Segmentation and Classification of Head and Neck Nodal Metastases and Primary Tumors in PET/CT

Vincent Andrearczyk[1], Valentin Oreiller[1,2], Mario Jreige[2], Joël Castelli[3], John O. Prior[2] and Adrien Depeursinge[1,2]

***Abstract*— The prediction of cancer characteristics, treatment planning and patient outcome from medical images generally requires tumor delineation. In Head and Neck cancer (H&N), the automatic segmentation and differentiation of primary Gross Tumor Volumes (GTVt) and malignant lymph nodes (GTVn) is a necessary step for large-scale radiomics studies to predict patient outcome such as Progression Free Survival (PFS). Detecting malignant lymph nodes is also a crucial step for Tumor-Node-Metastases (TNM) staging and to support the decision to resect the nodes. In turn, automatic TNM staging and patient outcome prediction can greatly benefit patient care by helping clinicians to find the best personalized treatment. We propose the first model to automatically individually segment GTVt and GTVn in PET/CT images. A bi-modal 3D U-Net model is trained for multi-class and multi-components segmentation on the multi-centric HECKTOR 2020 dataset containing 254 cases. The dataset has been specifically re-annotated by experts to obtain ground truth GTVn contours. The results show promising segmentation performance for the automation of radiomics pipelines and their validation on large-scale studies for which manual annotations are not available. An average test Dice Similarity Coefficients (DSC) of 0.717 is obtained for the segmentation of GTVt. The GTVn segmentation is evaluated with an aggregated DSC to account for the cases without GTVn, which is estimated at 0.729 on the test set.**

I. Introduction

Head and Neck (H&N) lymph nodes contain prognostically relevant information in PET/CT that can be used in radiomics analyses to predict patient outcomes such as Progression Free Survival (PFS). While tissue metabolism can be high in both the primary tumor (GTVt) and the lymph nodes (GTVn) when observed in PET, radiomics prognostic models should use distinct visual biomarkers from these two Volumes Of Interest (VOI) types as different tissular and metabolic profiles are expected. Both GTVt and GTVn carry complementary information for the prediction of patient outcome and staging. In clinical routine, radiologists and nuclear physicians will base decisions regarding treatment planning using both the GTVt and GTVn (lymph node positivity). Biopsy remains the gold standard to determine the malignancy of lymph nodes and is requested, for instance, for small primary tumors with suspicious lymph nodes. The multi-label segmentation of GTVt and GTVn is therefore a crucial step for the automation of patient outcome prediction to improve personalized medicine in H&N cancer. Segmenting GTVn also allows tackling the challenge of automatic Tumor-Node-Metastases (TNM) staging and could provide decision support for the resection of malignant lymph nodes.

In [1], [2], the authors proposed the segmentation, in PET/CT images of patients with H&N cancer, of GTVt and GTVn without distinction between the two VOI types (i.e. single label versus background). While this work was an important first step, its clinical relevance is limited as no differentiation is made between the two types of tumorous volumes. Various deep learning models were proposed in the context of the two editions of the HECKTOR challenges (2020 [3], [4] and 2021 [5]) to segment the GTVt only. The relevance of this type of automatic segmentation was revealed in [6], where radiomics features were extracted from automatically segmented GTVt, leading to promising results for fully automatic PFS prediction. In [7] and [8], multi-task CNNs were proposed for the prediction of PFS in H&N PET/CT images, guiding this task with the auxiliary task of GTVt segmentation. The existing literature specifically related to GTVn segmentation is limited. The automatic segmentation has been proposed in other regions including axillary and supraclavicular tumoral lymph nodes in [9], and thoracic lymph nodes in [10].

We propose the first method to segment GTVn and GTVt from PET/CT images in H&N patients, while being able to distinguish between the two types of tumoral volumes. Our approach is based on a bi-modal 3D U-Net model trained with a combination of standard Dice loss and a specific aggregated Dice loss to accommodate multi-class and multi-components GTVn segmentation. A similar batch-aggregated Dice loss was used in [11] for cardiac aorta and brain tumor segmentation to maintain a smooth, continuous and non-constant loss when facing no-target labels.

II. Methods

A. Dataset

We use the HECKTOR 2020 dataset [3], [4], which was originally released for the development of automatic GTVt segmentation from PET/CT images in the oropharyngeal region. Part of the data originates from [12]. The data were collected from five centers, of which four are used for the training, one for testing. An expert re-annotated this data to obtain high-quality annotations of both GTVt and GTVn,

This work was partially supported by the Swiss National Science Foundation (SNSF, grant 205320_179069), the Swiss Personalized Health Network (SPHN, via the IMAGINE project) and the Hasler Foundation.

[1]Institute of Information Systems, University of Applied Sciences Western Switzerland (HES-SO), Sierre, Switzerland.

[2]Department of Nuclear Medicine and Molecular Imaging, Centre Hospitalier Universitaire Vaudois (CHUV), Lausanne, Switzerland.

[3]Department of Radiation Oncology, Cancer Institute, Rennes, France, France.

which is detailed in [3]. Lymph nodes are considered malignant if they are pathologically confirmed, have a SUV greater than 2.5, or have a diameter greater than 1cm. The training and test split proposed in the HECKTOR 2020 challenge are used in the experiments. After splitting the training data for validation (60%, 40%), the training set contains 119 cases and the validation set 82 cases. The test set contains 52 cases. All of the test cases contain GTVt, but only 44 cases contain at least one GTVn. The average number of GTVn volumes (i.e. connected components) per case on the entire dataset is 1.8 (40 cases with 0, 83 with 1 and 130 with more than 1). The images are resampled using a trilinear interpolation to a 1mm^3 isotropic voxel size and further cropped using bounding boxes of size $144 \times 144 \times 144$ provided with the data and automatically obtained as described in [13]. Z-score normalization is applied to the PET images and the CT images are clipped in [-300, 300] and mapped to [0, 1].

Importantly, the annotations will be shared publicly, together with additional data for the HECKTOR 2022 challenge to allow other researchers to evaluate and compare their state of the art methods on this challenging task.

B. Segmentation Models

For these experiments, we use simple 3D U-Net models [14]. The encoder part of the U-Net is composed of five consecutive residual blocks with 4, 8, 16, 32 and 64 output feature maps. The decoder part contains four blocks with transpose convolution and skip-connections from the encoder and with 32, 16, 8 and 4 output feature maps. The output block is composed of a first convolution with 24 filters and a final layer depending on the configuration used, which we detail in the next paragraph. All the intermediate convolutions are ReLU activated and the residual blocks contain batch normalization. We test multiple variants to evaluate (i) the individual contributions and complementarity of the PET and CT modalities in a multi-modal approach (two input channels with a single encoder/decoder); (ii) the benefit of training the model for the segmentation of GTVt and GTVn simultaneously, as opposed to individual models.

For the single segmentation task, the models output a single channel with sigmoid activation in the final layer. The models trained for GTVt and GTVn in parallel output three channels for the background, GTVt and GTVn respectively, with softmax activation (there is no overlap between the classes). The implementation is available on our GitHub repository[1]. Note that ensembles of multiple deeper models trained for many more epochs reach higher performance on the same data for GTVt segmentation task (best Dice Similarity Coefficient, DSC, of 0.759 [15] in HECKTOR 2020). However, we limit our experiments to a single relatively lightweight model to maintain the computational time and energy footprint low as we evaluate various settings (uni-/multi- modal and uni-/multi- task).

[1]https://github.com/voreille/hecktor/tree/master/src/segmentation_gtvtn, as of April 2022.

C. Training Scheme

The GTVt classification models are trained using a simple Dice loss computed as follows.

$$\mathcal{L}_{Dice} = 1 - \frac{2\sum_k \hat{y}_k y_k + \epsilon}{\sum_k (\hat{y}_k + y_k) + \epsilon}, \tag{1}$$

where $\hat{y}_k \in [0, 1]$ is the softmax output for a voxel k, $y_k \in \{0, 1\}$ is the value of this voxel in the 3D ground truth[2] mask, $\epsilon = 10^{-7}$ is a smoothing term added so the loss is zero for empty sets of prediction and ground truth, and the sum is computed over all voxels of the image. This loss is averaged for all images in a training batch.

A main difference between the GTVt and GTVn annotations is that the former contains one single volume per case, while the latter may contain zero, one or more volumes. The classic Dice loss is therefore not suitable for the segmentation of GTVn since images with no GTVn will have a gradient of zero and therefore not contribute to learning. To counter this effect, we employ a batch-aggregated extension of (1). The intersections and unions are aggregated for the N images in a training batch to avoid empty sets of positive voxels for cases that do not contain GTVn as

$$\mathcal{L}_{AggDice} = 1 - \frac{2\sum_i^N \sum_k \hat{y}_{i,k} y_{i,k} + \epsilon}{\sum_i^N \sum_k (\hat{y}_{i,k} + y_{i,k}) + \epsilon}. \tag{2}$$

The multi-class models are trained using the sum of Dice losses of the GTVt and GTVn segmentations defined above, namely Dice loss (1) and aggregated Dice loss (2), respectively.

Standard hyperparameters are used, including a batch size of 4, a maximum number of epochs of 200 with an early stopping on the validation loss with a patience of 20 epochs. The initial learning rate is 10^{-3} which is scheduled using cosine decay with warm restart [16]. The models are implemented in TensorFlow and trained on an Nvidia V100 32GB.

D. Evaluation

The evaluation metrics are computed and reported for the individual labels (GTVt and GTVn). The DSC ranges from zero to one, with one reflecting a perfect similarity between predicted and ground truth contours. It is computed as

$$DSC = \frac{2TP}{2TP + FP + FN} = \frac{2\sum_k \bar{y}_k y_k}{\sum_k (\bar{y}_k + y_k)}, \tag{3}$$

where TP, FP and FN are the number of True Positive, False Positive and False Negative voxels, respectively, and $\bar{y}$ is the hard attribution to the class with highest softmax activation. The DSC measures the volumetric overlap between the predicted and ground truth contours. The reported DSCs for the GTVt segmentation are averaged across all test cases.

For the GTVn segmentation, we report the DSC_{agg}, where the intersections and unions are aggregated for the entire test set. This measure is performed to account for the FP

[2]With a slight misuse of language, we use the term ground truth to refer to the target annotations.

TABLE I

PERFORMANCE COMPARISON FOR THE SEGMENTATION OF GTVT AND GTVN. THE AVERAGE DSCS ARE REPORTED FOR GTVT SEGMENTATION ON THE TEST SET TOGETHER WITH STANDARD DEVIATIONS. THE AGGREGATED DICE IS REPORTED FOR THE GTVN SEGMENTATION (THERE IS NO STANDARD DEVIATION DUE TO AGGREGATTION OF INTERSECTIONS AND UNIONS ACROSS THE ENTIRE TEST SET). WE COMPARE MODELS TRAINED ONLY FOR GTVT SEGMENTATION, ONLY FOR GTVN, AND FOR BOTH.

mod.	training tasks	GTVn (DSC_{agg})	GTVt (DSC)
CT	GTVt only	-	$0.428_{\pm 0.245}$
	GTVn only	0.609	-
	GTVt & GTVn	0.600	$0.420_{\pm 0.267}$
PET	GTVt only	-	$0.651_{\pm 0.252}$
	GTVn only	0.647	-
	GTVt & GTVn	0.679	$0.658_{\pm 0.199}$
PET/CT	GTVt only	-	$0.694_{\pm 0.227}$
	GTVn only	0.722	-
	GTVt & GTVn	**0.729**	**0.717**$_{\pm 0.208}$

predictions in cases without ground truth positives (8 out of 52 test cases). This aggregation is similar to the division in (2), summing for all test images instead of all N images in a batch. For direct comparison of GTVt and GTVn, we also report the average DSC for the best model in the discussion section, using a smoothing term in (3), making the DSC of a case without GTVn ground truth equal to 1 if there is no false positive, 0 otherwise.

We also compute the confusion matrix of all voxels to evaluate misclassifications at the voxel level across the three classes: background, GTVt and GTVn. Finally, we report the average Surface Dice Similarity Coefficient (SDSC) at 1mm and the median Hausdorff Distance at 95% (HD95) as defined in [17] for the GTVt segmentation. These metrics are not well suited for the GTVn segmentation with less or more than a single volume since a single FP or FN lymph node makes the performances drop drastically.

III. RESULTS

A. Quantitative Results

The segmentation performance is reported in Table I. We also compute additional metrics for the best model, obtaining a SDSC of 0.622 and HD95 of 8.11 on the GTVt segmentation, and 0.433, 13.31 for the GTVn segmentation. The confusion matrix with three classes (background, GTVt, GTVn) is reported in Fig. 1 for the best model, i.e. the multi-modal model trained for GTVt and GTVn and corresponding to the last row of Table I.

A comparison of segmentation performance measured with the DSC_{agg} for different ranges of volumes of GTVn is reported in Fig. 2. In this evaluation, we partition the 44 test cases containing at least one GTVn volume into four sub-groups based on the quartiles of the distribution of all test GTVn volumes (Q1 = 1,454, Q2 (median) = 3,469 and Q3 = 12,384.5 mm^3). Each case is assigned to a range [Q0,Q1] (4 cases), [Q1-Q2] (10 cases), [Q2,Q3] (10 cases) or [Q3,inf] (20 cases) based on its largest GTVn volume.

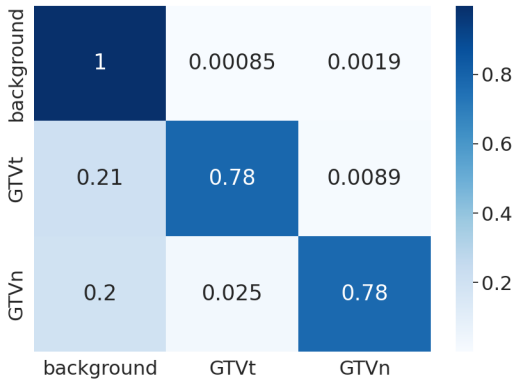


Fig. 1. Normalized confusion matrix of all test voxels predictions using the PET/CT model learning GTVt and GTVn in parallel. Left: grount-truth; bottom: predictions. The total numbers of ground truth voxels associated with the three classes are: background (153,901,994), GTVt (705,025) and GTVn (664,149).

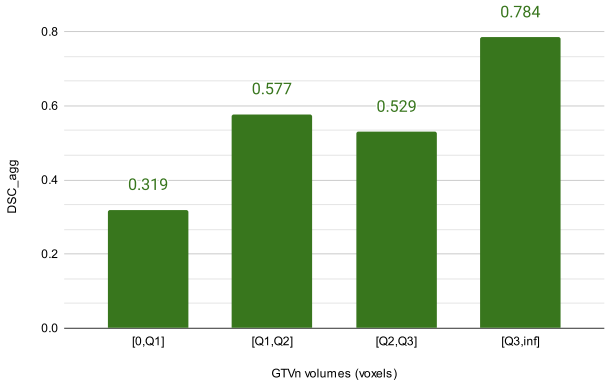


Fig. 2. Comparison of DSC_{agg} for different volumes of GTVn. The four ranges of volumes are based on the quartiles of the test volumes. There is no error bar due to the nature of the aggregated measure.

B. Qualitative Results

Qualitative results of the best model (multimodal U-Net trained on the GTVt ad GTVn tasks simultaneously), including correct segmentation, FPs and FNs of both GTVt and GTVn, are shown in Fig. 3.

C. Benefit of Aggregated Loss

The comparison of the proposed aggregated loss ($\mathcal{L}_{AggDice}$, Eq. 2) with those of identical models trained with a standard Dice loss ($\mathcal{L}_{Dice}$ in Eq. 1) is reported in Table II. The models are trained with the same splits and in a similar manner except for the GTVn segmentation loss.

IV. DISCUSSION AND CONCLUSIONS

This paper presented the first method for GTVn segmentation from PET/CT images and showed its feasibility. The best model obtained a DSC_{agg} of 0.729 for GTVn segmentation and average DSC of 0.717 for the GTVt segmentation. For comparison across tasks, the average DSCs for GTVn segmentation is 0.562, illustrating the difficulty of this task.

TABLE II

PERFORMANCE COMPARISON OF GTVN SEGMENTATION (DSC_{agg}) AND GTVT SEGMENTATION (DSC) ACROSS MODELS TRAINED WITH AND WITHOUT AGGREGATED DICE LOSS.

		$L_{AggDice}$		L_{Dice}	
mod.	training tasks	GTVn	GTVt	GTVn	GTVt
CT	GTVn only	0.609	-	0.646	-
	GTVt & GTVn	0.600	0.420	0.600	0.445
PET	GTVn only	0.647	-	0.665	-
	GTVt & GTVn	0.679	0.658	0.686	0.655
PET/CT	GTVn only	0.722	-	**0.735**	-
	GTVt & GTVn	0.729	**0.717**	0.726	0.693

Some non-malignant lymph nodes resulting from an inflammation can be visually similar to malignant ones. Besides this difficulty of discriminating between malignant and benign lymph nodes, the annotated GTVn volumes are smaller than the GTVt ones (median of 3,469 versus 7,408 voxels in the test set, see full histogram in Fig. 4), which biases the $DSCs$ towards lower values as demonstrated in [18]. The difference of performance due to volume size is illustrated in Fig. 2, showing a DSC_{agg} of 0.319 and 0.784 for the smallest and largest volumes respectively. In future work, metrics that measure detection accuracy instead of voxel-wise accuracy should also be considered for the evaluation of an automatic TNM staging method or to support the decision of lymph node resection.

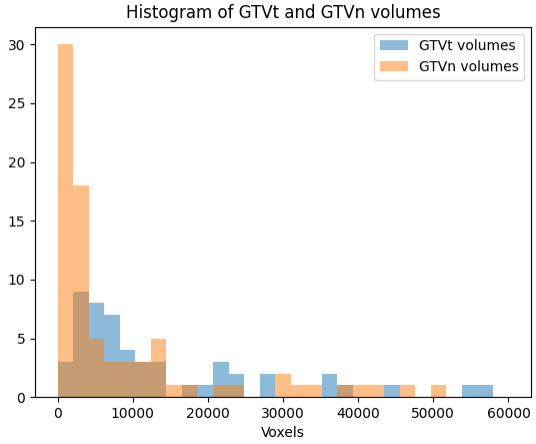


Fig. 4. Histogram of GTVt and GTVn volumes measured in voxels in the test set. Individual connected components are considered for the GTVn. Best viewed in color.

As reported in previous studies [1], most discriminative information for the segmentation of tumor volumes originate from the PET image showing the metabolic activity of the tumor. Yet, the best segmentation results are obtained with the multimodal models. An example of PET signal that expands outside the tumor area (in the trachea) is shown in Fig. 3(b) and is nevertheless correctly segmented thanks to the anatomical information of the CT image. Simultaneously training the segmentation of GTVt and GTVn tend to improve the segmentation quality of both regions types, as both tasks help each other thanks to their similar goal (i.e. segmentation of active cancer regions). The best GTVn segmentation, however, is obtained with a model trained only for this task, and with the standard dice loss.

The use of a batch-aggregated Dice loss improves the performance of the GTVt and GTVn segmentation as shown in Table II in the multi-task setting, where cases without positive GTVn voxels are better managed when compared to the classical Dice loss, limiting the FPs in the final predictions. Note that the number of cases without GTVn in the dataset is low (approx. 15%), resulting in a marginal performance difference between the two losses.

In future work, we will evaluate the benefit of using radiomics features extracted from both primary tumors and lymph nodes to predict patient outcome such as PFS [6]. This could be investigated on large patient cohorts without the need to manually annotate GTVt and GTVn contours, being a tedious task and with limited intra-observer consistency. Deep multi-task models will also be explored as in [1] to guide deep radiomics models with both GTVt and GTVn contours for the prediction of patient outcome.

REFERENCES

[1] Vincent Andrearczyk, Valentin Oreiller, Martin Vallières, Joel Castelli, Hesham Elhalawani, Mario Jreige, Sarah Boughdad, John O Prior, and Adrien Depeursinge, "Automatic segmentation of head and neck tumors and nodal metastases in PET-CT scans," in *Medical Imaging with Deep Learning*. PMLR, 2020, pp. 33–43.

[2] Yngve Mardal Moe, Aurora Rosvoll Groendahl, Martine Mulstad, Oliver Tomic, Ulf Indahl, Einar Dale, Eirik Malinen, and Cecilia Marie Futsaether, "Deep learning for automatic tumour segmentation in PET/CT images of patients with head and neck cancers," *arXiv preprint arXiv:1908.00841*, 2019.

[3] Valentin Oreiller, Vincent Andrearczyk, Mario Jreige, Sarah Boughdad, Hesham Elhalawani, Joel Castelli, Martin Vallières, Simeng Zhu, Juanying Xie, Ying Peng, Andrei Iantsen, Mathieu Hatt, Yading Yuan, Jun Ma, Xiaoping Yang, Chinmay Rao, Suraj Pai, Kanchan Ghimire, Xue Feng, Mohamed A. Naser, Clifton D. Fuller, Fereshteh Yousefirizi, Arman Rahmim, Huai Chen, Lisheng Wang, John O. Prior, and Adrien Depeursinge, "Head and neck tumor segmentation in PET/CT: The HECKTOR challenge," *Medical Image Analysis*, vol. 77, pp. 102336, Apr. 2022.

[4] Vincent Andrearczyk, Valentin Oreiller, Mario Jreige, Martin Vallières, Joel Castelli, Hesham Elhalawani, Sarah Boughdad, John O Prior, and Adrien Depeursinge, "Overview of the HECKTOR challenge at MICCAI 2020: Automatic head and neck tumor segmentation in PET/CT," in *3D Head and Neck Tumor Segmentation in PET/CT Challenge*. Springer, 2020, pp. 1–21.

[5] Vincent Andrearczyk, Valentin Oreiller, Sarah Boughdad, Catherine Chez Le Rest, Hesham Elhalawani, Mario Jreige, John O. Prior, Martin Vallières, Dimitris Visvikis, Mathieu Hatt, and Adrien Depeursinge, "Overview of the HECKTOR challenge at MICCAI 2021: Automatic head and neck tumor segmentation and outcome prediction in PET/CT images," in *LNCS proceedings*, 2022.

[6] Pierre Fontaine, Vincent Andrearczyk, Valentin Oreiller, Joel Castelli, Mario Jreige, John O Prior, and Adrien Depeursinge, "Fully automatic head and neck cancer prognosis prediction in PET/CT," in *Multimodal Learning and Fusion Across Scales for Clinical Decision Support (ML-CDS) at MICCAI*. PMLR, 2021.

[7] Vincent Andrearczyk, Pierre Fontaine, Valentin Oreiller, and Adrien Depeursinge, "Multi-task deep segmentation and radiomics for automatic prognosis in head and neck cancer," in *Workshop on PRedictive Intelligence in MEdicine (PRIME) at MICCAI*. PMLR, 2021.

[8] Mingyuan Meng, Bingxin Gu, Lei Bi, Shaoli Song, David Dagan Feng, and Jinman Kim, "DeepMTS: Deep multi-task learning for survival prediction in patients with advanced nasopharyngeal carcinoma using pretreatment PET/CT," *CoRR*, vol. abs/2109.07711, 2021.

[9] DL Farfan Cabrera, Nicolas Gogin, David Morland, Benoît Naegel, Dimitri Papathanassiou, and Nicolas Passat, "Segmentation of axillary and supraclavicular tumoral lymph nodes in PET/CT: A hybrid CNN/component-tree approach," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 6672–6679.

[10] Guoping Xu, Hanqiang Cao, Jayaram K Udupa, Yubing Tong, and Drew A Torigian, "Disegnet: A deep dilated convolutional encoder-decoder architecture for lymph node segmentation on PET/CT images," *Computerized Medical Imaging and Graphics*, vol. 88, pp. 101851, 2021.

[11] Phi Xuan Nguyen, Zhongkang Lu, Weimin Huang, Su Huang, Akie Katsuki, and Zhiping Lin, "Medical image segmentation with stochastic aggregated loss in a unified U-Net," in *2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*. IEEE, 2019, pp. 1–4.

[12] Martin Vallieres, Emily Kay-Rivest, Léo Jean Perrin, Xavier Liem, Christophe Furstoss, Hugo JWL Aerts, Nader Khaouam, Phuc Felix Nguyen-Tan, Chang-Shu Wang, Khalil Sultanem, et al., "Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer," *Scientific reports*, vol. 7, no. 1, pp. 1–14, 2017.

[13] Vincent Andrearczyk, Valentin Oreiller, and Adrien Depeursinge, "Oropharynx detection in PET-CT for tumor segmentation," in *Irish Machine Vision and Image Processing*, 2020.

[14] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger, "3D U-Net: learning dense volumetric segmentation from sparse annotation," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2016, pp. 424–432.

[15] Andrei Iantsen, Dimitris Visvikis, and Mathieu Hatt, "Squeeze-and-excitation normalization for automated delineation of head and neck primary tumors in combined PET and CT images," in *Lecture Notes in Computer Science (LNCS) Challenges*, 2021.

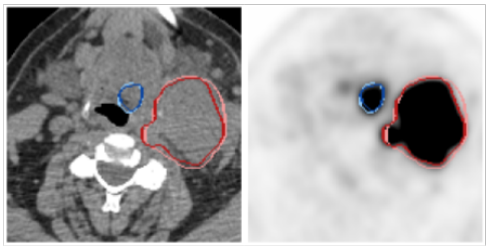
[16] Ilya Loshchilov and Frank Hutter, "SGDR: Stochastic gradient descent with warm restarts," in *International Conference on Learning Representations (ICLR)*, 2017.

[17] Stanislav Nikolov, Sam Blackwell, Alexei Zverovitch, Ruheena Mendes, Michelle Livne, Jeffrey De Fauw, Yojan Patel, Clemens Meyer, Harry Askham, Bernadino Romera-Paredes, et al., "Clinically applicable segmentation of head and neck anatomy for radiotherapy: Deep learning algorithm development and validation study," *Journal of Medical Internet Research*, vol. 23, no. 7, pp. e26151, 2021.

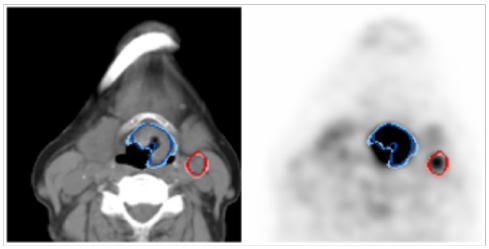
[18] Annika Reinke, Matthias Eisenmann, Minu Dietlinde Tizabi, Carole H Sudre, Tim Rädsch, Michela Antonelli, Tal Arbel, Spyridon Bakas, M Jorge Cardoso, Veronika Cheplygina, et al., "Common limitations of performance metrics in biomedical image analysis," in *Medical Imaging with Deep Learning*, 2021.

Compliance with Ethical Standards

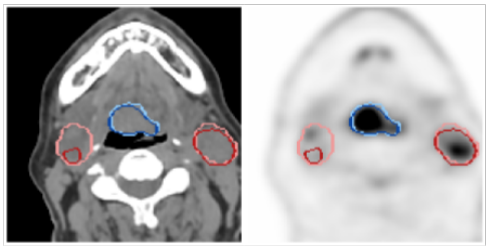
The ethics approvals were obtained from the institutions which provided the data to the HECKTOR challenge.



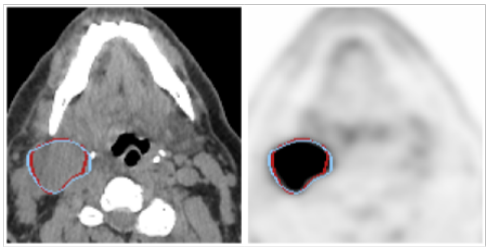
(a) GTVt: 0.872, GTVn: 0.847



(b) GTVt: 0.593, GTVn: 0.670



(c) GTVt: 0.886, GTVn: 0.549



(d) GTVt: 0, GTVn: 0

Fig. 3. Illustrations of 2D PET (SUV scale 0-7 mg/L) and CT slices overlayed with GTVt (blue) and GTVn (red) automatic segmentation. The corresponding DSC are reported in the captions. The ground truth is in bright color, the prediction in dark color. (a,b) Correctly detected and segmented GTVt and GTVn; (c) One GTVn correctly segmented (right), one largely undersegmented (left); (d) Standard DSC is reported for each case to evaluate the 3D segmentation of GTVt and GTVn (when there is a ground truth volume for the latter). Best viewed in color.