24 ± 2.22 mm), mandible (0.88 ± 0.03 and 3.73 ± 2.40 mm), right parotid (0.76 ± 0.07 and 12.32 ± 6.47 mm) and left parotid (0.68 ± 0.20 and 13.73 ± 6.98 mm). By utilizing the slice classification model, the average DSC significantly increased for all the OARs. Meanwhile, 95HD significantly decreased for all the OARs.

The delineations of OARs, including brainstem, eyes, optic nerves, temporal lobes, mandible and parotids for a representative patient generated by the two-step segmentation model and segmentation-only model, were visually compared with the ground truth in Fig. 6.

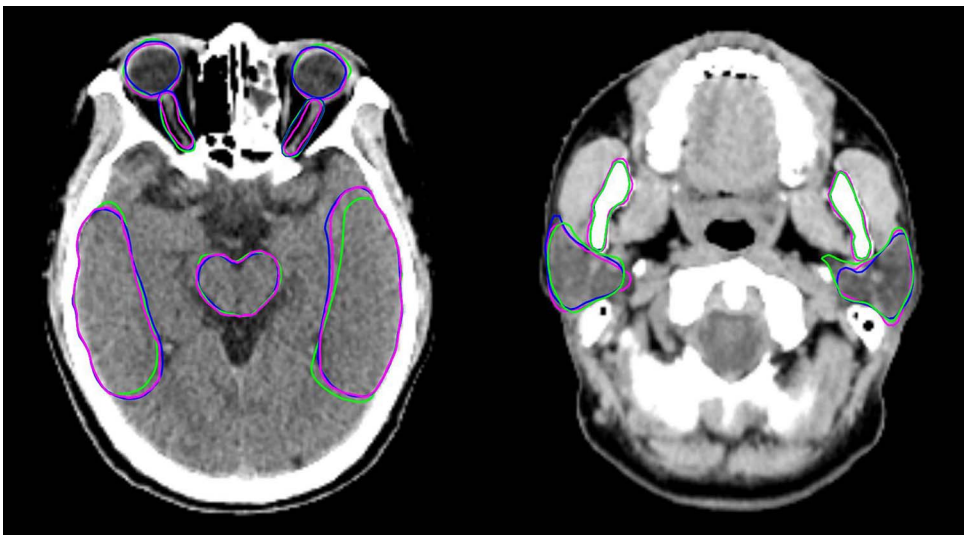

Fig. 6. Visual comparison of auto segmentation and ground truth of the brainstem, eyes, optic nerves, temporal lobes, mandible and parotids. The blue and green lines denote the delineation results generated by the two-step segmentation model and segmentation-only model, respectively. The purple lines indicate the ground truth labeled by experienced radiation oncologists.

Table 4. Results of segmentation time for the 3D encoder–decoder network with or without the slice classification model

	Time (s)	
Steps	With	Without
Slice classification	4.47	/
Segmentation	7.55	13.08
Post-processing	28.11	37.82
Total time	40.13	50.90

Segmentation time

The average time of segmenting all the twelve OARs for the segmentation-only model was 50.90 s (Table 4). By contrast, it was only 40.13 s for the two-step segmentation model, in which 4.47 s were spent on slice classification. The average segmentation time showed an efficiency improvement of 21.16%.

DISCUSSION

A slice classification-facilitated 3D encoder–decoder segmentation network for segmenting H&N OARs automatically was developed and validated in this study. The average accuracy of the slice classification model was up to 95.99%, which indicated that the slice classification network could be used as the prefixed model of the segmentation network. The average DSC and 95HD of all the OARs for the two-step segmentation model were superior to the segmentation-only model. These results demonstrated that the two-step segmentation model performed well with good consistency with ground truth. The representative example in Fig. 6 also demonstrated that the delineation of the two-step segmentation model was more consistent with ground truth than the segmentation-only model. Finally, by recruiting the preceding slice classification model, the total segmentation time decreased 21.16%, which indicated a remarkable increase in the segmentation efficiency of the two-step segmentation model. In this study, a 3D encoder–decoder network rather than 2D architecture was used in the segmentation model. The underlying assumption was that the 3D convolution kernel could capture the context information across the slices, which was helpful for the network to learn the organ shape in the *z*-direction. Based on the 3D encoder–decoder network architecture, we further employed dilated convolutions to aggregate multi-scale contextual information without systematically losing resolution.

To improve the accuracy of auto-segmentation, several methods have been proposed to optimize the segmentation model, such as interleaving CNNs [28], developing a new loss function [11], and adding an attention-based strategy [29]. Besides optimizing the models, the multi-model approaches aroused people's interest. They were used to improve the accuracy of auto-segmentation, such as the shape representation model combined FCNN [9], the detection model combined segmentation model [10], CNN cascades (simple region detector and fine segmentation unit) [13], hierarchical neural networks (coarse and fine stage CNN) [14], and the automatic anatomy recognition system [15,30]. All these studies consisted of two models, the first model detecting and/or coarsely segmenting the OARs and the second model segmenting OARs based on the results of the first model. However, all the first models in these studies worked in the axial plane (detecting and/or segmenting the OARs slice by slice). If the first model had a wrong detection on irrelevant slices, the second model had a high probability of a wrong delineation. The critical problem, false-positives on irrelevant slices, has not been properly solved. Also, the areas incorrectly segmented during the first model would enlarge the search space in the incorrect areas, increase unnecessary computation, and decrease the segmentation efficiency of the second model [14]. Although these false-positives on irrelevant slices could be somehow

alleviated using postprocessing [8], it was difficult to prevent false-positives on irrelevant slices [14]. Some studies also added a cropping step between the first and second models, which was used to crop target structure from the original image. Feng *et al.* [31] proposed a 3D U-Net to localize each organ in the axial plane. The original images were cropped to contain only one organ and served as the input to each individual organ segmentation network. They found that although the cropping step addressed GPU memory and efficiency issues, the inter-organ information was lost during cropping as the related information might help improve segmentation accuracy.

In this study, the developed two-step segmentation model had the main characteristic. The slice classification model was designed to classify CT slices into six categories in the craniocaudal direction, and only the target categories for different OARs were pushed to the corresponding 3D encoder–decoder segmentation network, respectively. The advantage of this method was to completely prevent the appearance of false-positives on irrelevant slices. The slice classification model was developed to provide a new direction to improve the accuracy and efficiency in segmenting OARs. We defined five boundaries in the *z*-direction (not the axial plane). If the number of categories was increased, the number of slices in each category would be decreased. Each OAR would cross more categories, which increased unnecessary classification. If the number of categories was decreased, the number of slices in each category would be increased. Each category contained a large range, which was not beneficial for OARs localization. According to the location characteristics of OARs, five boundaries were selected. The first and last boundaries were used to exclude CT slices that did not contain target OARs. The second and third boundaries were the top and bottom border of eyes, and a lot of OARs were in or adjacent to this category, such as eye, optic nerve, optic chiasm and pituitary. The fourth boundary was the boundary between intracranial and extracranial. The slice classification model achieved a satisfactory accuracy for distinguishing the six slice categories. The average accuracy was 95.99%. Although the accuracy of category IV was $<$90%, which was because its top and bottom boundary were both soft tissues, the maximum number of slices categorized incorrectly at the boundary was only 3, which were found in two patients. The potential classification error that may cause incomplete coverage of OARs could be well compensated by appending 5 extra slices at the top and bottom boundary. The slice classification model paid attention to the craniocaudal direction. Redundant slices were excluded by the slice classification model, which reduced the number of CT slices for the segmentation of the small OARs. With the help of the slice classification model to remove redundant slices before segmentation, the performance of the segmentation model was further improved. For example, the average DSC and 95HD of the right optic nerve for the two-step segmentation model were 0.72 ± 0.09 and 1.79 ± 1.01 mm, respectively. The average DSC and 95HD of the right optic nerve for the segmentation-only model were 0.69 ± 0.09 and 1.90 ± 1.11 mm, respectively. The average DSC and 95HD for the two-step segmentation model were superior to the segmentation-only model. Also, the postprocessing in the two-step segmentation model was simpler than in the segmentation-only model, which was because of using a multi-model approach [9]. Although 5 extra slices at the top and bottom boundary were appended to the original target CT slices, which seemed to increase the number of CT slices, the number of CT slices pushed to the segmentation model was still remarkably reduced. The average proportion of reduced CT slices for one OAR was 70.98%.

The two-step segmentation model in this study can be further optimized in the future. Several strategies, such as attention-based strategy (which was used to extract the most relevant features to identify organ boundaries), boosting-based strategy (which was used to improve the performance of a weak classifier) [32] and lifelong learning (which was used to transfer knowledge acquired on old tasks to new ones to improve generalization and facilitate model convergence) [33], could be added to the segmentation model to improve the accuracy of segmentation further. The main limitation was that each of the 12 OARs was segmented separately with different 3D networks. During the training phase of OAR-specific models, the spatial relationship among these OARs was excluded from the model learning phase. The segmentation results gained from these independent models may cause the delineated organ edges between the OARs to overlap. In the next step, we will use a multi-label segmentation model [34, 35], such as segmenting eyes and optic nerves together, to avoid this overlap problem. The segmentation performance can be expected to be further improved by learning the spatial context embedded among the global structures.

Overall, the refined 3D encoder–decoder network with the slice classification model demonstrated superior performance in accuracy and efficiency in automatically segmenting OARs in H&N CT images. It should be possible to significantly reduce the workload of OARs segmentation for radiation oncologists.

ACKNOWLEDGMENTS

We would like to thank Jingfei Wang (Beijing Linking Medical Technology Co., Ltd, China) for the data screening, and Dunrui Chang, Deqi Cui and Hua Zhang (Beijing Linking Medical Technology Co., Ltd, China) for their suggestions with regard to writing.

FUNDING

This work was partly supported by the National Natural Science Foundation of China (81372420), the Beijing Municipal Commission of Science and Technology Collaborative Innovation Project (Z201100005620012), Capital's Funds for Health Improvement and Research (2020-2Z-40919), and the Natural Science Foundation of Beijing (7202223).

CONFLICT OF INTEREST

The authors state that there are no conflicts of interest.

REFERENCES

1. Wang X, Eisbruch A. IMRT for head and neck cancer: Reducing xerostomia and dysphagia. *J Radiat Res* 2016;57:i69–75.
2. Hawkins PG, Kadam AS, Jackson WC et al. Organ-sparing in radiotherapy for head-and-neck cancer: Improving quality of life. *Semin Radiat Oncol* 2018;28:46–52.
3. Harari PM, Song S, Tomé WA. Emphasizing conformal avoidance versus target definition for IMRT planning in head-and-neck cancer. *Int J Radiat Oncol Biol Phys* 2010;77:950–8.

4. Lorenzen EL, Taylor CW, Maraldo M et al. Inter-observer variation in delineation of the heart and left anterior descending coronary artery in radiotherapy for breast cancer: A multi-Centre study from Denmark and the UK. *Radiother Oncol* 2013;108:254–8.
5. Sharp G, Fritscher KD, Pekar V et al. Vision 20/20: Perspectives on automated image segmentation for radiotherapy. *Med Phys* 2014;41:050902.
6. Kosmin M, Ledsam J, Romera-Paredes B et al. Rapid advances in auto-segmentation of organs at risk and target volumes in head and neck cancer. *Radiother Oncol* 2019;135:130–40.
7. Cardenas CE, Yang J, Anderson BM et al. Advances in auto-segmentation. *Semin Radiat Oncol* 2019;29:185–97.
8. Ibragimov B, Xing L. Segmentation of organs-at-risks in head and neck CT images using convolutional neural networks. *Med Phys* 2017;44:547–57.
9. Tong N, Gou S, Yang S et al. Fully automatic multi-organ segmentation for head and neck cancer radiotherapy using shape representation model constrained fully convolutional neural networks. *Med Phys* 2018;45:4558–67.
10. Liang S, Tang F, Huang X et al. Deep-learning-based detection and segmentation of organs at risk in nasopharyngeal carcinoma computed tomographic images for radiotherapy planning. *Eur Radiol* 2019;29:1961–7.
11. Zhu W, Huang Y, Zeng L et al. AnatomyNet: Deep learning for fast and fully automated whole-volume segmentation of head and neck anatomy. *Med Phys* 2019;46:576–89.
12. Thyreau B, Sato K, Fukuda H et al. Segmentation of the hippocampus by transferring algorithmic knowledge for large cohort processing. *Med Image Anal* 2018;43:214–28.
13. Men K, Geng H, Cheng C et al. Technical note: More accurate and efficient segmentation of organs-at-risk in radiotherapy with convolutional neural networks cascades. *Med Phys* 2018;46:289–92.
14. Tappeiner E, Pröll S, Hönig M et al. Multi-organ segmentation of the head and neck area: An efficient hierarchical neural networks approach. *Int J Comput Assist Radiol* 2019;14:745–54.
15. Wu X, Udupa JK, Tong Y et al. AAR-RT - a system for auto-contouring organs at risk on CT images for radiation therapy planning: Principles, design, and large-scale evaluation on head-and-neck and thoracic cancer cases. *Med Image Anal* 2019;54:45–62.
16. Brouwer CL, Steenbakkers RJ, Bourhis J et al. CT-based delineation of organs at risk in the head and neck region: DAHANCA, EORTC, GORTEC, HKNPCSG, NCIC CTG, NCRI, NRG oncology and TROG consensus guidelines. *Radiother Oncol* 2015;117:83–90.
17. Chollet, F. KERAS, GitHub. (2015) https://github.com/fchollet/keras
18. Abadi M, Agarwal A, Barham P et al. TensorFlow: Large-scale machine learning on heterogeneous systems. *Software* 2015; available from tensorflow.org.
19. Chollet F. Xception: Deep learning with Depthwise separable convolutions. *arXiv* 2017;1610:02357v3.
20. Zhang Z, Sabuncu MR. Generalized cross entropy loss for training deep neural networks with noisy labels. *arXiv* 2018;1805:07836.
21. Kingma DP, Ba JL, Adam A. Method for stochastic optimization. *arXiv* 2014;1412:6980v9.
22. Klambauer G, Unterthiner T, Mayr A et al. Self-normalizing neural networks. *Advances in neural information processing systems* 2017arXiv;1706:02515v5.
23. Sudre CH, Li W, Vercauteren T et al. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. *arXiv* 2017;1707:03237v3.
24. Crum WR, Camara O, Hill DLG. Generalized overlap measures for evaluation and validation in medical image analysis. *IEEE Trans Med Imaging* 2006;25:1451–61.
25. Huttenlocher DP, Klanderman GA, Rucklidge WJ. Comparing images using the Hausdorff distance. *IEEE Trans Pattern Anal Mach Intell* 1993;15:850–63.
26. Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool. *BMC Med Imaging* 2015;15:29.
27. Raudaschl PF, Zaffino P, Sharp GC et al. Evaluation of segmentation methods on head and neck CT: Auto-segmentation challenge 2015. *Med Phys* 2017;44:2020–36.
28. Ren X, Xiang L, Nie D et al. Interleaved 3D-CNNs for joint segmentation of small-volume structures in head and neck CT images. *Med Phys* 2018;45:2063–75.
29. Tang H, Chen X, Liu Y et al. Clinically applicable deep learning framework for organs at risk delineation in CT images. *Nat Mach Intell* 2019;1:480–91.
30. Udupa JK, Odhner D, Zhao L et al. Body-wide hierarchical fuzzy modeling, recognition, and delineation of anatomy in medical images. *Med Image Anal* 2014;18:752–71.
31. Feng X, Qing K, Tustison NJ et al. Deep convolutional neural network for segmentation of thoracic organs-at-risk using cropped 3D images. *Med Phys* 2019;46:2169–80.
32. Zhong T, Huang X, Tang F et al. Boosting-based cascaded convolutional neural networks for the segmentation of CT organs-at-risk in nasopharyngeal carcinoma. *Med Phys* 2019; [Online ahead of print].
33. Chan JW, Kearney V, Haaf S et al. A convolutional neural network algorithm for automatic segmentation of head and neck organs at risk using deep lifelong learning. *Med Phys* 2019;46:2204–13.
34. Fu H, Cheng J, Xu Y et al. Joint optic disc and cup segmentation based on multi-label deep network and polar transformation. *IEEE Trans Med Imaging* 2018;37:1597–605.
35. Nouranian S, Ramezani M, Spadinger I et al. Learning-based multi-label segmentation of transrectal ultrasound images for prostate brachytherapy. *IEEE Trans Med Imaging* 2016;35:921–32.