3D Lightweight Network for Simultaneous Registration and Segmentation of Organs-at-Risk in CT Images of Head and Neck Cancer

Bin Huang, Yufeng Ye, Ziyue Xu, *Member, IEEE*, Zongyou Cai, Yan He, Zhangnan Zhong, Lingxiang Liu, Xin Chen, Hanwei Chen, and Bingsheng Huang

***Abstract*—Image-guided radiation therapy (IGRT) is the most effective treatment for head and neck cancer. The successful implementation of IGRT requires accurate delineation of organ-at-risk (OAR) in the computed tomography (CT) images. In routine clinical practice, OARs are manually segmented by oncologists, which is time-consuming, laborious, and subjective. To assist oncologists in OAR contouring, we proposed a three-dimensional (3D) lightweight framework for simultaneous OAR registration and segmentation. The registration network was designed to align a selected OAR template to a new image volume for OAR localization. A region of interest (ROI) selection layer then generated ROIs of OARs from the registration results, which were fed into a multiview segmentation network for accurate OAR segmentation. To improve the performance of registration and segmentation networks, a centre distance loss was designed for the registration network, an ROI classification branch was employed for the segmentation network, and further, context information was incorporated to iteratively promote both networks' performance. The segmentation results were further refined with shape information for final delineation. We evaluated registration and segmentation performances of the proposed framework using three datasets. On the internal dataset, the Dice similarity coefficient (DSC) of registration and segmentation was 69.7% and 79.6%, respectively. In addition, our framework was evaluated on two external datasets and gained satisfactory performance. These results showed that the 3D lightweight framework achieved fast, accurate and robust registration and segmentation of OARs in head and neck cancer. The proposed framework has the potential of assisting oncologists in OAR delineation.**

***Index Terms*—Segmentation, registration, computed tomography, organ-at-risk, head and neck cancer, lightweight network.**

I. Introduction

HEAD and neck (HaN) cancer is one of the most common cancers and one major cause of cancer death in the world [1], [2]. Currently, image-guided radiation therapy (IGRT) is one of the most effective methods for the treatment of HaN cancer owing to its precise radiation dose delivery [3], [4]. To achieve accurate dose planning of IGRT, computed tomography (CT) conveying tissue density information is necessary, and accurate segmentation of gross tumour volumes and organ-at-risks (OARs) is a most critical step in IGRT planning [5].

In routine clinical practice, OARs are manually segmented by oncologists, which is time consuming, even for experienced oncologists. Such process may approximately take three hours

Manuscript received August 25, 2021; revised November 9, 2021; accepted November 10, 2021. Date of publication November 16, 2021; date of current version April 1, 2022. This work was supported in part by the Shenzhen Science and Technology Project under Grant JCYJ20160307114900292; in part by the Guangzhou Health Science and Technology Project under Grant 20211A011117; in part by the Science and Technology Plan Project of Panyu District, Guangzhou, under Grant 2021-Z04-001; and in part by the Guangzhou Science and Technology Project under Grant 201804010111. *(Bin Huang, Yufeng Ye, and Ziyue Xu contributed equally to this work.) (Corresponding authors: Bingsheng Huang; Hanwei Chen.)*

This work involved human subjects or animals in its research. The authors confirm that all human/animal subject research procedures and protocols are exempt from review board approval.

Bin Huang is with the Medical AI Laboratory, School of Biomedical Engineering, Health Science Center, Shenzhen University, Shenzhen 518060, China, and also with the National-Regional Key Technology Engineering Laboratory for Medical Ultrasound, Guangdong Key Laboratory for Biomedical Measurements and Ultrasound Imaging, School of Biomedical Engineering, Health Science Center, Shenzhen University, Shenzhen 518060, China (e-mail: 120339676@qq.com).

Yufeng Ye and Hanwei Chen are with the Department of Radiology, Guangzhou Panyu Central Hospital, Guangzhou 510000, China, and also with the Medical Imaging Institute of Panyu, Guangzhou 510000, China (e-mail: 838554325@qq.com; docterwei@sina.com).

Ziyue Xu is with NVIDIA Corporation, Bethesda, MD 20814 USA (e-mail: ziyue.xu@gmail.com).

Zongyou Cai, Zhangnan Zhong, and Bingsheng Huang are with the Medical AI Laboratory, School of Biomedical Engineering, Health Science Center, Shenzhen University, Shenzhen 518060, China (e-mail: 510962935@qq.com; 1322691896@qq.com; huangb@szu.edu.cn).

Yan He and Lingxiang Liu are with the Department of Oncology, Guangzhou Panyu Central Hospital, Guangzhou 510000, China, and also with the Cancer Institute of Panyu, Guangzhou 510000, China (e-mail: 396538984@qq.com; 740599037@qq.com).

Xin Chen is with the National-Regional Key Technology Engineering Laboratory for Medical Ultrasound, Guangdong Key Laboratory for Biomedical Measurements and Ultrasound Imaging, School of Biomedical Engineering, Shenzhen University, Shenzhen 518060, China (e-mail: chenxin@szu.edu.cn).

This article has supplementary downloadable material available at https://doi.org/10.1109/TMI.2021.3128408, provided by the authors.

Digital Object Identifier 10.1109/TMI.2021.3128408

for a single patient [6]. In addition, manual segmentation introduces inconsistencies owing to inter- and intra-observer variations [7]. Particularly, the determination of OARs' extent is dependent on the clinicians' experience and the specific imaging protocols, especially for soft tissue in a small volume.

It has been shown that an accurate registration method can assist clinicians in improving the accuracy of OAR localization [8], in which case experts can make a diagnosis based on a comparison between patients and healthy people with the registration results. Further, automated segmentation method can significantly reduce the time cost of OARs delineation. Therefore, an automatic, accurate, and efficient method for OAR registration and segmentation in the CT images is highly desirable.

However, it is challenging to achieve such a system, mainly due to the following four reasons. (1) The anatomical structures of OARs are complex with high variations. (2) The soft tissues in CT images, such as the optical nerve and chiasm, have low contrast which are hard to be recognized [9]. (3) The physical sizes of organs are highly imbalanced, making the model hard to be trained. (4) Planning CT with a high resolution may increase the inference time of the trained model.

To overcome these challenges, we proposed a cascade framework consisting of a registration network and a segmentation network for joint registration and segmentation of multiple OARs using CT image volumes. The registration network aimed to align the OARs of the template to the new CT image volume. Subsequently, a region of interest (ROI) selection layer generated the ROIs of each OAR in the CT image volumes based on the registration results. The ROIs of each OAR were then fed into the segmentation network to generate multiple segmentation probability maps. Furthermore, we incorporated context information into our proposed framework to improve the registration and segmentation accuracy iteratively. Compared with cascade multiple models for using context information, our strategy did not increase the number of parameters. Finally, shape information was utilized to refine the final output.

The remainder of this paper is organised as follows. In Section II, we list related works and their contributions / limitations, and then our contributions. In Section III, we introduce the details of our proposed framework. In Section IV, the experimental results and a comparison with other related results are presented. Finally, we discuss and draw the conclusion in Section V.

II. Related Work

Early works on the topic of OAR segmentation used traditional image analysis methods, which were primarily atlas-based. The segmentation results were generated by aligning a set of fixed and manually labelled templates to new image volumes [10]. Atlas-based segmentation methods typically include these steps: image pre-processing, atlas construction, image registration, and label fusion. Some studies applied atlas-based methods to CT or MRI for HaN OAR segmentation [5], [8], [11]–[15]. However, atlas-based methods have difficulty in achieving reliable performance, particularly for soft tissues. Meanwhile, the segmentation performance relies on the accuracy of the template. Therefore, some studies combined the atlas-based methods with other methods such as the active shape model, graph cut, active appearance model, and active contour model [16]–[22]. These methods utilized the anatomical knowledge of the template, but still, the segmentation performance relied heavily on the accuracy of the registration and the labeling of the atlas. In addition, it was time-consuming to complete each registration task by using traditional registration methods [23].

To address the limitations of atlas-based methods, conventional learning-based methods were employed for HaN OAR segmentation [24]–[27]. Because of the adaptive ability of learning-based methods, they improved the segmentation performance of soft tissues. However, these methods required both complex pre-processing steps and design of handcrafted features, which were also time-consuming. In addition, the performance of the learning-based methods was less robust than that of the atlas-based methods.

Recently, deep convolutional neural networks (CNNs) have shown significant success in image segmentation [28], including HaN OAR segmentation. Initially, the CNNs segmented the OARs on 2D/3D sliding windows or relied on the located auto/semi-auto OARs [29]–[34]. This design could not capture global features and required highly accurate atlas registration for localizing the OAR. To solve these problems, AnatomyNet, a deep learning model, was proposed in [35] for whole-volume image segmentation, which avoided complex pre-processing, and captured the global features. One limitation was that as the whole-volume image was fed into the network, the model required a large GPU memory. Therefore, some studies located the OARs and obtained ROI of OARs [36]–[41] in the CT image instead, to decrease the computation cost and the required GPU memory.

Previous deep learning studies generally used a segmentation or a detection network to locate the OARs in CT images, which led to huge network parameters. The anatomical information was not used as prior knowledge in these methods, which may lead to false segmentation results.

For faster and more accurate registration, deep learning methods have been applied in some studies. In these studies, the gold standard of OARs was used as landmark to optimise the registration results [42]–[44]. However, the gold standards had to be manually delineated on the image volumes, which was also time-consuming and laborious.

To address the problem of insufficient labels, joint registration and segmentation method has been proposed [45]–[47], in which only part of the images was labelled and used to establish the registration model, and then the segmentation results could be used to optimize the registration performance. However, these traditional methods were time-consuming and not robust [45]–[47]. Deep learning-based methods were thus proposed for joint registration and segmentation [48]–[53], in which the mask was used for loss computation to optimize the network, but ignored in the feature extraction stage.

Unlike previous works, our method combined the registration and segmentation in an iterative and organic way, such that the two work together and complement each other. In each iteration of our approach, the mask was used as a landmark in the feature extraction stage for more accurate registration. Subsequently, the more accurate registration further improved the segmentation performance.

It is worth highlighting our contributions as follows: 1) we proposed a lightweight framework for accurate OAR registration and segmentation; 2) by using context information in an iterative strategy, we simultaneously improved the registration and segmentation performance; 3) we designed an interpretable shape correction method for refining the segmentation results.

III. Methods

We proposed a cascade framework with mutual complement functionality for multiple OAR registration and segmentation in 3D CT images. The entire pipeline of the proposed framework is shown in Fig. 1. The framework consists of four stages. Stage ***A. Template Selection*** was designed to select a template for accurate registration. Stage ***B. 3D Registration Network*** aimed to align the template to a new image volume. Stage ***C. 3D Segmentation Network*** was designed for automatic OAR segmentation based on the registration results. The registration network and the segmentation network were involved in an iterative optimization process. The "Inner Loop" was designed for using context information. In "Outer Loop", the registration results were used for segmentation initialization, while the segmentation results were fed back into the registration network as a landmark to further improve the registration performance. Stage ***D. Shape Correction*** was designed to make further corrections to the shape of the segmentation result.

A. Template and ROI

To capture the global information of OARs and decrease the need for GPU memory, each OAR was located by the output of the registration network in the CT image volume, and the size of the bounding box was computed, and then the ROI of each OAR was obtained.

1) Template Selection: We used a simple strategy to select the best OAR template: the whole-volume image of one patient was used as a template and was registered to other patients. To assess the registration performance, the Dice similarity coefficient (DSC) and mean square error (MSE) were calculated between the registered template and other patients. Then, we combined DSC and MSE as difference indice (DI) to assess the difference between the registered template and other patients. The DSC and DI are described as:

$$DI = (1 - DSC) + 0.5 \times MSE, \tag{1}$$

$$DSC = \frac{2|P \cap G|}{|P| + |G|}, \tag{2}$$

where $\boldsymbol{P}$ is the segmentation result and $\boldsymbol{G}$ is the gold standard. The weight of MSE is lower than that of DSC because we believe that the ratio of overlap of OAR is more important.

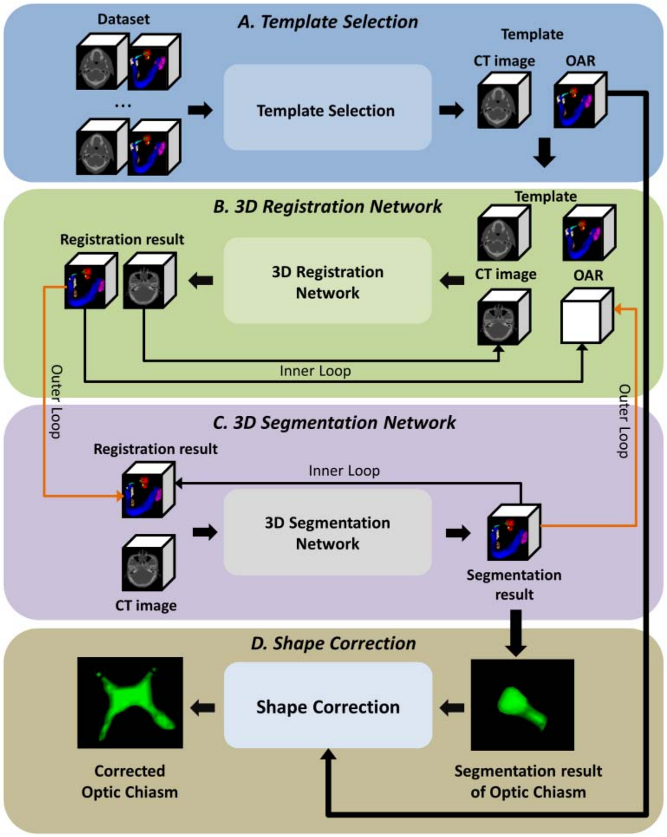

Fig. 1. Pipeline of our proposed joint registration and segmentation framework. Part A: the pipeline of template selection. Part B: the pipeline of registration. Part C: the pipeline of segmentation. Part D: the pipeline of shape correction. The long black arrows in Part B and Part C indicate the inner loop in the related stages, while the long orange arrows between Part B and Part C indicate the outer loop between the registration network and segmentation network. The two inner loops show the iterations in the 3D registration network and the 3D segmentation network for using context information, respectively. The outer loop indicates the iterations between the 3D registration network and the 3D segmentation network to improve each other iteratively.

After the template was registered for all other patients and the DI of each patient was calculated, we calculated the average DI. The patient with the lowest average DI was selected as the best template.

The registration method we used for template selection was a traditional registration algorithm [54]. This method used MSE as the loss function and the regular stepwise gradient descent method as the optimizer. The initial learning rate in our study was set as 1.0. To reduce the registration time, the maximum number of iterations was set to 30 and the minimum learning rate was 1×10^{-4}. We applied the 3D similarity transform [55] for rigid registration with seven parameters for image rotation (with three angles), translation (with three dimensions) and isotropic scaling (with one factor). The trilinear interpolation was used to interpolate the CT image volume, and the OAR gold standard was the nearest neighbour interpolation.

2) Size of ROI: The ROI should cover the entire target OAR volume and include sufficient surrounding information for OAR segmentation. The ROIs were divided into categories

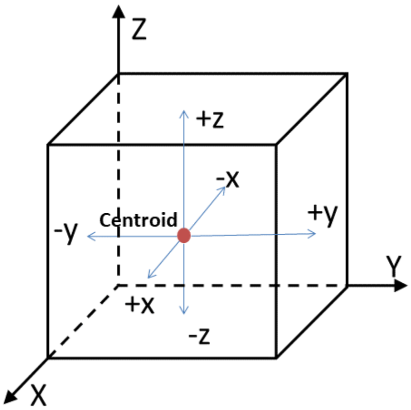

Fig. 2. Voxel number between the centroid and bounding box. Red point indicates the centroid of the organ. Rectangle is the bounding box containing the whole organ. Blue arrows and the x, y, z indicate the voxel number from the centroid to the bounding box.

depending on the location and volume of OARs. The first category was the nerve area, including the optic chiasm together with the left and right optic nerves. These organs were soft tissues, small in volume, and located in an adjacent area. The other organs, including mandible, left/right parotid gland, brainstem, and left/right submandibular gland were regarded as individual ROIs because these organs were large in volume and independent from other organs in locations.

The centroid of OARs was computed in the ROI regions, and the size of the bounding box was calculated. For example, in the nerve area, we computed the common centroid of the optic nerve and optic chiasm. Subsequently, the size of the bounding box was calculated such that it could contain both the optic nerve and optic chiasm. After the calculation of the centroid and bounding box, the voxel numbers from the centroid to the bounding box on the x-, y-, and z-coordinate axes were counted. As shown in Fig. 2, we obtained six values (−x, +x, −y, +y, −z, +z).

The voxel number from the centroid to the bounding box was calculated on the training set. Based on the statistical results of each bounding box in the training set, the average and standard deviation of −x, +x, −y, +y, −z, +z was calculated, respectively. To ensure that the ROIs covered all OARs and to ignore OAR differences between the patients, the size of ROI was expanded. The final ROI size was calculated as:

$$R(c) = [mean(D(c)) + 3 \times std(D(c))] \times 1.25, \qquad (3)$$

where $\boldsymbol{c}$ denotes −x, +x, −y, +y, −z, or +z. $\boldsymbol{R(.)}$ denotes the voxel number from the centroid of the ROI boundary. $\boldsymbol{D(x)}$ denotes the voxel number from the centroid to the bounding box.

B. 3D Registration Network

We designed a 3D registration network, inspired by Voxelmorph [42]. As shown in Fig. 3, our proposed registration network was a CNN that consisted of three encoder blocks, three decoder blocks, and a 3D spatial deformation layer. Each encoder block was followed by a down-sampling layer. The encoder block contained two 3D convolution blocks, a squeeze-and-excitation (SE) block [56], and a convolution layer with a kernel size of 1 × 1 × 1. The convolution block consisted of a 3 × 3 × 3 convolution layer, a switchable normalisation layer, and a leaky rectified linear unit (leaky ReLU) [57]. We applied two 3D convolution layers in the convolution block using kernel sizes of three and one, respectively. A convolution block with a kernel size of three was used to extract the image features, and a convolution block with a kernel size of one was designed to change the dimensions of the feature maps. After two layers of 3D convolution block, the feature maps were fed into an SE block. The feature maps were reweighted at the channel level using the outputs of the SE block. The reweighted feature maps were summed with the previous feature maps using the residual connection. Subsequently, the feature maps were fed into a 1 × 1 × 1 convolution layer to increase the dimensions of the feature maps. The decoder block contained a 3D convolution 1 × 1 × 1, and a residual connection. The decoder block was used to up-sample the feature maps of encoder stage and decreased the dimensions of feature maps.

The input of the registration network consisted of the target CT image volume, the template CT image volume, the template OAR mask, and the segmentation results. OARs in the target CT image volumes were segmented in the proposed framework. The template CT image volume was a selected CT image volume to be aligned to the target CT image volume. The template OAR mask was the manually labelled OAR of the template CT image volume. The details of template-selection method have been introduced in Subsection A. The registration result was identical to the previous registration result; hence, it was an empty matrix in the first iteration. All the input CT images were three-channel whole-volume images with multiple window widths and level settings. The three channels were set with the original window width (WW)/window level (WL), bone window (WW, 1,500 HU; WL, 300 HU), and soft tissue window (WW, 500 HU; WL, 50 HU), respectively. The first convolution layer was a group convolution layer whose group number was set to two, and the image features of the target image and template image could be separately extracted from the input image. The output layer, which was a convolution layer with a kernel size of 1 × 1 × 1, generated a deformation field with three channels from the high-level feature maps. The three channels represented the coordinate offsets in the x, y, and z dimensions, respectively. The spatial transformation layer calculated the new position of each voxel according to the coordinate offset of the deformation field.

C. 3D Segmentation Network

The 3D multiview segmentation network is shown in Fig. 3. It contained nine encoder and four decoder blocks. The structure of the encoder and decoder blocks was similar to that of the registration network, while the convolution kernel was different. In addition to the 3 × 3 × 3 convolution kernel, a 3 × 3 × 1 convolution kernel was added as a parallel branch to extract features from the axial view. To obtain high

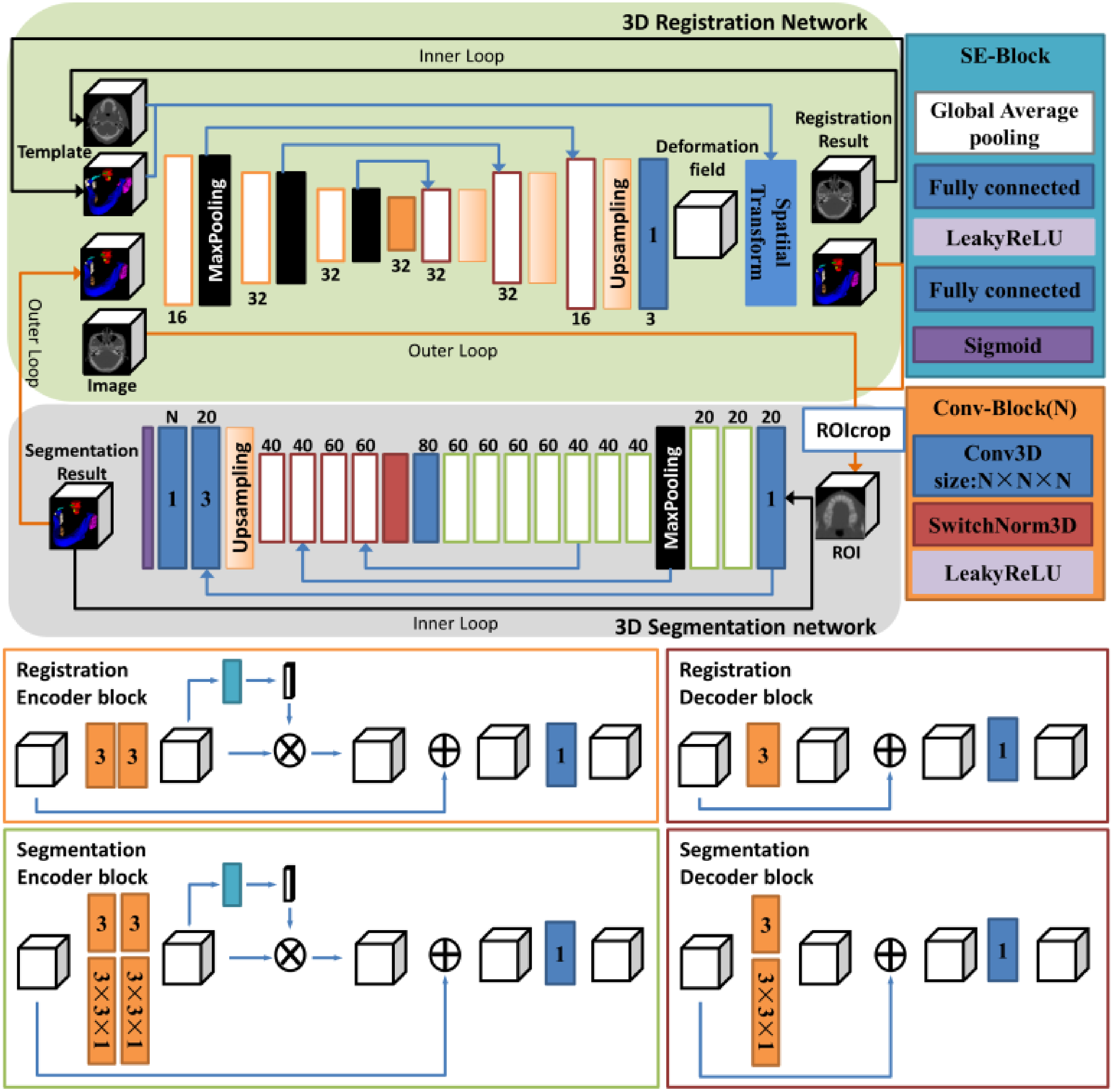

Fig. 3. Architecture of the registration and segmentation framework. White rectangles indicate the image matrixes in the framework. Black arrows indicate inner loops. Orange arrows indicate outer loops. The arrow from "Segmentation results" to the segmentation network's input indicates the inner loop. The previous segmentation results are fed into the segmentation network with the images again for making use of the context information.

segmentation performance for small OARs [35], we set only one up-sample layer and one down-sample layer.

We found that the performance of a single model on multiple OAR segmentations was unsatisfactory. The reason might be that the number of parameters of our proposed network was small, while the network had to segment multiple OARs from ROIs of different sizes. Therefore, an ROI classification branch was added to the single model network to improve the segmentation performance. The output of the last layer of the encoder structure was followed by a global average pooling and a fully connected layer. Finally, the softmax was set as the activation function to output the classification probability of the ROI category.

D. Cascade Network With Iterative Context Update

We proposed an ROI selection layer without training parameters to connect the proposed registration and segmentation modules. The registration result was fed into the ROI selection layer, and the centroid of the OAR was computed. The boundary of ROI was calculated using the centroid and the size of ROI. Then the ROIs of different OARs were generated from the CT image volumes as the input of the segmentation network.

To utilize the complementary context information between the two networks, we proposed a strategy to improve the registration and segmentation performance by iterative updates.

In the initial iteration of the proposed method, the input of the registration network consisted of a target CT image volume, template CT image volume, template OAR mask, and an empty image volume (all voxels had an image intensity of zero) with nine channels. The reason for inputting the empty image volume was that we had no initial segmentation results in the first iteration. For the same reason, the input of the segmentation network also contained an empty image volume. From the second iteration, the empty image volumes were replaced by the segmentation probability maps generated during the previous iteration.

This process was repeated for several times to iteratively improve both the registration and segmentation accuracy, similar to the auto-context model (ACM) [58]. The proposed strategy was different from the ACM strategy because we did not have to train multiple models for using the context information. In contrast to ACM, we fed the outputs into the same network again instead of using the cascade strategy, which decreased the number of trained models and parameters. In addition, the registration result was refined by using the segmentation results to avoid locating errors.

E. Loss Function

In the training stage of the registration network, the MSE loss, Dice loss, and centre distance (CD) loss were utilised to optimise the registration network. The MSE loss guaranteed the correct structure of OARs. The Dice loss ensured that the shape of each OAR was correct. The CD loss was used to decrease the locating bias of each OAR. In the training stage of the segmentation network, the Dice loss and cross-entropy (CE) loss were used to optimise the network parameters. These loss functions are defined as:

$$L_{mse} = (S_{ori} - S'_{ori})^2, \tag{4}$$

$$L_{Dice} = \frac{1}{k}\sum_{i=1}^{k} \frac{2\times(S'_i \cap S_i)}{(S'_i + S_i)}, \tag{5}$$

$$L_{CD} = \frac{1}{k}\sum_{i=1}^{k} |Cent(S'_i) - Cent(S_i)|, \tag{6}$$

$$L_{CE} = -\frac{1}{k}\sum_{i=1}^{k} S_i log S'_i + (1 - S_i) log(1 - S'_i), \tag{7}$$

$$L_{reg} = L_{Dice} + 0.5 \times L_{mse} + 0.1 \times L_{CD}, \tag{8}$$

$$L_{seg} = L_{Dice} + 0.25 \times L_{CE}, \tag{9}$$

$$L_{cascade} = 0.5 \times L_{reg} + L_{seg}, \tag{10}$$

where $\boldsymbol{L_{mse}}$, $\boldsymbol{L_{Dice}}$, $\boldsymbol{L_{CD}}$, and $\boldsymbol{L_{CE}}$ denote the MSE loss, Dice loss, CD loss, and CE loss, respectively. $\boldsymbol{L_{reg}}$ and $\boldsymbol{L_{seg}}$ denote the loss functions for the registration network and segmentation network, respectively. $\boldsymbol{L_{cascade}}$ is the loss function for the joint training of registration and segmentation networks. $\boldsymbol{S_{ori}}$ denotes the input CT image volume. $\boldsymbol{S'_{ori}}$ denotes the deformed template CT image volume after registration. $\boldsymbol{S_i}$ is the gold standard for the ith OAR. $\boldsymbol{S'_i}$ is the registration and segmentation result of the ith OAR. $\boldsymbol{Cent(.)}$ is used to compute the centroid coordinates.

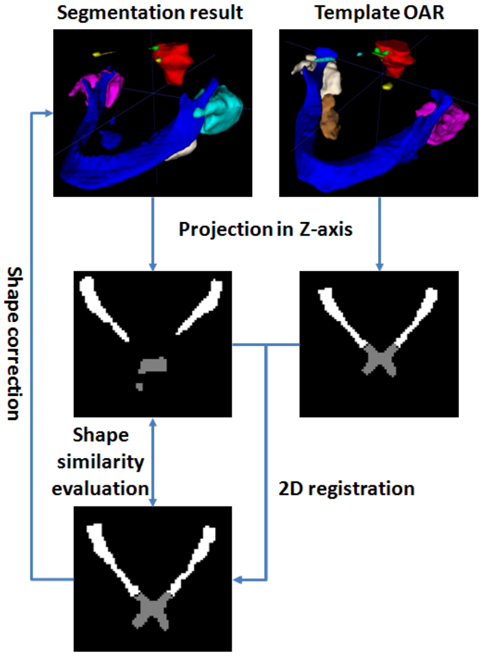

Fig. 4. Pipeline of shape correction. The first row shows the OARs (segmentation result or template). As shown in the second row, the shape of optic chiasm by automated segmentation is different from that of the template. Thus, 2D registration is applied to align the location of optic chiasm, and then shape similarity evaluation is used to compare the shapes of optic chiasm between the automated segmentation result and the template. Finally, shape correction is applied to modify the optic chiasm of segmentation result.

F. Shape Correction

For some organs of small size, especially optic chiasm, their segmentation results were sometimes incorrect even with a high DSC; therefore, we designed a post-processing method to make further corrections using the OAR template. We mainly targeted the optic chiasm, which contained only one or two slices in the CT image volume. In addition, the optic chiasm and optic nerves were adjacent to each other. Therefore, we designed out shape correction method by using the information of 2D slices and the relationship between optic chiasm and optic nerves.

The shape correction pipeline is illustrated in Fig. 4. First, the optic chiasm and optic nerves of the template/segmentation result were projected along the z-axis to be 2D images. Second, the projected template was registered to the projected segmentation result using a 2D traditional registration method, with MSE as the cost function and the regular stepwise gradient descent method as the optimizer. The maximum number of iterations was set to 10. To achieve rigid registration, a 2D similarity transform was used with four parameters for image rotation (with an angle), translation (with two dimensions) and scaling (with a factor).

Subsequently, the Hu moment [59] and DSC of optic chiasm were calculated between the 2D segmentation result and registered 2D template, and was used to evaluate the shape similarity (Hu moment) and the overlap ratio (DSC) of optic chiasm, respectively. When the DSC and Hu moment were both high, the segmentation result had a high overlap ratio with the gold standard; however, the shape of the segmentation result was incorrect. We thus corrected the shape by multiplying the optic chiasm of the template by the optic chiasm of the segmentation result. When the Hu moment was high and the DSC was low, the segmentation result had a low overlap ratio, and the shape was incorrect. Then, we replaced the segmentation result with the registered template.

G. Algorithm Details

We adopted Pytorch to implement the proposed framework. The registration, segmentation, and cascade networks were trained on an NVIDIA RTX TITAN with 24 GB memory.

First, we individually trained the segmentation network. Then, the registration network was individually trained by using the registration loss. To optimize the registration network with the segmentation result, we fixed the weights of the segmentation network and fed the output of the registration network into the segmentation network to obtain the segmentation results. The losses of segmentation and registration were both used to optimise the weights of the registration network. After training the registration network, we fine-tuned the segmentation and registration networks for joint optimisation.

In the training process, the training batch size was set to one because the sizes of the input ROIs were variable for different OARs. The initial learning rate was set to 5×10^{-4}. The total number of iterations was 300 epochs, and the learning rate was multiplied by 0.2/150 epochs. The optimizer was Adam. The decay was set to 0.05, and the L2 regular term was set to 1×10^{-5}.

In addition, we trained the model utilising online data augmentation [60] to prevent model overfitting, and the data augmentation methods included image rotation in the range of $-20°$ and $20°$; image scaling in the range of 0.9–1.1; image translation and translation voxels did not exceed 0.1 of the ROI size. It took approximately 36 hours to complete the entire training session.

IV. Experiment Results

A. Datasets

We evaluated the segmentation and registration performance of our proposed framework using three datasets, including an internal dataset and two external datasets. The internal dataset was the 2015 MICCAI Head and Neck Auto Segmentation Challenge dataset (hereinafter referred to as MICCAI 2015). The two external datasets were a public dataset from Automatic Structure Segmentation from Radiotherapy Planning Challenge 2019 (hereinafter referred to as StructSeg 2019), and a dataset collected from Guangzhou Panyu Central Hospital (hereinafter referred to as Panyu), respectively.

The MICCAI 2015 dataset [10] is a public database for computational anatomy (http://www.imagenglab.com/newsite/pddca). A total of 48 CT image volumes were also collected from the Radiation Therapy Oncology Group 0522 study (a multi-institutional clinical trial), together with manual segmentation of 9 OARs, including the left and right parotid glands, brainstem, optic chiasm, left and right optic nerves, mandible, and left and right submandibular glands. The CT image volumes have anisotropic voxel spacing ranging from 0.76 mm to 1.25 mm and inter-slice thickness ranging from 1.25 mm to 3 mm. The MICCAI 2015 dataset was used as the internal dataset. We compared the registration and segmentation performance of our proposed framework with that of state-of-the-art methods and performed ablation experiments on the MICCAI 2015 dataset.

To further evaluate the generalizability of our proposed framework, we used the StructSeg 2019 as an external dataset. To balance the sample size of test set, we randomly selected 15 patients for model performance evaluation, which was equal to the sample size of the test set of MICCAI 2015 dataset. 22 OARs were delineated in each CT image volume, but the submandibular glands were excluded. For this reason, we applied our proposed framework to evaluate the segmentation performance on the left and right parotid glands, brainstem, optic chiasm, left and right optic nerves, and mandible. The voxel spacing of the StructSeg 2019 dataset is around 1 mm and the inter-slice thickness is around 3 mm.

In addition, we collected the CT image volumes of 15 patients with HaN cancer from the Panyu as the second external dataset. The left and right parotid glands, brainstem, optic chiasm, left and right optic nerves, mandible, and left and right submandibular glands were manually segmented by an experienced radiologist referring to the delineated gold standard of MICCAI 2015, and double-checked by an experienced oncologist. The CT image volumes have anisotropic voxel spacing ranging from 0.85 mm to 1.19 mm and inter-slice thickness is 3 mm. The scanner is Philips Brilliance Big Bore.

We trained the model by using the internal dataset, and evaluated the trained model on the external datasets. The model showed good performance on both internal and external datasets.

B. Evaluation Metrics

We used three measurement metrics to quantitatively evaluate the accuracy of the automatic segmentation and registration.

1) DSC: the overlap ratio between the automatic and manual segmentations (the gold standard).

2) Average surface distance (ASD): the average distance between the surfaces of the automatic and manual segmentations (the gold standard). ASD is described as:

$$ASD = \frac{1}{2}\left(\frac{\sum_{z \in P} d(z, G)}{|P|} + \frac{\sum_{u \in G} d(u, P)}{|G|}\right), \tag{11}$$

where $\boldsymbol{d(z, G)}$ is the minimum distance of voxel $\boldsymbol{z}$ on the automatic segmentation organ surface $\boldsymbol{P}$ from all voxels on the gold standard surface $\boldsymbol{G}$. $\boldsymbol{d(u, P)}$ is the minimum distance of voxel $\boldsymbol{u}$ on the $\boldsymbol{G}$ surface from all voxels on the $\boldsymbol{P}$ surface. $| \bullet |$ denotes the total number of voxels in a set.

3) Hausdorff distance (HD): the maximum distance between a point on the automatic and manual segmentation surfaces.

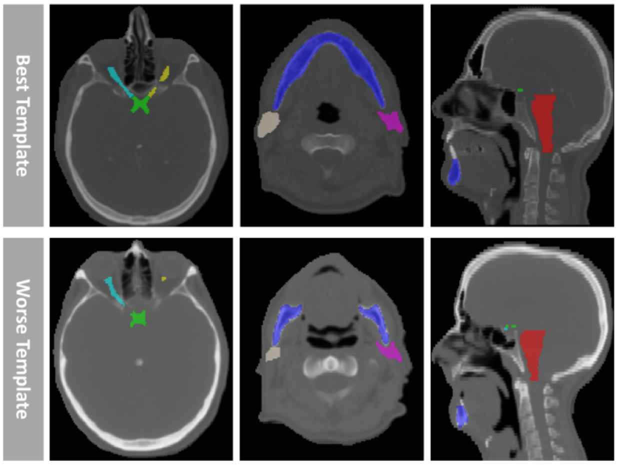

Fig. 5. CT images and OARs of the best and worse templates. The regions with different colors are the manual segmentation of OARs.

HD is described as:

$$HD = max_{a\in A}\left\{min_{b\in B}\left\{d(a,b)\right\}\right\}, \tag{12}$$

where $\boldsymbol{a}$ and $\boldsymbol{b}$ are points of sets $\boldsymbol{A}$ and $\boldsymbol{B}$, respectively, and $\boldsymbol{d(a,b)}$ is the Euclidean distance between these points.

Because HD is sensitive to outliers, such as noise, the upper 5% of distance measures are disregarded, and the 95% distance is often used to evaluate the segmentation accuracy.

C. Results

Using the MICCAI 2015 dataset, we evaluated our proposed method via ablation studies to investigate the contribution of each component. Additionally, we visually inspected the results and compared our proposed framework with the state-of-the-art methods.

1) Impact of Different Templates in the Registration Network: Following the template selection strategy, the templates were selected in the MICCAI 2015 dataset. As shown in Fig. 5, the best template of the MICCAI 2015 dataset was the image volume of patient No. 14 with an average DI of 0.912±0.087, while the worse template was those of patient No. 28 with an average DI of 1.051±0.122. We thought that the registration performance may be affected by different templates, so we compared the registration accuracy by using different templates.

To compare the impact of different templates, we used the registration network without iteratively using context information to generate the registration results. We calculated the average DSC of nine OARs on the test dataset with different templates. The average DSC on the test dataset was 52.6% with the best template and 44.2% with the worse template. As shown in Fig. 6, the results of the best template showed a more accurate registration performance than those of the worse template. For example, in the registration with the worse template, the overlap ratio of the mandible between the registration result and the gold standard was low.

We tried to analyze the registration performance of our method on the cases with abnormal OARs. We selected a case which had large difference from the template (Fig. 7). Note that the DSC of registration and segmentation was 67.4% and 78.6% on this case, respectively. This may indicate that our framework had stable performance even on the abnormal cases.

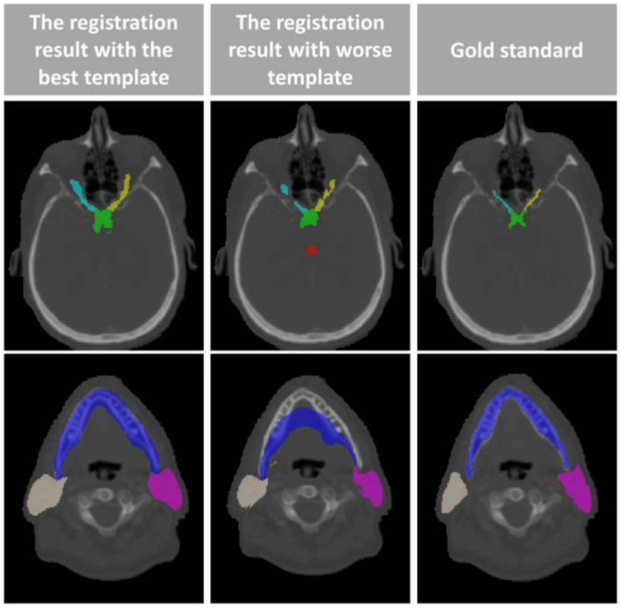
The registration result with the best template
The registration result with worse template
Gold standard

Fig. 6. Registration result with the best and worse templates. The regions with different colors are different OARs.

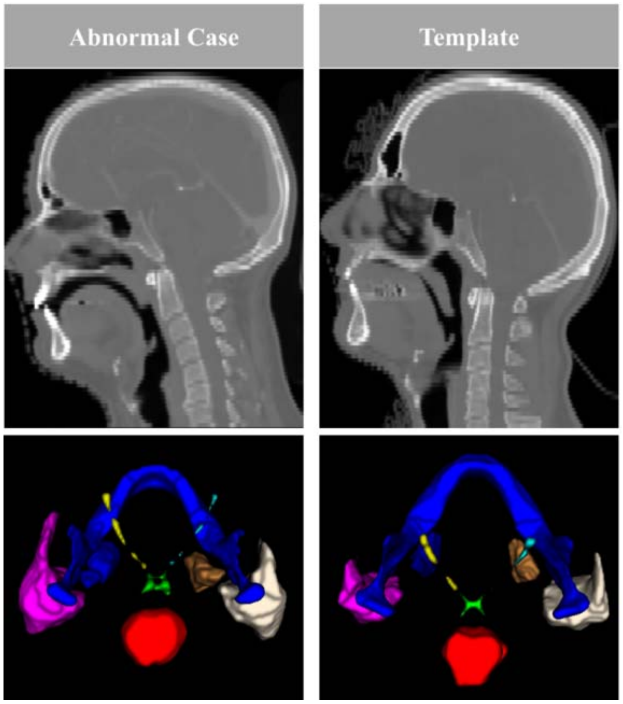

Fig. 7. CT images (the first row) and OARs' gold standards (the second row) of abnormal case and template.

2) Comparison With Common Registration Methods: We compared our registration network in terms of DSC with common registration methods including SyN [61], Voxelmorph [42], Hu *et al.*'s method [44] and U-ResNet [52].

TABLE I
QUANTITATIVE COMPARISON IN TERMS OF AVERAGE DSC FOR REGISTRATION OF NINE OARS BETWEEN OUR PROPOSED NETWORK AND OTHER METHODS

Method	DSC (%)
SyN[61]	55.7
Voxelmorph[42]	48.0
Hu[44]	27.3
U-ReSNet[52]	47.3
Proposed method w.o[a] context	52.6
Proposed method	***69.7***

[a] w.o, without.

TABLE II
SEGMENTATION RESULTS FOR NINE OARS IN TERMS OF AVERAGE/MEDIAN DSC AND AVERAGE 95%HD OF THE MULTI-MODELS, SINGLE-MODEL, AND SINGLE-ROI-MODEL

Model	Average/Median DSC (%)	Average 95%HD (mm)	p-value
Single-model	67.6/67.0	19.6	0.001
Multi-models	**77.4/77.8**	**3.9**	0.427
Single-ROI-model	76.2/76.7	**3.9**	—

The p-values indicate the significance level of Wilcoxon signed-rank tests on pairwise data, comparing the median DSCs of multi-models/single-model against that of the single-ROI model.

Because U-ResNet and Hu *et al.*'s method could not be used for multiple OARs registration, these methods were applied to each OAR separately. As shown in Table I, the average DSC of the nine OARs of our proposed registration method on MICCAI 2015 dataset was 69.7%. In contrast, the average DSC of SyN, Voxelmorph, Hu *et al.*'s method, and U-ReSNet was 55.7%, 48.0%, 27.3% and 43.0%, respectively. The results showed that our registration method exhibited the best performance with significant increase in DSC.

3) Impact of Different Convolution Kernel Sizes of Segmentation Network: Segmentation networks, trained with different kernel sizes, would extract different information from the input images. The networks generated different OAR segmentation masks because of this difference. To avoid the effect of registration errors in segmentation, we trained the segmentation network using the manually selected ROI. The segmentation results (in terms of DSC) with different kernel sizes are shown in Table SI of the Supplementary Materials. The p-values between the segmentation results with multiview and individual kernel sizes were calculated and reported in the rightmost column. The kernel size with joint 3 × 3 × 3 and 3 × 3 × 1 showed the best segmentation performance.

4) Impact of ROI Classification Branch in the Segmentation Network: We trained the segmentation network in three different ways to evaluate the performance of segmentation network with the ROI classification branch on MICCAI 2015 dataset. First, we trained a segmentation network without the ROI classification branch (hereinafter referred to as the single-model) to segment all nine OARs. Second, we trained nine segmentation models (hereinafter referred to as multi-models) to segment nine OARs separately. Third, the single segmentation network with the ROI classification branch (hereinafter referred to as the single-ROI model) was trained to segment nine OARs. The segmentation network was trained with the manually selected ROI to avoid the effect of registration error. As shown in Table II, the results showed that the segmentation performance of multi-models was the best, but the single-ROI model had no significant difference in DSC (p-value=0.427) compared to the multi-models. The DSC (67.6%) of the single-model was the lowest.

5) Impact of Iterative Context Information: Table III shows the OAR segmentation results by using the proposed segmentation network with different numbers of iterations in the context information-based refinement. In this experiment, we did not perform the joint registration and segmentation training, but only evaluated the performance of the segmentation network with individually trained registration model (automatically selected ROI) or with manually selected ROI. The DSC (76.2%) with the manually selected ROI was higher compared to that (75.5%) of the automatically selected ROI (p-value=0.041). After employing the context information, the average DSC of the nine OARs was significantly improved to 78.6% (p-value=0.003). After 2 iterations, the segmentation performance of manually selected ROI was slightly improved compared to the automatically selected ROI. In addition, we compared the segmentation performance between three channels input and single channel input, with manually selected ROIs. The results showed that the three channels input performed better with the DSC increased to 78.8% from 77.5%. These results showed that the context information effectively improved the segmentation performance. However, as the iteration number increased to 2 or more, the DSC and 95%HD were improved slightly.

TABLE III
SEGMENTATION RESULTS FOR NINE OARS IN TERMS OF AVERAGE/MEDIAN DSC AND AVERAGE 95%HD WITH CONTEXT INFORMATION IN DIFFERENT NUMBERS OF ITERATIONS (0, 1, 2, 3) OF THE PROPOSED SEGMENTATION NETWORK

Iterations	ROI selection	Average/Median DSC (%)	Average 95%HD (mm)	p-value
0	Manual	76.2/76.7	3.9	—
0	Automatic	75.5/75.8	4.2	0.041
1	Automatic	78.6/78.8	3.4	0.003
2	Manual, Single channel input	77.5/78.5	3.4	0.014
2	Automatic	78.7/79.4	3.4	0.003
2	Manual	**78.8/79.4**	**3.2**	0.002
3	Automatic	78.7/79.2	3.4	0.003

The p-values indicate the significance level of Wilcoxon signed-rank tests on pairwise data, comparing the median DSCs of other iterations/ROI selection methods against that of iteration 0 with manual ROI selection.

6) Impact of Shape Correction: As shown in Fig. 8, the optic chiasm segmentation result was incorrect in shape before shape correction. The anatomy of the optic chiasm segmentation result was not accepted by doctors because the optic chiasm should be X-shaped. After shape correction, the segmentation results were similar to the gold standard in shape. For the test set of MICCAI 2015 dataset, the average DSC of the optic chiasm increased from 61.7% to 64.3% (Table IV).

7) Comparison With the State-of-the-Art OARs Segmentation Methods: We jointly trained the registration network and the segmentation network, and compared the segmentation results with those of nine commonly used state-of-the-art methods,

TABLE IV

QUANTITATIVE COMPARISON IN TERMS OF DSC FOR SEGMENTATION OF NINE OARS BETWEEN NINE STATE-OF-THE-ART METHODS AND OUR PROPOSED FRAMEWORK. VALUES ARE GIVEN IN %

Method	Brain Stem	Mandible	Optic Nerves	Optic Chiasm	Parotid Glands	Submandibular Glands	Average DSC
Tong[33]	87.0±3.0	93.6±1.2	L[d]: 65.3±5.8 R[e]: 68.9±4.7	58.4±10.3	L: 83.9±2.9 R: 83.5±2.3	L: 76.7±7.3 R: 81.3±6.5	77.6
Wang[24]	90.3±3.8	94.4±1.3	— —	—	L: 82.3±5.2 R: 82.9±6.4	— —	—
Mannion[22]	88.0±3.5	92.5±1.0	70.5±4.0	40.0±22.5	84.0±5.0	78.0±8.5	75.5
Ren[31]	—	—	L: 72.0±8.0 R: 70.0±9.0	58.0±17.0	— —	— —	—
Tang[39]	87.5±2.5	95.0±0.8	L: 74.8±7.1 R: 72.3±5.9	61.5±10.2	**L: 88.7±1.9** R: 87.5±5.0	**L: 82.3±5.2** R: 81.5±4.5	81.2
Zhu[35]	86.7±2.0	92.5±2.0	L: 72.1±6.0 R:70.6±10.0	53.2±15.0	L: 88.1±2.0 R: 87.4±4.0	L: 81.4±4.0 R: 81.3±4.0	79.3
Guo[40]	87.6±2.8	**95.1±1.1**	**L: 75.3±7.1** **R: 74.6±5.2**	64.9±8.8	L: 88.2±3.2 **R: 88.2±5.2**	L: 84.2±7.3 R: 83.8±6.9	**82.4**
Wang[36]	87.5±2.2	93.0±1.9	L: 73.7±7.6 R: 73.6±8.8	45.1±17.2	L: 86.4±2.6 R: 84.8±7.0	L: 75.8±14.7 **R: 84.8±7.0**	78.3
Liang[38]	92.3±1.0	94.1±0.7	L: 73.8±4.6 R: 73.4±5.1	**71.3±8.3**	L: 88.2±1.3 R: 87.0±1.5	L: 81.5±2.9 R: 80.0±3.4	**82.4**
Liang[38] w.o[a] ACM[b]	89.5±1.1	91.6±1.4	L: 64.8±10.3 R: 65.4±6.8	65.9±12.7	L: 85.0±3.7 R: 84.8±4.1	L: 77.1±7.1 R: 75.6±6.3	77.7
Proposed method w.o SC[c]	87.9±2.4	91.6±2.1	L: 67.7±6.5 R: 70.6±7.1	61.7±17.6	L: 88.4±1.5 R: 87.8±2.0	L: 80.1±7.1 R: 77.6±4.5	79.3
Proposed method	*87.9±2.4*	*91.6±2.1*	*L: 67.7±6.5* *R: 70.6±7.1*	*64.3±13.7*	*L: 88.4±1.5* *R: 87.8±2.0*	*L: 80.1±7.1* *R: 77.6±4.5*	*79.6*

[a] w.o, without. [b] ACM, auto context model. [c] SC, shape correction. [d] L, left. [e] R, right.

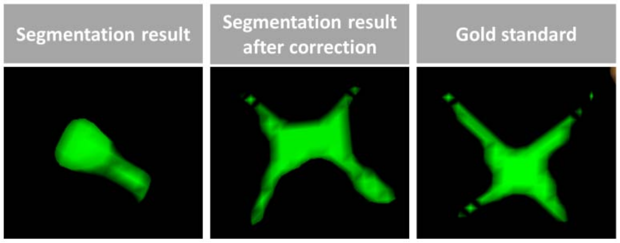

Fig. 8. Visualization of the optic chiasm segmentation results after shape correction.

as summarised in Tables IV and V. The DSC of our proposed method was slightly lower than that of Liang *et al.*'s method, Guo *et al.*'s method and Tang *et al.*'s method. However, the ACM strategy in Liang *et al.*'s method led to a significant increase in the model parameters and model number. As shown in Tables IV and V, the average DSC and average ASD of our proposed framework were both better than that of Liang *et al.*'s method without the ACM strategy. As shown in Table VI, the model parameters of Guo *et al.*'s method and Tang *et al.*'s method were much higher than our method. In addition, neural architecture search (NAS) was applied in Guo *et al.*'s method. To our best knowledge, NAS required a large GPU memory. Guo *et al.* [40] trained their proposed model on NVIDIA Quadro RTX 8000 with 48G GPU memory. Meanwhile, our method had the advantage of simultaneously performing registration and segmentation.

As shown in Table VI, the number of parameters and trained models of our proposed framework was the lowest except AnatomyNet. In addition, our framework was the fastest among the existing methods which realize registration task. Although Liang *et al.* did not present the inference time in their paper, we speculated that the inference time of Liang *et al.*'s method would be more than that of our proposed method because of the following reasons. 1) The 2D CNN must predict multiple times to complete 3D volume segmentation, and 3D CNN only must predict one time. 2) The multiview models of Liang *et al.*'s method needed three times the individual model time spent. 3) The ACM strategy involved multiple models. Liang *et al.* used nine models in total, which increased the inference time.

8) Segmentation Performance on the External Datasets: To evaluate the generalizability of our proposed method, we employed a model that was trained using the MICCAI 2015 dataset to segment the OARs of the external datasets. The segmentation performance is listed in Table VII. The average DSC and average ASD of Panyu dataset were closed to those of the MICCAI 2015 dataset. However, the performance of StructSeg 2019 was sharply dropped. The details for analysing the differences of gold standards between internal dataset and external datasets can be found in Section. B of the *Supplementary Materials*.

Lei *et al.* [41] mixed MICCAI 2015 dataset, StructSeg 2019 and locally collected dataset to train they proposed network. The segmentation performance is shown in Table VII. The DSCs of Lei *et al.*'s method were higher than our method. However, we only used MICCAI 2015 dataset to train our network and the StructSeg 2019 was used as an external evaluation dataset.

V. DISCUSSION AND CONCLUSION

We proposed a framework for multiple OAR registration and segmentation with single model using multiview image information from the CT image volumes. The registration network was used to locate OARs on the new image volumes

TABLE V
Quantitative Comparison in Terms of ASD for Segmentation of Nine OARS Between Three State-of-the-Art Methods and Our Proposed Framework. Values are Given in MM

Method	Brain Stem	Mandible	Optic Nerves	Optic Chiasm	Parotid Glands	Submandibular Glands	Average ASD
Tong[33]	1.17±0.56	0.37±0.11	L[c]: 1.14±0.75 R[d]: 1.15±0.65	0.65±0.21	L: 0.96±0.34 R: 1.12±0.56	L: 0.90±0.46 R: 1.33±0.57	0.92
Wang[24]	0.91±0.32	0.43±0.12	— —	—	L: 1.85±0.93 R: 1.81±0.63	— —	—
Liang[38]	**0.85±0.15**	**0.28±0.14**	**0.88±0.52**	**0.48±0.36**	**0.69±0.16**	**0.98±0.38**	**0.69**
Liang[38] w.o[a] ACM[b]	1.08±0.21	0.72±0.17	1.33±0.63	0.78±0.53	1.05±0.25	2.06±0.71	1.17
Proposed method	*1.28±0.45*	*0.56±0±27*	*L: 1.06±0.51* *R: 1.06±0.49*	*0.76±0.44*	*L: 0.86±0.24* *R: 1.02±0.38*	*L: 1.38±0.89* *R: 1.52±0.60*	*1.05*

[a] w.o, without. [b] ACM, auto context model. [c] L, left. [d] R, right.

TABLE VI
Brief Comparison of the Nine State-of-the-Art Methods and Our Proposed Framework for OARS Segmentation

Method	ACM[a]	Multiple OARs segmentation	Registration	Model structures	Model parameters (ppproximate)	Model number	Inference time
Tong[33]		✓		3D CNN[b] and shape represent model	120 million	2	9.5 s
Wang[24]	✓		✓	Vertex regression forests	—	—	6480 s
Mannion[22]	✓		✓	Active appearance model	—	—	1800 s
Ren[31]	✓			3D CNN	2.2 million	9	930 s
Tang[39]		✓		3D CNN	22.8 million	2	2.36 s
Zhu[35]		✓		AnatomyNet	0.7 million	1	4.52 s
Guo[40]		✓		NAS[c] CNN	30.5 million	4	20 s
Wang[36]		✓		Two stage cascade 3D CNN	36 million	2	108 s
Liang[38]	✓	✓		Multi-view 2D CNN	120 million	9	—
Liang[38] w.o[d] ACM		✓		Multi-view 2D CNN	40 million	3	—
Proposed method		✓	✓	*3D registration CNN + Multi-view segmentation CNN*	*11.2 million*	*2*	*8.6 s*

[a] ACM, auto context model. [b] CNN, convolutional neural network. [c] NAS, neural architecture search. [d] w.o, without.

TABLE VII
Segmentation Performance in Terms of DSC and ASD Obtained by Our Proposed Framework on OARS in the External Dataset and the Comparison With the Segmentation Performance of Another Method

Method	Dataset	Metric	Brain Stem	Mandible	Optic Nerves	Optic Chiasm	Parotid Glands	Submandibular Glands
Proposed method	Panyu	DSC (%)	95.7±3.7	84.8±3.0	L[a]: 82.4±13.0 R[b]: 84.3±14.1	43.4±15.3	L: 96.2±5.3 R: 94.6±6.0	L: 84.6±6.7 R: 80.8±9.7
		ASD (mm)	0.56±0.49	1.19±0.33	L: 0.26±0.19 R: 0.16±0.15	1.04±0.89	L: 0.38±0.35 R: 0.53±0.63	L: 1.00±1.01 R: 0.84±0.57
	StructSeg 2019	DSC (%)	76.9±4.9	80.7±4.9	L: 49.9±14.0 R: 53.4±10.4	21.1±13.1	L: 80.2±7.0 R: 80.2±5.5	— —
		ASD (mm)	1.83±0.21	1.67±0.54	L: 1.84±0.96 R: 2.36±1.57	1.73±1.45	L: 1.27±0.37 R: 1.41±0.57	— —
Lei[41]	StructSeg 2019, MICCAI 2015, Locally collected dataset	DSC (%)	87.4±2.6	90.0±4.2	L: 62.1±12.5 R: 64.2±13.0	29.0±23.3	L: 84.7±5.7 R: 84.6±4.5	— —

[a] L, left. [b] R, right.

using the location information of the template. The ROIs of each OAR were generated from the CT image volumes based on the registration results. The segmentation network generated segmentation results based on the ROIs. Furthermore, the registration and segmentation performances were both significantly improved by utilizing the context information with an iterative strategy. Finally, the shape correction method further improved the segmentation of optic chiasm.

The template selection strategy was used to select an optimal template in the training set. We adopted a 3D similarity transform, which only contained image rotation, translation, and scaling. In contrast to the affine transform, the 3D similarity transform did not change the shape and relative location of the OARs. Therefore, we compared the structural differences between different patients and position differences of various OARs. In addition, the designed metric DI was used to select the best template effectively. As shown in Fig. 6, the best template led to a more accurate result compared to the worse template with the same registration method.

The registration network achieved accurate OAR localisation in the whole-volume CT image. As shown in Table I, our proposed registration network obtained the highest DSC, which meant that the proposed registration method was more accurate in localizing the OAR compared to the commonly used methods. In addition, DSC increased from 52.6% to 69.7% by using context information. The OAR mask of the previous registration could be used as a landmark to improve the registration performance because the landmark alignment was easier compared to the CT image volume.

The segmentation network combined the 3D and 2D axial information of the CT image with two different convolution kernel sizes. As shown in Table SI of the *Supplementary Materials*, a kernel size of 3 × 3 × 1, which used the 2D axial CT information, showed the best segmentation performance compared to the kernels of 3 × 1 × 3 and 1 × 3 × 3. Therefore, the 2D axial information may be more useful for HaN OAR segmentation. This might be because the axial images had a higher spatial resolution, and the gold standards of OARs were manually contoured in the axial view. A kernel size of 3 × 3 × 3, which used the 3D information of CT, exhibited a better segmentation performance compared to the kernel size of 3 × 3 × 1. This might be because the 3D information included the context information in the z-dimension. Indeed, we simultaneously implemented kernel sizes of 3 × 3 × 3 and 3 × 3 × 1 in parallel to extract the 3D and 2D axial features, and achieved the best segmentation performance, as shown in Table SI of the *Supplementary Materials*.

Three channels input improved the segmentation performance compared to single channel input (Table III). Because the soft tissues showed low contrast in CT images, the edges of these organs were hard to be recognized. The oncologists usually delineate different OARs with different WW/WL for better recognition. For this reason, we fed the CT image volumes with three different WW/WL as three input channels into the segmentation network for more accurate segmentation.

The ROI classification branch also improved the segmentation performance. From Table II, the highest DSC was obtained from the multi-models because each OAR was segmented by individual model. Each model focused on the features of a type of OAR, and the size of inputted ROI was identical. The single-model had to recognise the nine different OARs with different ROI sizes. This turned out to be more difficult compared to the multi-models segmentation. As shown in Table II, the DSC of single-model was lower than that of the multi-models. Hence, we added an ROI classification branch to improve the feature identification among different ROIs. We found no statistically significant difference in terms of DSC (p-value=0.427) between the single-ROI model and the multi-models.

Compared with the ACM strategy by cascading multiple individual models for using context information, our strategy used context information to improve the accuracy of registration and segmentation iteratively without increasing the number of models and parameters. We fed the results from the previous iterations to use context information. With this strategy, the registration network could generate a deformable field based on the previous registration result, which could reduce the complexity of deformable field. The segmentation network could refine the OAR segmentation based on previous segmentation results. Simultaneously, the segmentation results from the previous iterations could be used to correct the registration results, and hence decreased the OAR locating errors.

Optic chiasm segmentation often faces two challenges: a wrong, rectangular or elliptical shape after automatic segmentation; and a severe under-segmentation. Our shape-correction method was designed to correct these by using prior information, and the results showed that the DSC increased from 61.7% to 64.3%.

Our proposed framework was a fast and lightweight framework for OAR registration and segmentation of HaN CT images. Compared with the previous methods, the segmentation performance of our proposed method was not the best but close to the best results. However, our proposed method was the most lightweight and the second fastest method. Although, AnatomyNet was faster and more lightweight than our method, it was hard to train well [40] and it needed three datasets to gain DSC of 79.3% which was slightly lower than that of our method. Moreover, our proposed method could also generate registration results, which could assist the less experienced oncologists in delineating OARs.

Our proposed framework had a stable generalizability. As shown in Table VII, the results showed a stable segmentation performance on the Panyu dataset, but a bad performance on the StructSeg 2019 dataset. The reason of the bad performance on StructSeg 2019 may be the different delineation standard, for which the details can be found in Section B. 2) of the *Supplementary Materials*. Discrepancies still existed between the segmentation results for MICCAI 2015 and Panyu dataset because of the different WW/WL in delineation stage, and the details of the analysis can be found in Section B. 1) of the *Supplementary Materials*.

Although our proposed method exhibited good segmentation and registration performances with a lightweight model, we found that it might not work appropriately in some cases. Further studies are warranted to improve our method. First, we only selected a patient as the template, which may not be the best strategy. Since different patients have different shapes of OARs, only one single patient may not reflect various OARs especially the soft tissues. For example, as shown in Fig. 5, the brain stem of worse template was longer than the best template. For this reason, we may construct an average template with more patient's scans. Second, we may design a strategy to double correct the segmentation results in the stage of using context information. We may correct the shape of segmentation results with the prior information in the loop, and the corrected segmentation results may be fed into the network as context information. Third, we may combine the shape information of OAR template to improve the accuracy of segmentation results further. Although our shape correction method could efficiently correct the shape of optic chiasm, it could not perform well on the other OARs. Fourth, the weights of multiple loss functions we used may not the best combination, though we have tried some different combinations. We may select a best combination by using a learning-based method in future studies.

References

[1] A. Jemal, F. Bray, M. M. Center, J. Ferlay, E. Ward, and D. Forman, "Global cancer statistics," *CA, Cancer J. Clinicians*, vol. 61, no. 2, pp. 69–90, 2011.

[2] C. Fitzmaurice *et al.*, "Global, regional, and national cancer incidence, mortality, years of life lost, years lived with disability, and disability-adjusted life-years for 29 cancer groups, 1990 to 2016: A systematic analysis for the global burden of disease study," *J. Amer. Med. Assoc. Oncol.*, vol. 4, no. 11, pp. 1553–1568, 2018.

[3] E. K. Hansen, M. K. Bucci, J. M. Quivey, V. Weinberg, and P. Xia, "Repeat CT imaging and replanning during the course of IMRT for head-and-neck cancer," *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 64, no. 2, pp. 355–362, 2006.

[4] W. F. A. R. Verbakel, J. P. Cuijpers, D. Hoffmans, M. Bieker, B. J. Slotman, and S. Senan, "Volumetric intensity-modulated arc therapy vs. conventional IMRT in head- and-neck cancer: A comparative planning and dosimetric study," *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 74, no. 1, pp. 252–259, 2009.

[5] X. Han *et al.*, "Atlas-based auto-segmentation of head and neck CT images," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2008*. Berlin, Germany: Springer, 2008, pp. 434–441.

[6] D. N. Teguh *et al.*, "Clinical validation of atlas-based auto-segmentation of multiple target volumes and normal tissue (swallowing/mastication) structures in the head and neck," *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 81, no. 4, pp. 950–957, Nov. 2011.

[7] J. Breunig *et al.*, "A system for continual quality improvement of normal tissue delineation for radiation therapy treatment planning," *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 83, no. 5, pp. e703–e708, Aug. 2012.

[8] R. Sims *et al.*, "A pre-clinical assessment of an atlas-based automatic segmentation tool for the head and neck," *Radiotherapy Oncol.*, vol. 93, no. 3, pp. 474–478, Dec. 2009.

[9] F. Berrino and G. Gatta, "Variation in survival of patients with head and neck cancer in Europe by the site of origin of the tumours," *Eur. J. Cancer*, vol. 34, no. 14, pp. 2154–2161, Dec. 1998.

[10] P. F. Raudaschl *et al.*, "Evaluation of segmentation methods on head and neck CT: Auto-segmentation challenge 2015," *Med. Phys.*, vol. 44, no. 5, pp. 2020–2036, 2017.

[11] T. Zhang, Y. Chi, E. Meldolesi, and D. Yan, "Automatic delineation of on-line head- and-neck computed tomography images: Toward on-line adaptive radiotherapy," *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 68, no. 2, pp. 522–530, Jun. 2007.

[12] A. Isambert *et al.*, "Evaluation of an atlas-based automatic segmentation software for the delineation of brain organs at risk in a radiation therapy clinical context," *Radiotherapy Oncol.*, vol. 87, no. 1, pp. 93–99, Apr. 2008.

[13] O. Commowick, V. Grégoire, and G. Malandain, "Atlas-based delineation of lymph node levels in head and neck computed tomography images," *Radiotherapy Oncol.*, vol. 87, no. 2, pp. 281–289, May 2008.

[14] C. Leavens *et al.*, "Validation of automatic landmark identification for atlas-based segmentation for radiation treatment planning of the head-and-neck region," *Proc. SPIE*, vol. 6914, Feb. 2008, Art. no. 69143G.

[15] P. C. Levendag *et al.*, "Atlas based auto-segmentation of CT images: Clinical evaluation of using auto-contouring in high-dose, high-precision radiotherapy of cancer in the head and neck," *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 72, no. 1, p. S401, Sep. 2008.

[16] A. Chen, M. A. Deeley, K. J. Niermann, L. Moretti, and B. M. Dawant, "Combining registration and active shape models for the automatic segmentation of the lymph node regions in head and neck CT images," *Med. Phys.*, vol. 37, no. 12, pp. 6338–6346, Nov. 2010.

[17] A. A. Qazi, V. Pekar, J. Kim, J. Xie, S. L. Breen, and D. A. Jaffray, "Auto-segmentation of normal and target structures in head and neck CT images: A feature-driven model-based approach," *Med. Phys.*, vol. 38, no. 11, pp. 6160–6170, Oct. 2011.

[18] V. Fortunati *et al.*, "Hyperthermia critical tissues automatic segmentation of head and neck CT images using atlas registration and graph cuts," in *Proc. 9th IEEE Int. Symp. Biomed. Imag. (ISBI)*, May 2012, pp. 1683–1686.

[19] V. Fortunati *et al.*, "Tissue segmentation of head and neck CT images for treatment planning: A multiatlas approach combined with intensity modeling," *Med. Phys.*, vol. 40, no. 7, 2013, Art. no. 071905.

[20] 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," *Med. Phys.*, vol. 41, no. 5, Apr. 2014, Art. no. 051910.

[21] C. Wachinger, K. Fritscher, G. Sharp, and P. Golland, "Contour-driven atlas-based segmentation," *IEEE Trans. Med. Imag.*, vol. 34, no. 12, pp. 2492–2505, Dec. 2015.

[22] 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.*, Jan. 2016. [Online]. Available: https://www.midasjournal.org/browse/publication/967

[23] H. Xu, A. A. Henry, M. Robillard, M. Amessis, and Y. M. Kirova, "The use of new delineation tool 'MIRADA' at the level of regional lymph nodes, step-by-step development and first results for early-stage breast cancer patients," *Brit. J. Radiol.*, vol. 91, no. 1090, 2018, Art. no. 20180095.

[24] Z. Wang, L. Wei, L. Wang, Y. Gao, W. Chen, and D. Shen, "Hierarchical vertex regression-based segmentation of head and neck CT images for radiotherapy planning," *IEEE Trans. Image Process.*, vol. 27, no. 2, pp. 923–937, Feb. 2018.

[25] C. Tam, X. Yang, S. Tian, X. Jiang, J. Beitler, and S. Li, "Automated delineation of organs-at-risk in head and neck CT images using multi-output support vector regression," *Proc. SPIE*, vol. 10578, Mar. 2018, Art. no. 1057824.

[26] X. Wu *et al.*, "Auto-contouring via automatic anatomy recognition of organs at risk in head and neck cancer on CT images," *Proc. SPIE*, vol. 10576, Mar. 2018, Art. no. 1057617.

[27] Y. Tong *et al.*, "Hierarchical model-based object localization for auto-contouring in head and neck radiation therapy planning," *Proc. SPIE*, vol. 10578, Mar. 2018, Art. no. 1057822.

[28] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*. Cham, Switzerland: Springer, 2015, pp. 234–241.

[29] K. Fritscher, P. Raudaschl, P. Zaffino, M. F. Spadea, G. C. Sharp, and R. Schubert, "Deep neural networks for fast segmentation of 3D medical images," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2016*. Cham, Switzerland: Springer, 2016, pp. 158–165.

[30] B. Ibragimov and L. Xing, "Segmentation of organs-at-risks in head and neck CT images using convolutional neural networks," *Med. Phys.*, vol. 44, no. 2, pp. 547–557, 2017.

[31] X. Ren *et al.*, "Interleaved 3D-CNNs for joint segmentation of small-volume structures in head and neck CT images," *Med. Phys.*, vol. 45, no. 5, pp. 2063–2075, 2018.

[32] A. Hänsch *et al.*, "Comparison of different deep learning approaches for parotid gland segmentation from CT images," *Proc. SPIE*, vol. 10575, Feb. 2018, Art. no. 1057519.

[33] 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," *Med. Phys.*, vol. 45, no. 10, pp. 4558–4567, 2018.

[34] S. Liang *et al.*, "Deep-learning-based detection and segmentation of organs at risk in nasopharyngeal carcinoma computed tomographic images for radiotherapy planning," *Eur. Radiol.*, vol. 29, no. 4, pp. 1961–1967, 2019.

[35] W. Zhu *et al.*, "AnatomyNet: Deep learning for fast and fully automated whole-volume segmentation of head and neck anatomy," *Med. Phys.*, vol. 46, no. 2, pp. 576–589, 2019.

[36] 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*, vol. 7, pp. 144591–144602, 2019.

[37] Y. Gao *et al.*, "FocusNet: Imbalanced large and small organ segmentation with an end-to-end deep neural network for head and neck CT images," in *Medical Image Computing and Computer Assisted Intervention—MICCAI 2019*. Cham, Switzerland: Springer, 2019, pp. 829–838.

[38] S. Liang, K.-H. Thung, D. Nie, Y. Zhang, and D. Shen, "Multi-view spatial aggregation framework for joint localization and segmentation of organs at risk in head and neck CT images," *IEEE Trans. Med. Imag.*, vol. 39, no. 9, pp. 2794–2805, Sep. 2020.

[39] H. Tang *et al.*, "Clinically applicable deep learning framework for organs at risk delineation in CT images," *Nature Mach. Intell.*, vol. 1, no. 10, pp. 480–491, Oct. 2019.

[40] D. Guo *et al.*, "Organ at risk segmentation for head and neck cancer using stratified learning and neural architecture search," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 4223–4232.

[41] W. Lei *et al.*, "Automatic segmentation of organs-at-risk from head-and-neck CT using separable convolutional neural network with hard-region-weighted loss," *Neurocomputing*, vol. 442, pp. 184–199, Jun. 2021.

[42] G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, and A. V. Dalca, "VoxelMorph: A learning framework for deformable medical image registration," *IEEE Trans. Med. Imag.*, vol. 38, no. 8, pp. 1788–1800, Aug. 2019.

[43] A. Hering, S. Kuckertz, S. Heldmann, and M. P. Heinrich, "Enhancing label-driven deep deformable image registration with local distance metrics for state-of-the-art cardiac motion tracking," in *Bildverarbeitung für die Medizin 2019*. Wiesbaden, Germany: Springer, 2019, pp. 309–314.

[44] Y. Hu *et al.*, "Weakly-supervised convolutional neural networks for multimodal image registration," *Med. Image Anal.*, vol. 49, pp. 1–13, Oct. 2018.

[45] S. Parisot, H. Duffau, S. Chemouny, and N. Paragios, "Joint tumor segmentation and dense deformable registration of brain MR images," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2012*. Berlin, Germany: Springer, 2012, pp. 651–658.

[46] A. Gooya *et al.*, "GLISTR: Glioma image segmentation and registration," *IEEE Trans. Med. Imag.*, vol. 31, no. 10, pp. 1941–1954, Oct. 2012.

[47] P. P. Wyatt and J. A. Noble, "MAP MRF joint segmentation and registration of medical images," *Med. Image Anal.*, vol. 7, no. 4, pp. 539–552, 2003.

[48] B. Li *et al.*, "A hybrid deep learning framework for integrated segmentation and registration: Evaluation on longitudinal white matter tract changes," in *Medical Image Computing and Computer Assisted Intervention—MICCAI 2019*. Cham, Switzerland: Springer, 2019, pp. 645–653.

[49] Z. Xu and M. Niethammer, "DeepAtlas: Joint semi-supervised learning of image registration and segmentation," in *Medical Image Computing and Computer Assisted Intervention—MICCAI 2019*. Cham, Switzerland: Springer, 2019, pp. 420–429.

[50] T. Estienne *et al.*, "Deep learning-based concurrent brain registration and tumor segmentation," *Frontiers Comput. Neurosci.*, vol. 14, p. 17, Mar. 2020.

[51] D. Mahapatra, Z. Ge, S. Sedai, and R. Chakravorty, "Joint registration and segmentation of Xray images using generative adversarial networks," in *Machine Learning in Medical Imaging*. Cham, Switzerland: Springer, 2020, pp. 73–80.

[52] T. Estienne *et al.*, "U-ReSNet: Ultimate coupling of registration and segmentation with deep nets," in *Medical Image Computing and Computer Assisted Intervention—MICCAI 2019*. Cham, Switzerland: Springer, 2019, pp. 310–319.

[53] L. Beljaards, M. S. Elmahdy, F. Verbeek, and M. Staring, "A cross-stitch architecture for joint registration and segmentation in adaptive radiotherapy," presented at the 3rd Conf. Med. Imag. Deep Learn., Montreal, QC, Canada, 2020. [Online]. Available: http://proceedings.mlr.press/v121/beljaards20a.html

[54] F. P. M. Oliveira and J. M. R. S. Tavares, "Medical image registration: A review," *Comput. Methods Biomech. Biomed. Eng.*, vol. 17, no. 2, pp. 73–93, 2014.

[55] J. Engel, T. Schöps, and D. Cremers, "LSD-SLAM: Large-scale direct monocular SLAM," in *Computer Vision—ECCV 2014*. Cham, Switzerland: Springer, 2014, pp. 834–849.

[56] J. Hu, L. Shen, and G. Sun, "Squeeze- and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, Jun. 2018, pp. 7132–7141.

[57] B. Xu, N. Wang, T. Chen, and M. Li, "Empirical evaluation of rectified activations in convolutional network," 2015, *arXiv:1505.00853*.

[58] Z. Tu and X. Bai, "Auto-context and its application to high-level vision tasks and 3D brain image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 10, pp. 1744–1757, Oct. 2010.

[59] M.-K. Hu, "Visual pattern recognition by moment invariants," *IRE Trans. Inf. Theory*, vol. 8, no. 2, pp. 179–187, 1962.

[60] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, 2012, pp. 1097–1105.

[61] B. B. Avants, C. L. Epstein, M. Grossman, and J. C. Gee, "Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain," *Med. Image Anal.*, vol. 12, no. 1, pp. 26–41, 2008.