Transfer Learning for Head and Neck CT Image Segmentation

Tianyu Zhang
New York University
tz904@nyu.edu

Abstract

Segmentation of organs-at-risk (OARs) is a key component of radiotherapy. Various deep learning methods have been attempted for automatic segmentation of head and neck organs from computed tomography (CT) images. I have jointly designed a Res3DUnet and UNEt TRansform-ers (UNETR) based model for head and neck segmentation with competitive performance (75.0 % in mean Dice Score) compared with state-of-the-art methods in the capstone project of last year. In this paper, as a continuation of capstone study, I incorporate the idea of transfer learning to further improve the model performance on previous OARs with more comprehensive evaluation metrics (77.1 % in mean Dice Score and 2.66 mm in mean 95% hausdorff distance). Meanwhile, with transfer learning, our model is able to not only shorten the training time, but also enhance the performance for 3 new small size OARs with limited data.

1 Introduction

Defining OARs through delineating them on the computed tomography (CT) scans is an essential task for the radiation oncologists. Accurate delineation of OARs can help protect normal organs from long term toxicity. While manually annotating hundreds of slices of 3D CT scans is very time-consuming, an automatic segmentation solution is able to save human efforts and improve the accuracy.

Results from previous studies have shown that the deep learning based method is able to achieve reasonably high accuracy on head and neck segmentation tasks. My project is a continuation of my last year's capstone project[1], which proposed an end-to-end Res3DUnet [2] and UNETER [5] based model solution with competitive performance to the winner of MICCAI imaging challenge in 2015 [3]. Through transfer learning with the help of other head and neck CT dataset, I am able to improve the previous model to have state-of-the-art model performance with more comprehensive evaluation metrics. Meanwhile, another contribution of this paper is to produce an end-to-end transfer learning solution on new OARs (PCM-Superior, PCM-Middle and PCM-Inferior), with shortened training period and superior performance to available solutions.

The remainder of this report is structured as follows:

- I provide an overview of related research work that inspired my project in Section 2.
- In Section 3, Section 4.1, and Section 4.2 I introduce the available data sets and my approach, with detailed experiment results.
- I present our final results and compare them with the state-of-the-art in Section 4.3 and Section 4.4.

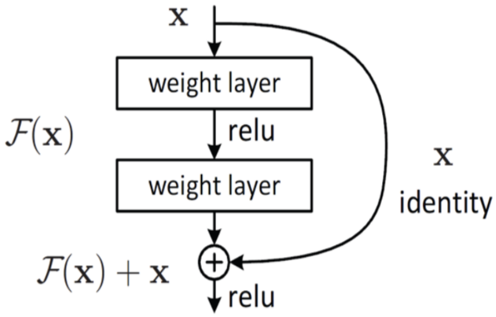

x
weight layer
$\mathcal{F}(\mathbf{x})$
relu
weight layer
x
identity
$\mathcal{F}(\mathbf{x}) + \mathbf{x}$
relu

Figure 1: Residual block

2 Related Work

MICCAI Head and Neck Auto Segmentation Challenge in 2015[3] MICCAI Head And Neck Auto-Segmentation Challenge is held in conjunction with the “Medical Image Computing and Computer Assisted Interventions” conference in Munich, Germany. The primary goal is to develop an benchmark of automatic segmentation performance on nine given OARs.

Residual Neural Network[4] The key component of residual neural network (ResNet) is the residual block (Figure 1). Residual blocks are designed to to alleviate the problem of the impact of vanishing gradients. Through setting the learning function to F(x) + x, it enables the networks to learn F(x) = H(x) - x instead of H(x), which is proven to be able to improve the generalization ability of the network.

3D Unet[2] 3d U-Net is a 3D-version of U-shaped network architecture[6] , containing a contractive and an expanding path. Through combination of convolution and pooling operations, the model condenses the large-scale features in the bottleneck. After the bottleneck, the volume is reconstructed by convolution and upsampling operations with small-scale features embedded. The pros of 3d U-Net compared with 2d U-Net are its abilities to include spatial and more global information. The cons of 3d U-Net are concerns about training time and memory issues.

UNEt TRansformers[5] UNEt TRansformers (UNETR) is a variant of 3D Unet which adopts the idea of Transformer. It utilizes a Transformer as the encoder to learn sequence representations of the input CT volume, while maintaining the "U-shaped" structure for the encoder and decoder. The Transformer encoder is directly connected to a CNN-based decoder through skip connections at various levels of resolution to construct the final semantic segmentation. This model is chosen as my comparison model because of its natural advantage of segmenting symmetric tasks compared with normal 3d U-net which has been proven in last year capstone project.

Transfer Learning[13] Transfer Learning is a method focusing on utilizing knowledge gained from one domain to solve a different but related problem, especially for new problem with limited data. Transferring information from learned tasks for new tasks has the potential to improve training efficiency. In the field of CT segmentation, Vu’s study [16] has shown that transfer learning can help shortened the training time significantly with fewer data and slightly worse performance.

3 Materials and methods

3.1 Datasets

MICCAI 2015 Head and Neck Auto-segmentation Challenge Dataset The data set contains 48 CT images corresponding to 48 patients with 9 OARs manually annotated. (brainstem, mandible, left and right optic nerves, optic chiasm, left and right parotid glands, left and right submandibular glands). 15 images were set as test set and the remaining images were randomly split into training (25 images) and validation (8 images). Each CT image is a 3D volume of dimension N x 512 × 512, having a height and width of 512 pixels and N (140 ± 30) slices.

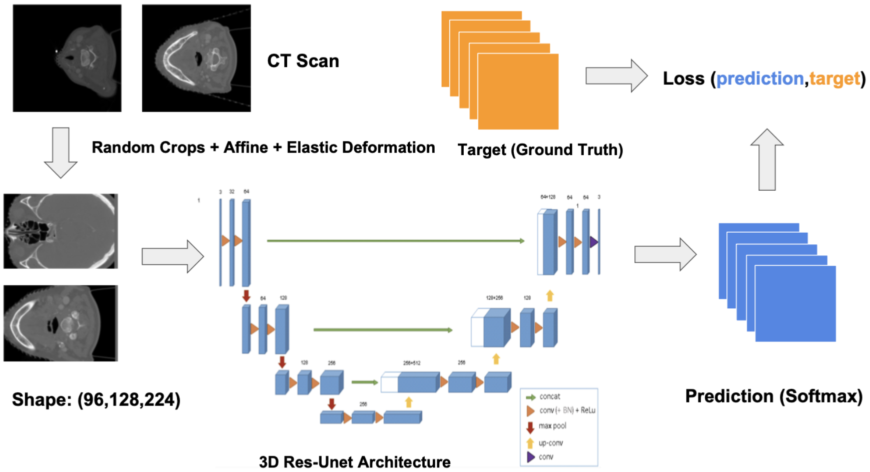


Figure 2: Overview of Our Training Framework

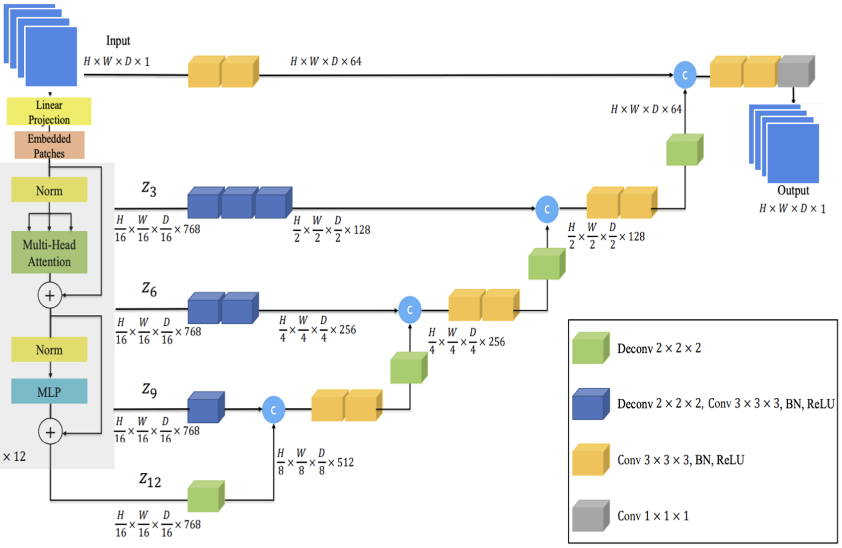

Input
$H \times W \times D \times 1$
$H \times W \times D \times 64$
Linear Projection
Embedded Patches
Norm
Multi-Head Attention
Norm
MLP
× 12
z_3
z_6
z_9
z_{12}
$\frac{H}{16} \times \frac{W}{16} \times \frac{D}{16} \times 768$
$\frac{H}{2} \times \frac{W}{2} \times \frac{D}{2} \times 128$
$\frac{H}{4} \times \frac{W}{4} \times \frac{D}{4} \times 256$
$\frac{H}{8} \times \frac{W}{8} \times \frac{D}{8} \times 512$
$H \times W \times D \times 64$
Output
$H \times W \times D \times 1$
Deconv 2 × 2 × 2
Deconv 2 × 2 × 2, Conv 3 × 3 × 3, BN, ReLU
Conv 3 × 3 × 3, BN, ReLU
Conv 1 × 1 × 1

Figure 3: UNEt TRansformers (UNETR) Architecture

NYU langone Dataset The data set consists of 311 CT images corresponding to 311 patients. The images are manually segmented by experts from NYU langone for 9 OARs (brainstem, mandible, left and right optic nerves, optic chiasm, left and right parotid glands, left and right submandibular glands). At the same time, 3 new OARs (PCM-Superior, PCM-Middle and PCM-Inferior) are annotated for 20 patients. Their region distributions account for 0.5%, 0.3%, 0.8% respectively in average of the whole input volumne, indicating their small size. The dataset is split into 80/20 percent split.

3.2 Network Architecture

The overview of the proposed training framework is illustrated in Figure 2 and Figure 3 . The given CT scan volume was first cropped into uniform size of (96,128,224) centering in the region of interest and then augmented through random affine and random elastic deformation for the input volume. The input volume was fed to model of our choice (3D Res-Unet and UNEt TRansformers). Two models were trained separately. I applied transfer learning here through locking weights dynamically from pre-trained models of different tasks or the same task with different datasets.

4 Experimental Evaluation

4.1 Methodology

This section discusses the methodologies that I use to achieve the final solution.

- Transfer Learning on 9 old OARs with the assist of pre-trained model from NYU langone Dataset of similar task.
- Transfer Learning on 3 new OARs with the assist of pre-trained model from NYU langone Dataset of different task.
- Introduce 95% hausdorff distance to evaluation metric together with dice score for more comprehensive analysis of segmentation results.

4.1.1 Loss Functions and Evaluation Metrics

I have used the combination of Weighted Cross Entropy, Generalized Dice Loss and Focal Loss as the training loss function due to their different focus on multi-task segmentation training. Weighted Cross Entropy helps the model locate the boundary of OARs fastly when Focal Loss helps the model weight more on "hard" tasks in the learning process. The generalized version of dice loss is able to alleviate the imbalanced data issue. The mean dice score performance is improved from 31.80% to 57.44% compared with using dice loss alone. Meanwhile, dynamic weighting of loss function is also used here to further deal with the class imbalance issue and strengthen the model's performance on small organs. Their weights are adjusted flexibly according to the model performance on validation set. More weight is assigned to Weighted Cross Entropy at the beginning stage to locate each OAR roughly. Afterwards, weight is shifted to Focal Loss to encourage the model focus more on hard tasks. With the help of dynamic weighting, the mean dice score performance is boosted to 73.24%. Dice Score and 95% hausdorff distance are used as my final evaluation metrics.

Weighted Cross Entropy[12] Cross Entropy is usually used for classification problems. The weighted variant of it can alleviate the data imbalance issue.

$$WeightedCrossEntropy = \begin{cases} -w\log p & y=1 \\ -\log(1-p) & otherwise \end{cases} ;\ w = c * \frac{N - \sum_n p_n}{\sum_n p_n} \tag{1}$$

note: w will be scaled by constant c such that weight for all classes sum up to 1.

Focal Loss[11] Focal loss is an refined version of cross entropy by weighting less on "easy" tasks and more on "hard" ones.

$$FocalLoss = \begin{cases} -(1-p)^{\gamma}\log p & y=1 \\ -(p)^{\gamma}\log(1-p) & otherwise \end{cases} ;\gamma = 2 \tag{2}$$

Dice Loss[10] Dice loss represents the volumetric overlap between two samples.

$$DiceLoss = 1 - \frac{2\sum_{pixels} p_i \cdot y_i}{\sum_{pixels} p_i + \sum_{pixels} y_i} \tag{3}$$

Generalized Dice Loss[12] Generalized Dice Loss is an extension of Dice Loss. Through considering pixel weight, it is useful in problem with imbalanced data.

$$GeneralizedDiceLoss = 1 - 2\frac{\sum_{labels} w_l \sum_{pixels} y_{li}p_{li}}{\sum_{labels} w_l \sum_{pixels} y_{li} + p_{li}} ; w_l = 1/\left(\sum_{pixels} y_{li}\right)^2 \tag{4}$$

95% Hausdorff Distance[7] The hausdorff distance is defined as the maximum distance between two structures. Figure 4 provides a vivid illustration. 95% hausdorff distance accounts for 95th percentile of the distances between X and Y. It aims to reduce the impact of small subset of the outliers.

$$d_H(X,Y) = max(d_{XY}, d_{YX}) = max(max_{x\in X} min_{y\in Y} d(x,y), max_{y\in Y} min_{x\in X} d(x,y)) \tag{5}$$

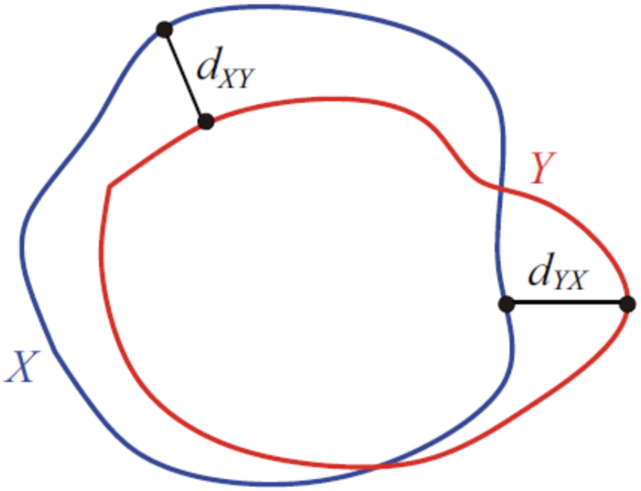


Figure 4: Hausdorff distance diagram

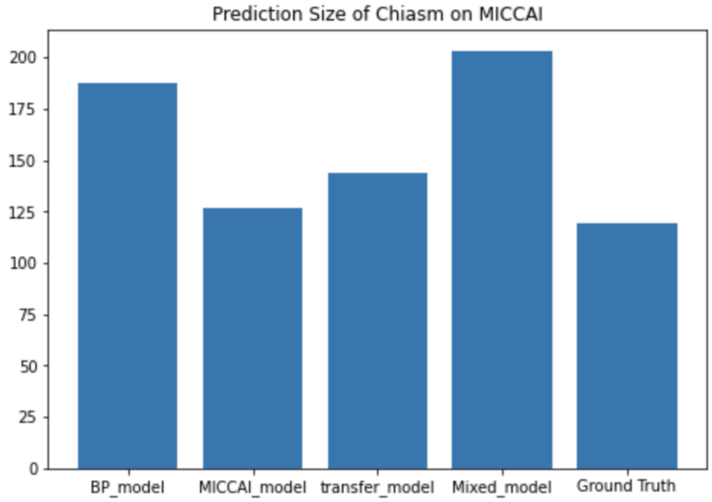


Figure 5: Size of Chiasm prediction on MICCAI test dataset. (Y-axis stands for the volume size of segmentation, BP-model stands for the model trained on NYU Langone Dataset, Mixed-model stand for the model trained on combination of NYU Langone Dataset and MICCAI Dataset)

4.1.2 Transfer Learning on old OARs

I constructed models from NYU Langone Dataset for the 9 OARs as pre-trained models using similar training framework with the selected two networks. The direct performance on the test set of MICCAI is relatively low compared with model trained directly on MICCAI dataset. For instance, the performance of segmentation on Chiasm dropped largely in terms of dice score and 95% hausdorff distance. Through further quantitative analysis on samples of Chiasm, it was caused by over predicting the region of Chiasm. Figure 5 shows us that, after applying transfer learning to the MICCAI dataset, the situation can be alleviated largely to obtain prediction of similar size to the ground truth. At the same time, performance of the other 8 OARs were all benefited similarly from transfer learning. The mean dice score of the best model was boosted to 77.1 while the mean 95% hausdorff distance of the best model was improved to 2.66 mm. The detailed performance is shown in Table 1 and Table 2.

Table 1: Dice score comparison on the test set of MICCAI 2015

	Res3DUnet (pre-trained)	Res3DUnet (transfer learning)	UNETR (transfer learning)	UNETR (pre-trained)
Chiasm	23.2	53.0	16.5	42.7
Mandible	83.8	**93.0**	92.2	82.2
BrainStem	86.0	**87.8**	84.6	80.5
OpticNerve-L	59.0	69.2	**69.6**	57.5
OpticNerve-R	56.4	**68.0**	67.4	53.2
Parotid-L	81.4	**85.8**	83.2	74.2
Parotid-R	80.6	**83.5**	79.4	71.7
Submandibular-L	72.4	**77.3**	70.9	57.5
Submandibular-R	72.0	**76.7**	73.0	57.9
Average	68.3	**77.1**	73.7	61.3

Table 2: 95% Hausdorff Distance (mm) comparison on the test set of MICCAI 2015

	Res3DUnet (pre-trained)	Res3DUnet (transfer learning)	UNETR (pre-trained)	UNETR (transfer learning)
Chiasm	6.06	**2.79**	6.95	4.41
Mandible	1.88	**1.29**	37.66	1.85
BrainStem	2.43	**2.17**	3.75	8.09
OpticNerve-L	3.42	**1.48**	5.58	**1.48**
OpticNerve-R	5.08	**1.29**	5.58	1.74
Parotid-L	7.38	4.99	10.85	**3.89**
Parotid-R	4.75	**3.34**	5.57	4.48
Submandibular-L	4.34	**3.29**	6.02	5.47
Submandibular-R	7.02	**3.28**	12.70	6.30
Average	4.71	**2.66**	10.52	4.14

4.1.3 Transfer Learning on new OARs

I also explored the use of transfer learning on completely new tasks with limited data. NYU Langone Dataset contains 3 new small OARS segmentation (PCM-Superior, PCM-Middle and PCM-Inferior) with 20 patients available. Following the same training methodology of previous 9 OARs, Res3DUnet performed superior to UNETER. Thus, I decided to focus on Res3DUnet only in this case. I tried various ways of transfer learning through modifying Res3DUnet for freezing different layers from pre-trained model, testing the effect of preserving different levels of features. The results are shown in Table 3 and Table 4. While only freezing the last few layers of the pre-trained model lead the model fail to converge, freezing the downward convolutional path of 3DUnet achieved the best performance in all three tasks. In terms of model structure, This method maintained the high level features of the pre-trained model. In addition, applying transfer learning can reduce the training time as well. Transfer learning converges around 110 epochs while training from scratch takes almost twice the time.

Table 3: Dice score comparison on the test set of new organs

	Res3DUnet (trained from sratch)	Res3DUnet (no freezing layer)	Res3DUnet (freezing the first half layer)
PCM-Superior	47.9	46.0	**53.8**
PCM-Middle	46.2	52.0	**57.5**
PCM-Inferior	48.3	53.7	**52.6**
Average	47.5	50.6	**54.6**

Table 4: 95% Hausdorff Distance (mm) comparison on the test set of MICCAI 2015

	Res3DUnet (trained from scratch)	Res3DUnet (no freezing layer)	Res3DUnet (freezing the first half layer)
PCM-Superior	3.15	3.29	**2.78**
PCM-Middle	4.97	4.47	**2.96**
PCM-Inferior	5.03	4.80	**4.83**
Average	4.38	4.19	**3.76**

Table 5: Dice score comparison on the test set of MICCAI 2015

	Capstone Model	Res3DUnet (transfer learning)	MICCAI 2015	FocusNet
Chiasm	50.2	53.0	55.7	**59.6**
Mandible	93.8	93.0	93.0	**93.5**
BrainStem	83.8	87.8	**88.0**	87.5
OpticNerve-L	67.6	69.2	64.4	**73.5**
OpticNerve-R	68.7	68.0	63.9	**74.4**
Parotid-L	82.9	85.8	82.7	**86.3**
Parotid-R	82.0	83.5	81.4	**87.9**
Submandibular-L	72.4	77.3	72.3	**79.8**
Submandibular-R	74.0	76.7	72.3	**80.1**
Average	75.0	77.1	74.9	**80.3**

4.2 Results

All models are trained until the validation performance does not improve after 50 consecutive epochs.

Comparison with State-of-the-Art Table 5 and Table 6 compares our models with the winner of MICCAI 2015 Challenge[3] and the current best state-or-art work: FocusNet[9]. Our transfer learning model is able to outperform in most of the tasks compared with the winner of MICCAI 2015 and my last year capstone model. Table 7 and Table 8 presents our results with available deep learning solutions to 3 new organs. My transfer learning solution show superiority in most evaluation metrics.

4.3 Discussion

Training giant neural networks like 3D U-Net and its variant for multiple classes segmentation is tricky, because it not only needs careful design of training methodology to deal with class imbalance issue, but also requires large amounts of data and time to train. My previous proposed training

Table 6: 95% Hausdorff Distance (mm) comparison on the test set of MICCAI 2015

	Capstone Model	Res3DUnet (transfer learning)	MICCAI 2015	FocusNet
Chiasm	3.04	2.79	**2.78**	3.16
Mandible	1.64	1.29	1.97	**1.18**
BrainStem	2,34	2.17	4.59	**2.14**
OpticNerve-L	1.52	**1.48**	2.76	3.76
OpticNerve-R	1.46	**1.29**	3.15	2.65
Parotid-L	5.71	4.99	5.11	**2.52**
Parotid-R	3.87	3.34	6.13	**2.07**
Submandibular-L	4.19	3.29	5.35	**2.67**
Submandibular-R	4.23	**3.28**	5.42	3.41
Average	3.11	2.66	4.14	**2.62**

Table 7: Dice score comparison on the test set of new OARs

	Stoyanov[14]	Gan[15]	Res3DUnet
PCM-Superior	**59**	30	53.8
PCM-Middle	48	30	**57.5**
PCM-Inferior	42	40	**52.6**
Average	49.7	33.3	**54.6**

Table 8: 95% Hausdorff Distance (mm) comparison on the test set of new OARs

	Gan[15]	Res3DUnet
PCM-Superior	7.5	**2.78**
PCM-Middle	12.5	**2.96**
PCM-Inferior	5	**4.83**
Average	8.33	**3.76**

methodology with various loss functions and dynamic weighting techniques is able to alleviate class imbalance issue largely. Transfer learning with pre-trained model of similar tasks can help us alleviate the issue of limited data. From Table 5 and Table 6, my best single model solution is able to outperform both the winner of MICCAI 2015 and my previous capstone ensemble model of 3DResNet and UNETER in both evaluation metrics in most of the tasks. We can see that my single model solution is a bit lower than the current state-of-the-art work (77.1 vs 80.3 in dice score, 2.66 vs 2.62 in 95% hausdorff distance). However, my model outperforms in certain regions like OpticNerve-L, OpticNerve-R, and Submandibular-R in terms of hausdorff distance, which intuitively means that our prediction is closer to the region of ground truth. Also, note that FocusNet is a complex model with multiple networks embedded while my solution is single model solution which is much easier to train.

Transfer learning with the pre-trained model of different tasks is beneficial in training new organs with limited data as well. With the assist of global scope features from the pre-trained model, adding as few as 12 cases for new organs can make my model achieve better performance than training from scratch (47.5 vs 54.6 in mean dice score, 4.38 vs 3.76 in hausdorff distance) and obtain superior results to available models. (Table 7 and Table 8) The training time is shortened significantly as well. The visualization of results of all OARs is shown in the Appendix.

5 Conclusions

In conclusion, I propose an end-to-end transfer learning single model solution with Res3DUnet, which achieves similar performance with the FocusNet (state-of-the-art work with complex model of multiple neural networks), and saves training time largely. I also leverage transfer learning to 3 new OARs tasks in order to tackle the limited data issue, boosting the performance to outperform the state-of-the-art work.

In the future, adjusting the inner structure of 3D U-Net for head and neck segmentation could be interesting in order to be more focused on the tasks. This work has shown the importance of having access to a large pre-trained model with relatively similar tasks to improve the model performance. Medical segmentation datasets other than head and neck segmentation could be taken into consideration to form a more giant pre-trained model with the potential for further improvement. In addition, I believe that more tailored use of pre-trained models and novel ways of using transfer learning can also be helpful.

Acknowledgement

I would like to thank Doctor Ye Yuan for mentoring and guiding me throughout the project. I also want to thank Doctor Narges Razavian, Doctor David Barbee, Doctor Ting Cheng for helping us in the project.

References

[1] Ding & Zhang (2021) "Deep Learning for Head and Neck CT Image Segmentation".

[2] Çiçek, Özgün, et al. "3D U-Net: learning dense volumetric segmentation from sparse annotation." *International conference on medical image computing and computer-assisted intervention*. Springer, Cham, 2016.

[3] Raudaschl, P. F., Zaffino, P., Sharp, G. C., Spadea, M. F., Chen, A., Dawant, B. M., ... Jung, F. (2017).Evaluation of segmentation methods on head and neck CT: Auto-segmentation challenge 2015.Medical Physics, 44(5), 2020-2036.

[4] He, Kaiming, et al. "Deep residual learning for image recognition." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

[5] Hatamizadeh, Ali, et al. "Uneter: Transformers for 3d medical image segmentation." *arXiv preprint arXiv:2103*.10504 (2021).

[6] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." *International Conference on Medical image computing and computer-assisted intervention*. Springer, Cham, 2015.

[7] D. P. Huttenlocher, G. A. Klanderman and W. J. Rucklidge, "Comparing images using the Hausdorff distance," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 15, no. 9, pp. 850-863, Sept. 1993, doi: 10.1109/34.232073.

[8] Raudaschl, Patrik F., et al. "Evaluation of segmentation methods on head and neck CT: auto-segmentation challenge 2015." *Medical physics* 44.5 (2017): 2020-2036.

[9] Gao, Yunhe, et al. "Focusnet: Imbalanced large and small organ segmentation with an end-to-end deep neural network for head and neck ct images." *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, Cham, 2019.

[10] Milletari, Fausto, Nassir Navab, and Seyed-Ahmad Ahmadi. "V-net: Fully convolutional neural networks for volumetric medical image segmentation." *2016 fourth international conference on 3D vision (3DV)*. IEEE, 2016.

[11] Lin, Tsung-Yi, et al. "Focal loss for dense object detection." *Proceedings of the IEEE international conference on computer vision*. 2017.

[12] Sudre, Carole H., et al. "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations." *Deep learning in medical image analysis and multimodal learning for clinical decision support*. Springer, Cham, 2017. 240-248.

[13] Torrey, Lisa, and Jude Shavlik. "Transfer learning." Handbook of research on machine learning applications and trends: algorithms, methods, and techniques. IGI global, 2010. 242-264.

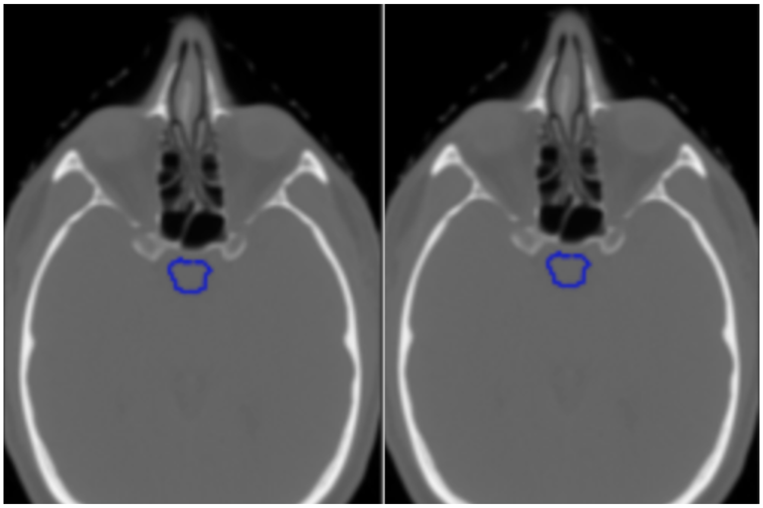
[14] Stoyanov, Danail, et al., eds. Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings. Vol. 11045. Springer, 2018.

[15]Gan, Yong, et al. "A novel semi auto-segmentation method for accurate dose and NTCP evaluation in adaptive head and neck radiotherapy." Radiotherapy and Oncology 164 (2021): 167-174.

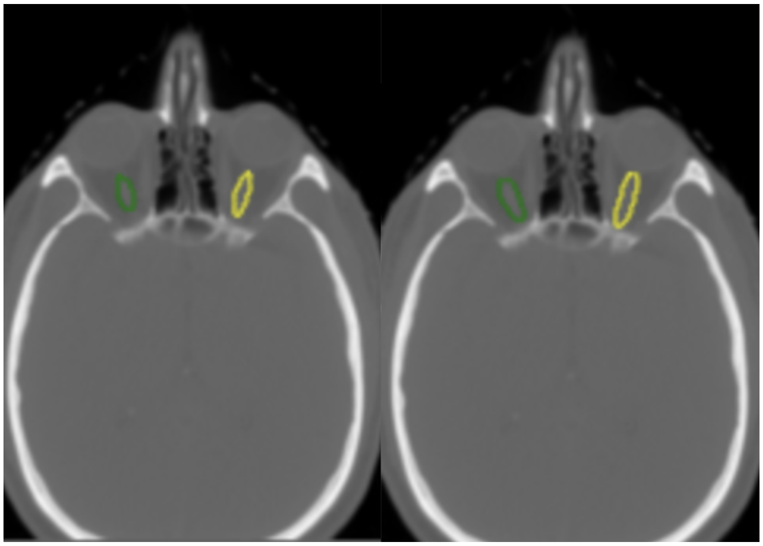
[16]Vu, Charles C., et al. "Deep convolutional neural networks for automatic segmentation of thoracic organs-at-risk in radiation oncology–use of non-domain transfer learning." Journal of Applied Clinical Medical Physics 21.6 (2020): 108-113.

Appendices

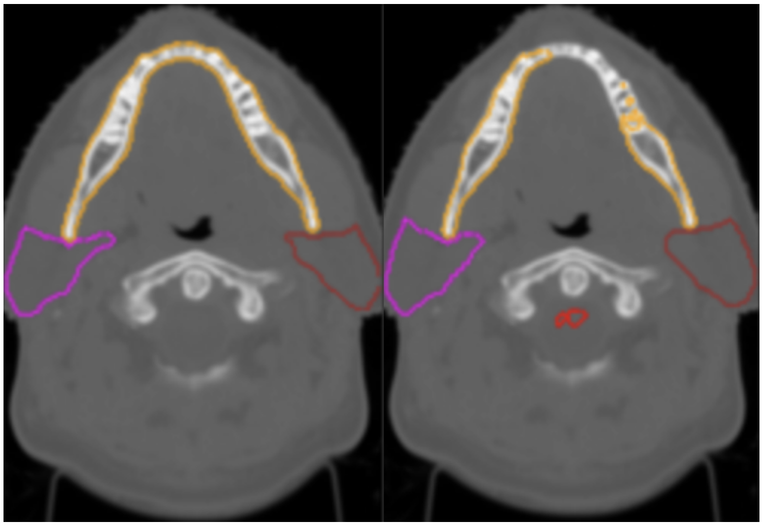
Ground-truth (left) and Prediction (right)



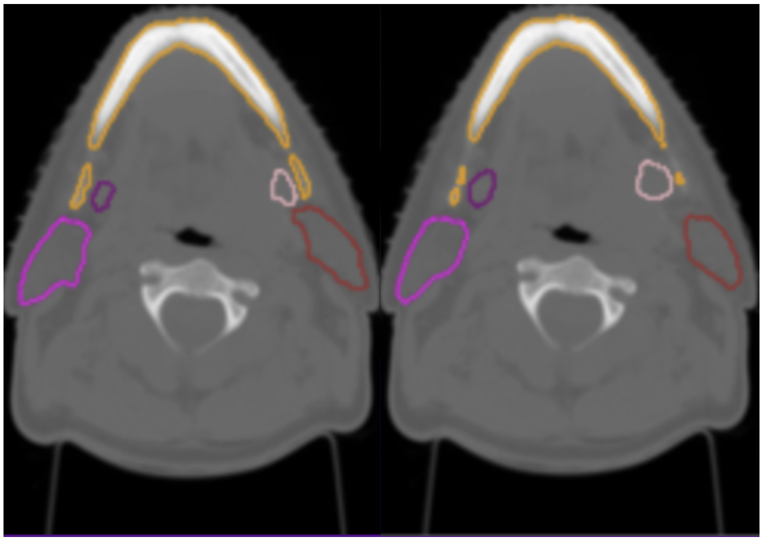
OpticChiasm (blue)
53% (mean dice score)



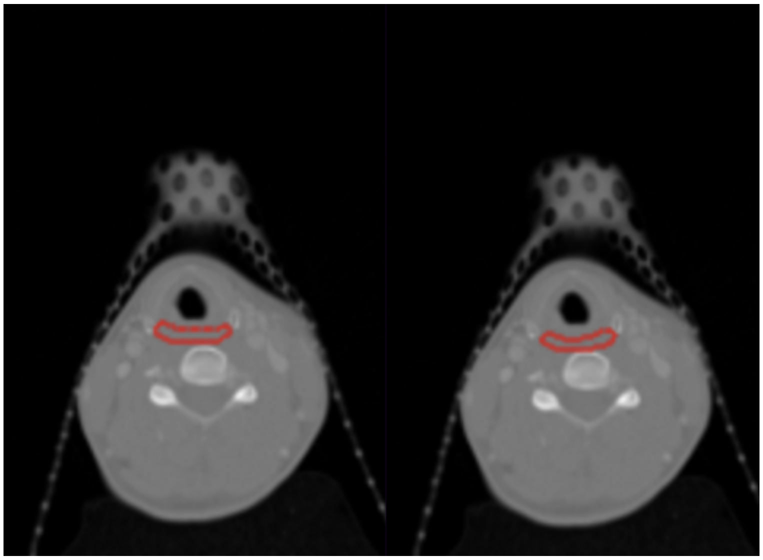
OpticNerve Left (green), OpticNerve Right (yellow)
69.2 % (mean dice score),68.0 % (mean dice score)



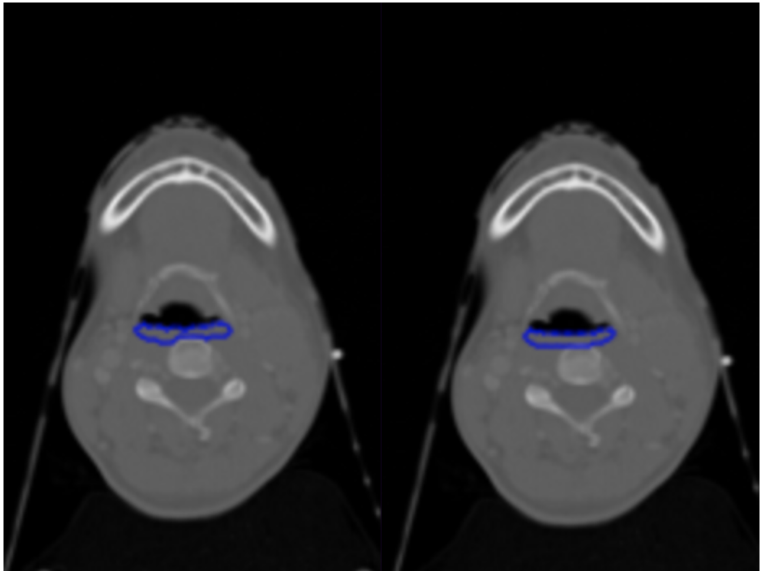
Parotid Left (margenta), Parotid Right (brown), BrainStem (red), Mandible (orange)
85.8 % (mean dice score),83.5 % (mean dice score), 87.8 % (mean dice score),93.0 % (mean dice score)



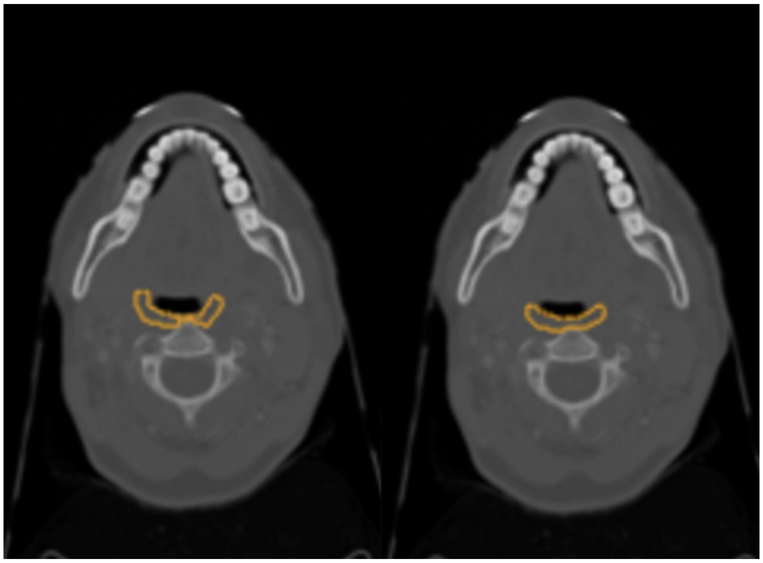
Mandible (orange), Submandibula Left (purple), Submandibula Right (pink),Parotid Left (margenta), Parotid Right (brown)
93.0 % (mean dice score),77.3 % (mean dice score),76.7 % (mean dice score),85.8 % (mean dice score),83.5 % (mean dice score)



PCM-Inferior(red)
52.6 % (mean dice score)



PCM-Middle(blue)
57.5 % (mean dice score)



PCM-Superior(organe)
53.8 % (mean dice score)