Title of the manuscript:

Comprehensive and Clinically Accurate Head and Neck Organs at Risk Delineation via Stratified Deep Learning: A Large-scale Multi-Institutional Study

Dazhou Guo[1†], Jia Ge[2†], Xianghua Ye[2†], Senxiang Yan[2], Yi Xin[3], Yuchen Song[2], Bing-shen Huang[4], Tsung-Min Hung[4], Zhuotun Zhu[5], Ling Peng[6], Yanping Ren[7], Rui Liu[8], Gong Zhang[9], Mengyuan Mao[10], Xiaohua Chen[11], Zhongjie Lu[2], Wenxiang Li[2], Yuzhen Chen[13], Lingyun Huang[3], Jing Xiao[3], Adam P. Harrison[1], Le Lu[1], Chien-Yu Lin[4,12*], Dakai Jin[1*], Tsung-Ying Ho[13*]

[1]PAII Inc., Bethesda, MD, USA
[2]Department of Radiation Oncology, The First Affiliated Hospital Zhejiang University, Hangzhou, China
[3]Ping An Technology, Shenzhen, China
[4]Department of Radiation Oncology, Chang Gung Memorial Hospital, Linkou, Taiwan, ROC
[5]Department of Computer Science, Johns Hopkins University, Baltimore, MD, USA
[6]Department of Respiratory Disease, Zhejiang Provincial People's Hospital, Hangzhou, Zhejiang, China
[7]Department of Radiation Oncology, Huadong Hospital Affiliated to Fudan University, Shanghai, China
[8]Department of Radiation Oncology, First Affiliated Hospital of Xi'an Jiaotong University, Xi'an, China
[9]Department of Radiation Oncology, People's Hospital of Shanxi Province, Shanxi, China
[10]Department of Radiation Oncology, Nanfang Hospital, Southern Medical University, Guangzhou, China
[11]Department of Radiation Oncology, The First Hospital of Lanzhou University, Lanzhou, Gansu, China
[12]Particle Physics and Beam Delivery Core Laboratory, Chang Gung Memorial Hospital and Chang Gung University, Taoyuan, Taiwan, ROC
[13]Department of Nuclear Medicine, Chang Gung Memorial Hospital, Linkou, Taiwan, ROC

D. Guo, J. Ge and X. Ye are co-first authors of this work.

C. Lin (cylin@cgmh.org.tw), D. Jin (dakai.jin@gmail.com) and T. Ho (tyho@cgmh.org.tw) are co-corresponding authors of this work.

Abstract (~150 words, but not a hard limit)

Accurate organ at risk (OAR) segmentation is critical to reduce the radiotherapy post-treatment complications. Consensus guidelines recommend a set of more than 40 OARs in the head and neck (H&N) region, however, due to the labor-prohibiting nature of this task, most institutions choose a substantially simplified protocol by delineating a (much) smaller subset of OARs and neglecting the dose distributions associated with other OARs. In this work we propose a novel, automated and highly effective stratified OAR segmentation (SOARS) system using deep learning to precisely delineate a comprehensive set of 42 H&N OARs. SOARS stratifies 42 OARs into anchor, mid-level, and small & hard subcategories, with specifically derived neural network architectures for each category by neural architecture search (NAS) principles. We built SOARS models using 176 training patients in an internal institution and independently evaluated on 1327 external patients across six different institutions. It consistently outperformed other state-of-the-art methods by at least 3-5% in Dice score for each institutional evaluation (up to 36% relative error reduction in other metrics). More importantly, extensive multi-user studies evidently demonstrated that 98% of the SOARS predictions need only very minor or no revisions for direct clinical acceptance (saving 90% radiation oncologists workload), and their segmentation and dosimetric accuracy are within or smaller than the inter-user variation. These findings confirmed the strong clinical applicability of SOARS for the OAR delineation process in H&N cancer radiotherapy workflows, with improved efficiency, comprehensiveness, and quality.

Head and neck (H&N) cancer is one of the most common cancers worldwide [1]. Radiation therapy (RT) is an important and effective treatment for H&N cancer [2]. In RT, the radiation dose to normal anatomical structures, i.e., organs at risk (OARs), needs to be limited to reduce post-treatment complications, such as dry mouth, swallowing difficulties, visual damage, and cognitive decline [3-6]. This requirement demands accurate OAR delineation on the planning computed tomography (pCT) images used to configure the radiation dosage treatment. Recent consensus guidelines recommend a set of more than 40 OARs in the H&N region [7]. Nevertheless, precise manual delineation of this quantity of OARs is an overwhelmingly demanding task that requires great clinical expertise and time efforts, e.g., > 3 hours for 24 OARs [8]. Due to the factors of patient overload and shortage of experienced physicians, long patient waiting times and/or undesirably inaccurate RT delineations are more common than necessary, reducing the treatment efficacy and safety [9]. To shorten time expenses, many institutions choose a simplified (sometimes overly simplified) OAR protocol by contouring a small subset of OARs (e.g., only the OARs closest to the tumor). Dosimetric information cannot be recorded for non-contoured OARs although it is clinically important to track for analysis of post-treatment side effects [10]. Automatic and accurate segmentation of a comprehensive set of H&N OARs is of great clinical benefit in this context.

OARs are spatially densely distributed in the H&N region and often have complex anatomical shapes, large size variations, and low CT contrasts. Conventional atlas-based methods previously enjoyed a prominent history [11-15], but significant amounts of editing efforts were found to be unavoidable [8,16]. Atlas-based methods heavily rely on the accuracy and reliability of deformable image registration that can be very challenging due to OARs' large shape variations, normal tissue removal, tumor growth, and image acquisition differences. Volumetric deformable registration methods often take many minutes or even hours to compute.

Deep learning approaches have shown substantial improvements for improving segmentation accuracy and efficiency as compared to atlas-based methods [17]. After early patch-based representation [18], fully convolutional network is the dominant formulation on segmentation [19-22] or adopting a segmentation-by-detection strategy [23] [24]when the number of considered OARs is often fewer than or around 20. With a greater number of OARs needed to be segmented, deep network optimization may become increasingly difficult. From an early preliminary version of this work[25], we introduced a novel stratified deep learning framework to segment a comprehensive set of H&N OARs by balancing the OARs' intrinsic spatial and appearance complexity with adaptive neural network architectures. The proposed system, stratified organ at risk segmentation (SOARS), divides OARs into three levels, i.e., anchor, mid-level, and small & hard (S&H) according to their complexity. Anchor OARs are high in intensity contrast and low in inter-user variability and can be segmented first to provide informative location references for the following harder categories. Mid-level OARs are low in contrast but not inordinately small. We use anchor-level predictions as additional input to guide the mid-level OAR segmentation. S&H OARs are very small in size or very poor in contrast. Hence, we use a detection by segmentation strategy to better manage the extremely unbalanced class distributions across the entire volume. Besides this processing stratification, we further deploy another stratification by using neural architecture search (NAS) to automatically determine the optimal network architecture for each OAR category since it is unlikely the same network architecture suits all categories equally. We specifically formulate this structure learning problem as differentiable NAS [26,27], allowing automatic selection across 2D, 3D or Pseudo-3D (P3D) convolutions with kernel sizes of 3 or 5 pixels at each convolutional block.

SOARS achieves the state-of-the-art performance in segmenting 42 OARs in a single institution cross-validation evaluation[25], but essential questions remain unclear regarding to its clinical applicability and generality: (1) does SOARS generalize well into a large-scale multi-institutional evaluation?; (2) how much manual editing effort is required before the predicted OARs can be considered as clinically accepted?; (3) how well does the segmentation accuracy of SOARS compare towards inter-user variation?; and more critically, (4) what are the dosimetric variations brought by OAR differences in the downstream RT planning stage? To adequately address these questions, we first enhance SOARS by replacing the segmentation backbone of P-HNN [28] with UNet [29] and conduct the NAS optimization based on the UNet architecture. Then, we extensively evaluate SOARS on an external set of 1327 unseen H&N cancer patients from six institutions (one internal and five external). Using another 30 randomly selected external patients, we further conducted three subjective user studies: (1) physician's assessment of the revision effort and time spent when editing on predicted OARs; (2) a comparison of contouring accuracy between SOARS and the inter-user variation; and (3) in the intensity modulated RT (IMRT) planning, a dosimetric accuracy comparison using different OAR contours (SOARS, SOARS + physician editing, and physician's manually labeling).

Results

Datasets for training and evaluation

In this multi-institutional retrospective study, we collected, in total, 1533 H&N cancer patients (each with a pCT scan and who received RT as their primary treatment) to develop and evaluate the performance of SOARS. Patients were collected from Chang Gung Memorial Hospital (CGMH), First Affiliated Hospital of Xi'an Jiaotong University (FAH-XJU), and First Affiliated Hospital of Zhejiang University (FAH-ZU), Gansu Provincial Hospital (GPH), Huadong Hospital Affiliated of Fudan University (HHA-FU), Southern Medical University (SMU). Detailed patient characteristics in each institution are shown in Table 1.

Training-validation dataset. First, we created a training-validation dataset to develop SOARS using 176 patients from CGMH between 2015 and 2018 (internal training dataset). Each patient had 42 OARs manually delineated by senior physicians (board-certified radiation oncologists) according to the consensus guideline [7]. Based on the OAR statistical shape, CT appearance and location characteristics (confirmed by the physicians), 42 OARs are divided into the following three categories. **Anchor OARs**: brainstem, cerebellum, eye (left and right), mandible (left and right), spinal cord, and temporomandibular joint (TMJoint, left and right). **Mid-level OARs**: brachial plexus (left and right), basal ganglia (left and right), constrictor muscle (inferior, middle, and superior), epiglottis, esophagus, glottic and supraglottic larynx (GSL), oral cavity, parotid (left and right), submandibular gland (SMG, left and right), temporal lobe (left and right), thyroid (left and right). **S&H OARs**: cochlea (left and right), hypothalamus, inner ear (left and right), lacrimal gland (left and right), lens (left and right), optic nerve (left and right), optic chiasm, pineal gland, and pituitary. These 42 OARs represent one of the most comprehensive H&N OAR sets and can serve as a superset when testing/evaluating patients in other institutions. We divided this dataset into two subgroups: 80% to train the segmentation model and 20% as a validation set for model selection and ablation study. The ablation performance of SOARS is depicted in Table 2.

Independent internal testing dataset. Next, for independent evaluation, we collected 326 patients from CGMH between 2012 and 2020 as another internal testing dataset besides the training-validation. OAR labels in this cohort were extracted from those generated during the clinical RT contouring process that senior physicians confirmed. Depending on the H&N cancer types or tumor locations, a range of 18 to 42 OAR contours were generally available for each patient in this cohort.

Multi-institutional external testing dataset. For quantitative external evaluation, 1001 patients were collected from five different institutions located in various areas of mainland China between 2013 and 2019 (external testing dataset). Each patient is accompanied by the clinical RT treatment OAR contours, ranging from 13 to 25 OARs, depending on their institutional-specific RT protocols. Senior physicians of each institution examined these clinical OAR contours to ensure they met the delineation consensus guidelines [7]. Detailed patient statistics and subject characteristics of these five external institution datasets are given in Table 1.

Multi-user testing dataset. To further evaluate the clinical applicability of SOARS, 30 nasopharyngeal cancer (NPC) patients were randomly selected from one external institution (FAH-ZU) to form a multi-user testing dataset. In this cohort, each patient contained 13 OAR contours, the tumor target volume contours, and the IMRT plan originally generated by the clinical team at FAH-ZU. First, two senior physicians (both with >10 years' experience in treating H&N cancers) edited the SOARS predicted 42 OARs (resulting in SOARS-revised contours) and recorded the editing time to assess the revision efforts required for making SOARS predicted OAR contours to be clinically accepted. One senior physician manually edited the 13 OARs used in FAH-ZU's RT protocol, while the other senior physician edited the other 29 OARs not included in FAH-ZU's RT protocol. Second, another physician with 4 years' experiences manually contoured the 13 OARs in FAH-ZU's protocol (denoted as human reader contours). Then, using the clinical treatment contours of the 13 OARs as gold-standard references, we compared the contouring accuracy of SOARS, SOARS-revised, and the human reader. Third, while keeping the original dose grid in RT plan, we replaced the clinical reference OAR contours by the SOARS, SOARS-revised, human reader contours respectively, to analyze the impact on OAR dose metrics. This helps determine if differences in OAR contouring would produce clinically relevant differences of radiation doses received by the OARs in the downstream dose planning stage. The overview of the multi-user evaluation is illustrated in Fig. 1.

Performance on the CGMH internal testing dataset

The quantitative performance of SOARS in the internal testing dataset is summarized in Table 3. SOARS achieved a mean Dice score coefficient (DSC), Hausdorff distance (HD) and average surface distance (ASD) of 74.8%, 7.9mm and 1.1mm, respectively, among 42 OARs. For stratified OAR categories, mean DSC, HD and ASD for anchor OARs were 86.9%, 5.0mm and 0.7mm, respectively; for mid-level OARs were 74.6%, 12.4mm and 1.9mm, respectively; and for S&H were 67.2%, 3.7mm and 0.7mm, respectively. In comparison, the previous state-of-the-art

H&N OAR segmentation approach UaNet [24] had a significantly worse performance (DSC: 69.8% vs 75.3%, HD: 8.8 vs 7.9mm, ASD: 1.6 vs 1.1mm; all $p<0.001$). UaNet adopted a modified version of 3D Mask R-CNN [30], which decoupled the whole task into detection followed by segmentation. Although UaNet achieved one of the previous best performances, it lacked dedicated stratified learning to adequately handle a larger number of OARs, possibly accounting for the markedly inferior segmentation accuracy compared to SOARS. Among three stratified OAR categories, S&H OARs exhibited the largest gap between SOARS and UaNet (DSC: 67.2% vs 59.4%, HD: 3.7 vs 4.7mm, ASD: 0.7 vs 1.2mm; all $p<0.001$). This result further confirmed the advantage of SOARS, which employed an adaptively tailored processing workflow and an optimized network architecture towards a particular category of OARs. Fig. 3 shows several qualitative comparisons on the internal testing dataset.

Performance on the multi-institutional external testing dataset

The overall quantitative external evaluation and the individual external institution evaluation results are shown in Table 4. SOARS achieved a mean DSC, HD95, and ASD of 78.0%, 6.7mm and 1.0mm, respectively, among 25 H&N OARs overall. These represented significant performance improvement ($p<0.001$) as compared against the UaNet (4% absolute DSC increase, 16% HD reduction, and 40% ASD reduction). For individual institutions, average DSC scores of SOARS ranged from 76.9% in FAH-XJU to 80.7% in GPH, while most institutions yielded approximately 78% DSC. HD values of SOARS were from 5.9mm in FAH-ZU to 8.1mm in SMU; and ASD obtained from 0.9mm in FAH-ZU and GPH to 1.3mm in SMU and FAH-XJU. Although the OAR numbers varied for external institutions (due to differences among institutional specific RT treatment protocols), these quantitative performance metrics are generally comparable against the internal testing performance levels, demonstrating that SOARS' generality and accuracy hold well to this large-scale external dataset. SOARS consistently and statistically significantly outperforms ($p<0.001$) UaNet in external evaluation

(UaNet had a mean DSC, HD and ASD of 74.3%, 8.0mm and 1.4mm, respectively). SOARS outperforms UaNet in 21 out of 25 OARs on all metrics, with an average DSC improvement of ~4% and relative distance error reductions of 17.5% for HD and 28.5% for ASD.

Assessment of editing effort in multi-user testing dataset

In 30 multi-user evaluation patients, assessment from two senior physicians showed that the vast majority (1237 of 1260 = 42 OAR types × 30, or 98%) of OAR instances produced by SOARS were clinically acceptable or required only very minor revision (no revision: 729 (58%); revision < 1 minute: 508 (40%)). Only 23 (2%) OAR instances had automated delineation or contouring errors that required 1-3 minutes of moderate modification efforts. None OAR instances required > 3 minutes of major revision. Fig. 3 details the assessment results. OAR types that needed the most frequent major revisions are hypothalamus, optic chiasm, esophagus, oral cavity, SMG, and temporal lobes. The average total editing time of all 42 OARs for each patient is 10.5 minutes. Using a random selection of 5 out of 30 patients, two senior physicians also annotated 42 OARs from scratch, which took averaged 106.4 minutes per patient. Thus, the contouring time was significantly reduced by 90% when editing based on SOARS predictions. This observation strongly confirms the added value of SOARS in clinical practice.

Inter-user contouring accuracy in multi-user testing dataset

The contouring accuracy of SOARS, SOARS-revised and human reader in the multi-user testing dataset is shown in Table 5. It is observed that SOARS consistently yielded higher or comparable performance in all 13 OARs (used in FAH-ZU's RT protocol) as compared to the performance of the human reader (a physician with 4 years' experience). Overall, SOARS achieved significantly improved quantitative results ($p<0.001$) in mean DSC (0.82 vs 0.79), HD

(4.3 vs 6.1mm) and ASD (0.6 vs 1.0mm). 11 out of 13 OAR types demonstrated remarked improvements when comparing SOARS with the human reader. On the other hand, by comparing the contouring accuracy between SOARS and SOARS-revised, they have showed very similar quantitative performance (mean DSC: 0.82 vs 0.83, HD: 4.3 vs 3.9mm, and ASD: 0.6 vs 0.5mm). Note that SOARS derived contours (both SOARS and SOARS revised) have significantly better performance as compared to those of the human reader, representing the inter-user segmentation variation. Results from the inter-user variation and the previous revision effort assessment validated that SOARS can be readily serving as an alternative “expert” to output high-quality automatically delineated OAR contours, where very minor or no manual efforts are usually required on further editing the SOARS’ predictions.

Dosimetric accuracy in multi-user testing dataset

Although OAR contouring accuracy reflects the OAR delineation quality, we can further examine its impact on the important downstream dose planning step. The quantitative dosimetric accuracy of various OAR sets, i.e., SOARS, SOARS-revised, and human reader, is illustrated in Table 5 and Fig. 5 (c), and the relationship between contouring accuracy and dosimetric accuracy is plotted in Supplementary Fig. 2 and Fig. 3. It was observed that, for SOARS, the dosimetric differences in mean dose and in maximum dose were 4.8% and 3.5%, respectively, averaged across all 13 OARs using 30 patients. These were statistically significantly smaller (p<0.001) than those of the human reader contours (6.1% and 5.0%), and comparable to those of SOARS-revised (4.7% and 3.4%). More specifically, using SOARS predictions, only 25 out of 390 (6%) OAR instances among 30 patients had a mean dose variation larger than 10%, and no OAR has a mean dose difference larger than 30%. In comparison, using the human reader contours, 49 out of 390 (12%) OAR instances among 30 patients had a mean dose variation larger than 10%, and 12 OAR instances with a mean dose difference larger than 30%. SOARS-revised contours generally had comparable performance with SOARS. Similar trends were observed for the differences in maximum dose. These results demonstrated that the high contouring accuracy of SOARS evidently leaded to high dosimetric accuracy in the dose planning stage. Fig. 5 (a, b) shows qualitative dosimetric examples and dose-volume histograms (DVH) for using three substitute OAR sets (SOARS, SOARS-revised, human reader). We observed that doses received by most OARs from SOARS and SOARS-revised matched more closely to the clinical reference doses than those from the human reader.

Discussion

In this multi-institutional study, we presented a novel Stratified OAR Segmentation deep learning model, SOARS, that can be used to automatically delineate 42 H&N OARs as the current most comprehensive clinical protocol. By stratifying the organs into three different OAR categories, the processing workflows and segmentation architectures (computed by NAS) were optimally tailored. As such, SOARS is a well-calibrated synthesis of organ stratification, multi-stage segmentation, and NAS. SOARS was trained using 176 patients from CGMH and extensively evaluated on 1327 unseen patients from six institutions (326 from CGMH and 1001 from five other external medical centers). It achieved a mean DSC and ASD of 74.9% and 1.3mm, respectively, in 42 OARs from the CGMH internal testing and generalized well to the external testing with a mean DSC of 78.0% and ASD of 1.0mm, respectively, in 25 OARs. SOARS consistently outperformed the previous state-of-the-art method UaNet [24] by 3-5% absolute DSC and 16-32% of relative ASD in all six institutions. In a multi-user study, 98% of SOARS-predicted OARs required no revision or very minor revision from physicians before they were clinically accepted , and the manual contouring time can be reduced by 90% (from 106.4 to 10.5 minutes). In addition, the segmentation and dosimetric accuracy of SOARS were comparable to or smaller than the inter-user variation.

Recent consensus guidelines recommended delineating more than 40 OARs in H&N cancer patients [7]. However, in practice, it is an unmet need. Most institutions only delineated a small subset of H&N OARs per their institutional specific RT protocol, or they can only afford to delineate OARs that are closest to tumor targets. The challenges of following the consensus guideline were probably due to the lack of efficient and accurate OAR delineation tools (most automated tools focused on segmenting less than or around 20 H&N OARs [18,21,23,31]). Manually contouring 40+ OARs was too time-consuming and expertise-demanding, unrealistic in practice. without assessment of the dosimetric results in the complete set of OARs, it was infeasible to track and analyze the organ-specific adverse effects after RT treatment in multi-institutional clinical trials. In addition, data pooling analysis of radiation therapy from different institutions was impeded by the inconsistency in OAR contouring guidance. The Global Quality Assurance of Radiation Therapy Clinical Trials Harmonization Group (CHG) has provided standardized nomenclature for clinical trial use to address this problem [32]. With the proposed SOARS, it is feasible to provide comprehensive OAR dose evaluation, further facilitating studies on post-treatment complications and quality assurance.

In this work, from the OAR contouring quality, we further analyzed the OAR dosimetric accuracy in the subsequent dose planning step. The dosimetric differences in mean dose and in maximum dose were used as dose metrics consistently with previous work [33]. Overall, the vast majority of SOARS-predicted OAR instances had the mean and maximum dose variance no larger than 10%, which was comparable to or smaller than the inter-user dose variations in our experiment. This variation was also smaller than the previously reported inter-user dose variations in six H&N OAR types [33], where quite few are larger than 30% or even above 50%. For individual OARs, we observed that the optic chiasm and optic nerve (left and right) exhibited increased dose variation (10-30%) in a small portion of patients (Supplementary Fig. 2 and Fig. 3). This phenomenon was consistently observed in SOARS, SOARS-revised, and the human reader contours. This indicated that dosimetries in areas consisting of these OARs are sensitive to the contouring differences, suggesting that more attention should be required to delineate the above OAR types for NPC patients.

Our study had several limitations. First, the external testing datasets do not have a complete set or the same amount of 42 OAR types. This reflects real-world situations among different institutions. Manually labeling 42 OARs for all 1001 external testing patients is impractical (estimated to require ≥3 hours per patient). Hence, we chose to use the existing clinically labeled OAR types to supplement for testing. Second, the multi-user testing dataset of FAH-ZU contains only 13 clinical reference OAR types according to its RT protocol. Thus we evaluated the inter-user variation of segmentation and dosimetric accuracy using these 13 OARs instead of the complete 42 OAR types. Nevertheless, these 13 OAR types included those from the three different OAR categories of anchor, mid-level, and S&H. We believe the performance from these would reflect the real inter-user variation with a larger number of OAR types. Third, to evaluate the dose metrics, we kept the original planning dose grid the same while replacing the original clinical OAR contours with substitute contours by SOARS, SOARS-revised, and the human reader. It would be interesting if we can further use the substitute OAR contours and original tumor target volumes to generate new planning dose grids to evaluate the OAR dose metrics, which might affect the tumor target dose distributions as well. It would also be helpful to conduct a randomized clinical trial comparing the side effects and life quality as outcomes of manual and SOARS assisted OAR contouring. This could further validate the clinical value of SOARS. We leave these for our future works.

To conclude, we introduced and developed a stratified deep learning method to segment the most comprehensive 42 H&N OAR types in radiotherapy planning. Through extensive multi-institutional validation, we demonstrated that our SOARS model achieved accurate and robust performance, and produced comparable or higher accuracy in OAR segmentation and the subsequent dose planning than the inter-user variation. Physicians needed very minor or no revision for 98% of the OAR instances (when editing on SOARS predicted contours) to warrant clinical acceptance. SOARS could be implemented and adopted in the clinical radiotherapy workflow for a more standardized, quantitatively accurate, and efficient OAR contouring process with high reproducibility.

Methods

The SOARS framework is illustrated in Fig. 2. It consists of three processing branches to stratify the anchor, mid-level, and S&H OAR segmentation, respectively. Stratification manifested first in the distinct processing workflow used for each OAR category. We next stratified neural network architectures by using differentiable neural architecture search (NAS) [26,27] to search a distinct network structure for each OAR category. We will explain each stratification process below.

Processing Stratification in SOARS

SOARS first segmented the anchor OARs. Then, with the help of predicted anchor OARs, mid-level and S&H OARs were segmented. For the most difficult category of S&H OARs, SOARS first detected their center locations and then zoomed in accordingly to segment the small OARs. For the backbone of all three branches, we adopted the UNet structure implemented in the nnUNet framework {nnUNet ref} [29], which has demonstrated leading performance in many medical image segmentation tasks. We tailored each UNet with NAS, which is explained in the subsequent subsection.

We denoted the training data of N instances as $S = \{X_i, Y_i^A, Y_i^M, Y_i^S\}_{i=1}^N$, where X_i, Y_i^A, Y_i^M, and Y_i^S were the input pCTs and ground-truth masks for anchor, mid-level, and S&H OARs, respectively. The indexing parameter i was dropped for clarity. We used boldface to denote vector-valued volumes and used vector concatenation as an operation across all voxel locations.

Anchor branch: Assuming there are C anchor classes, we first used the anchor branch to generate OAR prediction maps for every voxel location, j, and every output class, c:

$$\hat{Y}_c^A(j) = p^A(Y^A(j) = c \mid X; \mathbf{W}^A),\ \hat{\mathbf{Y}}^A = \left[\hat{Y}_1^A \ \cdots \ \hat{Y}_C^A\right] \tag{1}$$

where UNet functions, parameters, and the output prediction maps were denoted as $p^A(\cdot)$, $\mathbf{W}^{(\cdot)}$ and $\hat{\mathbf{Y}}^A$, respectively. Anchor OARs are easy and robust to segment based on their own CT image appearance and spatial context features. Consequently, they provided highly informative location and semantic cues to support the segmentation of other OARs.

Mid-level branch: Most mid-level OARs are primarily soft tissue, which have limited contrast and can be easily confused with other structures with similar intensities and shapes. Hence, we incorporated the anchor predictions into mid-level learning. Specifically, the anchor predictions and the pCT were concatenated to create a multi-channel input $\left[X, \hat{\mathbf{Y}}^A\right]$:

$$\hat{Y}_c^M(j) = p^M\left(Y^M(j) = c \mid X, \hat{\mathbf{Y}}^A ; \mathbf{W}^M\right) \tag{2}$$

Small & hard branch: Considering the low contrast and extremely unbalanced class distributions for S&H OARs across the entire CT volume, direct S&H OAR segmentation is challenging. Here, we further decoupled this branch into a detection followed by segmentation process. Because the H&N region has relatively stable anatomical spatial distribution, detecting rough locations of S&H OARs is a much easier and reliable task. Once the OAR center was approximately determined, a localized region can be cropped out to focus on segmenting the fine boundaries in a zoom-in fashion. The detection was implemented using a simple yet effective heat map regression approach and the heat map labels were generated at each organ center using a 3D Gaussian kernel [34,35]. Let $f(\cdot)$ denote the UNet function for the detection module, we also combined the anchor branch predictions with pCT as the detection input:

$$\hat{\mathbf{H}} = f\left(X, \hat{\mathbf{Y}}^A; \mathbf{W}^{\mathbf{D}}\right), \tag{3}$$

where $\widehat{\mathbf{H}}$ were the predicted heat maps of S&H OARs. Given the regressed heat map $\widehat{\mathbf{H}}$, the pixel location corresponding to the highest value was extracted to crop a volume of interest (VOI) using three times the extent of the maximum size of the OAR of interest. Then, SOARS segmented the fine boundaries of S&H OARs within the VOI. Let V denote the cropped VOI in pCT. The S&H OAR segmentation was implemented as:

$$\hat{Y}_c^S(j) = p^S(Y^S(j) = c \mid V; \mathbf{W}^S)\,. \tag{4}$$

Automatic Neural Architecture Search in SOARS

Considering the significant statistical variations in OAR appearance, shape, and size, it is unlikely that the same network architecture would suit each OAR category equally. Hence, SOARS automatically searches the more suitable network architectures for each branch, adding an additional dimension to the stratification. We conducted the differentiable NAS [26,27] on top of the network structure of UNet [29]. The NAS search space included 2D, 3D, and pseudo-3D convolutions with either kernel sizes of 3 or 5. Fig. 2 (b-c) demonstrates the network architecture and the search space of NAS. Let $\phi(\cdot\,; \omega_{x\times y\times z})$ denote a composite function of the following consecutive operations: batch normalization, a rectified linear unit, and a convolution with an $x \times y \times z$ dimension kernel. If one of the kernel dimensions is set to 1, it reduces to a 2D kernel. The search space $\mathbf{\Phi}$ can be represented as.

$$\begin{aligned}
\phi_{\mathrm{2D}_3} &= \phi(\cdot\,;\, \omega_{3\times3\times1}),\\
\phi_{\mathrm{2D}_5} &= \phi(\cdot\,;\, \omega_{5\times5\times1}),\\
\phi_{\mathrm{3D}_3} &= \phi(\cdot\,;\, \omega_{3\times3\times3}),\\
\phi_{\mathrm{3D}_5} &= \phi(\cdot\,;\, \omega_{5\times5\times5}),\\
\phi_{\mathrm{P3D}_3} &= \phi(\phi(\cdot\,;\, \omega_{3\times3\times1})\,;\, \omega_{1\times1\times3}),\\
\phi_{\mathrm{P3D}_5} &= \phi(\phi(\cdot\,;\, \omega_{5\times5\times1})\,;\, \omega_{1\times1\times5}),
\end{aligned}$$

$$\mathbf{\Phi} = \{\phi_{2\mathrm{D}_3}, \phi_{2\mathrm{D}_5}, \phi_{3\mathrm{D}_3}, \phi_{3\mathrm{D}_5}, \phi_{\mathrm{P3D}_3}, \phi_{\mathrm{P3D}_5}\}. \quad (5)$$

The architecture was learned in a differentiable fashion. We made the search space continuous by relaxing the selection of $\phi(\cdot\,; \omega_{x\times y\times z})$ to a softmax function over ϕ. For k operations, we define a set of α_k learnable logits for each. The weight γ_k for an operation is defined as $\gamma_k = \frac{\exp(\alpha_k)}{\sum_m \exp(\alpha_m)}$, and the combined output is $\phi' = \sum_k \gamma_k \phi_k$. As the result of NAS, we selected the operation with the top weight to be the searched operation. We used the same scheme to search the segmentation network architecture for all three branches (excluding the S&H detection module) and trained SOARS using the final auto-searched architecture. The searched network architectures for each branch are listed in Fig. 3. The implementation details are reported in the supplementary materials.

Quantitative evaluation of contouring accuracy

For the internal and external testing datasets, the contouring accuracy was quantitatively evaluated using three common segmentation metrics {add our ESOGTV arXiv ref} [37], i.e., Dice similarity coefficient (DSC), Hausdorff distance (HD) and average surface distance (ASD). Additionally, for quantitative comparison, we also trained and tested the previous state-of-the-art H&N OAR segmentation method, UaNet [24]. For the model development of UaNet, we used the default parameter setting from original authors[24] as these have been already specifically tuned for the head and neck OARs. We applied the same training-validation split as ours to ensure a fair comparison.

Human experts' assessment of revision efforts

An assessment experiment by human experts was conducted to evaluate the editing efforts needed for the predicted OARs to be clinically accepted. Specifically, using the multi-user testing dataset, two senior physicians (both >10 years of experience in treating H&N cancers) were asked to edit SOARS predictions of 42 OARs according to the consensus guideline [7]. Besides the pCT scans, other clinical information, and imaging modality such as MRI (if available) were also provided to physicians as reference. The edited OAR contours were denoted as SOARS-revised. Four manual revision categories were designated as no revision required, revision required in <1 minute (minor revision), revision required in 1–3 minutes (moderate revision), and revision required in >3 minutes (major revision

Inter-user contouring evaluation

Using the multi-user testing dataset, we further asked a board-certified radiation oncologist with 4 years' experience specialized in treating H&N cancers to delineate the 13 OAR types in FAH-ZU's RT protocol manually. Patients' pCT scans along with their clinical information and other available medical images (including MRI) were provided to the physician. The labeled OAR contours were denoted as human reader contours. Then, we compared the contouring accuracy between SOARS, SOARS-revised, and the human reader using the evaluation metrics of DSC, HD and ASD. The contouring performance of SOARS-revised and the human reader represents the inter-user variation in OAR contouring.

Inter-user dosimetric evaluation

Differences in the OAR contouring accuracy would not, by itself, indicate whether such differences are clinically relevant in terms of radiation doses received by the OARs. Therefore, we further quantified the dosimetric impact brought by the OAR contouring differences. Specifically, for each patient in the multi-user testing dataset, we first used the original clinical reference OARs and the corresponding dose grid (dose voxel sizes ranging from 2 to 4 mm) to compute the OAR dose metrics in terms of mean doses and max doses. Then, the same dose grid was combined with different OAR contour sets, i.e., SOARS, SOARS-revised, human reader, and the dose metrics of each OAR contour set were calculated. This design was to isolate the dose effects due strictly to contouring differences because the dose grid was fixed, and the dose metrics were quantified by replacing each clinical reference contours with the substitute contours. Following the work[33], we calculated the difference in mean dose and difference in maximum dose as follows:

$$\mathrm{Diff}_{\text{mean dose}} = \frac{\text{mean dose}(OAR_{substitute},\ Dose_{plan}) - \text{mean dose}(OAR_{ref},\ Dose_{plan})}{\text{mean dose}(OAR_{ref},\ Dose_{plan})} \times 100\% \tag{6}$$

$$\mathrm{Diff}_{\text{max dose}} = \frac{\text{max dose}(OAR_{substitute},\ Dose_{plan}) - \text{max dose}(OAR_{ref},\ Dose_{plan})}{\text{max dose}(OAR_{ref},\ Dose_{plan})} \times 100\% \tag{7}$$

where $\mathrm{OAR}_{\mathrm{substitute}}$ represents the OAR contours by SOARS, SOARS-revised, and the human reader, respectively, while $\mathrm{OAR}_{\mathrm{ref}}$ and $\mathrm{Dose}_{\mathrm{plan}}$ represent the original clinical OAR contours and the dose plans in the actual RT treatment, respectively. The dose-volume histogram (DVH) was also plotted for qualitative illustration. The dose/DVH statistics were generated using Eclipse 11.0 (Varian Medical Systems Inc., Palo Alto, CA).

Statistical Analysis

The Wilcoxon matched-pairs signed rank test was used to compare the evaluation metrics in paired data, while Manning-Whitney U test was used to compare the unpaired data. All analyses were performed by using **R** [38]. Statistical significance was set at two-tailed $p<0.05$.

Data availability

The data support the findings of this study are available from the corresponding author upon reasonable request. The image data utilized in this study on the head & neck organs at risk (OARs) segmentation is not publicly available due to the data privacy consideration and restricted permission of the current study.

Code availability

The baseline UNet used in this study is implemented in the nnUNet deep learning framework available at https://github.com/MIC-DKFZ/nnUNet. The codes used for inference and result performance evaluation cab be publicly available on GitHub after publication.

Acknowledgements

This work is partially supported by Maintenance Project of the Center for Artificial Intelligence in Medicine (Grant CLRPG3H0012, CMRPG3K1091, SMRPG3I0011) at Chang Gung Memorial Hospital.

Author contributions

For three first co-authors, D.G. was responsible for the data cleaning, deep learning model development and the internal and external evaluation, J.G. helped collect the external data, participate in the human assessment and dosimetric analysis in the multi-user studies, and X.Y. helped collect the external data, participate, and coordinate the human assessment and contouring analysis in the multi-user studies, and they all involved in the experimental design and drafted the manuscript. S.Y., Y.S., Z.L. and W.L. participated in the multi-user studies. Y.X. and Z.Z. aided in the deep learning model development and results interpretation. B.H and T.H approved the contours for validation of the internal institution. L.P., Y.R., R.L., G.Z., M.M. and X.C. helped collect, organize, and validated the external data. Y.C. collected and organized in internal data. L.H., and J.X. contributed to the design and implementation of the research. A.P.H aided in the interpreting the results and edited the manuscript. L. L. aided in the experimental design and interpreting the results and edited the manuscript. C.L. was responsible for reviewing and modifying the contours from internal institutions, and she also provided guidance and consulting in the multi-user study. D.J. was responsible for the data cleaning, development of the deep learning model, and overseeing the evaluation process. T.H. collected the internal training and evaluation data. D.J. and T.H. were responsible for conception and design of the experiments and oversaw overall direction and planning and drafted the manuscript.

Competing Interests

The authors declare no competing interests.

References

1 Sung, H. *et al.* Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians* **71**, 209-249 (2021).

2 Chua, M. L., Wee, J. T., Hui, E. P. & Chan, A. T. Nasopharyngeal carcinoma. *The Lancet* **387**, 1012-1024 (2016).

3 Yeh, S.-A., Tang, Y., Lui, C.-C., Huang, Y.-J. & Huang, E.-Y. Treatment outcomes and late complications of 849 patients with nasopharyngeal carcinoma treated with radiotherapy alone. *International Journal of Radiation Oncology* Biology* Physics* **62**, 672-679 (2005).

4 Hsiao, K.-Y. *et al.* Cognitive function before and after intensity-modulated radiation therapy in patients with nasopharyngeal carcinoma: a prospective study. *International Journal of Radiation Oncology* Biology* Physics* **77**, 722-726 (2010).

5 Lee, D. S. *et al.* Long-term outcome and toxicity of hypofractionated stereotactic body radiotherapy as a boost treatment for head and neck cancer: the importance of boost volume assessment. *Radiation Oncology* **7**, 1-9 (2012).

6 Machtay, M. *et al.* Factors associated with severe late toxicity after concurrent chemoradiation for locally advanced head and neck cancer: an RTOG analysis. *Journal of Clinical Oncology* **26**, 3582 (2008).

7 Brouwer, C. L. *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* **117**, 83-90, doi:10.1016/j.radonc.2015.07.041 (2015).

8 La Macchia, M. *et al.* Systematic evaluation of three different commercial software solutions for automatic segmentation for adaptive therapy in head-and-neck, prostate and pleural cancer. *Radiation Oncology* **7**, 1-16 (2012).

9 Round, C. *et al.* Radiotherapy demand and activity in England 2006–2020. *Clinical Oncology* **25**, 522-530 (2013).

10 Nguyen-Tan, P. F. *et al.* Randomized phase III trial to test accelerated versus standard fractionation in combination with concurrent cisplatin for head and neck carcinomas in the Radiation Therapy Oncology Group 0129 trial: long-term report of efficacy and toxicity. *Journal of Clinical Oncology* **32**, 3858 (2014).

11 Han, X. *et al.* Atlas-based auto-segmentation of head and neck CT images. *International Conference on Medical Image Computing and Computer-assisted Intervention.* 434-441 (Springer).

12 Isambert, A. *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 and oncology* **87**, 93-99 (2008).

13 Schreibmann, E., Marcus, D. M. & Fox, T. Multiatlas segmentation of thoracic and abdominal anatomy with level set-based local search. *Journal of applied clinical medical physics* **15**, 22-38 (2014).

14 Saito, A., Nawano, S. & Shimizu, A. Joint optimization of segmentation and shape prior from level-set-based statistical shape model, and its application to the automated segmentation of abdominal organs. *Medical image analysis* **28**, 46-65 (2016).

15 Raudaschl, P. F. *et al.* Evaluation of segmentation methods on head and neck CT: auto-segmentation challenge 2015. *Medical physics* **44**, 2020-2036 (2017).

16 Teguh, D. N. *et al.* Clinical validation of atlas-based auto-segmentation of multiple target volumes and normal tissue (swallowing/mastication) structures in the head and neck. *International Journal of Radiation Oncology* Biology* Physics* **81**, 950-957 (2011).

17 Chen, W. *et al.* Deep learning vs. atlas-based models for fast auto-segmentation of the masticatory muscles on head and neck CT images. *Radiat Oncol* **15**, 176, doi:10.1186/s13014-020-01617-0 (2020).

18 Ibragimov, B. & Xing, L. Segmentation of organs-at-risks in head and neck CT images using convolutional neural networks. *Medical physics* **44**, 547-557 (2017).

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

20 Nikolov, S. *et al.* Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy. *arXiv preprint arXiv:1809.04430* (2018).

21 Zhu, W. *et al.* AnatomyNet: deep learning for fast and fully automated whole-volume segmentation of head and neck anatomy. *Medical physics* **46**, 576-589 (2019).

22 Men, K. *et al.* More accurate and efficient segmentation of organs-at-risk in radiotherapy with convolutional neural networks cascades. *Medical physics* **46**, 286-292 (2019).

23 Gao, Y. *et al.* Focusnet: Imbalanced large and small organ segmentation with an end-to-end deep neural network for head and neck ct images. *International Conference on Medical Image Computing and Computer-Assisted Intervention.* 829-838 (Springer).

24 Tang, H. *et al.* Clinically applicable deep learning framework for organs at risk delineation in CT images. *Nature Machine Intelligence* **1**, 480-491 (2019).

25 Guo, D. *et al.* Organ at risk segmentation for head and neck cancer using stratified learning and neural architecture search. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 4223-4232.

26 Liu, H., Simonyan, K. & Yang, Y. Darts: Differentiable architecture search. *arXiv preprint arXiv:1806.09055* (2018).

27 Liu, C. *et al.* Auto-deeplab: Hierarchical neural architecture search for semantic image segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 82-92.

28 Harrison, A. P. *et al.* Progressive and multi-path holistically nested neural networks for pathological lung segmentation from CT images. *International conference on medical image computing and computer-assisted intervention.* 621-629 (Springer).

29 Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. *International Conference on Medical image computing and computer-assisted intervention.* 234-241 (Springer).

30 He, K., Gkioxari, G., Dollár, P. & Girshick, R. Mask r-cnn. *Proceedings of the IEEE international conference on computer vision.* 2961-2969.

31 Nikolov, S. *et al.* Clinically Applicable Segmentation of Head and Neck Anatomy for Radiotherapy: Deep Learning Algorithm Development and Validation Study. *J Med Internet Res* **23**, e26151, doi:10.2196/26151 (2021).

32 Mir, R. *et al.* Organ at risk delineation for radiation therapy clinical trials: Global Harmonization Group consensus guidelines. *Radiother Oncol* **150**, 30-39, doi:10.1016/j.radonc.2020.05.038 (2020).

33 Nelms, B. E., Tomé, W. A., Robinson, G. & Wheeler, J. Variations in the contouring of organs at risk: test case from a patient with oropharyngeal cancer. *International Journal of Radiation Oncology* Biology* Physics* **82**, 368-378 (2012).

34 Wei, S.-E., Ramakrishna, V., Kanade, T. & Sheikh, Y. Convolutional pose machines. *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition.* 4724-4732.

35 Xu, Z. *et al.* Less is more: Simultaneous view classification and landmark detection for abdominal ultrasound images. *International Conference on Medical Image Computing and Computer-Assisted Intervention.* 711-719 (Springer).

36 Liu, L. *et al.* On the variance of the adaptive learning rate and beyond. *arXiv preprint arXiv:1908.03265* (2019).

37 Lin, L. *et al.* Deep learning for automated contouring of primary tumor volumes by MRI for nasopharyngeal carcinoma. *Radiology* **291**, 677-686 (2019).

38 Team, R. C. R: A language and environment for statistical computing. (2013).

Table 1. Subject characteristics. CGMH: Chang Gung Memorial Hospital; FAH-XJU: First Affiliated Hospital of Xi'an Jiaotong University; FAH-ZU: First Affiliated Hospital of Zhejiang University; GPH: Gansu Provincial Hospital; HHA-FU: Huadong Hospital Affiliated of Fudan University; SMU: Southern Medical University.

Characteristics	Train/validation CGMH (n = 176)	Internal testing CGMH (n = 326)	External testing FAH-XJU (n = 82)	External testing FAH-ZJU (n = 447)	External testing GPH (n = 50)	External testing HHA-FU (n = 195)	External testing SMU (n = 227)
Sex							
Male	160 (91%)	284 (87%)	65 (79%)	321 (72%)	33 (66%)	145 (75%)	161 (71%)
Female	16 (9%)	42 (13%)	17 (21%)	126 (28%)	17 (34%)	50 (25%)	66 (29%)
Diagnostic age	54 [48-61]	54 [49-62]	57 [49-66]	57 [50-65]	58 [49-70]	56 [47-65]	50 [42-57]
Tumor site							
Nasopharynx	7 (4%)	90 (28%)	16 (19%)	349 (78%)	2 (4%)	94 (48%)	199 (88%)
Oropharynx	140 (80%)	86 (26%)	20 (24%)	26 (6%)	…	2 (1%)	9 (4%)
Hypopharynx	16 (9%)	115 (35%)	…	16 (4%)	…	8 (4%)	3 (1%)
Larynx	2 (1%)	12 (4%)	38 (47%)	11 (2%)	9 (18%)	25 (13%)	4 (2%)
Oral Cavity	9 (5%)	15 (5%)	9 (10%)	39 (9%)	3 (6%)	2 (1%)	5 (2%)
Salivary gland	…	…	…	…	4 (8%)	4 (2%)	3 (1%)
Others	2 (1%)	8 (2%)	…	6 (1%)	32 (64%)	60 (31%)	4 (2%)
Clinical T-stage							
cT1	23 (13%)	50 (15%)	10 (12%)	55 (12%)	12 (24%)	14 (7%)	20 (9%)
cT2	64 (36%)	82 (25%)	33 (41%)	181 (41%)	18 (36%)	64 (33%)	54 (24%)
cT3	42 (24%)	81 (25%)	25 (30%)	122 (27%)	12 (24%)	35 (18%)	101 (44%)
cT4	47 (27%)	113 (35%)	14 (17%)	89 (20%)	8 (16%)	82 (42%)	52 (23%)
OAR types annotated	42	42	13	13	17	13	25

Note: Others of tumor sites include tumors located at brain, nasal cavity, or lymph node metastasis.

Table 2. Quantitative results of the ablation studies of proposed SOARS using the validation set of the training-validation dataset. The best performance is highlighted in bold font.

	Anchor OARs			Mid-level OARs			S&H OARs			All OARs		
	DSC	HD	ASD	DSC	HD	ASD	DSC	HD	ASD	DSC	HD	ASD
Baseline UNet (nnUNet)	84.3%	12.4	1.0	71.4%	18.0	2.0	58.3%	4.7	1.1	70.4%	12.7	1.4
nnUNet + PS	86.7%	6.4	0.9	72.6%	11.4	1.9	73.7%	4.6	0.7	76.1%	8.2	1.3
nnUNet+PS+NAS	**87.4%**	**5.4**	**0.8**	**74.2%**	**10.4**	**1.7**	**76.2%**	**3.5**	**0.6**	**77.8%**	**7.2**	**1.2**

Note: PS, NAS represent processing stratification and neural architecture search, respectively. The unit for Hausdorff distance (HD) and average surface distance (ASD) is in mm.

Table 3. Quantitative comparisons on the internal testing dataset of 326 patients. Bold and highlighted values represent the best performance and significant improvement as compared between UaNet and SOARS, respectively.

	UaNet			SOARS		
Anchor OARs	DSC	HD (mm)	ASD (mm)	DSC	HD (mm)	ASD (mm)
BrainStem	81.6% ± 5.3%	**8.8 ± 3.3**	1.7 ± 0.7	**83.2% ± 5.8%**	9.0 ± 4.3	**1.6 ± 0.8**
Cerebellum	90.1% ± 9.4%	9.5 ± 6.6	1.2 ± 0.5	**92.9% ± 2.2%**	**7.9 ± 4.8**	**0.9 ± 0.3**
Eye_Lt	85.1% ± 13.2%	3.7 ± 1.4	0.8 ± 0.5	**88.5% ± 4.9%**	**2.9 ± 1.0**	**0.3 ± 0.4**
Eye_Rt	86.2% ± 9.8%	3.6 ± 1.2	0.8 ± 0.4	**88.7% ± 4.8%**	**2.8 ± 1.1**	**0.3 ± 0.4**
Mandible_Lt	85.0% ± 14.6%	14.7 ± 13.3	1.7 ± 4.4	**89.1% ± 2.9%**	**5.4 ± 4.4**	**0.5 ± 0.4**
Mandible_Rt	86.0% ± 12.3%	13.5 ± 11.6	1.5 ± 3.9	**89.0% ± 3.4%**	**5.5 ± 4.6**	**0.5 ± 0.4**
SpinalCord	81.5% ± 9.8%	17.1 ± 37.5	4.2 ± 14.5	**86.3% ± 4.0%**	**4.7 ± 1.4**	**0.7 ± 0.2**
TMJ_Lt	73.0% ± 7.0%	4.8 ± 1.6	1.2 ± 0.4	**81.0% ± 9.1%**	**3.5 ± 1.4**	**0.7 ± 0.5**
TMJ_Rt	75.5% ± 7.1%	4.4 ± 1.6	1.1 ± 0.4	**83.6% ± 6.6%**	**3.4 ± 1.1**	**0.6 ± 0.3**
Mid-level OARs						
BasalGanglia_Lt	**76.0% ± 7.7%**	**9.5 ± 3.0**	**1.8 ± 0.7**	70.9% ± 9.5%	11.4 ± 3.6	2.3 ± 1.0
BasalGanglia_Rt	**73.8% ± 9.1%**	**10.2 ± 2.8**	**2.0 ± 0.8**	71.4% ± 10.1%	10.7 ± 3.3	2.2 ± 0.9
Brachial_Lt	57.5% ± 8.1%	**19.8 ± 11.1**	**1.9 ± 1.4**	**60.8% ± 7.6%**	21.8 ± 13.2	**1.9 ± 1.7**
Brachial_Rt	54.6% ± 10.1%	**19.9 ± 9.7**	**2.0 ± 1.6**	**59.6% ± 7.8%**	24.8 ± 11.7	**2.0 ± 1.8**
Const_Inf	68.2% ± 11.9%	7.2 ± 2.9	1.3 ± 0.5	**70.2% ± 10.9%**	**5.9 ± 2.6**	**1.1 ± 0.5**
Const_Mid	61.3% ± 10.8%	11.5 ± 5.9	1.9 ± 0.8	**63.5% ± 8.7%**	**10.2 ± 5.5**	**1.7 ± 0.6**
Const_Sup	58.6% ± 10.3%	11.1 ± 4.3	2.0 ± 0.9	**61.2% ± 8.8%**	**10.6 ± 3.9**	**1.9 ± 0.7**
Epiglottis	68.1% ± 10.2%	8.9 ± 3.6	1.3 ± 0.7	**71.3% ± 9.2%**	**6.7 ± 3.0**	**1.1 ± 0.6**
Esophagus	**74.2% ± 10.5%**	**16.3 ± 11.4**	**2.0 ± 2.4**	72.7% ± 11.2%	28.9 ± 30.6	4.2 ± 7.2
GSL	58.6% ± 15.3%	9.5 ± 6.3	2.7 ± 2.1	**67.8% ± 10.8%**	**6.1 ± 1.9**	1.7 ± 0.6
OralCavity	73.4% ± 6.0%	21.2 ± 5.1	5.1 ± 1.3	**75.5% ± 7.4%**	**19.2 ± 5.2**	**4.0 ± 1.6**
Parotid_Lt	83.2% ± 5.8%	9.6 ± 3.3	1.4 ± 0.6	**88.4% ± 4.3%**	**7.8 ± 4.0**	**0.9 ± 0.4**
Parotid_Rt	82.7% ± 6.2%	10.6 ± 4.6	1.5 ± 0.7	**87.7% ± 3.9%**	**8.4 ± 4.5**	**1.0 ± 0.5**
SMG_Lt	79.2% ± 8.9%	7.7 ± 4.4	1.3 ± 0.6	**82.0% ± 7.8%**	**6.5 ± 4.1**	**1.0 ± 0.5**
SMG_Rt	77.7% ± 9.2%	7.9 ± 4.0	1.4 ± 0.8	**82.2% ± 6.6%**	**6.4 ± 2.8**	**1.0 ± 0.4**
TempLobe_Lt	80.9% ± 6.2%	13.9 ± 5.9	2.4 ± 0.9	**82.9% ± 5.2%**	**13.0 ± 5.0**	**2.2 ± 0.7**
TempLobe_Rt	81.4% ± 5.6%	13.9 ± 5.1	2.3 ± 0.8	**83.4% ± 5.2%**	**12.1 ± 4.6**	**2.1 ± 0.7**
Thyroid_Lt	80.0% ± 9.8%	**7.5 ± 4.7**	**1.0 ± 0.7**	**82.8% ± 8.7%**	7.7 ± 15.0	1.1 ± 3.5
Thyroid_Rt	80.6% ± 8.9%	7.4 ± 4.9	1.0 ± 0.9	**84.1% ± 5.8%**	**6.3 ± 4.0**	**0.8 ± 0.4**
S&H OARs						
Cochlea_Lt	62.8% ± 15.9%	2.8 ± 1.5	0.8 ± 0.7	**66.0% ± 11.4%**	**2.3 ± 0.7**	**0.6 ± 0.3**
Cochlea_Rt	61.7% ± 16.1%	2.9 ± 1.6	0.8 ± 0.7	**66.5% ± 10.7%**	**2.3 ± 0.7**	**0.6 ± 0.3**
Hypothalamus	37.5% ± 23.1%	9.2 ± 4.2	3.0 ± 1.9	**59.1% ± 11.5%**	**5.7 ± 2.2**	**1.4 ± 0.7**
InnerEar_Lt	65.6% ± 11.3%	4.2 ± 1.6	1.1 ± 0.6	**75.3% ± 7.9%**	**3.0 ± 0.7**	**0.6 ± 0.3**
InnerEar_Rt	66.0% ± 10.4%	4.2 ± 1.4	1.1 ± 0.5	**75.0% ± 7.8%**	**3.0 ± 0.7**	**0.7 ± 0.6**
LacrimalGland_Lt	45.9% ± 13.7%	5.7 ± 1.4	1.6 ± 0.5	**57.8% ± 9.5%**	**4.0 ± 0.9**	**0.9 ± 0.3**
LacrimalGland_Rt	43.6% ± 13.9%	5.6 ± 1.3	1.6 ± 0.5	**56.3% ± 10.2%**	**4.3 ± 1.2**	**1.0 ± 0.3**
Lens_Lt	70.9% ± 8.9%	2.8 ± 0.7	0.6 ± 0.3	**74.8% ± 9.7%**	**2.7 ± 0.8**	**0.4 ± 0.3**
Lens_Rt	72.4% ± 9.7%	2.8 ± 0.7	0.5 ± 0.3	**79.5% ± 8.3%**	**2.2 ± 0.8**	**0.3 ± 0.2**
OpticChiasm	59.8% ± 15.8%	6.5 ± 2.4	1.4 ± 0.7	**67.1% ± 11.4%**	**6.4 ± 2.1**	**0.8 ± 0.5**
OpticNerve_Lt	67.6% ± 8.6%	5.2 ± 2.6	0.8 ± 0.3	**69.8% ± 7.3%**	**4.8 ± 3.1**	**0.7 ± 0.3**
OpticNerve_Rt	67.0% ± 9.7%	5.4 ± 4.6	0.8 ± 0.5	**68.2% ± 7.1%**	**4.5 ± 3.0**	**0.7 ± 0.3**
PinealGland	50.6% ± 14.0%	4.0 ± 1.4	1.1 ± 0.5	**55.6% ± 10.1%**	**3.6 ± 1.3**	**0.9 ± 0.4**
Pituitary	60.2% ± 16.0%	4.1 ± 1.3	1.0 ± 0.4	**69.6% ± 12.1%**	**3.4 ± 1.2**	**0.6 ± 0.4**
Average Anchor	82.7%	8.9	1.6	**86.9%**	**5.0**	**0.7**
Average Mid-level	72.1%	11.8	1.9	**74.6%**	**12.4**	**1.8**
Average S&H	59.4%	4.7	1.2	**67.2%**	**3.7**	**0.7**
Average all	69.8%	8.8	1.6	**74.8%**	**7.9**	**1.2**

Table 4. Quantitative comparisons on the external testing dataset of 965 patient. The “#” and “OAR” in each parenthesis denote the number of patients and the number of annotated OARs, respectively. SOARS achieves the best average performance in all metrics amongst five external centers. DSC, HD and ASD represent Dice similarity coefficient, Hausdorff distance and average surface distance, respectively. Bold and highlighted values represent the best performance and significant improvement (calculated using Wilcoxon matched-pairs signed rank test) as compared between UaNet and SOARS, respectively.

OARs	UaNet			SOARS		
	DSC	HD (mm)	ASD (mm)	DSC	HD (mm)	ASD (mm)
BrainStem	77.7% ± 10.7%	11.4 ± 11.1	2.6 ± 2.4	**81.2% ± 9.8%**	**9.6 ± 11.3**	**2.0 ± 2.3**
Eye_Lt	86.8% ± 5.6%	3.9 ± 1.4	0.7 ± 0.4	**89.1% ± 4.8%**	**3.7 ± 1.5**	**0.5 ± 0.3**
Eye_Rt	86.6% ± 6.4%	4.1 ± 4.1	0.8 ± 3.1	**88.9% ± 4.2%**	**3.6 ± 1.0**	**0.5 ± 0.3**
InnerEar_Lt	55.1% ± 12.8%	8.0 ± 7.4	1.9 ± 1.0	**61.6% ± 14.0%**	**4.9 ± 2.0**	**0.9 ± 0.6**
InnerEar_Rt	54.0% ± 14.5%	9.4 ± 11.2	2.4 ± 2.4	**64.0% ± 13.8%**	**4.7 ± 1.9**	**0.8 ± 0.5**
Lens_Lt	74.4% ± 11.1%	2.6 ± 1.0	0.5 ± 0.4	**76.8% ± 9.7%**	**2.5 ± 1.0**	**0.4 ± 0.4**
Lens_Rt	74.7% ± 10.8%	2.6 ± 1.0	**0.4 ± 0.5**	**76.9% ± 9.4%**	**2.5 ± 0.9**	**0.4 ± 0.3**
Mandible_Lt	85.5% ± 12.2%	9.2 ± 8.8	1.5 ± 2.7	**88.9% ± 3.5%**	**7.6 ± 7.5**	**1.2 ± 1.0**
Mandible_Rt	85.8% ± 7.1%	9.3 ± 8.0	1.3 ± 1.2	**89.2% ± 3.3%**	**7.7 ± 7.6**	**1.2 ± 1.0**
OpticChiasm	55.1% ± 15.6%	9.1 ± 5.1	2.1 ± 1.4	**66.2% ± 12.3%**	**6.6 ± 4.2**	**1.0 ± 0.6**
OpticNerve_Lt	63.8% ± 12.8%	7.6 ± 5.2	1.1 ± 1.6	**66.8% ± 8.2%**	**5.3 ± 2.7**	**0.7 ± 0.4**
OpticNerve_Rt	65.5% ± 12.2%	6.7 ± 4.1	1.0 ± 0.9	**66.6% ± 8.3%**	**5.1 ± 2.3**	**0.7 ± 0.3**
OralCavity	66.4% ± 5.6%	**23.6 ± 3.8**	5.7 ± 1.0	**68.5% ± 7.2%**	25.7 ± 4.5	**4.8 ± 1.4**
Parotid_Lt	83.2% ± 5.9%	11.6 ± 6.9	1.4 ± 0.8	**85.7% ± 5.0%**	**10.0 ± 6.9**	**1.1 ± 0.6**
Parotid_Rt	82.8% ± 6.4%	11.9 ± 8.7	1.6 ± 2.1	**85.2% ± 5.1%**	**10.6 ± 8.2**	**1.2 ± 1.6**
Pituitary	67.5% ± 15.4%	4.1 ± 1.4	0.9 ± 0.7	**74.7% ± 10.6%**	**3.6 ± 1.1**	**0.5 ± 0.4**
SpinalCord	81.2% ± 10.1%	10.6 ± 19.4	1.3 ± 4.6	**83.8% ± 7.1%**	**7.2 ± 15.7**	**1.1 ± 4.6**
SMG_Lt	72.0% ± 2.0%	9.4 ± 4.9	2.4 ± 0.3	76.8% ± 4.9%	6.4 ± 2.3	**1.3 ± 0.2**
SMG_Rt	**75.1% ± 3.2%**	**8.2 ± 4.9**	1.5 ± 0.4	74.8% ± 5.6%	9.1 ± 4.3	**0.9 ± 0.1**
TempLobe_Lt	75.9% ± 4.3%	22.5 ± 6.7	2.6 ± 1.1	**78.7% ± 3.0%**	**20.7 ± 5.8**	**2.2 ± 0.9**
TempLobe_Rt	78.2% ± 4.3%	**20.2 ± 5.8**	**2.1 ± 0.9**	**79.1% ± 3.3%**	20.4 ± 7.1	**2.1 ± 0.9**
Thyroid_Lt	73.1% ± 10.0%	14.8 ± 15.3	2.2 ± 2.3	**74.5% ± 10.4%**	**14.5 ± 16.0**	**2.1 ± 2.6**
Thyroid_Rt	73.6% ± 10.8%	10.4 ± 5.5	1.6 ± 1.2	**76.4% ± 9.7%**	**9.2 ± 4.6**	**1.4 ± 1.0**
TMJ_Lt	63.7% ± 12.6%	6.0 ± 4.3	1.6 ± 1.1	**75.3% ± 9.2%**	**4.1 ± 1.5**	**0.7 ± 0.4**
TMJ_Rt	64.9% ± 12.0%	6.0 ± 4.7	1.6 ± 1.2	**74.0% ± 9.4%**	**4.3 ± 1.9**	**0.8 ± 0.5**
FAH-XJU (#82, OAR 13)	74.8%	7.2	1.2	**77.3%**	**6.4**	**1.0**
FAH-ZU (#447, OAR 13)	73.7%	7.5	1.3	**77.4%**	**5.9**	**0.9**
GPH (#50, OAR17)	76.0%	7.6	1.4	**80.7%**	**6.8**	**0.9**
HHA-FU (#195, OAR 13)	73.5%	8.0	1.5	**77.7%**	**6.4**	**1.0**
SMU (#227, OAR 25)	73.4%	9.5	1.8	**76.9%**	**8.1**	**1.3**
Average all	74.3%	8.0	1.4	**78.0%**	**6.7**	**1.0**

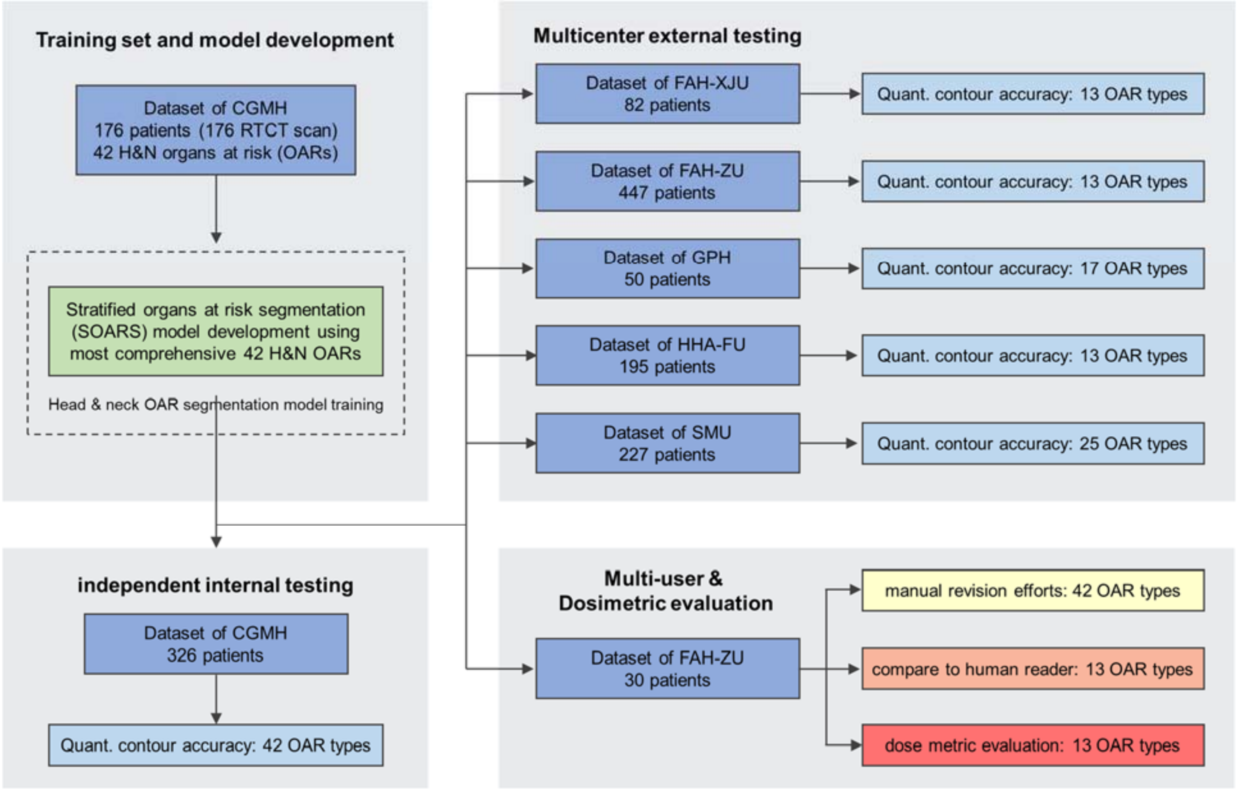
Table 5. Quantitative comparisons between SOARS, SOARS-revised and human reader contour accuracy on the multi-user testing dataset of 30 patient. DSC, HD and ASD represent Dice similarity coefficient, Hausdorff distance, and average surface distance, respectively. Difference in mean dose and difference in maximum dose are calculated using the equation (6) and (7), respectively. DSC higher the better, while HD, difference in mean dose and difference in maximum dose lower the better. SOARS and SOARS-revised results are compared to human reader results using Wilcoxon matched-pairs signed rank test, and bold and highlighted values represent the best performance and significant improvement (calculated using Wilcoxon matched-pairs signed rank test) as compared between UaNet and SOARS, respectively.

	Segmentation accuracy					
OARs	human reader		SOARS		SOARS-revised	
	DSC	HD (mm)	DSC	HD (mm)	DSC	HD (mm)
BrainStem	86.3% ± 3.4%	6.1 ± 1.7	**88.3% ± 2.8%**	**5.3 ± 1.5**	**89.7% ± 2.1%**	**4.4 ± 1.0**
Eye_Lt	91.0% ± 1.9%	3.2 ± 0.4	**91.4% ± 1.6%**	**2.9 ± 0.3**	91.4% ± 1.6%	**2.9 ± 0.3**
Eye_Rt	89.1% ± 6.3%	3.4 ± 0.7	**90.3% ± 4.9%**	**3.1 ± 0.3**	**90.3% ± 4.9%**	**3.1 ± 0.4**
Lens_Lt	74.7% ± 10.2%	2.5 ± 0.8	74.0% ± 8.7%	**2.3 ± 0.6**	**74.8% ± 8.3%**	**2.3 ± 0.6**
Lens_Rt	71.6% ± 11.0%	2.7 ± 0.9	**76.7% ± 6.1%**	**2.2 ± 0.6**	**76.8% ± 6.1%**	**2.2 ± 0.6**
OpticChiasm	69.3% ± 16.1%	5.1 ± 1.6	**78.1% ± 7.8%**	**4.1 ± 1.0**	**80.2% ± 7.3%**	**3.9 ± 1.0**
OpticNerve_Lt	65.9% ± 8.0%	9.8 ± 3.2	**75.3% ± 6.8%**	**4.8 ± 2.2**	**76.4% ± 6.2%**	**3.6 ± 0.9**
OpticNerve_Rt	65.0% ± 11.6%	8.9 ± 4.7	**72.1% ± 7.8%**	**4.5 ± 2.0**	**72.5% ± 8.3%**	**3.9 ± 1.2**
Parotid_Lt	85.1% ± 2.6%	13.8 ± 5.6	**90.2% ± 2.2%**	**7.3 ± 2.7**	**90.2% ± 2.2%**	**7.3 ± 2.7**
Parotid_Rt	84.5% ± 4.2%	13.1 ± 5.9	**89.9% ± 2.5%**	**7.0 ± 3.3**	**90.1% ± 2.2%**	**6.5 ± 1.6**
SpinalCord	82.5% ± 6.9%	12.9 ± 8.7	**84.4% ± 2.5%**	**4.6 ± 1.5**	**84.9% ± 2.4%**	**3.7 ± 0.6**
TMJ_Lt	74.7% ± 11.2%	3.3 ± 0.9	**79.7% ± 7.9%**	3.2 ± 0.6	**82.0% ± 9.8%**	2.9 ± 0.6
TMJ_Rt	72.4% ± 15.4%	3.4 ± 1.4	**77.4% ± 7.3%**	3.5 ± 0.5	**80.3% ± 11.6%**	2.9 ± 0.7
Average	77.8%	7.8	**82.1%**	**4.2**	**83.0%**	**3.8**

	Dosimetric accuracy					
OARs	human reader		SOARS		SOARS-revised	
	diff in mean dose	diff in max dose	diff in mean dose	diff in max dose	diff in mean dose	diff in max dose
BrainStem	3.2% ± 3.1%	4.3% ± 3.7%	**2.4% ± 2.6%**	**3.3% ± 2.6%**	**2.3% ± 2.7%**	**3.3% ± 2.5%**
Eye_Lt	3.4% ± 3.8%	5.7% ± 7.0%	**3.3% ± 3.6%**	**4.9% ± 5.2%**	**3.3% ± 3.6%**	**4.8% ± 5.2%**
Eye_Rt	4.7% ± 5.2%	5.8% ± 5.5%	**4.5% ± 4.9%**	**5.7% ± 5.5%**	**4.5% ± 4.9%**	**5.7% ± 5.5%**
Lens_Lt	2.2% ± 3.0%	3.1% ± 3.3%	**1.9% ± 2.5%**	3.4% ± 3.0%	**1.6% ± 1.4%**	**2.9% ± 2.6%**
Lens_Rt	2.9% ± 3.6%	5.1% ± 6.3%	**1.9% ± 2.6%**	**3.9% ± 6.5%**	**1.9% ± 2.6%**	**3.8% ± 6.5%**
OpticChiasm	6.3% ± 8.5%	5.5% ± 9.7%	**3.7% ± 5.0%**	**2.4% ± 5.8%**	**4.0% ± 6.2%**	**3.1% ± 7.2%**
OpticNerve_Lt	12.7% ± 10.1%	5.2% ± 5.6%	**9.4% ± 8.4%**	**3.1% ± 6.2%**	**9.5% ± 8.1%**	**1.9% ± 5.3%**
OpticNerve_Rt	**10.4% ± 8.5%**	4.7% ± 4.6%	10.9% ± 11.4%	**1.9% ± 3.8%**	11.8% ± 11.5%	**2.5% ± 5.0%**
Parotid_Lt	3.9% ± 3.3%	1.8% ± 2.0%	**1.9% ± 1.5%**	**1.2% ± 1.5%**	**1.9% ± 1.5%**	**1.2% ± 1.5%**
Parotid_Rt	4.2% ± 3.6%	1.8% ± 1.6%	**2.0% ± 2.0%**	**0.6% ± 0.9%**	**2.0% ± 2.0%**	**0.6% ± 0.9%**
SpinalCord	13.3% ± 13.5%	2.2% ± 3.5%	**1.7% ± 1.4%**	**2.0% ± 2.4%**	**1.6% ± 1.3%**	**2.0% ± 2.7%**
TMJ_Lt	2.1% ± 2.0%	2.3% ± 2.2%	2.2% ± 1.7%	2.9% ± 2.1%	**1.6% ± 1.5%**	**2.6% ± 2.0%**
TMJ_Rt	**1.6% ± 1.3%**	1.9% ± 2.2%	2.1% ± 1.2%	**1.9% ± 1.7%**	1.9% ± 1.0%	**1.7% ± 1.5%**
Average	5.5%	3.8%	**3.7%**	**2.9%**	**3.7%**	**2.8%**

Note: diff in mean dose and diff in max dose represent the difference in mean dose and difference in maximum dose, respectively.

Figures



Training set and model development
Dataset of CGMH
176 patients (176 RTCT scan)
42 H&N organs at risk (OARs)
Stratified organs at risk segmentation (SOARS) model development using most comprehensive 42 H&N OARs
Head & neck OAR segmentation model training
Multicenter external testing
Dataset of FAH-XJU
82 patients
Quant. contour accuracy: 13 OAR types
Dataset of FAH-ZU
447 patients
Quant. contour accuracy: 13 OAR types
Dataset of GPH
50 patients
Quant. contour accuracy: 17 OAR types
Dataset of HHA-FU
195 patients
Quant. contour accuracy: 13 OAR types
Dataset of SMU
227 patients
Quant. contour accuracy: 25 OAR types
independent internal testing
Dataset of CGMH
326 patients
Quant. contour accuracy: 42 OAR types
Multi-user &
Dosimetric evaluation
Dataset of FAH-ZU
30 patients
manual revision efforts: 42 OAR types
compare to human reader: 13 OAR types
dose metric evaluation: 13 OAR types

Fig. 1. The study flow diagram. We totally collected 1533 head & neck cancer patients to develop and evaluate the performance of stratified organ at risk segmentation (SOARS). The training patients were collected from Chang Gung Memorial Hospital (CGMH), while the testing patients were collected from the internal institution CGMH and other five external institutions including First Affiliated Hospital of Xi'an Jiaotong University (FAH-XJU), and First Affiliated Hospital of Zhejiang University (FAH-ZU), Huadong Hospital Affiliated of Fudan University (HHA-FU), Gansu Provincial Hospital (GPH), Southern Medical University (SMU). We further randomly collected 30 nasopharyngeal cancer patients from FAH-ZU to form a multi-user testing dataset to evaluate the clinical applicability of SOARS, including the effort for manual revision, comparison to the inter-user OAR segmentation accuracy and comparison to the inter-user OAR dosimetric accuracy.

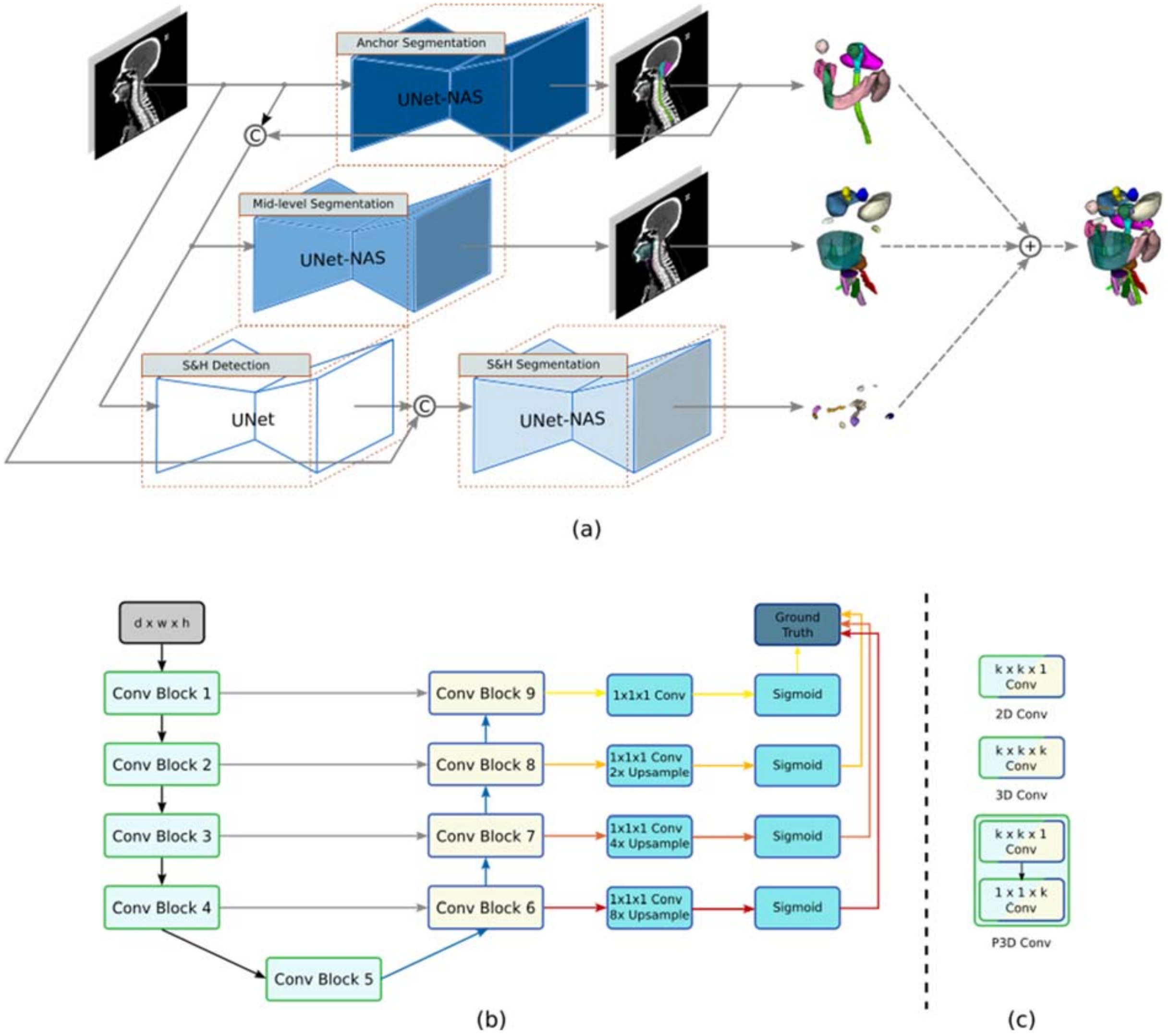


Fig. 2. (a) stratified organ at risk segmentation (SOARS) stratifies OAR into anchor, mid-level, and small & hard (S&H) categories and uses the anchor OARs to guide the mid-level and S&H OAR segmentation. (b) The backbone network UNet with neural architecture search (NAS), which permits an automatic selection across 2D, 3D, and P3D convolution blocks. (c) The NAS convolution setting.

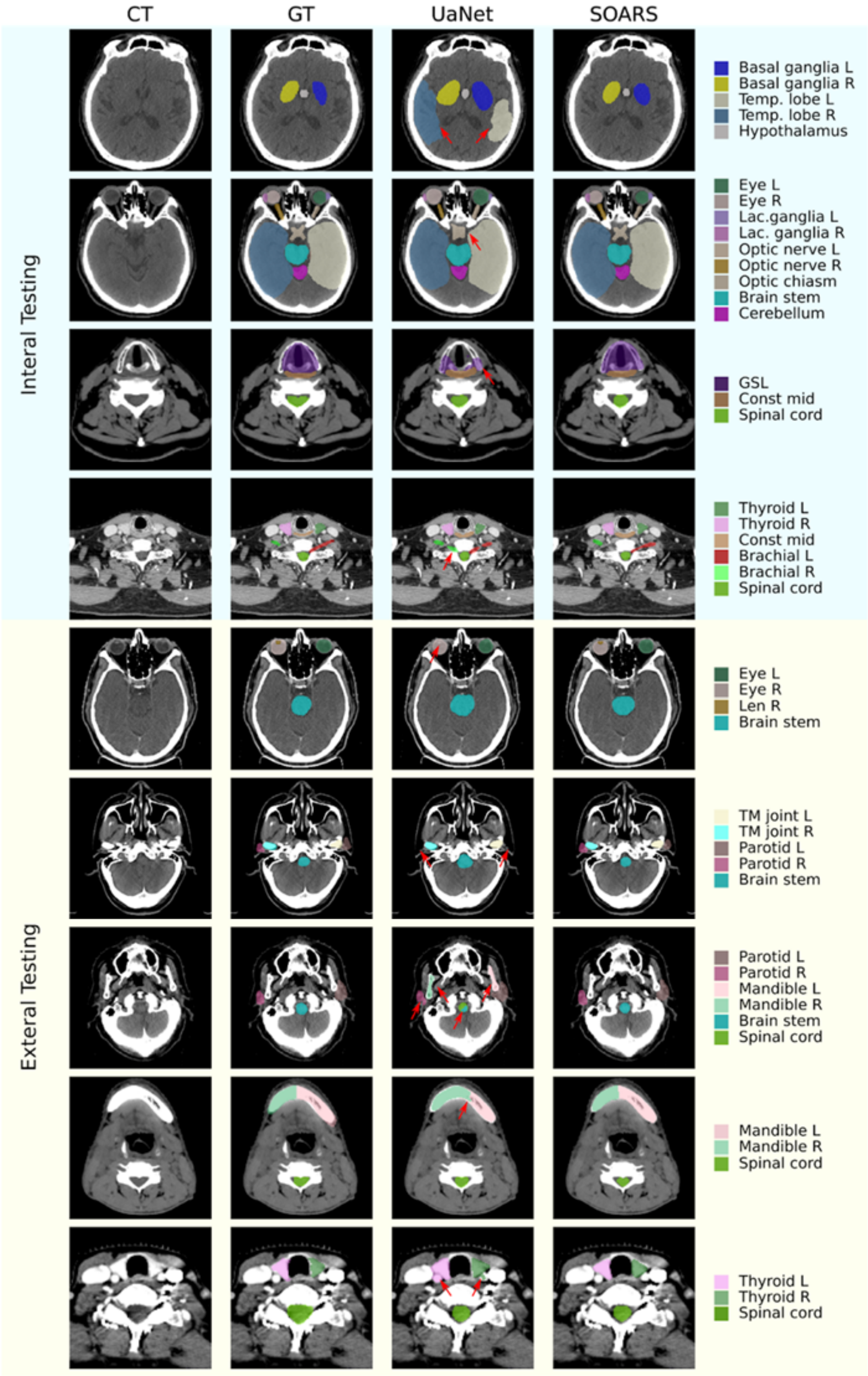


Fig. 3. Qualitative 42-OAR segmentation using UaNet and SOARs on internal (upper 4 rows) & external (lower 5 rows) datasets. Rows 5-9 are sample images from GPH, FAH-ZU, HHA-FU, SMU, and FAH-XJU, respectively. The 1-4 columns are pCT image, pCT with manual OAR delineations, pCT with UaNet predictions, pCT with SOARS predictions, respectively. The five external centers have different OAR delineation protocols -- a subset of 42 OARs is manually labeled. For better comparison, we only show the ground truth associated predictions and use red arrows to indicate the improvements.

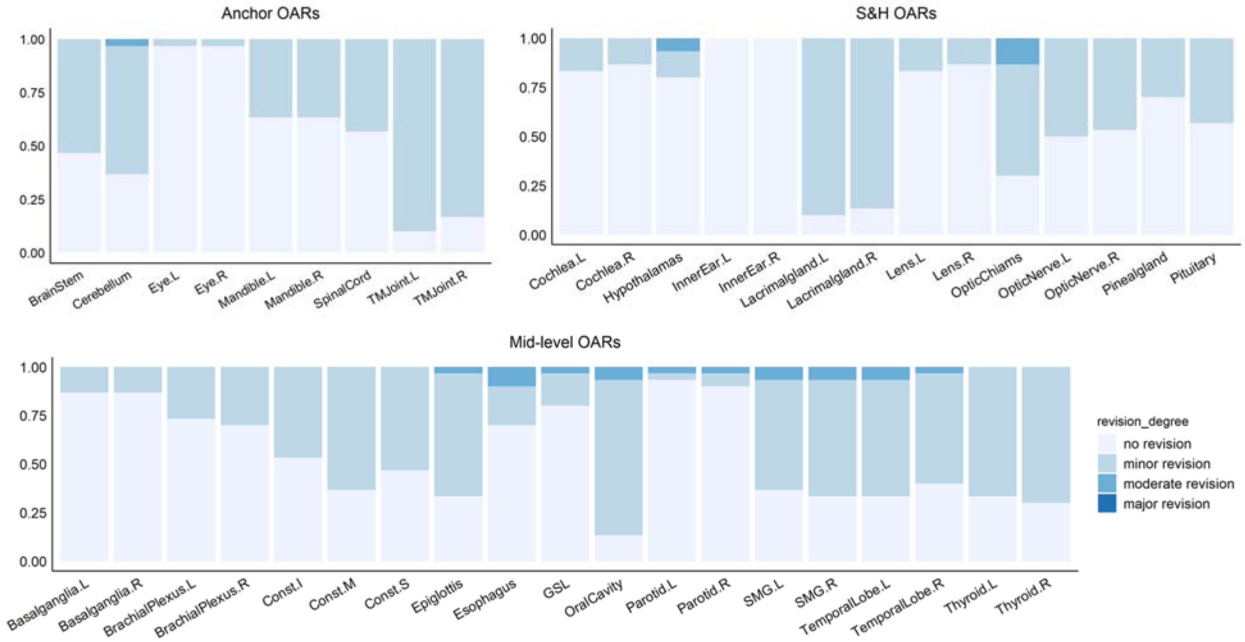


Fig. 4. Summary of human experts' assessment of revision effort on SOARS predicted 42 OARs. Anchor, mid-level and S&H OAR categories are shown separately. Vast majority of SOARS predicted OARs only required minor revision or no revision from expert's editing before they can be clinically accepted. Only a very small amount of OARs need moderate revision and no OARs need major revision. Minor revision: editing required in <1 minute; moderate revision: editing required in 1–3 minutes; and major revision: editing required in >3 minutes.

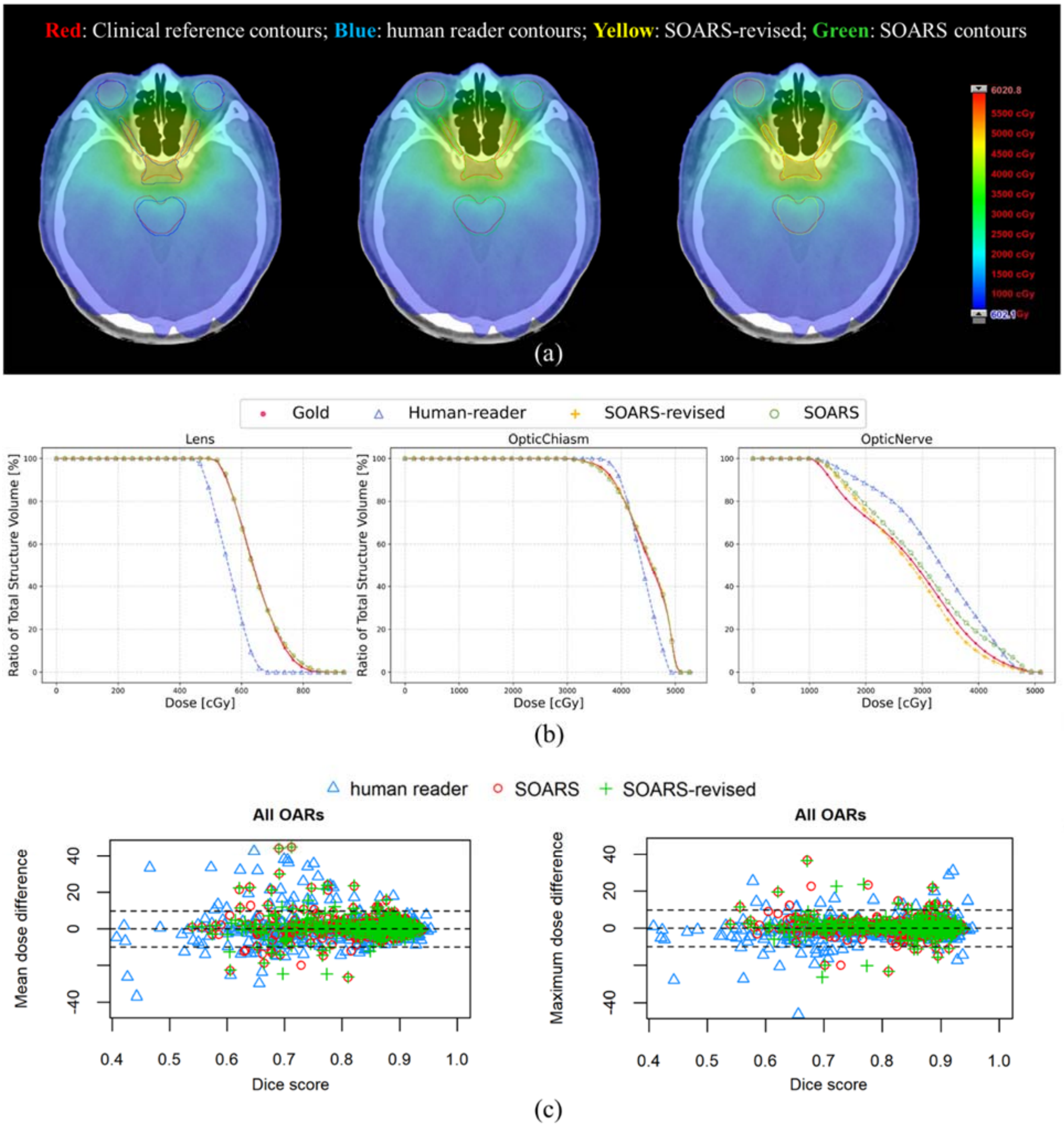


Fig. 5. Using a specific patient, we show a qualitatively dosimetric example (a) in axial views of two anatomic locations. Clinical OAR reference: red; human reader: blue; SOARS OAR: green; SOARS-revised: yellow. (b) the dose–volume histograms (DVH) plot of OARs in this patient. (c) The scatter plot of differences in mean dose and differences in maximum dose brought by various OAR contour sets of human reader, SOARS, and SOARS-revised among 30 multi-user testing patients. Blue triangle, green cross and red circle represent results of human reader, SOARS-revised and SOARS, respectively.

Supplementary materials

Supplementary Method

Detailed segmentation baseline and auto-searched network architecture

The UNet backbone adopted in our work, i.e., the nnUNet {nnUNet ref}, includes a 5-block encoding path and a 4-block decoding path. Each encoding block consists of the following consecutive operations with residual connection: a convolution, a instance normalization {D. Ulyanov 2016 arXiv}, a Leaky ReLu unit, followed by a 2x2x2 max-pooling operator. Each decoding block is composed of a 2x2x2 up-sampling process, followed by consecutive operations similar to the ones in the encoding block. The specific convolution operation in each block is automatically determined using network architecture search (NAS) {Liu, 2018 #102}{ Liu, 2019 #57} with the search space defined by 2D, 3D, pseudo-3D (P3D) convolutions with kernel size of 3 or 5. The detailed convolutional neural network architectures for each organ at risk (OAR) segmentation branch are described in Supplementary Fig 1.

Implementation details

Image preprocessing. A windowing of [-500, 1000] HU to every pCT scan is applied covering the intensity range of our target OARs. VOIs of 256×256×64 voxels are randomly extracted around the OAR foreground as training samples for NAS. The heat map labels in the detection module are 3D Gaussian distributions (zero mean with standard deviation of 8mm) centered at the center of each S&H OAR.

NAS training. We exploit NAS to search the optimal network architecture for each stratified OAR segmentation branch. The combined Dice and Cross-Entropy losses are adopted and the RAdam optimizer [36] is used with a momentum of 0.9 and a weight decay of 0.005. To train the NAS parameter α_k, we first fix α_k to 1/9 for 200 epochs. Then we alternatively update α_k and the network weights for another additional 300 epochs. The batch size is set to 2 for NAS training. Only the validation set is used for updating *α*. The ratio between the training set and the validation set is 2:1. The initial learning rates are set to 0.005 for the anchor and mid-level branches, and 0.001 for the S&H branch, respectively.

Final segmentation network training. After NAS is completed, we retrain the searched segmentation network from scratch. Data augmentation is applied {nnUNet ref}, e.g., horizontal flipping, random rotations in the x-y plane within ±10 degrees, intensity scaling with a ratio between [0.75, 1.25], adding Gaussian noise with zero mean and (0, 0.1) variance. The batch size is 2. The optimizer is stochastic gradient descent with a Polynomial learning rate policy. The initial learning rate is 0.01 with a Nesterov momentum of 0.99. The S&H detection branch is trained using L2 loss with a 0.01 learning rate. The total number of training epochs for each module is 1000. The average training time is 9~10 GPU days. For inference, the average running time is normally less than 3 minutes per patient. All deep models are developed using PyTorch and trained on one NVIDIA Quadro RTX 8000 GPU.

Quantitative ablation results of SOARS in the training-validation dataset

Effect of processing stratification in SOARS. Processing stratification played a key role to improve the OAR segmentation performance. The processing stratification ablation results are shown in Table 2. The baseline is using 3D UNet model (implemented in the nnUNet framework {nnUNet ref}) trained on all 42 OARs together. When anchor OARs were stratified to train only on themselves, there was a 2.4% Dice similarity coefficient (DSC) improvement as compared to the baseline models. When focusing on mid-level OARs, with the help of anchor OAR guidance, there was a significant 37% Hausdorff distance (HD) error reduction (11.4 versus 18.0mm) as compared to the baseline model of training on all OARs. This demonstrated the intrinsic difficulty in segmenting a large number of various organs without explicitly taking their differences into account. It simultaneously indicated that anchor OARs served as effective references to better delineate the hard-to-discern boundaries of mid-level organs (most are soft-tissue organs). For S&H OARs, by cropping the volume of interest (VOI) using the detection module and with the support of anchor OAR predictions, there were remarkable accuracy improvements in segmenting S&H OARs, boosting DSC from 58.3% to 73.7%, as compared against directly segmenting from CT. This further demonstrated the merits and advantages of our stratified learning approach that adapted to provide the optimal handling of OAR categories with different characteristics. Fig. 3 depicts qualitative examples of segmenting anchor, mid-level and S&H OARs.

Effect of neural architecture search (NAS) associated with SOARS. Table 2 also outlines the performance improvements provided by NAS. As can be seen, all three branches trained with NAS consistently produced more accurate segmentation results than those trained using the baseline 3D UNet. This validated the effectiveness of NAS on more complicated segmentation tasks. For the three branches, mid-level and S&H OAR categories showed considerable performance improvements, from 72.6% to 74.2% and 73.7% to 76.2% in DSC scores respectively, while the anchor branch provides a marginal but consistent improvement (0.7% in DSC). Considering that anchor OARs are already relatively easy to segment, the fact that NAS can further boost the performance attested to its benefits.

The NAS searched neural network architectures are depicted in Supplementary Fig. 1. It is observed that, for the encoding path, the mid-level and S&H branches gradually involve more 3D or P3D convolution kernels as compared to the anchor branch. This indicates that 3D kernels may not always be the best choice for segmenting objects with reasonable size or contrast, as 2D kernels dominate the anchor branch. Consequently, appropriate 2D and P3D kernels can reduce the computation cost and memory consumption. For the S&H branch, our findings are consistent with the intuition that small or low contrast objects rely more on the 3D spatial information and context for better segmentation. As for the decoding path, all three branches are mainly equipped with 3D or P3D convolution kernels. This is an interesting result, as it implies that the decoding path tries to incorporate the convolutional features in a more 3D fashion for all three OAR categories.

References

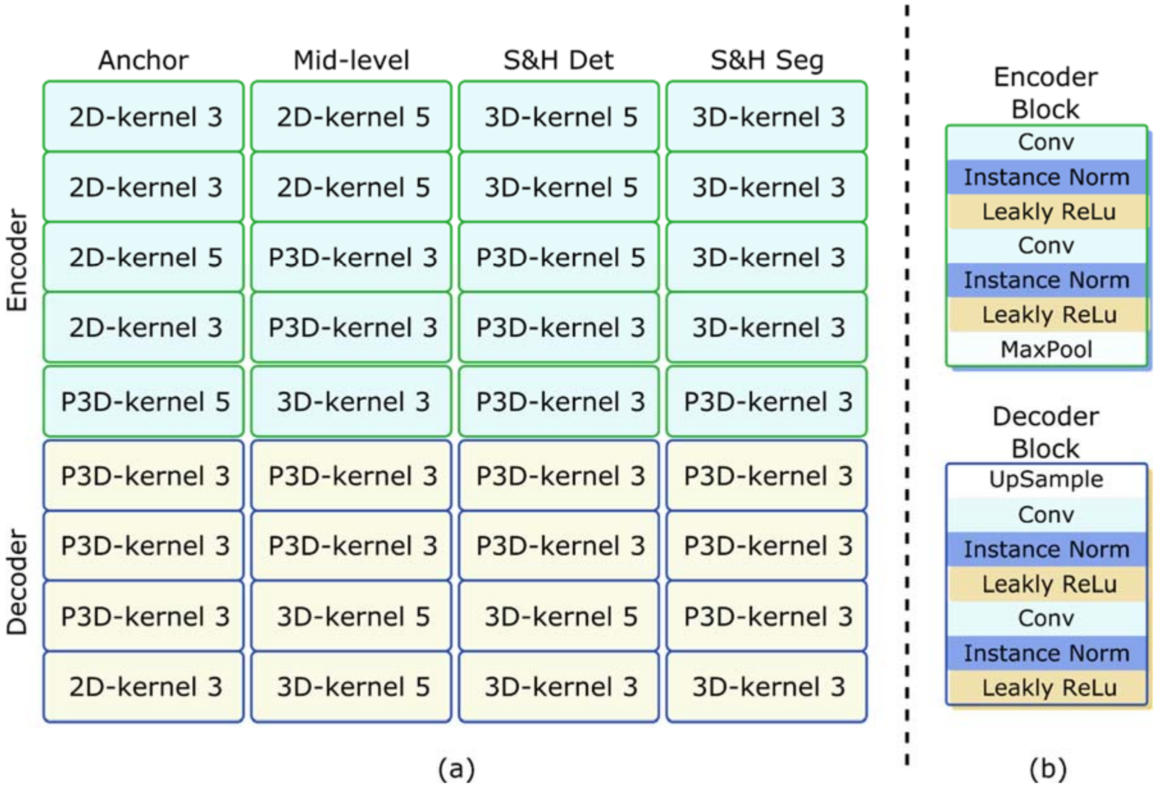
Isensee, Fabian, et al. "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation." *Nature methods* 18.2 (2021): 203-211

D. Ulyanov, A. Vedaldi, and V. Lempitsky, “Instance normalization: The missing ingredient for fast stylization,” arXiv preprint arXiv:1607.08022, 2016.

Liu, H., Simonyan, K. & Yang, Y. Darts: Differentiable architecture search. *arXiv preprint arXiv:1806.09055* (2018).

Liu, C. *et al.* Auto-deeplab: Hierarchical neural architecture search for semantic image segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 82-92.

Supplementary Figures



	Anchor	Mid-level	S&H Det	S&H Seg
Encoder	2D-kernel 3	2D-kernel 5	3D-kernel 5	3D-kernel 3
	2D-kernel 3	2D-kernel 5	3D-kernel 5	3D-kernel 3
	2D-kernel 5	P3D-kernel 3	P3D-kernel 5	3D-kernel 3
	2D-kernel 3	P3D-kernel 3	P3D-kernel 3	3D-kernel 3
	P3D-kernel 5	3D-kernel 3	P3D-kernel 3	P3D-kernel 3
Decoder	P3D-kernel 3	P3D-kernel 3	P3D-kernel 3	P3D-kernel 3
	P3D-kernel 3	P3D-kernel 3	P3D-kernel 3	P3D-kernel 3
	P3D-kernel 3	3D-kernel 5	3D-kernel 5	P3D-kernel 3
	2D-kernel 3	3D-kernel 5	3D-kernel 3	3D-kernel 3

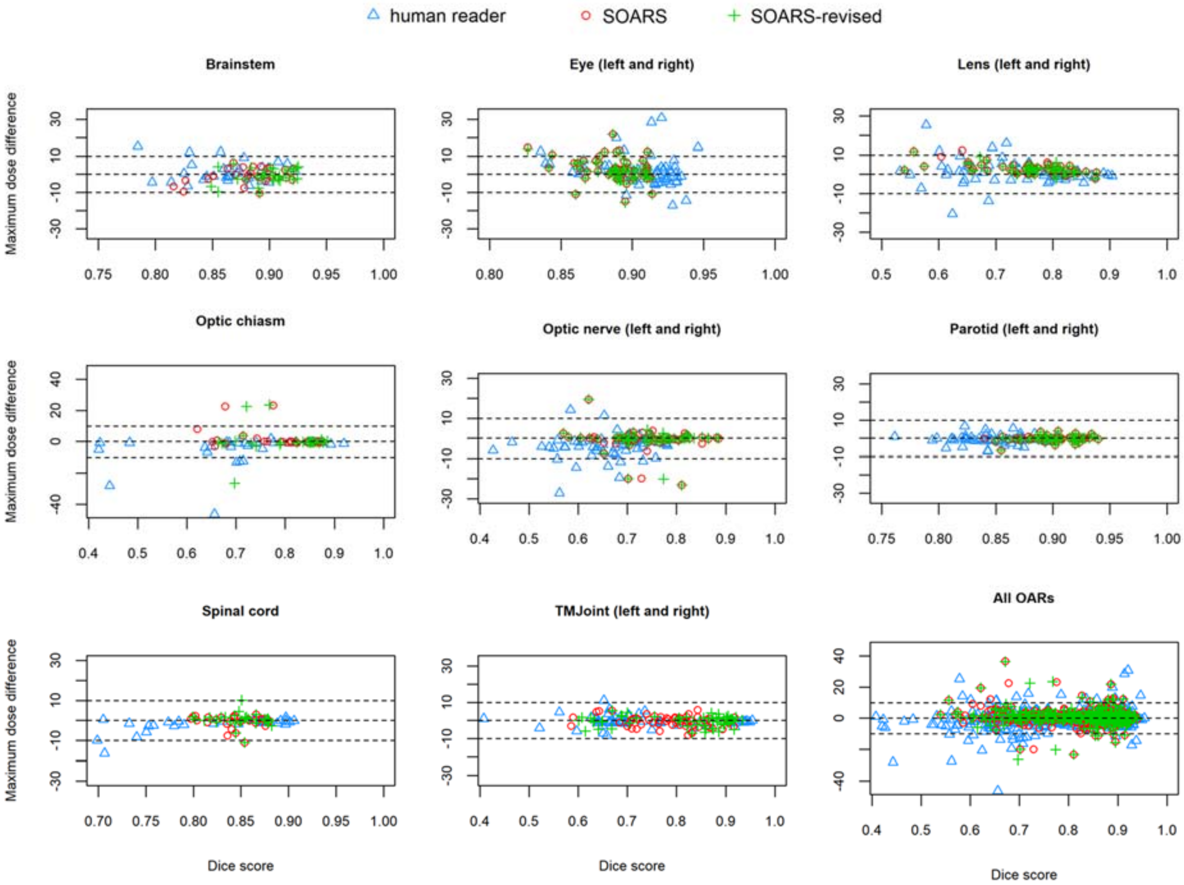
(a)

Encoder Block
Conv
Instance Norm
Leakly ReLu
Conv
Instance Norm
Leakly ReLu
MaxPool

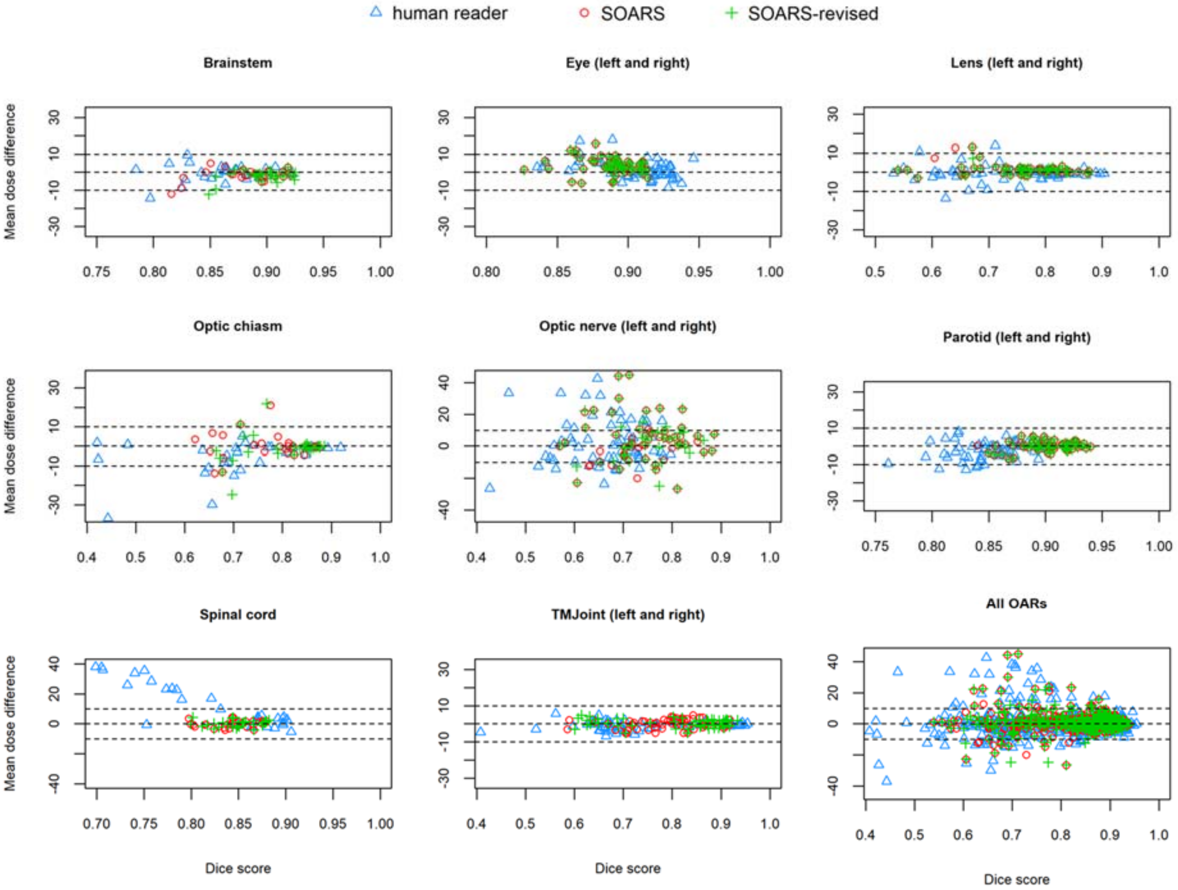
Decoder Block
UpSample
Conv
Instance Norm
Leakly ReLu
Conv
Instance Norm
Leakly ReLu

(b)

Supplementary Fig. 1 (a) illustrates the auto-searched backbone network architecture (UNet) for the anchor, mid-level, and small & hard (S&H) branches. The search space of the convolution operation includes 2D, 3D, and pseudo-3D (P3D) with either kernel size of 3 or 5. (b) lists the detailed operations in the encoder and decoder blocks. The auto-searched two convolution operations within each block are of the same type.



Supplementary Fig. 2. The scatter plot of differences in maximum dose brought by various OAR contour sets of SOARS, SOARS-revised, and human reader. Each OAR and all OAR results are plotted, respectively. Blue triangle, green cross and red circle represent results of human reader, SOARS-revised and SOARS, respectively.



Supplementary Fig. 3. The scatter plot of differences in mean dose brought by various OAR contour sets of SOARS, SOARS-revised, and human reader. Each OAR and all OAR results are plotted, respectively. Blue triangle, green cross and red circle represent results of human reader, SOARS-revised and SOARS, respectively.

Supplementary Tables

Supplementary Table 1. Quantitative comparisons on the external FAH-XJU testing dataset of 82 patients. The proposed SOARS outperforms the previous leading approach UaNet in almost all metrics across different OARs. DSC, HD and ASD represent Dice similarity coefficient, Hausdorff distance, and average surface distance, respectively. Bold and highlighted values represent the best performance and significant improvement calculated using Wilcoxon matched-pairs signed rank test as compared between UaNet and SOARS, respectively.

OARs	UaNet			SOARS		
	DSC	HD (mm)	ASD (mm)	DSC	HD (mm)	ASD (mm)
BrainStem	78.4% ± 6.4%	9.2 ± 3.1	1.9 ± 0.8	**80.5% ± 6.7%**	**8.2 ± 3.0**	**1.8 ± 0.9**
Eye_Lt	**86.7% ± 5.3%**	**5.0 ± 5.3**	**0.8 ± 0.6**	85.8% ± 5.7%	5.4 ± 5.4	0.9 ± 0.7
Eye_Rt	**87.5% ± 2.6%**	**4.1 ± 1.0**	**0.7 ± 0.2**	86.1% ± 3.5%	4.3 ± 1.2	**0.7 ± 0.3**
Lens_Lt	68.4% ± 9.4%	2.8 ± 0.6	0.6 ± 0.3	**69.6% ± 7.0%**	**2.7 ± 0.6**	**0.6 ± 0.2**
Lens_Rt	70.8% ± 8.2%	2.8 ± 0.7	0.5 ± 0.3	**72.9% ± 7.2%**	**2.4 ± 0.7**	**0.5 ± 0.2**
OpticChiasm	57.6% ± 14.0%	6.7 ± 2.4	1.5 ± 0.8	**68.1% ± 7.0%**	**6.5 ± 3.0**	**1.0 ± 0.5**
OpticNerve_Lt	66.0% ± 7.4%	**5.0 ± 2.6**	**0.8 ± 0.4**	**67.9% ± 6.9%**	6.2 ± 3.0	**0.8 ± 0.3**
OpticNerve_Rt	65.5% ± 8.5%	**4.3 ± 1.1**	**0.8 ± 0.3**	**66.4% ± 6.0%**	4.8 ± 1.2	**0.8 ± 0.2**
Parotid_Lt	78.2% ± 5.2%	11.7 ± 3.0	1.9 ± 0.6	**79.7% ± 5.0%**	**10.5 ± 3.2**	**1.8 ± 0.6**
Parotid_Rt	77.6% ± 6.2%	12.4 ± 4.5	2.0 ± 0.8	**79.4% ± 5.2%**	**10.9 ± 3.8**	**1.8 ± 0.6**
Pituitary	62.4% ± 12.8%	4.4 ± 1.5	1.2 ± 0.6	**75.6% ± 11.1%**	**3.7 ± 1.6**	**0.6 ± 0.5**
SpinalCord	79.2% ± 14.2%	7.5 ± 9.0	1.0 ± 1.2	**82.3% ± 4.9%**	**6.7 ± 9.4**	**0.9 ± 1.4**
TMJ_Lt	76.7% ± 5.2%	12.5 ± 6.5	1.9 ± 1.5	**77.8% ± 9.0%**	**10.9 ± 7.0**	**1.5 ± 1.5**
TMJ_Rt	72.8% ± 12.4%	11.3 ± 8.4	2.1 ± 1.6	**81.4% ± 5.2%**	**6.9 ± 1.5**	**0.9 ± 0.3**
Average	74.8%	7.2	1.2	77.3%	6.4	1.0

Supplementary Table 2. Quantitative comparisons on the external FAH-ZU testing dataset of 447 patients. The proposed SOARS outperforms the previous leading approach UaNet in almost all metrics across different OARs. DSC, HD and ASD represent Dice similarity coefficient, Hausdorff distance, and average surface distance, respectively. Bold and highlighted values represent the best performance and significant improvement calculated using Wilcoxon matched-pairs signed rank test as compared between UaNet and SOARS, respectively.

OARs	UaNet			SOARS		
	DSC	HD (mm)	ASD (mm)	DSC	HD (mm)	ASD (mm)
BrainStem	77.8% ± 10.9%	10.1 ± 6.2	2.6 ± 2.4	**82.4% ± 11.3%**	**8.3 ± 5.7**	**1.8 ± 2.2**
Eye_Lt	**87.9% ± 3.0%**	3.8 ± 0.9	0.6 ± 0.2	**87.9% ± 3.6%**	**3.5 ± 1.0**	**0.5 ± 0.3**
Eye_Rt	86.8% ± 5.6%	4.3 ± 5.7	0.9 ± 4.4	**87.3% ± 2.2%**	**3.7 ± 1.0**	**0.6 ± 0.2**
Lens_Lt	69.6% ± 10.5%	3.0 ± 1.1	0.7 ± 0.5	**71.4% ± 9.0%**	**3.0 ± 1.0**	**0.6 ± 0.4**
Lens_Rt	70.5% ± 10.7%	2.9 ± 1.2	0.7 ± 0.5	**72.0% ± 8.4%**	**2.9 ± 0.9**	**0.6 ± 0.4**
OpticChiasm	53.0% ± 15.7%	9.9 ± 5.8	2.3 ± 1.6	**65.9% ± 12.8%**	**6.6 ± 4.9**	**1.1 ± 0.6**
OpticNerve_Lt	**66.4% ± 9.7%**	8.9 ± 5.0	1.1 ± 1.9	66.3% ± 8.1%	**5.4 ± 2.9**	**0.7 ± 0.5**
OpticNerve_Rt	**68.3% ± 8.4%**	7.6 ± 3.9	0.8 ± 0.4	66.1% ± 7.8%	**5.3 ± 2.3**	**0.7 ± 0.3**
Parotid_Lt	82.2% ± 4.8%	12.8 ± 5.1	1.7 ± 0.7	**85.4% ± 4.6%**	**10.6 ± 4.7**	**1.2 ± 0.5**
Parotid_Rt	82.8% ± 5.2%	12.1 ± 6.0	1.6 ± 0.8	**84.8% ± 4.5%**	**11.1 ± 5.3**	**1.3 ± 0.7**
SpinalCord	83.8% ± 7.6%	13.1 ± 22.6	1.5 ± 6.0	**86.3% ± 7.4%**	**8.6 ± 22.0**	**1.3 ± 6.5**
SMG_Lt	64.3% ± 8.5%	4.4 ± 1.0	1.3 ± 0.4	**76.2% ± 7.7%**	**3.7 ± 0.9**	**0.7 ± 0.4**
SMG_Rt	63.5% ± 10.1%	4.5 ± 1.8	1.3 ± 0.5	**74.6% ± 7.8%**	**3.9 ± 1.7**	**0.8 ± 0.5**
Average	73.5%	7.5	1.3	**77.4%**	**5.9**	**0.9**

Supplementary Table 3. Quantitative comparisons on the external GPH testing dataset of 50 patients. The proposed SOARS outperforms the previous leading approach UaNet in almost all metrics across different OARs. DSC, HD and ASD represent Dice similarity coefficient, Hausdorff distance, and average surface distance, respectively. Bold and highlighted values represent the best performance and significant improvement calculated using Wilcoxon matched-pairs signed rank test as compared between UaNet and SOARS, respectively.

OARs	UaNet			SOARS		
	DSC	HD (mm)	ASD (mm)	DSC	HD (mm)	ASD (mm)
BrainStem	77.1% ± 14.6%	12.6 ± 7.2	2.4 ± 1.1	**78.9% ± 10.5%**	**11.7 ± 9.5**	**2.1 ± 1.4**
Eye_Lt	85.6% ± 3.5%	4.0 ± 0.8	0.8 ± 0.3	**92.1% ± 3.9%**	**3.5 ± 0.7**	**0.4 ± 0.3**
Eye_Rt	85.3% ± 4.6%	4.4 ± 1.2	0.8 ± 0.4	**91.5% ± 4.2%**	**3.3 ± 0.9**	**0.4 ± 0.3**
Lens_Lt	78.1% ± 8.2%	2.1 ± 0.7	0.4 ± 0.2	**82.2% ± 5.2%**	**2.1 ± 0.5**	**0.3 ± 0.2**
Lens_Rt	79.6% ± 9.4%	2.7 ± 0.9	**0.3 ± 0.2**	**81.9% ± 7.5%**	**2.0 ± 0.6**	**0.3 ± 0.2**
Mandible_Lt	89.8% ± 1.4%	**6.7 ± 2.7**	0.8 ± 0.2	**91.7% ± 1.1%**	**6.7 ± 2.7**	**0.7 ± 0.1**
Mandible_Rt	88.8% ± 1.2%	9.1 ± 2.1	0.8 ± 0.2	**91.8% ± 1.2%**	**6.2 ± 2.8**	**0.7 ± 0.1**
OpticChiasm	51.5% ± 16.0%	9.1 ± 2.1	2.2 ± 1.0	**60.1% ± 9.8%**	**7.7 ± 2.2**	**1.1 ± 0.5**
OpticNerve_Lt	57.9% ± 16.7%	6.4 ± 4.0	1.6 ± 1.4	**69.9% ± 6.1%**	**4.8 ± 1.5**	**0.6 ± 0.3**
OpticNerve_Rt	57.4% ± 18.5%	6.5 ± 3.5	1.6 ± 1.5	**69.2% ± 8.3%**	**4.6 ± 1.6**	**0.6 ± 0.3**
OralCavity	69.0% ± 3.1%	**23.6 ± 4.5**	5.3 ± 0.8	**72.2% ± 4.7%**	26.9 ± 4.9	**4.1 ± 0.8**
Parotid_Lt	87.1% ± 4.3%	11.3 ± 6.2	0.9 ± 0.5	**87.6% ± 4.4%**	**9.8 ± 5.8**	**0.8 ± 0.6**
Parotid_Rt	86.5% ± 4.5%	9.7 ± 5.8	0.9 ± 0.6	**87.1% ± 4.4%**	**8.8 ± 4.6**	**0.8 ± 0.5**
Pituitary	88.8% ± 3.3%	2.3 ± 0.8	**0.1 ± 0.1**	**89.0% ± 3.3%**	**2.0 ± 0.3**	**0.1 ± 0.1**
SpinalCord	78.7% ± 5.4%	6.6 ± 2.5	**1.1 ± 0.5**	**78.9% ± 5.1%**	**6.5 ± 2.5**	**1.1 ± 0.5**
TMJ_Lt	65.8% ± 17.1%	8.1 ± 4.6	1.5 ± 0.8	**73.1% ± 20.0%**	**4.1 ± 1.7**	**0.9 ± 0.9**
TMJ_Rt	65.0% ± 17.2%	7.2 ± 3.7	1.5 ± 0.8	**75.3% ± 23.2%**	**4.1 ± 1.7**	**0.8 ± 1.0**
Average	76.0%	7.6	1.4	**80.7%**	**6.8**	**0.9**

Supplementary Table 4. Quantitative comparisons on the external HHA-FU testing dataset of 195 patients. The proposed SOARS outperforms the previous leading approach UaNet in almost all metrics across different OARs. DSC, HD and ASD represent Dice similarity coefficient, Hausdorff distance, and average surface distance, respectively. Bold and highlighted values represent the best performance and significant improvement calculated using Wilcoxon matched-pairs signed rank test as compared between UaNet and SOARS, respectively.

OARs	UaNet			SOARS		
	DSC	HD (mm)	ASD (mm)	DSC	HD (mm)	ASD (mm)
BrainStem	75.8% ± 13.2%	13.3 ± 7.5	2.9 ± 1.8	**78.4% ± 8.9%**	**10.6 ± 6.8**	**2.3 ± 1.4**
Eye_Lt	85.3% ± 7.0%	4.0 ± 1.2	0.9 ± 0.5	**90.6% ± 5.7%**	**3.7 ± 1.1**	**0.5 ± 0.4**
Eye_Rt	86.3% ± 6.6%	3.8 ± 1.1	0.8 ± 0.5	**90.9% ± 5.6%**	**3.5 ± 1.1**	**0.5 ± 0.4**
Lens_Lt	78.4% ± 9.0%	2.3 ± 0.7	0.4 ± 0.3	**82.3% ± 6.9%**	**2.1 ± 0.6**	**0.3 ± 0.2**
Lens_Rt	78.2% ± 8.0%	2.3 ± 0.6	0.4 ± 0.2	**82.4% ± 6.7%**	**2.1 ± 0.6**	**0.3 ± 0.2**
OpticChiasm	50.1% ± 15.1%	10.7 ± 3.5	2.5 ± 1.2	**57.2% ± 10.2%**	**9.5 ± 2.9**	**1.5 ± 0.7**
OpticNerve_Lt	52.4% ± 14.7%	7.2 ± 4.1	1.6 ± 1.2	**62.3% ± 8.5%**	**6.2 ± 2.9**	**1.0 ± 0.4**
OpticNerve_Rt	56.2% ± 13.7%	**6.0 ± 3.1**	1.2 ± 0.8	**61.9% ± 9.7%**	6.2 ± 3.1	**1.0 ± 0.4**
Parotid_Lt	85.1% ± 6.0%	8.6 ± 4.1	1.1 ± 0.7	**85.6% ± 6.0%**	**7.7 ± 3.7**	**1.0 ± 0.7**
Parotid_Rt	84.1% ± 6.5%	10.1 ± 12.9	1.6 ± 4.2	**85.5% ± 5.8%**	**9.2 ± 12.9**	**1.3 ± 4.0**
SpinalCord	74.9% ± 12.6%	12.2 ± 26.7	1.9 ± 4.6	**78.9% ± 6.8%**	**7.6 ± 2.7**	**1.2 ± 0.6**
SMG_Lt	70.7% ± 5.2%	13.5 ± 2.1	2.2 ± 0.2	**78.3% ± 7.9%**	**7.9 ± 1.5**	**1.3 ± 0.3**
SMG_Rt	**76.2% ± 4.5%**	10.5 ± 6.4	1.6 ± 0.7	**76.2% ± 9.1%**	**7.5 ± 2.2**	**1.4 ± 0.9**
Average	73.2%	8.0	1.5	**77.7%**	**6.4**	**1.0**

Supplementary Table 5. Quantitative comparisons on the external SMU testing dataset of 227 patients. The proposed SOARS outperforms the previous leading approach UaNet in almost all metrics across different OARs. DSC, HD and ASD represent Dice similarity coefficient, Hausdorff distance, and average surface distance, respectively. Bold and highlighted values represent the best performance and significant improvement calculated using Wilcoxon matched-pairs signed rank test as compared between UaNet and SOARS, respectively.

OARs	UaNet			SOARS		
	DSC	HD (mm)	ASD (mm)	DSC	HD (mm)	ASD (mm)
BrainStem	78.7% ± 7.9%	12.6 ± 19.1	2.4 ± 3.2	**81.2% ± 7.2%**	**11.4 ± 19.9**	**2.1 ± 3.3**
Eye_Lt	85.8% ± 8.4%	3.8 ± 0.9	0.7 ± 0.3	**90.8% ± 4.7%**	**3.6 ± 0.9**	**0.5 ± 0.3**
Eye_Rt	86.6% ± 8.5%	3.7 ± 0.9	0.7 ± 0.3	**90.5% ± 4.7%**	**3.6 ± 0.9**	**0.5 ± 0.3**
InnerEar_Lt	55.1% ± 12.8%	8.0 ± 7.4	1.9 ± 1.0	**61.6% ± 14.0%**	**4.9 ± 2.0**	**0.9 ± 0.6**
InnerEar_Rt	54.0% ± 14.5%	9.4 ± 11.2	2.4 ± 2.4	**64.0% ± 13.8%**	**4.7 ± 1.9**	**0.8 ± 0.5**
Lens_Lt	81.1% ± 8.9%	2.1 ± 0.8	0.3 ± 0.2	**83.8% ± 5.9%**	**2.0 ± 0.7**	**0.2 ± 0.2**
Lens_Rt	80.1% ± 9.4%	**2.1 ± 0.8**	**0.3 ± 0.2**	**82.5% ± 7.6%**	2.1 ± 0.8	0.3 ± 0.2
Mandible_Lt	85.3% ± 12.5%	9.4 ± 9.0	1.5 ± 2.7	**88.8% ± 3.5%**	**7.7 ± 7.6**	**1.2 ± 1.0**
Mandible_Rt	85.7% ± 7.2%	9.5 ± 8.2	1.3 ± 1.2	**89.1% ± 3.3%**	**7.8 ± 7.8**	**1.2 ± 1.0**
OpticChiasm	53.0% ± 15.3%	6.6 ± 2.0	1.4 ± 0.7	**69.1% ± 10.9%**	**5.8 ± 2.1**	**0.6 ± 0.4**
OpticNerve_Lt	63.9% ± 13.9%	5.7 ± 5.4	1.0 ± 1.2	**69.0% ± 7.6%**	**4.8 ± 2.4**	**0.6 ± 0.4**
OpticNerve_Rt	64.7% ± 14.7%	5.5 ± 4.6	1.0 ± 1.4	**68.8% ± 8.1%**	**4.6 ± 1.8**	**0.6 ± 0.3**
OralCavity	48.2% ± 6.9%	29.4 ± 7.3	9.0 ± 1.7	**50.9% ± 6.5%**	**28.2 ± 5.0**	**7.5 ± 1.4**
Parotid_Lt	85.0% ± 6.7%	10.6 ± 10.3	1.0 ± 0.9	**87.4% ± 4.3%**	**9.6 ± 10.9**	**0.7 ± 0.5**
Parotid_Rt	83.3% ± 8.0%	12.4 ± 11.7	1.4 ± 2.7	**87.6% ± 4.5%**	**10.4 ± 11.1**	**0.8 ± 0.8**
Pituitary	66.7% ± 15.2%	4.2 ± 1.3	0.9 ± 0.7	**73.2% ± 10.1%**	**3.7 ± 1.0**	**0.5 ± 0.4**
SpinalCord	80.3% ± 11.4%	6.2 ± 5.4	0.8 ± 0.6	**83.0% ± 4.8%**	**4.6 ± 1.3**	**0.7 ± 0.2**
SMG_Lt	70.9% ± 2.1%	5.4 ± 0.8	2.6 ± 0.2	**75.3% ± 0.7%**	**4.9 ± 1.9**	**1.3 ± 0.1**
SMG_Rt	**74.2% ± 2.7%**	**5.9 ± 3.4**	1.3 ± 0.1	73.4% ± 1.7%	6.6 ± 0.7	**1.0 ± 0.0**
TempLobe_Lt	75.6% ± 4.1%	22.5 ± 6.8	2.6 ± 1.1	**78.8% ± 3.1%**	**20.6 ± 5.5**	**2.2 ± 0.9**
TempLobe_Rt	78.4% ± 4.0%	**19.8 ± 5.6**	**2.0 ± 0.9**	**79.2% ± 3.1%**	20.1 ± 6.4	2.1 ± 0.8
Thyroid_Lt	72.8% ± 10.3%	12.8 ± 6.2	1.9 ± 1.3	**74.2% ± 10.6%**	**12.2 ± 6.9**	**1.8 ± 1.4**
Thyroid_Rt	73.7% ± 10.9%	10.0 ± 4.4	1.6 ± 1.1	**75.9% ± 10.0%**	**9.4 ± 4.8**	**1.5 ± 1.1**
TMJ_Lt	68.9% ± 15.6%	10.0 ± 6.5	2.3 ± 1.6	**73.2% ± 11.9%**	**5.1 ± 2.1**	**0.8 ± 0.5**
TMJ_Rt	68.7% ± 15.5%	10.0 ± 7.3	2.5 ± 1.9	**72.4% ± 12.3%**	**5.2 ± 2.1**	**0.9 ± 0.6**
Average	72.4%	9.5	1.8	**76.9%**	**8.1**	**1.3**