Organ at Risk Segmentation for Head and Neck Cancer using Stratified Learning and Neural Architecture Search

Dazhou Guo[1]   Dakai Jin[1]   Zhuotun Zhu[3]   Tsung-Ying Ho[2]   Adam P. Harrison[1]   Chun-Hung Chao[4]
Jing Xiao[5]   Le Lu[1]

[1]PAII Inc.   [2]Chang Gung Memorial Hospital   [3]The Johns Hopkins University   [4]National Tsing Hua University   [5]Ping An Technology

Abstract

Organ at risk (OAR) segmentation is a critical step in radiotherapy of head and neck (H&N) cancer, where inconsistencies across radiation oncologists and prohibitive labor costs motivate automated approaches. However, leading methods using standard fully convolutional network workflows that are challenged when the number of OARs becomes large, e.g. > 40. For such scenarios, insights can be gained from the stratification approaches seen in manual clinical OAR delineation. This is the goal of our work, where we introduce stratified organ at risk segmentation (SOARS), an approach that stratifies OARs into anchor, mid-level, and small & hard (S&H) categories. SOARS stratifies across two dimensions. The first dimension is that distinct processing frameworks are used for each OAR category. In particular, inspired by clinical practices, anchor OARs are used to guide the mid-level and S&H categories. The second dimension is that distinct network architectures are used to manage the significant contrast, size, and anatomy variations between different OARs. We use differentiable neural architecture search (NAS), allowing the network to choose among 2D, 3D or Pseudo-3D convolutions. Extensive 4-fold cross-validation on 142 H&N cancer patients with 42 manually labeled OARs, the most comprehensive OAR dataset to date, demonstrates that both framework- and NAS-stratification significantly improves quantitative performance over the state-of-the-art (from 70.44% to 75.14% in absolute Dice scores). Thus, SOARS provides a powerful and principled means to manage the highly complex segmentation space of OARs.

1. Introduction

Head and neck (H&N) cancer is one of the most common cancers worldwide [17]. High-precision radiation therapy, *e.g.* intensity-modulated radiotherapy, has been widely used
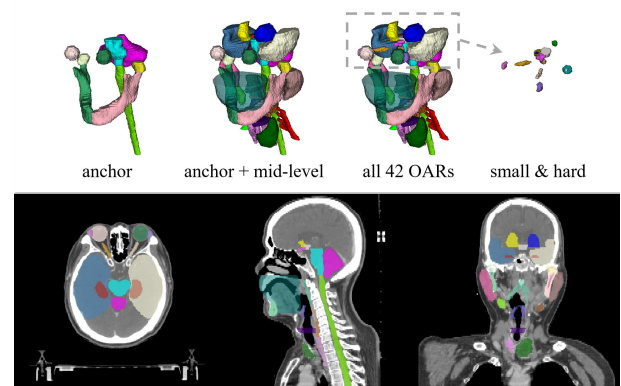

Figure 1. Illustration of 42 OARs in 3D demonstrating their various contrasts, sizes, and shapes in RTCT.

for H&N cancer treatment because of its ability for highly conformal dose delivery. In this process, the radiation dose to normal anatomical structures, *i.e.* organs at risk (OARs), should be controlled to minimize post-treatment complications [11]. This requires accurate delineation of tumors and OARs in radiotherapy computed tomography (RTCT) images [2, 18, 19, 22, 27, 37]. Clinically, OAR segmentation is predominantly carried out manually by radiation oncologists. Manual delineation is not only time consuming, *e.g.* > 2 hrs for 9 OARs, but also suffers from large interpractitioner variability [11]. Unsurprisingly, with more OARs included, time requirements increase significantly, limiting the number of patients who may receive timely radiotherapy [26]. These issues have spurred efforts toward automatic OAR segmentation in H&N cancer [29]. Despite this progress, performance gaps remain, calling for approaches better tailored to this distinct and challenging problem. This is the goal of our work.

By their nature, H&N OARs are 1) complex in anatomical shapes, 2) dense in spatial distributions, 3) large in size variations, and 4) low in RTCT image contrast. Currently, deep convolutional neural networks (CNNs) are a dominant approach[1, 9, 15, 27, 37–39, 41, 46]. How-

ever, existing methods either perform whole volume segmentation [27, 46] or segmentation-by-detection [9, 37]. Yet, model optimization becomes increasingly difficult as greater numbers of OARs need to be segmented. Leveraging insights from clinical practices can help ease the corresponding difficulties.

Within the clinic, radiation oncologists typically refer to easy OARs when delineating harder ones, *e.g.* the eyes, brain stem, and mandible, to serve as anchors to segment hard OARs, such as different types of soft-tissue H&N glands [36]. Figure 1 visually illustrates this stratification. As such, this process suggests that automated solutions could benefit from also stratifying OARs, both to create anchors and to create tailor-made analysis workflows for each stratification. Indeed, Gao *et al*. [9] showed that exploiting two branches for OAR segmentation boosts overall performance. However, large OARs did not serve as support to small OARs in that work. Moreover, the network architecture was manually crafted and fixed across OAR stratifications. Yet, given their highly distinct natures, different OARs likely require different network architectures for optimal performance. It is difficult to see how regular CNNs can meet these needs.

Our work fills this gap by introducing stratified organ at risk segmentation (SOARS), a novel stratified learning framework to segment OARs. SOARS divides OARs into three levels, *i.e.* anchor, mid-level, and small & hard (S&H). Emulating clinical practice, each is processed using tailored workflows. *Anchor OARs* are high in intensity contrast and low in inter- and intra-reader variability. Thus these can be segmented first to provide informative location references to the harder categories. *Mid-level OARs* are low in contrast, but not inordinately small. We provide anchor-level predictions as additional input for mid-level segmentation as guidance and reference-based grounding. *S&H OARs* are very poor in contrast and very small. Similar to mid-level OARs, we use anchor OARs to guide S&H segmentation. However, we use a detection followed by segmentation strategy [9], to better manage the extremely unbalanced class distributions across the entire volume. While this workflow provides specialized frameworks for each OAR category, data processing could be even better tailored, as it is unlikely the same network architecture suits each stratification equally. Thus, we deploy an additional dimension of stratification, using neural architecture search (NAS) to automatically search the optimal architecture for each category. Concretely, we formulate the structure learning as a differentiable NAS [23, 24, 47], allowing for an automatic selection across 2D, 3D or Pseudo-3D (P3D) convolutions with kernel sizes of 3 or 5 at each convolutional block.

Using four-fold cross-validation, we evaluate SOARS on 142 RTCT images with 42 annotated OARs, *the most comprehensive H&N OAR dataset to date*. We demonstrate

that both dimensions of our stratification, *i.e.* category-specific processing and NAS, significantly impact performance. We achieve an average Dice score (DSC) and Hausdorff distance (HD) of 75.14% and 6.98mm, respectively, which corresponds to improvements of 7.51% and 2.41mm, respectively over a non-stratified baseline. Compared to the state-of-the-art, a 3D Mask R-CNN based UaNet method [37], we produce improvements of 4.70% and 2.22mm, in DSC and HD, respectively. Validation on a public dataset (the MICCAI 2015 OAR Challenge [29]), further confirms these compelling performance improvements. In summary, the contributions and novelty of this paper are three folds:

- Segmenting a comprehensive set of OARs is essential and critical for radiotherapy treatment planing in head and neck cancer. We work on the most clinically complete and desirable set of 42 OARs as compared to previous state-of-the-art work.

- Our main methodological contribution is the proposed whole framework on stratifying different organs into different categories of OARs which to be dealt respectively with tailored segmentors (achieved by NAS). Our method is a well-calibrated framework of integrating organ stratification, multi-stage segmentation and NAS in a synergy.

- Our idea of stratifying the 42 OARs into three levels comes from the combination of emulation of oncologists manual OAR contouring knowledge and the OAR's size distributions. To our best knowledge, this simple yet effective organ stratification scheme has not been studied for such a complex segmentation and parsing task like ours, by previous work.

2. Related Works

OAR Segmentation There is a large body of work on OARs segmentation. Atlas-based approaches [10, 16, 33–35, 40] enjoy a prominent history [20, 29]. Their main disadvantage is a reliance on accurate and efficient image registration [46], which is challenged by shape variations, normal tissue removal, abnormal tissue growth, and image acquisition differences [44]. Registration often also take many minutes or even hours to complete. Another common approach is statistical shape or appearance models [5, 6, 32]. These have shown promise, but a prominent issue is that they can be limited to specific shapes described by the statistical model, which makes them less flexible when the number of OARs is large [8]. Of note is Tong *et al*. [39], who applied intensity and texture-based fuzzy models using a hierarchical stratification.

Recently, deep CNN based approaches have proven capable of delivering substantially better performance.
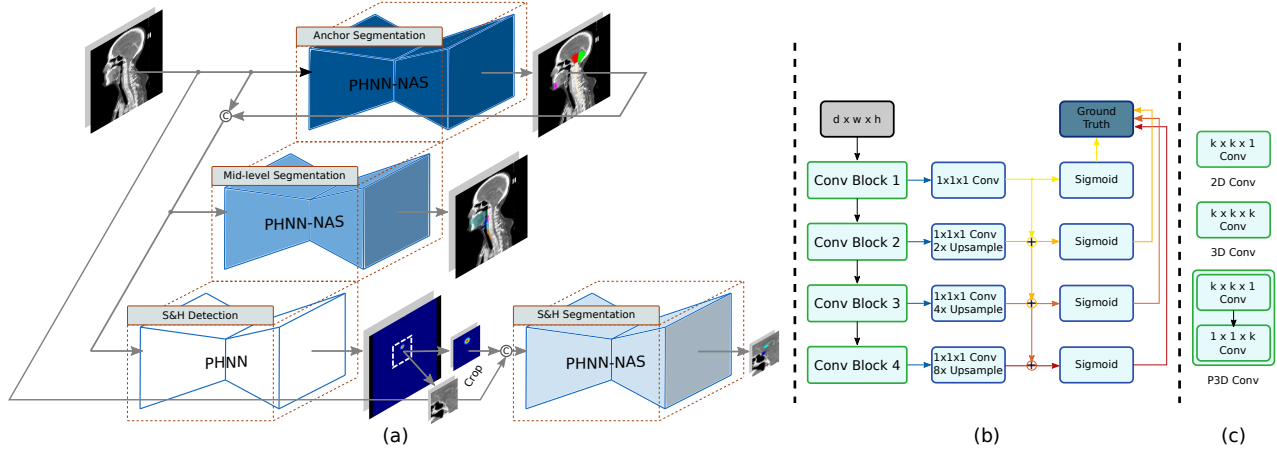
Figure 2. (a) SOARS stratifies OAR segmentation across two dimensions: distinct processing frameworks and distinct architectures. We execute the latter using differentiable NAS. (b) depicts illustrates the backbone network (P-HNN) with NAS, which allows for an automatic selection across 2D, 3D, P3D convolutions. (c) demonstrates the NAS search space setting.

Apart from early efforts [15], fully convolutional networks (FCNs) have quickly becomes the mainstream method [18, 27, 38, 46]. To address data imbalance issues when faced with S&H OARs, FocusNet [9] and UaNet [37] adopt a segmentation-by-detection strategy to achieve better segmentation accuracy. However, both approaches do not stratify OARs, and hence, cannot use easier OARs as support to more difficult ones. Moreover, when the number of OARs is large, *e.g.* > 40, optimization becomes more difficult. Finally, their network architecture remains manually fixed, which is less optimized for the distinct OARs categories.

Stratified Learning Stratification is an effective strategy to decouple a complicated task into easier sub-tasks. Computer vision has a long history using this strategy. Several contextual learning models have been used to assist general object detection [14, 28] within the conditional random field framework [21]. Instance learning, *i.e.* instance localization, segmentation and categorization, often stratifies the problem into multiple sub-tasks [3, 7, 13]. Within medical imaging, stratified statistical learning has also been used to recognize whether a candidate nodule connects to any other major lung anatomies [43]. Yet, the use of stratified learning for semantic segmentation, particularly in the deep-learning era, is still relatively understudied in medical imaging. Within OAR segmentation, Tong *et al.* [39] have applied a hierarchical stratification, but this used a non-deep fuzzy-connectedness model. We are the first to execute stratified learning for deep OAR segmentation.

Neural Architecture Search This is the process of automatically discovering better network architectures. Many NAS methods exploit reinforcement learning [48] or evolutionary algorithms [30]. However, both strategies are extremely computationally demanding. Differentiable NAS [23, 24, 47] realize all candidate architectures simultaneously during optimization, limiting the allowable or feasi-

ble search spaces. Nonetheless, these approaches are highly practical means to tailor architectures. In this work, we follow the differentiable NAS formulation [24, 47] to search the architectures for each of the three OAR stratifications. We explore the optimal kernel size and combination of 2D, 3D, and P3D configurations. As such, we are the first to apply NAS to OAR segmentation.

3. Methods

Figure 2 depicts the SOARS framework, which uses three processing branches to stratify anchor, mid-level and S&H OARs segmentation. A first stratification dimension is distinct processing frameworks. SOARS first segments the anchor OARs. Then, with the help of predicted anchors, mid-level and S&H OARs are segmented. For the most difficult category of S&H, we first detect center locations and then zoom-in to segment the small OARs. The deeply-supervised 3D P-HNN [12] is adopted as the backbone for all three branches, which uses deep supervision to progressively propagate lower-level features to higher-levels ones using a parameter-less pathway. We opt for this backbone due to its good reported performance in other RTCT works [18, 19]. A second dimension of stratification uses differentiable NAS to search distinct P-HNN convolutional blocks for each OAR category.

3.1. Processing Stratification

As mentioned, SOARS segments OARs using three distinct frameworks, where OARs are divided according to clinician recommendations (the details for our 42 OAR dataset is reported in Sec. 4.1). We denote the training data of N data instances as $\mathbb{S} = \left\{ X_n, Y_n^{\mathrm{A}}, Y_n^{\mathrm{M}}, Y_n^{\mathrm{S}} \right\}_{n=1}^{N}$, where X_n, Y_n^{A}, Y_n^{M}, and Y_n^{S} denote the input RTCT and ground-truth masks for anchor, mid-level, and S&H OARs,

respectively. Here, we drop n, when appropriate, for clarity. Throughout, we will abuse matrix/vector notation, using boldface to denote vector-valued volumes and use vector concatenation as an operation across all voxel locations. **Anchor branch:** Assuming we have C classes, SOARS first uses the anchor branch to generate OAR prediction maps for every voxel location, j, and every output class, c:

$$\hat{Y}_c^A(j) = p^A\left(Y^A(j) = c \,|\, X; \mathbf{W}^A\right), \tag{1}$$

$$\hat{\mathbf{Y}}^A = \left[\hat{Y}_1^A \dots \hat{Y}_C^A\right], \tag{2}$$

where $p^A(\cdot)$ and $\hat{\mathbf{Y}}_c^A$ denote the CNN functions and output segmentation maps, respectively. Here, predictions are vector valued 3D masks as they provide a pseudo-probability for every class. $\mathbf{W}^{(\cdot)}$ represents the corresponding CNN parameters.

Anchor OARs have high contrast compared to surrounding tissue or are in easy-to-locate regions; hence, it is relatively easy to segment them directly and robustly based on pure appearance and context features. Consequently, they are ideal candidates to support the segmentation of other OARs.

Mid-level branch: Most mid-level OARs are primarily soft tissue, which have low contrast and can be easily confused with other structures with similar intensities and shapes. Direct segmentation can lead to false-positives or over/under-segmentations. This can be addressed by using processing stratification to directly incorporate anchor predictions into mid-level learning, since the anchor predictions are robust and provide highly informative location and semantically-based cues. As demonstrated in Figure 2, we combine the anchor predictions with the RTCT to create a multi-channel input: $\left[X, \ \hat{\mathbf{Y}}^A\right]$:

$$\hat{Y}_c^M(j) = p^M\left(Y^M(j) = c \,|\, X, \hat{\mathbf{Y}}^A; \mathbf{W}^M\right). \tag{3}$$

In this way, the mid-level branch leverages both the computed tomography (CT) intensities as well as the anchor OAR guidance, which can be particularly helpful in managing regions with otherwise similar CT appearance. Like (2), we can collect mid-level predictions into a vector-valued entity $\hat{\mathbf{Y}}^M$.

Small & hard branch: In this branch, we further decouple segmentation into a detection followed by segmentation process. Directly segmenting the fine boundaries of S&H OARs from CT is very challenging due to the poor contrast and the extremely imbalanced foreground and background distributions when considering the entire volume. In contrast, the detection of center regions of S&H OARs is a much easier problem, since the H&N region has relatively stable anatomical spatial distribution. This means

that the *rough* locations of S&H OARs can be inferred from the CT context with confidence. Once the center location is detected, a localized region can be cropped out to focus on segmenting the fine boundaries in a zoom-in fashion. This has similarities to Gao *et al.*'s [9] approach to segment small organs.

For detecting S&H OAR centers, we adopt a simple yet effective heat map regression method [42, 45], where the heat map labels are created at each organ center using a 3D Gaussian kernel. Similar to the mid-level branch, to increase detection robustness and accuracy we also combine the anchor branch predictions with the RTCT as the detection input channels:

$$\hat{\mathbf{H}} = f(X, \hat{\mathbf{Y}}^A; \mathbf{W}^D) \tag{4}$$

where $\hat{\mathbf{H}}$ denotes the predicted heat maps for every S&H OAR. Like the segmentation networks, we use the same P-HNN backbone for $f(\cdot)$. Given the resulting regressed heat map, we choose the pixel location corresponding to the highest value, and crop a volume of interest (VOI) using three-times the extent of the maximum size of the OAR of interest. With the VOI cropped, SOARS can then segment fine boundaries of the S&H OARs. As illustrated in Figure 2, we concatenate the output from Eq. (4) together with the cropped RTCT image as the input to the S&H OAR segmentation network:

$$\hat{Y}_c^S(j) = p^S\left(Y^S(j) = c \,|\, X, \hat{\mathbf{H}}'; \mathbf{W}^S\right), \tag{5}$$

where here it's understood that (5) is only operating on the cropped region.

3.2. Architectural Stratification

While stratifying OARs into different processing frameworks with distinct inputs and philosophies is key to pushing performance, more can be done. Namely, considering the significant variations in OAR appearance, shape, and size, it is likely that each OAR type would benefit from segmentation branch architectures tailored to their needs. To do this, SOARS automatically searches network architectures for each branch, adding an additional dimension to the stratification. Throughout, we use P-HNN [12] as the base backbone. The whole network structure is illustrated in Figure 2, in which the architecture is learned in a differentiable way [24].

Let $\phi(\cdot; \boldsymbol{\omega}_{x \times y \times z})$ 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 dimensions of the kernel is set to 1, it reduces to a 2D kernel. As shown in Eq. (6), we search a set of possible architectures that include: 2D convolutions, 3D convolutions, or pseudo-3D convolution with

either kernel sizes of 3 or 5:

$$\phi_{2D_3} = \phi\left(\cdot\,;\boldsymbol{\omega}_{3\times3\times1}\right),$$
$$\phi_{2D_5} = \phi\left(\cdot\,;\boldsymbol{\omega}_{5\times5\times1}\right),$$
$$\phi_{3D_3} = \phi\left(\cdot\,;\boldsymbol{\omega}_{3\times3\times3}\right),$$
$$\phi_{3D_5} = \phi\left(\cdot\,;\boldsymbol{\omega}_{5\times5\times5}\right), \quad (6)$$
$$\phi_{P3D_3} = \phi\left(\phi\left(\cdot\,;\boldsymbol{\omega}_{3\times3\times1}\right);\boldsymbol{\omega}_{1\times1\times3}\right),$$
$$\phi_{P3D_5} = \phi\left(\phi\left(\cdot\,;\boldsymbol{\omega}_{5\times5\times1}\right);\boldsymbol{\omega}_{1\times1\times5}\right),$$
$$\boldsymbol{\Phi} = \left\{\phi_{2D_3}, \phi_{2D_5}, \phi_{3D_3}, \phi_{3D_5}, \phi_{P3D_3}, \phi_{P3D_5}\right\},$$

where $\boldsymbol{\Phi}$ denotes the search space of possible architectures. For simplicity, instead of a layer-by-layer architecture search, we use only one type of convolutional kernel to build each P-HNN convolutional block.

Similar to [24, 47], we make the search space continuous by relaxing the categorical choice of a particular operation to a softmax over all 6 possible operations. More formally, if we index each possibility in (6) by k, then we can define a set of 6 learnable logits for each, denoted a_k. A softmax can then be used to aggregate all possible architectures into one combined output, ϕ':

$$\gamma_k = \frac{\exp\left(\alpha_k\right)}{\sum_m \exp\left(\alpha_m\right)} \quad (7)$$

$$\phi' = \sum_k \gamma_k \phi_k, \quad (8)$$

where we have dropped dependence on the input images for convenience. As Zhu *et al*. demonstrated [47], this type of NAS scheme can produce significant gains within medical image segmentation. This creates a sort of super network that comprises all possible manifestations of (6). This super network can be optimized in the same manner as standard networks. At the end of the NAS, the chosen network architecture of each block, $\tilde{\phi}$, can be determined by selecting the ϕ corresponding to the largest α_k value. If the index to this maximum is denoted $\tilde{k}$, then $\tilde{\phi} = \phi_{\tilde{k}}$. If we have b blocks, then based on (8), the searched network can be represented as $\tilde{p}\left(\cdot\,;\tilde{\mathbf{W}}\right) = \tilde{\phi}^b\left(\tilde{\phi}^{b-1}\left(\cdots\tilde{\phi}^1\left(\cdot\,;\tilde{\boldsymbol{\omega}}^1\right);\tilde{\boldsymbol{\omega}}^{b-1}\right);\tilde{\boldsymbol{\omega}}^b\right)$, where $\tilde{(\cdot)}$ denotes the searched network architecture. For consistency, we use the same strategy to search the network architecture for each branch of SOARS.

4. Experiments

4.1. Datasets and Preprocessing

To evaluate performance, we collected 142 anonymized non-contrast RTCT images in H&N cancer patients, where 42 OARs are delineated during the target contouring process for radiotherapy (hereafter denoted as H&N 42 dataset). Extensive 4-fold cross validation, split at the patient level, was conducted on the H&N 42 dataset to report

results. We compare against other state-of-the-art methods including P-HNN [12], UNet [4], and UaNet [37]. To evaluate the effectiveness of SOARS, we conducted two ablation studies using 1 fold of the dataset. Furthermore, we examined our performance using the public MICCAI 2015 head and neck auto-segmentation challenge data[1] (referred hereafter as MICCAI2015). This external testing set contains 9 OARs with 15 test cases. **Evaluation metrics:** We report the segmentation performance using DSC in percentage, HD and ASD in mm. Note that we use HD metric instead of HD95 as reported in some previous works.

H&N 42 OARs dataset: Each CT scan is accompanied by 42 OAR 3D masks annotated by an experienced oncologist. The average CT size is $512\times512\times360$ voxels with an average resolution of $0.95\times0.95\times1.9$ mm. The specific OARs stratification is as follows. **Anchor OARs:** brain stem, cerebellum, eye (left and right), mandible (left and right), spinal cord and temporomandibular joint (left and right). **Mid-level OARs:** brachial plexus (left and right), basal ganglia (left and right), constrictor muscle (inferior, middle and superior), epiglottis, esophagus, hippocampus (left and right), larynx core, oralcavity, parotid (left and right), submandibular gland (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), optic nerve (left and right), optic chiasm, pineal gland, and pituitary.

MICCAI2015 dataset: This dataset has been extensively used by researchers to evaluate atlas and deep learning based H&N OAR segmentation. It contains 33 training cases and 15 test cases with 9 OARs annotated. The 9 OARs include brain stem, mandible, optic chiasm, optic nerve (left and right), parotid (left and right) and submandibular gland (left and right).

Image preprocessing: We apply a windowing of [-500, 1000] HU to every CT scan covering the intensity range of our target OARs, from which we extract $128\times128\times64$ VOIs as training samples for the anchor and mid-level branches as well as the detection module in the S&H branch. The heat map labels in the detection module is a 3D Gaussian distribution with a standard deviation of 8mm. The training VOIs are sampled in two manners: (1) we randomly extract VOIs centered within each of the OARs to ensure sufficient positive samples. (2) we randomly sample additional 15 VOIs from the whole volume to obtain sufficient negative examples. This results in on average 70 VOIs per CT scan. We further augment the training data by applying random scaling between $0.8-1.2$. In testing, 3D sliding windows with sub-volumes of $128\times128\times64$ and strides of $96\times96\times32$ voxels are used. The probability maps of sub-volumes are aggregated to obtain the whole volume

---

[1]http://www.imagenglab.com/wiki/mediawiki/index.php?title=2015_MICCAI_Challenge

<table>
<tr><td colspan="4">Anchor OARs</td><td colspan="4">Mid-level OARs</td><td colspan="4">S&H OARs</td></tr>
</table>

	DSC	HD	ASD		DSC	HD	ASD		DSC	HD	ASD
Baseline	84.02	5.98	0.82	Baseline	63.68	12.97	3.48	Baseline	60.97	4.86	0.98
CT Only	84.14	5.25	0.79	CT Only	67.31	12.03	3.97	CT Only	62.09	4.19	1.06
CT+NAS	85.73	4.77	0.77	CT+Anchor	70.73	10.34	1.67	CT+Heat map	71.75	2.93	0.52
				CT+Anchor+NAS	72.55	9.05	1.31	CT+Heat map +NAS	72.57	2.94	0.49

Table 1. Quantitative results of the ablation studies of the proposed method using 1 fold of the dataset. The baseline network is a 3D P-HNN. For S&H OARs, all methods, except the baseline, segment on *predicted* VOIs. The performance is measured by DSC (unit: %), HD (unit: mm), and ASD (unit: mm).
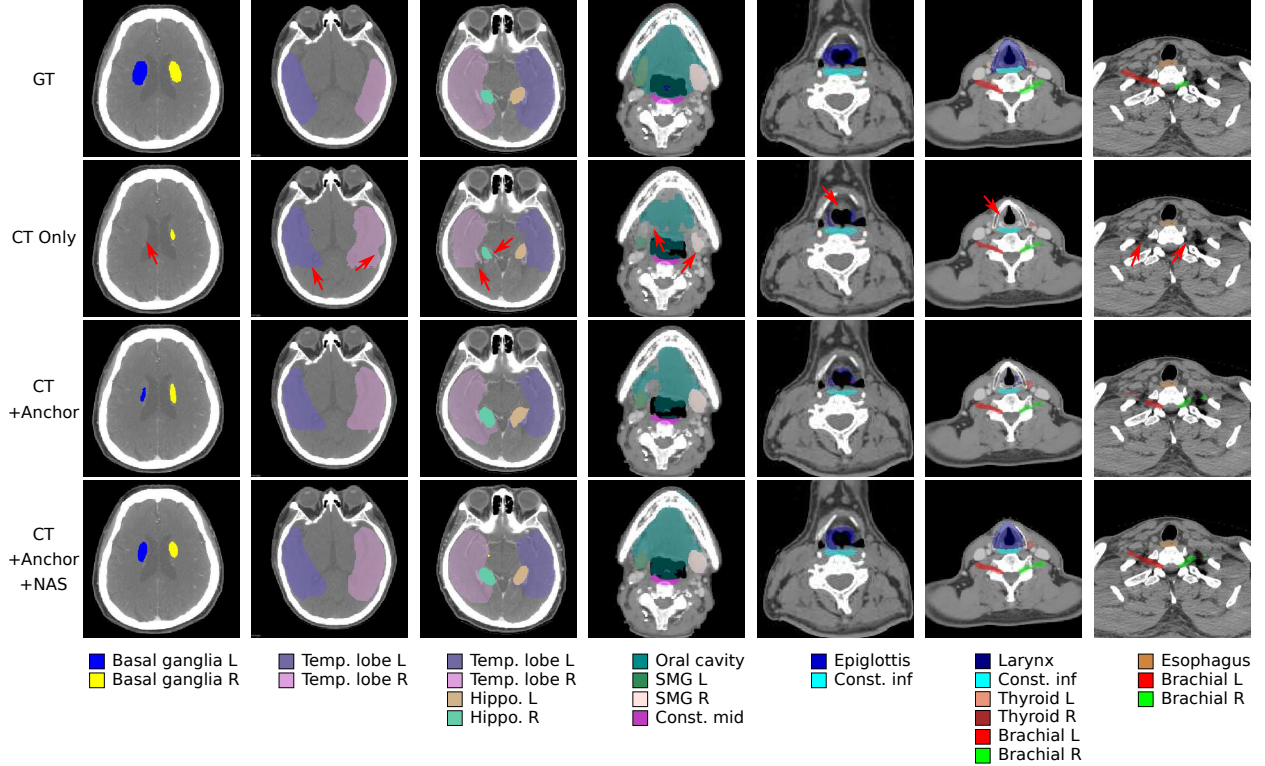

Figure 3. Qualitative mid-level OAR segmentation using different setups. The seven columns are seven representative axial slices in the RTCT image. For better comparison, we use red arrows to indicate the improvements. The 1st row is the RTCT image with OAR delineations of a radiation oncologist. The 3rd row shows the impact of using anchor OARs, which can help the segmentation of soft-tissue mid-level OARs. The 4th demonstrates the impact of NAS, indicating the necessity of adapting network architectures for different OARs.

	Dist (mm)
CT Only	3.25±2.34
CT+Anchor	2.91±1.74

Table 2. S&H OAR detection results measuring the average distance between regressed and true center points.

prediction, taking on average 20s to process one input volume using a single GPU.

4.2. Implementation Details

We implemented SOARS in PyTorch[2], and trained it on an NVIDIA Quadro RTX 8000. The RAdam solver [25] is used to optimize all models with a momentum of 0.9 and a weight decay of 0.005. The DSC loss is used for the segmentation task training. The S&H detection branch is trained using L2 loss with a 0.01 learning rate.

We exploit NAS to search the optimal network architecture for each branch. For the NAS parameter α_k, we first fix α_k for 20 epochs. Then we update α_k and the network weights for an additional 30 epochs. The batch size for NAS training is set to 2. Note that we use only the validation set for α updating. The ratio between the training set and the validation set is 2:1. The initial learning rate is set to 0.005 for the anchor and mid-level branches, and 0.001 for the S&H branch.

After NAS is completed, we retrain the searched network from scratch with a batch size of 12. The batch size is set to be 12. The initial learning rate is set to 0.01 for the anchor and mid-level branches, and 0.005 for the S&H

branch. The detailed training strategy is described as follows: 1) We train the anchor branch for 50 epochs; 2) We fix the parameters of the anchor branch and concatenate its output to the original RTCT, followed by further training the mid-level and S&H branches for 50 epochs; 3) Finally we fine-tune the whole framework in an end-to-end manner for 10 epochs.

4.3. Processing Stratification

We first evaluate the effectiveness of the processing stratification of SOARS. The ablation results for segmenting the anchor, mid-level and S&H OARs are shown in Table 1. The baseline comparison is the 3D P-HNN model trained on all 42 OARs together. When anchor OARs are stratified to train only on themselves, there is a slight improvement as compared to the baseline model, consistent with the observation that anchor OARs generally have good contrast and are easy to optimize. However, when focusing on mid-level OARs, there is a marked DSC score improvement (3.63%) when only training on mid-level OARs instead of training on all. This demonstrates the difficulty in segmenting a large number of organs together without considering their differences. When further adding anchor OAR predictions as support, both DSC scores and the ASD experience large improvements, *i.e.* from 67.31% to 70.73% in DSC and 3.97 to 1.67mm in ASD. These significant error reductions indicate that anchor OARs serve as effective references to better delineate the hard-to-discern boundaries of mid-level organs (most are soft-tissue). Figure 3 depicts qualitative examples of segmenting mid-level OARs. As can be seen, our method achieves much better visual results.

For the S&H branch, we first report the accuracy of the regressed center-point using the detection-by-segmentation network. As Table 2 demonstrates, the center points of S&H OARs can be detected with high robustness. Moreover, when using the anchor OARs as support, the distance errors between regressed and true center points are further reduced. In our experiments, no S&H OAR was missed by our detection-by-segmentation strategy, demonstrating the robustness of our approach. Now focusing on the segmentation results of Table 1, by cropping the VOI using the detection module, there is remarkable improvement in segmenting the S&H OARs, moving DSC from 62.09% to 71.75%, as compared against directly segmenting from the CT. This further demonstrates the value of our processing-based stratification method, which provides for optimal treatment of OAR categories with different characteristics. As the examples of Figure 4 demonstrate, the benefits of processing stratification for S&H OARs is clearly shown in the optic chiasm, hypothalamus, and pineal gland, which are insufficiently segmented/missed when using only RTCT for prediction.
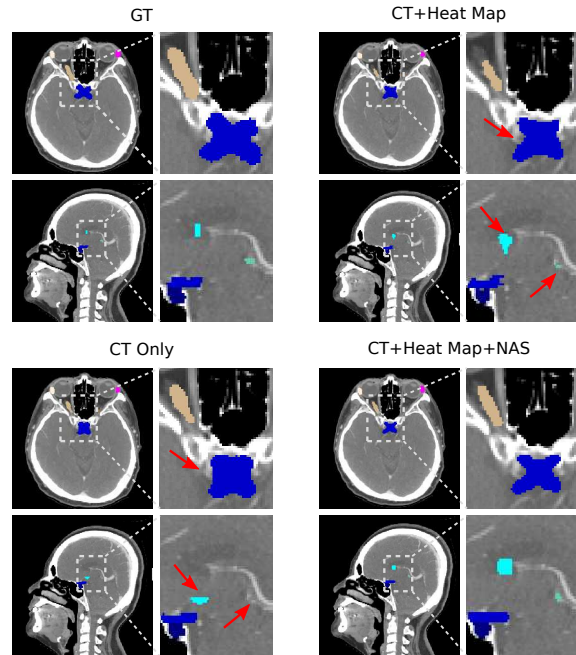

Figure 4. Examples of S&H OAR segmentation using different setups. For visualiztion purpose, the dashed rectangles are enlarged for highlighting improvements. As indicated using the red arrows, the proposed method achieves visually better optic chiasm, hypothalamus, and pineal gland segmentation.

4.4. Architectural Stratification

Table 1 also outlines the performance improvements provided by NAS. As can be seen, all three branches trained with NAS consistently produce more accurate segmentation results than those trained with the baseline 3D P-HNN network. This validates the effectiveness of NAS on complicated segmentation tasks. For the three branches, the anchor and mid-level branches have considerable performance improvement, from 84.14% to 85.73% and 70.73% to 72.55% in DSC scores respectively, while the S&H branch provides a marginal improvement (0.82% in DSC score). For segmenting the S&H OARs, the strong priors of detected heat maps may have already made the segmentation task much easier. Nonetheless, considering the dramatic improvements already provided by the stratified approach in Sec. 4.3, the fact that NAS is able to boost performance even further attests to its benefits. Some qualitative examples demonstrating the effectiveness of NAS are shown in Figure 3 and Figure 4.

The searched network architectures for the anchor branch are 2D-kernel3, 2D-kernel5, 2D-kernel3 and 3D-kernel5 for the four convolution blocks, while for the mid-level branch they are 2D-kernel3, 2.5D-kernel5, 2D-kernel3 and 2.5D-kernel5. This is an interesting result, as it indicates that 3D kernels may not always be the best choice for segmenting objects with reasonable size, as mixed 2D or P3D kernels dominate both branches. Consequently, it is

	Anchor OARs			Mid-level OARs			S & H OARs			All OARs		
	DSC	HD	ASD	DSC	HD	ASD	DSC	HD	ASD	DSC	HD	ASD
UNet [4]	82.97	8.90	1.06	63.61	11.06	1.92	59.64	6.38	1.31	66.62	9.26	1.86
P-HNN [12]	84.26	6.12	1.18	65.19	13.15	2.97	59.42	5.23	0.82	67.62	9.39	2.23
UaNet [37]	84.30	8.89	1.72	69.40	11.57	2.06	61.85	5.28	1.53	70.44	9.20	1.83
SOARS	**85.04**	**5.08**	**0.98**	**72.75**	**10.10**	**1.66**	**71.90**	**2.93**	**0.53**	**75.14**	**6.98**	**1.12**

Table 3. Quantitative results of different approaches on segmenting the 42 H&N OARs using the 4-fold cross validation. Our proposed SOARS achieves the best performance in all metrics (indicated in bold).

	Brain Stem	Mandible	Optic Chiasm	Optic Nerve		Parotid		SMG		All OARs
				Lt	Rt	Lt	Rt	Lt	Rt	
Ren *et al.* [31]	-	-	58.0±17.0	72.0±8.0	70.0±9.0	-	-	-	-	-
Wang *et al.* [41]	**90.0±4.0**	94.0±1.0		-	-	83.0±6.0	83.0±6.0	-	-	-
AnatomyNet [46]	86.7±2.0	92.5±2.0	53.2±15.0	72.1±6.0	70.6±10.0	88.1±2.0	87.3±4.0	81.4±4.0	81.3±4.0	79.2
FocusNet [9]	87.5±2.6	93.5±1.9	59.6±18.1	73.5±9.6	74.4±7.2	86.3±3.6	87.9±3.1	79.8±8.1	80.1±6.1	80.3
UaNet [37]	87.5±2.5	95.0±0.8	61.5±10.2	74.8±7.1	72.3±5.9	**88.7±1.9**	87.5±5.0	82.3±5.2	81.5±4.5	81.2
SOARS	87.6±2.8	**95.1±1.1**	64.9±8.8	**75.3±7.1**	74.6±5.2	88.2±3.2	**88.2±5.2**	**84.2±7.3**	**83.8±6.9**	**82.4**

Table 4. For MICCAI 2015 9 OARs segmentation challenge, the proposed method achieves 7 (in bold) best performance and 2 (in grey box) second best performance.

possible that much computation and memory used for 3D networks could be avoided by using an appropriately designed 2D or P3D architecture. For the S&H branch, the search architecture is 2D-kernel3, 3D-kernel5, 2D-kernel3 and 3D-kernel5 for the four convolution blocks. As can be seen, more 3D kernels are used, consistent with the intuition that small objects with low contrast rely more on the 3D spatial information for better segmentation.

Intuitively, it would be interesting to let the network search the OAR levels. However, NAS becomes computationally unaffordable since automatically stratifying anchor OARs alone is at the complexity of $C_{42}^9 \times$ more expensive.

4.5. Comparison to State-of-the-art

Table 3 compares SOARS against 3 state-of-the-art (SOTA) OAR segmentation methods, *i.e.* UNet [4], P-HNN [12], and UaNet [37], using the 4-fold cross-validation on the H&N 42 OARs dataset. We also tested anatomyNet [46], but it consistently missed very small organs, so we do not report its results. Although P-HNN [12] achieves comparable performance on the anchor and S&H OARs segmentation with UaNet [37], it has decreased performance for mid-level OARs. UaNet is a modified version of 3D Mask R-CNN [13], which conducts object segmentation within the detected boxes. Hence, it decouples the whole complicated task into detection followed by segmentation, possibly accounting for the better segmentation accuracy for the mid-level OARs as compared to P-HNN [12]. Nonetheless, despite being much simpler, P-HNN is still able to match or beat UaNet on the S&H OARs, demonstrating its effectiveness as a baseline and backbone method for SOARS. When considering SOARS, consistent improvements can be observed in all metrics as compared to all competitors, with 4.70% absolute DSC increases and $2.22mm$ HD error reduction as compared to UaNet [37].

4.6. MICCAI2015 Challenge

We use the MICCAI2015 dataset as an external dataset to further demonstrate the generalizability of SOARS. Similar to other comparison methods, we trained our framework from scratch using the MICCAI2015 training set. We get an average DSC of 82.4%, which has 1.2% improvement as compared to [37], or 2.1% over [9]. Compared to competitor methods, we achieve 7 best performance and 2 second best performance on all 9 OARs, especially the most difficult optic chiasm, where we have a 3.4% improvement on DSC as compared to the best previous result achieved by UaNet [37]. These results on the MICCAI2015 dataset further validate the effectiveness and consistency of our method, reinforcing its value.

5. Conclusion

This work presented SOARS, a novel framework that stratifies H&N OARs segmentation into two dimensions. Inspired by clinical practices, we stratify OARs into three categories of anchor, mid-level and S&H, providing customized processing frameworks for each. Importantly, the mid-level and S&H branches build off of the anchor branch's more reliable predictions. Additionally, we stratify network architectures, executing an effective NAS for each. We test on the most comprehensive H&N dataset to date that comprises 42 different OARs. Comparing to SOTA methods, the improvements are most significant for the mid-level and S&H OARs. With this, we demonstrate that our proposed SOARS can outperform all state-of-the-art baseline networks, including the most recent representative work UaNet [37], by margins as high as 4.70% in DSC. Thus, our work represents an important step forward toward reliable and automated H&N OAR segmentation.

References

[1] M Agn, P M af Rosenschöld, O Puonti, M J Lundemann, L Mancini, A Papadaki, S Thust, J Ashburner, I Law, and K Van Leemput. A modality-adaptive method for segmenting brain tumors and organs-at-risk in radiation therapy planning. *Medical Image Analysis*, 54:220–237, 2019. 1

[2] C E Cardenas, B M Anderson, M Aristophanous, J Yang, D J Rhee, R E McCarroll, A SR Mohamed, M Kamal, B A Elgohari, H M Elhalawani, et al. Auto-delineation of oropharyngeal clinical target volumes using 3d convolutional neural networks. *Physics in Medicine & Biology*, 63(21):215026, 2018. 1

[3] K Chen, J Pang, J Wang, Y Xiong, X Li, S Sun, W Feng, Z Liu, J Shi, W Ouyang, et al. Hybrid task cascade for instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4974–4983, 2019. 3

[4] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention*, pages 424–432. Springer, 2016. 5, 8

[5] T F Cootes, G J Edwards, and Christopher J T. Active appearance models. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (6):681–685, 2001. 2

[6] T F Cootes, C J Taylor, D H Cooper, and J Graham. Active shape models-their training and application. *Computer vision and image understanding*, 61(1):38–59, 1995. 2

[7] J Dai, K He, and J Sun. Instance-aware semantic segmentation via multi-task network cascades. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3150–3158, 2016. 3

[8] K D Fritscher, M Peroni, P Zaffino, M F Spadea, R Schubert, and G Sharp. Automatic segmentation of head and neck ct images for radiotherapy treatment planning using multiple atlases, statistical appearance models, and geodesic active contours. *Medical physics*, 41(5):051910, 2014. 2

[9] Y Gao, R Huang, M Chen, Z Wang, J Deng, Y Chen, Y Yang, J Zhang, C Tao, and H Li. Focusnet: Imbalanced large and small organ segmentation with an end-to-end deep neural network for head and neck ct images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 829–838. Springer, 2019. 1, 2, 3, 4, 8

[10] X Han, M S Hoogeman, P C Levendag, L S Hibbard, D N Teguh, P Voet, A C Cowen, and T K Wolf. Atlas-based auto-segmentation of head and neck ct images. In *International Conference on Medical Image Computing and Computer-assisted Intervention*, pages 434–441. Springer, 2008. 2

[11] P M Harari, S Song, and W A Tomé. Emphasizing conformal avoidance versus target definition for imrt planning in head-and-neck cancer. *International Journal of Radiation Oncology* Biology* Physics*, 77(3):950–958, 2010. 1

[12] A P Harrison, Z Xu, K George, and et al. Progressive and multi-path holistically nested neural networks for pathological lung segmentation from ct images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 621–629. Springer, 2017. 3, 4, 5, 8

[13] K He, G Gkioxari, P Dollár, and R Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 3, 8

[14] G Heitz and D Koller. Learning spatial context: Using stuff to find things. In *European conference on computer vision*, pages 30–43. Springer, 2008. 3

[15] B Ibragimov and L Xing. Segmentation of organs-at-risks in head and neck ct images using convolutional neural networks. *Medical physics*, 44(2):547–557, 2017. 1, 3

[16] A Isambert, F Dhermain, F Bidault, O Commowick, P Bondiau, G Malandain, and D Lefkopoulos. 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(1):93–99, 2008. 2

[17] A Jemal, F Bray, M M Center, J Ferlay, E Ward, and D Forman. Global cancer statistics. *CA: a cancer journal for clinicians*, 61(2):69–90, 2011. 1

[18] D Jin, D Guo, T Ho, A P Harrison, J Xiao, C Tseng, and L Lu. Accurate esophageal gross tumor volume segmentation in pet/ct using two-stream chained 3d deep network fusion. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 182–191. Springer, 2019. 1, 3

[19] D Jin, D Guo, T Ho, Adam P Harrison, J Xiao, C Tseng, and L Lu. Deep esophageal clinical target volume delineation using encoded 3d spatial context of tumors, lymph nodes, and organs at risk. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 603–612. Springer, 2019. 1, 3

[20] M Kosmin, J Ledsam, B Romera-Paredes, R Mendes, S Moinuddin, D de Souza, L Gunn, C Kelly, CO Hughes, A Karthikesalingam, et al. Rapid advances in auto-segmentation of organs at risk and target volumes in head and neck cancer. *Radiotherapy and Oncology*, 135:130–140, 2019. 2

[21] J Lafferty, A McCallum, and F CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001. 3

[22] L Lin, Q Dou, Y Jin, G Zhou, Y Tang, W Chen, B Su, F Liu, C Tao, N Jiang, et al. Deep learning for automated contouring of primary tumor volumes by mri for nasopharyngeal carcinoma. *Radiology*, 291(3):677–686, 2019. 1

[23] C Liu, L Chen, F Schroff, H Adam, W Hua, A L Yuille, and F Li. Auto-deeplab: Hierarchical neural architecture search for semantic image segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 82–92, 2019. 2, 3

[24] H Liu, K Simonyan, and Y Yang. Darts: Differentiable architecture search. *arXiv preprint arXiv:1806.09055*, 2018. 2, 3, 4, 5

[25] L Liu, H Jiang, P He, W Chen, X Liu, J Gao, and J Han. On the variance of the adaptive learning rate and beyond. *arXiv preprint arXiv:1908.03265*, 2019. 6

[26] J S Mikeljevic, R Haward, C Johnston, A Crellin, D Dodwell, A Jones, P Pisani, and D Forman. Trends in postoperative radiotherapy delay and the effect on survival in breast cancer patients treated with conservation surgery. *British*

journal of cancer, 90(7):1343, 2004. 1

[27] S Nikolov, S Blackwell, R Mendes, J De Fauw, C Meyer, C Hughes, H Askham, B Romera-Paredes, A Karthikesalingam, C Chu, et al. Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy. *arXiv preprint arXiv:1809.04430*, 2018. 1, 2, 3

[28] A Rabinovich, A Vedaldi, C Galleguillos, E Wiewiora, and S J Belongie. Objects in context. In *ICCV*, volume 1, page 5. Citeseer, 2007. 3

[29] P F Raudaschl, P Zaffino, G C Sharp, M F Spadea, A Chen, B M Dawant, T Albrecht, T Gass, C Langguth, M Lüthi, et al. Evaluation of segmentation methods on head and neck ct: auto-segmentation challenge 2015. *Medical physics*, 44(5):2020–2036, 2017. 1, 2

[30] E Real, A Aggarwal, Y Huang, and Q V Le. Regularized evolution for image classifier architecture search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4780–4789, 2019. 3

[31] X Ren, L Xiang, D Nie, Y Shao, H Zhang, D Shen, and Q Wang. Interleaved 3d-cnn s for joint segmentation of small-volume structures in head and neck ct images. *Medical physics*, 45(5):2063–2075, 2018. 8

[32] D Rueckert, A F Frangi, and J A Schnabel. Automatic construction of 3-d statistical deformation models of the brain using nonrigid registration. *IEEE transactions on medical imaging*, 22(8):1014–1025, 2003. 2

[33] A Saito, S Nawano, and A Shimizu. 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. 2

[34] E Schreibmann, D M Marcus, and T Fox. Multiatlas segmentation of thoracic and abdominal anatomy with level set-based local search. *Journal of applied clinical medical physics*, 15(4):22–38, 2014.

[35] R Sims, A Isambert, V Grégoire, F Bidault, L Fresco, J Sage, J Mills, J Bourhis, D Lefkopoulos, O Commowick, et al. A pre-clinical assessment of an atlas-based automatic segmentation tool for the head and neck. *Radiotherapy and Oncology*, 93(3):474–478, 2009. 2

[36] D A Tamboli, M A Harris, J P Hogg, T Realini, and J A Sivak-Callcott. Computed tomography dimensions of the lacrimal gland in normal caucasian orbits. *Ophthalmic Plastic & Reconstructive Surgery*, 27(6):453–456, 2011. 2

[37] H Tang, X Chen, Y Liu, Z Lu, J You, M Yang, S Yao, G Zhao, Y Xu, T Chen, et al. Clinically applicable deep learning framework for organs at risk delineation in ct images. *Nature Machine Intelligence*, pages 1–12, 2019. 1, 2, 3, 5, 8

[38] N Tong, S Gou, S Yang, D Ruan, and K Sheng. Fully automatic multi-organ segmentation for head and neck cancer radiotherapy using shape representation model constrained fully convolutional neural networks. *Medical physics*, 45(10):4558–4567, 2018. 3

[39] Y Tong, J K Udupa, X Wu, D Odhner, G Pednekar, C B Simone, D McLaughlin, C Apinorasethkul, G Shammo, P James, et al. Hierarchical model-based object localization for auto-contouring in head and neck radiation therapy planning. In *Medical Imaging 2018: Biomedical Applications in Molecular, Structural, and Functional Imaging*, volume 10578, page 1057822. International Society for Optics and Photonics, 2018. 1, 2, 3

[40] P WJ Voet, M LP Dirkx, D N Teguh, M S Hoogeman, P C Levendag, and B JM Heijmen. Does atlas-based autosegmentation of neck levels require subsequent manual contour editing to avoid risk of severe target underdosage? a dosimetric analysis. *Radiotherapy and Oncology*, 98(3):373–377, 2011. 2

[41] 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 Transactions on Image Processing*, 27(2):923–937, 2017. 1, 8

[42] S Wei, V Ramakrishna, T Kanade, and Y Sheikh. Convolutional pose machines. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4724–4732, 2016. 4

[43] D Wu, L Lu, J Bi, Y Shinagawa, K Boyer, A Krishnan, and M Salganicoff. Stratified learning of local anatomical context for lung nodules in ct images. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2791–2798. IEEE, 2010. 3

[44] X Wu, J K Udupa, Y Tong, D Odhner, G V Pednekar, C B Simone II, D McLaughlin, C Apinorasethkul, O Apinorasethkul, J Lukens, et al. Aar-rt–a system for auto-contouring organs at risk on ct images for radiation therapy planning: Principles, design, and large-scale evaluation on head-and-neck and thoracic cancer cases. *Medical image analysis*, 54:45–62, 2019. 2

[45] Z Xu, Y Huo, J Park, B Landman, A Milkowski, S Grbic, and S Zhou. Less is more: Simultaneous view classification and landmark detection for abdominal ultrasound images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 711–719. Springer, 2018. 4

[46] W Zhu, Y Huang, L Zeng, X Chen, Y Liu, Z Qian, N Du, W Fan, and X Xie. Anatomynet: Deep learning for fast and fully automated whole-volume segmentation of head and neck anatomy. *Medical physics*, 46(2):576–589, 2019. 1, 2, 3, 8

[47] Z Zhu, C Liu, D Yang, A Yuille, and D Xu. V-nas: Neural architecture search for volumetric medical image segmentation. *arXiv preprint arXiv:1906.02817*, 2019. 2, 3, 5

[48] B Zoph, V Vasudevan, J Shlens, and Q V Le. Learning transferable architectures for scalable image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8697–8710, 2018. 3