



OPEN ACCESS

EDITED BY

Xiaodong Wu,
The University of Iowa, United States

REVIEWED BY

Jiajun Qiu,
Sichuan University, China
Lichun Zhang,
Stanford University, United States

*CORRESPONDENCE

Shujun Liang

✉ lsj123@smu.edu.cn

Yu Zhang

✉ yuzhang@smu.edu.cn

Yikai Xu

✉ yikai.xu@163.com

[†]These authors have contributed equally to this work

RECEIVED 27 January 2024

ACCEPTED 15 May 2024

PUBLISHED 14 June 2024

CITATION

Yang K, Dong X, Tang F, Ye F, Chen B, Liang S, Zhang Y and Xu Y (2024) A transformer-based multi-task deep learning model for simultaneous T-stage identification and segmentation of nasopharyngeal carcinoma. *Front. Oncol.* 14:1377366. doi: 10.3389/fonc.2024.1377366

COPYRIGHT

© 2024 Yang, Dong, Tang, Ye, Chen, Liang, Zhang and Xu. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

A transformer-based multi-task deep learning model for simultaneous T-stage identification and segmentation of nasopharyngeal carcinoma

Kaifan Yang^{1†}, Xiuyu Dong^{2,3,4†}, Fan Tang^{5†}, Feng Ye⁵, Bei Chen⁵, Shujun Liang^{2,3,4*}, Yu Zhang^{2,3,4*} and Yikai Xu^{1*}

¹Department of Medical Imaging Center, Nanfang Hospital, Southern Medical University, Guangzhou, Guangdong, China, ²School of Biomedical Engineering, Southern Medical University, Guangzhou, Guangdong, China, ³Guangdong Provincial Key Laboratory of Medical Image Processing, Southern Medical University, Guangzhou, Guangdong, China, ⁴Guangdong Province Engineering Laboratory for Medical Imaging and Diagnostic Technology, Southern Medical University, Guangzhou, Guangdong, China, ⁵Department of Radiation Oncology, Nanfang Hospital, Southern Medical University, Guangzhou, Guangdong, China

Background: Accurate tumor target contouring and T staging are vital for precision radiation therapy in nasopharyngeal carcinoma (NPC). Identifying T-stage and contouring the Gross tumor volume (GTV) manually is a laborious and highly time-consuming process. Previous deep learning-based studies have mainly been focused on tumor segmentation, and few studies have specifically addressed the tumor staging of NPC.

Objectives: To bridge this gap, we aim to devise a model that can simultaneously identify T-stage and perform accurate segmentation of GTV in NPC.

Materials and methods: We have developed a transformer-based multi-task deep learning model that can perform two tasks simultaneously: delineating the tumor contour and identifying T-stage. Our retrospective study involved contrast-enhanced T1-weighted images (CE-T1WI) of 320 NPC patients (T-stage: T1-T4) collected between 2017 and 2020 at our institution, which were randomly allocated into three cohorts for three-fold cross-validations, and conducted the external validation using an independent test set. We evaluated the predictive performance using the area under the receiver operating characteristic curve (ROC-AUC) and accuracy (ACC), with a 95% confidence interval (CI), and the contouring performance using the Dice similarity coefficient (DSC) and average surface distance (ASD).

Results: Our multi-task model exhibited sound performance in GTV contouring (median DSC: 0.74; ASD: 0.97 mm) and T staging (AUC: 0.85, 95% CI: 0.82–0.87) across 320 patients. In early T category tumors, the model achieved a median DSC of 0.74 and ASD of 0.98 mm, while in advanced T category tumors, it reached a median DSC of 0.74 and ASD of 0.96 mm. The accuracy of automated T staging was 76% (126 of 166) for early stages (T1-T2) and 64% (99 of 154) for advanced stages (T3-T4). Moreover, experimental results show that our multi-task model outperformed the other single-task models.

Conclusions: This study emphasized the potential of multi-task model for simultaneously delineating the tumor contour and identifying T-stage. The multi-task model harnesses the synergy between these interrelated learning tasks, leading to improvements in the performance of both tasks. The performance demonstrates the potential of our work for delineating the tumor contour and identifying T-stage and suggests that it can be a practical tool for supporting clinical precision radiation therapy.

KEYWORDS

nasopharyngeal carcinoma, deep learning, tumor segmentation, T-stage identification, multi-task

1 Introduction

Nasopharyngeal carcinoma (NPC), a frequently occurring malignant tumor in the head and neck region (1), primarily receives intensity-modulated radiation therapy (IMRT) as the standard treatment (2). The close proximity of NPC to critical neural and other organs necessitates precise delineation of the primary gross tumor volume (GTV) to minimize the risk of radiation-induced toxicities (3). Typically, early-stage NPC is managed with radiotherapy alone, whereas advanced local-regionally advanced disease often requires a combined approach of radiotherapy and chemotherapy (1, 4). Significantly, errors in contouring and T-staging adversely affect survival rates in head and neck cancer patients (5). Therefore, accurate tumor target contouring and T staging are vital for precision radiation therapy in NPC.

The complex anatomy of the nasopharynx and skull base complicates the precise identification of tumor invasion areas through surgical resection. NPC staging predominantly relies on physician interpretations of diagnostic images (1). Consequently, GTV contouring and T-staging for NPC are not only laborious but also highly time-consuming, requiring meticulous review of multi-modal or multi-parametric imaging data slice-by-slice, even by experienced physicians. Furthermore, GTV contouring and T-staging are prone to errors, particularly since NPC can infiltrate adjacent skull base and neural structures with often subtle signal variations in MRI scans. Thus, the NPC radiation therapy planning process heavily depends on the expertise of the physicians. Therefore, there is a significant clinical need for automatic segmentation and identification of GTV to aid in clinical decision-making, treatment planning, and ongoing tumor monitoring.

Recent advancements in deep-learning-based automatic methods have facilitated automatic NPC GTV segmentation (3, 6–9) and NPC staging (10–12). Lin et al. (3) pioneered the application of deep learning techniques for automating the contouring of primary tumor volumes in NPC patients using MRI. Chen et al. (6) introduced a multi-modality MRI fusion network to achieve precise segmentation of NPC by leveraging T1, T2, and contrast-enhanced T1 MRI images

(CE-T1WI). Tang et al. (7) proposed a dual attention mechanism for feature refinement to enhance NPC segmentation accuracy. Tao et al. (8) devised a sequential method for NPC segmentation in MR images, effectively addressing inherent background dominance issues at both the instance and feature levels. Regarding NPC staging, Yang et al. (10) proposed a weakly supervised deep learning model for NPC T-staging. Wong et al. (11) employed a CNN to automatically discriminate early-stage NPC from benign hyperplasia based solely on a non-contrast-enhanced MRI sequence. Huang et al. (12) proposed a two-stage classification framework for predicting locoregionally advanced NPC (stages III and IVa). In clinical practice, sizes and shapes of tumors are various and heterogeneous, assist the radiologists in identifying the four stages of T staging. This important characteristic of positional information is useful for both tumor segmentation and classification. However, existing deep learning-based studies have predominantly focused on either tumor segmentation or tumor staging of NPC alone, overlooking the inherent correlation between tumor segmentation and classification. Recent studies have highlighted the potential of deep learning models, especially those with refined multi-task features, in managing complex clinical tasks related to diseases like breast cancer (13), COVID-19 (14), gastric cancer (15), and thyroid nodules (16). Consequently, an integrated approach for training GTV contouring and T-staging within a unified network, promoting information sharing between these tasks, holds promise.

Convolutional Neural Network (CNN) are particularly adept at capturing spatial hierarchies in images, which makes them highly effective for image segmentation tasks (3, 6–8). The local connectivity and weight sharing in CNNs allow them to excel in extracting features from images that are spatially correlated, making them suitable for delineating complex structures like tumors in medical images. Additionally, the Transformer architecture, equipped with its Multi-Head Self-Attention (MHSA) mechanism (17), has demonstrated outstanding performance in analyzing long sequence correlations. This ability allows Transformers to learn expressive representations and filter out irrelevant signals, thereby significantly enhancing image analysis capabilities. Given the intricacy and distinctiveness of our task, we aim to develop a multi-task model that leverages the strengths of

both architectures using contrast-enhanced T1-weighted images (CE-T1WI) that can simultaneously identify T-stage and perform accurate segmentation of GTV in NPC. Specifically, we employ the 3D U-Net architecture for the GTV segmentation task. For T-stage identification, we design a Transformer-style prediction network that is adept at processing complex sequential information. To connect these two distinct tasks, we introduce a feature-fusion-aware module which is designed to analyze correlations among long sequences and learn expressive representations, improving the aggregation of features between the segmentation and staging tasks.

2 Materials and methods

2.1 Dataset

This retrospective study received approval from the ethics committee of our institution (No. NFEC-2023-417), and the requirement for informed consent was waived. The research was conducted in strict adherence to the guidelines and regulations set by the ethics committee.

Two independent datasets were used in this study. The inclusion criteria was as follows: (1) a definitive histopathological diagnosis of NPC; (2) absence of distant metastasis prior to treatment; (3) no previous history of radiotherapy; and (4) availability of a CE-T1WI sequence.

The first dataset included 356 cases of nasopharyngeal carcinoma. All patients were imaged using MRI platforms from different manufacturers using dissimilar imaging protocols: Optima MR360, GE Medical Systems, USA (n=280); Signa HDxt, GE Medical Systems, USA (n=43); Ingenia, Philips Medical Systems, USA (N=4); Achieva, Philips Medical Systems, USA (n=5); and Avanto, Siemens, Germany (n=24). The image size is $(256 - 672) \times (256 - 672) \times (16 - 120)$, with in-plane resolution as 0.34 - 1.09 mm, and slice thickness as 3.5 - 7 mm. We excluded 36 cases because of poor image quality (n = 16) or missing data (n = 20). Finally, 320 cases were retrospectively included as Dataset 1. These images were collected using different scanners with different image sizes and resolutions, thus increasing the diversity of the acquired MR images, and further improving the generalization of models.

The second dataset included 172 cases of nasopharyngeal carcinoma. The image size is $512 \times 512 \times (40 - 76)$, with in-plane resolution as 0.4688 - 0.5078 mm, and slice thickness as 3 - 6.7 mm. We excluded 22 cases because of poor image quality (n = 16) or missing data (n = 6). Finally, 150 cases were retrospectively included as Dataset 2.

In this study, a three-fold cross-validation strategy was used in Dataset 1 to select the best hyper-parameters. All subjects within Dataset 1 were divided into three subsets with the same proportion of each T-stage. Specifically, in each round, one subset was used for validation and the remaining two subsets were used for training. We repeated this process three times until all three data subsets had served as the validation set. Finally, the best hyper-parameters were obtained, and the models obtained by this three-fold cross-validation strategy were tested on the independent testing set. The research design is illustrated in Figure 1.

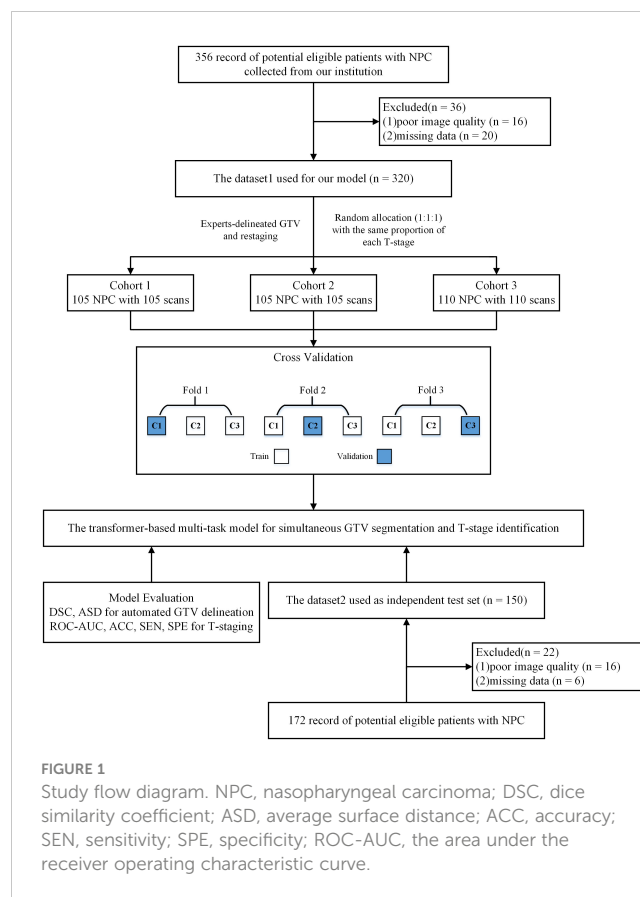


FIGURE 1
Study flow diagram. NPC, nasopharyngeal carcinoma; DSC, dice similarity coefficient; ASD, average surface distance; ACC, accuracy; SEN, sensitivity; SPE, specificity; ROC-AUC, the area under the receiver operating characteristic curve.

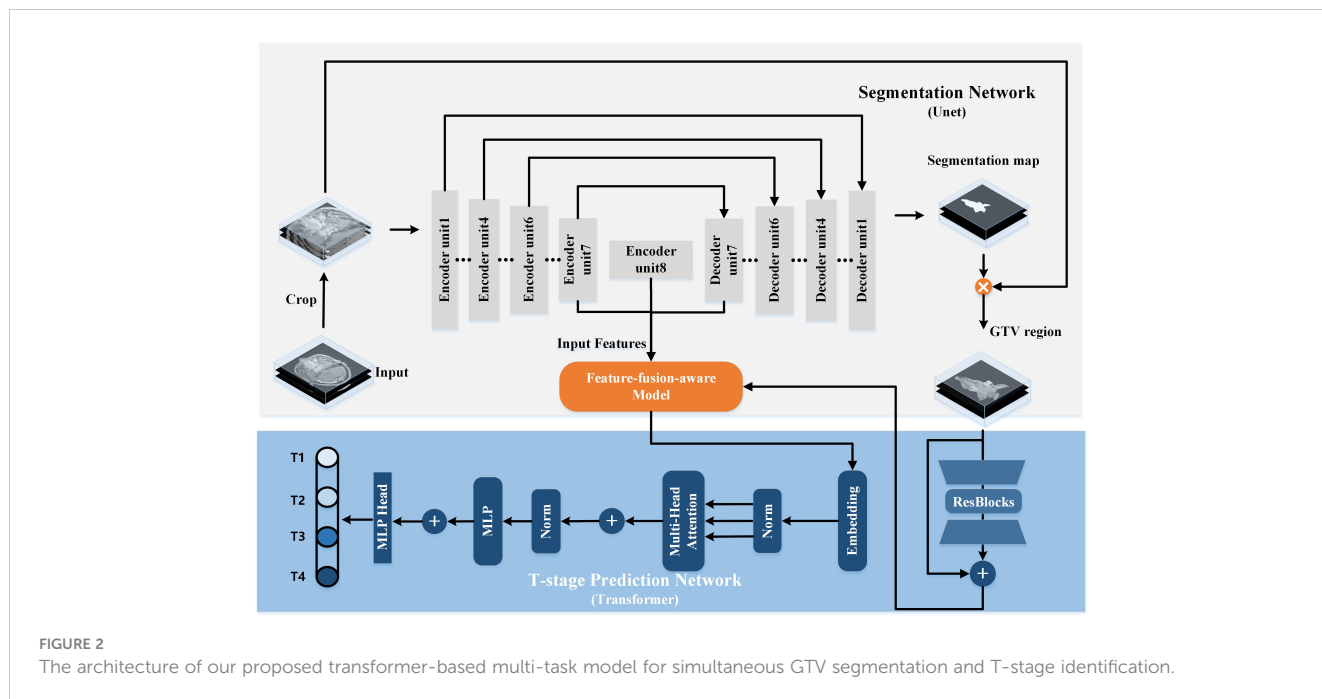
2.2 GTV manual delineation and patient re-staging

Manual segmentation of GTVs on CE-T1WI images were performed by a radiologist with 7 years of experience and checked by a radiation oncologist with more than 10 years of experience. All delineations were performed using the ITK-SNAP software (version 3.8.0, <http://www.itksnap.org>). These Manual segmentation of GTVs served as the basis for training, validating and testing our multi-task model.

To enhance data consistency and mitigate inter-observer variability, two experienced radiation oncologists (more than 10 years of experience in the field) independently reviewed the MR images and clinical data. Any discrepancies were resolved through a consensus-based approach. The T stages of the patients with NPC were reassessed according to the 8th edition of the Union for International Cancer Control/American Joint Committee on Cancer (UICC/AJCC) staging system (18).

2.3 Network architectures

We propose a multi-task model designed to simultaneously identify T-stage and accurately segment GTV in NPC. Illustrated in Figure 2, the proposed model comprises two primary components: a segmentation network for delineating the regions of interest (ROI) of the GTV, and a T-stage predictor for automated prediction of T-



stage probability scores. Within the segmentation network, a modified 3D U-Net is employed to perform GTV segmentation. Additionally, the segmented map is multiplied by the original MR image to generate a GTV region, serving as input for the T-stage predictor. In the T-stage prediction network, the tumor region is taken as input, initiating with a Bottleneck Transformer as the backbone to extract relevant features pertaining to the tumor. Notably, we adapt the network to accommodate and process 3D inputs. Furthermore, a feature-fusion-aware module is devised to connect and synchronize features derived from both GTV segmentation and T-staging processes. Finally, a classifier is employed to predict the probability scores of the four T stages.

2.3.1 3D segmentation network

The architecture of the 3D segmentation network encompasses a U-shaped encoding-decoding network comprising eight encoder units and seven decoder units. Traditionally, lower network layers extract feature maps delineating spatial structure information, while higher network layers furnish rich semantic information. Consequently, we posit that the most representative features learned by the segmentation network reside at the bottom of the U-shaped structure, specifically the seventh and final encoder units and the initial decoder unit. These features are subsequently fed into the T-stage prediction network.

The encoder part constitutes a contracting path facilitating down-sampling of the input image through eight encoder units, each succeeded by a 3D max pooling or average pooling operation with a stride of 2. Conversely, the decoder part embodies an expansive path, where feature maps are up-sampled through seven decoder units employing up-convolution layers. Notably, skip connections play a pivotal role. By establishing connections between feature maps from the compression path to the expansion

path, the network assimilates spatial information and refines segmentation predictions.

Subsequently, the output of the last decoder unit undergoes processing by a $1 \times 1 \times 1$ convolution followed by a sigmoid function. The generated segmentation map is then multiplied by the original MR image, yielding the input for the T-stage prediction network.

2.3.2 T-stage prediction network

The Transformer architecture, with its Multi-Head Self-Attention (MHSA) mechanism (17), has shown exceptional performance in analyzing long sequence correlations, learning expressive representations, and filtering out irrelevant signals, thus enhancing image analysis capabilities. Motivated by these findings, we design a transformer-style prediction network for T staging of NPC. As illustrated in Figure 2, the structure of T-stage prediction network consists of three modules: the 3D bottleneck transformer, the feature-fusion-aware module and a classifier.

3D Bottleneck Transformer. We adopt a backbone architecture called Bottleneck Transformer (19), which integrates self-attention mechanisms tailored for our T-stage prediction task. Its conception is simple: replace the final three spatial convolutions in a ResNet (20) with MHSA layers. As we all known, the ResNet structure typically has four stages. Our network just replaces the final three bottleneck blocks of ResNet-50 with MHSA layers.

Feature-fusion-aware Module. In Figure 3A, our feature-fusion-aware module combines the convolutional layer in CNNs with the MHSA layer in Transformer, which can take into account the features across different tasks. The MHSA layer consists of four parallel self-attention layers, each referred to as a head. For each head, query, key, and value are mapped into three current layers before attention calculation. The output of the four attention heads is then concatenated and input into the last linear layer for integration. In the self-attention layer (Figure 3B), the input is

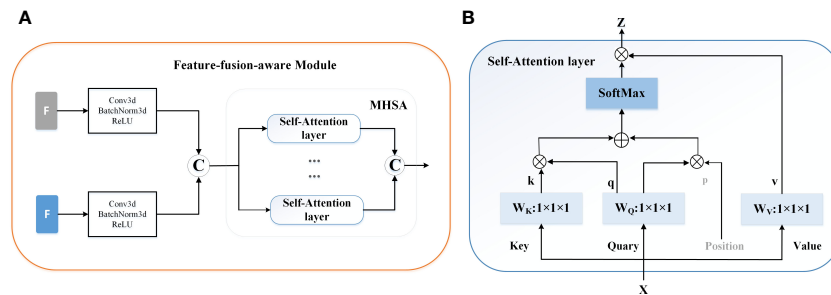


FIGURE 3
 The mechanism of the feature-fusion-aware module. **(A)** The gray 'F' represent the features extracted from the segmentation network and the blue 'F' indicates the output of the 3D backbone of prediction network. ⊕ represents the concatenation operation. **(B)** The detailed structure of Self-Attention layer.

mapped to q, k, and v. Subsequently, the spatial position encoding vector is multiplied by q as a spatial prior, and content information is acquired by multiplying q and k. Then, the spatially sensitive similarity feature is obtained by adding the two. Finally, the corresponding attention coefficient of the feature is multiplied by v to yield the final output z. Note that, the symbols ⊕ and ⊗ represent element-wise sum and matrix multiplication respectively, while 1 × 1 × 1 represents a point-wise convolution. And q, k, v and p represent query, key, value and position respectively.

In the final stage of the network, a classifier is deployed to accurately estimate the probability scores for the four distinct T stages.

2.3.3 Loss function

Our proposed method is a multi-task learning model that involves T-staging and tumor segmentation. T-staging can be considered as a classification task. As such, the overall loss function in Equation 1 consists of two components: the classification loss L_{Cls} for T-staging and the segmentation loss L_{Seg} for GTV segmentation:

$$L = \lambda_{Seg} L_{Seg} \{p, \hat{p}\} + \lambda_{Cls} L_{Cls} \{y, \hat{y}\} \quad (1)$$

where λ_{Cls} and λ_{Seg} are flexible hyper-parameters updated during network training. In this study, the parameters λ_{Cls} and λ_{Seg} are initially set as two tensors with a value of 1, and they are subsequently iteratively updated during the training phase. L_{Seg} represents a hybrid loss function which is combined by the Jaccard loss and the Focal loss, p denotes the ground-truth mask and \hat{p} denotes the predicted segmentation map; L_{Cls} is defined as cross-entropy loss aimed to compare the predicted probability of class with the ground-truth label, y denotes the ground-truth label and \hat{y} is the prediction probability.

2.4 Implementation details

Image processing: We resampled the images to a targeted resolution of 0.4688×0.4688×6 mm³, which represents the median spacing of all image data. This resampling utilized third-order spline interpolation for in-plane adjustments and the nearest neighbor approach for out-of-plane interpolation, as per the

methodology described in (21). To address potential issues in deep network signal propagation caused by large signals, we normalized each input by subtracting the mean and then dividing by the standard deviation. For the inputs to our multi-task model, we cropped the original MR images to a size of 21×224×224.

Training detail: The model was trained on Pytorch, utilizing an NVIDIA RTX 3090 Ti graphics processing unit (GPU), with axial T1CE. The training employed the 'Adam' optimizer, which was tasked with optimizing the combined loss. Training parameters included a batch size of four and an initial learning rate of 2×10^{-4} . To enhance the model's generalizability and stability, data augmentation techniques (22) and a ReduceLROnPlateau learning rate scheduling strategy (23) were employed. The training process was concluded either when the learning rate fell below 10^{-9} or after completing 300 epochs.

2.5 Performance evaluation

The performance of the models was assessed using the Dice similarity coefficient (DSC) and the average surface distance (ASD). The DSC quantifies the spatial overlap between the model-generated contour (A) and the ground truth contour (G), which is defined as $DSC(A, G) = \frac{2|A \cap G|}{|A| + |G|}$. ASD measures the average distance between the surfaces of two contours. Additionally, The predicted accuracy (ACC), sensitivity (SEN), specificity (SPE) of the model and the radiologists for identifying T-stage were evaluated by calculating the 95% confidence intervals (CIs) using the Clopper-Pearson method. Moreover, the area under the receiver operating characteristic curve (ROC-AUC) was employed to evaluate the algorithm's capability in predicting the T-stage. This metric is crucial for determining the effectiveness of our multi-task model in classifying the stages of nasopharyngeal carcinoma accurately.

2.6 Statistical analyses

For our study, categorical variables across the combined three cross-validation cohorts were compared using the chi-square test. Numeric variables, including comparisons of the DSC and ASD between different subgroups, were analyzed with the Mann-Whitney

U test. Furthermore, the Wilcoxon matched-pairs signed rank test was employed for contrasting DSC and ASD values obtained from our model against both manual delineation and other deep learning methods. The DeLong test was utilized for comparing the ROC-AUC metrics, while the McNemar test was applied for assessing the ACC.

Statistical analyses were conducted using the Statistical Product and Service Solutions (IBM SPSS, version 26.0) and open-access Python (version 3.6.10) statistical packages, namely “sklearn”, “scipy”, and “statsmodels”. For all tests, statistical significance was determined based on a two-tailed $P < 0.05$.

3 Results

3.1 Patient demographics

Patient demographics of Dataset1 is shown in Table 1. The Dataset1 included 320 patients, characterized by a median age of 48 years (range: 19-80 years), comprising 235 men (73.3%) and 85 women (26.6%). Stratification according to the 8th edition of the AJCC/UICC staging system identified 26 (8.1%), 140 (43.8%), 87 (27.2%), and 67 (20.9%) patients with T1-T4 stage disease, respectively. No significant differences in sex, age, and T category were observed between the combined three cross-validation cohorts.

3.2 Model performance in delineating GTV contour

The segmentation performance of our method is summarized in Table 2; Figure 4 illustrates the concordance level for GTV contours between our method and other segmentation methods. The median DSC was 0.74 (IQR, 0.11; 95% CI: 0.73, 0.75), and the median ASD was 0.97 mm (IQR, 0.75 mm; 95% CI: 0.9, 1.02 mm), less than the commonly accepted 3-mm margin of systematic and random error in radiation therapy for head and neck cancers (24). These results indicate strong concordance with human experts in GTV contouring.

In different T categories, our multi-task model achieved a median DSC of 0.74 (IQR, 0.11; 95% CI: 0.73, 0.75) and a median ASD of 0.98 mm (IQR, 0.83 mm; 95% CI: 0.87, 1.16 mm) in early T category tumors, and a median DSC of 0.74 (IQR, 0.11; 95% CI: 0.72, 0.75) and a median ASD of 0.96 mm (IQR, 0.68 mm; 95% CI: 0.89, 1.02 mm) in advanced T category tumors. No differences in DSC and ASD were observed across T categories, suggesting robustness of our multi-task model in contouring tumors of various sizes.

We also compared GTV contour accuracy of our multi-task model against other models like 3D U-Net (25), V-Net (26), and Med3D Net (27). As shown in Table 3, 3D U-Net showed the best DSC and ASD compared to V-Net, and Med3D Net. However, our multi-task model-generated contours had superior median DSC and ASD compared to those generated with 3D U-Net (median DSC, 0.74 vs 0.73; median ASD, 0.97 vs 1.05 mm; $P = 0.013 < 0.05$ for DSC; Table 3).

3.3 Model performance in identifying T-stage

The outputs of our multi-task model are the probabilities of stages T1, T2, T3, and T4. Confusion matrices and ROC curves are presented in Figures 5, 6, respectively. We observed an AUC of 0.85 (95% CI: 0.82, 0.87), an ACC of 0.7 (225 of 320; 95% CI: 0.65, 0.75), a SPE of 0.89 (95% CI: 0.88, 0.9), and a SEN of 0.62 (95% CI: 0.61, 0.62) as per Table 4. Three-fold cross-validation revealed ACCs of 0.72, 0.7, and 0.68, respectively (Figure 5), and AUCs of 0.88, 0.86, and 0.81, respectively (Figures 6A–C). For early T1-2 stages, automated T staging accuracy was 0.76 (126 of 166), while for advanced T3-4 stages, it was 0.64 (99 of 154), indicating a significant difference between early and advanced stages ($P < 0.05$).

Additionally, we compared our multi-task model’s predictive performance against other classification models like 3D ResNet-50 (20), RAN (28), DenseNet (29), and ResNext (30). The ROC curves and results are presented in Figure 7; Table 4. As indicated in Table 4, our multi-task model achieved the best AUC, ACC, SPE, and SEN ($P < 0.05$ for all).

TABLE 1 Characteristics of the study population of Dataset1.

Characteristic	Total (n = 320)	Cohort 1 (n = 105)	Cohort 2 (n = 105)	Cohort 3 (n = 110)	P Value		
					C1 vs C2	C1 vs C3	C2 vs C3
Sex, No (%)					0.541	0.407	0.148
Male	235 (73.3)	77 (73.4)	73 (69.5)	85 (77.3)			
Female	85 (26.6)	28 (26.7)	32 (30.5)	25 (22.7)			
Age, median (range), y	48 (19-80)	48 (19-80)	49 (20-74)	48 (19-80)	0.696	0.633	0.925
T stage, No (%)					1	0.962	0.962
T1	26 (8.1)	8 (7.6)	8 (7.6)	10 (9.1)			
T2	140 (43.8)	47 (44.8)	47 (44.8)	46 (41.8)			
T3	87 (27.2)	28 (26.7)	28 (26.7)	31 (28.2)			
T4	67 (20.9)	22 (20.9)	22 (20.9)	23 (20.9)			

Data are either number of patients, with the percentage in parentheses, or median, with the range in parentheses. We calculated P values by using the χ^2 test for category variables and the Mann-Whitney U test for numeric variables. Two-tailed $P < 0.05$ indicated a significant difference. C1, Cohort1; C2, Cohort2; C3, Cohort3.

TABLE 2 Performance of our multi-task model in delineating GTV contour using Dataset1.

	Total (n = 320)	T Category		P Value
		T1 and T2 (n = 166)	T3 and T4 (n = 154)	
DSC				0.648
Median	0.74	0.74	0.74	
IQR	0.11	0.11	0.11	
95% CI	0.73, 0.75	0.73, 0.75	0.72, 0.75	
ASD (mm)				0.984
Median	0.97	0.98	0.96	
IQR	0.75	0.83	0.68	
95% CI	0.90, 1.02	0.87, 1.16	0.89, 1.02	

We calculated the P value by using Mann-Whitney U test. Two-tailed P < 0.05 indicates a significant difference. DSC, Dice similarity coefficient; ASD, average surface distance; IQR, Inter-quartile range; CI, confidence interval.

3.4 Ablation study

In investigating the contribution of our multi-task design for simultaneously identifying T-stage and segmenting GTV, we have designed two ablation studies: 1) we adopted a 2-step training process (i.e., by first training the segmentation network, then training the T-stage prediction network using the tumor region generated by the segmentation map) and 2) we compared the multi-task model with its degraded version lacking the feature-fusion-

aware module (wo-FFA module). In the degraded version, features extracted from segmentation and T-stage prediction tasks were directly concatenated. Employing separate training reduces T-stage prediction performance greatly (see Separate-training in Table 5). This reduction may be attributed to the independent nature of segmentation of tumor region, which limits the information transfer between the two related learning tasks, thereby potentially hindering prediction performance. We can observe that the multi-task model trained without feature-fusion-aware

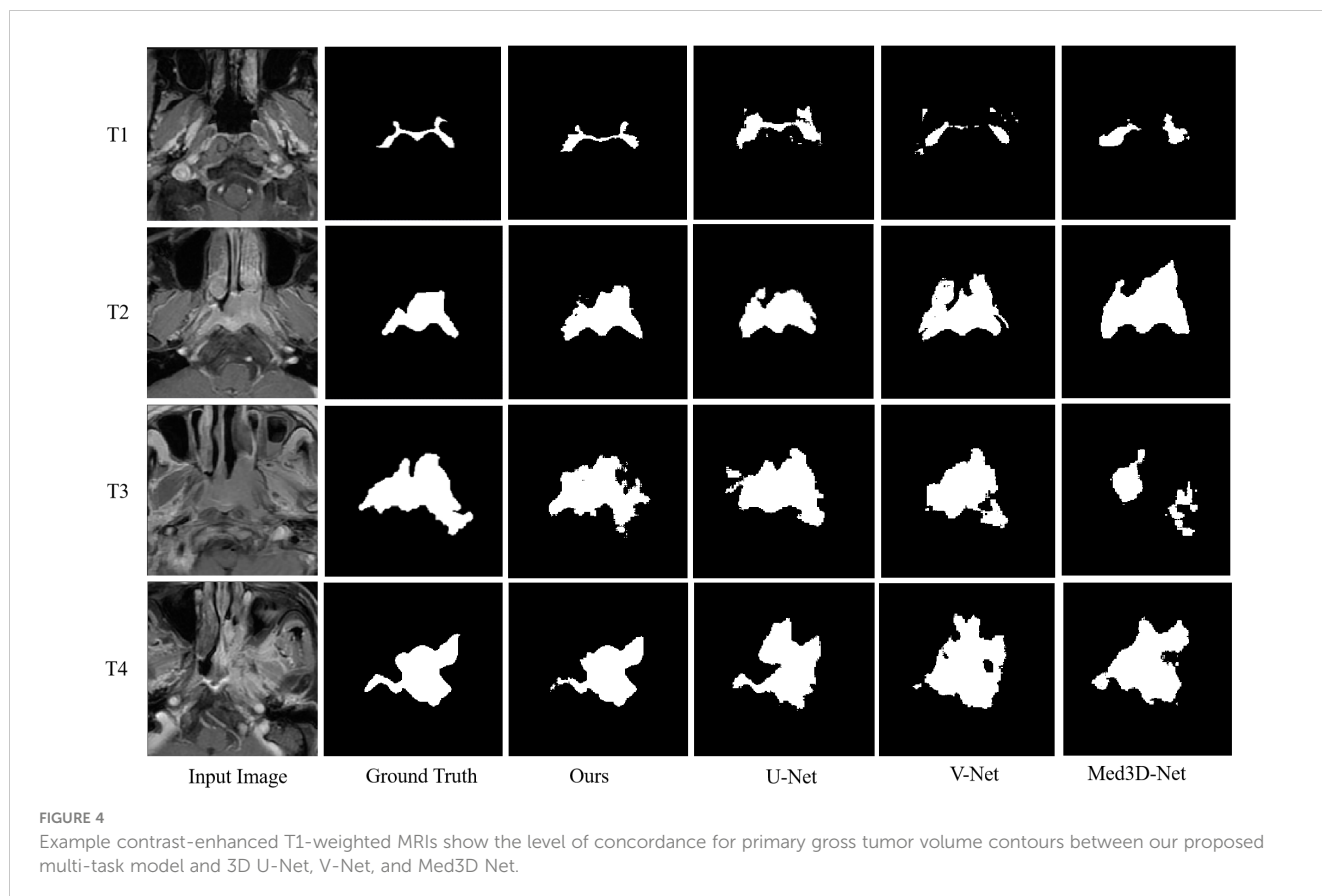


TABLE 3 Accuracy of contours generated by our proposed multi-task model and the 3D U-Net, V-Net, and Med3D Net.

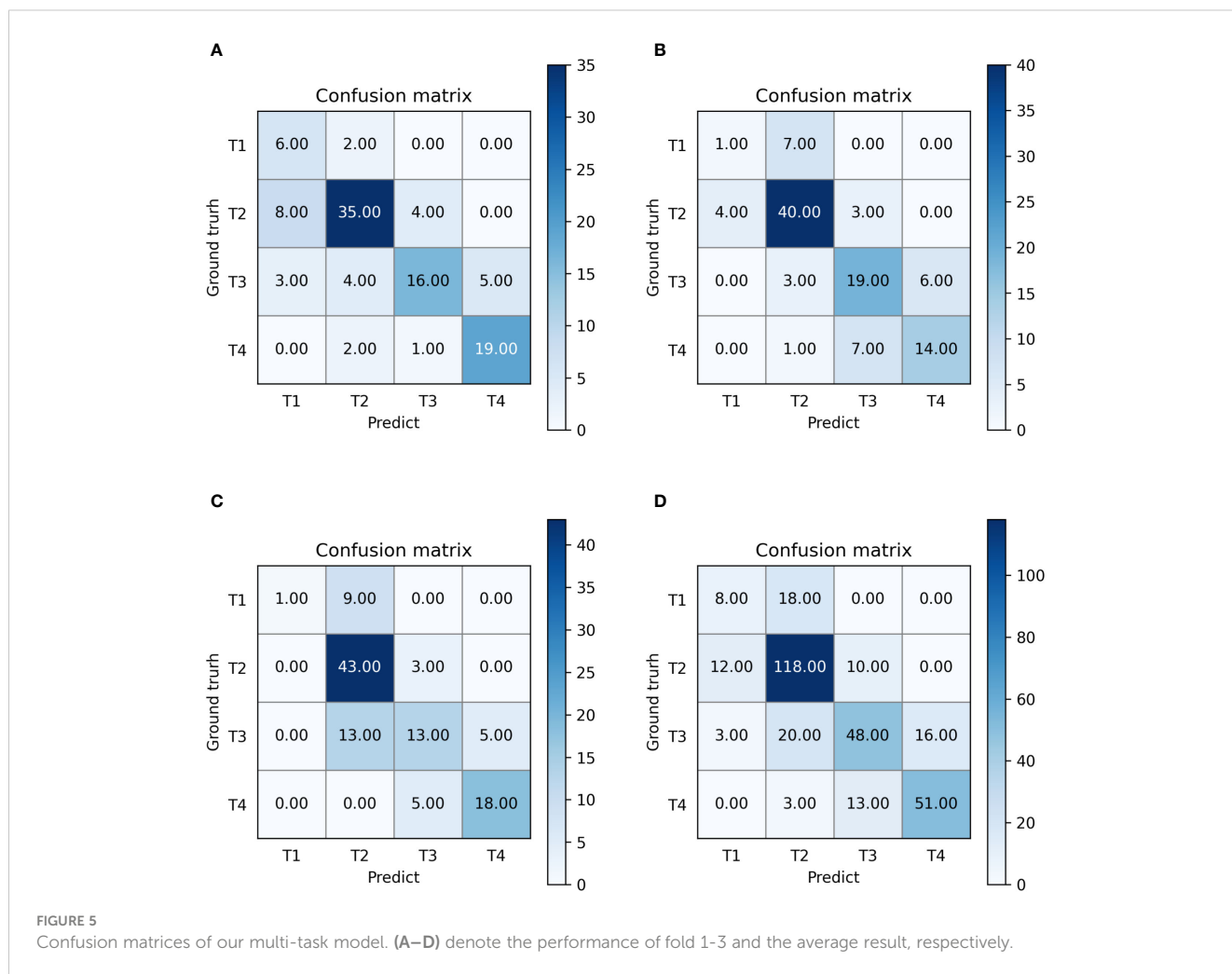
	DSC			ASD (mm)			P Value	
	Median	95% CI	IQR	Median	95% CI	IQR	DSC	ASD
U-Net	0.73	0.72, 0.74	0.12	1.05	0.96, 1.13	0.86	0.013	0.144
V-Net	0.69	0.68, 0.71	0.13	1.26	1.16, 1.36	1.01	< 0.05	< 0.05
Med3D Net	0.65	0.63, 0.67	0.18	1.51	1.42, 1.65	1.55	< 0.05	< 0.05
Ours	0.74	0.73, 0.75	0.11	0.97	0.9, 1.02	0.75		

We calculated P value using Wilcoxon matched-pairs signed-rank test, two-tailed $P < 0.05$ indicates significant difference. DSC, Dice similarity coefficient; ASD, average surface distance; IQR, inter-quartile range; CI, confidence interval.

module improve the performance compared with the Separate-training in Table 5, which indicated the effectiveness of our multi-task design for simultaneously identifying T-stage and segmenting GTV. Upon incorporating the proposed feature-fusion-aware module, our multi-task model achieved best performance (ACC: 0.7 vs 0.61, AUC: 0.85 vs 0.80, SPE: 0.89 vs 0.86, SEN: 0.62 vs 0.57; median DSC: 0.74 vs 0.73, median ASD: 0.97 vs 0.98). This underscores the advantage of the feature-fusion-aware module in refining and fusing features across different tasks and scales.

3.5 Result on the independent dataset

To show the generalization ability of our proposed method, we further conducted experiments on the independent dataset with other 150 subjects. Note that, we used the 150 subjects for testing three proposed multi-task models which were trained on three-fold cross-validation using 320 subjects, and obtained the average result. To quantitatively evaluate classification performance, we reported the results in terms of ACC, AUC, SPE, SEN for T-stage prediction



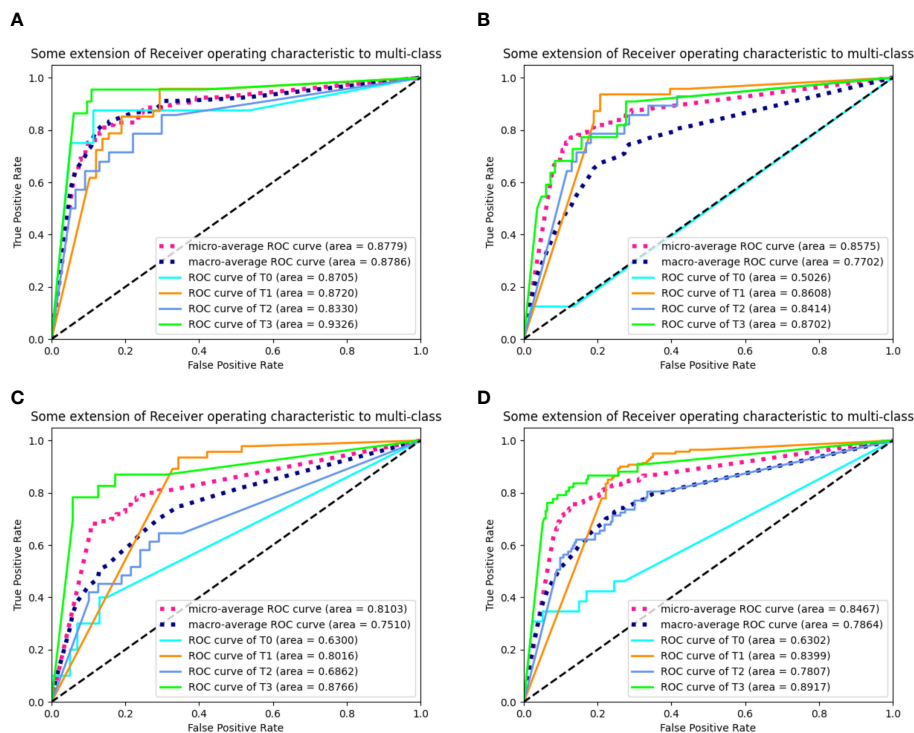


FIGURE 6 The receiver operating characteristic curves (ROCs) of our model. (A–D) denote the performance of fold 1–3 and the average result, respectively. ROC, receiver operating characteristic; AUC, area under curve; Micro-average, calculate metrics globally by considering each element of the label indicator as a label; Macro- average, calculate metrics for each label, and find their un-weighted mean.

and median DSC, median ASD for segmentation in Table 6. As shown in Table 6, our proposed method achieved reliable performance, illustrated the good generalization ability of the proposed multi-task model.

4 Discussion

In this research, we developed a multi-task deep learning model that achieves simultaneous GTV segmentation and T-stage identification in an end-to-end framework. Specifically, we tackle GTV segmentation task with the 3D U-Net architecture and design a transformer-style prediction network for T staging. To integrate the tasks of GTV segmentation and T-stage identification, we introduced a feature-fusion-aware module. We evaluated our proposed model

using two sets of CE-T1WI images. The proposed multi-task model has demonstrated remarkable competency in delineating and diagnosing primary GTV in the nasopharynx, achieved comparable contouring results, with a median DSC of 0.74 and ASD of 0.97 mm. It also proved effective in automated T staging, exhibiting an AUC of 0.85 (95% CI: 0.82, 0.87) across 320 patient cases. In comparative studies, our multi-task model outperformed the single-task segmentation network U-Net, achieving better DSC and ASD values (median DSC: 0.74 vs. 0.73; median ASD: 0.97 mm vs. 1.05 mm). Additionally, our multi-task model outperformed single-task classification networks in T-stage prediction, achieving the best results in AUC, ACC, SPE, and SEN. These extensive experimental results underscore that by integrating CNN for GTV segmentation and Transformer for T-staging, our multi-task model leverages the strengths of both architectures, optimizing performance for each

TABLE 4 Accuracy of T-staging by our proposed multi-task model and the 3D ResNet-50, RAN, DenseNet and ResNext.

	AUC	ACC	SPE	SEN	P*	P#
ResNet	0.80 (0.77, 0.83)	0.59 (0.54, 0.64)	0.85 (0.84, 0.86)	0.48 (0.45, 0.51)	<0.05	<0.05
RAN	0.80 (0.77, 0.83)	0.57 (0.51, 0.62)	0.83 (0.78, 0.88)	0.46 (0.36, 0.55)	<0.05	<0.05
DenseNet	0.78 (0.75, 0.82)	0.57 (0.52, 0.63)	0.83 (0.81, 0.85)	0.46 (0.39, 0.54)	<0.05	<0.05
ResNext	0.80 (0.76, 0.83)	0.59 (0.54, 0.65)	0.85 (0.83, 0.87)	0.48 (0.47, 0.50)	<0.05	<0.05
Ours	0.85 (0.82, 0.87)	0.70 (0.65, 0.75)	0.89 (0.88, 0.9)	0.62 (0.61, 0.62)		

We calculated P value using Wilcoxon matched-pairs signed-rank test, two-tailed P < 0.05 indicates significant difference. ACC, accuracy; SEN, sensitivity; SPE, specificity; AUC, the area under the receiver operating characteristic curve, P* for AUC; P# for ACC.

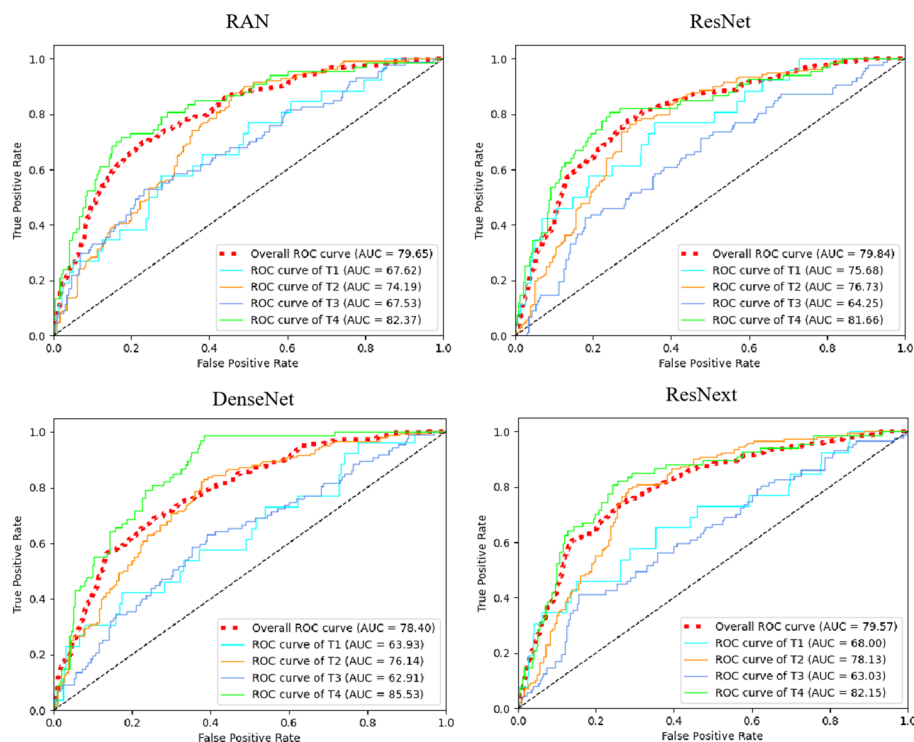


FIGURE 7 The receiver operating characteristic curves (ROCs) of ResNet, RAN, DenseNet, and ResNext. ROC, receiver operating characteristic; AUC, area under curve. Red dotted line means micro-average AUC, which calculates metrics globally by considering each element of the label indicator as a label.

specific task, providing a robust and reliable tool for NPC diagnosis and treatment planning.

The Chinese Society of Clinical Oncology (CSCO) 2022 Guidelines for NPC diagnosis and treatment advocate for “stratified administration”, suggesting less intense treatment for patients with lower risk of recurrence or metastasis and more aggressive treatment for those at higher risk (31). Timely diagnosis and personalized treatment plans are crucial for improving the long-term survival of NPC patients. Prior studies have shown that customized target setting for different T-stages can minimize radiation exposure to surrounding structures without increasing recurrence risk (32, 33). Given the scarcity of oncologists/radiologists and the labor-intensive nature of interpreting vast amounts of imaging data, there is a pressing need for more accurate tools to assist with the multitude of follow-up MR scans in NPC patients. To address this need, our objective was to develop a

computer aided design (CAD) tool using MR images for simultaneous tumor delineation and T-staging in NPC. Notably, the accuracy of our multi-task model in contouring GTV of NPC on CE-T1WI MR scan was lower than that for detecting primary tumor on multi-sequence MR scans in term of median DSC (0.74 vs 0.79), but achieved higher median ASD (0.97 mm vs 2.0 mm) (3). We hypothesize that integrating data from multiple sequences in future iterations of our multi-task model could further enhance its GTV contouring performance. In previous study focusing on T-stage prediction, a multi-perspective information aggregation framework demonstrated a slightly higher AUC (0.88 vs 0.85) for automatic T-staging in NPC than our multi-task model (34). However, it is important to note that this framework employed a multi-branch architecture with outputs from three parallel branches integrated through major voting, which may have contributed to its enhanced performance. In contrast, our

TABLE 5 Comparison of performance by our multi-task model and its degraded counterparts.

	Classification				Segmentation	
	ACC	AUC	SPE	SEN	DSC	ASD
Separate-training	0.54	0.78	0.84	0.53	0.73	1.05
wo-FFA module	0.61	0.80	0.86	0.57	0.73	0.98
Ours	0.7	0.85	0.89	0.62	0.74	0.97

DSC, Dice similarity coefficient; ASD, average surface distance; ACC, accuracy; SEN, sensitivity; SPE, specificity; AUC, the area under the receiver operating characteristic curve. Separate-training, a 2-step training process (i.e., by first training the segmentation network, then training the T-stage prediction network using the tumor region generated by the segmentation map); wo-FFA module, lacking the feature-fusion aware module.

TABLE 6 The performance for tumor segmentation and T-staging by our proposed multi-task model in second dataset.

Folds	Classification				Segmentation	
	ACC	AUC	SPE	SEN	DSC	ASD
Fold_1	0.69	0.83	0.87	0.48	0.76	1.23
Fold_2	0.65	0.84	0.85	0.49	0.72	1.68
Fold_3	0.62	0.81	0.86	0.75	0.71	1.50
Avg	0.65	0.83	0.86	0.53	0.73	1.42

DSC, Dice similarity coefficient; ASD, average surface distance; ACC, accuracy; SEN, sensitivity; SPE, specificity; AUC, the area under the receiver operating characteristic curve; Avg, Average.

multi-task model was designed as an end-to-end, multi-task network that synchronously outputted tumor region maps and T-stage probabilities, offering a streamlined and efficient approach to NPC diagnosis and staging.

Despite the promising results, our study does have several limitations. Primarily, our reliance solely on CE-T1WI MR scans may have limited the model's capability by not incorporating the discriminatory power of other MR sequences. To address this, future developments of our multi-task model will involve the integration of data from multiple sequences, aiming to enhance both GTV contouring and T-staging accuracy. Additionally, although our proposed method demonstrated reliable performance on an independent dataset, the generalizability of the model requires further enhancement through studies involving larger and more diverse datasets. Moreover, our model exhibited relatively weaker performance in identifying T1 stage disease compared to other stages. This discrepancy may be attributed in part to the category imbalance among the four T-stages in our dataset. Addressing this imbalance in future iterations could enhance the model's performance across all stages. Another limitation is that our current model does not fully align with the latest TNM staging system (8th edition), which underscores the importance of considering different combinations of T, N, and M classifications for NPC treatment. Consequently, there is a critical need to expand the model to include nodal gross tumor volume (GTVn) contouring and N-staging, which would offer a more comprehensive approach to NPC staging and treatment planning.

In summary, our study devises a model that can simultaneously identify T-stage and perform accurate segmentation of GTV in NPC. The findings emphasized the potential of multi-task model for simultaneously delineating the tumor contour and identifying T-stage. The multi-task model harnesses the synergy between these interrelated learning tasks, leading to improvements in the performance of both tasks. The impressive performance demonstrates the potential of our work for delineating the tumor contour and identifying T-stage and suggests that it can be a practical tool for supporting clinical precision radiation therapy.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material. Further inquiries can be directed to the corresponding authors.

Ethics statement

The studies involving humans were approved by Medical ethics committee of NanFang hospital of Southern Medical University. The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required from the participants or the participants' legal guardians/next of kin in accordance with the national legislation and institutional requirements.

Author contributions

KY: Data curation, Formal analysis, Investigation, Methodology, Validation, Visualization, Writing – original draft. XD: Formal analysis, Investigation, Methodology, Software, Validation, Writing – original draft. FT: Data curation, Formal analysis, Investigation, Visualization, Writing – original draft. FY: Data curation, Investigation, Validation, Writing – original draft. BC: Data curation, Investigation, Validation, Writing – original draft. SL: Conceptualization, Funding acquisition, Resources, Supervision, Writing – review & editing. YZ: Conceptualization, Funding acquisition, Resources, Software, Supervision, Writing – review & editing. YX: Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This study was supported by the National Natural Science Foundation of China [grant numbers 62001206; 61971213; U22A20350; 82271939] and Guangzhou Science and Technology Project [grant number 2023A04J2262].

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated

organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

- Chen Y, Chan ATC, Le QT, Blanchard P, Sun Y, Ma J. Nasopharyngeal carcinoma. *Lancet*. (2019) 394:64–80. doi: 10.1016/S0140-6736(19)30956-0
- Lee AW, Ma BB, Ng WT, Chan AT. Management of nasopharyngeal carcinoma: current practice and future perspective. *J Clin Oncol*. (2015) 33:3356–64. doi: 10.1200/JCO.2015.60.9347
- Lin L, Dou Q, Jin Y, Zhou G, Tang Y, Chen W, et al. Deep learning for automated contouring of primary tumor volumes by MRI for nasopharyngeal carcinoma. *Radiology*. (2019) 291:677–86. doi: 10.1148/radiol.2019182012
- Du X, Wang G, Zhu X, Han Y, Lei F, Shen L, et al. Refining the 8th edition TNM classification for EBV related nasopharyngeal carcinoma. *Cancer Cell*. (2024) 42:1–10. doi: 10.1016/j.ccell.2023.12.020
- Chen A, Chin R, Beron P, Yoshizaki T, Mikaeilian A, Cao M. Inadequate target volume delineation and local-regional recurrence after intensity-modulated radiotherapy for human papillomavirus-positive oropharynx cancer. *Radiother Oncol*. (2017) 123:412–8. doi: 10.1016/j.radonc.2017.04.015
- Chen H, Qi Y, Yin Y, Li T, Liu X, Li X, et al. MMFNet: A multi-modality MRI fusion network for segmentation of nasopharyngeal carcinoma. *Neurocomputing*. (2020) 394:27–47. doi: 10.1016/j.neucom.2020.02.002
- Tang P, Zu C, Hong M, Yan R, Peng X, Xiao J, et al. DA-DSUnet: Dual Attention-based Dense SU-net for automatic head-and-neck tumor segmentation in MRI images. *Neurocomputing*. (2021) 435:103–13. doi: 10.1016/j.neucom.2020.12.085
- Tao G, Li H, Huang J, Bree R, Lee AWM, Lee VHF, et al. SeqSeg: A sequential method to achieve nasopharyngeal carcinoma segmentation free from background dominance. *Med Image Anal*. (2022) 78:102381. doi: 10.1016/j.media.2022.102381
- Ng WT, But B, Choi HCW, et al. Application of artificial intelligence for nasopharyngeal carcinoma management—a systematic review. *Cancer Manage Res*. (2022) 14:339–66. doi: 10.2147/CMAR.S341583
- Yang Q, Guo Y, Ou XM, Wang JZ, Hu CS. Automatic T staging using weakly supervised deep learning for nasopharyngeal carcinoma on MR images. *J Magn Reson Imaging*. (2020) 52:1074–82. doi: 10.1002/jmri.27202
- Wong LM, King AD, Ai QYH, Lam WKJ, Poon DMC, Ma BBY, et al. Convolutional neural network for discriminating nasopharyngeal carcinoma and benign hyperplasia on MRI. *Eur Radiol*. (2021) 31:3856–63. doi: 10.1007/s00330-020-07451-y
- Huang J, He R, Chen J, Li S, Deng Y, Wu X. Boosting advanced nasopharyngeal carcinoma stage prediction using a Two-stage classification framework based on deep learning. *Int J Comput Intell Syst*. (2021) 14:1–14. doi: 10.1007/s44196-021-00026-9
- Zhou Y, Chen H, Li Y, Liu Q, Xu X, Wang S, et al. Multi-task learning for segmentation and classification of tumors in 3D automated breast ultrasound images. *Med Image Anal*. (2021) 70:101918. doi: 10.1016/j.media.2020.101918
- Wang X, Jiang L, Li L, Xu M, Deng X, Dai L, et al. Joint learning of 3D lesion segmentation and classification for explainable COVID-19 diagnosis. *IEEE Trans Med Imaging*. (2021) 40:2463–76. doi: 10.1109/TMI.2021.3079709
- Zhang Y, Li H, Du J, Qin J, Wang T, Chen Y, et al. 3D multi-attention guided multi-task learning network for automatic gastric tumor segmentation and lymph node classification. *IEEE Trans Med Imaging*. (2021) 40:1618–31. doi: 10.1109/TMI.2021.3062902
- Zhao S, Chen Y, Yang K, Luo Y, Ma B, Li Y. A local and global feature disentangled network: toward classification of benign-malignant thyroid nodules from ultrasound image. *IEEE Trans Med Imaging*. (2022) 41:1497–509. doi: 10.1109/TMI.2022.3140797
- Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. *Adv Neural Inf Process Syst*. (2017) 30:5998–6008. doi: 10.48550/arXiv.1706.03762
- Lydiatt WM, Patel SG, O'Sullivan B, Brandwein MS, Ridge JA, Migliacci JC, et al. Head and neck cancers-major changes in the American Joint Committee on cancer eighth edition cancer staging manual. *CA: A Cancer J Clin*. (2017) 67:122–37. doi: 10.3322/caac.21389
- Srinivas A, Lin T, Parmar N, Shlens J, Abbeel P, Vaswani A. (2021). Bottleneck transformers for visual recognition, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (Nashville, TN: IEEE), pp. 16519–29. doi: 10.1109/CVPR46437.2021.01625
- He K, Zhang X, Ren S, Sun J. (2016). Deep residual learning for image recognition, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (Las Vegas, NV: IEEE), pp. 770–8. doi: 10.1109/CVPR.2016.90
- Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods*. (2021) 18:203–11. doi: 10.1038/s41592-020-01008-z
- Zaheer M, Reddi S, Sachan D, Kale S, Kumar S. Adaptive methods for nonconvex optimization. *Adv Neural Inform Process Syst*. (2018) 31:9793–803. doi: 10.5555/3327546.3327647
- Tang Z, Xu Y, Jin L, Aibaidula A, Lu J, Jiao Z, et al. Deep learning of imaging phenotype and genotype for predicting overall survival time of glioblastoma patients. *IEEE Trans Med Imaging*. (2020) 39:2100–9. doi: 10.1109/TMI.2020.02.002
- Karam I, Lee NY, Le QT, O'Sullivan B, Lu JJ, Poon I. Nasopharyngeal carcinoma. In: *Target Volume Delineation and Field Setup: A Practical Guide for Conformal and Intensity-Modulated Radiation Therapy, 2nd ed.* Springer, Berlin Germany (2022). p. 1–13.
- Ronneberger O, Fischer P. (2015). Brox T. U-Net: Convolutional networks for biomedical image segmentation, in: *2015 Conference on Medical image computing and computer-assisted intervention- MICCAI*, Cham. pp. 234–41. (Munich, Germany: Springer, Cham). doi: 10.1007/978-3-319-24574-4_28
- Milletari F, Navab N. (2016). Ahmadi SA. V-Net: Fully convolutional neural networks for volumetric medical image segmentation, in: *Proceedings of the International Conference on 3D vision (3DV)*, (Stanford, CA: IEEE), 565–71. doi: 10.1109/3DV.2016.79
- Chen S, Ma K, Zheng Y. Med3D: Transfer learning for 3D medical image analysis. *arXiv preprint arXiv*. (2019) 1904.00625. doi: 10.48550/arXiv.1904.00625
- Wang F, Jiang M, Qian C, Yang S, Li C, Zhang H, et al. (2017). Residual attention network for image classification, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (Honolulu, HI: IEEE), 3156–64. doi: 10.1109/CVPR.2017.683
- Huang G, Liu Z, van der Maaten L, Weinberger KQ. (2017). Densely connected convolutional networks, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (Honolulu, HI: IEEE), 2261–9. doi: 10.1109/CVPR.2017.243
- Xie S, Girshick R, Dollár P, Tu Z, He K. (2017). Aggregated residual transformations for deep neural networks, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (Honolulu, HI: IEEE), 5987–95. doi: 10.1109/CVPR.2017.634
- Tang L, Chen Y, Chen C, Chen M, Chen N, Chen X, et al. The Chinese society of clinical oncology (CSCO) clinical guidelines for the diagnosis and treatment of nasopharyngeal carcinoma. *Cancer Commun*. (2021) 41:1195–227. doi: 10.1002/cac2.12218
- Miao J, Di M, Chen B, Wang L, Cao Y, Xiao W, et al. A prospective 10-year observational study of reduction of radiation therapy clinical target volume and dose in early-stage nasopharyngeal carcinoma. *Int J Radiat Oncol Biol Phys*. (2020) 107:672–82. doi: 10.1016/j.ijrobp.2020.03.029
- Sanford N, Lau J, Lam M, Juliano AF, Adams JA, Goldberg SL, et al. Individualization of clinical target volume delineation based on stepwise spread of nasopharyngeal carcinoma: outcome of more than a decade of clinical experience. *Int J Radiat Oncol Biol Phys*. (2019) 103:654–68. doi: 10.1016/j.ijrobp.2018.10.006
- Liang S, Dong X, Yang K, Chu Z, Tang F, Ye F, et al. A multi-perspective information aggregation network for automated T-staging detection of nasopharyngeal carcinoma. *Phys Med Biol*. (2022) 67:245007. doi: 10.1088/1361-6560/aca516