RPE | Radiation Physics & Engineering
A Peer-Reviewed Journal
Published by K.N. Toosi University of Technology
http://rpe.kntu.ac.ir
Research Article

Radiation Physics and Engineering 2026; ?(?):?–?

Improving head and neck organs at risk segmentation in CT using residual U-Net with slice-based preprocessing

Khashayar Heshmati Jannat Magham[a], Laleh Rafat-Motavalli[a,*], Hashem Miri-Hakimabad[a], Mahdieh Dayyani[b,c]

[a] *Department of Physics, Faculty of Science, Ferdowsi University of Mashhad, Mashhad, Iran*
[b] *Research and Education Department, Reza Radiotherapy and Oncology Center, Mashhad, Iran*
[c] *Radiation Oncology Department, Reza Radiotherapy and Oncology Center, Mashhad, Iran*

HIGHLIGHTS

-
-
-
-
-

ABSTRACT

Background. Accurate delineation of organs at risk (OARs) in head and neck CT images is essential for safe and effective radiotherapy. While U-Net-based deep learning models perform well, segmenting small, complex OARs is still difficult due to low contrast and class imbalance. Purpose. This study investigates the impact of a slice-based (cropping) preprocessing strategy on the segmentation accuracy, consistency, and computational efficiency of a Residual U-Net framework for head and neck OAR delineation. Methods. A total of 63 CT scans covering 41 OARs were used. Networks were trained separately for each organ with full images and organ-specific crops. Segmentation accuracy was assessed using Dice Similarity Coefficient (DSC) and Intersection over Union (IoU). Paired Wilcoxon tests compared full-size and slice-based models. Additional experiments with dropout and extended training targeted challenging structures like the optic chiasm and optic nerves. Results. Slice-based preprocessing improved IoU by 4.1% and Dice score by 3.2% across 41 OARs ($p < 0.001$). For 11 small, complex organs, gains were 10.9% in IoU and 9.0% in Dice ($p < 0.001$). It also reduced performance variability, indicating better consistency. Training time dropped to less than half, about 2.3 times faster, while inference speed increased eightfold. Dropout and extended training slightly improved optic pathway metrics, but not significantly. Conclusions. The proposed slice-based Residual U-Net framework improves segmentation accuracy and computational efficiency for head and neck OAR delineation. The approach is particularly beneficial for small and anatomically complex structures and may provide a practical solution for integration into radiotherapy planning workflows.

KEYWORDS

Auto-Segmentation
Residual U-Net
Head and Neck CT
Organs at Risk

HISTORY

Received:
Revised:
Accepted:
Published:

1 Introduction

Head and neck cancers, including malignancies of the salivary glands, nasopharynx, and larynx, represent a major group of cancers worldwide. Modern radiotherapy techniques such as intensity-modulated radiation therapy (IMRT), volumetric-modulated arc therapy (VMAT), and proton therapy allow precise dose delivery to tumors while sparing surrounding normal tissue. However, accurate delineation of organs at risk (OARs) remains essential to minimize radiation-induced complications (Ibragimov and Xing, 2017).

The efficacy and safety of radiotherapy are critically dependent on the accurate delineation of both the target

*Corresponding author: rafat@um.ac.ir

https://doi.org/

volume and the Organs at Risk. OARs in the head and neck region are numerous, anatomically complex, and often in close proximity to the target, making their precise segmentation a crucial yet challenging step in the treatment planning process. Traditional manual contouring of OARs is not only labor-intensive and time-consuming but also subject to significant inter- and intra-observer variability, which can lead to inconsistencies in treatment plans and clinical outcomes (Harari et al., 2010; Sharp et al., 2014). This has driven the pursuit of automated segmentation methods to improve the efficiency, consistency, and accuracy of radiotherapy planning.

While Computed Tomography (CT) is the primary imaging modality for radiotherapy planning due to its geometric accuracy and electron density information for dose calculation, it presents specific challenges for automated segmentation. The inherently low soft-tissue contrast of CT can make it difficult to distinguish between adjacent OARs with similar attenuation, such as muscles and glands. Furthermore, the presence of metal artifacts from dental fillings or implants can obscure anatomical boundaries and degrade image quality, further complicating the segmentation task (Ibragimov and Xing, 2017).

Traditional automated segmentation approaches, including thresholding (Otsu et al., 1979), edge detection (Kass et al., 1988), region-growing (Adams and Bischof, 1994), and atlas-based registration (Han et al., 2008), often lacked the robustness and accuracy required for clinical adoption. These methods were largely based on handcrafted features and heuristic rules, meaning their performance was heavily dependent on pre-defined parameters and assumptions about the image data. They struggled with the high degree of anatomical variability across patients and were frequently unable to handle pathological alterations or image artifacts effectively.

This field was transformed by the shift from these rule-based systems to machine learning (ML). In a typical ML approach for segmentation, an algorithm is trained on a dataset of images and corresponding labels (masks). Instead of relying on fixed rules, the algorithm learns to identify patterns and relationships between image pixels and the desired output. Classical ML models, such as random forests (Breiman, 2001) or support vector machines (Cortes and Vapnik, 1995), often still required a feature extraction step where domain experts would identify relevant characteristics (e.g., texture, intensity, shape) for the model to learn from. While a significant improvement, the performance of these models was ultimately bounded by the quality and completeness of the human-designed features.

This limitation was overcome by the rise of deep learning (DL), a subfield of machine learning that uses artificial neural networks with many layers. The revolutionary power of deep learning lies in its ability to perform end-to-end learning; the model automatically learns the most relevant features directly from the raw input data, simultaneously with learning how to use those features for the task at hand, such as segmentation. This eliminates the need for manual feature engineering and allows the model to discover complex, hierarchical patterns that may be imperceptible to a human expert.

The most successful deep learning architecture for image analysis is the Convolutional Neural Network (CNN). CNNs are specifically designed to process pixel data by using mathematical operations called convolutions. These operations apply a set of learnable filters across the image, allowing the network to build a hierarchical representation of features.

The U-Net architecture (Ronneberger et al., 2015), a seminal CNN variant, has become the dominant paradigm for biomedical image segmentation. Its symmetric encoder-decoder structure is uniquely suited for this task. The encoder (or contracting path) progressively down-samples the image, extracting and condensing its features into a compact representation. The decoder (or expanding path) then up-samples this representation to reconstruct a full-resolution segmentation map. The key innovation of U-Net is its skip connections, which bypass the bottleneck of the network by forwarding high-resolution feature maps from the encoder to the corresponding layers in the decoder. This allows the model to combine the coarse, semantic information from the deeper layers with the fine, spatial details from the earlier layers, enabling precise localization of organ boundaries.

A natural evolution from 2D to 3D CNNs promised to leverage valuable volumetric context from medical scans. However, 3D models like the 3D U-Net (Çiçek et al., 2016) are notoriously memory-intensive and computationally expensive, as they process entire 3D volumes. They often require significant hardware resources or aggressive image down-sampling, which can sacrifice the high-resolution details necessary for segmenting small, intricate OARs.

While 2D U-Nets mitigate these computational demands and are highly effective for slice-wise analysis, they process entire CT slices, which often contain a substantial amount of irrelevant background information. This extraneous data not only increases the computational load but may also introduce noise and distract the model from learning the salient features of small, low-contrast OARs, potentially hindering peak performance. Our study builds upon the 2D U-Net by integrating Residual Networks (ResNet) (He et al., 2016) as the encoder backbone. ResNets address the vanishing gradient problem in very deep networks through skip connections that learn residual functions, enabling the training of more powerful and accurate models without degradation in performance.

Some studies have implemented strategies analogous to the slice-based approach discussed in this work, primarily targeting computational efficiency in medical image segmentation. Notable examples include are Liang et al. (Liang et al., 2019), Van Rooij et al. (Van Rooij et al., 2019) and Chan et al. (Chan et al., 2019). Liang et al. delineated OARs in head and neck CT images. they utilized a detect-then-segment approach, where the target organ was first localized, and segmentation was subsequently performed only within that detected region. To reduce memory usage during the training of their models for head and neck CT delineation, Van Rooij et al. and Chan et al. initially cropped the images, thereby feeding smaller inputs into the network.

Table 1: Number of slices for each organ at risk.

#	OAR	Public Dataset	Institute Dataset	Total
1	Left Carotid	4049	0	4049
2	Right Carotid	3638	0	3638
3	Arytenoid	166	0	166
4	Bone Mandible	1535	420	1955
5	Left Brachial Plexus	0	452	452
6	Right Brachial Plexus	0	468	468
7	Brain	0	463	463
8	Brainstem	999	306	1305
9	Buccal Mucosa	591	0	591
10	Cavity Oral	1139	307	1446
11	Left Cochlea	116	36	152
12	Right Cochlea	137	14	151
13	Cricopharyngeus	385	0	385
14	Esophagus	332	844	1176
15	Left Eye	0	102	102
16	Left Eye Posterior	521	0	521
17	Right Eye Posterior	516	0	516
18	Right Eye	0	95	95
19	Left Gland Lacrimal	280	73	353
20	Right Gland Lacrimal	281	68	349
21	Left Submandibular Gland	707	182	889
22	Right Submandibular Gland	714	173	887
23	Thyroid Gland	983	106	1089
24	Glottis	351	0	351
25	Larynx	680	199	879
26	Left Lens	259	36	295
27	Right Lens	275	32	307
28	Lips	827	109	936
29	Left Lung	0	435	435
30	Right Lung	0	430	430
31	Lungs	0	197	197
32	Muscle Constrictor	0	416	416
33	Optic Chiasm	94	33	127
34	Left Optic Nerve	201	41	242
35	Right Optic Nerve	205	41	246
36	Left Parotid	1119	303	1422
37	Right Parotid	1130	319	1449
38	Pituitary	140	30	170
39	Spinal Canal	0	567	567
40	Spinal Cord	2875	71	2946
41	Trachea	0	424	424

Most of the previous studies have employed cropping or region-of-interest (ROI) localization to manage computational demands (Chan et al., 2019; Van Rooij et al., 2019). However, a comprehensive investigation into the quantitative impact of systematic, organ-specific slice determination and quantification across 41 organs-at-risk (OARs) on both segmentation accuracy and computational efficiency, using a modern Residual U-Net architecture, remains underexplored.

This study therefore evaluates a Residual U-Net architecture for the automatic segmentation of 41 OARs in head and neck CT images and systematically investigates how a slice-based preprocessing strategy influences not only segmentation accuracy and consistency but also computational performance during training and inference.

2 Materials and Methods

This study utilized a NVIDIA Tesla V100S GPU (32 GB VRAM) with 32 GB RAM, accessed via a virtual machine, for training the deep learning models. The implementation relied on Python 3.10, TensorFlow 2.11, and the segmentation-models 1.0 package (Iakubovskii, 2019) to construct and evaluate the networks.

The dataset provided by Podobnik et al. (Podobnik et al., 2023), (hereafter referred to as the public dataset) includes 42 CT images, each accompanied by 30 masks for different OARs, provided in NRRD format. These 42 CT images and their corresponding masks were stored slice-by-slice for use in a 2D U-Net, resulting in 7,581 slices.

Additionally, 21 CT images from Reza Radiotherapy and Oncology Center (RROC) in DICOM format, (hereafter referred to as the institute dataset) contoured by specialists, were used to increase the quantity and diversity of the training data. The contours from RTSTRUCTUREs were converted to binary masks for network training. These 21 CT images consist of 2,495 slices with spacing and thickness (0.9766, 0.9766) and 3.0 mm (CT device model: Siemens SOMATOM Definition AS). In total, 63 CT images covering 41 OARs in the head and neck region from the two aforementioned datasets were used for training. For training the network for each specific organ (one network for each organ), only the slices containing masks for that organ were used from the total of 10,076 slices. The number of slices containing masks for OARs varies depending on the organ.

The Segmentation Models package includes several widely used convolutional neural network architectures, such as VGG, MobileNet, DenseNet, ResNet, ResNeXt, and SEResNet. Each of these networks, due to its unique depth, connectivity patterns, and computational complexity, can be suited for specific applications or purposes in computer vision in general.

VGG-based architectures are characterized by their simple and uniform design, relying on sequential stacks of convolutional layers with small receptive fields. While effective for feature extraction, their lack of skip connections can limit training efficiency in deeper networks and increase susceptibility to vanishing gradients (Simonyan and Zisserman, 2014). MobileNet architectures, in contrast, utilize depthwise separable convolutions to significantly reduce computational cost and memory usage, making them suitable for resource-constrained environments, albeit sometimes at the expense of representational capacity for complex anatomical structures (Howard et al., 2017). DenseNet models introduce dense connectivity between layers, encouraging feature reuse and improved gradient flow, but at the cost of increased memory consumption due to feature map concatenation (Huang et al., 2017).

In this study, ResNet, ResNeXt, and SEResNet architectures were selected as encoder backbones for the U-Net framework. These residual networks employ identity-based skip connections that enable the network to learn residual functions, thereby stabilizing the training of deep models and mitigating the vanishing gradient problem (He et al., 2016). ResNeXt extends this concept through

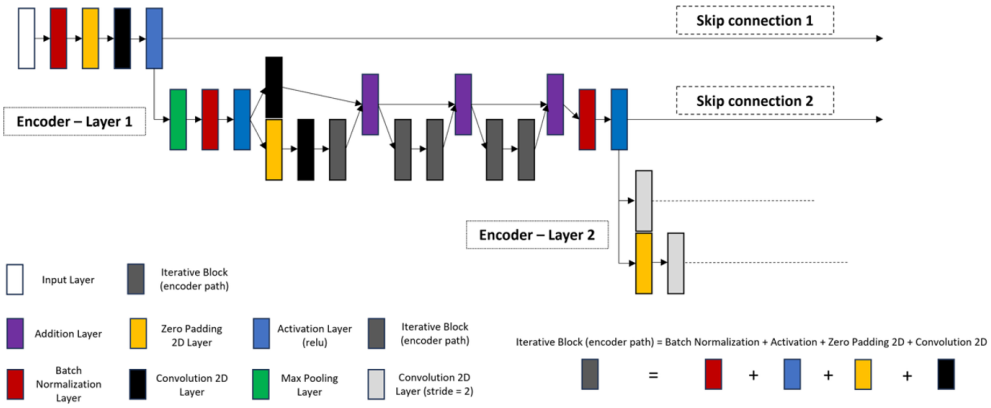


Figure 1: A schematic view of the first encoder layer of the U-Net network with ResNet34 backbone. In convolutional layers that are not preceded by a zero-padding layer, same padding is applied to prevent size-related issues in the network.

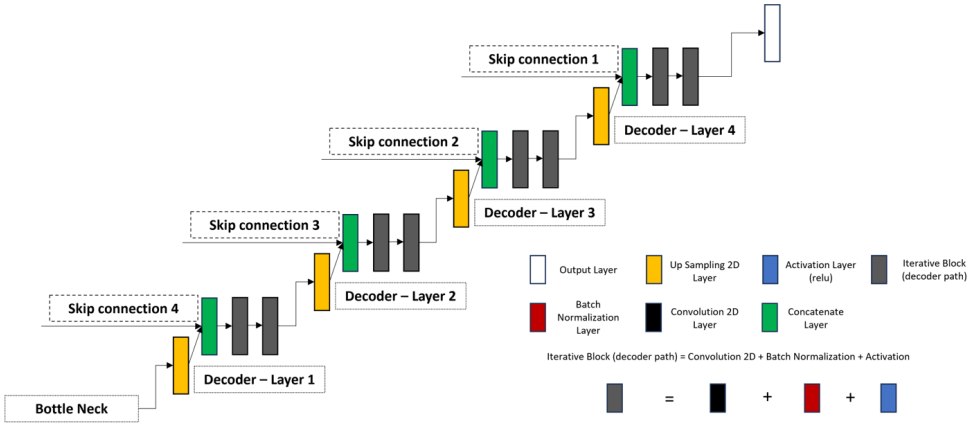


Figure 2: A schematic view of the decoder pathway of the U-Net network.

grouped convolutions, allowing more expressive feature representations without a proportional increase in parameters, while SEResNet incorporates channel-wise attention mechanisms to adaptively recalibrate feature responses. The robustness and training stability of residual-based backbones make them particularly well suited for the segmentation of anatomically complex and low-contrast organs at risk in head and neck CT images (Hu et al., 2018; Xie et al., 2017).

The selected CNN backbones were integrated as encoder networks within a U-Net architecture. All encoder weights were trained from scratch, and no pretrained weights were employed. Feature maps extracted at predefined encoder stages were connected to the corresponding decoder layers through skip connections, following the standard U-Net design to preserve spatial information across scales (Figs. 1 and 2) (Iakubovskii, 2019).

Segmentation accuracy was quantified using, the Dice Similarity Coefficient (Sudre et al., 2017) and Intersection over Union (Rahman and Wang, 2016).

1. Dice Similarity Coefficient:

$$\text{Dice Score} = \frac{2 \times |X \cap Y|}{|X| + |Y|} \tag{1}$$

2. Intersection over Union:

$$\text{IOU} = \frac{|X \cap Y|}{|X \cup Y|} \tag{2}$$

where X represents the predicted mask and Y represents the ground truth mask. The intersection and union of the two masks are calculated based on the number of pixels. While high-contrast structures in CT images, such as the mandible, are readily distinguishable from surrounding tissues, segmenting low-contrast soft tissues like the optic chiasm remains a significant challenge. To address this, we implemented a "slice-based preprocessing" strategy designed to focus the network's attention on the target Volume of Interest while minimizing extraneous data and computational workload.

In this pipeline, organ-specific 2D regions of interest are extracted from CT slices by cropping the image around the centroid of the ground-truth mask. Using the NumPy library, we determined the pixel coordinate range for each OAR across the dataset. Based on the maximum anatomical diameter of each organ, we assigned standardized patch sizes of 64×64, 128×128, or 256×256 pixels. These sizes ensure that the entire organ, along with a narrow margin of surrounding anatomical context, is captured while excluding unnecessary background voxels.

By archiving the specific slicing parameters (coordi-

nates and patch sizes), the predicted masks can be precisely mapped back into the original volumetric image space after inference (see Fig. 12). This approach not only minimizes computational overhead but also improves segmentation by saving the model from train on unnecessary background pixels.

Statistical significance of performance differences between slice-based and full-size networks was assessed using the paired Wilcoxon signed-rank test (Wilcoxon, 1945). This non-parametric test was chosen due to the paired nature of the data and the absence of normality assumptions for Dice and IoU distributions across organs. Statistical analyses were performed both across all 41 organs at risk and, separately, for a subset of organs exhibiting larger performance gains.

All the 41 OARs and their slice(s) size are

- 64×64: Arytenoid, Left Carotid, Right Carotid, and Cricopharyngeus.
- 128×128: Left Carotid, Right Carotid, Arytenoid, Cricopharyngeus, Glottis, Buccal Mucosa, Left Cochlea, Right Cochlea, Esophagus, Larynx, Lips, Optic Chiasm, Left Optic Nerve, Right Optic Nerve, Pituitary, Trachea, Left Eye, Right Eye, Left Eye (Posterior), Right Eye (Posterior), Left Lens, Right Lens, Left Lacrimal Gland, Right Lacrimal Gland, Left Submandibular Gland, Right Submandibular Gland, Thyroid Gland, and Constrictor Muscle.
- 256×256: Mandible, Brain, Brainstem, Oral Cavity, Left Lung, Right Lung, Lungs, Spinal Canal, Spinal Cord, Left Parotid, Right Parotid, Left Brachial Plexus, and Right Brachial Plexus.

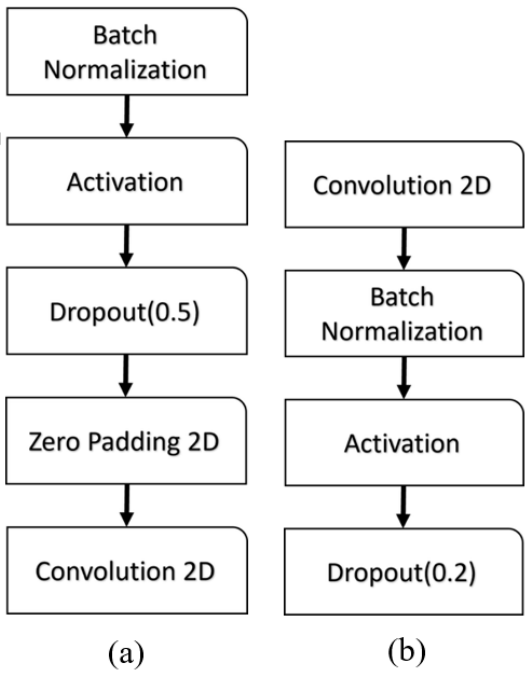


Figure 3: (a) Iterative block in the encoder path of modified ResNet34 (Dropout addition). (b) Iterative block in the decoder path of modified ResNet34 (Dropout addition).

Although the primary objective of this study is to investigate the differences in the network's performance when reducing the dimensions of input images, additional efforts were made to improve the networks performance for three organs at risk, the optic chiasm and optic nerves, which exhibited the lowest segmentation accuracies. To address this, we transitioned from using the segmentation-models package and instead implemented a U-Net with a ResNet34 backbone using Keras. Dropout layers were incorporated after the activation layers in Iterative blocks across both the encoder and decoder paths (Fig. 3), with a dropout rate of 50% in the backbone and 20% in the upsampling path (Kamnitsas et al., 2017; Srivastava et al., 2014). An Iterative block refers to a set of repeated layers that appear multiple times throughout the network architecture. The specific dropout percentages were empirically determined through comparative experiments evaluating IoU on validation data.

For the optic chiasm and optic nerve experiments, initial evaluations indicated that SEResNet34 and ResNeXt50 backbones achieved higher baseline performance compared to ResNet34, which means they had higher IoU score on validation and test data, with greater emphasis placed on test set performance. Nevertheless, ResNet34 was deliberately chosen for the dropout experiments to investigate whether explicit regularization could compensate for its relatively lower baseline accuracy. The objective was not to maximize absolute performance, but to isolate the effect of explicit regularization on a baseline residual architecture. Apart from the addition of dropout layers and extended training epochs, all other experimental conditions, including dataset splits, optimizer, learning rate, and loss function, remained consistent across models.

3 Results

Networks for the automatic segmentation of OARs were trained using several backbones available in the segmentation-models package. Ultimately, three backbones were selected for the U-Net architecture: ResNet34, SEResNet34, and ResNeXt50 (Table 2). The final backbone selection was based on the IoU achieved on both the validation and test datasets, with greater emphasis placed on test set performance. Inference time was also taken into consideration during the selection process. The results presented here reflect the backbone that yielded the best overall performance according to these criteria. A total of 10,076 slices were used for training the networks. CT images in the dataset varied in size between 1024×1024 and 512×512 pixels. The variability in image resolution within the public dataset arises from the use of two different CT scanners during data acquisition, as reported in the referenced study (Podobnik et al., 2024). For consistency, all images were resized to 512×512, with cubic interpolation applied to images and nearest neighbor interpolation applied to masks. The number of slices available for training varied from 95 to 4,049, depending on the number of masks assigned to each organ (Table 1). The available images were divided into test, validation, and training sets with a ratio of 0.15, 0.15, and 0.70,

Table 2: The selected backbone for each OAR.

Backbone	OARs	#
ResNet34	Left Carotid - Right Carotid - Bone Mandible - Left Brachial Plexus - Right Brachial Plexus - Brain - Brainstem - Cavity Oral - Right Cochlea - Cricopharyngeus - Esophagus - Left Eye - Left Eye Posterior - Right Eye Posterior - Right Eye - Right Submandibular Gland - Glottis - Right Lens - Left Lung - Muscle Constrictor - Left Parotid - Trachea	22
SeResNet34	Arytenoid - Left Cochlea - Left Gland Lacrimal - Right Gland Lacrimal - Left Submandibular Gland - Thyroid Gland - Larynx - Lips - Right Lung - Lungs - Optic Chiasm - Right Parotid - Pituitary - Spinal Canal - Spinal Cord	15
ResNeXt50	Buccal Mucosa - Left Lens - Left Optic Nerve - Right Optic Nerve	4

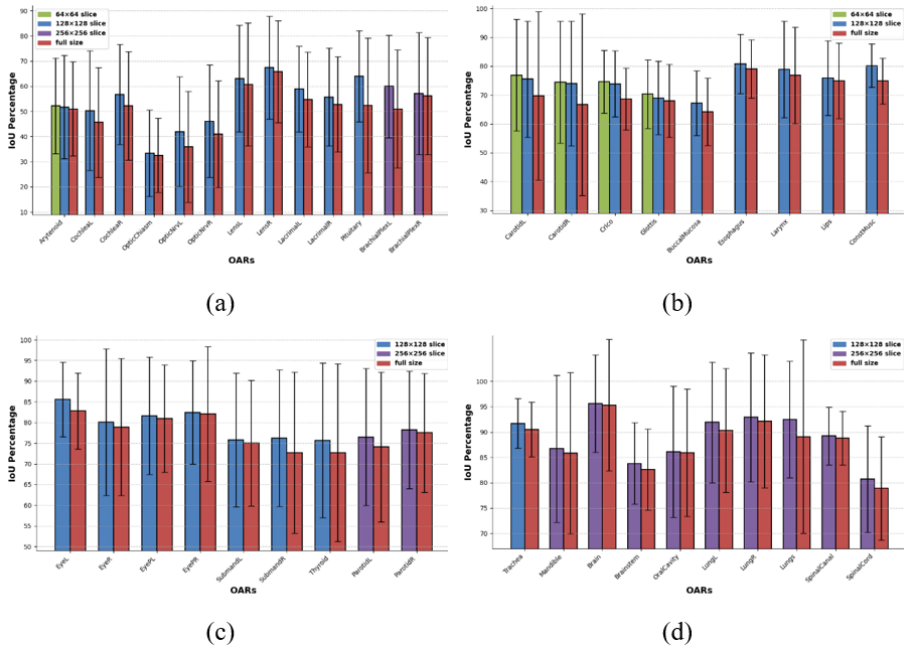


Figure 4: Bar plot with standard deviation error bars showing IoU scores for all the OARs for sliced and full models. (a) Arytenoid, Left Cochlea, Right Cochlea, Optic Chiasm, Left Optic Nerve, Right Optic Nerve, Left Lens, Right Lens, Left Lacrimal Gland, Right Lacrimal Gland, Pituitary, Left Brachial Plexus, and Right Brachial Plexus. (b) Left Carotid, Right Carotid, Cricopharyngeus, Glottis, Buccal Mucosa, Esophagus, Larynx, Lips, and Constrictor Muscle. (c) Left Eye, Right Eye, Left Eye (Posterior), Right Eye (Posterior), Left Submandibular Gland, Right Submandibular Gland, Thyroid Gland, Left Parotid, and Right Parotid. (d) Trachea, Mandible, Brain, Brainstem, Oral Cavity, Left Lung, Right Lung, Lungs, Spinal Canal, and Spinal Cord.

respectively.

All networks (one for each organ) were trained for 50 epochs using the Adam optimizer (learning rate = 0.0005) and Dice loss function. The slice-based preprocessing strategy led to consistent improvements in segmentation performance. As shown in Fig. 4, the sliced models achieved higher IoU scores across all OARs compared to the full-size models, a trend that was similarly reflected in the Dice scores (Fig. 5). Quantitatively, the slice-based networks achieved an average increase of 4.1% in IoU and 3.2% in Dice score across all 41 OARs.

Furthermore, the consistency of segmentation was greatly enhanced. The standard deviation of IoU and Dice scores decreased by 8.1% and 13.0%, respectively, indicating more reliable and stable performance across the test set (Figs. 6-b and 6-d).

The performance gains were even more pronounced for a subset of 11 anatomically challenging structures: the pituitary gland, left and right brachial plexuses, left and right carotid arteries, cricopharyngeus, left and right optic nerves, and left and right cochleae. For these OARs, the slice-based networks showed a remarkable average improvement of 10.9% in IoU and 9.0% in Dice score.

Paired Wilcoxon signed-rank testing revealed highly statistically significant improvements in Dice and IoU scores for slice-based networks compared to full-size networks across all 41 organs at risk (paired Wilcoxon signed-rank test, $p < 0.001$ for both metrics). A separate analysis focusing on the subset of organs with larger reported improvements also demonstrated highly statistically significant gains (paired Wilcoxon signed-rank test, $p < 0.001$ for Dice and IoU), confirming the robustness of the observed performance improvements. Furthermore, this test was also applied to the reduction in the standard deviation of IoU and Dice, which yielded statistically significant reductions (paired Wilcoxon signed-rank test, $p < 0.01$ for Dice and $p = 0.018$ for IoU).

In addition to accuracy gains, the slicebased approach

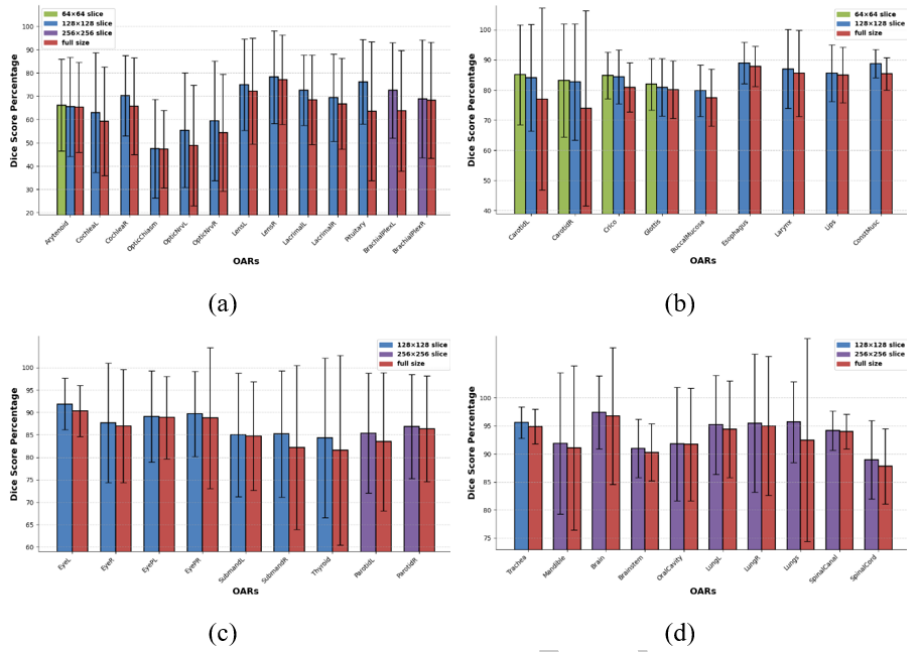


Figure 5: Bar plot with standard deviation error bars showing Dice scores for all the OARs for sliced and full models. (a) Arytenoid, Left Cochlea, Right Cochlea, Optic Chiasm, Left Optic Nerve, Right Optic Nerve, Left Lens, Right Lens, Left Lacrimal Gland, Right Lacrimal Gland, Pituitary, Left Brachial Plexus, and Right Brachial Plexus. (b) Left Carotid, Right Carotid, Cricopharyngeus, Glottis, Buccal Mucosa, Esophagus, Larynx, Lips, and Constrictor Muscle. (c) Left Eye, Right Eye, Left Eye (Posterior), Right Eye (Posterior), Left Submandibular Gland, Right Submandibular Gland, Thyroid Gland, Left Parotid, and Right Parotid. (d) Trachea, Mandible, Brain, Brainstem, Oral Cavity, Left Lung, Right Lung, Lungs, Spinal Canal, and Spinal Cord.

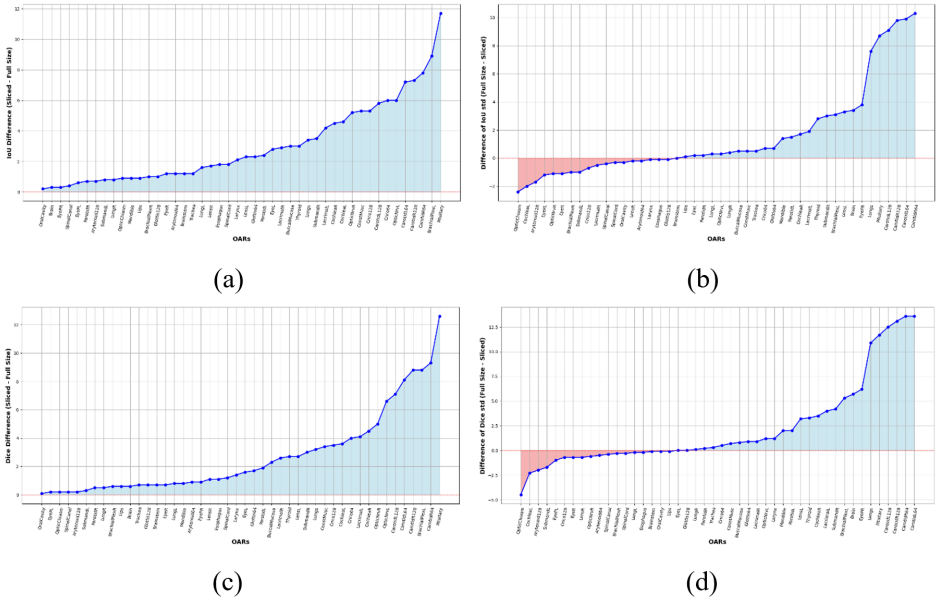


Figure 6: (a) Difference in IoU scores between sliced and full-size networks across all OARs (sliced - full size). (b) Comparison of standard deviation of IoU scores between sliced and full-size networks across all OARs (full size - sliced). (c) Difference in Dice scores between sliced and full-size networks across all OARs (sliced - full size). (d) Comparison of standard deviation of Dice scores between sliced and full-size networks across all OARs (full size - sliced).

yielded substantial computational benefits. The faster training time of slicebased networks is naturally attributed to the smaller input arrays used during training, which result from the slice-based preprocessing. The training time for slicebased networks was reduced to less than half the time required for fullsize networks, representing an improvement of over 130% in training speed.

During prediction on a CT scan, the slicebased networks performed about 3 to 30 times faster (8 times on average) than the fullsize networks. Figure 7 illustrates the inference time comparison between the two network types via box plots. The outliers observed in the fullsize network

Table 3: Effects of adding dropout layers to the U-Net architecture and increasing epoch for three selected OARs.

OAR	50 epochs / sliced / without dropout		50 epochs / sliced / with dropout		300 epochs / sliced / with dropout		300 epochs / full size / with dropout	
	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice
Optic Chiasm	33.4±17.2	47.4±21.2	37.9±17.9	52.3±20.8	45.5±18.2	60.2±18.8	36.7±12.1	52.4±14.1
Left Optic Nerve	41.9±21.8	55.3±24.7	45.0±23.7	57.8±26.7	51.1±22.6	64.1±23.9	48.7±21.7	62.2±23.4
Right Optic Nerve	46.1±22.3	59.3±25.7	49.4±22.8	62.3±25.8	51.2±22.2	64.1±25.2	50.0±23.1	62.9±25.4

plot correspond to models with the ResNeXt50 backbone, which has higher architectural complexity compared to the other two backbones tested. Furthermore, Fig. 8 demonstrates the influence of both backbone type and slice size on inference time. As illustrated in the figure, the greater number of parameters and layers in ResNeXt50 increases its inference time. Inferences were performed for all OARs using a CPU on a 198slice CT volume.

To further improve performance on the most challenging OARs, dropout layers were incorporated and training duration was extended. As summarized in Table 3, the introduction of dropout resulted in average improvements of 3.6% in IoU and 3.5% in Dice across the optic chiasm and optic nerves at 50 training epochs. Extending the training to 300 epochs while maintaining the same dropout configuration further increased the average improvement to 8.8% for both metrics.

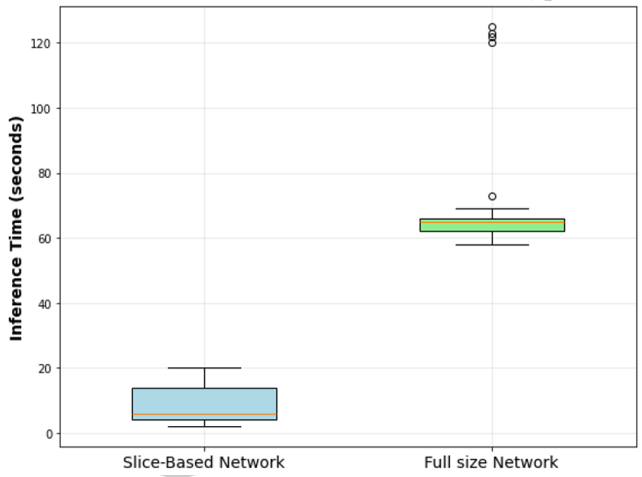


Figure 7: Box plot of Inference time for slice-based and fullsize networks.

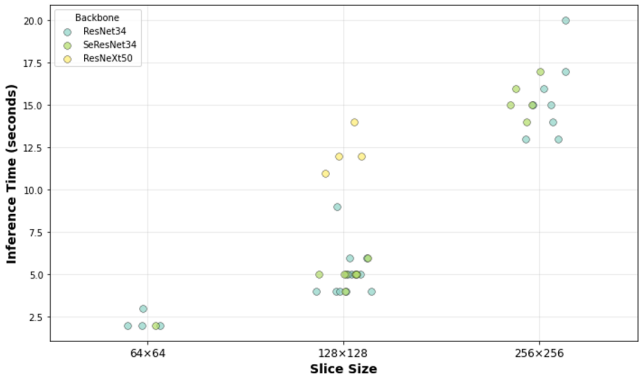


Figure 8: This scatter plot shows the impact of backbones and slice sizes on the inference time of slice-based networks.

A paired Wilcoxon signed-rank test was performed to assess the statistical significance of the performance differences between the baseline configuration and the models trained with dropout. Although numerical improvements were observed, these differences did not reach statistical significance ($p = 0.25$ for both Dice and IoU)), likely due to limited sample size for these structures. These results should therefore be interpreted cautiously. Similar results were obtained when comparing models trained for 50 and 300 epochs.

To compare the impact of using two different datasets for network training, new networks were trained separately with each dataset for organs whose related networks had previously been trained with a mixture of images from both datasets. This was done for the OARs that have a minimum sufficient number of images available. The results are presented in Fig. 9. Additionally, Figs. 10 and 11 illustrate the differences in outcomes between these networks and the previous ones.

To ensure comparability, the slice size for all organs was consistent with the dimensions specified in material and methods section. However, for the public dataset, smaller slice sizes could also be applied depending on the organ:

- 128 × 128 slices: Brainstem, Oral Cavity, Left and Right Parotids, and Spinal Cord.
- 64 × 64 slices: Left and Right Cochlea, Esophagus, Left and Right Lacrimal Glands, Left and Right Submandibular Glands, Thyroid Gland, Larynx, Left and Right Lenses, Optic Chiasm, and Pituitary.

Figure 12 illustrate a sample for Esophagus predicted mask generated by the slice-based network. For comparison with some of the results presented in this study, Table 4 displays the outcomes obtained using the commercial Limbus AI software on two CT scans ($n = 2$), one from each dataset, across seven OARs, as processed at the institute. This comparison should be interpreted as qualitative rather than statistically conclusive.

Table 4: IoU and Dice scores from the evaluation of contours generated by the commercial software Limbus AI.

#	OAR	IoU ± std	Dice Score ± std
1	Mandible	74.8±15.7	84.3±13.4
2	Brainstem	74.0±14.3	84.2±10.3
3	Oral Cavity	83.7±9.7	90.8±6.4
4	Esophagus	66.5±12.1	80.4±4.4
5	Optic Chiasm	35.8±6.6	52.2±7.5
6	Left Optic Nerve	22.1±4.3	35.1±5.7
7	Right Optic Nerve	32.3±13.1	46.3±14.7

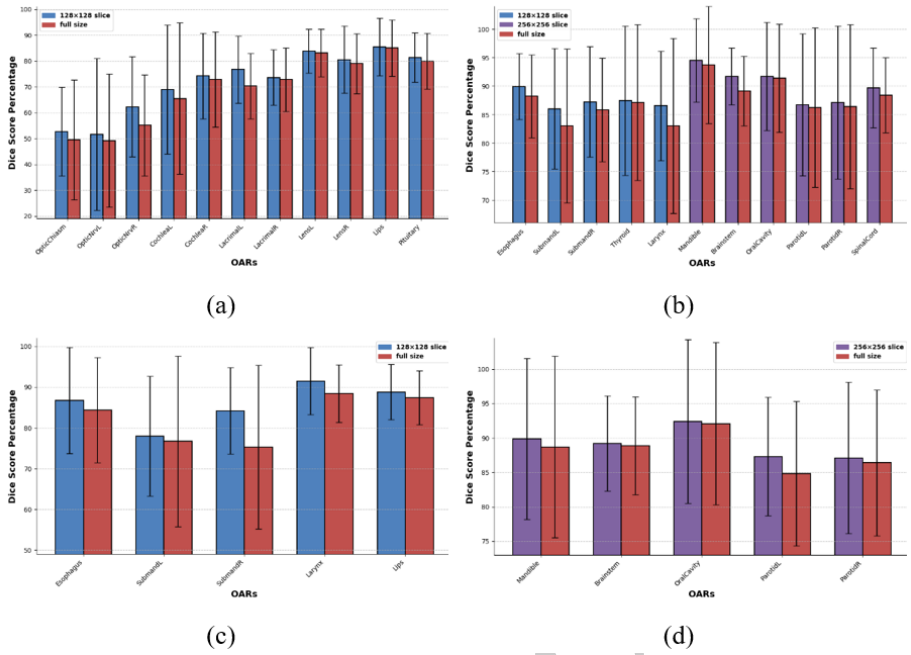


Figure 9: (a) and (b) are Dice score percentage for sliced and full-size networks trained on the public dataset. (c) and (d) are Dice score percentage for sliced and full-size networks trained on the institute dataset. (a) Optic Chiasm, Left Optic Nerve, Right Optic Nerve, Left Cochlea, Right Cochlea, Left Lacrimal Gland, Right Lacrimal Gland, Left Lens, Right Lens, Lips, and Pituitary. (b) Esophagus, Left Submandibular Gland, Right Submandibular Gland, Thyroid Gland, Larynx, Mandible, Brainstem, Oral Cavity, Left Parotid, Right Parotid, and Spinal Cord. (c) Esophagus, Left Submandibular Gland, Right Submandibular Gland, Larynx, and Lips. (d) Mandible, Brainstem, Oral Cavity, Left Parotid, and Right Parotid.

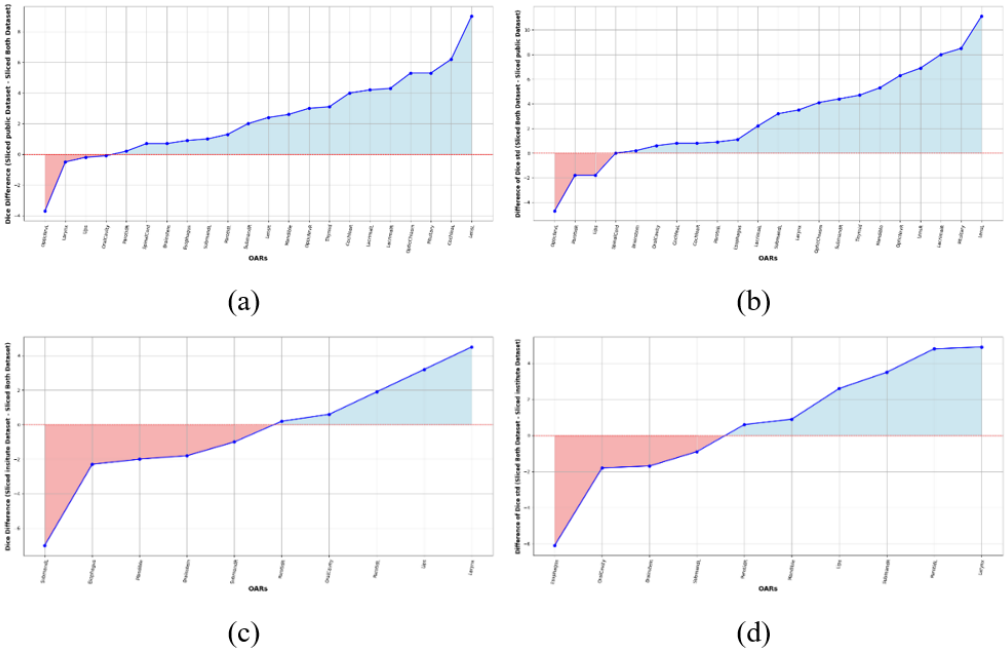


Figure 10: (a) Comparison of Dice score percentage between the public dataset and complete (both) dataset for slice-based networks. (b) Comparison of standard deviations of Dice scores between the public dataset and complete (both) dataset for slice-based networks. (c) Comparison of Dice score percentage between the institute dataset and complete (both) dataset for slice-based networks. (d) Comparison of standard deviations of Dice scores between the institute dataset and complete (both) dataset for slice-based networks.

4 Discussion

Deep learning has revolutionized medical image segmentation, and this field is advancing rapidly. Extensive research, including numerous papers and studies, has been conducted and continues to drive progress. For instance, Gibbons et al. (Gibbons et al., 2023) used a dataset of 90 images, comprising 30 CT scans each from the Head

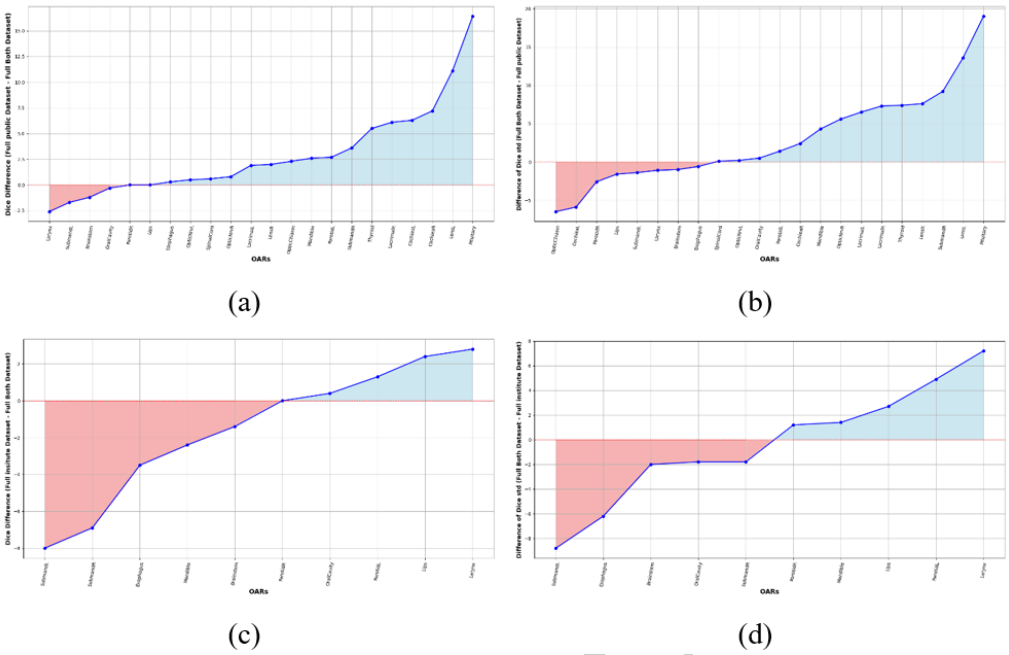


Figure 11: (a) Comparison of Dice score percentage between the public dataset and complete (both) dataset for full-size networks. (b) Comparison of standard deviations of Dice scores between the public dataset and complete (both) dataset for full-size networks. (c) Comparison of Dice score percentage between the institute dataset and complete (both) dataset for full-size networks. (d) Comparison of standard deviations of Dice scores between the institute dataset and complete (both) dataset for full-size networks.

and Neck (H&N), Thorax, and Pelvis regions. To compare two segmentation methods, Atlas-based segmentation and deep learning, Gibbons et al. employed the commercial software, Miranda DLCExpert. The study involved segmenting organs at risk, including the parotids, brainstem, spinal cord, mandible, oral cavity, and submandibular glands (H&N region); lungs, heart, and esophagus (Thorax); and rectum, bladder, and femoral heads (Pelvis). The DSC and Hausdorff Distance (HD) were used to evaluate the results. The deep learning model demonstrated superior performance. Its best and worst DSC scores ranged from 0.80 to 0.94 (H&N), 0.74 to 0.98 (Thorax), and 0.87 to 0.98 (Pelvis), compared to the atlas-based model's ranges of 0.68 to 0.91, 0.48 to 0.98, and 0.77 to 0.96 for the respective regions.

Mikhailov et al. (Mikhailov et al., 2024) implemented a deep learning-based interactive segmentation network. The network was trained using distinct sets of CT and MRI images, 97 for cancers in the pelvis region, 131 for the liver region, and 244 for the pancreas region. An Nvidia P40 GPU with 24 GB of VRAM served as the training hardware. Performance metrics varied by region:

- Pelvis: For tumors, the mean Dice and IoU scores were 88.3 and 79.0, respectively. For other organs within the pelvis, the scores ranged from 73.3-93.0 (Dice) and 57.8-87.0 (IoU).
- Liver: The Dice score achieved was 0.90 for the tumor and 0.96 for the liver organ itself.
- Pancreas: Correspondingly, the Dice scores were 0.85 for the tumor and 0.84 for the pancreas organ.

Putz et al. (Putz et al., 2025) utilized the BraTS 2020 challenge dataset to train a network for segmenting glioma brain tumors. This dataset included 396 MRI images, comprising 16,744 slices, for the training process. The hardware employed was an NVIDIA RTX A6000 GPU with 48 GB of VRAM, and the study utilized Meta AI's Segment Anything model. The reported result, based on the mean best Intersection over Union (IoU), was 0.762.

The aforementioned studies (Gibbons et al., 2023; Mikhailov et al., 2024; Putz et al., 2025), along with many others in this field, demonstrate that training powerful segmentation networks typically requires significant computational resources and may involve costly commercial software. These factors often entail substantial costs. To address these challenges, our study employs a slice-based preprocessing method. By reducing the input data size, this approach lowers hardware demands, potentially decreasing training time and allowing for the use of larger datasets without needing more powerful equipment.

Furthermore, the performance results reported in these prior articles provide valuable benchmarks, offering insight into the current ranges of accuracy for medical image segmentation. For illustrative context, we also include results from Limbus AI in Table 4, though it is important to note that these were obtained from a very limited sample size and should be considered indicative only, not statistically generalizable. Most results obtained for the automatic segmentation of head and neck OARs (Figs. 4 and 5) are comparable to or even surpass of those reported in studies, such as Gibbons et al. (Gibbons et al., 2023), as well as results of Limbus AI in Table 4.

The performance gains observed with the slice-based

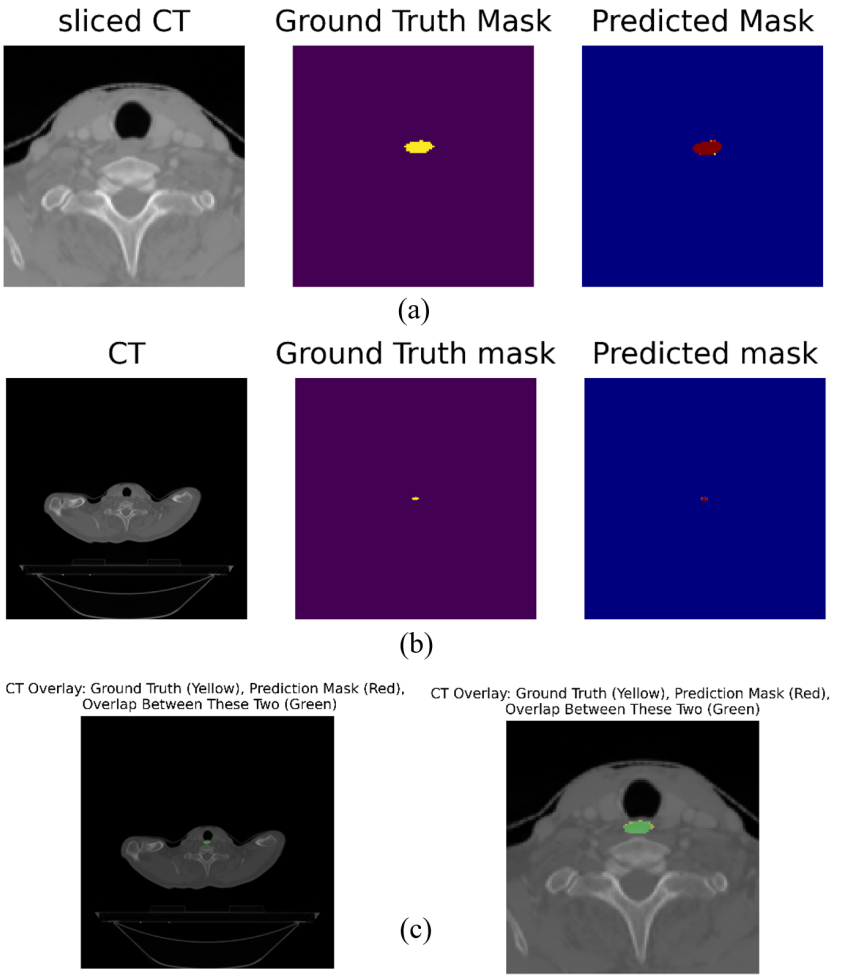


Figure 12: (a) A sample Esophagus predicted mask generated by the slice-based network, displayed alongside the corresponding CT image and ground truth mask. (b) The predicted mask from the slice-based network was reconstructed to its original size. (c) Overlay of the ground truth and predicted mask on the CT image, shown for both the original and sliced sizes.

approach can be attributed to several factors. By cropping the image to the relevant region, the network's effective receptive field is better aligned with the scale of the target OAR. This forces the model to dedicate its capacity to learning discriminative features from the local context, reducing the risk of being influenced by irrelevant anatomical variations or image artifacts distant from the organ of interest. This is particularly can be beneficial for smaller structures, which can be overwhelmed by the vast background in a full-size image. Furthermore, the reduction in input size inherently decreases the model's workload, allowing it to train more efficiently and focus its computational resources on the most relevant pixels.

Across all OARs, the slice-based networks achieved consistent improvements over full-size networks, with increases of 4.1% in IoU and 3.2% in Dice score and decrease of 8.1% in IoU standard deviation, and 13.0% in Dice score standard deviation. Notably, for anatomically challenging structures, including the pituitary gland, brachial plexuses, carotid arteries, cricopharyngeus, optic nerves, cochleae, and constrictor muscle, performance gains were even more pronounced, with IoU and Dice scores improving by 10.9% and 9.0%, respectively.

The findings suggest that utilizing smaller input patch sizes within the slice-based approach can enhance network performance, aligning with expectations regarding computational efficiency. Additionally, for OARs that are typically well-defined and exhibit high contrast in CT images (e.g., mandible, lungs), the performance difference between the proposed slice-based network and full-size network was minimal. This contrasts with structures that are often less distinct, such as the carotid arteries or pituitary gland, where differences might be more apparent.

Beyond improvements in the mean IoU and Dice scores, the consistency of the segmentation also improved, as indicated by a reduction in the standard deviation of these metrics across the dataset. In terms of computational speed, the slice-based network demonstrated significantly faster segmentation during prediction on a CT, performing about 3 to 30 times faster than the full-size networks. The average of 6.68 seconds inference time for slice networks versus 69.68 seconds for full-size networks. The exact speed-up factor varied depending on parameters such as the backbone, and the input patch size employed. Additionally, slicing the images significantly reduces the training time of the models. In this study, in total, the training time of the slice-based networks was approximately 130% shorter than the training time of their

corresponding full-size networks.

Moreover, the integration of dropout layers into the U-Net architecture led to noticeable performance gains for three specific OARs, the optic chiasm and the optic nerves (left and right). This improvement was particularly evident with increased training epochs and also conferred benefits to the full-size network's performance. This enhancement is likely attributable to dropout's effectiveness in reducing model overfitting (Srivastava et al., 2014).

A potential limitation of the proposed methodology, as implemented in this study, is its primary design for networks focused on the segmentation of a single organ at a time. While this approach yields high accuracy for individual OARs, it may be less optimal for clinical workflows requiring the simultaneous contouring of all OARs within a single pass.

Furthermore, the number of training slices varied substantially across organs, which may influence segmentation performance, particularly for underrepresented structures. In this study, no explicit data augmentation or balancing strategies were employed, as the primary objective was to conduct a controlled comparative analysis between full-size and slice-based training frameworks under identical data distributions and preprocessing conditions. While such a design facilitates relative comparison between methods, data imbalance may still affect the absolute performance and generalization capability for organs with limited training data. Future work will investigate stratified sampling or class-balanced loss to address this limitation.

Resizing and resampling are known to influence the effective spatial resolution of medical images, particularly when voxel spacing is not explicitly incorporated into the preprocessing pipeline. As discussed in prior work, differences in voxel spacing can affect the representation of anatomical structures and should be carefully considered in segmentation studies. In the present study, voxel spacing was not explicitly accounted for, and images were resized to a fixed in-plane resolution. However, all network configurations were trained and evaluated using identical preprocessing steps. As previously reported, when comparative analyses are conducted under consistent preprocessing conditions, relative performance differences between models remain interpretable despite such limitations (Brudfors et al., 2022; Haque and Neubert, 2020). Accordingly, the reported results should be interpreted within this consistent preprocessing framework.

An important consideration in this study is the impact of inter-observer variability in the training data. The manual segmentations used as ground truth inherently contain variations between different clinical experts, reflecting the subjective nature of anatomical boundary delineation in medical imaging. This inter-observer variability represents a fundamental performance ceiling for any segmentation model, as the algorithm cannot achieve consensus accuracy beyond the level of agreement among human experts (Brouwer et al., 2014). Consequently, some portion of the observed segmentation errors may be attributable to these inherent ambiguities in the reference standard rather than model deficiencies alone (Podobnik et al., 2024).

In this study, two different types of datasets were used, which inherently increased the likelihood of requiring larger slicing ranges. Based on the results obtained, it is evident that smaller slices tend to yield better outcomes. However, due to anatomical differences and variations in CT images produced by different devices, using smaller slices is not always feasible. Nevertheless, if the slicing range for images from a specific device at a particular medical center is determined individually, the chances of achieving better and more consistent results with this method will be improved.

Training the networks on the separate datasets yielded several findings. For the public dataset, segmentation performance improved for most organs compared to the combined dataset results. Furthermore, employing smaller slice sizes is feasible for certain organs within the public dataset, and doing so can further enhance network performance. This combined approach, utilizing standardized data and potentially smaller slices, appears particularly suitable for implementation within a specific clinic or medical center where consistent data acquisition protocols can be maintained. However, for general use cases requiring the segmentation of new images from potentially diverse sources, the primary network (trained on the larger, combined dataset) is expected to offer more robust and generalizable performance due to its broader exposure during training.

Finally, regarding the performance on the institute dataset, its weaker results compared to the public dataset are likely attributable to its smaller number of images. This underscores the critical importance of dataset size for effective deep learning model training.

5 Conclusions

This study systematically evaluated the effect of slice-based preprocessing on the performance of Residual U-Net models for head and neck OAR segmentation. The results demonstrate that restricting the network input to organ-specific regions improves segmentation accuracy and consistency while substantially reducing computational cost.

Across 41 organs at risk, slice-based models achieved statistically significant improvements in both Dice and IoU metrics compared to full-size training. The performance gains were particularly pronounced for small and anatomically challenging structures, including the pituitary gland, brachial plexuses, carotid arteries, optic nerves, cochleae, and cricopharyngeus. In addition to accuracy improvements, the proposed approach resulted in markedly faster training and inference times, supporting its practical applicability in clinical environments.

Additional experiments incorporating dropout regularization and extended training for optic pathway structures suggested potential performance benefits; however, these improvements were not statistically significant and should therefore be interpreted cautiously. While the present study trained independent models for each organ and did not explicitly incorporate voxel spacing normalization or advanced data balancing strategies, all comparative experiments were conducted under identical preprocessing

conditions, enabling consistent relative evaluation. Future work may explore multi-organ segmentation frameworks, adaptive cropping strategies, and data augmentation techniques to further enhance generalization performance.

Overall, the slice-based Residual U-Net method provides a computationally efficient and practically implementable solution for improving OAR segmentation in head and neck radiotherapy planning.

Acknowledgements

The authors would like to thank the Reza Radiotherapy and Oncology Center (RROC) for providing clinical imaging data for this study.

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work.

Funding

The authors declare that no funds, grants, or other financial support were received during the preparation of this manuscript.

References

Adams, R. and Bischof, L. (1994). Seeded region growing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(6):641–647.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Brouwer, C. L., Steenbakkers, R. J., Gort, E., et al. (2014). Differences in delineation guidelines for head and neck cancer result in inconsistent reported dose and corresponding NTCP. *Radiotherapy and Oncology*, 111(1):148–152.

Brudfors, M., Balbastre, Y., Ashburner, J., et al. (2022). Fitting segmentation networks on varying image resolutions using splatting. In *Annual Conference on Medical Image Understanding and Analysis*, pages 271–282. Springer.

Chan, J. W., Kearney, V., Haaf, S., et al. (2019). A convolutional neural network algorithm for automatic segmentation of head and neck organs at risk using deep lifelong learning. *Medical Physics*, 46(5):2204–2213.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., et al. (2016). 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.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.

Gibbons, E., Hoffmann, M., Westhuyzen, J., et al. (2023). Clinical evaluation of deep learning and atlas-based auto-segmentation for critical organs at risk in radiation therapy. *Journal of Medical Radiation Sciences*, 70:15–25.

Han, X., Hoogeman, M. S., Levendag, P. C., et al. (2008). 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.

Haque, I. R. I. and Neubert, J. (2020). Deep learning approaches to biomedical image segmentation. *Informatics in Medicine Unlocked*, 18:100297.

Harari, P. M., Song, S., and Tomé, W. A. (2010). 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.

He, K., Zhang, X., Ren, S., et al. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778.

Howard, A. G., Zhu, M., Chen, B., et al. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

Hu, J., Shen, L., and Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141.

Huang, G., Liu, Z., Van Der Maaten, L., et al. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708.

Iakubovskii, P. (2019). Segmentation models pytorch. GitHub. *GitHub repository https://github.com/qubvel/segmentation_models. pytorch*.

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

Kamnitsas, K., Ledig, C., Newcombe, V. F., et al. (2017). Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Medical Image Analysis*, 36:61–78.

Kass, M., Witkin, A., and Terzopoulos, D. (1988). Snakes: Active contour models. *International Journal of Computer Vision*, 1(4):321–331.

Liang, S., Tang, F., Huang, X., et al. (2019). Deep-learning-based detection and segmentation of organs at risk in nasopharyngeal carcinoma computed tomographic images for radiotherapy planning. *European Radiology*, 29(4):1961–1967.

Mikhailov, I., Chauveau, B., Bourdel, N., et al. (2024). A deep learning-based interactive medical image segmentation framework with sequential memory. *Computer Methods and Programs in Biomedicine*, 245:108038.

Otsu, N. et al. (1979). A threshold selection method from gray-level histograms. *Automatica*, 11(285-296).

Podobnik, G., Ibragimov, B., Peterlin, P., et al. (2024). OARiability: Interobserver and intermodality variability analysis in OAR contouring from head and neck CT and MR images. *Medical Physics*, 51(3):2175–2186.

Podobnik, G., Strojan, P., Peterlin, P., et al. (2023). HaN-Seg: The head and neck organ-at-risk CT and MR segmentation dataset. *Medical Physics*, 50(3):1917–1927.

Putz, F., Beirami, S., Schmidt, M. A., et al. (2025). The Segment Anything foundation model achieves favorable brain tumor auto-segmentation accuracy in MRI to support radiotherapy treatment planning. *Strahlentherapie und Onkologie*, 201(3):255–265.

Rahman, M. A. and Wang, Y. (2016). Optimizing intersection-over-union in deep neural networks for image segmentation. In *International Symposium on visual Computing*, pages 234–244. Springer.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer.

Sharp, G., Fritscher, K. D., Pekar, V., et al. (2014). Vision 20/20: perspectives on automated image segmentation for radiotherapy. *Medical Physics*, 41(5):050902.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

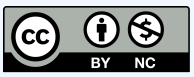
Srivastava, N., Hinton, G., Krizhevsky, A., et al. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

Sudre, C. H., Li, W., Vercauteren, T., et al. (2017). Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *International Workshop on Deep Learning in Medical Image Analysis*, pages 240–248. Springer.

Van Rooij, W., Dahele, M., Brandao, H. R., et al. (2019). Deep learning-based delineation of head and neck organs at risk: geometric and dosimetric evaluation. *International Journal of Radiation Oncology* Biology* Physics*, 104(3):677–684.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics bulletin*, 1(6):80–83.

Xie, S., Girshick, R., Dollár, P., et al. (2017). Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500.



To cite this article:

K. Heshmati Jannat Magham, L. Rafat-Motavalli, H. Miri-Hakimabad, M. Dayyani. Improving head and neck organs at risk segmentation in CT using residual U-Net with slice-based preprocessing. *Radiation Physics and Engineering*, In Press.

DOI:

To link to this article: