

University of Groningen

Swin UNETR for Tumor and Lymph Node Segmentation Using 3D PET/CT Imaging

Chu, Hung; De la O Arévalo, Luis Ricardo; Tang, Wei; Ma, Baoqiang; Li, Yan; De Biase, Alessia; Both, Stefan; Langendijk, Johannes Albertus; van Ooijen, Peter; Sijtsma, Nanna Maria

Published in:

Head and Neck Tumor Segmentation and Outcome Prediction - 3rd Challenge, HECKTOR 2022, Held in Conjunction with MICCAI 2022, Proceedings

DOI:

[10.1007/978-3-031-27420-6_12](https://doi.org/10.1007/978-3-031-27420-6_12)

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version

Publisher's PDF, also known as Version of record

Publication date:

2023

[Link to publication in University of Groningen/UMCG research database](#)

Citation for published version (APA):

Chu, H., De la O Arévalo, L. R., Tang, W., Ma, B., Li, Y., De Biase, A., Both, S., Langendijk, J. A., van Ooijen, P., Sijtsma, N. M., & van Dijk, L. V. (2023). Swin UNETR for Tumor and Lymph Node Segmentation Using 3D PET/CT Imaging: A Transfer Learning Approach. In V. Andrearczyk, V. Oreiller, A. Depaeys, & M. Hatt (Eds.), *Head and Neck Tumor Segmentation and Outcome Prediction - 3rd Challenge, HECKTOR 2022, Held in Conjunction with MICCAI 2022, Proceedings* (pp. 114-120). (Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Vol. 13626 LNCS). Springer Science and Business Media Deutschland GmbH. Advance online publication. https://doi.org/10.1007/978-3-031-27420-6_12

Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.












Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.



Swin UNETR for Tumor and Lymph Node Segmentation Using 3D PET/CT Imaging: A Transfer Learning Approach

Hung Chu^(✉) , Luis Ricardo De la O Arévalo^(✉) , Wei Tang^(✉) ,
Baoqiang Ma^(✉) , Yan Li^(✉) , Alessia De Biase^(✉) , Stefan Both^(✉) ,
Johannes Albertus Langendijk^(✉) , Peter van Ooijen^(✉) ,
Nanna Maria Sijtsma^(✉) , and Lisanne V. van Dijk^(✉) 

University Medical Center Groningen (UMCG), 9700 RB Groningen, The Netherlands
{d.h.chu,l.r.de.la.o.arevalo,w.tang,b.ma,y.li05,a.de.biase,s.both,
j.a.langendijk,p.m.a.van.ooijen,n.m.sijtsma,l.v.van.dijk}@umcg.nl

Abstract. Delineation of Gross Tumor Volume (GTV) is essential for the treatment of cancer with radiotherapy. GTV contouring is a time-consuming specialized manual task performed by radiation oncologists. Deep Learning (DL) algorithms have shown potential in creating automatic segmentations, reducing delineation time and inter-observer variation. The aim of this work was to create automatic segmentations of primary tumors (GTV_p) and pathological lymph nodes (GTV_n) in oropharyngeal cancer patients using DL. The organizers of the HECKTOR 2022 challenge provided 3D Computed Tomography (CT) and Positron Emission Tomography (PET) scans with ground-truth GTV segmentations acquired from nine different centers. Bounding box cropping was applied to obtain an anatomic based region of interest. We used the Swin UNETR model in combination with transfer learning. The Swin UNETR encoder weights were initialized by pre-trained weights of a self-supervised Swin UNETR model. An average Dice score of 0.656 was achieved on a test set of 359 patients from the HECKTOR 2022 challenge. Code is available at: https://github.com/HC94/swin_unetr_hecktor_2022.

Aicrowd Group Name: RT_UMCG

Keywords: Head and neck cancer · Deep learning · Swin UNETR · HECKTOR 2022 · Radiotherapy · Tumor segmentation · Lymph node segmentation · Auto contouring · Image processing

1 Introduction

Head and neck cancers (HNC) are among the most common worldwide (5th leading cancer by incidence) [8]. Radiation therapy (RT) is pivotal in the treatment of HNC patients, however more than one out of four of all HNC patients in Europe did not receive RT due to limited trained personnel and equipment [9]. Accurate delineation of the tumor contour is important for delivering high dose

in the tumor area without damaging surrounding normal tissues. However, delineation of the tumor contour is usually performed by experts and is susceptible to inter-observer variability. Treatment planning would benefit from automatic analysis of medical imaging data, as automatic segmenting tumors can reduce delineation time and interobserver variability.

In recent years, Deep Learning (DL) models have shown to be great potential for the medical field. More specifically, DL-based algorithms using fluorodeoxyglucose (FDG) Positron Emission Tomography (PET) and Computed Tomography (CT) as inputs have been explored in previous HECKTOR challenges for auto-segmenting GTV contour of the primary tumor, herewith showing promising results in terms of Dice scores [1, 2].

The aim of this paper is to segment Head and Neck (H&N) primary tumors and lymph nodes in FDG-PET/CT images using a DL algorithm. We propose the Swin UNETR model, which showed top performance results for 3D semantic segmentation of brain tumors in Magnetic Resonance Imaging (MRI) images [4]. Furthermore, we performed transfer learning by using weights from a self-supervised Swin UNETR model [10].

2 Methods and Materials

2.1 Data

The training dataset available consisted of 524 HNC patients with histologically proven oropharyngeal cancer who underwent radiotherapy and/or radiochemotherapy treatment planning. The data was collected from seven different medical centers and provided by the organizers of the HECKTOR (HEAd and neCK TumOR) 2022 challenge [1, 7] (Table 1). For each patient a 3D FDG-PET scan, a 3D CT scan and the GTV_p and GTV_n segmentations (RTSTRUCT) were available. The GTV_p and GTV_n contours, used as ground-truth during training, were manually delineated by an annotator and cross checked by another annotator. Delineation guidelines were elaborated to ensure unification. The FDG-PET and low-dose non-contrast-enhanced CT images were acquired with combined PET/CT scanners. The independent test set (i.e. not used in model training) was a cohort of 359 HNC patients with FDG-PET and CT scans collected from three different centers (Table 1).

Table 1. Number of patients from each center.

	CHUM	CHUP	CHUS	CHUV	MDA	HGJ	HMR	USZ	CHB	<i>Total</i>
Training	56	72	72	53	198	55	18	0	0	524
Test	0	0	0	0	200	0	0	101	58	359

All files were provided in Nifti format. More information about medical data centers, scanners and data availability can be found at the following link: <https://hecktor.grand-challenge.org/Data/>.

2.2 Data Preprocessing

As data was collected by different centers, we preprocessed the data to obtain unification and adapted it to the input type required by our model. Firstly, we resampled the FDG-PET, CT and segmentations to an isotropic voxel spacing of $1 \times 1 \times 1 \text{ mm}^3$. The FDG-PET and CT scans were resampled with spline interpolation of degree 3, and the segmentations with nearest neighbor interpolation. Then we cropped a bounding box region using the automatic bounding box extraction algorithm from last year’s HECKTOR challenge [3]: firstly, the brain is detected as the largest component containing SUV larger than 3, and secondly a rigid sized bounding box was placed at anatomic midpoints voxels in the x and y -axis, and at the lowest brain voxel in the z -axis. To increase the field of view, we increased the bounding box size from $144 \times 144 \times 144$ (*height* (H) \times *width* (W) \times *depth* (D)) to $192 \times 192 \times 192$. The FDG-PET and CT intensity values were expressed in Standard Uptake Value (SUV) and Hounsfield Units (HU) respectively, and were clipped between $[0, 25]$ SUV and $[-200, 400]$ HU. Lastly, we normalized the values to $[0, 1]$ as per $x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}}$, where $x_{min} = 0$, $x_{max} = 25$ SUV for the PET modality and $x_{min} = -200$, $x_{max} = 400$ HU for the CT modality.

2.3 Model Architecture

We used the Swin UNETR model [10]. An overview of the original Swin UNETR model architecture is depicted in Fig. 1. Firstly, the model projected the multi-modal input data into a 1D embedding sequence. Secondly, the embedding sequence was used as input for the Swin UNETR encoder, which was composed of a stack of Swin Transformer blocks. The output of each block was used in a U-Net style.

We tailored the Swin UNETR model for our task. Our model accepted FDG-PET and CT as inputs with combined size $96 \times 96 \times 96 \times 2$, and generated a segmentation map of size $96 \times 96 \times 96 \times 3$ for background, GTVp and GTVn combined. The Swin UNETR encoder weights were initialized with pre-trained weights from a self-supervised Swin UNETR encoder that was trained on a cohort of 5050 CT scans from publicly available datasets [10]. The encoder was pre-trained on three different tasks: inpainting, contrastive learning and rotation prediction. To be able to use the pre-trained encoder weights, we used the same embedding size as the pre-trained model (i.e. 48 features). The reason for this is that an embedding of size 48 was used as input by the pre-trained encoder.

An ensemble model was created from the seven models of a 7-fold cross-validation (CV) (‘leave-one-center-out’-approach): one model from each CV fold. More specifically, for each patient in the test set we averaged the probability segmentation map of the seven models, and then discretized by applying arguments of the maxima ($\arg \max$) to obtain the ensemble segmentation map.

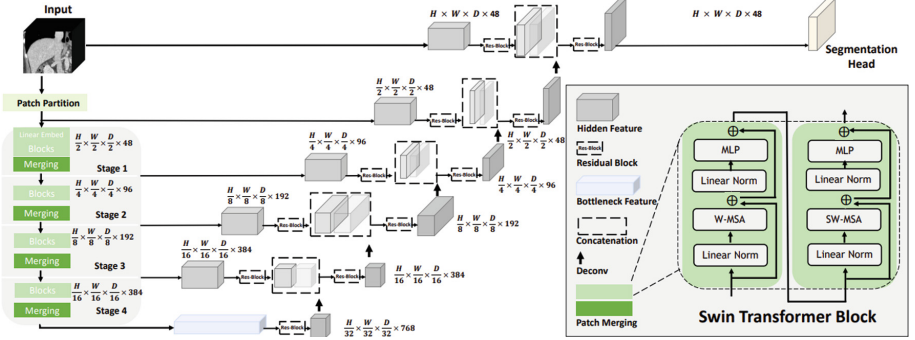


Fig. 1. Overview of the Swin UNETR model architecture, taken with permission from [10]. Firstly, the input data was projected into a 1D embedding sequence, which consequently was used as input for the encoder. A linear embedding of 48 features was used, same as the pre-trained model, to be able to use the pre-trained encoder weights.

2.4 Experiments

In each training iteration we randomly selected two fixed sized regions of size $96 \times 96 \times 96$ from a full input of size $192 \times 192 \times 192$. Since the majority of background voxels imbalanced the data, we selected the regions such that half of the all selected regions had a foreground (either from GTVp or GTVn) voxel in the center of the region, and the other half had a background voxel in the center of the region. The cropped regions were used as input batch of size two for the model. For model inferences we performed a sliding window approach as depicted in Fig. 2.

The model was trained for 200 epochs using the Dice + Cross-Entropy (DiceCE) loss function and the AdamW optimizer [5], and validated on the held-out fold with multi-class mean Dice score. The model weights were saved at the epoch with the highest validation score. The learning rate was updated using cosine annealing schedule with warm restarts [6]. Data augmentation techniques were adopted such as random translation, zooming, flipping, rotating and intensity shifting¹. Each data augmentation technique was independently applied with 0.5 probability. The data augmentation and modeling were implemented using Project MONAI 0.9 in PyTorch 1.10. A comprehensive list of all training methodology is summarized in Table 2. The experiments were conducted on NVIDIA V100 GPU with 32 GB GPU memory.

2.5 Quantitative Evaluation

The experimental results were evaluated in terms of the aggregated Dice similarity coefficient $DSC_{agg} = \frac{2 \sum_i^N \sum_k \hat{y}_{i,k} \cdot y_{i,k}}{\sum_i^N \sum_k (\hat{y}_{i,k} + y_{i,k})}$, where N is the total number of test images, $y_{i,k}$ is the ground truth (either GTVp or GTVn) for voxel k of image i , and $\hat{y}_{i,k}$ is the prediction.

¹ <https://docs.monai.io/en/stable/transforms.html>.

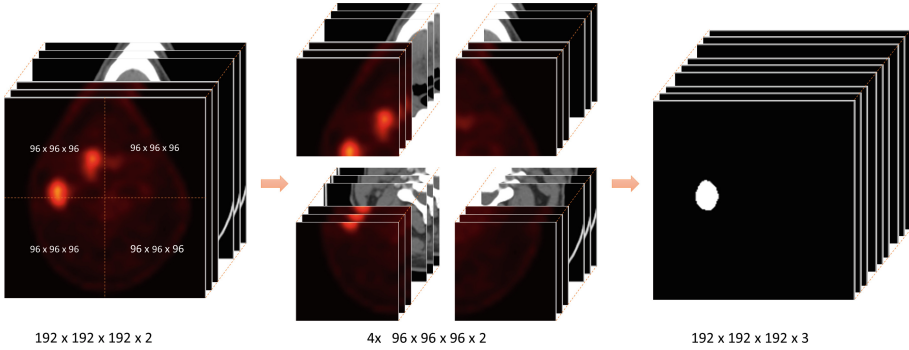


Fig. 2. Sliding window approach. Firstly, divide the full input of size $192 \times 192 \times 192 \times 2$ uniformly into four windows of size $96 \times 96 \times 96 \times 2$. Secondly, perform model inference on each windows. Finally, aggregate the output into a single segmentation map of size $192 \times 192 \times 192 \times 3$ for background, GTVp and GTVn combined.

Table 2. Training methodology and hyper-parameters.

Component	Value
Epochs	200
Batch size	2
Initial learning rate	$1e^{-4}$
Loss function	DiceCE
Optimizer	AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$)
Scheduler	Cosine ($T_0 = 40$)
Weight decay	$1e^{-5}$
Data augmentation	Translating $[-10, 10]$, zooming $[90, 110]\%$, flipping, rotating $[-180^\circ, 180^\circ]$, intensity shifting $[-0.1, 0.1]$

3 Results

Table 3 presents the results of the 7-fold CV for each fold separately as well as the average over all folds. We observe that the DSC_{agg} of GTVp is always higher than that of GTVn. Moreover, the scores can differ significantly across folds.

For this challenge we submitted predictions from the ensemble model and the model from fold 5, which had the highest average DSC_{agg} in CV. Table 4 presents the results on the test set. The average Dice of GTVp and GTVn contouring is higher on the test set than the average CV result. Interestingly, we observed a significantly lower Dice score of GTVp segmentation on the test set than in CV, and the opposite for GTVn segmentation. Furthermore, the performance of the ensemble and fold 5 model are similar.

Table 3. Evaluation performance of each CV fold during training as well as the average over all folds.

	$DSC_{agg} \text{ GTV}_p$	$DSC_{agg} \text{ GTV}_n$	$\frac{DSC_{agg} \text{ GTV}_p + DSC_{agg} \text{ GTV}_n}{2}$
Fold 1	0.675	0.600	0.638
Fold 2	0.753	0.504	0.629
Fold 3	0.613	0.580	0.596
Fold 4	0.711	0.536	0.623
Fold 5	0.752	0.583	0.667
Fold 6	0.688	0.587	0.637
Fold 7	0.758	0.419	0.589
<i>Average</i>	0.707	0.582	0.626

Table 4. Evaluation performance of the ensemble model and the model from fold 5 on the test set.

	$DSC_{agg} \text{ GTV}_p$	$DSC_{agg} \text{ GTV}_n$	$\frac{DSC_{agg} \text{ GTV}_p + DSC_{agg} \text{ GTV}_n}{2}$
Ensemble	0.642	0.670	0.656
Fold 5	0.633	0.673	0.653

4 Discussion and Conclusion

In this paper we proposed to use Swin UNETR model in conjunction with transfer learning. We combined FDG-PET and CT images into a single input for our end-to-end model.

The self-supervised pre-trained Swin UNETR model was trained on CT scans and no FDG-PET imaging. Therefore the pre-trained weights may not be helpful for the FDG-PET modality. Another limitation is the computational time: training the model for 200 epochs for a single CV fold iteration took about two days. Therefore we did not do any hyperparameter tuning. With one GPU we recommend to apply training-validation split instead of CV and train for more than 200 epochs, because the validation performance in the CV iterations was still improving at 200 epochs.

Also, we observed unexpected values in the provided training and test data. The CT input intensity values (i.e. after cropping the bounding boxes and resampling) in the training and test data was in the interval $[-17.200, 32.636]$ and $[-10.223, 38.010]$, respectively. However, CT intensity values in HU should be in the interval $[-1.024, 3.000]$. In fact, only 6% of the training patients complied with that, and 66% of the training patients had CT intensity value in $[-4.000, 4.000]$. This holds similarly for the test data. These findings suggest that most CT scans were not represented in HU. Different data normalization techniques should have been applied to obtain data unification. We did not deal with these issues due to late discovery of these issues and therefore lack of time.

The ensemble and fold 5 model have similar Dice scores on the test set. On top of that, DSC_{agg} of GTVp is higher than that of GTVn in CV, while this is opposite for the test set. Therefore we suspect that the data across centers may differ significantly, possibly due to different PET/CT scanners, and require additional data preprocessing to obtain data unification.

For future work we suggest to use pre-trained weights trained on FDG-PET imaging, perform hyperparameter tuning, train for more than 200 epochs, and improve uniformity of the data modalities.

Acknowledgements. We thank the Center for Information Technology of the University of Groningen for their support and for providing access to the Peregrine high performance computing cluster.

References

1. Andrearczyk, V., et al.: Overview of the HECKTOR challenge at MICCAI 2022: automatic head and neck tumor segmentation and outcome prediction in PET/CT. In: Andrearczyk, V., Oreiller, V., Hatt, M., Depeursinge, A. (eds.) HECKTOR 2022. LNCS, vol. 13626, pp. 1–30. Springer, Cham (2023)
2. De Biase, A., et al.: Skip-SCSE multi-scale attention and co-learning method for oropharyngeal tumor segmentation on multi-modal PET-CT images. In: Andrearczyk, V., et al. (eds.) HECKTOR 2021. LNCS, vol. 13209, pp. 109–120. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-98253-9_10
3. Andrearczyk, V., Oreiller, V., Depeursinge, A.: Oropharynx detection in PET-CT for tumor segmentation. In: Irish Machine Vision and Image Processing (2020)
4. Hatamizadeh, A., et al.: Swin UNETR: swin transformers for semantic segmentation of brain tumors in MRI images (2022). <https://doi.org/10.48550/arXiv.2201.01266>
5. Loshchilov, I., et al.: Decoupled weight decay regularization (2017). <https://doi.org/10.48550/arXiv.1711.05101>
6. Loshchilov, I., et al.: SGDR: stochastic gradient descent with warm restarts (2016). <https://doi.org/10.48550/arXiv.1608.03983>
7. Oreiller, V., et al.: Head and neck tumor segmentation in PET/CT: the HECKTOR challenge. *Med. Image Anal.* **77**, 102336 (2022)
8. Parkin, D.M., et al.: Global cancer statistics, 2002. *CA Cancer J. Clin.* **55**(2), 74–108 (2005)
9. Lievens, Y.: Provision and use of radiotherapy in Europe. *Mol. Oncol.* **14**(7), 1461–1469 (2020). <https://doi.org/10.1002/1878-0261.12690>
10. Tang, Y., et al.: Self-supervised pre-training of swin transformers for 3D medical image analysis (2021). <https://doi.org/10.48550/arXiv.2111.14791>