Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

SegRap2023: A Benchmark of Organs-at-Risk and Gross Tumor Volume *Seg*mentation for *Ra*diotherapy *P*lanning of Nasopharyngeal Carcinoma

Xiangde Luo[a,b,c], Jia Fu[a], Yunxin Zhong[d], Shuolin Liu[d], Bing Han[d], Mehdi Astaraki[e], Simone Bendazzoli[f], Iuliana Toma-Dasu[e], Yiwen Ye[g], Ziyang Chen[g], Yong Xia[g], Yanzhou Su[c], Jin Ye[c], Junjun He[c], Zhaohu Xing[h], Hongqiu Wang[h], Lei Zhu[h], Kaixiang Yang[i], Xin Fang[i], Zhiwei Wang[i], Chan Woong Lee[j], Sang Joon Park[k], Jaehee Chun[l], Constantin Ulrich[m], Klaus H. Maier-Hein[m], Nchongmaje Ndipenoch[n], Alina Miron[n], Yongmin Li[n], Yimeng Zhang[o], Yu Chen[o], Lu Bai[o], Jinlong Huang[p], Chengyang An[p], Lisheng Wang[p], Kaiwen Huang[q], Yunqi Gu[q], Tao Zhou[q], Mu Zhou[c], Shichuan Zhang[b], Wenjun Liao[b], Guotai Wang[a,c,*], Shaoting Zhang[a,c,*]

[a] *School of Mechanical and Electrical Engineering, University of Electronic Science and Technology of China, Chengdu, China.*
[b] *Department of Radiation Oncology, Sichuan Cancer Hospital & Institute, Chengdu, China.*
[c] *Shanghai AI Lab, Shanghai, China.*
[d] *Canon Medical Systems (China) Co. Ltd., Beijing, China.*
[e] *Department of Medical Radiation Physics, Stockholm University, Solna, Sweden.*
[f] *Department of Biomedical Engineering and Health Systems, KTH, Huddinge, Sweden.*
[g] *National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application Technology, School of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an, China.*
[h] *Department of Systems Hub, Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China.*
[i] *Wuhan National Laboratory for Optoelectronics and with MoE Key Laboratory for Biomedical Photonics, School of Engineering Sciences, Huazhong University of Science and Technology.*
[j] *Medical Physics and Biomedical Engineering Lab (MPBEL), Yonsei University College of Medicine, Seoul, South Korea.*
[k] *Department of Radiation Oncology, Yonsei Cancer Center, Heavy Ion Therapy Research Institute, Yonsei University College of Medicine, Seoul, South Korea.*
[l] *Oncosoft Inc. Seoul, South Korea.*
[m] *German Cancer Research Center (DKFZ) Heidelberg, Division of Medical Image Computing, Germany*
[n] *Department of Computer Science, Brunel University London, Uxbridge, United Kingdom.*
[o] *MedMind Technology Co. Ltd., Beijing, China.*
[p] *Department of Automation, Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai, China.*
[q] *School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China.*

ARTICLE INFO

Article history:

Keywords: Nasopharyngeal carcinoma, organ-at-risk, gross tumor volume, Segmentation

ABSTRACT

Radiation therapy is a primary and effective treatment strategy for NasoPharyngeal Carcinoma (NPC). The precise delineation of Gross Tumor Volumes (GTVs) and Organs-At-Risk (OARs) is crucial in radiation treatment, directly impacting patient prognosis. Despite that deep learning has achieved remarkable performance on various medical image segmentation tasks, its performance on OARs and GTVs of NPC is still limited, and high-quality benchmark datasets on this task are highly desirable for model development and evaluation. To alleviate this problem, the SegRap2023 challenge was organized in conjunction with MICCAI2023 and presented a large-scale benchmark for OAR and GTV segmentation with 400 Computed Tomography (CT) scans from 200 NPC patients, each with a pair of pre-aligned non-contrast and contrast-enhanced CT scans. The challenge aimed to segment 45 OARs and 2 GTVs from the paired CT scans per patient, and received 10 and 11 complete submissions for the two tasks, respectively. In this paper, we detail the challenge and analyze the solutions of all participants. The average Dice similarity coefficient scores for all submissions ranged from 76.68% to 86.70%, and 70.42% to 73.44% for OARs and GTVs, respectively. We conclude that the segmentation of relatively large OARs is well-addressed, and more efforts are needed for GTVs and small or thin OARs. The benchmark remains available at: SegRap2023.

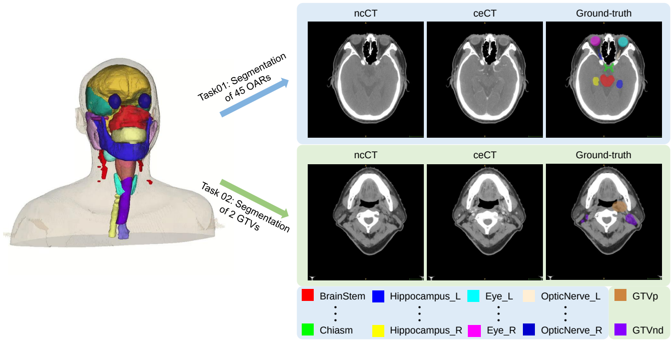


Fig. 1. Overview of two sub-tasks in the SegRap2023 challenge.

1. Introduction

1.1. Clinical background

NasoPharyngeal Carcinoma (NPC), a malignant tumor originating in the nasopharyngeal region, is particularly prevalent in Southeast Asia and North Africa (Lee et al., 2015; Chua et al., 2016; Sun et al., 2019). The primary treatment modality for NPC relies heavily on radiation therapy, especially Intensity-Modulated Radiation Therapy (IMRT) (Xia et al., 2000; Kam et al., 2003). In IMRT, the accurate delineation of the Gross Tumor Volumes (GTVs) and the surrounding Organs-At-Risk (OARs) is crucial for treatment effectiveness. Accurately identifying the target area is essential to ensure that high doses of radiation precisely cover the tumor while protecting the adjacent normal tissues (Tang et al., 2019). Proper delineation of the GTVs enhances local control rates of the treatment and reduces the risk of recurrence. NPC is located near several vital structures, such as the skull base, internal carotid arteries, and optic nerves (Wang and Kang, 2021). Inaccurate delineation may expose these OARs to unnecessarily high doses of radiation, increasing the risk of acute and delayed radiation-induced damage (Lin et al., 2019).

Accurate delineation of OARs and GTVs is a significant challenge for junior radiation oncologists and automated delineation methods (Chen et al., 2021). Firstly, the anatomical structure of the nasopharyngeal region is inherently complex, being near critical organs and neural structures such as the skull base, internal carotid arteries, and optic nerves. This complexity makes the accurate delineation of the target area and OARs extremely challenging and prone to errors (Tang et al., 2019). Secondly, the tumor size, shape, and location vary among NPC patients, coupled with individual anatomical differences, which further complicates the delineation process (Lee et al., 2018). Additionally, the low contrast and ambiguous boundary between OAR or GTV and other soft tissues in CT images lead to difficulties in the delineation of OAR and GTV, radiation oncologists usually require other modality images for complementary guidelines to perform delineation. Moreover, the reliance on the experience and judgment of physicians for delineating the target area and OARs introduces potential variability and subjectivity among different practitioners, potentially leading to inconsistencies in treatment planning. In past clinical practices, the delineation of OARs and GTVs in NPC was predominantly conducted by experienced radiation oncologists. However, according to the clinical treatment guideline, each patient has more than 40 OARs and 2 GTVs need to be delineated accurately (Ye et al., 2022; Guo et al., 2020). It requires the radiation oncologists to spend much time performing delineation, increasing the annotator's burden and patient waiting time. It's desirable to develop efficient and accurate automatic segmentation tools to assist and accelerate the clinical delineation workflow and reduce the annotator's burden and patient waiting time.

1.2. Technical challenges

Deep learning-based segmentation methods have shown promising performance on certain medical segmentation datasets, such as abdominal organ segmentation (Luo et al., 2021b; Isensee et al., 2021; Gibson et al., 2018; Bilic et al., 2019) and thoracic organ segmentation (Dong et al., 2019; Feng et al., 2019). However, there remains a notable scarcity of studies reporting automatic segmentation tools for OARs and GTVs in NPC that achieve clinically applicable performance on large-scale datasets. The automation of OAR and GTV segmentation remains challenging due to inherent characteristics, including size, shape, and location variations among NPC patients, compounded by individual anatomical differences and ambiguous boundaries. Moreover, creating and annotating a large-scale, high-quality dataset for OAR and GTV segmentation is a resource-intensive process, demanding both expertise and time to generate accurate delineations. Consequently, there is still a lack of large-scale and high-quality annotated datasets for developing automatic segmentation models for NPC OARs and GTVs.

Recently, few studies have reported in detail the segmentation results of GTVs and OARs of NPC (Liu et al., 2021; Lin et al., 2019; Luo et al., 2023, 2021a; Liao et al., 2022; Ye et al., 2022; Guo et al., 2020; Shi et al., 2022; Tang et al., 2019; Wu et al., 2024). Most of them only focused on the segmentation of part of the OARs or the GTVs of head and neck cancers. For example, Shi et al. (2022) and Ye et al. (2022) evaluated the performance on 27 head OARs and 42 head and neck OARs, respectively. In addition, few works investigated the model segmentation performance on multiple inputs, such as non-contrast or contrast-enhancement CT scans (Wang et al., 2020; Oreiller et al., 2022). The limited number of OARs and using single-modality in these existing works limited the performance and clinical application of the segmentation models. Therefore, a large-scale benchmark with exhausted and high-quality annotations and multiple modalities is highly desired for boosting the development of OAR and GTV segmentation models for the radiation treatment of NPC.

1.3. Contribution

To comprehensively evaluate the performance of state-of-the-art (SOTA) algorithms for automatic OAR and GTV segmentation in the radiation treatment planning of NPC, we or-

*Corresponding authors
e-mail: `guotai.wang@uestc.edu.cn` (Guotai Wang), `Rutgers.shaoting@gmail.com` (Shaoting Zhang)

Table 1. Summary of several publicly available organ-at-risk segmentation Computed Tomography (CT) datasets. ceCT is the contrast-enhanced Computed Tomography. ncCT means the non-contrast Computed Tomography.

Dataset	Modality	No. of categories	Scans (Training/Testing)	Year	Link
PDDCA	ncCT	9 OARs	48 (33/15)	2015	`www.imagenglab.com/newsite/pddca`
HNC	ncCT	28 OARs	35 (18/17)	2015	`https://wiki.cancerimagingarchive.net/x/xwxp`
HNPETCT	ncCT	28 OARs	105 (52/53)	2017	`https://doi.org/10.7937/K9/TCIA.2017.8oje5q00`
StrucSeg2019	ncCT	22 OARs	60 (50/10)	2019	`https://structseg2019.grand-challenge.org`
HaN-Seg2023	ncCT and MRI	30 OARs	56 (42/16)	2023	`https://han-seg2023.grand-challenge.org`
SegRap2023	ncCT and ceCT	**45 OARs**	**200 (140/60)**	**2023**	`https://segrap2023.grand-challenge.org`

ganized the SegRap2023 challenge in conjunction with MICCAI2023. The key contributions of this work can be summarized as three-fold. Firstly, we built the first large-scale public dataset of 200 NPC patients where each patient has pre-aligned non-contrast and contrast-enhanced CT scans with high-quality manual annotations of 45 OARs and 2 GTVs. Secondly, the SegRap2023 challenge was successfully organized during MICCAI2023 via the grand challenge platform which attracted a total of 387 teams registered during the model development phase. In the final evaluation phase, 10 and 11 teams successfully submitted their solutions for the OARs and GTVs tasks, respectively. Thirdly, we evaluated, ranked, summarized, analyzed, and discussed the results of all submissions. The results demonstrated that the large-size OAR segmentation is well-addressed, and more attention needs to be paid to GTV and small-size or thin-structure OAR segmentation. We believe this dataset and challenge can bring benefits to the whole community.

This paper summarizes the SegRap2023 challenge and is organized as follows. Section 2 reviews the existing datasets and methods for OAR and GTV segmentation. Then, Section 3 presents the details of the challenge in the aspects of data collection and annotation, challenge organization and evaluation. Details of all submitted methods are illustrated in Section 4. Afterwards, the analysis and description of the results are presented in Section 5. Finally, we conclude and discuss the SegRap2023 challenge in Section 6 and 7, respectively.

2. Related Works

2.1. OAR segmentation in head and neck cancers

2.1.1. Benchmarks and datasets

OAR segmentation plays an irreplaceable role in radiation therapy planning of Head and Neck Cancers (HNC). Developing an accurate and robust automatic segmentation model always relies on large-scale annotated datasets. However, publicly available datasets are very limited because collecting and annotating a large-scale dataset are challenging due to high expenses and data privacy protection. (Wang et al., 2023a; Simpson et al., 2019). Table 1 summarizes several public datasets for OAR segmentation in the head and neck region. PDDCA (Raudaschl et al., 2017) provides 48 CT scans with 9 OARs annotated for the Head and Neck Auto Segmentation MICCAI Challenge (2015). HNC (Ang et al., 2014) and HNPETCT (Vallieres et al., 2017) consist of 35 and 105 CT scans from head and neck cancer patients, respectively, and all of them have annotations of 28 OARs. Tang et al. (2019) selected 35 CT scans from HNC and 105 CT scans from HNPETCT for further annotation and released all masks for public research, where each patient has 28 OAR labels. StructSeg2019 (Podobnik et al., 2023) organized a head and neck OAR segmentation from CT and Magnetic Resonance Imaging (MRI) challenge conjoint with MICCAI2023. The HaN-Seg2023 consists of 56 patients with head and neck cancer and each patient has a CT and a T1-weighted MRI scan and a reference annotation with 30 OARs.

Although these datasets have facilitated the methods research of head and neck OAR segmentation in the community, they may be still not enough to develop clinically applicable segmentation tools and provide comprehensive evaluations due to the small number of cases and annotated OARs. In other medical image segmentation tasks, such as abdominal organ segmentation (Luo et al., 2021b; Gibson et al., 2018; Bilic et al., 2019), many large-scale datasets have been available for foundation model development and evaluation, and also advance the automatic segmentation methods to be applied in clinical practice (Chen et al., 2021; Kirillov et al., 2023; Huang et al., 2023; Wang et al., 2023d,b). Therefore, for the head and neck OAR segmentation, it is desirable to build a large-scale dataset and benchmark to boost technical improvements and clinical application development.

2.1.2. HNC OAR segmentation methods

Recently, deep learning-based segmentation methods have shown superiority in producing more accurate and robust than previous atlas-based counterparts (Tang et al., 2019; Kosmin et al., 2019; Chen et al., 2021). FocusNet (Gao et al., 2019) incorporates densely connected atrous spatial pyramid pooling and squeeze-and-excitation modules into the main segmentation network for OAR segmentation. FocusNetV2 (Gao et al., 2021) presents a two-stage framework to locate and segment OARs progressively by combining the multi-scale convolutional neural network and a shape adversarial constraint. It was evaluated on a large-scale private nasopharyngeal cancer dataset with 1164 CT scans and 22 OARs and the public PDDCA dataset and showed a mean dice score of 82.98% and 84.50%, respectively. UaNet (Tang et al., 2019) proposes a combination framework to detect OARs and segment them step-by-step, which was trained on a private dataset with 215 CT scans and 28 OARs and tested on 100 CT scans with a mean dice score of 78.34%.

Recently, Guo et al. (2020) and Ye et al. (2022) developed an auto-contouring system (SOARS) by combining the neural architecture search strategy and an organ-level stratification learning. The proposed SOARS was trained on an internal private dataset with 176 CT scans and 42 OARs and independently evaluated on several external cohorts with a total of 1327 CT scans with mean dice scores ranging from 74.80% to 78.00%. Additionally, He et al. (2024) introduced a statistical deformation model-based data augmentation strategy to boost the training set's diversity and realism and further advance the model performance. The proposed was trained and tested on the HN-PETCT dataset and achieved a mean dice score of 79.49%. Lei (Lei et al., 2021) proposed a segmental linear function to make organs more distinguishable and introduced a hardness-aware loss function to emphasize the learning of hard voxels. It was evaluated on StructSeg 2019 challenge data and achieved a weighted average Dice of 80.52%. These reported results show that the performance of existing OAR segmentation methods varies significantly on different datasets. Especially the results on private datasets were higher than those on the public datasets (Zhu et al., 2019; Tang et al., 2019; Ye et al., 2022; Gao et al., 2021; He et al., 2024; Chen et al., 2021). Therefore, building a large-scale public benchmark for a fair comparison across multiple state-of-the-art methods is essential.

2.2. *NPC GTV segmentation*

2.2.1. *Benchmarks and datasets*

For the GTV segmentation of HNC, the public dataset HECKTOR [1] was available for model development and evaluation. HECKTOR (Oreiller et al., 2022) challenge has been organized in conjunction with MICCAI in recent three years, which aims to encourage all participants to develop cut-edge primary gross tumor volume (GTVp) and the lymph node gross tumor volume (GTVnd) segmentation models from CT and FDG-PET scans. The total number of patients increased from 254 patients just with GTVp annotation in HECKTOR2020 to more than 880 patients with both GTVp and GTVnd annotations in HECKTOR2022. For NPC GTV segmentation, the StructSeg2019provides 60 nasopharyngeal carcinoma patients' CT scans and each patient had a GTVp annotation. Although the HECKTOR challenge provides a large-scale dataset for GTVp and GTVnd segmentation, they focus on head and neck cancer rather than nasopharyngeal carcinoma, so the SegRap2023 is still an important dataset for the GTVp and GTVnd of NPC segmentation.

2.2.2. *SOTA NPC GTV segmentation methods*

Unlike OAR segmentation, GTV segmentation has traditionally been conducted by experienced radiation oncologists in clinical practice. This is attributed to the intricate nature of GTV structures and their significant correlation with prognosis. Moreover, the scarcity of publicly available datasets has been a notable challenge in the field. Many prior studies have reported GTV segmentation outcomes based on private datasets, posing difficulties for both reproducibility and equitable comparisons in the whole community. Li et al. (2019) trained a basic U-Net (Ronneberger et al., 2015) to segment GTVp and GTVnd using a large-scale private dataset with 502 CT scans and achieved a mean dice of 65.86% and 74.00% for GTVp and GTVnd, respectively. Lin et al. (2019) developed a 3D segmentation model on an MRI dataset with 1021 patients to segment the GTVp and reported the performance with a mean dice score of 79.00%. Mei et al. (2021) proposed a 2.5D segmentation network with multi-scale and spatial attention to segment GTVp from CT scans and won second place in the StructSeg2019 challenge with a mean dice of 65.66%.

Table 2. Clinical characteristics of the SegRap2023 training, validation and testing sets. * means the values are presented as median (range). T and N stages denote the tumor and lymph node staging according to the AJCC2017 standardized classification system (Amin et al., 2017).

Characteristics	Training (n=120)	Validation (n=20)	Testing (n=60)
Sex			
Male	81 (67.5%)	12 (60%)	37 (61.7%)
Female	39 (32.5%)	8 (40%)	23 (38.3%)
Age* (years)	48 (22-74)	50 (36-69)	47 (22-70)
T stage			
T1	12 (10%)	2 (10%)	7 (11.7%)
T2	27 (22.5%)	5 (25%)	13 (21.7%)
T3	62 (51.7%)	11 (55%)	32 (53.3%)
T4	19 (15.8%)	2 (10%)	8 (13.3%)
N stage			
N0	10 (8.3%)	1 (5%)	4 (6.7%)
N1	24 (20%)	3 (15%)	11 (18.3%)
N2	54 (45%)	11 (55%)	31 (51.7%)
N3	32 (26.7%)	4 (20%)	14 (23.3%)
Resolution (mm)			
Inter-plane	3.0	3.0	3.0
Intra-plane*	0.55 (0.43-1.13)	0.54 (0.49-0.60)	0.59 (0.45-1.34)

In addition, Luo et al. (2021a) proposed a multi-scale consistency-based semi-supervised learning framework to utilize the unlabeled data for GTVp and GTVnd segmentation performance improvement, and further demonstrated the applicable in the clinical delineation flow on a private MRI dataset with 258 patients (Liao et al., 2022), where the mean dice scores of GTVp and GTVnd were 83.00% and 80.00%, respectively. Recently, Luo et al. (2023) conducted a comprehensive evaluation of GTVp segmentation using a total number of 1057 patients from 5 hospitals and achieved a mean dice score of 88.00% on the multi-center testing cohorts. According to these observations, it can be noted that there is a substantial variation in segmentation results across different datasets. Meanwhile, despite that MRI provides a higher soft tissue contrast for GTVs than CT, the current radiotherapy treatment method is mostly based on CT scans, so accurately contouring the GTVs of NPC from CT scans is still challenging and urgent (Sahbaee et al., 2017).

3. SegRap2023 challenge setup

3.1. *Challenge overview*

To evaluate existing methods and boost the development of novel ones for OAR and GTV segmentation, we organized the SegRap2023 challenge in conjunction with MICCAI2023. The challenge released 400 CT scans from 200 NPC patients where each patient has a pre-aligned pair of ncCT and ceCT scans. Figure 1 shows an overview of the SegRap2023 challenge. The challenge consists of two sub-tasks. The first one (Task01) is to segment 45 OARs, and the second task (Task02) is to segment 2 GTVs.

[1] https://hecktor.grand-challenge.org

3.2. Data description

The SegRap2023 dataset consists of 200 NPC patients from Sichuan Cancer Hospital & Institute, Sichuan Cancer Center, Chengdu, China. The data acquisition was approved by the Sichuan Cancer Hospital & Institute ethics board[2] and the private information of each patient has been anonymized and shared with the license of Creative Commons license Attribution-Noncommercial (CC BY-NC). Each patient has a ncCT scan and a ceCT scan. All CT scans are collected by Siemens CT scanners with the following scanning conditions: bulb voltage, 120 kV; current, 300 mA; scan thickness, 3.0 mm; matrix size, 1024 × 1024 or 512 × 512; injected contrast agent, iohexol (volume, 60−80 mL; rate, 2 mL/s; delay, 50 s). Table 2 lists the clinical characteristics of the training, validation, and testing sets. It can be found that there is a similar distribution of clinical characteristics in the training, validation, and testing sets (age, sex, T and N stages, and inter- or intra-plane spacings). To build the dataset, we retrospectively collected 200 newly treated NPC patients from December 2018 to December 2019. The inclusion criteria were defined as (a) Patients who were histologically confirmed as NPC in the M.D. S.C. Zhang treatment group; (b) The treatment strategy included radiotherapy; (c) The radiotherapy planning had ncCT and ceCT scans that were acquired before the first radiation therapy for each patient and 45 OARs and 2 GTVs annotations; (d) Patients who are alive and not recurrent until December 2022.

The initial contours of OARs and GTVs were delineated by S.C. Zhang (MD, with more than twenty years of experience in oncology radiation therapy) and their team (mainly including M.D. W. Liao, M.D. Y. Zhao, and M.D. C. Li, all of them are with more than ten years of experience in oncology radiation therapy) using MIM Software [3] according to the latest radiation therapy delineation guideline published by Radiation Therapy Oncology Group [4]. The MIM software is a widely used commercial radiotherapy planning software for OARs and GTVs delineations, which provides the Atlas-based automatic OARs segmentation algorithms (Iglesias and Sabuncu, 2015) and allows the oncologists to edit the contours. In the real clinical workflow, the radiation oncologists will adjust or re-contour the Atlas-generated initial OARs' contours and delineate the GTVs' contours manually until these contours are acceptable for radiotherapy planning. Besides, during the initial delineation stage, the radiation oncologists referred to other images (MRI, PET) for clear contours, especially for the GTV delineation. To ensure high-quality annotations, we invited W. Liao and S.C. Zhang to check and refine these annotations using ITK-SNAP (Yushkevich et al., 2006). Here, we also presented the performance between initial Atlas-based automatic OARs segmentation (Iglesias and Sabuncu, 2015) and the final ground truth in the testing set on Table 8 and 9, the significant performance gaps mean that the annotation quality is not subject to the Atlas segmentation bias. Note that some small, challenging and uncommon organs can not be segmented using Atlas-based methods, so we just listed the performance of successfully segmented organs. These annotated 45 OARs are the Brain, BrainStem, Chiasm, Cochlea left (Cochlea_L), Cochlea right (Cochlea_R), Esophagus, Eustachian tube bone left (ETbone_L), Eustachian tube bone right (ETbone_R), Eye left (Eye_L), Eye right (Eye_R), Hippocampus left (Hippocampus_L), Hippocampus right (Hippocampus_R), Internal auditory canal left (IAC_L), Internal auditory canal right (IAC_R), Larynx, Larynx glottic (Larynx_Glottic), Larynx supraglottic (Larynx_Supraglot), Lens left (Len_L), Lens right (Len_R), Mandible left (Mandible_L), Mandible right (Mandible_R), Mastoid left (Mastoid_L), Mastoid right (Mastoid_R), Middle Ear left (MiddleEar_L), Middle ear right (MiddleEar_R), Optic nerve left (OpticNerve_L), Optic nerve right (OpticNerve_R), Oral cavity, Parotid left (Parotid_L), Parotid right (Parotid_R), Pharyngeal constrictor muscle (PharynxCont), Pituitary, SpinalCord, Submandibular left (Submandibular_L), Submandibular right (Submandibular_R), Temporal lobe left (TemporalLobe_L), Temporal lobe right (TemporalLobe_R), Thyroid, Temporomandibular joint left (TMjoint_L), Temporomandibular joint right (TMjoint_R), Trachea, Tympanic cavity left (TympanicCavity_L), Tympanic cavity right (TympanicCavity_R), Vestibular semicircular canal left (VestibulSemi_L), Vestibular semicircular canal right (VestibulSemi_R). Note that, different classes may have an overlap, for example, Brain and BrainStem, Larynx and Larynx_Glottic. The 2 annotated GTVs are GTVp and GTVnd. Afterwards, we provided a random split including training, validation, and testing sets with 120, 20, and 60 patients, respectively, according to clinical characteristics, as detailed in Table 2.

3.3. Evaluation and rank strategies

The challenge employed two widely used evaluation metrics to measure the performance of each submission: (1) a region overlap-based metric, Dice Similarity Coefficient (DSC) that ranges from 0.0 to 1.0, and (2) a distance-aware metric, Normalized Surface Dice (NSD) that ranges from 0.0 to 1.0 (Nikolov et al., 2021):

$$DSC(P, Y) = \frac{2\,|V_P \cap V_Y|}{|V_P| + |V_Y|} \tag{1}$$

$$NSD(P, Y) = \frac{\left|S_P \cap S_Y^{(\tau)}\right| + \left|S_Y \cap S_P^{(\tau)}\right|}{|S_P| + |S_Y|} \tag{2}$$

where V_P and V_Y in Eq. 1 denote the predicted segmentation results and the ground truth, respectively. In Eq. 2, S_P and S_Y denote two sets of nearest-neighbor distances, and $S_P^{(\tau)}$ and $S_Y^{(\tau)}$ denote the subsets of distances that are not larger than the acceptable distance τ, which is set as 1 *mm* according to the median intra-plane spacing for all classes in the test phase of the SegRap2023 challenge except for Larynx is set as 2 *mm*. If a submission has some missing target OARs or GTVs on test cases, the corresponding DSC and NSD will be set to 0. Then, we calculated the average DSC and NSD of each OAR or GTV across all testing patients, respectively. Afterwards,

[2] https://github.com/HiLab-git/SegRap2023/blob/main/ethics.pdf

[3] https://www.mimsoftware.com

[4] https://www.rtog.org

we followed Bakas et al. (2018) to rank all the participants according to the value of each metric on each segmentation class respectively, and each team has 45 × 2 and 2 × 2 ranking scores for OAR and GTV segmentation tasks, respectively. Finally, for each task, we employed the average ranking of each team for the final ranking.

3.4. Challenge setup

In the SegRap2023 challenge, we designed two sub-tasks: Segmentation of 45 OARs (Task01) and GTVs (Task02). The challenge consists of three phases (training, validation and testing) and all of them were hosted in the grand challenge platform [5]. During the training stage, the training set can be accessed for all participants by signing and sending back an end-user agreement file, which has been made still publicly available in the community after the challenge. The validation phase was open from July 10th, 2023 to August 20th, 2023 and each team was allowed to submit 5 times. In addition, we also provided the evaluation on our local machine if the participants sent their predictions for the validation set to us. That is because some participants can not submit their evaluation docker successfully and are also limited by the computation costs, which are too high to afford their evaluation online many times. It is worth noting that this process evaluates the model performance on the validation set, and no participant can access the test set to ensure a fair comparison.

In the final testing phase, due to the testing set is not accessible (Maier-Hein et al., 2020), each team was required to submit their solution Docker container for evaluation and ranking. We provided a tutorial [6] to containerize the algorithm with Docker. Each team was only allowed to submit the Docker container once successfully. All submitted Docker containers were run on the grand challenge platform after being submitted successfully. The segmentation performance was calculated online using an automatic evaluation Docker container with two public Python packages (*Evalutils* [7] and *MedPy* [8]). Finally, the final leaderboard was announced in the MICCAI2023 challenge event after the organization team carefully reviewed and excluded the teams without submitting their technical reports.

4. Overview of participating methods

A total of 387 teams registered for the SegRap2023 Challenge, allowing them to download the training data. During the testing phase, there were 10 and 11 teams that successfully submitted the containerized algorithms and met the submission requirements for Task01 and Task02, respectively. In this section, we summarize the methods employed by the participating teams (two teams were excluded due to the lack of their technical report). Table 3 and Table 4 summarize the key techniques of benchmarked algorithms for Task01 and Task02, respectively. Table 5 and Table 6 summarize the training details of benchmarked algorithms for task01 and task02, respectively. More details and references can be found at: `https://github.com/HiLab-git/SegRap2023`.

4.1. Task01: OAR segmentation

Almost all teams submitted deep learning-based methods based on nnUNet (Isensee et al., 2021) structure. All teams used similar loss functions (mainly the combination of Dice and CE loss), and six of them used an ensemble learning method. Two of the top five teams used two-stage approaches, and one team used a pre-trained model. In this task, we provided a baseline based on the nnUNet (Isensee et al., 2021) for model training, docker preparation and inference evaluation. When establishing the baseline, we noticed that nnUNet, with its default data augmentation strategies, did not achieve promising performance on symmetrical, small, and complex organs. Upon further investigation, we found that spatial augmentations, such as mirror/flipping, disrupt spatial symmetry, while elastic transformations increase the training time and do not lead to performance gain. So, we modified the default nnUNet as the baseline by removing the mirror/flipping, and elastic transformations from the default augmentation strategy to train the model and also removing the test-time augmentation for inference.

(1st place, Y. Zhong *et al.*) Zhong *et al.* proposed a two-stage approach to segment OARs: structure-specific label generation and boundary refinement. For structure-specific label generation, 45 organs are divided into 29 distinct classes considering the left and right counterparts and label overlapping in the ear and oral cavity. The segmentation model was built based on nnUNetV2 (Isensee et al., 2021) and trained with paired ncCT and ceCT scans. For boundary refinement, ROIs with a size of 128 × 128 × 128 were extracted based on the segmentation result and refined using a model with a shared encoder-decoder architecture, but different output layers for each organ. Finally, the refined ROI was then integrated back into the original segmentation results.

(2nd place, Y. Ye *et al.*) Ye *et al.* employed the UniSeg (Ye et al., 2023), a supervised pre-trained nnUNet model trained on multiple segmentation datasets. To fine-tune the UniSeg model to OAR segmentation, the images were first pre-processed following nnUNet (Isensee et al., 2021) and then resampled to match the median spacing. Then, UniSeg was trained with 1500 epochs and 2000 epochs using paired ncCT and ceCT images. During inference, the image was pre-processed with nnUNet's pre-processing step, then segmented into patches using a sliding window approach, and the two predictions for each patch from two fine-tuned UniSeg models were averaged to form the final segmentation map.

(3rd place, Y. Su *et al.*) Su *et al.* used a vanilla nnUNet (Isensee et al., 2021) for OAR segmentation, incorporating data augmentation techniques, including additive brightness, gamma correction, rotation, scaling, and elastic deformation. Given the symmetry of head and neck organs, mirror operation was not used. The model was trained with an increased patch size (48 × 256 × 256) to improve segmentation performance.

[5] `https://grand-challenge.org`
[6] `https://github.com/HiLab-git/SegRap2023`
[7] `https://evalutils.readthedocs.io/en/latest`
[8] `https://loli.github.io/medpy`

Table 3. Summary of the benchmarked algorithms for Task01. IN means intensity normalization. IH means intensity harmonization. SA means simple augmentation techniques, including random rotation, random scaling, ransom shifting, random cropping, and random warping. CC means Connected component-based post-processing and CDA means Connectivity Domain Algorithm for splitting the paired organs into left and right parts.

Team	Pre-processing	Pre-train	Two-stage	Data augmentation	Post-processing
Y. Zhong *et al.*	Crop, IN, resample	×	✓	Rotation, scaling, Gaussian noise, Gaussian blur, brightness, contrast, gamma, elastic	CC, CDA
Y. Ye *et al.*	Crop, IN, resample	✓	×	Rotation, scaling, Gaussian noise, Gaussian blur, brightness, contrast, gamma, elastic	None
Y. Su *et al.*	Crop, IN, resample	×	×	Rotation, scaling, Gaussian noise, Gaussian blur, brightness, contrast, gamma, elastic	None
K. Yang *et al.*	Crop, IN, resample	×	×	Rotation, scaling, Gaussian noise, Gaussian blur, brightness, contrast, gamma, elastic	CC
C. Lee *et al.*	Crop, resample	×	✓	Rotation, scaling, Gaussian noise, Gaussian blur, contrast	None
M. Astaraki *et al.*	IH, crop	×	×	Rotation, scaling, Gaussian noise, Gaussian blur, brightness, contrast, gamma	None
Z. Xing *et al.*	Crop, IN, resample	×	×	SA, mirror, Gaussian noise, Gaussian blur, brightness, contrast, gamma	None
Y. Zhang *et al.*	Crop, IN, Resample	×	×	SA, mirror, Gaussian noise, Gaussian blur, brightness, contrast, gamma	None
J. Huang *et al.*	Crop, IN, Resample	×	✓	Rotation, scaling, Gaussian noise, Gaussian blur, brightness, contrast, gamma	None
K. Huang *et al.*	Crop, IN	×	×	SA, brightness, contrast	None

Table 4. Summary of the benchmarked algorithms for Task02. IH means intensity harmonization. IN means intensity normalization. SA means simple augmentation techniques, including random rotation, random scaling, ransom shifting, random cropping, random warping.

Team	Pre-processing	Pre-train	Two-stage	Data augmentation
M. Astaraki et al.	IH, crop	×	×	Rotation, scaling, Gaussian noise, Gaussian blur, brightness, contrast, gamma, mirror
Y. Ye et al.	Crop, IN	✓	×	Rotation, scaling, Gaussian noise, Gaussian blur, brightness, contrast, gamma, mirror
Z. Xing et al.	Crop, IN, resample	×	×	SA, gaussian noise, gaussian blur, brightness, contrast, gamma
K. Yang et al.	Crop, IN, resample	×	×	Rotation, scaling, Gaussian noise, Gaussian blur, brightness, contrast, gamma, mirror
C. Ulrich et al.	Crop, IN, resample	✓	×	Rotation, scaling, Gaussian noise, Gaussian blur, brightness, contrast, simulation of low resolution, gamma, mirror
N. Ndipenoch et al.	Crop, IN, resample	×	×	Rotation, scaling, Gaussian noise, Gaussian blur, brightness, contrast, gamma, mirror
Y. Su et al.	Crop, IN, resample	×	×	Rotation, scaling, Gaussian noise, Gaussian blur, brightness, contrast, gamma, elastic
J. Huang et al.	Crop, IN, resample	×	✓	Rotation, scaling, Gaussian noise, Gaussian blur, brightness, contrast, gamma
Y. Zhang et al.	Crop, IN, resample	×	×	SA, Gaussian noise, Gaussian blur, brightness, contrast, gamma
C. Lee et al.	Crop, resample	×	✓	Rotation, scaling, Gaussian noise, Gaussian blur, contrast, mirror
K. Huang et al.	Crop, IN	×	×	SA, brightness

Table 5. Network architectures and training details of the benchmarked algorithms for Task01. CE and BCE mean cross-entropy and binary cross-entropy, respectively. × (*) refers to the number of ensemble models.

Team	Architecture	Ensemble (size)	Batch size	Patch Size	Loss function	Optimizer	Learning rate	Device
Y. Zhong et al.	nnUNetV2, nnUnetV1	× (5)	4	56×192×160 128×128×128	Dice and CE	SGD	0.01	NVIDIA A800
Y. Ye et al.	nnUNet	× (2)	2	32×192×192	Dice and CE	SGD	0.01	NVIDIA Geforce RTX 2080Ti
Y. Su et al.	nnUNetV2	None	2	48×256×256	Dice and CE	SGD	0.01	NVIDIA A100
K. Yang et al.	nnUNet	None	2	28×224×224	Dice and CE	SGD	0.01	TITAN RTX 24G
C. Lee et al.	yolo-v7 + UNet	× (5)	4	32×96×96 32×128×128	Dice and CE	AdamW	1e-4	NVIDIA A5000
M. Astaraki et al.	nnUNetV2	× (5)	2	64×192×160	BCE and Dice	SGD	0.01	Nvidia DGX-1 Cluster
Z. Xing et al.	nnUNet	× (3)	2	64×256×256	Dice and CE	SGD	0.01	NVIDIA A100 GPU
Y. Zhang et al.	nnUNet	None	2	64×192×160	Dice and CE	SGD	0.01	NVIDIA Geforce RTX 3090
J. Huang et al.	nnUNetV2	× (4)	2	40×256×160	Dice and CE	SGD	0.01	NVIDIA Geforce RTX 3090
K. Huang et al.	nnUNetV2	None	2	24×224×224	Soft-dice and CE	AdamW	1e-3	NVIDIA Geforce RTX 2080Ti

Table 6. Network architectures and training details of the benchmarked algorithms for Task02. CE and BCE mean cross-entropy and binary cross-entropy, respectively. SE means Squeeze-and-Excitation. × (*) refers to the number of ensemble models.

Team	Architecture	Ensemble (size)	Batch size	Patch Size	Loss function	Optimizer	Learning rate	Device
M. Astaraki et al.	nnUNetV2	× (5)	2	80×192×160	Dice and BCE	SGD	0.01	Nvidia DGX-1 cluster
Y. Ye et al.	nnUNet	× (5)	2	64×192×192	Dice and CE	SGD	0.01	NVIDIA Geforce RTX 2080Ti
Z. Xing et al.	nnUNet	× (3)	2	64×256×256	Dice and CE	SGD	0.01	NVIDIA A100 GPU
K. Yang et al.	nnUNet	None	2	28×256×256	Dice and Focal	SGD	0.01	TITAN RTX 24G
C. Ulrich et al.	nnUNetV2	× (5)	4	32×320×256	Soft-dice and CE	SGD	0.01	Nvidia V100, Nvidia A100, Titan RTX
N. Ndipenoch et al.	nnUNet_SE	× (10)	2	64×192×192	Dice and CE	SGD	0.01	NVIDIA RTX A6000 48GB
Y. Su et al.	nnUNetV2	None	2	48x256x256	Dice and CE	SGD	0.01	NVIDIA A100
J. Huang et al.	nnUNetV2	× (4)	2	40×256×160	Dice and CE	SGD	0.01	NVIDIA Geforce RTX 3090
Y. Zhang et al.	nnUNet	None	2	64×192×160	Dice and CE	SGD	0.01	NVIDIA Geforce RTX 3090
C. Lee et al.	yolo-v7 + UNet	× (5)	4	32×96×96 32×128×128	Dice and CE	AdamW	1e-4	NVIDIA A5000
K. Huang et al.	nnUNetV2	None	2	24×224×224	Soft-dice and CE	AdamW	1.e-3	NVIDIA Geforce RTX 2080Ti

Table 7. Rankings of methods in DSC/NSD scores for OAR segmentation.

Team	Y. Zhong *et al.*	Y. Ye *et al.*	Y. Su *et al.*	K. Yang *et al.*	C. Lee *et al.*	M. Astaraki *et al.*	Z. Xing *et al.*	Y. Zhang *et al.*	J. Huang *et al.*	K. Huang *et al.*
Brain	4/3	2/2	1/1	3/4	7/7	5/5	8/6	9/9	6/8	10/10
BrainStem	1/1	3/3	5/4	7/6	10/9	8/8	2/2	4/5	6/7	9/10
Chiasm	4/3	2/1	8/7	7/6	3/8	6/5	5/4	1/2	10/10	9/9
Cochlea_L	1/1	3/3	2/2	6/5	4/4	5/6	9/9	8/8	7/7	10/10
Cochlea_R	1/1	3/3	2/2	6/4	5/6	4/5	9/8	8/9	7/7	10/10
Esophagus	2/2	4/4	1/1	3/3	5/5	6/6	8/7	9/9	7/8	10/10
ETbone_L	1/1	3/3	2/2	7/5	4/4	6/6	8/8	5/7	9/9	10/10
ETbone_R	1/1	3/2	2/3	6/4	5/6	4/5	9/9	8/8	7/7	10/10
Eye_L	1/1	3/3	2/2	6/5	5/6	4/4	9/8	8/9	7/7	10/10
Eye_R	1/1	3/3	2/4	4/2	7/7	5/5	8/8	6/6	9/9	10/10
Hippocampus_L	1/1	2/2	3/3	5/4	6/6	4/5	8/8	7/7	9/9	10/10
Hippocampus_R	1/1	3/2	5/5	4/3	2/4	7/7	8/8	6/6	10/10	9/9
IAC_L	1/1	2/2	5/5	7/7	6/6	4/4	8/8	3/3	10/10	9/9
IAC_R	1/1	3/3	2/2	5/5	4/4	6/6	8/8	7/7	10/10	9/9
Larynx	1/1	4/3	2/2	5/5	3/4	6/6	8/8	7/7	10/9	9/10
Larynx_Glottic	1/2	2/1	3/3	5/5	4/4	6/6	8/9	7/8	10/7	9/10
Larynx_Supraglot	1/1	2/2	4/4	5/5	3/3	6/6	8/9	7/7	9/8	10/10
Lens_L	1/1	2/3	4/2	5/4	3/6	6/5	7/7	8/8	10/10	9/9
Lens_R	1/1	2/2	4/3	5/5	3/4	6/6	8/8	7/7	10/10	9/9
Mandible_L	1/1	2/2	4/4	3/3	5/5	6/6	8/8	7/9	9/7	10/10
Mandible_R	1/1	2/2	7/4	4/5	3/3	5/6	8/8	6/7	10/9	9/10
Mastoid_L	2/3	1/2	3/1	4/4	5/5	6/6	7/8	8/9	9/7	10/10
Mastoid_R	1/1	2/3	6/2	4/4	3/5	5/6	8/9	7/7	10/8	9/10
MiddleEar_L	1/1	2/2	3/3	4/4	5/5	6/6	8/7	7/8	10/9	9/10
MiddleEar_R	2/2	5/4	1/1	7/7	3/3	6/6	8/8	4/5	10/9	9/10
OpticNerve_L	1/2	3/3	7/5	4/4	5/7	8/9	2/1	9/6	10/10	6/8
OpticNerve_R	1/1	3/3	2/2	5/4	4/5	7/7	8/8	10/10	6/6	9/9
OralCavity	1/1	4/5	3/3	7/7	6/4	5/6	8/8	10/10	2/2	9/9
Parotid_L	4/4	2/1	3/2	7/7	6/6	5/5	8/8	10/10	1/3	9/9
Parotid_R	6/3	2/4	1/1	5/5	7/7	4/6	8/8	10/10	3/2	9/9
PharynxConst	3/2	2/3	1/1	5/4	8/8	4/6	7/7	10/10	6/5	9/9
Pituitary	3/4	2/2	1/1	4/3	7/7	6/6	8/9	10/10	5/5	9/8
SpinalCord	5/4	2/2	3/3	1/1	7/6	4/5	8/8	10/10	6/7	9/9
Submandibular_L	1/1	3/3	2/2	4/4	5/5	6/6	8/7	10/10	7/8	9/9
Submandibular_R	2/2	1/1	4/4	3/3	6/6	5/5	8/8	9/9	7/7	10/10
TemporalLobe_L	2/2	4/4	1/1	3/3	6/6	5/5	7/7	8/8	10/10	9/9
TemporalLobe_R	1/1	3/3	2/2	4/4	6/6	5/5	7/7	8/8	10/10	9/9
Thyroid	1/2	2/1	3/3	5/5	9/9	4/4	6/7	8/8	7/6	10/10
Trachea	1/1	6/6	5/5	4/4	2/2	3/3	9/8	10/10	7/7	8/9
TympanicCavity_L	1/1	3/3	2/2	5/5	6/6	4/4	7/7	8/8	10/10	9/9
TMjoint_L	2/1	3/2	5/4	6/3	4/5	8/7	7/8	9/9	10/10	1/6
TMjoint_R	1/1	4/2	2/3	5/5	7/7	6/6	3/4	8/9	10/10	9/8
TympanicCavity_R	1/1	3/2	4/4	6/6	7/7	5/5	2/3	9/9	10/10	8/8
VestibulSemi_L	1/1	2/3	3/2	4/4	6/5	5/6	7/7	10/10	8/8	9/9
VestibulSemi_R	3/4	1/1	4/3	2/5	9/9	5/7	6/2	7/6	10/10	8/8
Average	1.7/1.6	2.7/2.6	3.1/2.8	4.8/4.4	5.2/5.6	5.4/5.7	7.3/7.2	7.7/7.9	8.1/7.9	9/9.3
Overall	1	2	3	4	5	6	7	8	9	10

(4th place, K. Yang *et al.*) Yang *et al.* used nnUNet (Isensee et al., 2021) and region-based training mode for accurate and efficient segmentation. In the training stage, the images were augmented by elastic deformation without flipping. To address the issue of missing labels in some training cases, such as MiddleEar ETbone Overlap, a masked loss function was used, where the channels of label missing were ignored to correct model training. For overlapping ragions, a region-based training mode was used to segment areas that are merged by more than one class. During inference, a sliding window strategy and a connect component-based post-processing were adopted to obtain final segmentation results.

(5th place, C. Lee *et al.*) Lee *et al.* proposed a two-stage method consisting of organ localization followed by segmentation. In the localization stage, a 2D-based object detection network powered by the YOLO-v7 model (Wang et al., 2022) was used for identifying a bounding box around the OARs. For segmentation, different window widths and levels were used for multi-channel input generation. A segmentation network with DynUNet architecture was trained using these multi-channel inputs, employing single organ training and symmetrical OARs Flipped-Unification. For OARs Flipped-Unification, the training data was from one of the symmetrical OARs while utilizing a flipped version of the same to represent its counterpart because of the symmetry in the head and neck area. During inference, ROIs were first extracted, and then all predictions from five segmentation models were averaged as final results.

(6th place, M. Astaraki *et al.*) Astaraki *et al.* utilized intensity distribution harmonization and efficient cropping strategies. To better distinguish the overlapping OARs from each other, the HU values of the ceCT and ncCT volumes were clamped into the range of [-400, 2000] and [-300, 800] for pre-processing, respectively. The pre-processed paired full-resolution CT images were used to train a segmentation network based on the nnUNetV1 (Isensee et al., 2021) framework with 2000 epochs using five-fold cross-validation. During inference, volumes were cropped based on the TotalSegmentor (Wasserthal et al., 2023) model and a connected component analysis before being segmented by the trained segmentation network.

(7th place, Z. Xing *et al.*) Xing *et al.* focused on using cropping and test-time augmentation strategies to perform OAR segmentation. To save training time, the images were cropped

Table 8. Summary of the average DSC (%) score of OAR segmentation by the ten teams.

Team	Y. Zhong *et al.*	Y. Ye *et al.*	Y. Su *et al.*	K. Yang *et al.*	C. Lee *et al.*	M. Astaraki *et al.*	Z. Xing *et al.*	Y. Zhang *et al.*	J. Huang *et al.*	K. Huang *et al.*	Baseline	Atlas
Brain	98.62±0.26	98.63±0.30	**98.65±0.32**	98.62±0.31	98.58±0.25	98.61±0.35	98.54±0.22	98.44±0.18	98.60±0.27	98.42±0.22	98.47±0.27	98.23±0.31
BrainStem	**92.45±2.76**	92.28±2.67	91.97±2.82	91.88±2.62	91.57±4.45	91.75±2.74	92.32±2.73	92.06±2.77	91.92±2.75	91.72±2.85	91.84±3.01	88.24±4.36
Chiasm	70.55±14.41	71.08±13.67	69.49±13.34	69.67±13.72	70.67±15.60	70.03±14.41	70.53±14.68	**71.76±13.05**	64.57±16.07	69.13±14.21	70.12±12.31	52.32±23.93
Cochlea_L	**94.91±1.36**	94.77±1.27	94.83±1.47	94.54±2.13	94.76±1.27	94.55±1.41	87.10±19.12	89.02±9.36	94.26±1.59	83.54±26.02	93.27±1.66	-
Cochlea_R	**95.32±1.28**	94.93±1.53	94.99±1.53	94.63±2.52	94.71±1.42	94.84±1.38	87.65±18.36	88.93±10.50	94.52±1.58	80.58±30.46	94.38±1.73	-
Esophagus	77.32±8.09	76.60±7.95	**77.63±7.81**	76.69±8.15	76.05±8.59	75.71±8.10	73.53±16.30	73.51±9.88	73.83±11.55	67.91±23.08	73.34±9.36	63.87±21.49
ETbone_L	**79.18±8.19**	78.19±8.20	78.98±8.37	76.82±12.69	77.97±7.91	77.38±8.07	76.07±16.24	77.47±6.59	74.55±12.71	68.27±26.12	77.07±6.88	-
ETbone_R	**94.04±2.09**	93.91±2.01	93.99±2.19	93.53±4.74	93.69±2.09	93.74±2.23	88.11±21.67	90.00±11.70	92.89±4.76	84.23±26.48	93.14±1.87	-
Eye_L	**93.30±2.08**	93.17±1.90	93.24±2.11	91.60±11.29	92.72±2.32	92.82±2.07	87.92±20.51	89.23±10.42	90.71±12.00	81.23±29.61	92.52±2.02	71.38±12.35
Eye_R	**72.34±7.78**	78.02±8.12	78.18±8.21	77.72±8.99	74.78±12.34	77.41±8.28	73.43±20.77	75.10±13.17	70.34±15.76	67.93±22.86	71.08±10.38	70.27±14.45
Hippocampus_L	**75.83±8.52**	75.54±7.88	75.31±7.30	74.88±12.74	73.31±10.89	75.02±7.95	71.74±18.55	71.95±14.31	67.18±18.18	64.19±24.61	75.29±6.91	-
Hippocampus_R	**79.99±7.71**	78.99±8.05	78.43±8.86	78.60±9.48	79.44±7.34	77.48±9.19	75.73±18.94	77.75±12.85	65.90±20.71	69.79±25.48	78.49±8.13	-
IAC_L	**81.94±7.23**	81.75±7.50	80.50±8.92	79.26±13.27	80.24±7.85	80.57±7.43	78.18±16.92	81.01±7.93	65.89±24.97	71.09±25.13	78.59±8.60	-
IAC_R	**88.42±5.18**	87.38±5.32	87.45±4.72	86.78±5.93	87.16±5.12	85.25±7.49	82.46±17.02	82.68±17.50	69.85±7.35	76.40±24.44	84.85±5.09	-
Larynx	**89.25±5.02**	87.37±5.28	87.98±5.08	86.62±7.24	87.47±6.55	85.98±7.74	83.10±16.66	84.07±15.80	68.19±8.82	74.56±30.97	87.26±4.35	82.68±6.81
Larynx_Glottic	**84.94±8.45**	84.54±8.13	83.82±8.01	82.80±9.32	83.70±7.62	82.36±8.66	74.23±17.56	79.66±17.29	72.73±18.33	73.46±23.09	83.50±8.22	-
Larynx_Supraglot	**85.34±7.34**	84.72±7.27	84.17±7.33	82.28±13.03	84.60±6.18	81.25±8.88	75.82±17.41	79.70±19.63	70.28±23.08	67.76±31.81	82.58±8.15	-
Lens_L	**81.95±7.28**	81.39±7.41	80.77±8.17	80.64±7.51	81.00±7.30	80.27±8.49	76.96±16.47	74.80±20.48	52.98±11.99	71.39±23.57	78.62±9.20	46.42±23.56
Lens_R	**84.18±7.22**	83.58±7.15	82.83±7.63	82.33±7.76	83.57±7.16	81.57±8.06	78.96±16.66	79.39±16.33	55.07±13.34	70.78±28.94	82.47±7.64	44.76±24.39
Mandible_L	**83.79±8.80**	83.42±8.51	82.68±8.47	82.75±9.03	82.38±7.77	81.67±11.81	77.55±17.46	77.98±20.02	73.33±14.70	71.63±24.32	82.39±8.03	69.46±28.64
Mandible_R	**83.49±9.06**	83.19±8.55	79.35±10.92	82.25±9.04	82.65±7.60	81.07±12.63	77.98±15.78	79.48±16.41	66.84±18.83	67.28±29.47	82.49±8.14	67.19±32.47
Mastoid_L	84.10±8.21	**84.50±7.72**	84.04±7.42	83.49±8.23	82.56±8.01	81.81±12.57	78.98±16.85	78.25±20.06	72.57±18.13	71.46±24.56	82.92±8.47	-
Mastoid_R	**83.35±9.43**	82.85±9.47	80.43±11.63	81.50±13.63	81.97±8.31	80.98±12.45	76.76±16.70	79.54±16.75	68.09±22.35	68.15±29.63	82.52±9.48	-
MiddleEar_L	**82.14±5.72**	82.06±5.49	81.46±5.72	80.92±6.77	80.46±7.23	79.77±7.90	77.36±15.87	77.64±17.39	66.98±16.86	72.24±23.18	70.65±8.31	-
MiddleEar_R	78.99±10.86	76.35±9.74	**79.12±9.46**	74.61±12.70	78.06±9.95	74.78±10.87	74.40±16.81	76.54±14.28	61.84±18.19	67.83±25.41	74.82±9.83	-
OpticNerve_L	**77.70±13.86**	77.27±13.6	75.78±17.65	76.58±16.14	76.58±16.31	75.52±14.98	77.65±14.04	75.35±17.87	64.44±23.53	75.78±13.26	75.81±16.44	56.29±16.41
OpticNerve_R	**95.04±1.56**	94.96±1.61	94.98±1.64	94.94±1.60	94.95±1.59	94.79±1.58	94.63±1.61	94.15±1.70	94.85±1.57	94.28±1.79	93.89±1.78	57.08±16.43
OralCavity	**95.02±1.88**	94.92±1.84	94.99±1.87	92.60±3.74	94.35±2.04	94.67±1.89	90.47±15.26	72.19±19.30	95.01±1.90	85.46±22.03	93.38±2.30	87.21±10.01
Parotid_L	94.27±3.30	94.39±3.23	94.36±3.33	91.73±6.08	93.76±3.32	94.16±3.22	91.10±12.85	73.17±18.57	**94.41±3.15**	84.57±22.07	93.41±3.41	71.52±17.55
Parotid_R	88.99±9.85	89.63±6.48	**89.74±6.26**	89.00±7.61	87.94±9.33	89.13±7.73	86.70±13.62	67.10±20.60	89.30±7.19	83.82±18.29	88.31±7.54	72.39±16.63
PharynxConst	87.27±11.50	87.59±9.24	**87.82±9.18**	86.46±12.04	85.11±13.41	87.23±9.48	85.65±13.63	66.49±21.86	85.94±15.49	81.91±18.79	86.99±9.12	-
Pituitary	90.26±4.41	90.28±4.51	**90.36±4.66**	90.25±4.48	89.09±5.32	89.89±4.52	83.04±16.22	70.21±24.19	90.23±4.49	81.62±24.54	88.36±5.28	57.81±28.56
SpinalCord	88.26±7.46	88.68±6.50	88.63±6.33	**88.99±5.73**	86.41±11.02	88.46±6.46	82.39±16.29	71.96±20.22	87.40±7.22	78.44±24.69	86.32±7.56	78.42±18.32
Submandibular_L	**92.90±2.40**	92.79±2.58	92.84±2.53	92.55±2.70	92.36±2.50	92.33±2.67	84.56±19.04	79.26±23.89	86.69±4.59	81.31±26.09	90.62±3.92	63.42±15.87
Submandibular_R	92.47±3.52	**92.49±3.49**	92.30±3.40	92.35±3.60	92.00±3.63	92.05±3.64	84.46±19.79	82.66±16.45	87.95±4.30	77.68±30.37	91.62±3.69	61.89±13.14
TemporalLobe_L	89.23±7.20	88.84±7.08	**89.32±6.80**	88.88±7.36	88.45±7.40	88.54±7.06	81.76±21.86	79.91±23.90	73.35±19.82	79.19±25.61	88.37±6.81	83.42±9.82
TemporalLobe_R	**90.37±4.72**	89.72±5.17	89.95±4.69	89.43±5.55	88.78±6.09	89.21±5.89	83.88±15.17	83.32±15.93	67.09±22.58	75.22±31.10	89.22±4.53	84.57±6.49
Thyroid	**89.69±4.29**	89.54±3.85	89.44±3.98	89.27±4.05	88.80±4.14	89.28±4.12	89.17±4.21	88.90±3.96	88.95±3.66	88.32±4.00	88.52±3.31	73.38±14.38
TMjoint_L	82.34±8.16	82.25±8.01	82.21±8.00	81.86±8.01	82.21±8.00	81.31±8.51	81.41±7.98	81.26±7.56	34.91±25.87	82.42±7.97	**84.33±10.96**	74.59±12.67
TMjoint_R	**89.74±3.97**	89.28±4.18	89.35±3.91	89.14±4.19	88.75±3.89	88.90±3.98	89.32±3.95	88.35±3.95	63.13±23.27	88.08±3.45	89.59±4.41	75.48±11.69
Trachea	**85.01±2.66**	83.98±2.15	84.07±2.26	84.08±2.20	84.81±2.91	84.10±2.09	82.57±3.50	82.28±3.11	82.89±3.47	82.70±2.94	79.65±4.65	73.29±16.72
TympanicCavity_L	**89.66±2.21**	89.37±2.18	89.55±2.32	89.23±2.38	89.21±2.07	89.25±2.43	89.03±2.37	88.80±2.17	81.43±4.94	88.76±2.20	88.43±2.03	-
TympanicCavity_R	**85.17±4.83**	84.53±4.66	84.36±4.89	84.04±4.92	83.03±4.57	84.08±4.85	84.71±4.89	81.05±6.05	71.91±6.89	82.13±5.85	81.77±3.83	-
VestibulSemi_L	**91.27±3.34**	90.90±3.11	90.90±3.15	90.59±3.12	90.25±3.44	90.30±3.07	90.10±3.16	88.66±3.83	89.91±3.36	88.84±3.19	79.46±9.08	-
VestibulSemi_R	85.18±9.46	**85.56±8.55**	85.11±8.96	85.48±7.87	84.46±9.47	84.96±8.85	84.94±9.22	84.73±8.50	77.13±7.37	84.69±8.46	84.27±6.97	-
Average	**86.70±9.30**	86.36±9.15	86.14±9.58	85.62±10.48	85.68±9.87	85.44±10.17	82.51±16.48	80.57±16.52	76.68±19.62	78.14±23.65	84.65±9.95	-

based on regions with intensity values in the range of [-175, 250]. Extensive data augmentation techniques, including spatial (with random mirror) and intensity transforms, were used to improve the robustness segmentation model. An ensemble of five UNet-based segmentation models, each with varying batch sizes, parameter scales, and normalization methods, was used to generate a robust prediction. During inference, test-time augmentation based on mirror operation and sliding window with overlap was used to improve the robustness of the prediction.

(8th place, Y. Zhang *et al.*) Zhang *et al.* employed nnUNet (Isensee et al., 2021) framework, clipping the HU values of the CT images to the [0.5, 99.5] percentiles of these intensity values. Data augmentation methods, including spatial (with random mirror), intensity, and label-based transformation, were used to enhance data diversity and richness. Paired CT images were randomly cropped into patches of size [28, 224, 224] and used to train a 3D full-resolution UNet based on nnUNet (Isensee et al., 2021). During inference, the patch size was equal to the patch size during training, and the sliding window with a step size was half of the window size.

(9th place, J. Huang *et al.*) J. Huang *et al.* used a two-step method for OAR segmentation, consisting of coarse and fine segmentation. The intensity values of paired CT images were clipped to [-300, 1500] and then normalized to [-1, 1] by min-max normalization. Symmetrical organs on the left and right sides were treated as separate tags for model training, incorporating data augmentation methods like random flipping and rotation. In the coarse segmentation stage, pre-processed images were used to train a 3D UNet to get the position and size of the target areas, after which the corresponding ROIs were cropped based on the coarse segmentation results. In the fine stage, a 3D UNet was trained based on paired CT images and corresponding ROIs to refine the coarse segmentation results. During inference, segmentation results are generated through these two progressive stages and then divided into left and right parts based on spatial position.

(10th place, K. Huang *et al.*) K. Huang *et al.* proposed a method based on the nnUNetV2 framework (Isensee et al., 2021). The paired CT volumes were resampled, cropped, and normalized following Isensee et al. (2021). Data augmentation strategies, including spatial transform, intensity transform, and simulated low-resolution transform, were used to improve the diversity of data. Five-fold cross-validation was used to train segmentation networks. During inference, various augmentations like different region cropping and adjustments in scaling were applied to enhance the stability of results, and the average of predictions was taken as the final results.

4.2. Task02: GTV segmentation

Almost all teams submitted deep learning-based methods based on nnUNet (Isensee et al., 2021) structure. Nine of the submitted teams used end-to-end methods, in which two teams

Table 9. Summary of the average NSD (%) score of OAR segmentation by the ten teams.

Team	Y. Zhong *et al.*	Y. Ye *et al.*	Y. Su *et al.*	K. Yang *et al.*	C. Lee *et al.*	M. Astaraki *et al.*	Z. Xing *et al.*	Y. Zhang *et al.*	J. Huang *et al.*	K. Huang *et al.*	Baseline	Atlas
Brain	89.68±4.75	89.77±5.25	**89.79±5.28**	89.64±5.29	88.92±4.84	89.39±5.81	88.92±4.80	87.57±4.56	88.89±5.11	87.08±5.06	88.02±4.93	87.86±4.98
BrainStem	**82.00±10.57**	81.54±10.29	80.57±10.65	80.27±9.95	79.28±12.01	79.82±10.12	81.55±10.60	80.38±10.28	80.17±10.40	79.16±11.02	79.13±11.03	78.36±14.38
Chiasm	77.07±15.58	**77.50±14.65**	75.98±14.4	76.18±14.36	75.84±17.35	76.38±15.59	76.77±16.05	77.24±13.85	72.55±15.29	75.18±14.71	75.79±12.76	49.89±26.79
Cochlea_L	**79.99±7.71**	79.43±7.04	79.76±7.43	78.90±7.67	79.26±7.24	78.31±7.73	69.98±16.97	70.37±10.53	77.11±7.19	66.35±22.97	73.14±8.17	-
Cochlea_R	**80.61±7.80**	78.60±8.85	78.91±8.78	78.21±9.29	77.57±8.87	77.99±8.46	69.33±16.95	68.59±12.74	76.28±8.81	63.14±25.63	76.50±8.57	-
Esophagus	68.31±12.25	67.57±12.03	**68.92±11.76**	68.06±11.69	66.24±12.48	66.14±11.74	64.93±16.76	62.79±13.27	64.89±13.97	59.43±21.29	60.88±12.99	56.48±25.29
ETbone_L	**71.62±13.33**	70.06±12.92	71.31±13.82	68.81±15.40	68.88±13.05	68.69±12.98	68.02±18.13	68.48±11.05	65.67±14.79	61.14±24.81	68.08±11.17	-
ETbone_R	**91.16±8.21**	90.87±7.64	90.81±8.41	90.73±9.33	89.84±8.07	90.01±8.49	84.88±21.83	85.43±14.38	88.78±9.28	79.67±26.71	88.49±7.27	-
Eye_L	**88.71±8.11**	88.40±7.77	88.66±9.12	87.15±11.96	86.89±8.99	87.43±8.16	83.00±20.97	82.79±13.17	83.94±13.79	74.40±28.73	86.37±8.03	73.49±14.18
Eye_R	**90.12±8.76**	89.88±8.71	89.64±8.66	89.90±9.64	85.80±12.36	89.07±8.74	84.35±22.90	86.61±13.24	82.47±17.22	76.96±26.45	87.61±10.22	72.58±13.27
Hippocampus_L	**86.58±10.94**	86.42±10.33	86.38±8.96	85.63±15.20	83.46±12.23	85.53±10.29	81.77±21.68	83.15±13.58	78.56±20.00	71.97±28.12	86.11±8.85	-
Hippocampus_R	**87.96±9.00**	86.73±9.32	86.20±10.42	86.41±10.69	86.30±8.23	84.95±10.63	83.16±20.83	85.39±14.38	76.88±17.87	77.39±27.81	85.54±9.59	-
IAC_L	**89.19±7.66**	89.16±7.88	87.84±9.28	86.74±14.40	86.88±8.54	87.86±8.16	84.90±18.33	88.01±8.65	75.94±23.33	78.10±26.91	84.83±9.74	-
IAC_R	**91.71±6.54**	90.28±6.99	90.43±5.98	89.68±7.50	89.94±6.61	87.84±9.26	85.11±17.75	85.59±17.44	75.54±5.58	78.33±25.03	87.46±6.33	-
Larynx	**98.10±3.54**	97.03±3.92	97.53±2.84	96.54±5.56	96.63±5.00	96.09±6.84	92.2±17.52	93.75±14.35	86.65±4.65	85.11±30.30	97.38±2.48	91.83±12.37
Larynx_Glottic	95.38±6.24	**95.40±6.06**	95.07±6.26	94.19±8.04	94.43±6.41	93.72±7.39	86.18±18.62	90.37±18.57	90.94±13.52	84.29±23.52	94.63±6.46	-
Larynx_Supraglot	**96.17±5.44**	95.88±5.85	95.67±5.53	93.90±13.54	95.86±4.75	93.48±7.57	87.69±18.65	90.73±19.27	87.72±20.38	78.02±35.78	94.59±6.69	-
Lens_L	**92.05±6.83**	91.69±7.02	91.71±6.91	91.17±6.82	90.42±6.95	90.57±8.62	86.53±17.63	85.38±20.46	67.51±10.57	80.40±26.61	88.52±7.95	63.47±13.24
Lens_R	**92.27±7.17**	91.66±7.46	91.31±7.94	91.10±8.19	91.18±8.32	90.24±8.91	86.63±18.03	87.47±15.81	67.26±12.83	78.62±29.98	90.19±7.99	61.88±16.29
Mandible_L	**94.99±7.19**	94.97±6.98	94.72±7.16	94.83±7.22	93.93±7.46	93.66±9.64	88.82±17.91	88.80±21.98	89.49±10.39	83.58±25.60	94.55±6.85	82.39±23.47
Mandible_R	**94.94±6.94**	94.85±6.74	94.58±6.38	94.23±7.32	94.65±6.60	93.68±9.61	90.06±16.37	91.26±15.45	84.97±14.03	78.89±32.05	94.84±6.58	76.47±18.49
Mastoid_L	95.43±6.70	95.69±6.14	**95.84±5.79**	95.11±7.39	93.97±6.63	93.90±10.61	90.75±18.45	88.93±21.90	92.12±10.52	83.16±27.24	94.35±6.83	-
Mastoid_R	**95.10±7.91**	94.70±8.12	94.88±7.00	93.99±11.00	93.84±7.46	93.19±10.38	89.04±17.92	91.50±15.78	89.27±14.63	80.41±30.46	94.53±8.08	-
MiddleEar_L	**95.01±4.37**	94.93±4.47	94.87±4.52	94.19±5.78	93.93±6.02	93.39±6.28	90.40±17.2	90.33±18.99	87.38±13.56	84.86±25.40	86.70±6.76	-
MiddleEar_R	92.60±9.06	91.16±7.60	**93.41±7.43**	89.55±11.08	91.37±9.40	89.85±9.24	88.44±17.09	90.55±13.75	84.19±14.82	81.35±27.67	89.93±7.94	-
OpticNerve_L	86.61±11.99	85.90±11.99	84.73±16.59	84.78±16.08	84.56±15.9	84.34±13.38	**86.69±12.24**	84.69±14.77	79.65±21.20	84.43±11.96	84.59±15.79	74.49±19.21
OpticNerve_R	**75.79±10.36**	75.32±10.46	75.48±10.55	75.09±10.34	74.85±10.81	74.26±9.98	72.94±10.54	69.66±10.86	74.84±10.29	70.94±11.69	69.01±10.45	75.07±16.54
OralCavity	**99.79±0.56**	99.72±0.47	99.74±0.47	96.88±3.37	99.74±0.60	99.72±0.52	95.43±15.02	76.82±13.87	99.75±0.39	91.04±21.56	98.77±1.48	93.28±2.25
Parotid_L	91.87±9.84	**92.17±9.33**	92.07±9.67	88.57±10.41	89.93±11.59	91.72±9.53	87.82±15.78	65.97±17.64	92.02±9.50	80.89±23.70	89.58±9.51	66.31±16.57
Parotid_R	82.91±14.67	82.79±12.91	**83.53±12.02**	82.42±13.51	79.86±13.89	82.00±13.95	79.70±17.06	58.44±17.61	83.40±12.46	75.35±19.66	78.75±15.08	67.39±14.32
PharynxConst	79.34±18.10	79.08±16.74	**79.66±16.66**	78.32±18.06	74.14±18.84	77.87±17.67	76.52±18.75	55.21±20.20	78.30±19.54	72.43±21.29	76.45±16.94	-
Pituitary	74.12±15.93	74.14±16.30	**74.51±16.71**	74.12±16.19	70.20±16.50	72.50±16.29	65.99±19.65	56.07±19.93	73.74±16.30	66.91±24.45	68.53±16.03	56.09±22.56
SpinalCord	70.25±19.22	71.21±18.31	70.84±18.64	**71.53±17.16**	68.13±18.61	69.95±18.42	62.98±20.18	53.59±18.92	66.39±20.17	60.69±24.24	64.62±18.47	58.35±19.35
Submandibular_L	**90.06±6.36**	89.68±7.09	89.86±6.94	89.25±7.04	88.87±6.23	88.65±7.18	79.47±18.82	75.79±22.20	79.26±8.48	76.41±26.28	84.18±8.71	59.29±25.18
Submandibular_R	88.93±9.12	**89.13±8.97**	88.68±8.89	88.69±9.14	87.71±9.61	87.96±9.33	78.94±20.22	77.12±18.07	80.10±8.82	73.53±28.56	86.79±9.11	58.27±23.49
TemporalLobe_L	87.63±12.37	87.18±12.19	**87.95±11.72**	87.27±12.41	86.22±12.68	86.69±12.00	79.67±22.85	78.71±23.43	74.70±17.31	77.11±26.57	86.27±11.54	82.33±13.24
TemporalLobe_R	**89.89±8.22**	88.93±9.00	89.26±8.24	88.38±9.64	86.60±11.24	87.81±10.27	81.53±17.32	81.52±17.72	72.94±17.02	74.25±29.92	87.79±8.14	83.04±14.26
Thyroid	86.53±11.01	**86.72±10.41**	86.39±10.92	86.09±10.95	84.34±10.97	86.13±11.00	85.36±11.15	84.96±10.89	86.06±10.09	83.76±11.24	84.62±10.06	75.31±16.32
TMjoint_L	**80.14±12.74**	80.05±12.56	79.65±12.71	79.67±12.52	79.51±12.85	78.90±12.90	78.45±13.18	77.54±11.69	35.72±23.12	78.98±12.27	77.97±14.18	72.57±17.32
TMjoint_R	**88.36±7.82**	87.89±7.74	87.88±7.31	87.33±7.75	86.88±7.63	87.14±7.45	87.54±7.66	85.69±7.51	60.06±21.32	86.12±7.07	86.81±6.75	71.81±18.14
Trachea	**78.04±5.72**	75.18±5.83	75.29±5.99	75.30±5.88	77.00±6.11	75.43±6.20	72.45±6.98	71.51±6.74	73.97±8.95	71.76±7.44	68.10±7.98	63.16±12.94
TympanicCavity_L	**75.71±9.08**	74.86±8.49	75.12±9.40	73.72±9.52	72.54±8.73	74.25±8.92	72.38±9.43	71.31±8.80	60.59±8.78	71.30±9.25	69.36±7.68	-
TympanicCavity_R	**86.41±9.70**	85.86±9.24	85.70±9.89	84.77±9.92	81.92±9.81	85.15±9.46	85.76±9.97	79.19±12.11	72.22±10.29	81.37±12.48	80.91±9.54	-
VestibulSemi_L	**89.19±9.27**	88.36±8.94	88.58±9.16	87.78±9.05	87.08±9.42	86.94±9.34	86.50±9.37	83.03±10.55	86.02±9.27	83.27±10.44	67.15±12.31	-
VestibulSemi_R	75.36±17.45	**75.87±15.43**	75.51±16.01	74.97±14.70	72.40±18.46	74.68±16.49	75.72±16.48	74.74±15.35	58.30±12.54	74.66±15.13	71.12±14.15	-
Average	**86.53±12.85**	86.09±12.64	86.12±12.79	85.33±13.42	84.62±13.62	84.96±13.21	81.67±18.56	79.18±18.69	77.85±18.04	76.94±24.31	82.88±14.01	-

used pre-trained models, and the other two used two-stage approaches. Only one team used Dice and Focal loss, the others used similar loss functions that are Dice and CE loss. In this task, we employed the default nnUNet (Isensee et al., 2021) without the test-time-augmentation strategy as the baseline, as this task does not have symmetrical and complex structure organs. So, the most noticeable difference between the data augmentation strategies of Task02 and Task01 baselines was the presence or absence of the mirror and flipping transformations.

(1^{st} place, M. Astaraki *et al.*) Astaraki *et al.* used intensity distribution harmonization and efficient cropping. The HU values of the ceCT and ncCT volumes were clamped into the range of [-1000, 1000] and [-600, 600], respectively, to better distinguish the cancer regions from nearby healthy tissues. To discard the background and irrelevant anatomical structures, the paired CT volumes were cropped based on TotalSegmentor (Wasserthal et al., 2023) model and a connected component analysis. The cropped paired CT images were used to train a segmentation network based on the nnUNetV1 (Isensee et al., 2021) framework with 600 epochs using five-fold cross-validation. During inference, the test volumes were harmonized and cropped as training data and then sent to the segmentation network for segmentation labels over the cropped images.

(2^{nd} place, Y. Ye *et al.*) Ye *et al.* employed the UniSeg (Ye et al., 2023) model and ensemble strategy. In the training stage, each image was divided into multiple 3D patches of identical size using a sliding window approach, and then these patches were pre-processed following nnUNet (Isensee et al., 2021). Then, UniSeg was trained using paired patches with 1000 epochs. During inference, the entire image was segmented into overlapping patches, and then each patch was sent to the fine-tuned UniSeg to predict its corresponding segmentation map, and these individual patch-based predictions were aggregated as the final prediction.

(3^{rd} place, Z. Xing *et al.*) Xing *et al.* used crop and test-time augmentation strategies. Regions with HU values of [-175, 250] were cropped for training. To improve the robustness of the segmentation model, spatial- and intensity-based transforms are used. An ensemble of five segmentation models based on UNet structure with different batch sizes, parameter scales, and normalization methods was used to generate a robust prediction. During inference, test-time augmentation was used to improve the robustness of the prediction.

(4^{th} place, K. Yang *et al.*) Yang *et al.* used nnUNet (Isensee et al., 2021) for GTV segmentation, employing Dice loss and Focal loss (Lin et al., 2017) to address the challenges of segmenting difficult GTVs. Due to the variance in GTVs among patients, the sliding window strategy was not used. During inference, test-time augmentation based on flipping was used to improve the segmentation performance.

(5^{th} place, C. Ulrich *et al.*) Ulrich *et al.* employed MultiTalent (Ulrich et al., 2023) model that is trained with multiple partially labeled datasets. The model was initially pre-trained following the target spacing, normalization scheme, and network topology suggested by nnUNet experiment planning for the SegRap2023. After pre-training, the MultiTalent model was fine-tuned with paired CT images by only updating the segmentation heads for 10 epochs, and the whole network was updated for a 50 epoch warm-up period. Finally, a Residual Encoder UNet was initialized using the MultiTalent model and trained for 2000 epochs to generate the final segmentation results.

(6^{th} place, N. Ndipenoch *et al.*) Ndipenoch *et al.* proposed a nnUNet with squeeze and excitation block (nnUNet_SE) model (Isensee et al., 2021), where residual blocks were introduced to mitigate the problem of vanishing gradients, and the squeeze-and-excitation block was introduced to capture global features. The nnUNet_SE model was trained with paired ncCT and ceCT scans, and each of the GTVs was trained separately as binary segmentation tasks to improve the performance.

(7^{th} place, Y. Su *et al.*) Su *et al.* used a vanilla nnUNet (Isensee et al., 2021) to perform GTV segmentation. Almost all settings were the same as those automatically generated by Isensee et al. (2021), except for the patch size. A large patch size (48 × 256 × 256) was used to improve the model's performance. During inference, test-time augmentation strategy was applied for robust segmentation results.

(8^{th} place, J. Huang *et al.*) J. Huang *et al.* used two progressive steps for GTV segmentation: coarse segmentation and fine segmentation. The HU values of paired CT images were clipped to [-300, 1500] and then normalized to [-1, 1] by min-max normalization. In the coarse segmentation stage, the recall rate was maximized to effectively identify tumor areas, after which the corresponding tumor regions were cropped based on these initial results.. In the fine stage, a 3D UNet was trained based on paired CT images and corresponding ROIs to refine the coarse segmentation results.

(9^{th} place, Y. Zhang *et al.*) Zhang *et al.* employed nnUNet (Isensee et al., 2021) framework, incorporating cropping data and corresponding label based on body bounding box. Data augmentation methods, including spatial-, intensity- and label-based transformation, were used to enhance data diversity. Paired CT images were randomly cropped into patches of size 28 × 224 × 224 and used to train a 3D full-resolution UNet based on nnUNet (Isensee et al., 2021). During inference, the patch size was equal to the patch size during training, and the sliding window with a step size was half of the window size.

(10^{th} place, C. Lee *et al.*) Lee *et al.* proposed a two-step methods, consisting of localization and segmentation. In the localization stage, a 2D-based object detection network powered by the YOLO-v7 model (Wang et al., 2022) was used to identify a bounding box encompassing the GTVs. In the segmentation stage, different window widths and levels were used for multi-channel input generation. A segmentation network with DynUNet architecture was trained with these multi-channel inputs to enhance the ability to distinguish detailed features. During inference, ROIs were first extracted, and the segmentation network was used to generate the final predictions.

(11^{th} place, K. Huang *et al.*) K. Huang *et al.* employed nnUNetV2 (Isensee et al., 2021) framework, with settings consistent with those used for Task01. During inference, various augmentations were applied, including different region cropping and adjustments in scaling. The final results were obtained by averaging the predictions under different augmentations.

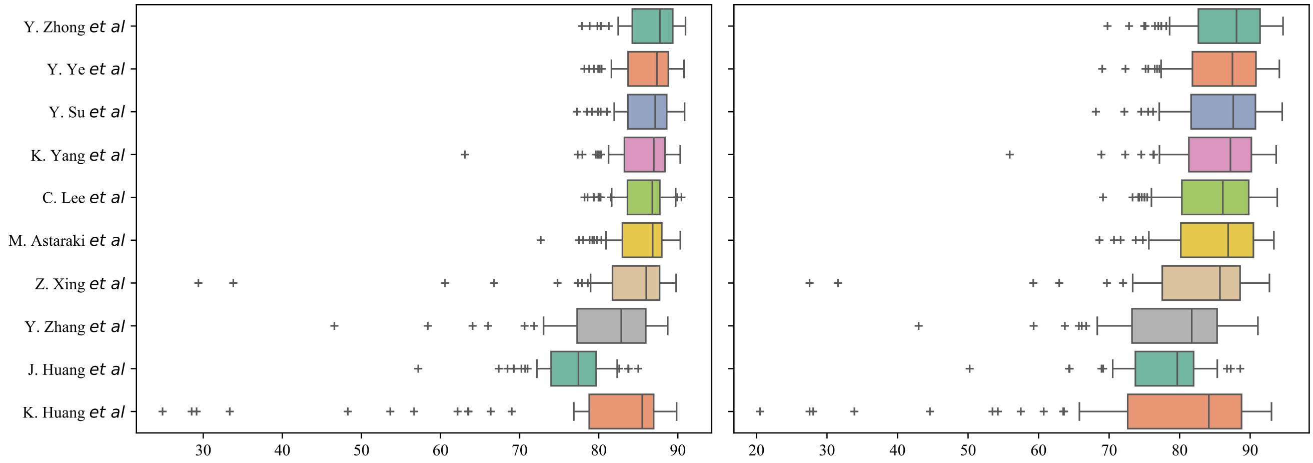
Table 10. Summary of statistical significance analysis (p-value) for the top 3 teams on the OAR segmentation task.

Team	DSC			NSD		
	Y. Zhong *et al.*	Y. Ye *et al.*	Y. Su *et al.*	Y. Zhong *et al.*	Y. Ye *et al.*	Y. Su *et al.*
Brain	0.19	0.46	0.19	0.54	0.94	0.53
BrainStem	0.10	0.04	0.61	0.18	0.08	0.62
Chiasm	0.49	0.07	0.80	0.59	0.07	0.78
Cochlea_L	0.10	0.54	0.23	0.23	0.51	0.11
Cochlea_R	9e-4	0.60	0.20	3e-4	0.56	0.24
Esophagus	0.03	0.04	0.08	0.23	0.09	0.21
ETbone_L	0.01	0.04	0.14	0.02	0.04	0.07
ETbone_R	0.22	0.55	0.39	0.37	0.90	0.88
Eye_L	0.18	0.66	0.26	0.30	0.63	0.19
Eye_R	0.57	0.80	0.61	0.70	0.69	0.76
Hippocampus_L	0.53	0.74	0.77	0.77	0.96	0.64
Hippocampus_R	0.02	0.25	0.78	0.02	0.34	0.75
IAC_L	0.62	0.03	0.42	0.94	0.05	0.48
IAC_R	4e-7	0.86	0.16	2e-7	0.75	0.18
Larynx	4e-11	0.25	0.07	3e-6	0.21	0.11
Larynx_Glottic	0.18	0.07	0.10	0.90	0.23	0.13
Larynx_Supraglot	0.03	0.16	0.13	0.21	0.52	0.23
Lens_L	0.14	0.21	0.81	0.21	0.95	0.21
Lens_R	0.13	0.12	0.30	0.11	0.46	0.64
Mandible_L	0.24	0.07	0.90	0.94	0.45	0.79
Mandible_R	0.34	8e-5	4e-3	0.72	0.50	0.41
Mastoid_L	0.26	0.33	0.39	0.37	0.64	0.21
Mastoid_R	0.21	0.04	0.44	0.32	0.69	0.30
MiddleEar_L	0.80	0.16	0.40	0.77	0.82	0.25
MiddleEar_R	5e-6	4e-6	2e-4	9e-3	5e-6	4e-4
OpticNerve_L	0.54	0.20	0.38	0.30	0.36	0.94
OpticNerve_R	0.13	0.82	0.39	0.11	0.68	0.19
OralCavity	7e-2	0.29	4e-8	0.08	0.51	7e-9
Parotid_L	0.02	0.65	8e-5	0.06	0.51	3e-10
Parotid_R	0.34	0.74	0.13	0.86	0.22	0.08
PharynxConst	0.60	0.27	0.20	0.74	0.32	0.11
Pituitary	0.89	0.54	0.42	0.96	0.39	0.38
SpinalCord	0.07	0.68	0.06	0.11	0.40	0.23
Submandibular_L	0.18	0.66	0.02	0.10	0.54	0.05
Submandibular_R	0.71	0.03	0.68	0.33	0.06	0.95
TemporalLobe_L	0.18	0.35	0.44	0.35	0.35	0.49
TemporalLobe_R	0.09	0.49	0.35	0.11	0.55	0.35
Thyroid	0.17	0.33	0.08	0.33	0.13	0.17
TMjoint_L	0.81	0.91	0.34	0.87	0.45	0.97
TMjoint_R	4e-3	0.66	0.29	0.14	0.98	0.16
Trachea	3e-5	0.50	0.92	2e-8	0.70	1.00
TympanicCavity_L	5e-3	0.08	4e-3	0.04	0.55	2e-3
TympanicCavity_R	4e-5	0.35	0.04	0.06	0.61	3e-3
VestibulSemi_L	2e-3	0.96	0.02	8e-3	0.39	0.02
VestibulSemi_R	0.30	0.16	0.36	0.35	0.37	0.34
Average	1e-6	0.08	0.15	2e-5	0.88	0.03

5. Results

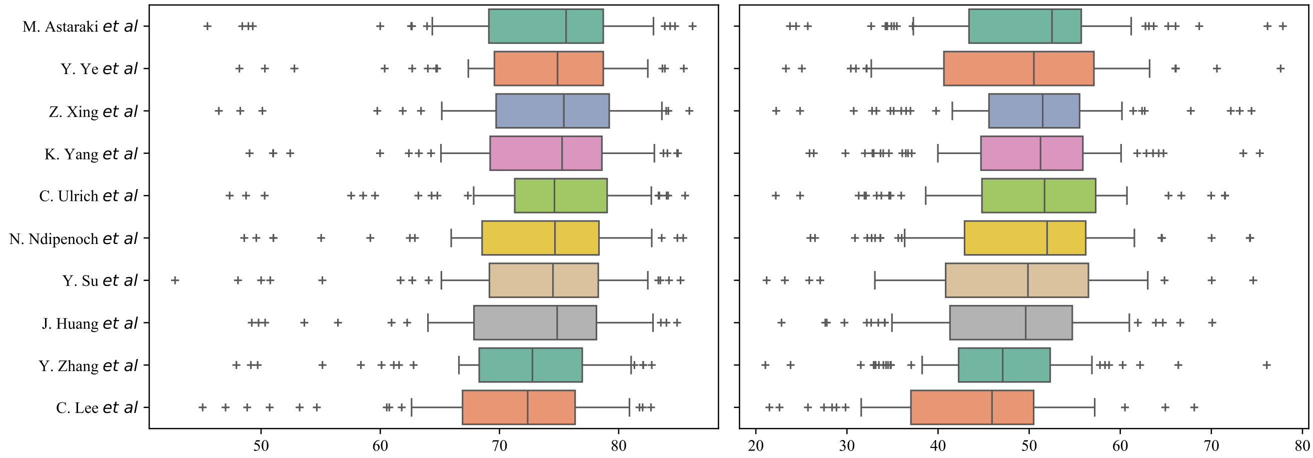
5.1. Results of Task01

The final ranking results of Task01 are listed in Table 7 sorted by their scores. Table 8 and Table 9 present the detailed performance of each team and the baseline on the OARs in terms of DSC and NSD, respectively. It can be observed that the baseline achieved average DSC and NSD scores of 84.65% and 82.88%, respectively. A total of six teams exceeded the baseline in terms of average DSC and NSD scores. The winner (Y. Zhong *et al.*) achieved the best performance on more than 30 OARs and ranked top 3 for most of the rest OARs. The top 3 teams achieved promising performance with average DSC and NSD scores over 86.14%±9.58% and 86.12%±12.79%, respectively. Figure 4 and Figure 5 (a)-(e) show the DSC and NSD score distributions of the top 5 easiest OARs obtained by all the teams, suggesting that the large-scale organs segmentations are well-solved consistently. However, these methods still perform poorly on some small, complex organs as shown in (f) to (j) Figure 4 and Figure 5. Previous works (Tang et al., 2019; Chen et al., 2021; Liao et al., 2022) performed clinical assessments



(a) Average DSC of Task01 (b) Average NSD of Task01

Fig. 2. Boxplot of the patient-level average segmentation performance for OARs in terms of DSC and NSD.



(a) Average DSC of Task02 (b) Average NSD of Task02

Fig. 3. Boxplot of the patient-level average segmentation performance for GTVs in terms of DSC and NSD.

Table 11. Rankings of methods in terms of DSC and NSD scores for GTV segmentation

Method	DSC Rank			NSD Rank			Overall
	GTVp	GTVnd	Average	GTVp	GTVnd	Average	
M. Astaraki *et al.*	3	4	3.5	1	4	2.5	1
Y. Ye *et al.*	2	3	2.5	2	6	4	2
Z. Xing *et al.*	7	1	4	3	2	2.5	3
K. Yang *et al.*	1	5	3	4	5	4.5	4
C. Ulrich *et al.*	8	2	5	6	1	3.5	5
N. Ndipenoch *et al.*	5	6	5.5	5	3	4	6
Y. Su *et al.*	6	7	6.5	7	7	7	7
J. Huang *et al.*	4	8	6	8	8	8	8
Y. Zhang *et al.*	10	9	9.5	9	9	9	9
C. Lee *et al.*	9	11	10	10	11	10.5	10
K. Huang *et al.*	11	10	10.5	11	10	10.5	11

Table 12. Summary of the quantitative evaluation results of GTVp and GTVnd segmentation by the eleven teams.

Team	DSC (%)			NSD (%)		
	GTVp	GTVnd	Average	GTVp	GTVnd	Average
M. Astaraki *et al.*	78.56±7.54	67.75±14.64	73.15±12.83	**36.61±12.17**	63.15±16.24	49.88±19.55
Y. Ye *et al.*	78.76±7.16	68.10±12.17	73.43±11.31	36.45±11.70	62.26±15.57	49.36±18.87
Z. Xing *et al.*	78.07±7.82	**69.28±12.12**	**73.68±11.11**	36.44±12.25	64.04±14.37	**50.24±19.20**
K. Yang *et al.*	**78.76±6.60**	67.41±13.78	73.09±12.21	35.92±11.05	63.08±15.37	49.50±19.07
C. Ulrich *et al.*	77.71±7.79	69.18±12.80	73.44±11.42	35.60±11.66	**64.76±15.04**	50.18±19.84
N. Ndipenoch *et al.*	78.25±7.54	67.21±14.52	72.73±12.82	35.90±11.87	63.31±15.78	49.61±19.56
Y. Su *et al.*	78.13±7.27	66.91±14.54	72.52±12.79	35.21±11.11	62.24±16.00	48.73±19.30
J. Huang *et al.*	78.36±7.09	66.36±14.09	72.36±12.66	34.18±10.26	61.96±15.48	48.07±19.12
Y. Zhang *et al.*	76.89±7.37	66.25±12.74	71.57±11.69	33.22±10.66	60.30±13.94	46.76±18.37
C. Lee *et al.*	77.46±7.53	63.39±13.85	70.42±13.18	32.96±10.69	55.62±14.51	44.29±17.05
K. Huang *et al.*	76.71±6.85	65.97±12.04	71.34±11.17	32.76±9.61	59.70±13.34	46.23±17.79
Baseline	75.80±7.28	66.83±11.48	71.32±10.61	33.41±11.61	61.49±13.06	47.45±18.70

and found that most clinically acceptable segmentations have a good DSC score (DSC > 80%). However, in this challenge, the average DSC and NSD of the chiasm and esophagus are around 72% and 77% respectively, which may be not clinically applicable without user revision.

Figure 2 provides the boxplots of DSC and NSD scores of each team based on patient-level average segmentation. The best average Dice and NSD scores were both achieved by Y. Zhong *et al.*. In general, the patient-level average DSC and NSD scores achieved promising results that are larger than 80%. In addition, to show the significance among the top 3 teams with others, we calculated the paired *t-test* between the ranking *n-th* team and the ranking *(n+1)-th* team (*n* ranges from

Table 13. Summary of statistical significance analysis (p-value) for the top 3 teams on the GTV segmentation task.

Team	DSC			NSD		
	M. Astaraki *et al.*	Y. Ye *et al.*	Z. Xing *et al.*	M. Astaraki *et al.*	Y. Ye *et al.*	Z. Xing *et al.*
GTVp	0.55	0.16	0.18	0.81	0.99	0.54
GTVnd	0.68	0.17	0.12	0.30	0.04	0.34
Average	0.55	0.60	0.41	0.38	0.13	0.32

1 to 3). Table 10 presents the statistical analysis results of the top 3 teams. It can be observed that the winner is significantly superior (p-value < 0.05) to the second place in terms of average DSC and NSD scores. However, there are no significant differences between the second and third teams, which averaged DSC scores are 86.36%±9.15% and 86.14%±9.58%, and NSD scores are 86.09% and 86.12%, respectively. Compared with the fourth team which achieved average DSC and NSD scores of 85.62%±10.48% and 85.33%±13.42%, the third team achieved significantly better NSD scores (86.12%±12.79%) and comparable DSC scores (86.14%±9.58%).

5.2. Results of Task02

Table 11 presents the final ranking scores of the GTV segmentation. It can be seen that M. Astaraki *et al.* won first place with an average ranking score of 3. Y. Ye *et al.* and Z. Xing *et al.* achieved the same average ranking score of 3.25, but the standard deviation of Y. Ye *et al.* was smaller, so the final ranking results were that Y. Ye *et al.* and Z. Xing *et al.* won the second and third places, respectively. The detailed performance of all teams and the baseline (pure nnUNet with a default setting of 3d_fullres) is shown in Table 12 and Figure 3. A total of 10 and 8 teams outperformed the baseline in terms of average DSC and NSD scores, respectively, as shown in (k) and (l) in Figure 4) and Figure 5. Four teams obtained encouraging performance with average DSC scores greater than 73%. In addition, all submissions of Task02 performed well on the GTVp segmentation with DSC higher than 76.71%±6.85%, and the DSC scores in GTVnd segmentation have a larger variability ranging from 63.39%±13.85% to 69.28%±12.12%. In addition, we also found that most of the methods can not achieve promising performances on both GTVp and GTVnd segmentation at the same time. These results demonstrated that the automatic GTVp and GTVs contouring is still a challenging and unsolved problem, and more attention should be paid to improve the segmentation performance further.

Different from the results of Task01, these teams that used nnUNet or its variants achieved similar results on the GTV segmentation task. The average performance gap between the winner and the 11-*th* ranking team was nearly 2 and 3 percentage points in terms of DSC and NSD scores. Compared with the pure nnUNet baseline (the last line in Table 12), eight teams achieved better results in both terms of DSC and NSD scores. Although the segmentation results are consistent and robust, there are huge performance gaps between these methods and real clinical requirements according to previously reported user studies and clinical assessments (Lin et al., 2019; Liao et al., 2022; Luo et al., 2023), where the DSC of the clinically applicable results ranged from 80% to 90%.

Figure 3 shows the boxplots of DSC and NSD from the patient-level GTV segmentation of each team. M. Astaraki *et al.* achieved the best average DSC and NSD scores. It can be seen that the median of patient-level average DSC and NSD scores of Y. Ye *et al.* were both lower than that of Z. Xing *et al.*. The fourth place achieved similar performances with Z. Xing *et al.* at patient-level. Table 13 presents a detailed statistical analysis of the top 3 teams. The results show that there are no significant performance differences in terms of DSC and NSD scores between the winner and the second-place method except for the numerical values and the ranking scores. Similar trends can be found in the pair of the second and third places, no significant performance differences were found except for the NSD score in GTVnd segmentation. Besides, it can be noticed from Table 12 and Table 13 that Z. Xing *et al* obtained the best average performance in both terms of DSC and NSD, but this team ranked on the third place due to the low overall ranking score. In addition, C. Ulrich *et al* achieved the best NSD and second DSC in GTVnd segmentation and were not even included in the top 3 teams yet caused by the insufficient results in GTVp segmentation. These results show the ranking scheme of this challenge (rank-then-aggregate (Dorent et al., 2023)) is robust and alleviates the impact of some extremely good or bad results.

5.3. Visualization

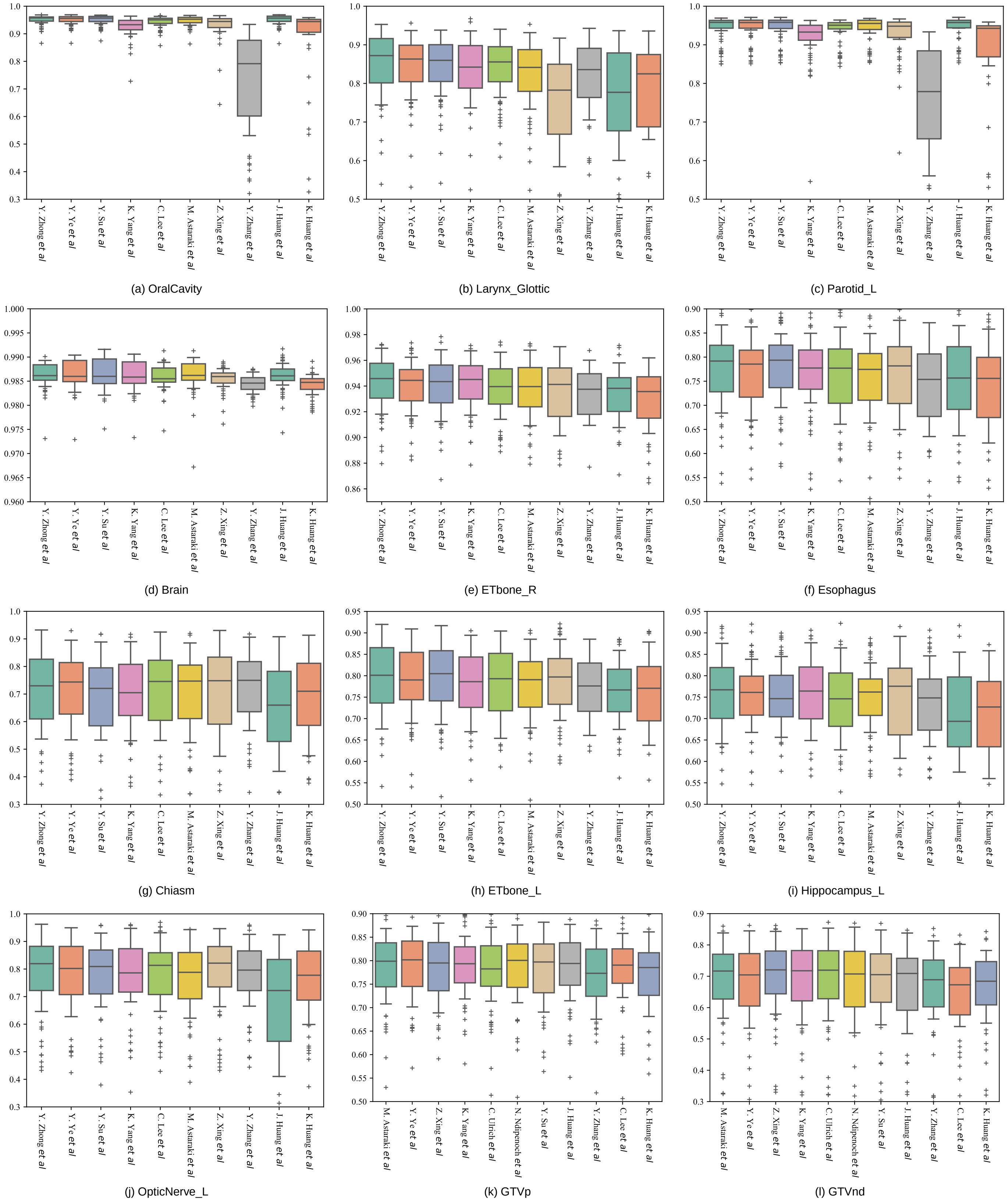
Figure 6 visually presents the OAR segmentation outcomes from the top three performing teams. To show segmentation differences, we selected three patients based on the lower quartile (LQ), median quartile (MQ), and high quartile (HQ) of the average DSC and NSD scores across the top three teams and the 45 OARs. The results highlight that these methods achieve accurate segmentations for larger organs such as BrainStem, Parotid_L, and Parotid_R. However, challenges persist in accurately segmenting small and intricate organs. For instance, the Chiasm exhibits under-segmentation, particularly in the case of the LQ patient. Figure 7 visualizes the GTV segmentation results of the top 3 teams. These results show that the GTVp and GTVnd segmentation are still challenging. Specifically, most GTVp segmentation results suffer from under-segmentation (in HQ, MQ and LQ patients). Additionally, some GTVnd cannot even be identified and segmented in the case of the LQ patient. These findings highlight the challenge of achieving precise and automated GTV segmentation, which warrants heightened attention and further investigation.

6. Discussion

In this section, we discuss the potential solutions, limitations, and future directions of automatic segmentation in radiation therapy planning and provide some insights about the clinically applicable OAR and GTV segmentation.

6.1. OAR segmentation in head and neck

All submitted algorithms demonstrated that supervised learning can achieve promising mean performance (> 80%) in terms



(a) OralCavity

(b) Larynx_Glottic

(c) Parotid_L

(d) Brain

(e) ETbone_R

(f) Esophagus

(g) Chiasm

(h) ETbone_L

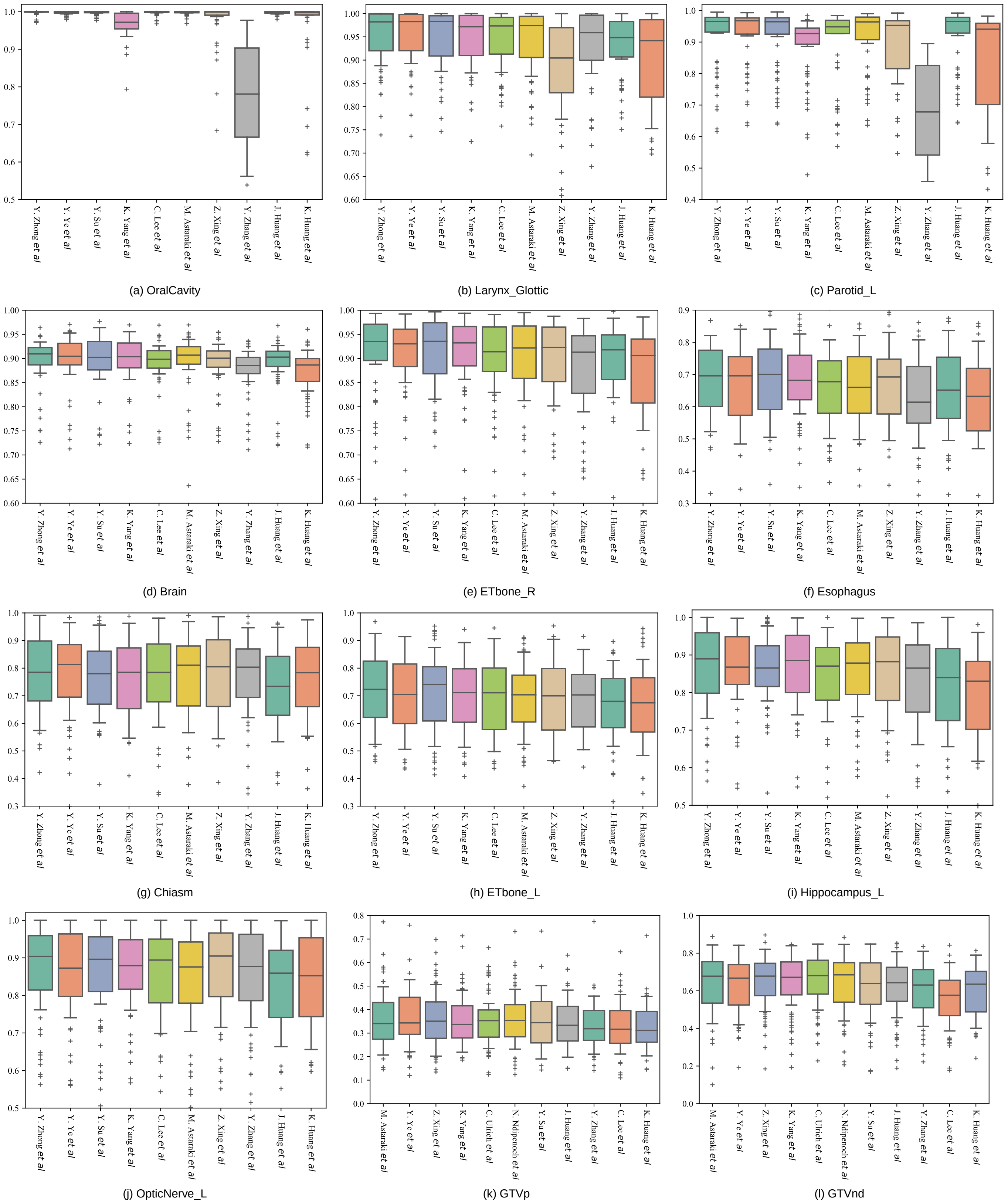
(i) Hippocampus_L

(j) OpticNerve_L

(k) GTVp

(l) GTVnd

Fig. 4. Boxplot of the patient-level average segmentation performance for top 5 easiest and hardest OARs and 2 GTVs in terms of DSC. (a)-(e): top 5 easiest OARs, (f)-(j): top 5 hardest OARs.



(a) OralCavity

(b) Larynx_Glottic

(c) Parotid_L

(d) Brain

(e) ETbone_R

(f) Esophagus

(g) Chiasm

(h) ETbone_L

(i) Hippocampus_L

(j) OpticNerve_L

(k) GTVp

(l) GTVnd

Fig. 5. Boxplot of the patient-level average segmentation performance for top 5 easiest and hardest OARs and 2 GTVs in terms of NSD. (a)-(e): top 5 easiest OARs, (f)-(j): top 5 hardest OARs.

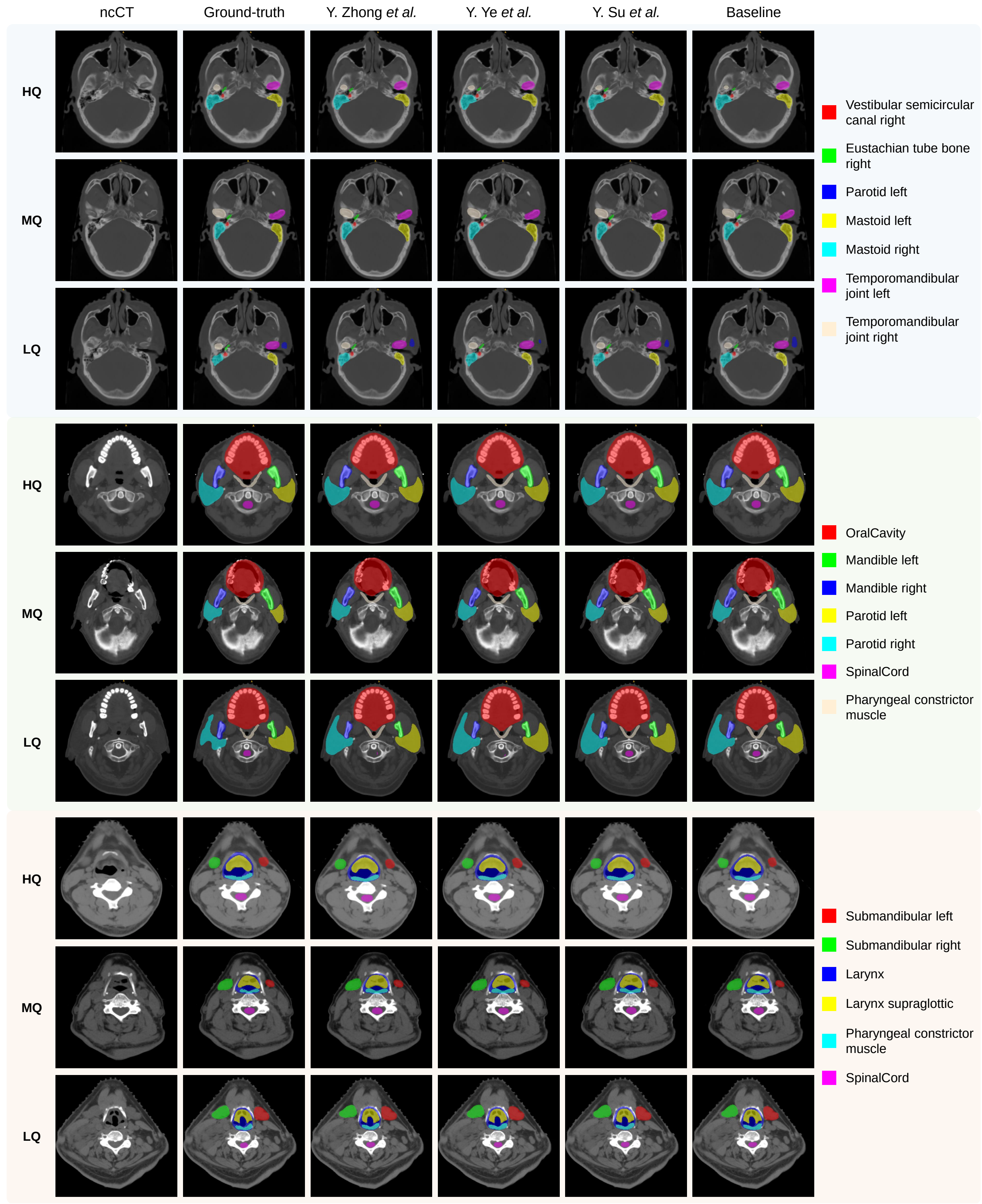


Fig. 6. Qualitative OAR segmentation using the Top3 teams and baseline on the SegRap2023 testing set.

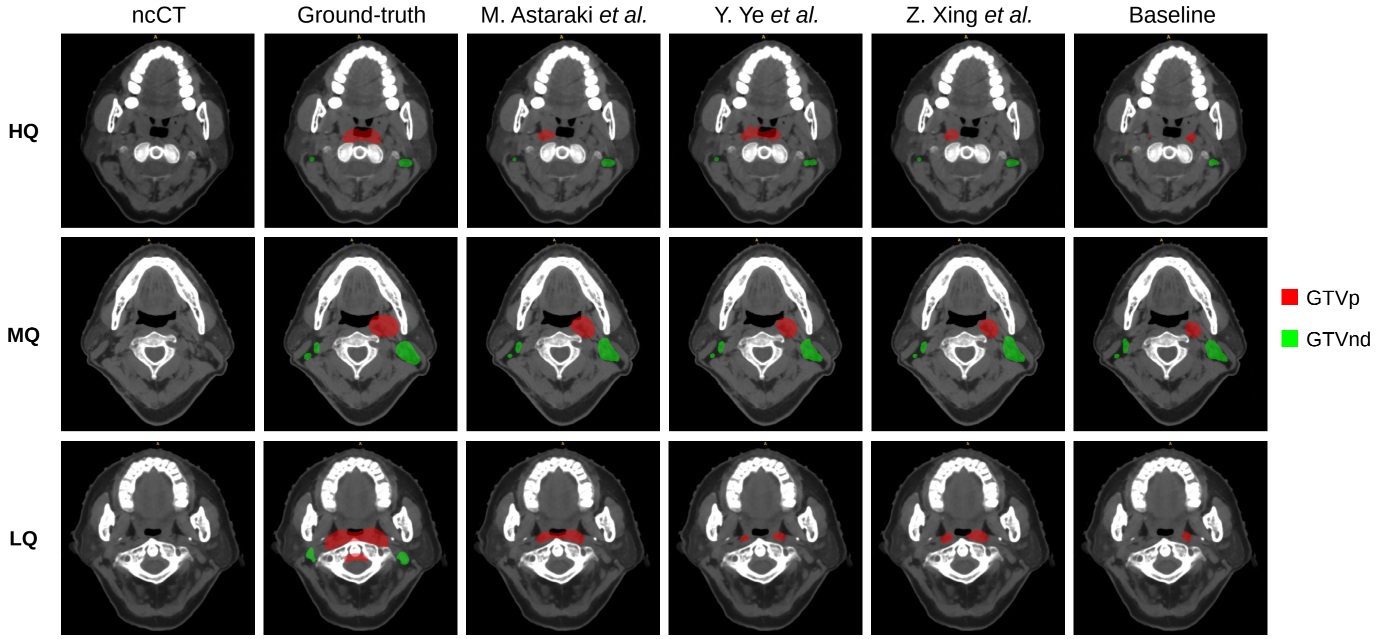


Fig. 7. Qualitative GTV segmentation using the Top3 teams and baseline on the SegRap2023 testing set.

of DSC and NSD scores. However, the results of some complex OARs are still not good enough (< 80%). The reason may be most of these solutions are based on one-stage segmentation and do not apply specific designs for complex or small organs. The winner's solution demonstrated that structure-specific label generation and boundary refinement can obtain encouraging performance improvement over the baseline. Meanwhile, imbalance problems and inequality optimization exist when segmenting 45 OARs directly. Applying the balance loss (Lin et al., 2017) and stratified optimization (Ye et al., 2022) may bring benefits to improve the segmentation performance of the small and complex OAR, but there are no participants that have investigated the performance of these methods.

Interestingly, almost all teams used nnUNet (Isensee et al., 2021) or its variants as the baseline, but their performances were hugely different. For example, the performance of the winner and the K. Huang *et al.* methods is significantly different, 86.70%±9.30% *vs* 78.14%±23.65% in terms of DSC score. Meanwhile, four teams performed worse than the baseline, removing some spatial data augmentations and highlighting the necessity of designing specific data-processing strategies, network modules, training, or testing approaches for this task according to the data characteristics. Specifically, the data process and augmentations significantly impact performance, such as the winner merging the left and right counterparts into one and removing the mirror augmentation strategy, leading to the most significant improvement on the original nnUNetV2. Besides, the model ensemble also leads to performance differences, but it does not mean it can consistently improve performance by increasing the model numbers. These findings can provide some insights for powerful OAR segmentation model development where some appropriate data augmentation and pre- or post-processing are important and should be tuned based on the data characteristics.

Recently, the universal model with transfer learning has shown promising performance on multiple medical image segmentation tasks (Liu et al., 2023; Ye et al., 2023; Wang et al., 2023b). The second place solution shows the transferable ability of the universal model (Ye et al., 2023) from other tasks to the head and neck OAR segmentation. The third place method proved that large patch size and simple task-driven data processing methods except for mirror operation can boost segmentation performance. Note that although with different datasets, the top 3 teams reached a promising performance with an average DSC of above 86.14%±9.58%, which is superior to previous head and neck OAR segmentation studies with an average DSC of below 84.5% (Tang et al., 2019; Gao et al., 2021; Lei et al., 2021). These results also provided a fair baseline and benchmarking results for further research.

6.2. NPC GTV segmentation

All submitted methods for GTV segmentation obtained comparable results. The top 3 teams applied the two-stage segmentation with intensity distribution harmonization, transfer learning, and test-time augmentation strategies to handle the inherent and challenging problems in GTV segmentation, respectively. However, none of the top 3 teams surpassed 80% in terms of DSC or NSD scores, and the visualization in Figure 7 shows there are under-segmentation and even targets missing. Besides, the results show that the training strategies do not lead to significant performance differences except the intensity-based data augmentations, suggesting that we should choose suitable intensity-based augmentation methods when developing high-performance GTV segmentation models. In addition, there are still huge segmentation performance gaps between the challenge benchmarks (average DSC of 75.8%±7.28% and

66.83%±11.48% for GTVp and GTVnd) and previous works (average DSC of 79.0% and 74.0% for GTVp and GTVnd) (Luo et al., 2023; Li et al., 2022; Liao et al., 2022; Lin et al., 2019; Wang et al., 2023c). The main reason caused the performance gaps is that these works segmented the GTV from multi-sequence MRI, where the MRI has higher quality and clearer contrast between normal tissue and GTV. However, most planning, dose estimation and radiation treatment were performed based on CT and MRI was just used as a delineation reference modality. Recently, Mei et al. (2021) reported the performance of NPC GTV segmentation from CT is 65.66% (won the second place in StructSeg2019) which conforms to the findings of this challenge. These results highlight the urgency of developing an accurate GTV segmentation method to handle the inherent challenges and further evaluate in the clinical practice.

There are some potential directions to enhance the GTV segmentation performance: 1) exploiting the position and boundary-aware feature attention method to describe the variable location and irregular boundary of GTV (Li et al., 2022); 2) investigating the performance improvement by using the OAR segmentation to provide the anatomical information (Yan et al., 2023); 3) mining the complementary information across ncCT and ceCT scans to highlight the target representation, which not be noticed by recent works; 4) employing pre-trained models to capture comprehensive common semantic features for targets (Ye et al., 2023).

6.3. The gap between clinically applicable segmentation

The ultimate goal of developing automatic OAR and GTV segmentation methods is to accelerate the clinical delineation workflow and reduce the radiation oncologists' burden. In clinical practice, most automatic segmentation methods can not be applied directly and need radiation oncologists to refine, especially for the online IMRT system (Luo et al., 2021c). Recent studies Tang et al. (2019) claimed that the deep learning-based automatic contouring system with a mean DSC of 78.34% over 28 OARs was clinically applicable after minor revision. Some studies (Liao et al., 2022) and (Luo et al., 2023) also performed clinical studies on GTVp and GTVnd segmentation and showed that the deep learning segmentation system can be clinically accepted with few refinements when the DSC of GTVp and GTVnd are greater than 83% and 80%. According to these studies, most solutions for the SegRap2023 challenge have achieved clinically applicable results for most OARs. However, there are still huge gaps between the performance of these methods and the clinically acceptable results for the GTVs.

6.4. Limitation and future direction

Compared with the abdominal organ and tumor segmentation (Luo et al., 2021b; Gibson et al., 2018; Isensee et al., 2021), there are very few works that have built large-scale datasets and comprehensively evaluated the performance of recent methods for the OARs and GTVs of head and neck cancer. Although this work has developed a large-scale dataset and evaluated more than ten cut-edge methods, it still faces limitations in terms of robustness and generalization evaluation, primarily attributed to the absence of a multi-center dataset. Additionally, the dataset exclusively focuses on NPC patients, overlooking the diverse range of patients encompassed by head and neck cancer. Despite the inclusion of annotations for 45 OARs and 2 GTVs in the SegRap2023 challenge, there is an omission of several radiotherapy-required Clinical Target Volumes (CTV). To address these shortcomings, we plan to enlarge the scale of the dataset and data source and further extend the segmentation tasks to more categories in the future.

7. Conclusion

This work summarizes the submitted methods from the SegRap2023 challenge, which provides 200 paired CT scans for the segmentation of 45 OARs and 2 GTVs for NPC patients. To the best of our knowledge, SegRap2023 has the most comprehensive and exhausted labeled dataset among existing OAR and GTV segmentation challenges so far. A total of ten and eleven algorithms successfully submitted their solutions that met the challenge requirements. They were benchmarked for comparisons in the OAR and GTV segmentation, respectively, and their methods and results were analyzed. The results demonstrate that most large-size OARs can be segmented accurately and can be seen as a well-solved problem. However, for the small-size OARs and GTVs, there are still huge gaps between segmentation performance and clinical applicability, suggesting that future research should focus on these unsolved problems more. In the future, we plan to extend this challenge in the aspect of data scale, source, and categories to be more suitable for the clinical requirement.

8. Acknowledgment

This work was supported by the National Natural Science Foundation of China [Grant 62271115, Grant 82203197], the Sichuan Science and Technology Program, China (Grant 2022YFSY0055, 2023NSFSC1852), the Sichuan Provincial Cadre Health Research Project (Grant/award number: 2023-803) and the Radiation Oncology Key Laboratory of Sichuan Province Open Fund (2022ROKF04). We would like to thank M.D. S.C. Zhang, M.D. W. Liao, M.D. Y. Zhao, M.D. C. Li and their team members for data collection, annotation, and checking. We also would like to thank the support team of the Grand Challenge Platform and the MICCAI challenge organization team for their sincere help while hosting the challenge. We also would like to thank all participants for their active participation and working hard.

References

Amin, M.B., Greene, F.L., Edge, S.B., Compton, C.C., Gershenwald, J.E., Brookland, R.K., Meyer, L., Gress, D.M., Byrd, D.R., Winchester, D.P., 2017. The eighth edition ajcc cancer staging manual: continuing to build a bridge from a population-based to a more "personalized" approach to cancer staging. CA: a cancer journal for clinicians 67, 93–99.

Ang, K.K., Zhang, Q., Rosenthal, D.I., Nguyen-Tan, P.F., Sherman, E.J., Weber, R.S., Galvin, J.M., Bonner, J.A., Harris, J., El-Naggar, A.K., et al., 2014. Randomized phase iii trial of concurrent accelerated radiation plus cisplatin with or without cetuximab for stage iii to iv head and neck carcinoma: Rtog 0522. Journal of clinical oncology 32, 2940.

Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., Shinohara, R., Berger, C., Ha, S., Rozycki, M., et al., 2018. Identifying the best machine learning algorithms for brain tumor segmentation. progression assessment, and overall survival prediction in the BRATS challenge 10.

Bilic, P., Christ, P.F., Vorontsov, E., Chlebus, G., Chen, H., Dou, Q., Fu, C.W., Han, X., Heng, P.A., Hesser, J., et al., 2019. The liver tumor segmentation benchmark (lits). arXiv preprint arXiv:1901.04056 .

Chen, X., Sun, S., Bai, N., Han, K., Liu, Q., Yao, S., Tang, H., Zhang, C., Lu, Z., Huang, Q., et al., 2021. A deep learning-based auto-segmentation system for organs-at-risk on whole-body computed tomography images for radiation therapy. Radiotherapy and Oncology 160, 175–184.

Chua, M.L., Wee, J.T., Hui, E.P., Chan, A.T., 2016. Nasopharyngeal carcinoma. The Lancet 387, 1012–1024.

Dong, X., Lei, Y., Wang, T., Thomas, M., Tang, L., Curran, W.J., Liu, T., Yang, X., 2019. Automatic multiorgan segmentation in thorax ct images using u-net-gan. Medical physics 46, 2157–2168.

Dorent, R., Kujawa, A., Ivory, M., Bakas, S., Rieke, N., Joutard, S., Glocker, B., Cardoso, J., Modat, M., Batmanghelich, K., et al., 2023. Crossmoda 2021 challenge: Benchmark of cross-modality domain adaptation techniques for vestibular schwannoma and cochlea segmentation. Medical Image Analysis 83, 102628.

Feng, X., Qing, K., Tustison, N.J., Meyer, C.H., Chen, Q., 2019. Deep convolutional neural network for segmentation of thoracic organs-at-risk using cropped 3d images. Medical physics 46, 2169–2180.

Gao, Y., Huang, R., Chen, M., Wang, Z., Deng, J., Chen, Y., Yang, Y., Zhang, J., Tao, C., Li, H., 2019. Focusnet: imbalanced large and small organ segmentation with an end-to-end deep neural network for head and neck ct images, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part III 22, Springer. pp. 829–838.

Gao, Y., Huang, R., Yang, Y., Zhang, J., Shao, K., Tao, C., Chen, Y., Metaxas, D.N., Li, H., Chen, M., 2021. Focusnetv2: Imbalanced large and small organ segmentation with adversarial shape constraint for head and neck ct images. Medical Image Analysis 67, 101831.

Gibson, E., Giganti, F., Hu, Y., Bonmati, E., Bandula, S., Gurusamy, K., Davidson, B., Pereira, S.P., Clarkson, M.J., Barratt, D.C., 2018. Automatic multi-organ segmentation on abdominal ct with dense v-networks. TMI 37, 1822–1834.

Guo, D., Jin, D., Zhu, Z., Ho, T.Y., Harrison, A.P., Chao, C.H., Xiao, J., Lu, L., 2020. Organ at risk segmentation for head and neck cancer using stratified learning and neural architecture search, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4223–4232.

He, W., Zhang, C., Dai, J., Liu, L., Wang, T., Liu, X., Jiang, Y., Li, N., Xiong, J., Wang, L., et al., 2024. A statistical deformation model-based data augmentation method for volumetric medical image segmentation. Medical Image Analysis 91, 102984.

Huang, Z., Wang, H., Deng, Z., Ye, J., Su, Y., Sun, H., He, J., Gu, Y., Gu, L., Zhang, S., et al., 2023. Stu-net: Scalable and transferable medical image segmentation models empowered by large-scale supervised pre-training. arXiv preprint arXiv:2304.06716 .

Iglesias, J.E., Sabuncu, M.R., 2015. Multi-atlas segmentation of biomedical images: a survey. Medical image analysis 24, 205–219.

Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H., 2021. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods 18, 203–211.

Kam, M.K., Chau, R.M., Suen, J., Choi, P.H., Teo, P.M., 2003. Intensity-modulated radiotherapy in nasopharyngeal carcinoma: dosimetric advantage over conventional plans and feasibility of dose escalation. International Journal of Radiation Oncology* Biology* Physics 56, 145–157.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al., 2023. Segment anything. arXiv preprint arXiv:2304.02643 .

Kosmin, M., Ledsam, J., Romera-Paredes, B., Mendes, R., Moinuddin, S., de Souza, D., Gunn, L., Kelly, C., Hughes, C., Karthikesalingam, A., et al., 2019. Rapid advances in auto-segmentation of organs at risk and target volumes in head and neck cancer. Radiotherapy and Oncology 135, 130–140.

Lee, A., Ma, B., Ng, W.T., Chan, A., et al., 2015. Management of nasopharyngeal carcinoma: current practice and future perspective. J Clin Oncol 33, 3356–3364.

Lee, A.W., Ng, W.T., Pan, J.J., Poh, S.S., Ahn, Y.C., AlHussain, H., Corry, J., Grau, C., Grégoire, V., Harrington, K.J., et al., 2018. International guideline for the delineation of the clinical target volumes (ctv) for nasopharyngeal carcinoma. Radiotherapy and Oncology 126, 25–36.

Lei, W., Mei, H., Sun, Z., Ye, S., Gu, R., Wang, H., Huang, R., Zhang, S., Zhang, S., Wang, G., 2021. Automatic segmentation of organs-at-risk from head-and-neck ct using separable convolutional neural network with hard-region-weighted loss. Neurocomputing 442, 184–199.

Li, S., Xiao, J., He, L., Peng, X., Yuan, X., 2019. The tumor target segmentation of nasopharyngeal cancer in ct images based on deep learning methods. Technology in cancer research & treatment 18, 1533033819884561.

Li, Y., Dan, T., Li, H., Chen, J., Peng, H., Liu, L., Cai, H., 2022. Npcnet: jointly segment primary nasopharyngeal carcinoma tumors and metastatic lymph nodes in mr images. IEEE Transactions on Medical Imaging 41, 1639–1650.

Liao, W., He, J., Luo, X., Wu, M., Shen, Y., Li, C., Xiao, J., Wang, G., Chen, N., 2022. Automatic delineation of gross tumor volume based on magnetic resonance imaging by performing a novel semisupervised learning framework in nasopharyngeal carcinoma. International Journal of Radiation Oncology* Biology* Physics 113, 893–902.

Lin, L., Dou, Q., Jin, Y.M., Zhou, G.Q., Tang, Y.Q., Chen, W.L., Su, B.A., Liu, F., Tao, C.J., Jiang, N., et al., 2019. Deep learning for automated contouring of primary tumor volumes by mri for nasopharyngeal carcinoma. Radiology 291, 677–686.

Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017. Focal loss for dense object detection, in: ICCV, pp. 2980–2988.

Liu, J., Zhang, Y., Chen, J.N., Xiao, J., Lu, Y., A Landman, B., Yuan, Y., Yuille, A., Tang, Y., Zhou, Z., 2023. Clip-driven universal model for organ segmentation and tumor detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 21152–21164.

Liu, Y., Yuan, X., Jiang, X., Wang, P., Kou, J., Wang, H., Liu, M., 2021. Dilated adversarial u-net network for automatic gross tumor volume segmentation of nasopharyngeal carcinoma. Applied Soft Computing 111, 107722.

Luo, X., Liao, W., Chen, J., Song, T., Chen, Y., Zhang, S., Chen, N., Wang, G., Zhang, S., 2021a. Efficient semi-supervised gross target volume of nasopharyngeal carcinoma segmentation via uncertainty rectified pyramid consistency, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24, Springer. pp. 318–329.

Luo, X., Liao, W., He, Y., Tang, F., Wu, M., Shen, Y., Huang, H., Song, T., Li, K., Zhang, S., et al., 2023. Deep learning-based accurate delineation of primary gross tumor volume of nasopharyngeal carcinoma on heterogeneous magnetic resonance imaging: A large-scale and multi-center study. Radiotherapy and Oncology 180, 109480.

Luo, X., Liao, W., Xiao, J., Chen, J., Song, T., Zhang, X., Li, K., Metaxas, D.N., Wang, G., Zhang, S., 2021b. Word: A large scale dataset, benchmark and clinical applicable study for abdominal organ segmentation from ct image. arXiv preprint arXiv:2111.02403 .

Luo, X., Wang, G., Song, T., Zhang, J., Aertsen, M., Deprest, J., Ourselin, S., Vercauteren, T., Zhang, S., 2021c. Mideepseg: Minimally interactive segmentation of unseen objects from medical images using deep learning. Medical image analysis 72, 102102.

Maier-Hein, L., Reinke, A., Kozubek, M., Martel, A.L., Arbel, T., Eisenmann, M., Hanbury, A., Jannin, P., Müller, H., Onogur, S., et al., 2020. Bias: Transparent reporting of biomedical image analysis challenges. Medical image analysis 66, 101796.

Mei, H., Lei, W., Gu, R., Ye, S., Sun, Z., Zhang, S., Wang, G., 2021. Automatic segmentation of gross target volume of nasopharynx cancer using ensemble of multiscale deep neural networks with spatial attention. Neurocomputing 438, 211–222.

Nikolov, S., Blackwell, S., Zverovitch, A., Mendes, R., Livne, M., De Fauw, J., Patel, Y., Meyer, C., Askham, H., Romera-Paredes, B., et al., 2021. Clinically applicable segmentation of head and neck anatomy for radiotherapy: deep learning algorithm development and validation study. Journal of medical Internet research 23, e26151.

Oreiller, V., Andrearczyk, V., Jreige, M., Boughdad, S., Elhalawani, H., Castelli, J., Vallieres, M., Zhu, S., Xie, J., Peng, Y., et al., 2022. Head and neck tumor segmentation in pet/ct: the hecktor challenge. Medical image analysis 77, 102336.

Podobnik, G., Strojan, P., Peterlin, P., Ibragimov, B., Vrtovec, T., 2023. Hanseg: The head and neck organ-at-risk ct and mr segmentation dataset. Medical physics 50, 1917–1927.

Raudaschl, P.F., Zaffino, P., Sharp, G.C., Spadea, M.F., Chen, A., Dawant, B.M., Albrecht, T., Gass, T., Langguth, C., Lüthi, M., et al., 2017. Evaluation of segmentation methods on head and neck ct: auto-segmentation challenge 2015. Medical physics 44, 2020–2036.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, Springer. pp. 234–241.

Sahbaee, P., Abadi, E., Segars, W.P., Marin, D., Nelson, R.C., Samei, E., 2017. The effect of contrast material on radiation dose at ct: Part ii. a systematic evaluation across 58 patient models. Radiology 283, 749–757.

Shi, F., Hu, W., Wu, J., Han, M., Wang, J., Zhang, W., Zhou, Q., Zhou, J., Wei, Y., Shao, Y., et al., 2022. Deep learning empowered volume delineation of whole-body organs-at-risk for accelerated radiotherapy. Nature Communications 13, 6566.

Simpson, A.L., Antonelli, M., Bakas, S., Bilello, M., Farahani, K., Van Ginneken, B., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., et al., 2019. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. arXiv preprint arXiv:1902.09063 .

Sun, X.S., Liu, S.L., Luo, M.J., Li, X.Y., Chen, Q.Y., Guo, S.S., Wen, Y.F., Liu, L.T., Xie, H.J., Tang, Q.N., et al., 2019. The association between the development of radiation therapy, image technology, and chemotherapy, and the survival of patients with nasopharyngeal carcinoma: a cohort study from 1990 to 2012. International Journal of Radiation Oncology* Biology* Physics 105, 581–590.

Tang, H., Chen, X., Liu, Y., Lu, Z., You, J., Yang, M., Yao, S., Zhao, G., Xu, Y., Chen, T., et al., 2019. Clinically applicable deep learning framework for organs at risk delineation in ct images. Nature Machine Intelligence 1, 480–491.

Ulrich, C., Isensee, F., Wald, T., Zenk, M., Baumgartner, M., Maier-Hein, K.H., 2023. MultiTalent: A multi-dataset approach to medical image segmentation, in: MICCAI, pp. 648–658.

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., 2017. Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer. Scientific reports 7, 10117.

Wang, D., Wang, X., Wang, L., Li, M., Da, Q., Liu, X., Gao, X., Shen, J., He, J., Shen, T., et al., 2023a. Medfmc: A real-world dataset and benchmark for foundation model adaptation in medical image classification. arXiv preprint arXiv:2306.09579 .

Wang, G., Wu, J., Luo, X., Liu, X., Li, K., Zhang, S., 2023b. Mis-fm: 3d medical image segmentation using foundation models pretrained on a large-scale unannotated dataset. arXiv preprint arXiv:2306.16925 .

Wang, H., Chen, J., Zhang, S., He, Y., Xu, J., Wu, M., He, J., Liao, W., Luo, X., 2023c. Dual-reference source-free active domain adaptation for nasopharyngeal carcinoma tumor segmentation across multiple hospitals. arXiv preprint arXiv:2309.13401 .

Wang, H., Guo, S., Ye, J., Deng, Z., Cheng, J., Li, T., Chen, J., Su, Y., Huang, Z., Shen, Y., et al., 2023d. Sam-med3d. arXiv preprint arXiv:2310.15161 .

Wang, R., Kang, M., 2021. Guidelines for radiotherapy of nasopharyngeal carcinoma. Precision Radiation Oncology 5, 122–159.

Wang, X., Yang, G., Zhang, Y., Zhu, L., Xue, X., Zhang, B., Cai, C., Jin, H., Zheng, J., Wu, J., et al., 2020. Automated delineation of nasopharynx gross tumor volume for nasopharyngeal carcinoma by plain ct combining contrast-enhanced ct using deep learning. Journal of Radiation Research and Applied Sciences 13, 568–577.

Wang, Y., Wang, H., Xin, Z., 2022. Efficient detection model of steel strip surface defects based on yolo-v7. IEEE Access 10, 133936–133944.

Wasserthal, J., Breit, H.C., Meyer, M.T., Pradella, M., Hinck, D., Sauter, A.W., Heye, T., Boll, D.T., Cyriac, J., Yang, S., Bach, M., Segeroth, M., 2023. TotalSegmentator: Robust segmentation of 104 anatomic structures in CT images. Radiology: Artificial Intelligence 5, e230024.

Wu, Y., Luo, X., Xu, Z., Guo, X., Ju, L., Ge, Z., Liao, W., Cai, J., 2024. Diversified and personalized multi-rater medical image segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11470–11479.

Xia, P., Fu, K.K., Wong, G.W., Akazawa, C., Verhey, L.J., 2000. Comparison of treatment plans involving intensity-modulated radiotherapy for nasopharyngeal carcinoma. International Journal of Radiation Oncology* Biology* Physics 48, 329–337.

Yan, K., Jin, D., Guo, D., Xu, M., Shen, N., Hua, X.S., Ye, X., Lu, L., 2023. Anatomy-aware lymph node detection in chest ct using implicit station stratification. arXiv preprint arXiv:2307.15271 .

Ye, X., Guo, D., Ge, J., Yan, S., Xin, Y., Song, Y., Yan, Y., Huang, B.s., Hung, T.M., Zhu, Z., et al., 2022. Comprehensive and clinically accurate head and neck cancer organs-at-risk delineation on a multi-institutional study. Nature Communications 13, 6137.

Ye, Y., Xie, Y., Zhang, J., Chen, Z., Xia, Y., Xia, Y., 2023. UniSeg: A prompt-driven universal segmentation model as well as a strong representation learner, in: MICCAI, p. 508—518.

Yushkevich, P.A., Piven, J., Hazlett, H.C., Smith, R.G., Ho, S., Gee, J.C., Gerig, G., 2006. User-guided 3d active contour segmentation of anatomical structures: significantly improved efficiency and reliability. Neuroimage 31, 1116–1128.

Zhu, W., Huang, Y., Zeng, L., Chen, X., Liu, Y., Qian, Z., Du, N., Fan, W., Xie, X., 2019. Anatomynet: deep learning for fast and fully automated whole-volume segmentation of head and neck anatomy. Medical physics 46, 576–589.