Radiotherapy and Oncology 180 (2023) 109480

Contents lists available at ScienceDirect

Radiotherapy and Oncology

journal homepage: www.thegreenjournal.com

Original Article

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

Xiangde Luo [a,c], Wenjun Liao [a,b,*], Yuan He [f], Fan Tang [g], Mengwan Wu [b], Yuanyuan Shen [b], Hui Huang [h], Tao Song [d], Kang Li [e], Shichuan Zhang [a,b], Shaoting Zhang [a,c], Guotai Wang [a,c,*]

[a] University of Electronic Science and Technology of China, Chengdu 611731, China; [b] Department of Radiation Oncology, Sichuan Cancer Hospital & Institute, Sichuan Cancer Center, School of Medicine, University of Electronic Science and Technology of China, Chengdu 610041, China; [c] Shanghai AI Laboratory, Shanghai 200030, China; [d] SenseTime Research, Shanghai 200233, China; [e] West China Biomedical Big Data Center, West China Hospital, Sichuan University, Chengdu 610041, China; [f] Department of Radiation Oncology, The First Affiliated Hospital of USTC, Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei, Anhui 23000, China; [g] Department of Radiation Oncology, Nanfang Hospital, Southern Medical University, Guangzhou 510515, China; and [h] Cancer center, Sichuan Provincial People's Hospital, University of Electronic Science and Technology of China, Chengdu 610072, China

ARTICLE INFO

Article history:
Received 23 August 2022
Received in revised form 7 January 2023
Accepted 8 January 2023
Available online 16 January 2023

Keywords:
Nasopharyngeal carcinoma
Automatic delineation
Primary gross tumor volume
Generalization
Deep learning

ABSTRACT

Background and purpose: The problem of obtaining accurate primary gross tumor volume (GTVp) segmentation for nasopharyngeal carcinoma (NPC) on heterogeneous magnetic resonance imaging (MRI) images with deep learning remains unsolved. Herein, we reported a new deep-learning method than can accurately delineate GTVp for NPC on multi-center MRI scans.

Material and methods: We collected 1057 patients with MRI images from five hospitals and randomly selected 600 patients from three hospitals to constitute a mixed training cohort for model development. The resting patients were used as internal (n = 259) and external (n = 198) testing cohorts for model evaluation. An augmentation-invariant strategy was proposed to delineate GTVp from multi-center MRI images, which encouraged networks to produce similar predictions for inputs with different augmentations to learn invariant anatomical structure features. The Dice similarity coefficient (DSC), 95 % Hausdorff distance (HD95), average surface distance (ASD), and relative absolute volume difference (RAVD) were used to measure segmentation performance.

Results: The model-generated predictions had a high overlap ratio with the ground truth. For the internal testing cohorts, the average DSC, HD95, ASD, and RAVD were 0.88, 4.99 mm, 1.03 mm, and 0.13, respectively. For external testing cohorts, the average DSC, HD95, ASD, and RAVD were 0.88, 3.97 mm, 0.97 mm, and 0.10, respectively. No significant differences were found in DSC, HD95, and ASD for patients with different T categories, MRI thickness, or in-plane spacings. Moreover, the proposed augmentation-invariant strategy outperformed the widely-used nnUNet, which uses conventional data augmentation approaches.

Conclusion: Our proposed method showed a highly accurate GTVp segmentation for NPC on multi-center MRI images, suggesting that it has the potential to act as a generalized delineation solution for heterogeneous MRI images.

 Radiotherapy and Oncology 180 (2023) 1–8

Intensity-modulated radiation therapy (IMRT) has become a preferred radiation technique in the treatment of nasopharyngeal carcinoma (NPC). It has improved the 5-year locoregional control rate and reduced radiation-related toxicities of patients with NPC [1–3]. Due to the distinctly dosimetric characteristics of IMRT, which contains sharp drops in dosages between tumor margins and normal tissues [4,5], delineation inaccuracies may compromise the survival of patients and lead to severe side effects. Hence, accurate delineation for primary gross tumor volume (GTVp) is important in the era of IMRT. Currently, the majority of GTVp delineation is performed manually. However, manual delineation for NPC is tiring, error-prone, and also subject to great inter-observer variability [6].

Benefitting from the advantages of feature learning, deep learning (DL)-based automatic delineation has provided a promising solution to solve the problems of manual delineation [7,8]. Recently, many sophisticated automatic segmentation models have been proposed to segment GTVp for NPC. Encouraging

* Corresponding authors.
E-mail addresses: lwjpsy@163.com (W. Liao), guotai.wang@uestc.edu.cn (G. Wang).

https://doi.org/10.1016/j.radonc.2023.109480
0167-8140/

segmentation results have been obtained using these models [6,9,10]. Due to the good contrast for soft tissues, magnetic resonance imaging (MRI) is the most common imaging modality for the diagnosis and treatment of NPC [11]. As a result, many automatic segmentation-based protocols and experiments were performed using MRI [12–14].

However, most existing MRI-based DL models have been developed based on single institutional data [6,9,10]. Specifically, these models were trained and tested using uniform MRI images. These images had strict inclusion and exclusion criteria, such as uniform thickness and scanning protocols [6,15]. It is unclear whether the method of segmentation used on these datasets would produce the same results using more heterogeneous data. As indicated in the study of Zhang et al, direct deployment of a trained MRI-based model from a single center to the unseen data from multi-centers led to an average Dice similarity coefficient (DSC) decrease of more than 10 % [16]. Therefore, it is necessary to construct a model with powerful generalizability so it can adapt to more heterogeneous data. This is exactly what the clinical practice needs.

The purpose of this study is to construct a generalizable DL model for the robust delineation of GTVp for NPC on multi-center heterogeneous MRI. To achieve this goal, we first proposed an augmentation-invariant training strategy to encourage DL models to pay more attention to invariant anatomical structure features rather than intensity distribution to boost the generalizability of DL models. Afterwards, we applied this method to train the DL model on a mixed cohort from three hospitals. Then we evaluated the trained model on three seen and two unseen MRI cohorts. Finally, we investigated the clinical applicability of the proposed framework.

Material and methods

Data

Patients from five tertiary hospitals were retrospectively collected. The inclusion criteria were as follows: (i) Patients who were histologically confirmed as NPC; (ii) Patients who underwent MRI examinations for nasopharynx and neck before anticancer treatment; (iii) The MRI sets included the contrast-enhanced T1-weighted sequence. The patients were excluded if their images had a low resolution that affected GTVp delineation. Finally, a total of 1057 patients were enrolled. We had 367 cases from Sothern Medical University (SMU), 284 cases from West China Hospital (WCH), 208 cases from Sichuan Provincial People's Hospital (SPH), 146 cases from Anhui Provincial Hospital (APH), and 52 cases from Sichuan Cancer Hospital (SCH). Similar to previous works [15,17], we used the contrast-enhanced T1-weighted sequence images as network input for network training and testing.

Afterward, 256, 198, and 146 patients were randomly selected from SMU, WCH, and SPH, respectively. This was done with a training-to-testing ratio of 7:3, and these patients constituted a mixed training cohort (n = 600). The resting 111, 86, and 62 patients from SMU, WCH, and SPH, respectively, made up the three internal testing cohorts (Table 1). Additionally, all patients (n = 146) from APH and all patients (n = 52) from SCH were used as the two external testing cohorts (Table 1). The flow chart is illustrated in Fig. 1.

Table 1 summarizes the main clinical characteristics of the mixed training, the internal, and the external testing cohorts. The characteristics included: sex, age, T category, and primary tumor size. MRI acquisition parameters of these cohorts are presented in Table 1. These parameters included: MRI scanners, the magnetic field, the echo time, the repetition time, the field of views, the flip angle, slice thickness, and in-plane spacing. Additionally, baseline characteristics and MRI acquisition parameters of the separated training cohorts are shown in Supplementary Table 1.

This study was approved by the Ethics Committee on Biomedical Research of these hospitals, and the informed consent was waived (Number 1085). All patients were restaged according to the eighth edition of the American Joint Committee on Cancer [18].

Ground truth GTVp delineation

The protocol for GTVp contouring was consistent with many previous studies [6,15]. Two oncologists from WCH with over ten years of experience in the treatment of NPC were invited to delineate GTVp for all patients coming from the five hospitals. Another oncologist with over 20 years of experience was consulted in case of disagreement. The criteria used to determine GTVp delineation were taken from the International Commission on Radiation Units and Measurements (ICRU) report 83 [19].

Augmentation-invariant framework for GTVp segmentation.

In this work, an augmentation-invariant framework was introduced to train DL models to generate accurate predictions of GTVp from heterogeneous MRI images (Fig. 2). Specifically, for the same images, two different intensity transformation strategies were employed for data augmentation, and two augmented images were obtained. Then the two augmented images were sent to a DL model to produce two predictions. Afterward, a loss function was used to encourage the two predictions to be similar, which could impose the DL model to focus more on invariant features (anatomical structures) rather than appearance differences (intensity distributions). The proposed framework was implemented by extending the widely-used and powerful nnUNet [20]. Detailed image processing, network training, and testing were described in the Supplementary material and methods.

Experimental setting

To investigate the effectiveness of our proposed framework, we performed comparison experiments and result analysis based on the 1057 patients. Firstly, we compared the proposed framework with three extensions of nnUNet to demonstrate the effectiveness of the proposed framework. The three extensions were nnUNet without any data augmentation methods (nnUNet wo-DA), nnUNet with default data augmentation methods (nnUNet w-DDA), and nnUNet with all the proposed framework used data augmentation methods (nnUNet w-ADA). After that, we analyzed the performance of the proposed framework on the internal and external cohorts to investigate its generalization for the heterogeneous MRI images. Then, we investigated the impact of the different tumor stages and imaging resolutions to further evaluate the strength of the proposed framework.

Qualitative analysis was also an important part of our research. Similar to a previous work [21], two senior radiation oncologists with 15 years of experience (not involved in ground truth GTVp delineation) were invited to subjectively evaluate each automatic contouring result together. We used 4-grade criteria to estimate the potential curative effect. These criteria included: (i) no revision (the prediction can be directly used without any revision), (ii) minor revision (the prediction needs a few refinements to be clinically acceptable), (iii) major revision (the prediction needs to be substantially revised), and (iv) re-delineation (the prediction is unacceptable and needs to be delineated manually).

Evaluation metrics

Following several previous studies [22–24], we employed four widely-used metrics to measure the DL-model-generated segmen-

Table 1
Baseline characteristics and MRI parameters in the training and testing cohorts.

Variable	Mixed training cohort (n = 600)	Internal testing cohorts (n = 259)			External testing cohorts (n = 198)	
	SMU + WCH + SPH (n = 600, %)	SMU (n = 111, %)	WCH (n = 86, %)	SPH (n = 62, %)	APH (n = 146, %)	SCH (n = 52, %)
Sex						
Male	411 (68.5)	84 (75.7)	60 (69.8)	47 (75.8)	107 (73.3)	37 (71.2)
Female	189 (31.5)	27 (24.3)	26 (30.2)	15 (24.2)	39 (26.7)	15 (28.8)
Age (median (range))	48(12–80)	46 (19–77)	49 (18–79)	49 (28–69)	50 (17–79)	46 (26–74)
T category						
T1	60 (10.0)	10 (9.0)	11 (12.8)	4 (6.5)	18 (12.3)	11 (21.1)
T2	135 (22.5)	26 (23.4)	22 (25.6)	11 (17.7)	27 (18.5)	8 (15.4)
T3	286 (47.7)	51 (45.9)	35 (40.7)	32 (51.6)	72 (49.3)	28 (53.8)
T4	119 (19.8)	24 (21.7)	18 (20.9)	15 (24.2)	29 (19.9)	5 (9.7)
Tumor size* (cm^3)	31.9 (2.4–249.1)	36.1 (8.9–163.7)	24.2 (1.3–147.9)	32.9 (7.8–179.5)	24.1 (9.7–187.9)	23.6 (8.5–105.6)
Vendor						
GE	432 (72.0)	100 (90.1)	86 (100.0)	0 (0.0)	113 (77.4)	0 (0.0)
Siemen	162 (27.0)	7 (6.3)	0 (0.0)	62 (100.0)	27 (18.5)	52 (100.0)
Philips	6 (1.0)	4 (3.6)	0 (0.0)	0 (0.0)	6 (4.1)	0 (0.0)
Magnetic field (T)						
1.5	239 (39.8)	106 (95.5)	0 (0.0)	1 (1.6)	119 (81.5)	23 (44.2)
3.0	361 (60.2)	5 (4.5)	86 (100.0)	61 (98.4)	27 (18.5)	29 (55.8)
Echo time* (ms)	9.58 (2.05–19.0)	2.66 (2.03–18.0)	9.47 (3.79–10.39)	9.9 (7.8–9.9)	14.9 (10.0–16.0)	11 (2.03–190.0)
Repetition time* (ms)	692 (9.54–2331)	455 (90–2334)	721 (9.54–931)	788 (650–2190)	610 (309–1223)	756 (4.39–860)
FOV	100 (75–115)	100 (75–127)	100 (95–100)	100 (78.0–100)	90 (87.0–109.0)	85 (80–100)
Flip Angle	90 (12–180)	80 (15–180)	111 (12–111)	120 (90–120)	90 (90–160)	150 (9–160)
Thickness (mm)						
1–3	199 (33.2)	1 (0.9)	84 (97.6)	1 (1.6)	108 (74.0)	17 (32.7)
> 3–5	21 (3.5)	2 (1.8)	1 (1.2)	3 (4.8)	38 (26.0)	35 (67.3)
> 5–8	380 (63.3)	108 (97.3)	1 (1.2)	58 (93.6)	0 (0.0)	0 (0.0)
In-plane spacing (mm)						
0.3–0.5	563 (93.8)	104 (93.7)	86 (100.0)	56 (90.3)	106 (72.9)	10 (19.2)
> 0.5–0.8	37 (6.2)	7 (6.3)	0 (0.0)	6 (9.7)	40 (27.4)	42 (80.8)

Abbreviations.
ms = millisecond.
mm = millimeter.
FOV = field of view.
*Data were denoted as median with range. The mixed training cohort included patients from SMU (n = 256), WCH (n = 198), and SPH (n = 146).

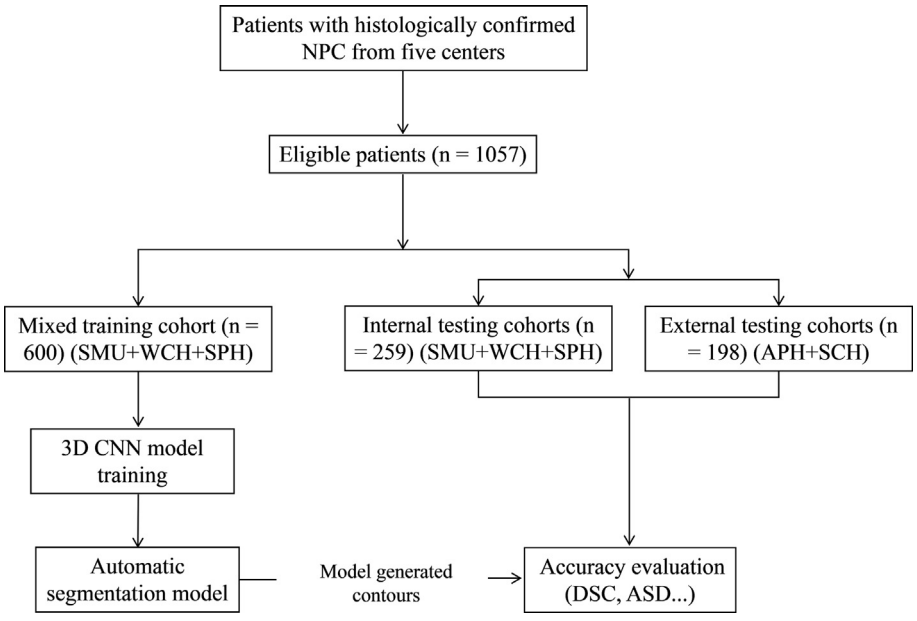

Fig. 1. The flow chart of this study. CNN = Conventional Neural Network.

tation of GTVp. The four metrics were the DSC, the average surface distance (ASD), the 95 % Hausdorff distance (HD95), and the relative absolute volume difference (RAVD). DSC and RAVD are pixel-wise metrics that measure the region overlap between the network's predictions and the ground truth. HD95 and ASD are distance-based metrics that aim to measure the boundary distance between the predictions and the ground truth. All these metrics were calculated by a public package (https://github.com/loli/medpy).

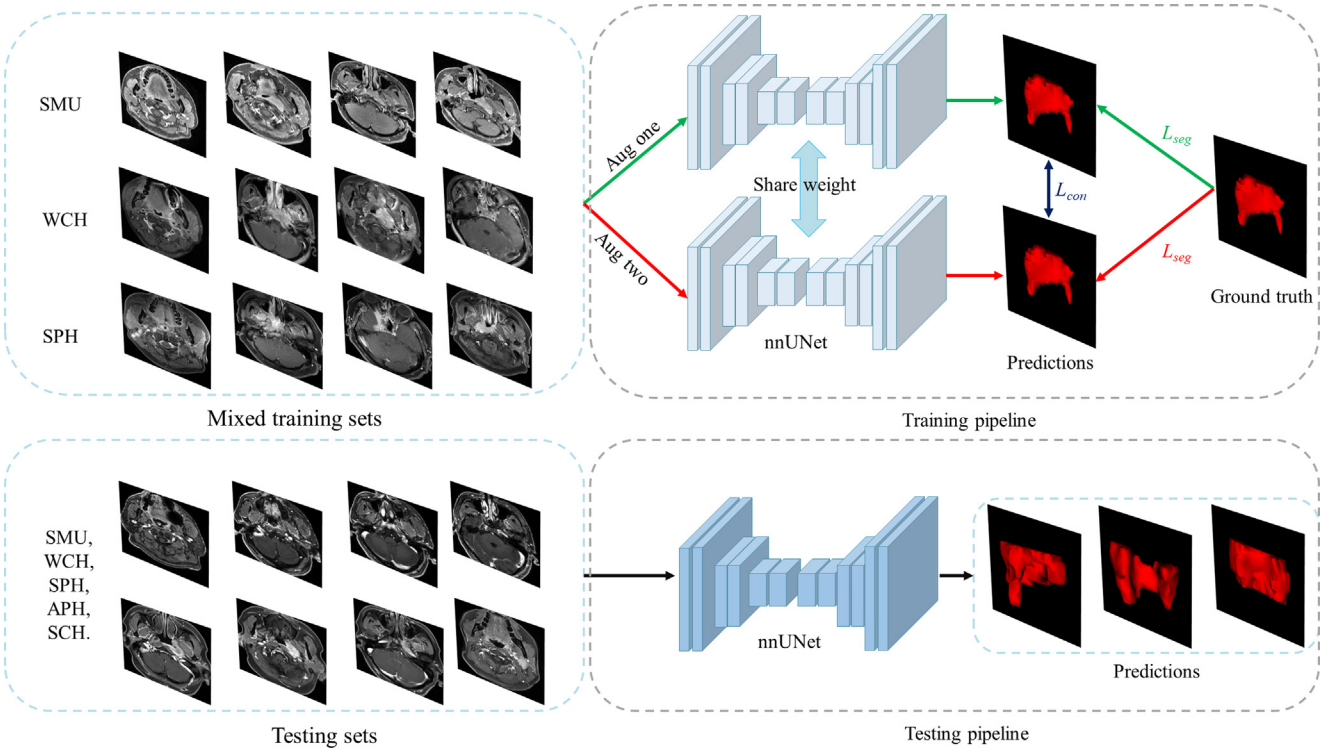

Fig. 2. Overview of the augmentation-invariant framework for accurate and robust NPC delineation. Aug = data augmentation; L_{seg} = the loss function of segmentation; L_{con} = the proposed augmentation-invariant loss function.

Statistical analysis

Statistical analysis was performed with the SPSS software package (Version 22.0, IBM SPSS Inc). Numeric variables were denoted as mean ± standard deviation. The Friedman test was used for multiple comparisons. This was followed by post hoc tests for pairwise comparisons. In these comparisons, the Bonferroni method was utilized to adjust the significance level. The correlation between DSC, HD95, and ASD and the clinical parameters was examined with the Pearson correlation coefficient. A two-tailed p-value < 0.05 was considered significant.

Results

The quantitative performance of our proposed DL model covering both the internal and external cohorts is summarized in Table 2. For the internal testing cohorts, the average DSC, RAVD, HD95, and

Table 2
Accuracy comparison for model-generated GTVp in the testing cohorts by different methods.

Variables	nnUNet without any data augmentation		nnUNet with default data augmentation		nnUNet with all the used augmentation transformations		nnUNet with our proposed augmentation-invariant strategy	
Volume-based metrics	DSC	RAVD	DSC	RAVD	DSC	RAVD	DSC	RAVD
Internal testing cohorts (n = 259)								
SMU (n = 111)	0.87 ± 0.06***	0.14 ± 0.11**	0.87 ± 0.06***	0.12 ± 0.11	0.87 ± 0.05***	0.12 ± 0.09	**0.89 ± 0.06**	**0.11 ± 0.11**
WCH (n = 86)	0.85 ± 0.08**	0.18 ± 0.44	0.86 ± 0.10**	0.14 ± 0.35	0.87 ± 0.05**	0.15 ± 0.34	**0.88 ± 0.10**	0.16 ± 0.53
SPH (n = 62)	0.83 ± 0.09	0.18 ± 0.17	0.85 ± 0.07	0.14 ± 0.14	0.85 ± 0.08	0.17 ± 0.16	0.86 ± 0.07	0.15 ± 0.17
Whole (n = 259)	0.85 ± 0.08***	0.16 ± 0.27***	0.86 ± 0.08***	0.13 ± 0.23	0.86 ± 0.06***	0.14 ± 0.22	**0.88 ± 0.08**	**0.13 ± 0.33**
External testing cohorts (n = 198)								
APH (n = 146)	0.84 ± 0.08***	0.14 ± 0.15***	0.87 ± 0.05***	0.11 ± 0.12*	0.87 ± 0.06***	0.11 ± 0.13*	**0.89 ± 0.05**	**0.08 ± 0.10**
SCH (n = 52)	0.83 ± 0.07*	0.16 ± 0.14	0.83 ± 0.07*	0.15 ± 0.16	0.84 ± 0.08	0.15 ± 0.18	**0.86 ± 0.07**	0.13 ± 0.21
Whole (n = 198)	0.84 ± 0.08***	0.15 ± 0.14***	0.86 ± 0.06***	0.12 ± 0.13*	0.86 ± 0.06***	0.15 ± 0.14*	**0.88 ± 0.06**	**0.10 ± 0.14**
Distance-based metrics	HD95 (mm)	ASD (mm)	HD95 (mm)	ASD (mm)	HD95 (mm)	ASD (mm)	HD95 (mm)	ASD (mm)
Internal testing cohorts (n = 259)								
SMU (n = 111)	5.76 ± 4.67	1.21 ± 1.29***	5.50 ± 3.59	1.07 ± 1.13**	5.76 ± 6.30	1.03 ± 1.23*	5.62 ± 13.96	**1.00 ± 2.69**
WCH (n = 86)	4.98 ± 3.32**	1.36 ± 1.19*	4.58 ± 3.77	1.35 ± 1.59*	4.71 ± 3.30	1.42 ± 1.49*	**4.18 ± 3.86**	**1.13 ± 1.45**
SPH (n = 62)	6.90 ± 4.28***	1.25 ± 0.80***	5.93 ± 7.31	1.24 ± 1.45	6.45 ± 5.09*	1.23 ± 0.93**	**4.99 ± 3.18**	**0.92 ± 0.72**
Whole (n = 259)	5.78 ± 4.21***	1.27 ± 1.15***	5.30 ± 4.80**	1.20 ± 1.38***	5.58 ± 5.20***	1.20 ± 1.27***	**4.99 ± 9.53**	**1.03 ± 1.97**
External testing cohorts (n = 198)								
APH (n = 146)	5.99 ± 5.36***	1.81 ± 1.53***	5.56 ± 6.03***	1.60 ± 1.40***	5.85 ± 4.87***	1.69 ± 1.33***	**3.75 ± 2.61**	**0.08 ± 0.10**
SCH (n = 52)	7.80 ± 4.86**	1.98 ± 1.77***	7.35 ± 6.26	1.84 ± 1.58***	6.41 ± 4.75	1.77 ± 1.18***	**4.58 ± 2.84**	**0.91 ± 0.92**
Whole (n = 198)	6.47 ± 5.28***	1.86 ± 1.60***	6.03 ± 6.13***	1.66 ± 1.45***	6.00 ± 4.83***	1.71 ± 1.29***	**3.97 ± 2.69**	**0.97 ± 0.85**

Data were denoted as mean ± standard deviation. The bold font shows significant improvement when compared our method with others by multiple comparisons followed by post hoc tests. Adjust p-values were obtained by Bonferroni method. *, **, *** mean adjusted p < 0.05, <0.01, <0.001, respectively.

ASD were 0.88, 0.13, 4.99 mm, and 1.03 mm, respectively. For external testing cohorts, the average DSC, RAVD, HD95, and ASD were 0.88, 0.10, 3.97 mm, and 0.97 mm, respectively. No significant difference was found between the internal testing and external testing cohorts when it came to these metrics (all p-values > 0.05) (Fig.S1).

Moreover, we analyzed the false positive and the false negative Dice (FPD and FND) to quantify the potential for miss- or over-treatment [25]. The average FPD and FND for internal testing cohorts were 0.14 and 0.11, respectively (Supplementary Table 2). The average FPD and FND for the external testing cohorts were 0.12 and 0.11, respectively. There was no significant difference between the internal and external testing cohorts with regard to FPD (p = 0.223) and FND (p = 0.340) (Fig.S1). A visualization of the model-generated contours is illustrated in Fig. 3. This was done according to the best, the mean, the median, and the worst DSC measures of the external testing cohorts.

The proposed augmentation-invariant framework was compared with three extensions of the nnUNet with different data augmentation strategies for both the internal and the external testing cohorts (Table 2). For the internal testing cohorts, the average DSC for nnUNet wo-DA, nnUNet w-DDA, nnUNet w-ADA, and ours was 0.85, 0.86, 0.86, and 0.88, and the average RAVD for each measured out to 0.16, 0.13, 0.14, and 0.13, respectively (Table 2). By multiple comparisons, our proposed framework achieved a significant improvement of DSC for the SMU, the WCH, and the whole internal testing cohorts (all p - values < 0.05). Our framework also achieved a significant reduction of RAVD for the SMU and the whole internal testing cohorts (all p - values < 0.05). For all external testing cohorts, the average DSC for nnUNet wo-DA, nnUNet w-DDA, nnUNet w-AAT, and ours was 0.84, 0.86, 0.86, and 0.88, and the average RAVD for each of these was 0.15, 0.12, 0.15, and 0.10, respectively (Table 2). The average DSC produced by our method was significantly higher than that produced by other methods (all p - values < 0.05). And the averages of RAVD were significantly lower (all p - values < 0.05) in the APH and the whole external testing cohorts.

Among the whole internal testing cohorts, the average HD95 for nnUNet wo-DA, nnUNet w-DDA, nnUNet w-ADA, and ours was 5.78 mm, 5.30 mm, 5.58 mm, and 4.99 mm, respectively (Table 2). For the whole external testing cohorts, the averages were 6.47 mm, 6.03 mm, 6.00 mm, and 3.97 mm, respectively. Our framework obtained a significant reduction of HD95 for all external testing cohorts (all p - values < 0.05) and all internal testing cohorts (all p - values < 0.05) with the exception of the SMU. For the whole internal testing cohorts, the average ASD for nnUNet wo-DA, nnUNet w-DDA, nnUNet w-ADA, and ours was 1.27 mm, 1.20 mm, 1.20 mm, and 1.03 mm, respectively. Among the whole external testing cohorts, the averages were 1.86 mm, 1.66 mm, 1.71 mm, and 0.97 mm, respectively. The average ASD produced by our method was significantly lower than that produced by other methods. This was true not only for both internal and external testing cohorts but the whole internal and external testing cohorts as well (all p - values < 0.05).

To further examine the accuracy and robustness of our model, quantitative comparison was conducted of patients with different T categories in these testing cohorts. We found that there was no significant difference among patients with various T categories for DSC (p = 0.057), HD95 (p = 0.097), or ASD (p = 0.125) (Fig. 4a to 4c). The average DSC and HD95 for T1, T2, T3, and T4 were

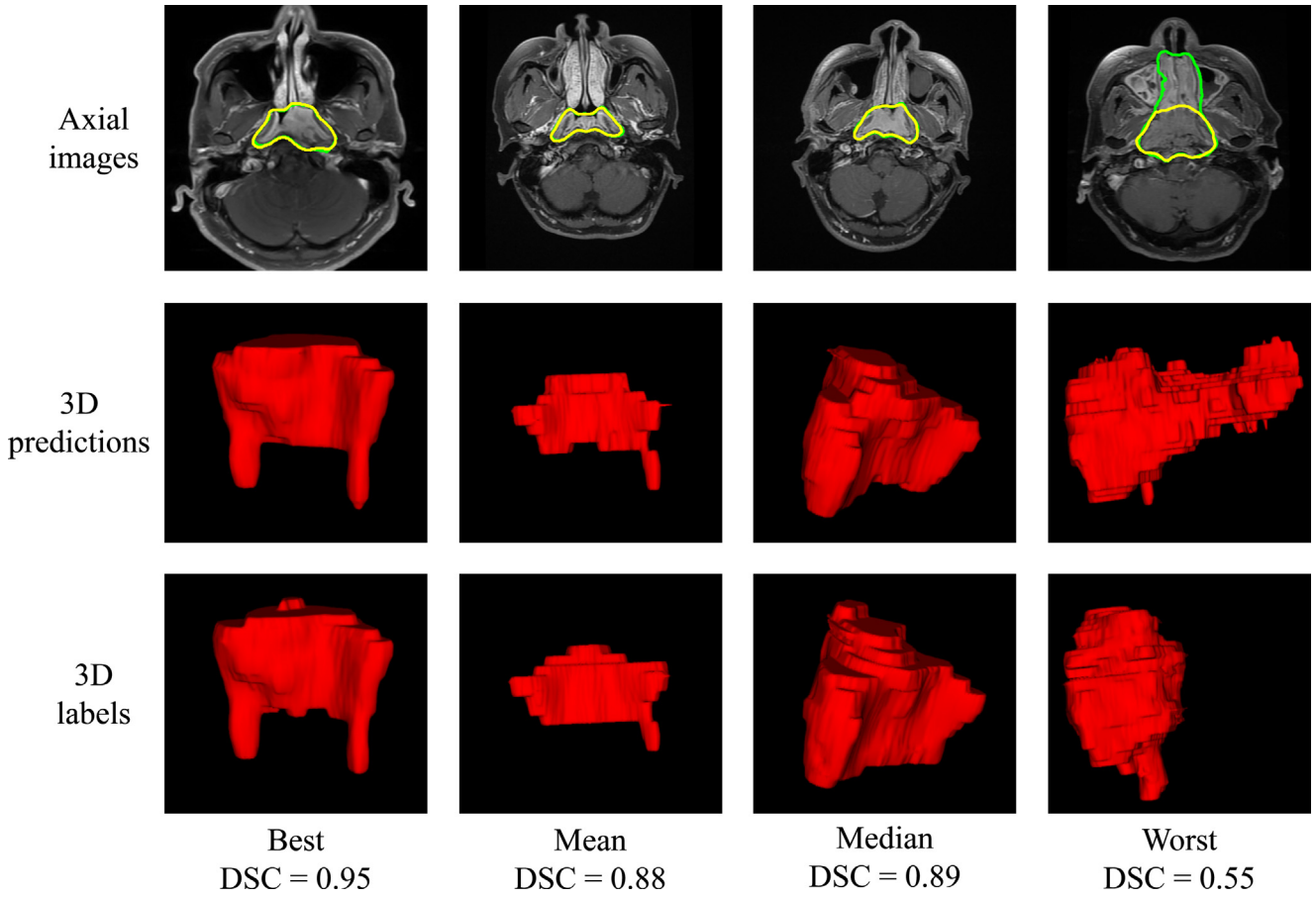

Fig. 3. Visual examples of the model-generated contours in external testing cohorts. We ranked all patients in external testing cohorts and selected the best/mean/median/worst scores of patients in term of DSC. Yellow lines denoted the experts' delineated contours, and green lines denoted the model-predicted contours.

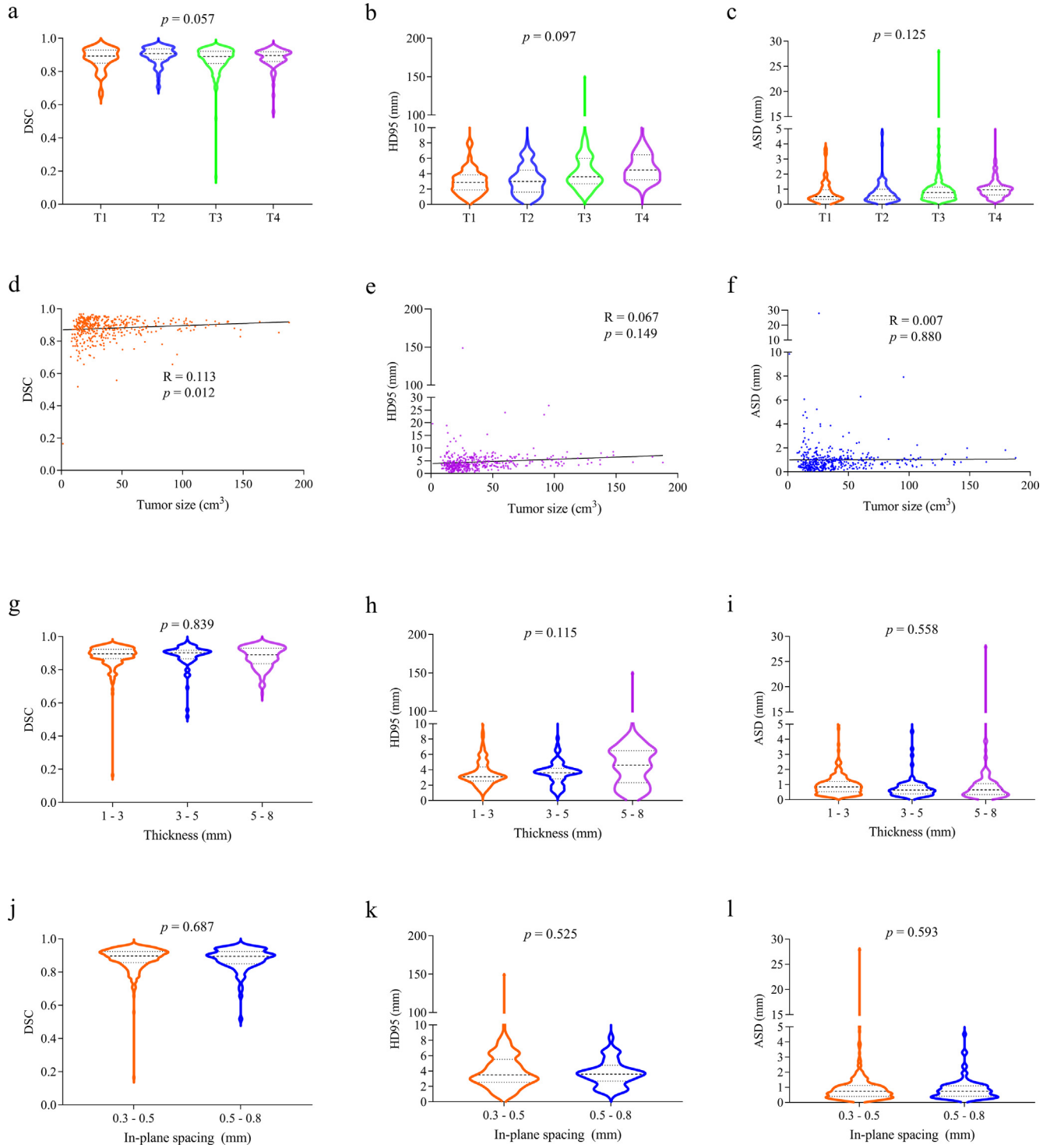

Fig. 4. Accuracy of the model in subgroup patients. (a-c), Accuracy of the model in patients with different T categories. (d-f), The relationship between the accuracy of the model and the primary tumor sizes. (g-i), Accuracy of the model in patients with different thickness. (j-l) Accuracy of the model in patients with different in-plane spacing.

0.88, 0.90, 0.87, and 0.88, respectively, and 3.30 mm, 3.33 mm, 5.00 mm, and 5.47 mm, respectively. The average ASD for T1, T2, T3, and T4 was 0.74 mm, 0.75 mm, 1.12 mm, and 1.14 mm, respectively. When considering primary tumor volumes, we found that the DSC increased with increasing primary tumor sizes, showing a significant positive correlation (R = 0.113, p = 0.012) (Fig. 4d). However, no significant correlation was observed between primary tumor sizes and the HD95 (p = 0.149) or the ASD (p = 0.880) (Fig. 4e and 4f). Four visual examples of model-generated contours (made according to T categories) are illustrated in Fig.S2. Several model-generated contours for patients with difficulty in GTVp delineation are presented in Fig.S3.

Since several key MRI parameters had the potential to influence our results, we studied each of them carefully to determine whether or not they had affected the model's performance. The slice thickness of the T1-weighted sequence was calculated for each patient and separated into three subgroups as shown in Table 1. We found that there was no statistical difference in MRI

images with different slice thicknesses for DSC (p = 0.839), HD95 (p = 0.115), and ASD (p = 0.558) (Fig. 4g to 4i). The average DSC for 1–3 mm thickness, > 3–5 mm thickness, and > 5–8 mm thickness was 0.88, 0.88, and 0.88. The average HD95 for these same thicknesses was 4.00 mm, 4.01 mm, and 5.50 mm, respectively. The average ASD for these thicknesses was 1.06 mm, 0.84 mm, and 1.00 mm, respectively.

In a similar manner, the in-plane spacing of the T1-weighted sequence was calculated for each patient and divided into two subgroups, as indicated in Table 1. The average DSC, HD95, and ASD were compared for these two subgroups, and no significant differences were found (all p - values > 0.5) (Fig. 4j to 4 l). The average DSC for the 0.3–0.5 mm in-plane spacing and the > 0.5–0.8 mm in-plane spacing was 0.88 and 0.88, respectively. The average HD95 for these same in-plane spacings was 4.67 mm and 4.12 mm, respectively. The mean ASD for these in-plane spacings was 1.02 mm and 0.93 mm, respectively.

Two experts together subjectively evaluated the 457 GTVp predictions from the internal and external testing cohorts (Supplementary Table 3). For the internal and external cohorts respectively, approximately 24 % and 15 % of the proposed framework-generated predictions can be seen as clinically acceptable (no revision). For the internal cohort, 51.74 % of predictions were evaluated as "minor revision", and 20.85 % were evaluated as "major revision". 3.47 % of predictions were seen as re-delineation. For the external cohort, 50.51 % of predictions were evaluated as "minor revision", and 32.32 % were evaluated as "major revision". 2.02 % of predictions were seen as re-delineation.

Finally, a comparison was performed between our segmentation results and previously published results produced by DL models regarding the delineation of GTVp for NPC. Although there are some differences in datasets and experimental settings, the results demonstrated that our framework is reasonable and comparable. The mean DSC for all patients in the testing cohorts produced by our model was higher than that produced by any other DL model used in these studies, with the exception of Li et al., 2018 (Supplementary Table 4).

Discussion

In this multi-institutional study, we developed a new DL model based on an intensity augmentation-invariant framework to segment GTVp for NPC. This framework was comprehensively evaluated by using three seen internal testing cohorts and two unseen external testing cohorts. The results showed that our model performed well for the internal testing cohorts and generalized well for the external testing cohorts. Afterward, we extensively examined the segmentation performance of the model for various subgroups of patients, demonstrating that the model could acquire uniform and high-accuracy GTVp delineation for patients with different T categories and image resolutions. Moreover, we found that the vast majority of our model-predicted contours were clinically acceptable after some refinements.

Previous works have shown the great potential for the clinical applicability of DL models that delineate GTVp for NPC and other solid tumors [6,26,27]. For instance, a DL-based on semi-supervised learning was proposed by us to segment GTVp and metastatic lymph nodes (GTVnd). This had an average DSC of 0.81 and 0.76 for the GTVp and GTVnd, respectively [17]. Similarly, in the study of Li et al [6], the proposed DL tool achieved comparable accuracy for GTVp delineation with expert-generated ground truth contours. It also considerably reduced inter-user variations on MRI images. However, there are very few studies that focus on the problems of generalizability when it comes to GTVp segmentation on multi-center heterogeneous MRI images.

Compared with the other three methods (nnUNet and its extensions), our method achieved significant improvement in all metrics among the majority of internal and external cohorts. It is suggested that our framework is a better method to segment GTVp from multi-center heterogeneous MRI images. We also observed that in the other methods the performance gaps between internal cohorts and external cohorts were not very significant in terms of DSC and RAVD. But the proposed framework alleviated the performance gap between internal and external testing data in terms of HD95 and ASD (Table 2). The potential reason may be that this study trained these DL models on mixed multi-center datasets rather than single-center datasets. This may have boosted the data diversity and alleviated the domain gap. In addition, DSC may not be as efficient at showing the domain gap. After all, it just considered the region-level overlap and ignored anatomical structure similarity and clinical practice. Although there appear to be no giant generalization gaps between the proposed framework and the nnUNet or its extensions, the proposed framework still improved the segmentation results for both internal and external cohorts significantly. This shows it could help reduce the GTVp delineation burden.

In this study, we first demonstrate that our model has a similar performance for both internal and external testing cohorts. There were no significant differences in any metric, which shows the powerful generalizability of our model. Then, subgroup analysis in all testing cohorts was conducted. At first, we found that T categories had little influence on the accuracy of the model. There was comparable DSC, HD95, and ASD for these patients. Since patients with more advanced T categories often had increased tumor volumes, a correlation between primary tumor sizes and DSC, HD95, and ASD was analyzed. It was indicated that HD95 and ASD were not impacted by primary tumor sizes. And even though the DSC was positively associated with primary tumor sizes, the R-value was quite small. From a statistical point of view, it could not be considered clinically significant. Taken together, these results indicated that the model performed well regardless of T categories or tumor volumes.

Similar to a previous study [6], two crucial parameters for image quality were selected to investigate the impact of image characteristics on the model's accuracy. Our results showed that no significant differences were found in DSC, HD95, and ASD for patients with different MRI thickness or in-plane spacings, suggesting that our model is robust to those differences. It may benefit from the way we trained the segmentation network, where we simulated the image with different thicknesses and in-plane spacing via spatial transformation.

In clinical evaluation, our results showed that although the DL model can achieve good performance in some evaluation metrics (DSC, HD95...), it is still hard to apply it directly with the clinical flow in mind. The ratio of no revision was only 23.94 % and 15.15 % for the internal and external cohorts, respectively. However, more than 50 % of predictions only need a minor revision. So, the model could still play an essential role in DL-assisted GTVp delineation to reduce oncologists' delineation burden and save time. These observations demonstrated that the proposed framework might be the potential solution for accurate and generalizable delineation of GTVp for NPC from multiple hospitals' MRI images.

This study had several limitations. First, although standard protocols for GTVp delineation were established, the ground truth contours were generated manually and could suffer from subjective variations or errors. However, similar to most previous studies, this is both the most commonly used and most reliable method currently [6,9,22]. Second, this work mainly focused on the domain generalization problem, which requires the collection and annotation of large-scale images from many hospitals for network train-

ing. Collecting and annotating a large-scale dataset is very expensive, time-consuming, and limited by privacy protection. It is more desirable to develop a method that just requires a small dataset from one hospital for training and can still apply to test for multiple hospitals directly. In the future, we will extend this work to be more generalizable and employ small-scale images from a single hospital dataset for network training. Third, although we accurately delineated GTVp for NPC on multi-center MRI images, the current terminal treatment planning method is still based on simulation CT. Hence, similar to other studies [6,15], we are unsure whether the model will perform in a similar manner using CT-based data. Much work is still needed.

Conclusion

In summary, the augmentation-invariant framework could boost the generalization and robustness of the DL model. Using the proposed framework and a mixed training set for network training produced more accurate segmentation results of GTVp for both the internal and the external testing cohorts. These results show that the proposed framework is a potential solution for accurate and generalizable GTVp delineation of NPC from multiple hospitals' MRI images.

Funding statement

This work was supported by the National Natural Science Foundation of China under Grant 81771921, Grant 61901084, and Grant 82203197.

Data statement

The code of the proposed method is available: https://github.com/Luoxd1996/RobustNPC. Other data generated and analyzed during this study can be obtained by contacting the corresponding author with reasonable requirements.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.radonc.2023.109480.

References

[1] Sun XS, Liu SL, Luo MJ, et al. 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. Int J Radiat Oncol Biol Phys 2019;105:581–90. https://doi.org/10.1016/j.ijrobp.2019.06.2549.
[2] Lee AW, Ma BB, Ng WT, Chan AT. Management of nasopharyngeal carcinoma: current practice and future perspective. J Clin Oncol 2015;33:3356–64. https://doi.org/10.1200/jco.2015.60.9347.
[3] Chen YP, Chan ATC, Le QT, Blanchard P, Sun Y, Ma J. Nasopharyngeal carcinoma. Lancet 2019;394:64–80. https://doi.org/10.1016/s0140-6736(19)30956-0.
[4] Xia P, Fu KK, Wong GW, Akazawa C, Verhey LJ. Comparison of treatment plans involving intensity-modulated radiotherapy for nasopharyngeal carcinoma. Int J Radiat Oncol Biol Phys 2000;48:329–37. https://doi.org/10.1016/s0360-3016(00)00585-x.
[5] Kam MK, Chau RM, Suen J, Choi PH, Teo PM. Intensity-modulated radiotherapy in nasopharyngeal carcinoma: dosimetric advantage over conventional plans and feasibility of dose escalation. Int J Radiat Oncol Biol Phys 2003;56:145–57. https://doi.org/10.1016/s0360-3016(03)00075-0.
[6] Lin L, Dou Q, Jin YM, et al. Deep learning for automated contouring of primary tumor volumes by MRI for nasopharyngeal carcinoma. Radiology 2019;291:677–86. https://doi.org/10.1148/radiol.2019182012.
[7] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436–44. https://doi.org/10.1038/nature14539.
[8] Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. Med Image Anal 2017;42:60–88. https://doi.org/10.1016/j.media.2017.07.005.
[9] Li Y, Dan T, Li H, et al. NPCNet: jointly segment primary nasopharyngeal carcinoma tumors and metastatic lymph nodes in MR images. IEEE Trans Med Imaging 2022. https://doi.org/10.1109/tmi.2022.3144274.
[10] Li Y, Peng H, Dan T, Hu Y, Tao G, Cai H. Coarse-to-fine Nasopharyngeal carcinoma Segmentation in MRI via Multi-stage Rendering. IEEE International Conference on Bioinformatics and Biomedicine (BIBM)2020: IEEE, 2020:623-628. https://doi.org/10.1109/BIBM49941.2020.9313574.
[11] King AD. MR imaging of nasopharyngeal carcinoma. Magn Reson Imaging Clin N Am 2022;30:19–33. https://doi.org/10.1016/j.mric.2021.06.015.
[12] Ma Z, Wu X, Song Q, Luo Y, Wang Y, Zhou J. Automated nasopharyngeal carcinoma segmentation in magnetic resonance images by combination of convolutional neural networks and graph cut. Exp Ther Med 2018;16:2511–21. https://doi.org/10.3892/etm.2018.6478.
[13] Ke L, Deng Y, Xia W, et al. Development of a self-constrained 3D DenseNet model in automatic detection and segmentation of nasopharyngeal carcinoma using magnetic resonance images. Oral Oncol 2020;110:. https://doi.org/10.1016/j.oraloncology.2020.104862104862.
[14] Huang W, Chan KL, Zhou J. Region-based nasopharyngeal carcinoma lesion segmentation from MRI using clustering-and classification-based methods with learning. J Digit Imaging 2013;26:472–82. https://doi.org/10.1007/s10278-012-9520-4.
[15] Liao W, He J, Luo X, et al. Automatic delineation of gross tumor volume based on magnetic resonance imaging by performing a novel semisupervised learning framework in nasopharyngeal carcinoma. Int J Radiat Oncol Biol Phys 2022;113:893–902. https://doi.org/10.1016/j.ijrobp.2022.03.031.
[16] Zhang L, Wang X, Yang D, et al. Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation. IEEE Trans Med Imaging 2020;39:2531–40. https://doi.org/10.1109/tmi.2020.2973595.
[17] Luo X, Liao W, Chen J, et al. Efficient Semi-Supervised Gross Target Volume of Nasopharyngeal Carcinoma Segmentation via Uncertainty Rectified Pyramid Consistency. Medical Image Computing and Computer Assisted Intervension (MICCAI) 2021:318-329. https://doi.org/10.1007/978-3-030-87196-3_30.
[18] Pan JJ, Ng WT, Zong JF, et al. Proposal for the 8th edition of the AJCC/UICC staging system for nasopharyngeal cancer in the era of intensity-modulated radiotherapy. Cancer 2016;122:546-558. https://doi.org/10.1002/cncr.29795.
[19] Hodapp N. The ICRU Report 83: prescribing, recording and reporting photon-beam intensity-modulated radiation therapy (IMRT). Strahlenther Onkol 2012;188:97–9. https://doi.org/10.1007/s00066-011-0015-x.
[20] Isensee F, Jaeger PF, Kohl S, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods 2021;18:203–11. https://doi.org/10.1038/s41592-020-01008-z.
[21] Liu Z, Liu X, Guan H, et al. Development and validation of a deep learning algorithm for auto-delineation of clinical target volume and organs at risk in cervical cancer radiotherapy. Radiother Oncol 2020;153:172–9. https://doi.org/10.1016/j.radonc.2020.09.060.
[22] Tang H, Chen X, Liu Y, Lu Z, Xie XJ. Clinically applicable deep learning framework for organs at risk delineation in CT images. Nat Mach Intell 2019;1:480–91. https://doi.org/10.1038/s42256-019-0099-z.
[23] Wang G, Liu X, Li C, et al. A noise-robust framework for automatic segmentation of COVID-19 pneumonia lesions from CT images. IEEE Trans Med Imaging 2020;39:2653–63. https://doi.org/10.1109/tmi.2020.3000314.
[24] Vrtovec T, Močnik D, Strojan P, Pernuš F, Ibragimov B. Auto-segmentation of organs at risk for head and neck radiotherapy planning: From atlas-based to deep learning methods. Med Phys 2020;47:e929-e950. https://doi.org/10.1002/mp.14320.
[25] Cardenas CE, Beadle BM, Garden AS, et al. Generating high-quality lymph node clinical target volumes for head and neck cancer radiation therapy using a fully automated deep learning-based approach. Int J Radiat Oncol Biol Phys 2021;109:801–12. https://doi.org/10.1016/j.ijrobp.2020.10.005.
[26] Ye X, Guo D, Tseng CK, et al. Multi-institutional validation of two-streamed deep learning method for automated delineation of esophageal gross tumor volume using planning CT and FDG-PET/CT. Front Oncol 2021;11:. https://doi.org/10.3389/fonc.2021.785788785788.
[27] Boers TGW, Hu Y, Gibson E, et al. Interactive 3D U-net for the segmentation of the pancreas in computed tomography scans. Phys Med and Biol 2020;65:. https://doi.org/10.1088/1361-6560/ab6f99065002.