Overview of the HECKTOR Challenge at MICCAI 2020: Automatic Head and Neck Tumor Segmentation in PET/CT

Vincent Andrearczyk*[1], Valentin Oreiller*[1,2], Mario Jreige[2], Martin Vallières[3], Joel Castelli[4,5,6], Hesham Elhalawani[7], Sarah Boughdad[2], John O. Prior[2], and Adrien Depeursinge[1,2]

[1] Institute of Information Systems, School of Management, HES-SO Valais-Wallis University of Applied Sciences and Arts Western Switzerland, Sierre, Switzerland
[2] Centre Hospitalier Universitaire Vaudois (CHUV), Lausanne, Switzerland
[3] Department of Computer Science, University of Sherbrooke, Sherbrooke, Québec, Canada
[4] Radiotherapy Department,Cancer Institute Eugène Marquis, Rennes, France
[5] INSERM, U1099, Rennes, France
[6] University of Rennes 1, LTSI, Rennes, France
[7] Cleveland Clinic Foundation, Department of Radiation Oncology, Cleveland, OH, USA

Abstract. This paper presents an overview of the first HEad and neCK TumOR (HECKTOR) challenge, organized as a satellite event of the 23rd International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) 2020. The task of the challenge is the automatic segmentation of head and neck primary Gross Tumor Volume in FDG-PET/CT images, focusing on the oropharynx region. The data were collected from five centers for a total of 254 images, split into 201 training and 53 testing cases. The interest in the task was shown by the important participation with 64 teams registered and 18 team submissions. The best method obtained a Dice Similarity Coefficient (DSC) of 0.7591, showing a large improvement over our proposed baseline method with a DSC of 0.6610 as well as inter-observer DSC agreement reported in the literature (0.69).

Keywords: Automatic segmentation· Challenge· Medical Imaging · Head and Neck Cancer · Oropharynx.

1 Introduction: Research Context

The prediction of disease characteristics using quantitative image biomarkers from medical images (i.e. radiomics) has shown tremendous potential to optimize patient care, particularly in the context of Head and Neck (H&N) tumors [20]. FluoroDeoxyGlucose (FDG)-Positron Emission Tomography (PET)

* equal contribution

and Computed Tomography (CT) imaging are the modalities of choice for the initial staging and follow-up of H&N cancer. Yet, radiomics analyses rely on an expensive and error-prone manual annotation process of Volumes of Interest (VOI) in three dimensions. The automatic segmentation of H&N tumors from FDG-PET/CT images could therefore enable the validation of radiomics models on very large cohorts and with optimal reproducibility. Besides, automatic segmentation algorithms could enable a faster clinical workflow. By focusing on metabolic and morphological tissue properties respectively, PET and CT modalities include complementary and synergistic information for cancerous lesion segmentation. The HEad and neCK TumOR (HECKTOR)[1] challenge aims at identifying the best methods to leverage the rich bi-modal information in the context of H&N primary tumor segmentation. This precious knowledge will be transferable to many other cancer types where PET/CT imaging is relevant, enabling large-scale and reproducible radiomics studies.

The potential of PET information for automatically segmenting tumors has been long exploited in the literature. For an in-depth review of automatic segmentation of PET images in the pre-deep learning era, see [5] covering methods such as thresholding, active contours and mixture models. The first challenge on tumor segmentation in PET images was proposed at MICCAI 2016[2] by Hatt et al. [8]. The need for a standardized evaluation of PET automatic segmentation methods and a comparison study between all the current algorithms was highlighted in [9]. Multi-modal analyses of PET and CT images have also recently been proposed for different tasks, including lung cancer segmentation in [11,12,25,26] and bone lesion detection in [22]. In [2], we developed a baseline Convolutional Neural Network (CNN) approach based on a leave-one-center-out cross-validation on the training data of the HECKTOR challenge. Promising results were obtained with limitations that motivated additional data curation, data cleaning and the creation of this challenge. This challenge builds upon these works by comparing, on a publicly available dataset, recent segmentation architectures as well as the complementarity of the two modalities on a task of primary Gross Tumor Volume (GTVt) segmentation of H&N tumor in the oropharynx region. The proposed dataset comprises data from five centers. Four centers are used for the training data and one for testing. The task is challenging due to, among others, the variation in image acquisition and quality across centers (test set from an unseen center) and the presence of lymph nodes with high metabolic responses in the PET images.

The critical consequences of the lack of quality control in challenge designs were shown in [14], including reproducibility and interpretation of the results often hampered by the lack of provided relevant information, and non-robust ranking of algorithms. Solutions were proposed in the form of the Biomedical Image Analysis challengeS (BIAS) [15] guidelines for reporting the results. This paper presents an overview of the challenge following these guidelines.

[1] `www.aicrowd.com/challenges/hecktor`, as of October 2020.

[2] `https://portal.fli-iam.irisa.fr/petseg-challenge/overview#_ftn1`, as of October 2020.

Individual participants' papers were submitted to the challenge organizers, reporting methods and results. Reviews were organized by the organizers and the papers of the participants are published in the LNCS challenges proceedings [10,4,13,18,21,27,6,23,24,17].

The paper is organized as follows. The challenge design and data description are described in Section 2. The main results of the challenge are reported in Section 3 and discussed in Section 4. Finally, Section 5 concludes this paper.

2 Methods: Reporting of Challenge Design

A summary of the information on the challenge organization is provided in Appendix 1, following the BIAS recommendations.

2.1 Mission of the Challenge

Biomedical application
The participating algorithms target the following fields of application: diagnosis, prognosis and research. The participating teams' algorithms were designed for image segmentation, more precisely, classifying voxels as either tumor or background.

Cohorts
As suggested in [15], we refer to the patients from whom the image data were acquired as the cohort. The target cohort[3] comprises patients received for initial staging of H&N cancer. The clinical goals are two-fold; the automatically segmented regions can be used as a basis for (i) treatment planning in radiotherapy, (ii) further radiomics studies to predict clinical outcomes such as overall patient survival, disease-free survival, tumor aggressivity. In the former case (i), the regions will need to be further refined or extended for optimal dose delivery and control. The challenge cohort[4] includes patients with histologically proven H&N cancer who underwent radiotherapy treatment planning. The data were acquired from five centers (four for the training and one for the testing) with variation in the scanner manufacturers and acquisition protocols. The data include PET and CT imaging modalities as well as patient information including age, sex and acquisition center. A detailed description of the annotations is provided in Section 2.2.

Target entity
The data origin, i.e. the region from which the image data were acquired, varied from the head region only to the whole body. While we provided the data

[3] The target cohort refers to the subjects from whom the data would be acquired in the final biomedical application. It is mentioned for additional information as suggested in BIAS, although all data provided for the challenge are part of the challenge cohort.
[4] The challenge cohort refers to the subjects from whom the challenge data were acquired.

as acquired, we limited the analysis to the oropharynx region and provided an automatically detected bounding box locating the oropharynx region [1], as illustrated in Figure 1.

Assessment aim
The assessment aim is the following; evaluate the feasibility of fully automatic GTVt segmentation for H&N cancers in the oropharyngeal region via the identification of the most accurate segmentation algorithm. The performance of the latter is identified by computing the Dice Similarity Coefficient (DSC) between prediction and manual expert annotations. The individual DSC scores are averaged for all test patients and the ranking is based on this average score. DSC measures volumetric overlap between segmentation results and annotations. It is a good measure of segmentation for imbalanced segmentation problems, i.e. the region to segment is small as compared to the image size. DSC is commonly used in the evaluation and ranking of segmentation algorithms and particularly tumor segmentation tasks [7,16].

Missing values (i.e. missing predictions on one or multiple patients), did not occur in the submitted results, but will be treated as DSC of zero if it occurs in future submissions on the open leaderboard. In case of tied rank, very unlikely due to the computation of the results (average of 53 DSCs), we will consider the precision as the second ranking metric.

A statistical analysis is performed to statistically compare the performance of the algorithms using a Wilcoxon-signed rank test.

2.2 Challenge Dataset

Data source
The data were acquired from five centers as listed in Table 1. It consists of PET/CT images of patients with H&N cancer located in the oropharynx region. The scanners (devices) and imaging protocols used to acquire the data are described in Table 2. Additional information about the image acquisition is provided in Appendix 2.

Table 1: List of the hospital centers in Canada (CA) and Switzerland (CH) and number of cases, with a total of 201 training and 53 test cases.

Center	Split	# cases
HGJ: Hôpital Général Juif, Montréal, CA	Train	55
CHUS: Centre Hospitalier Universitaire de Sherbooke, Sherbrooke, CA	Train	72
HMR: Hôpital Maisonneuve-Rosemont, Montréal, CA	Train	18
CHUM: Centre Hospitalier de l'Université de Montréal, Montréal, CA	Train	56
Total	Train	201
CHUV: Centre Hospitalier Universitaire Vaudois, CH	Test	53

Data preprocessing

Table 2: List of scanners used in the different centers.

Center	Device
HGJ	hybrid PET/CT scanner (Discovery ST, GE Healthcare)
CHUS	hybrid PET/CT scanner (GeminiGXL 16, Philips)
HMR	hybrid PET/CT scanner (Discovery STE, GE Healthcare)
CHUM	hybrid PET/CT scanner (Discovery STE, GE Healthcare)
CHUV	hybrid PET/CT scanner (Discovery D690 TOF, GE Healthcare)

Training and test case characteristics
The training data comprise 201 cases from four centers (HGJ, HMR[5], CHUM and CHUS). Originally, the dataset in [20] contained 298 cases, among which we selected the cases with oropharynx cancer. The test data comprise 53 cases from another fifth center (CHUV). Examples of PET/CT images of each center are shown in Figure 1. Each case comprises a CT image, a PET image and a GTVt mask (for the training cases) in the Neuroimaging Informatics Technology Initiative (NIfTI) format, as well as patient information (age, sex) and center. A bounding box locating the oropharynx region was also provided (details of the automatic region detection can be found in [1]).

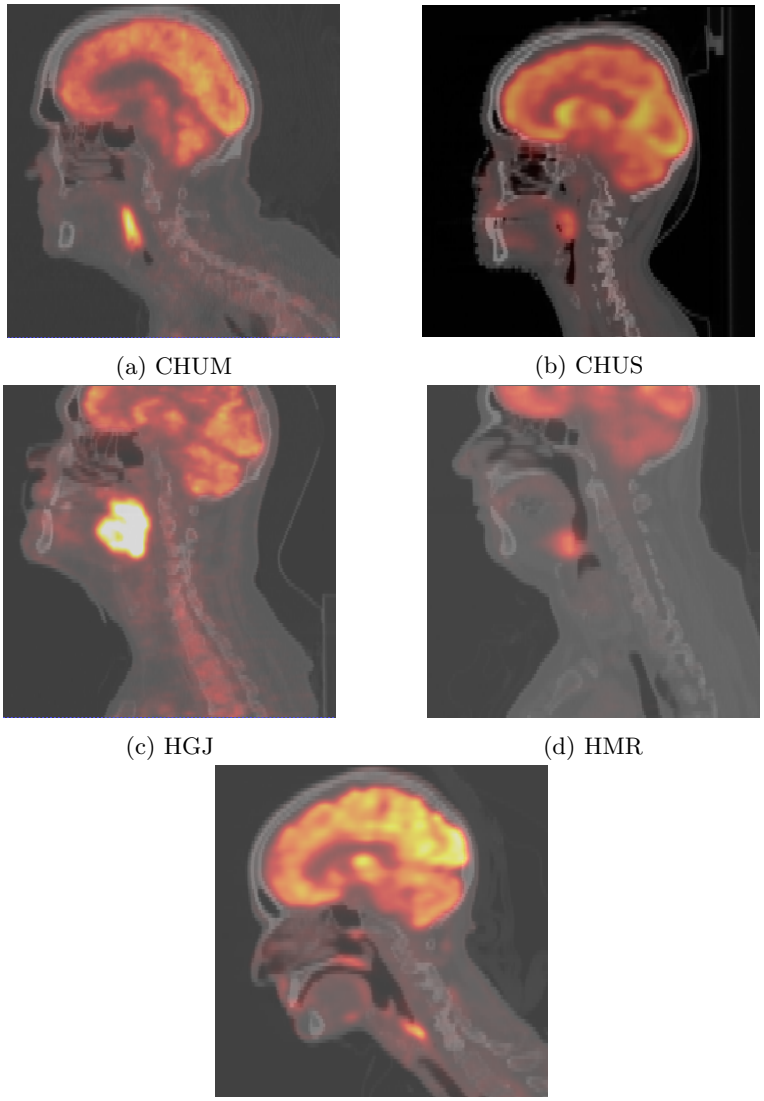
Finally, to provide a fair comparison, participants who wanted to use additional external data for training were asked to also report results using only the HECKTOR data and discuss differences in the results.

Annotation characteristics
Initial annotations, i.e. 3D contours of the GTVt, were made by expert radiation oncologists and were later modified by a VOI quality control and correction as described later. Details of the initial annotations of the training set can be found in [20]. In particular, 40% (80 cases) of the training radiotherapy contours were directly drawn on the CT of the PET/CT scan and thereafter used for treatment planning. The remaining 60% of the training radiotherapy contours were drawn on a different CT scan dedicated to treatment planning and were then registered to the FDG-PET/CT scan reference frame using intensity-based free-form deformable registration with the software MIM (MIM software Inc., Cleveland, OH). The initial contours of the test set were all directly drawn on the CT of the PET/CT scan.

VOI quality control and correction were supervised by an expert who is both radiologist and nuclear medicine physician. Two non-experts (organizers of the challenge) made an initial cleaning in order to facilitate the expert's work. The expert either validated or edited the VOIs. The Siemens Syngo.Via RT Image Suite was used to edit the contours in 3D with fused PET/CT images. The main points corrected during the data curation are listed in the following.

– All annotations were originally performed in a radiotherapy context. They were potentially inadequate for radiomics studies as too large, often including

[5] For simplicity, these centers were renamed CHGJ and CHMR during the challenge.



(a) CHUM (b) CHUS

(c) HGJ (d) HMR

(e) CHUV

Fig. 1: Case examples of 2D sagittal slices of fused PET/CT images from each of the five centers.

air in the trachea and various tissues surrounding the tumor. The primary tumors were delineated as close as possible to the real tumoral volume.
- Some annotations were originally drawn on a distinct CT scan dedicated to treatment planning. In this case, the contours were registered to the PET/CT scans (more details in [20]). Some registrations failed and had to be corrected.
- Some VOIs included both the primary tumor (GTVt) and lymph nodes (GTVn) without distinction. Separating the primary tumor from the lymph nodes was essential as they carry different information and should not be grouped.
- One contour (HGJ069) was flipped symmetrically on the A/P plane.
- Some annotations were missing and had to be drawn from scratch by the expert.

Data preprocessing methods
No preprocessing was performed on the images to reflect the diversity of clinical data and to leave full flexibility to the participants. However, we provided various pieces of code to load, crop, resample the data, train a baseline CNN (NiftyNet) and evaluate the results on our GitHub repository[6]. This code was provided as a suggestion to help the participants and to maximize transparency, but the participants were free to use other methods.

Sources of errors
According to Gudi et al. [7], in the context of radiotherapy planning, one can expect an inter-observer DSC in tumor segmentation of 0.57 and 0.69 on CT and PET/CT respectively, highlighting the difficulty of the task. A source of error therefore originates from the degree of subjectivity in the annotation and correction of the expert. For most patients, the tumors were contoured on another CT scan, then the two CTs were registered and the annotations were transformed according to the registrations. Thus, a major source of error came from this registration step.

Another source of error comes from the lack of CT images with a contrast agent for a more accurate delineation of the primary tumor.

Institutional review boards
Institutional Review Boards (IRB) of all participating institutions permitted the use of images and clinical data, either fully anonymized or coded, from all cases for research purposes, only. Retrospective analyses were performed following the relevant guidelines and regulations as approved by the respective institutional ethical committees with protocol numbers: MM-JGH-CR15-50 (HGJ, CHUS, HMR, CHUM) and CER-VD 2018-01513 (CHUV).

2.3 Assessment Method

Participants were given access to the test cases without the ground truth annotations and were asked to submit the results of their algorithms on the test cases on the AIcrowd platform.

[6] `github.com/voreille/hecktor`, as of October 2020.

Results were ranked using the (3D) Dice Similarity Coefficient (DSC) computed on images cropped using the provided bounding boxes (see Section 2.2) in the original CT resolution as:

$$DSC = \frac{2TP}{2TP + FP + FN}, \tag{1}$$

where TP, FP and FN are the number of true positive, false positive and false negative pixels, respectively. If the submitted results were in a resolution different from the CT resolution, we applied nearest-neighbor interpolation before evaluation. We also computed other metrics for comparison, namely precision ($\frac{TP}{TP+FP}$) and recall ($\frac{TP}{TP+FN}$) to investigate whether the method was rather providing a large FP or FN rate. The evaluation implementation can be found on our GitHub repository[7] and was provided to maximize transparency.

Each participating team had the opportunity to submit up to five (valid) runs. The best result of each team was used in the final ranking, which is detailed in Section 3 and discussed in Section 4.

3 Results: Reporting of Challenge Outcome

3.1 Participation

We received and approved, as of Sept. 10 2020 (submission deadline), 85 signed end-user-agreements. At the same date, the number of registered teams was 64. A team is made of at least one participant and not all participants that signed the end-user-agreement registered a team. Each team could submit up to five valid submissions. By the submission deadline, we had received 83 results submissions, including valid and invalid ones (i.e. non graded due to format errors). For the first iteration of the challenge, these numbers are high and show an important interest in the task.

3.2 Algorithms Summary

Organizers' baselines
We trained several baseline models using standard 3D and 2D U-Nets [19] as in our preliminary results in [2] (the data were different). We trained on multi-modal PET/CT as well as individual modalities with a non-weighted Dice (i.e. based on DSC) and cross-entropy loss and without data augmentation.

Participants' methods
In [10], Iantsen et al. proposed a model based on a U-Net architecture with residual layers and supplemented with 'Squeeze and Excitation' (SE) normalization, previously developed by the same authors for brain tumor segmentation. An unweighted sum of soft Dice loss and Focal Loss was used for training. The

[7] `github.com/voreille/hecktor/tree/master/src/evaluation`, as of October 2020.

test results were obtained as an ensemble of eight models trained and validated on different splits of the training set. No data augmentation was performed.

In [13], Ma and Yang used a combination of U-Nets and hybrid active contours. First, 3D U-Nets are trained to segment the tumor (with a cross-validation on the training set). Then, the segmentation uncertainty is estimated by model ensembles on the test set to select the cases with high uncertainties. Finally, the authors used a hybrid active contour model to refine the high uncertainty cases. The U-Nets were trained with an unweighted combination of Dice loss and top-K loss. No data augmentation was used.

In [27], Zhu et al. used a two steps approach. First, a classification network (based on ResNet) selects the axial slices which may contain the tumor. These slices are then segmented using a 2D U-Net to generate the binary output masks. Data augmentation was applied by shifting the crop around the provided bounding boxes and the U-Net was trained with a soft Dice loss. The preprocessing includes clipping the CT and the PET, standardizing the HU within the cropped volume and scaling the range of the PET to correspond to the CT range by dividing it by a factor of 10.

In [24], Yuan proposed to integrate information across different scales by using a dynamic Scale Attention Network (SA-Net), based on a U-Net architecture. Their network incorporates low-level details with high-level semantics from feature maps at different scales. The network was trained with standard data augmentation and with a Jaccard distance loss, previously developed by the authors. The results on the test set were obtained as an ensemble of ten models.

In [4], Chen et al. proposed a three-step framework with iterative refinement of the results. In this approach, multiple 3D U-Nets are trained one-by-one using a Dice loss without data augmentation. The predictions and features of previous models are captured as additional information for the next one to further refine the segmentation.

In [6], Ghimire et al. developed a patch-based approach to tackle the memory issue associated with 3D images and networks. They used an ensemble of conventional convolutions (with small receptive fields capturing fine details) and dilated convolutions (with a larger receptive field of capturing global information). They trained their model with a weighted cross-entropy and dice loss and random left-right flips of the patches were applied for data augmentation. Finally, an ensemble of the best two models selected during cross-validation was used for predicting the segmentation of the test data.

In [23], Yousefirizi and Rahmim proposed a deep 3D model based on SegAN, a generative adversarial network (GAN) for medical image segmentation. An improved polyphase V-net (to help preserve boundary details) is used for the generator and the discriminator network has a similar structure to the encoder part of the former. The networks were trained using a combination of Mumford-Shah (MS) and multi-scale Mean Absolute Error (MAE) losses, without data augmentation.

In [21], Xie and Peng proposed a 3D scSE nnU-Net model, improving upon the 3D nnU-Net by integrating the spatial and channel 'Squeeze and Excitation' (scSE) blocks. They trained the model with a weighted combination of Dice and cross-entropy losses, together with standard data augmentation techniques (rotation, scaling etc.). To preprocess the CT images an automated level-window-like clipping of intensity values is performed based on the 0.5 and 99.5th percentile of these values. The intensity values of the PET are standardized by subtracting the mean and then, by dividing by the standard deviation of the image.

In [17], Naser et al. used a variant of 2D and 3D U-Net (we report the best result, with the 3D model). The models were trained with a combination of Dice and cross-entropy losses with standard data augmentation.

In [18], Rao et al. proposed an ensemble of two methods, namely a 3D U-Net and another 2D U-Net variant with 3D context. A top-k loss was used to train the models without data augmentation.

In Table 3, we summarize some of the main components of the participants' algorithms, including model architecture, preprocessing, training scheme and postprocessing.

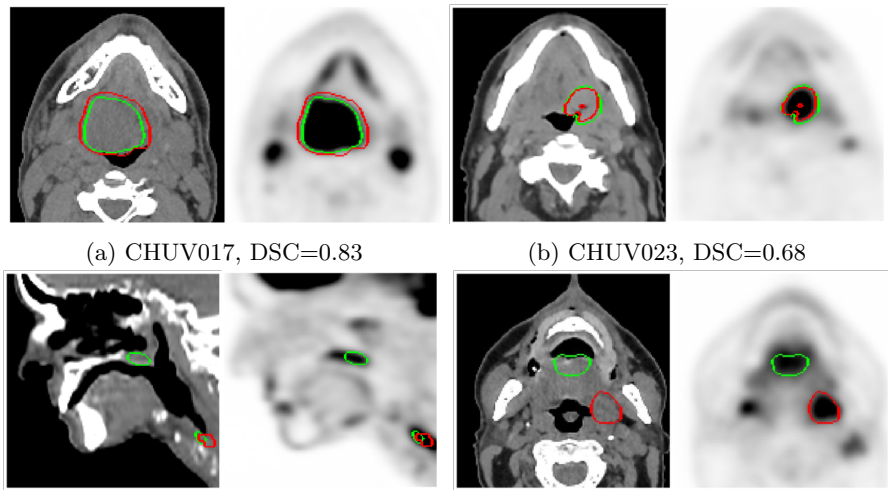
3.3 Results

The results, including average DSC, precision, recall and challenge rank are summarized in Table 4. Our baseline method, developed in [2] and provided to participants as an example on our GitHub repository, obtains an average DSC of 0.6588 and 0.6610 with the 2D and 3D implementations respectively. Results on individual modalities are also reported for comparison. The results from the participants range from an average DSC of 0.5606 to 0.7591. Iantsen et al. [10] (participant *andrei.iantsen*) obtained the best overall results with an average DSC of 0.7591, an average precision of 0.8332 and an average recall of 0.7400. These results (DSCs) are not significantly higher than the second best participant [13] (p-value 0.3501 with a one-tail Wilcoxon signed-rank test) and are significantly higher than the third best participant (p-value 0.0041 with the same test). Across all participants, the average precision ranges from 0.5850 to 0.8479. The recall ranges from 0.5022 to 0.8534, with the latter surprisingly obtained by the 3D PET/CT baseline (although with low precision, reflecting an over-segmentation as compared to other algorithms' outputs). Note that two participants decided to withdraw their submissions due to very low scores. We allowed them to do so since their low scores were due to incorrect postprocessing (e.g setting incorrect pixel spacing, or image origin) and were not representative of the performance of their algorithms.

Examples of segmentation results (true positives on top row, and false positives on bottom row) are shown in Figure 2.

Table 3: Summary of the algorithms with some of the main components: 2D or 3D U-Net, resampling, preprocessing, training or testing data augmentation, loss used for optimization, ensemble of multiple models for test prediction and postprocessing of the results. We use the following abbreviations for the preprocessing: clipping (C), standardization (S), and if it is applied only to one modality, it is specified in parentheses. For the image resampling, we specify whether the algorithms use isotropic (I) or anisotropic (A) resampling and nearest neighbor (NN), linear (L) or cubic (Cu) interpolation. We use the following abbreviation for the loss: Cross-Entropy (CE), Mumford-Shah (MS) and Mean Absolute Error (MAE). More details can be found in the respective participants' publications.

Team	2D/3D	preproc.	resampling	augm.	loss	ensemble	postproc.
`andrei.iantsen` [10]	3D	C+S	I/L	✓	soft Dice+Focal	✓	✗
`junma` [13]	3D	S(PET)	I/Cu	✗	Dice+Top-K	✓	✓
`badger` [21]	3D	C(CT)+S(PET)	A/Cu	✓	Dice+CE	✗	✗
`deepX` [24]	3D	C(CT)+S	I/L	✓	Jaccard distance	✓	✗
`AIView_sjtu` [4]	3D	C+S	A/NN	✓	Dice	✗	✗
`xuefeng` [6]	3D	C(CT)+S	A/L	✓	Dice+CE	✓	✓
`QuritLab` [23]	3D	S	I/L	✗	MS+MAE	✗	✗
`HFHSegTeam` [27]	2D	C+S	I/L	✓	soft Dice	✗	✗
`Fuller_MDA_Lab` [17]	3D	C+S	A/Cu	✓	Dice+CE	✗	✗
`Maastro-Deep-Learning` [18]	2D/3D	C	A/Cu	✗	Top-K	✓	✓
`Our baseline 3D PET/CT`	3D	C+S	I/Cu	✗	Dice+CE	✗	✗
`Our baseline 2D PET/CT`	2D	C+S	I/Cu	✗	Dice+CE	✗	✗



(a) CHUV017, DSC=0.83 (b) CHUV023, DSC=0.68

(c) CHUV001, DSC=0.11 (d) CHUV019, DSC=0.00

Fig. 2: Examples of results of the winning team (andrei.iantsen [10]). The automatic segmentation results (green) and ground truth annotations (red) are displayed on an overlay of 2D slices of PET (right) and CT (left) images. The reported DSC is computed on the whole image (see Eq 1). (a), (b) Excellent segmentation results, detecting the GTVt of the primary oropharyngeal tumor localized at the base of the tongue and discarding the laterocervical lymph nodes despite high FDG uptake on PET. (c) Incorrect segmentation of the top volume at the level of the soft palate; (d) Incorrect segmentation of the tongue due to an abnormal FDG uptake (possible reasons include prior surgical intervention of the tongue, chewing gum and involuntary movements of the tongue).

Table 4: Summary of the challenge results. The average DSC, precision and recall are reported for the baseline algorithms and for the different teams (best result of each team). The participant names are reported when no team name was provided. The ranking is only provided for teams that presented their method in a paper submission.

Team	DSC	precision	recall	rank
andrei.iantsen [10]	**0.7591**	0.8332	0.7400	1
junma [13]	0.7525	0.8384	0.7471	2
badger [21]	0.7355	0.8326	0.7023	3
deepX [24]	0.7318	0.7851	0.7319	4
AIView_sjtu [4]	0.7241	**0.8479**	0.6701	5
DCPT	0.7049	0.7651	0.7047	-
xuefeng [6]	0.6911	0.7525	0.6928	6
ucl_charp	0.6765	0.7231	0.7256	-
QuritLab [23]	0.6677	0.7290	0.7164	7
Unipa	0.6674	0.7143	0.7039	-
Our baseline 3D PET/CT	0.6610	0.5909	**0.8534**	-
Our baseline 2D PET/CT	0.6588	0.6241	0.7629	-
HFHSegTeam [27]	0.6441	0.6938	0.7014	8
UESTC_501	0.6381	0.6455	0.6874	-
Fuller_MDA_Lab [17]	0.6373	0.7546	0.6283	9
Our baseline 3D PET	0.6306	0.5768	0.8214	-
Our baseline 2D PET	0.6284	0.6470	0.6666	-
Maastro-Deep-Learning [18]	0.5874	0.6560	0.6141	10
Yone	0.5737	0.6606	0.5590	-
SC_109	0.5633	0.7652	0.5022	-
Roque	0.5606	0.5850	0.6843	-
Our baseline 2D CT	0.3071	0.3477	0.3574	-
Our baseline 3D CT	0.2729	0.2154	0.5874	-

4 Discussion: Putting the Results into Context

4.1 Outcome and Findings

A major benefit of this challenge is to compare various algorithms developed by teams from all around the world on the same dataset and task, with held-out test data.

We distinguish here between the technical and biomedical impact. The main technical impact of the challenge is the comparison of state-of-the-art algorithms on the provided data. We identified key elements for addressing the task: 3D U-Net, preprocessing, normalization, data augmentation and ensembling, as summarized in Table 3. The main biomedical impact of the results is the opportunity to generate large cohorts with automatic tumor segmentation for comprehensive radiomics studies.

The best methods obtain excellent results with DSCs above 0.75, better than[8] reported inter-observer variability (DSCs of 0.57 and 0.69 on CT and PET-CT respectively) [7]. Note that without injected contrast CT, delineating the exact contour of the tumor is very difficult. Thus, the inter-observer DSC could be low only due to disagreements at the border of the tumor, without taking into account the error rate due to the segmentation of non-malignant structures (if any). For that reason, defining the task as solved solely based on the DSC is not sufficient. In the context of this challenge, we can therefore define the task as solved if the algorithms follow these three criteria:

1. Higher or similar DSC than inter-observers agreement.
2. Detect all the primary tumors in the oropharynx region (i.e. segmentation not evaluated at the pixel level, rather at the occurrence level).
3. Similarly, detect only the primary tumors in the oropharynx region (discarding lymph nodes and other potentially false positives).

According to these criteria, the task is partially solved. The first criterion, evaluating the segmentation at the pixel level, is fulfilled. At the occurrence level (criteria 2 and 3), however, even the algorithms with the highest DSC output FP and FN regions. These errors are generally made in very difficult cases and we should further evaluate their source, e.g. Figure 2c and 2d. Besides, there is still a lot of work to do on highly related tasks, including the segmentation of lymph nodes, the development of super-annotator ground truth as well as the agreement of multiple annotators, and, finally, the prediction of patient outcome following the tumor segmentation.

Following the analysis of poorly segmented cases, we identified several key elements that cause the algorithms to fail. These elements are as follows; low FDG uptake on PET, primary tumor that looks like a lymph node, abnormal uptake in the tongue and tumor present at the border of the oropharynx region. Some examples are illustrated in Figure 1. Understanding these errors will lead to better methods and to a more targeted task for the next iteration of this challenge.

4.2 Limitations of the Challenge

The dataset provided in this challenge suffers from several limitations. First, the contours were mainly drawn based on the PET/CT fusion which is not sufficient to clearly delineate the tumor. Other methods such as MRI with gadolinium or contrast CT are the gold standard to obtain the true contours for radiation oncology. Since the target clinical application is radiomics, however, the precision of the contours is not as important as for radiotherapy planning.

[8] These values are reported only to give an idea of inter-observer variability on a similar task reported in the literature. The datasets are different and the comparison is limited. In future work, we will compute the inter-observer agreement on the challenge data.

Another limitation comes from the definition of the task, only one segmentation was drawn on the fusion of PET and CT. For radiomic analysis, it could be beneficial to consider one segmentation per modality since the PET signal is often not contained in the fusion-based segmentation due to the poor spatial resolution of this modality.

4.3 Lessons Learned and Next Steps

The guidelines, requirements and review process of the MICCAI submission helped us to design the challenge and to consider as much as possible the potential difficulties that could arise during the challenge.

Feedback from participants
Relevant feedback was provided by the participants in the form of discussion and survey. They overall rated the quality of the data as good and the timing (see "Challenge schedule" in Appendix 1) appropriate. Participants were particularly interested in extending the challenge task to the segmentation of lymph nodes and, more moderately, in a radiomics task.

Future of the challenge
The leaderboard remains open on the AIcrowd platform[9]. Participants can continue to develop new segmentation algorithms and compare their results with the existing ones.

Potentially, in the next edition (HECKTOR 2021), the participants will be asked to segment also the lymph nodes and/or to perform a radiomics study. We will also try to increase the size of the dataset with new training and test cases from other centers.

Finally, radiomics studies were proposed in [20,3] to predict the prognosis of patients with H&N cancer in a non-invasive fashion. A limitation of these studies is that they were validated on 100 to 400 patients. Larger cohorts are required for estimating the generalization in real clinical settings. Manual annotations in 3D are tedious and error-prone, and the automatic tumor segmentation is an important step for large scale radiomics studies. To evaluate the feasibility of using automatic segmentation for radiomics studies, we will compare the automatic annotations to the manually delineated ones in a future work.

5 Conclusions

This paper presented a general overview of the HECKTOR challenge including the data, the participation, main results and discussions. The proposed task was the segmentation of the primary tumor in oropharyngeal cancer. The participation was relatively good with 18 results submissions and 10 participant's papers. This participation in the first edition of the HECKTOR challenge showed a high interest in automatic lesion segmentation for H&N cancer.

[9] `www.aicrowd.com/challenges/miccai-2020-hecktor/leaderboards`

The task proposed this year was to segment the primary tumor in PET/CT images. This task is not as simple as thresholding the PET image since we target only the primary tumor and the region covered by high PET activation is often too large, going beyond the limits of the tumor tissues. Deep learning methods based on U-Net models were mostly used in the challenge. Interesting ideas were implemented to combine PET and CT complementary information. Model ensembling, as well as data preprocessing and augmentation, seem to have played an important role in achieving top-ranking results.

Acknowledgments

The organizers thank all the teams for their participation and valuable work. This challenge and the winner prize are sponsored by Siemens Healthineers Switzerland. This work was also partially supported by the Swiss National Science Foundation (SNSF, grant 205320_179069) and the Swiss Personalized Health Network (SPHN, via the IMAGINE and QA4IQI projects).

Appendix 1: Challenge Information

In this appendix, we list important information about the challenge as suggested in the BIAS guidelines [15].

Challenge name
HEad and neCK TumOR segmentation challenge (HECKTOR) 2020

Organizing team
(Authors of this paper) Vincent Andrearczyk, Valentin Oreiller, Martin Vallières, Joel Castelli, Mario Jreige, John O. Prior and Adrien Depeursinge

Life cycle type
A fixed submission deadline was set for the challenge results. Open online leaderboard following the conference.

Challenge venue and platform
The challenge is associated with MICCAI 2020. Information on the challenge is available on the website, together with the link to download the data, the submission platform and the leaderboard[10].

Participation policies
(a) Algorithms producing fully-automatic segmentation of the test cases were allowed.
(b) The data used to train algorithms was not restricted. If using external data (private or public), participants were asked to also report results using only the HECKTOR data.
(c) Members of the organizers' institutes could participate in the challenge but were not eligible for awards.
(d) The award was 500 euros, sponsored by Siemens Healthineers Switzerland.
(e) Policy for results announcement: The results were made available on the AIcrowd leaderboard and the best three results were announced publicly. Once participants submitted their results on the test set to the challenge organizers via the challenge website, they were considered fully vested in the challenge, so that their performance results (without identifying the participant unless permission is granted) became part of any presentations, publications, or subsequent analyses derived from the challenge at the discretion of the organizers.
(f) Publication policy: This overview paper was written by the organizing team's members. The participating teams were encouraged to submit a paper describing their method. The participants can publish their results separately elsewhere when citing the overview paper, and (if so) no embargo will be applied.

[10] `www.aicrowd.com/challenges/hecktor`

Submission method
Submission instructions are available on the website[11] and are reported in the following. Results should be provided as a single binary mask (1 in the predicted GTVt) *.nii.gz* file per patient in the CT original resolution and cropped using the provided bounding boxes. The participants should pay attention to saving NIfTI volumes with the correct pixel spacing and origin with respect to the original reference frame. The .nii files should be named [PatientID].nii.gz, matching the patients' file names, e.g. *CHUV001.nii.gz* and placed in a folder. This folder should be zipped before submission. If results were submitted without cropping and/or resampling, we employed nearest-neighbor interpolation given that the coordinate system is provided. Participants were allowed five valid submissions. The best result was reported for each team.

Challenge schedule
The schedule of the challenge, including modifications, is reported in the following.

- the release date of the training cases: ~~June 01 2020~~ June 10 2020
- the release date of the test cases: Aug. 01 2020
- the results submission date(s): opens Sept. 01 2020 closes Sept. 10 2020
- paper submission deadline: ~~Sept. 18 2020~~ Sept. 15 2020
- the release date of the results: Sept. 15 2020
- associated workshop days: Oct. 04 2020, 9:00-13:00 UTC

Ethics approval
Training dataset: The ethics approval was granted by the Research Ethics Committee of McGill University Health Center (Protocol Number: MM-JGH-CR15-50). Test dataset: The ethics approval was obtained from the Commission cantonale (VD) d'éthique de la recherche sur l'être humain (CER-VD) with protocol number: 2018-01513.

Data usage agreement
The participants had to fill out and sign an end-user-agreement in order to be granted access to the data. The form can be found under the Resources tab of the HECKTOR website.

Code availability
The evaluation software was made available on our github page[12]. The participating teams decided whether they wanted to disclose their code (they were encouraged to do so).

Conflict of interest
No conflict of interest applies. Fundings are specified in the acknowledgments. Only the organizers had access to the test cases ground truth contours.

[11] `www.aicrowd.com/challenges/hecktor#results-submission%20format`
[12] `github.com/voreille/hecktor/tree/master/src/evaluation`

Author contributions

Vincent Andrearczyk:
Design of the task and of the challenge, writing of the proposal, development of baseline algorithms, development of the AIcrowd website, writing of the overview paper, organization of the challenge event, organization of the submission and reviewing process of the participants' papers.

Valentin Oreiller:
Design of the task and of the challenge, writing of the proposal, development of the AIcrowd website, development of the evaluation code, writing of the overview paper, organization of the challenge event, organization of the submission and reviewing process of the papers.

Mario Jreige:
Design of the task and of the challenge, quality control/annotations, annotations for inter-annotator agreement, revision of the paper and accepted the last version of the submitted paper.

Martin Vallières:
Design of the task and of the challenge, provided the initial data and annotations for the training set [20], revision of the paper and accepted the last version of the submitted paper.

Joel Castelli:
Design of the task and of the challenge, annotations for inter-annotator agreement.

Hesham Elhalawani:
Design of the task and of the challenge, annotations for inter-annotator agreement.

Sarah Boughdad:
Design of the task and of the challenge, annotations for inter-annotator agreement.

John O. Prior:
Design of the task and of the challenge, revision of the paper and accepted the last version of the submitted paper.

Adrien Depeursinge:
Design of the task and of the challenge, writing of the proposal, writing of the overview paper, organization of the challenge event.

Appendix 2: Image Acquisition Details

HGJ: All patients had FDG-PET and CT scans done on a hybrid PET/CT scanner (Discovery ST, GE Healthcare) within 37 days before treatment (median: 14 days). For the PET portion of the FDG-PET/CT scan, a median of 584 MBq (range: 368-715) was injected intravenously. Imaging acquisition of the head and neck was performed using multiple bed positions with a median of 300 s (range: 180-420) per bed position. Attenuation corrected images were reconstructed using an ordered subset expectation maximization (OSEM) iterative algorithm and a span (axial mash) of 5. The FDG-PET slice thickness resolution was 3.27 mm

for all patients and the median in-plane resolution was 3.52 x 3.52 mm^2 (range: 3.52-4.69). For the CT portion of the FDG-PET/CT scan, an energy of 140 kVp with an exposure of 12 mAs was used. The CT slice thickness resolution was 3.75 mm and the median in-plane resolution was 0.98x0.98 mm^2 for all patients.

CHUS: All 102 eligible patients had FDG-PET and CT scans done on a hybrid PET/CT scanner (GeminiGXL 16, Philips) within 54 days before treatment (median: 19 days). For the PET portion of the FDG-PET/CT scan, a median of 325 MBq (range: 165-517) was injected intravenously. Imaging acquisition of the head and neck was performed using multiple bed positions with a median of 150 s (range: 120-151) per bed position. Attenuation corrected images were reconstructed using a LOR-RAMLA iterative algorithm. The FDG-PET slice thickness resolution was 4 mm and the median in-plane resolution was 4x4 mm^2 for all patients. For the CT portion of the FDG-PET/CT scan, a median energy of 140 kVp (range: 12-140) with a median exposure of 210 mAs (range: 43-250) was used. The median CT slice thickness resolution was 3 mm (range: 2-5) and the median in-plane resolution was 1.17x1.17 mm^2 (range: 0.68-1.17).

HMR: All patients had FDG-PET and CT scans done on a hybrid PET/CT scanner (Discovery STE, GE Healthcare) within 60 days before treatment (median: 34 days). For the PET portion of the FDG-PET/CT scan, a median of 475 MBq (range: 227-859) was injected intravenously. Imaging acquisition of the head and neck was performed using multiple bed positions with a median of 360 s (range: 120-360) per bed position. Attenuation corrected images were reconstructed using an ordered subset expectation maximization (OSEM) iterative algorithm and a median span (axial mash) of 5 (range: 3-5). The FDG-PET slice thickness resolution was 3.27 mm for all patients and the median in-plane resolution was 3.52 x 3.52 mm^2 (range: 3.52-5.47). For the CT portion of the FDG-PET/CT scan, a median energy of 140 kVp (range: 120-140) with a median exposure of 11 mAs (range: 5-16) was used. The CT slice thickness resolution was 3.75 mm for all patients and the median in-plane resolution was 0.98 x 0.98 mm^2 (range: 0.98-1.37).

CHUM: All patients had FDG-PET and CT scans done on a hybrid PET/CT scanner (Discovery STE, GE Healthcare) within 66 days before treatment (median: 12 days). For the PET portion of the FDG-PET/CT scan, a median of 315 MBq (range: 199-3182) was injected intravenously. Imaging acquisition of the head and neck was performed using multiple bed positions with a median of 300 s (range: 120-420) per bed position. Attenuation corrected images were reconstructed using an ordered subset expectation maximization (OSEM) iterative algorithm and a medianspan (axial mash) of 3 (range: 3-5). The median FDG-PET slice thickness resolution was 4 mm (range: 3.27-4) and the median in-plane resolution was 4 x 4 mm^2 (range: 3.52-5.47). For the CT portion of the FDG-PET/CT scan, a median energy of 120 kVp (range: 120-140) with a median exposure of 350 mAs (range: 5-350) was used. The median CT slice thickness resolution was 1.5 mm (range: 1.5-3.75) and the median in-plane resolution was 0.98 x 0.98 mm^2 (range: 0.98-1.37). All patients received their FDG-PET/CT

scan dedicated to the head and neck area right before their planning CT scan, in the same position with the immobilization device.

CHUV (test): All patients underwent FDG PET/CT for staging before treatment. Blood glucose levels were checked before the injection of (18F)-FDG. After a 60-min uptake period of rest, patients were imaged with the Discovery D690 TOF PET/CT (General Electric Healthcare, Milwaukee, WI, USA). First, a CT (120 kV, 80 mA, 0.8-s rotation time, slice thickness 3.75 mm) was performed from the base of the skull to the mid-thigh. PET scanning was performed immediately after acquisition of the CT. Images were acquired from the base of the skull to the mid-thigh (2 min/bed position). PET images were reconstructed after time-of-flight and point-spread-function recovery corrections by using an ordered-subset expectation maximization iterative reconstruction (OSEM) (two iterations, 28 subsets) and an iterative fully 3D (Discovery ST). CT data were used for attenuation calculation.

References

1. Andrearczyk, V., Oreiller, V., Depeursinge, A.: Oropharynx detection in PET-CT for tumor segmentation. In: Irish Machine Vision and Image Processing (2020)
2. Andrearczyk, V., Oreiller, V., Vallières, M., Castelli, J., Elhalawani, H., Jreige, M., Boughdad, S., Prior, J.O., Depeursinge, A.: Automatic segmentation of head and neck tumors and nodal metastases in PET-CT scans. In: International Conference on Medical Imaging with Deep Learning (MIDL) (2020)
3. Bogowicz, M., Tanadini-Lang, S., Guckenberger, M., Riesterer, O.: Combined CT radiomics of primary tumor and metastatic lymph nodes improves prediction of loco-regional control in head and neck cancer. Scientific reports **9**(1), 1–7 (2019)
4. Chen, H., Chen, H., Wang, L.: Iteratively refine the segmentation of head and neck tumor in FDG-PET and CT images. In: Lecture Notes in Computer Science (LNCS) Challenges (2021)
5. Foster, B., Bagci, U., Mansoor, A., Xu, Z., Mollura, D.J.: A review on segmentation of positron emission tomography images. Computers in biology and medicine **50**, 76–96 (2014)
6. Ghimire, K., Chen, Q., Feng, X.: Patch-based 3D UNet for head and neck tumor segmentation with an ensemble of conventional and dilated convolutions. In: Lecture Notes in Computer Science (LNCS) Challenges (2021)
7. Gudi, S., Ghosh-Laskar, S., Agarwal, J.P., Chaudhari, S., Rangarajan, V., Paul, S.N., Upreti, R., Murthy, V., Budrukkar, A., Gupta, T.: Interobserver variability in the delineation of gross tumour volume and specified organs-at-risk during IMRT for head and neck cancers and the impact of FDG-PET/CT on such variability at the primary site. Journal of medical imaging and radiation sciences **48**(2), 184–192 (2017)
8. Hatt, M., Laurent, B., Ouahabi, A., Fayad, H., Tan, S., Li, L., Lu, W., Jaouen, V., Tauber, C., Czakon, J., Drapejkowski, F., Dyrka, W., Camarasu-Pop, S., Cervenansky, F., Girard, P., Glatard, T., Kain, M., Yao, Y., Barillot, C., Kirov, A., Visvikis, D.: The first MICCAI challenge on PET tumor segmentation. Medical Image Analysis **44**, 177–195 (February 2018)
9. Hatt, M., Lee, J.A., Schmidtlein, C.R., Naqa, I.E., Caldwell, C., De Bernardi, E., Lu, W., Das, S., Geets, X., Gregoire, V., et al.: Classification and evaluation strategies of auto-segmentation approaches for pet: Report of aapm task group no. 211. Medical physics **44**(6), e1–e42 (2017)
10. Iantsen, A., Visvikis, D., Hatt, M.: 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)
11. Kumar, A., Fulham, M., Feng, D., Kim, J.: Co-learning feature fusion maps from PET-CT images of lung cancer. IEEE Transactions on Medical Imaging (2019)
12. Li, L., Zhao, X., Lu, W., Tan, S.: Deep learning for variational multimodality tumor segmentation in PET/CT. Neurocomputing (2019)
13. Ma, J., Yang, X.: Combining CNN and hybrid active contours for head and neck tumor segmentation. In: Lecture Notes in Computer Science (LNCS) Challenges (2021)
14. Maier-Hein, L., Eisenmann, M., Reinke, A., Onogur, S., Stankovic, M., Scholz, P., Arbel, T., Bogunovic, H., Bradley, A.P., Carass, A., et al.: Why rankings of biomedical image analysis competitions should be interpreted with care. Nature communications **9**(1), 1–13 (2018)

15. Maier-Hein, L., Reinke, A., Kozubek, M., Martel, A.L., Arbel, T., Eisenmann, M., Hanbury, A., Jannin, P., Müller, H., Onogur, S., et al.: BIAS: Transparent reporting of biomedical image analysis challenges. Medical Image Analysis p. 101796 (2020)
16. Moe, Y.M., Groendahl, A.R., Mulstad, M., Tomic, O., Indahl, U., Dale, E., Malinen, E., Futsaether, C.M.: Deep learning for automatic tumour segmentation in PET/CT images of patients with head and neck cancers. In: Medical Imaging with Deep Learning (2019)
17. Naser, M., van Dijk, L., He, R., Wahid, K., Fuller, C.: Tumor segmentation in patients with head and neck cancers using deep learning based-on multi-modality PET/CT images. In: Lecture Notes in Computer Science (LNCS) Challenges (2021)
18. Rao, C., Pai, S., Hadzic, I., Zhovannik, I., Bontempi, D., Dekker, A., Teuwen, J., Traverso, A.: Oropharyngeal Tumour Segmentation using Ensemble 3D PET-CT Fusion Networks for the HECKTOR Challenge. In: Lecture Notes in Computer Science (LNCS) Challenges (2021)
19. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
20. Vallieres, M., Kay-Rivest, E., Perrin, L.J., Liem, X., Furstoss, C., Aerts, H.J., Khaouam, N., Nguyen-Tan, P.F., Wang, C.S., Sultanem, K., et al.: Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer. Scientific reports **7**(1), 1–14 (2017)
21. Xie, J., Peng, Y.: The head and neck tumor segmentation using nnU-Net with spatial and channel 'squeeze & excitation' blocks. In: Lecture Notes in Computer Science (LNCS) Challenges (2021)
22. Xu, L., Tetteh, G., Lipkova, J., Zhao, Y., Li, H., Christ, P., Piraud, M., Buck, A., Shi, K., Menze, B.H.: Automated whole-body bone lesion detection for multiple myeloma on 68Ga-pentixafor PET/CT imaging using deep learning methods. Contrast Media & Molecular Imaging (2018)
23. Yousefirizi, F., Rahmim, A.: GAN-based bi-modal segmentation using mumford-shah loss: Application to head and neck tumors in PET-CT images. In: Lecture Notes in Computer Science (LNCS) Challenges (2021)
24. Yuan, Y.: Automatic head and neck tumor segmentation in PET/CT with scale attention network. In: Lecture Notes in Computer Science (LNCS) Challenges (2021)
25. Zhao, X., Li, L., Lu, W., Tan, S.: Tumor co-segmentation in PET/CT using multi-modality fully convolutional neural network. Physics in Medicine & Biology **64**(1), 015011 (2018)
26. Zhong, Z., Kim, Y., Zhou, L., Plichta, K., Allen, B., Buatti, J., Wu, X.: 3D fully convolutional networks for co-segmentation of tumors on PET-CT images. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). pp. 228–231. IEEE (2018)
27. Zhu, S., Dai, Z., Ning, W.: Two-stage approach for segmenting gross tumor volume in head and neck cancer with CT and PET imaging. In: Lecture Notes in Computer Science (LNCS) Challenges (2021)