IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 3, 2020, accepted January 20, 2021, date of publication February 1, 2021, date of current version February 8, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3056130

Combining Images and T-Staging Information to Improve the Automatic Segmentation of Nasopharyngeal Carcinoma Tumors in MR Images

MINGWEI CAI, JIAZHOU WANG, QING YANG, YING GUO, ZHEN ZHANG, HONGMEI YING, WEIGANG HU, AND CHAOSU HU

Department of Radiation Oncology, Fudan University Shanghai Cancer Center, Shanghai 200032, China
Department of Oncology, Shanghai Medical College, Fudan University, Shanghai 200032, China

Corresponding authors: Weigang Hu (jackhuwg@gmail.com) and Chaosu Hu (hucsu62@163.com)

ABSTRACT The accurate and reproducible delineation of tumors from uninvolved tissue is essential for radiation oncology. However, the tumor margin may be challenging to identify from magnetic resonance (MR) images of nasopharyngeal carcinomas (NPCs). Additionally, clinical diagnoses such as T-staging may already provide some information on tumor invasion. To use this information and improve the performance of tumor segmentation, we propose a novel deep learning neural network architecture that can incorporate both T-staging and image information. Based on U-Net, our model adds a T-channel composed of T-staging information and uses the attention mechanism. Since the T-staging information is defined by the extent of tumor invasion, the T-channel using T-staging information can improve the segmentation accuracy at different stages. Additionally, the addition of an attention mechanism allows our model to retain the most valuable pixels of the image, thus further improving the delineation accuracy. In our experiments, the proposed network was trained and validated based on records from 251 clinical patients using 10-fold cross-validation. The dice similarity coefficient (DSC) and average symmetric surface distance (ASSD) were used to evaluate our network's results. The average DSC and ASSD and their standard deviation (SD) values are 0.841 ± 0.011 and 0.747 ± 0.199 mm. The unique T-channel effectively utilizes T-staging information to improve the results. With the combination of the T-channel module and the attention module, we significantly improved NPC tumor delineation performance.

INDEX TERMS Deep learning, magnetic resonance images, autosegmentation, nasopharyngeal carcinoma.

I. INTRODUCTION

NPC is epithelial cancer with a worldwide distribution. It is endemic in Southeast China, especially in the Guangzhou area's Cantonese population (up to 80 cases per 100,000 people per year). A medium incidence is found in other parts of North Africa and Southern Asia and among indigenous people in Greenland and Alaska (8–12 cases per 100,000 people per year) [1]. Radiotherapy (RT) for NPC is the primary treatment and has achieved 5-year survival rates of 90% and 84% for early-stage I and IIA diseases, respectively [2].

Radiotherapy planning is significantly affected by NPC delineation accuracy, and MR is a preferred technique for NPC delineation because it uses nonionizing radiation, is non-invasive, and has superior soft-tissue contrast [3]. In clinical practice, NPC is manually delineated by experienced physicians. However, it is a time-consuming and subjective process that has an enormous influence on RT planning.

Although many studies have investigated the autosegmentation of NPC tumors, it remains a challenging task due to the complicated anatomical structures involved, and medical images contain much more information than what can be observed by humans [4]. Additionally, the patient's T-staging, based on tumor invasion, is assessed by an experienced radiologist during an appropriate diagnosis [5]. The classification of malignant tumor TNM is a globally accepted classification standard for cancer spread. It is the classification system of tumor anatomy. T-staging describes the size of the original tumor and whether it has invaded nearby tissue [6].

The associate editor coordinating the review of this manuscript and approving it for publication was Jenny Mahoney.

Therefore, we considered combining the T-staging information with deep learning model to improve NPC tumor segmentation performance. There are no related works to our knowledge.

In this study, we propose a deep learning model that incorporates T-staging and image information to further improve NPC tumor delineation accuracy.

Based on U-Net [7], our model adds a T-channel composed of T-staging information and uses the attention mechanism. Since the T-staging information is defined by the extent of tumor invasion, the T-channel using T-staging information can improve different stages' effect. Simultaneously, the addition of an attention mechanism allows the model to retain the most valuable pixels of the image, further improving the delineation accuracy. In our experiments, the proposed network was trained and validated on 251 clinical patients using 10-fold cross-validation. The MRIs used in our study consisted of T1-weighted (T1W), T2-weighted (T2W), and contrast-enhanced T1-weighted (CE-T1W) images. Different types of MRIs show different information on the same tissue, which is helpful for tumor segmentation. Additionally, we utilize T-staging information as input to a deep learning model designed as an additional channel of input images.

This paper is organized as follows. Section II summarizes the related research progress of NPC tumor automatic delineation. The details of our proposed automatic segmentation model are described in Section III. Section IV introduces the experimental details, and the results are presented in Section V. Finally, we discuss the research results and deficiencies in Section VI.

II. RELATED WORKS

A. TRADITIONAL SEGMENTATION MODELS

Several automatic delineation algorithms have been developed over the past 20 years. Jolly [8] cultivated a segmentation model with registrations and minimum surfaces in medical images. The deformable model used to be representative of medical image segmentation. Billet *et al.* [9] developed the deformable model and achieved good performance. The use of a shape prior had also been a trend in the past. Lin *et al.* [10] utilized a shape prior model to finish segmentation tasks. Active shape and appearance models (ASM/AAM) [11] represent shape and texture variability in medical images.

Subsequently, Zhang *et al.* [12] developed an autosegmentation algorithm with the AAM/ASM model. An atlas shows the different structures present in a given image [13]. Lötjönen *et al.* [14] used an atlas to complete segmentation in medical images.

B. DEEP LEARNING FOR MEDICAL IMAGE SEGMENTATION

With the development of deep learning in recent years, significant progress has been made in the automatic segmentation of medical images. A fully convolutional network was developed for pixel-level prediction [15], making it possible to use deep learning to complete segmentation tasks. The subsequent development of U-Net further improved segmentation accuracy, and U-Net achieved better results in the field of medical image segmentation. Kumar *et al.* [16] used U-Net to complete the real-time segmentation of breast masses. Yang *et al.* [17] used ResNet [18] and U-Net to achieve brain tumor segmentation.

In recent studies, various modified U-Net models have further improved the accuracy of delineation. Milletari *et al.* [19] cultivated a 3D convolution and a residual module to address 3D tasks. Nie *et al.* [20] changed the concatenation of skip connections into convolutions to improve performance. Wang *et al.* [21] combined local and global information and obtained better results.

In the abovementioned studies, deep learning has shown unique advantages in medical image segmentation, making it necessary for future progress.

C. DEEP LEARNING FOR NPC SEGMENTATION

Li *et al.* [22] proposed an automatic segmentation method for NPC based on a CNN with dynamic contrast-enhanced MRI, which significantly improved NPC delineation accuracy. Zhao *et al.* [23] used fully convolutional networks with auxiliary paths to achieve the NPC's automatic segmentation on PET-CT images. During training, they implemented a deep supervision technique by adding auxiliary paths to improve the network's capability. Guo *et al.* [24] designed a 3D CNN with a long-range skip connection and a multiscale pyramid for NPC segmentation. This network can perceive the multiscale features of tumors as well as hierarchical semantic and contextual information. They also used deep supervision to generate auxiliary segmentation prediction and added the weighted loss of the auxiliary segmentation map to the total loss, which helped accelerate the network's convergence. Ye *et al.* [25] proposed and verified an accurate and efficient automatic NPC segmentation method based on dense connectivity embedding U-Net (DEU) and dual-sequence MRI images. The DEU extracted the features of T1W and T2W in different paths automatically and fused the features with dense connectivity blocks, which contributed to the increased accuracy.

There are considerable differences in NPC patients' tumors, such as the shape and dimension, making delineation difficult. By using a CNN, the accuracy of NPC automatic segmentation can be significantly improved.

D. OUR CONTRIBUTIONS

In this work, we proposed a deep learning model that incorporated both T-staging information and MR image information. Additionally, we added an attention module into our network that made the model more robust and accurate. The contributions of this study can be summarized as follows:

(1) We developed a novel T-channel module for our network to incorporate T-staging information and a deep learning model.

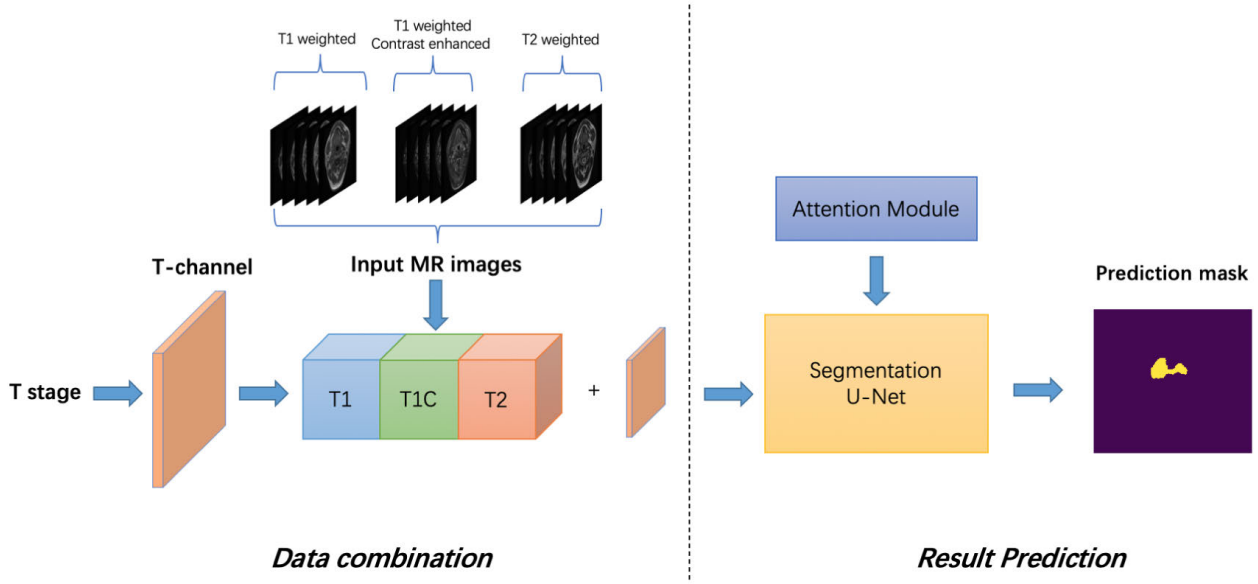


FIGURE 1. The flowchart of our proposed method. In the data combination part, we transfer T-staging information into the T-channel to the MR images. In the result prediction part, we develop an attention U-Net to complete segmentation.

(2) We proposed a novel modified U-Net with an attention module and a T-channel module that can be used in autosegmentation for NPC.

(3) We used 251 patients to train and validate our model and obtained a robust result.

III. METHODS

A. INPUT DATA PROCESSING

The data combination part is shown in Figure 1. First, we took five slices of each type of MR image and composed them into different channels (T1W-channel, CE-T1W-channel, and T2W-channel). Then, we merged these three channels through a concatenation operation and obtained a 512 × 512 × 15 image volume. This operation helped the model learn the same features from the input images in different states and be applied to different MR images.

To utilize the T-staging information for improving the performance of a CNN model, we transferred its digital information to the T-channel by:

$$F(n) = n\alpha x \tag{1}$$

where n represents the T-staging information (1 to 4 represents the T1 to T4 stages) and α denotes the weight of the background; here, we set α to 0.25 due to the four different stages and x is a background image, which is a 512 × 512 image with a pixel value of 255. Through $F(n)$, we obtained the T-channel of each stage shown in Figure 3. We concatenated the T-channel with the T1W-channel, CE-T1W-channel, and T2W-channel and obtained the final input as a 512 × 512 × 16 input volume that contained different type and stage information.

B. ATTENTION U-NET

We trained a U-Net with an attention module to accomplish our task. The basic structure of our model is shown in Figure 2. The input volume (512 × 512 × 16) was obtained from the data combination part, and then the encoder part extracted highly representative features and reduced the input volume size. The decoder part utilized a deconvolution operation, an upsampling transposed convolution, to rebuild an image of the same size as the input image from the extracted features.

The encoder part was composed of four downsampling blocks and a bottom block. Each downsampling block consisted of two 3 × 3 convolutional layers, two batch normalization (BN) [26] layers, and double rectified linear unit (ReLU) [27] layers. The BN layer was designed to prevent gradient explosion and vanishing, and a ReLU activation function followed each BN layer. The ReLU is defined as:

$$y = \begin{cases} 0, & x < 0 \\ x, & x \geq 0 \end{cases} \tag{2}$$

where x represents the input and y denotes the output.

The decoder part was composed of three upsampling blocks. Each upsampling block consisted of a 3 × 3 deconvolutional layer, a concatenation layer, two BN layers, two 3 × 3 convolutional layers, and double ReLU. High-resolution images may lose information because of the deconvolution operation.

The skip connection operation was used to address this problem by fusing the feature maps from the downsampling blocks with the feature maps in the deconvolutional layer. These concatenation layers can obtain more contextual information on multiple scales to improve delineation

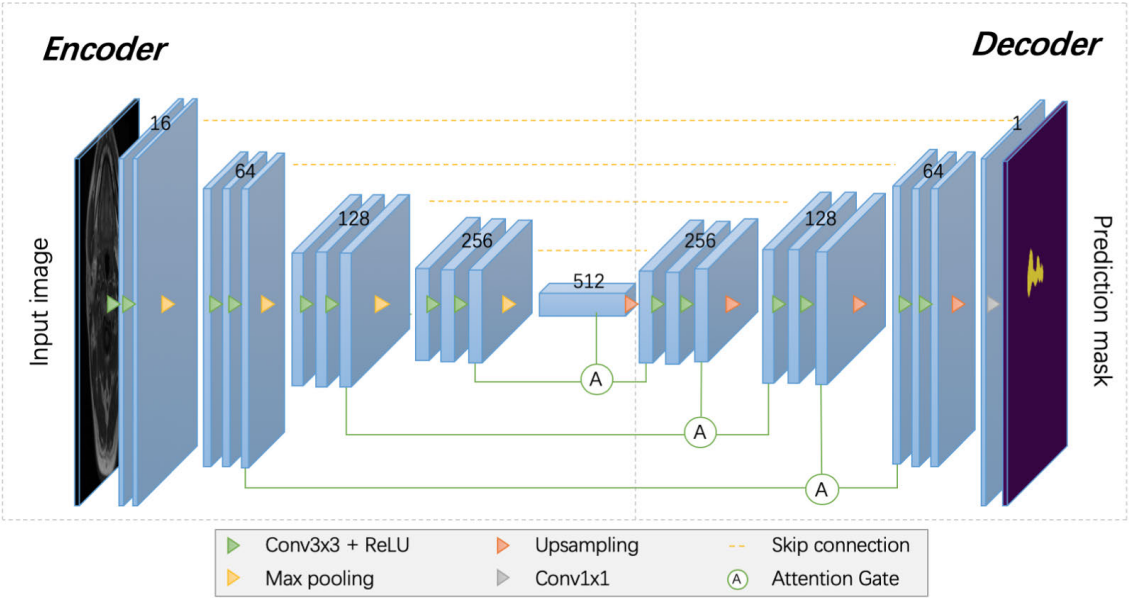


FIGURE 2. Illustration of our proposed convolutional neural network architecture. The network includes two phases of encoder and decoder. The encoder part extracted highly representative features and reduced the input volume size and the decoder part utilized a deconvolution operation, to rebuild an image of the same size as the input image from the extracted features.

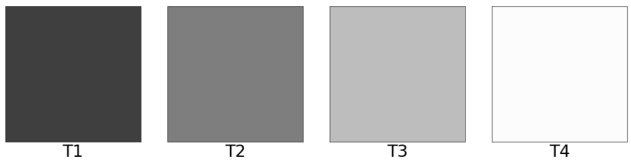


FIGURE 3. The T-channel of each stage. We use different pixel values to represent each stage.

performance. Finally, the feature maps were computed by a 1×1 convolutional layer with a sigmoid function.

With all the upsampling blocks, the model output a 512×512 image that was rebuilt by the decoder part, which was the same size as the input images. The dice loss [19] between the ground truth and the prediction mask was computed as a loss function for our network. In the encoder and decoder parts, an attention mechanism was applied to optimize the extracted spatial information of the feature maps [28], [29].

In our study, a mask with pixel values between 0 and 1 was generated by transformation, and then the feature maps were multiplied by the mask. The region of interest stayed unchanged, and the rest of the feature map was set to zero. Finally, the attention mechanism ensured that the useful information in the feature maps was preserved. Figure 4 shows our attention gating signal unit. A gating signal unit that consists of a 3×3 convolutional layer, a BN layer, a ReLU, a 1×1 convolutional layer, and a sigmoid function used to produce the signal information of the input feature map that kept the region of interest unchanged. The attention gating was responsible for synthesizing feature maps from different parts. The entire attention module process is shown in figure 5. First, the module receives two feature maps as input, x^d is from downsampling blocks, and x^u is from upsampling blocks, then we obtain the corresponding gating signal information through $G(x)$. An element-wise product was applied for the input and their corresponding gating signal, which preserved the image's essential information. Finally, we got two feature maps with the same size through a deconvolution operation and added them together to obtain the attention module output.

IV. EXPERIMENTS

A. DATASETS AND EXPERIMENTS

A total of 251 NPC patients diagnosed and histologically confirmed at the Fudan University Shanghai Cancer Center from February 2010 to January 2012 were selected for this study. The patients included 183 males and 68 females who ranged from 13 to 75 (average 46.19) years old. The patients' T-staging information based on the 8th edition of the UICC/AJCC staging system was extracted from Electronic Medical Records (EMRs) and reviewed by one radiation oncologist. The NPC tumor boundary's ground truth was manually delineated by physicians with five years of experience with MR images.

The diagnostic MR images included T1-weighted (T1W), T2-weighted (T2W), and contrast-enhanced T1-weighted (CE-T1W) images. We obtained T1W and T2W images on a 1.5 T MRI system (GE, Milwaukee, WI) with an 8-channel phased-array joint coil. T1W scans (echo time [TE]: 9-15 ms, repetition time [TR]: 600-800 ms) in the sagittal and transverse planes and T2W scans (TE: 80-100 ms and TR: 3000-4000 ms) in the transverse plane were obtained before the injection of a contrast agent. Gadolinium-diethylene triamine pentaacetic acid (Gd-DTPA) was applied as the

TABLE 1. Distribution of patient and slice numbers in each group.

Fold	1	2	3	4	5	6	7	8	9	10
Patient Number	25	25	25	25	25	25	25	25	25	26
Total Slices	426	425	427	430	430	429	428	431	425	443

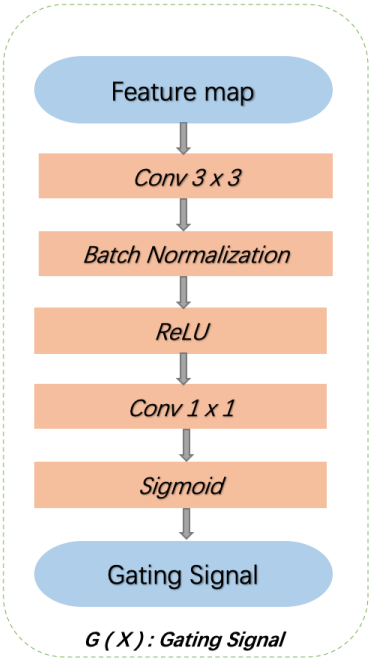


FIGURE 4. The structure of gating signal unit. This unit includes a 3 × 3 convolutional layer, a BN layer, a ReLU, a 1 × 1 convolutional layer, and a sigmoid function used to produce the gating signal.

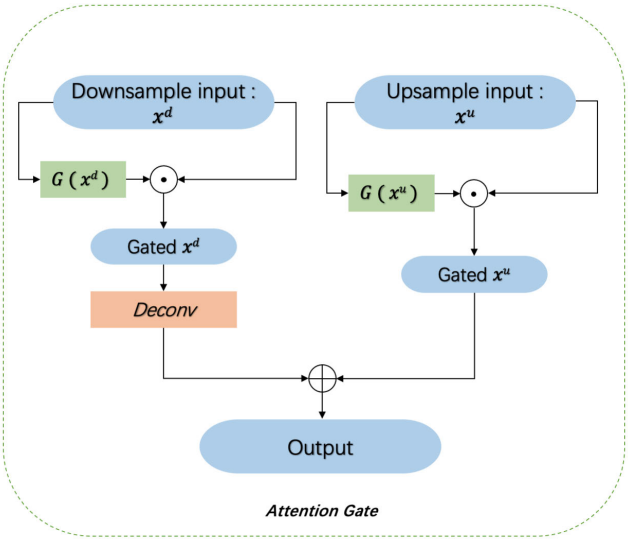


FIGURE 5. The workflow of our attention gate. The attention gate was responsible for synthesizing feature maps from different parts.

contrast-enhanced agent at a dose of 0.1 mmol/kg. The CE-T1W ([TE]:1.9-2.5 ms and [TR]:185-215 ms) in the coronal and transverse planes was obtained after the injection of the contrast agent. The matrix size was 512 × 512, and the in-plane resolution was 0.468-0.523 mm.

The patients were randomly divided into ten groups. Our models were validated with one group, and the other nine groups were used as the training dataset. The distribution of the numbers of patients in the ten groups is shown in Table 1.

To verify the effectiveness of our model, we performed a 10-fold cross-validation experiment. A baseline model was designed based on our proposed model without a T-channel. Then, we evaluated our T-channel model and baseline model by 10-fold cross-validation. Additionally, we verified the effect of the attention module and T-channel module in comparing different module experiments.

B. IMPLEMENTATION AND EVALUATION

In our 10-fold cross-validation experiment, we randomly divided all the patients into ten groups. Each model was validated with one group, and the other nine groups were used as the training dataset. The network was implemented in PyTorch [30] and trained on two NVIDIA Geforce GTX 1080 Ti GPUs for 600 epochs. An Adam [31] optimizer was applied with $\beta 1 = 0.9$ and $\beta 2 = 0.999$, and the learning rate, which was initially set to 1e-4, was reduced to 1e-5 after 400 epochs. In our attention experiment, we utilized the same training and validation dataset as fold1, which is shown in Table 1. A total of 600 epochs were trained for this experiment, and the parameters were the same as those in the 10-fold cross-validation experiment.

We used the validation dataset to evaluate the segmentation performance of all the models by calculating the dice similarity coefficient (DSC) [32] and average symmetric surface distance (ASSD). Specifically, we let A and B represent the ground truth and the prediction mask, respectively. Then, the DSC and ASSD are computed by:

$$DSC = \frac{2\,|A \cap B|}{(|A| + |B|)} \tag{3}$$

where $|\cdot|$ denotes the number of 1s in the ground truth or prediction mask and $|A \cap \acute{E}B|$ indicates the number of 1s shared by A and B. A larger DSC indicates a more accurate result.

$$ASSD = \frac{1}{2}\left\{\frac{\sum_{a\in\hat{E}A} min_{b\in\hat{E}B} d\,(a,b)}{|A|} + \frac{\sum_{b\in\hat{E}B} min_{a\in\hat{E}A} d\,(b,a)}{|B|}\right\} \tag{4}$$

TABLE 2. The average DSC and ASSD of each fold between our T-channel model and without T-channel model.

Model		1	2	3	4	5	6	7	8	9	10	Ave
T-channel	DSC	0.845	0.863	0.833	0.847	0.841	0.820	0.844	0.843	0.833	0.845	0.841
	ASSD (mm)	0.533	0.594	0.679	0.709	1.201	1.020	0.609	0.816	0.658	0.647	0.747
Without T-channel	DSC	0.815	0.847	0.800	0.847	0.824	0.815	0.814	0.804	0.810	0.807	0.818
	ASSD (mm)	0.830	0.813	0.974	0.718	0.941	1.070	0.875	0.874	0.762	0.915	0.877

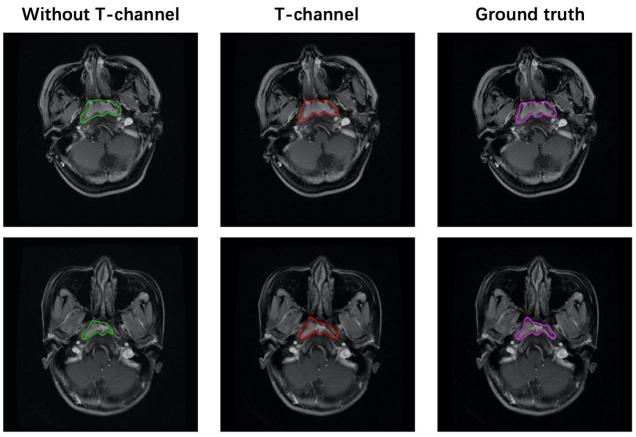


FIGURE 6. Qualitative results of our proposed method. For each sub-figure, the left column indicates the results with T-channel. The middle column, as comparative results, shows the results without T-channel. And the right column represents the ground truth.

where A and B represent the surface voxels of the ground truth and the predicted segmentation results, respectively, and $d(a, b)$ represents the Euclidean distance between a and b. A smaller ASSD denotes a more accurate result.

V. RESULTS

A. TENFOLD CROSS-VALIDATION EXPERIMENT

As shown in Table 2, the average DSC and ASSD of the T-channel models were 0.841 and 0.747 mm, respectively, in the 10-fold cross-validation experiment. The average DSC and ASSD at our model without T-channel were 0.818 and 0.877 mm, respectively. Our proposed model performed better than the model without T-channel, which indicated the T-channel's effectiveness. Figure 6 shows the qualitative results of our experiments. For each sub-figure, the left column, as comparative results, shows the results without T-channel. The middle column indicates the results with T-channel. And the right column represents the ground truth.

B. COMPARISON OF DIFFERENT MODULES

Table 3 shows comparisons of the segmentation performance among the different modules in our experiments. The U-Net model's DSC and ASSD reached 0.811 and 0.830 mm, respectively. Next, we evaluated the effects of the attention module and the T-channel. The U-Net and attention module combined model ultimately achieved a DSC of 0.815 and an ASSD of 0.778 mm, while U-Net with T-channel performed better, with DSC and ASSD values of 0.839 and 0.614 mm, respectively. Finally, we combined the attention and T-channel modules and obtained our proposed model, with a DSC of 0.845 and an ASSD of 0.533 mm, performing better than using a single module alone.

TABLE 3. Comparisons of segmentation performance for different models (95% CI).

Model	DSC	ASSD (mm)
U-Net	0.811 (0.719, 0.882)	0.830 (0.212, 1.939)
U-Net + Attention	0.815 (0.707, 0.892)	0.778 (0.213, 1.873)
U-Net + T-channel	0.839 (0.775, 0.897)	0.614 (0.187, 1.703)
U-Net + Attention + T-channel	**0.845 (0.791, 0.897)**	**0.533 (0.174, 1.254)**

C. COMPARISON WITH THE START-OF-THE-ART METHODS

We compared our model with four state-of-the-art methods:

(1) PSPNet [33] that uses Resnet50 [18] as the backbone; (2) SegNet [34]; (3) U-Net [7] and a variant of it that uses DenseNet-201 as the backbone; (4) DeepLabv3+ [35] that uses Resnet101 as the backbone. We trained all these networks with the same training and validation dataset as fold1's dataset. Quantitative comparison results of these methods are presented in Table 4. It shows that all state-of-the-art methods have good performance in terms of Dice score and ASSD. Our approach yielded the best results on both dice and ASSD with a dice score of 0.845 and an ASSD of 0.533 mm, which is considerably improved compared with the other methods.

VI. DISCUSSION

We proposed an automated NPC segmentation method based on the combination of clinical diagnosis information and a CNN. To address NPC segmentation's difficulty, we convert

TABLE 4. Quantitative evaluations of the state-of-the-art methods and our proposed network for NPC segmentation (95% CI).

Network	DSC	ASSD (mm)
PSPNet [33]	0.737 (0.625, 0.822)	0.875 (0.429, 1.587)
SegNet [34]	0.785 (0.713, 0.870)	0.744 (0.259, 1.974)
DeepLabV3+ [35]	0.803 (0.683, 0.884)	0.668 (0.226, 1.488)
U-Net [7]	0.811 (0.719, 0.882)	0.830 (0.212, 1.939)
DenseUnet	0.824 (0.715, 0.889)	0.604 (0.212, 1.411)
Proposed	**0.845 (0.791, 0.897)**	**0.533 (0.174, 1.254)**

the T-staging information into an additional image channel and concatenate it with different types of MR images, which provides more reliable staging information for NPC segmentation and improves the accuracy of tumor segmentation. Besides, we added an attention mechanism to the model to retain the image's essential information, further improving our model's accuracy. We achieved better performance with our model, as shown· in Table 2 and Figure 6. The performance on the validation dataset proved the robustness of our model.

As shown in Table 2, in our 10-fold cross-validation experiment, there were ten models in total, of which the model using our proposed T-channel module performed better. The T-channel model significantly improved the average DSC for fold3, fold8, and fold10, which was more than 0.030 higher than that of the model without T-channel. The other models also achieved higher results than those without T-channel. In the end, we achieved average DSCs of 0.841 (T-channel model) and 0.815 (without T-channel model). The T-channel module we proposed effectively improved the delineation performance, and ultimately, the average DSC increased by 0.023, and the ASSD decreased by 0.130 mm.

Table 3 shows our comparative experiments for verifying the effectiveness of our attention module and T-channel module. We used the fold1 dataset in the 10-fold cross-validation for training, controlling all the training parameters to be the same. When using U-Net, we obtained an average DSC of 0.811 (95% CI = 0.719 to 0.882) and an ASSD of 0.830 mm (95% CI = 0.212 mm to 1.939 mm). Then, we combined U-Net with the attention module and T-channel module separately. When the attention module was used alone, the average DSC and ASSD reached 0.815 (95% CI = 0.707 to 0.892) and 0.778 mm (95% CI = 0.213 mm to 1.873 mm), respectively. The average DSC was 0.004 higher than that of the original U-Net, and the ASSD declined by 0.052 mm. When the T-channel module was used alone, the model's performance was significantly improved compared to that of the original U-Net. The DSC reached 0.839 (95% CI = 0.775 to 0.897), the ASSD reached 0.614 mm (95% CI = 0.187 mm to 1.703 mm), the DSC was 0.028 higher than U-Net, and the ASSD was 0.216 mm lower than U-Net. When we combined these two modules, we obtained better performance than when using one of the modules alone and ultimately obtained a DSC of 0.845 (95% CI = 0.791 to 0.897) and an ASSD of 0.533 mm (95% CI = 0.174 mm to 1.254 mm). Figure 7 shows a scatterplot of the DSCs of U-Net and the other models, which show better results than the basic U-Net. In Figure 8, we can see the effect of each module in improving segmentation performance. The model we proposed combining the T-channel module and the attention module obtained the best results among all the models and was also more stable than the others.

From the experimental results, we drew the following two conclusions: (a) both the attention module and the T-channel

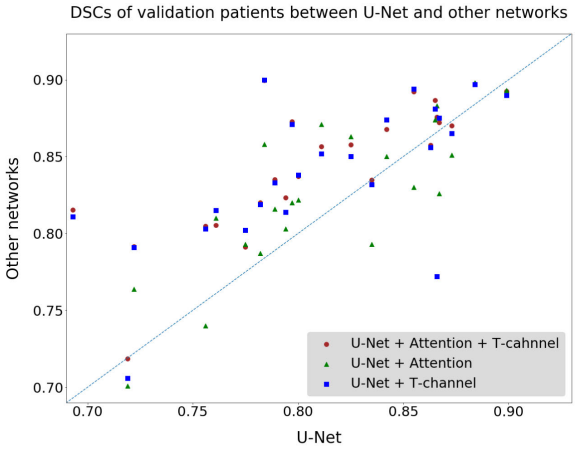


FIGURE 7. A scatterplot that shows the differences between U-Net and other networks.

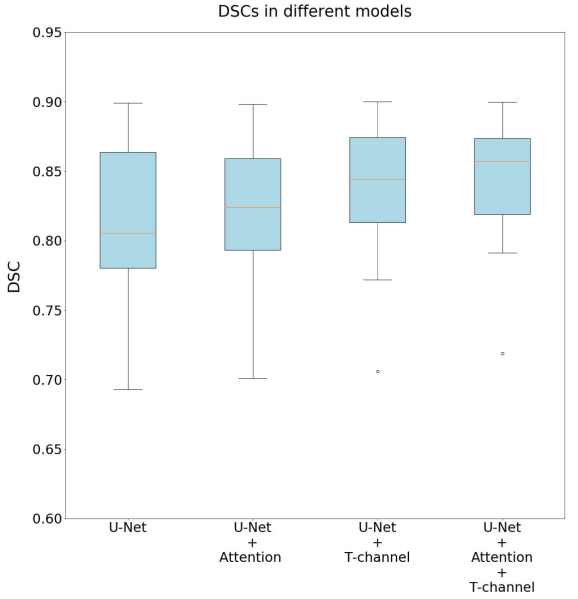


FIGURE 8. A boxplot showing the DSCs of different models. The model with attention gate and T-channel exclaims less variance and a higher average DSC.

module can improve the performance of the model, and the combination of these components can further improve the accuracy of delineation; (b) the improvement in the model from the T-channel module is more evident than that from the attention module, proving the importance and necessity of T-channel in our study. The comparison results between our model and other methods are shown in Table 4. We used 251 patients to train and evaluate our experiments to ensure that our results were reliable. Among the studies shown in Table 4, our method has the best results on dice and ASSD. Our research has the following limitations: (a) due to the GPU memory limitation and to maintain the original image without compressing it, we ultimately set the batch size to 6 in our experiments; a larger batch size may help improve the performance. Using group normalization [36] may solve this problem and increase accuracy. (b) We trained a total of 20 models in T-channel model and baseline model in the 10-fold cross-validation experiment. To reduce the training time, we set the number of epochs to 600 and obtained a convergence result. A larger number of epochs may improve the model's performance. (c) The T-channel occupied only one channel of the input volume; compared with the 15 channels of the original image, the proportion was tiny. Increasing the proportion of T-channel may improve the performance of the model. In this study, we proposed and evaluated an accurate and useful automatic NPC segmentation method based on the combination of clinical diagnosis information and CNN.

T-staging information indicates the extent of tumor invasion. We employ different pixel values to represent each stage, which is equivalent to setting a specific background color for each stage. The CNN can distinguish the stage of the patient by the background color, thereby further promoting the precision of segmentation.

Although U-Net has been applied widely in tumor segmentation tasks, no research has added T-staging information into a CNN model. We proposed this method for the first time for the NPC's automatic segmentation and achieved more reliable and better performance than other methods. With the combination of the T-channel module and attention module, we successfully improved NPC tumor delineation performance. The 10-fold cross-validation results showed that our proposed method displayed better performance with T-channel. Future studies may aim to improve segmentation accuracy with more kinds of clinical diagnosis information.

VII. CONCLUSION

Our proposed T-staging network performs better than a network using image information only, when using the same dataset under the same test conditions. The unique T-channel effectively utilizes T-staging information to improve the result. With the combination of the T-channel module and the attention module, we significantly improved NPC tumor delineation performance.

ETHICS APPROVAL

This retrospective study was approved by the Fudan University Shanghai Cancer Center Institutional Review Board and all methods were performed in accordance with the guidelines and regulations of this ethics board.

REFERENCES

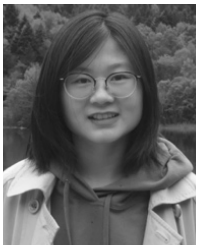
[1] P. P. Claudio and D. A. Denning, "Nasopharyngeal carcinoma," *Lancet*, vol. 365, no. 9476, pp. 2041–2054, 2011.
[2] A. T. C. Chan, "Nasopharyngeal carcinoma," *Ann. Oncol.*, vol. 21, no. 7, pp. 308–312, 2010.
[3] S. H. Ng, T. C. Chang, S. F. Ko, P. S. Yen, Y. L. Wan, L. M. Tang, and M. H. Tsai, "Nasopharyngeal carcinoma: MRI and CT assessment," *Neuroradiology*, vol. 39, no. 10, pp. 741–746, Oct. 1997.
[4] Z. Chang, "Will AI improve tumor delineation accuracy for radiation therapy?" *Radiology*, vol. 291, no. 3, pp. 687–688, Jun. 2019.
[5] K. Sakata, M. Hareyama, M. Tamakawa, A. Oouchi, M. Sido, H. Nagakura, H. Akiba, K. Koito, T. Himi, and K. Asakura, "Prognostic factors of nasopharynx tumors investigated by MR imaging and the value of MR imaging in the newly published TNM staging," *Int. J. Radiat. Oncol., Biol., Phys.*, vol. 43, no. 2, pp. 273–278, 1999.
[6] The Union for International Cancer Control. *TNM History, Evolution and Milestones*. [Online]. Available: http://www.uicc.org/sites/main/files/private/History_Evolution_Milestones_0.pdf
[7] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*, 2015, pp. 234–241.
[8] M. Jolly, "Fully automatic left ventricle segmentation in cardiac cine MR images using registration and minimum surfaces," *MIDAS J.-Cardiac MR Left Ventricle Segmentation Challenge*, vol. 4, p. 59, Aug. 2009.
[9] F. Billet, M. Sermesant, H. Delingette, and N. Ayache, "Cardiac motion recovery and boundary conditions estimation by coupling an electromechanical model and cine-MRI data," in *Proc. FIMH*, 2009, pp. 376–385.
[10] X. Lin, B. Cowan, and A. Young, "Model-based graph cut method for segmentation of the left ventricle," in *Proc. IEEE Eng. Med. Biol. 27th Annu. Conf.*, Jan. 2005, pp. 3059–3062.
[11] C. Petitjean and J. Dacher, "A review of segmentation methods in short axis cardiac MR images," *Med. Image Anal.*, vol. 15, no. 2, pp. 169–184, 2011.
[12] H. Zhang, A. Wahle, R. K. Johnson, T. D. Scholz, and M. Sonka, "4-D cardiac MR image analysis: Left and right ventricular morphology and function," *IEEE Trans. Med. Imag.*, vol. 29, no. 2, pp. 350–364, Feb. 2010.
[13] T. Rohlfing, R. Brandt, R. Menzel, D. B. Russakoff, and C. R. Maurer, "Quo vadis, atlas-based segmentation?" in *The Handbook of Medical Image Analysis*, vol. 3. 2005, ch. 11, pp. 435–486.
[14] J. Lötjönen, S. Kivistö, J. Koikkalainen, D. Smutek, and K. Lauerma, "Statistical shape model of atria, ventricles and epicardium from short- and long-axis MR images," *Med. Image Anal.*, vol. 8, no. 3, pp. 371–386, 2004.
[15] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.
[16] V. Kumar, J. Webb, A. Gregory, M. Denis, D. D. Meixner, M. Bayat, D. H. Whaley, N. Fatemi, and A. Alizad, "Automated and real-time segmentation of suspicious breast masses using convolutional neural network," *PLoS ONE*, vol. 13, no. 5, pp. 1569–1571, 2018.
[17] C. Yang, X. Guo, T. Wang, Y. Yang, N. Ji, D. Li, H. Lv, and T. Ma, "Automatic brain tumor segmentation method based on modified convolutional neural network," in *Proc. 41st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2019, pp. 998–1001.
[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[19] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 4th Int. Conf. 3D Vis. (3DV)*, Oct. 2016, pp. 565–571.
[20] D. Nie, L. Wang, E. Adeli, C. Lao, W. Lin, and D. Shen, "3-D fully convolutional networks for multimodal isointense infant brain image segmentation," *IEEE Trans. Cybern.*, vol. 49, no. 3, pp. 1123–1136, Mar. 2019.
[21] Z. Wang, N. Zou, D. Shen, and S. Ji, "Non-local U-Net for biomedical image segmentation," in *Proc. AAAI*, 2020, pp. 6315–6322.

[22] Q. Li, Y. Xu, Z. Chen, D. Liu, S.-T. Feng, M. Law, Y. Ye, and B. Huang, "Tumor segmentation in contrast-enhanced magnetic resonance imaging for nasopharyngeal carcinoma: Deep learning with convolutional neural network," *BioMed Res. Int.*, vol. 2018, pp. 1–7, Oct. 2018.
[23] L. Zhao, Z. Lu, J. Jiang, Y. Zhou, Y. Wu, and Q. Feng, "Automatic nasopharyngeal carcinoma segmentation using fully convolutional networks with auxiliary paths on dual-modality PET-CT images," *J. Digit. Imag.*, vol. 32, no. 3, pp. 462–470, Jun. 2019.
[24] F. Guo, C. Shi, X. Li, X. Wu, J. Zhou, and J. Lv, "Image segmentation of nasopharyngeal carcinoma using 3D CNN with long-range skip connection and multi-scale feature pyramid," *Soft Comput.*, vol. 24, no. 16, pp. 12671–12680, Aug. 2020.
[25] Y. Ye, Z. Cai, B. Huang, Y. He, P. Zeng, G. Zou, W. Deng, H. Chen, and B. Huang, "Fully-automated segmentation of nasopharyngeal carcinoma on dual-sequence MRI using convolutional neural networks," *Frontiers Oncol.*, vol. 10, p. 166, Feb. 2020.
[26] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015, *arXiv:1502.03167*. [Online]. Available: http://arxiv.org/abs/1502.03167
[27] M. Biehl and H. Schwarze, "Learning by on-line gradient descent," *J. Phys. A, Math. Gen.*, vol. 28, no. 3, pp. 643–656, Feb. 1995.
[28] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, "Attention to scale: Scale-aware semantic image segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3640–3649.
[29] J. Schlemper, O. Oktay, M. Schaap, M. Heinrich, B. Kainz, B. Glocker, and D. Rueckert, "Attention gated networks: Learning to leverage salient regions in medical images," *Med. Image Anal.*, vol. 53, pp. 197–207, Apr. 2019.
[30] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. Devito, Z. M. Lin, A. Desmaison, L. Antiga, and A. Leter, "Automatic differentiation in PyTorch," Tech. Rep., 2017.
[31] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980
[32] W. R. Crum, O. Camara, and D. L. G. Hill, "Generalized overlap measures for evaluation and validation in medical image analysis," *IEEE Trans. Med. Imag.*, vol. 25, no. 11, pp. 1451–1461, Nov. 2006.
[33] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6230–6239.
[34] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.
[35] L. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 801–808.
[36] Y. Wu and K. He, "Group normalization," *Int. J. Comput. Vis.*, vol. 128, no. 3, pp. 742–755, Mar. 2020.

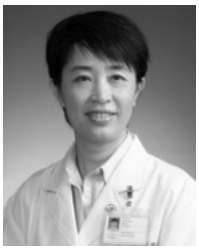
MINGWEI CAI received the B.S. degree in bioinformatics from Soochow University, Suzhou, China, in 2019. He is currently pursuing the M.S. degree in biomedical engineering from Fudan University, Shanghai, China. His research interests include medical image processing and deep learning.

JIAZHOU WANG is currently pursuing the Ph.D. degree from Fudan University Shanghai Cancer Center, Shanghai, China. His research interests include artificial intelligence and medical image processing.

QING YANG is currently pursuing the Ph.D. degree in radiation therapy from Fudan University, Shanghai, China. Her research interest includes radiotherapy for head and neck tumors.



YING GUO received the B.S. degree in physics from Fudan University, Shanghai, China, in 2017, where she is currently pursuing the M.S. degree in biomedical engineering. Her research interests include medical image processing and auto segmentation.



ZHEN ZHANG received the B.S. degree from the Shanghai Medical College, Fudan University, Shanghai, China, in 1988, and the Ph.D. degree in oncology from Fudan University Shanghai Cancer Center, Shanghai, in 2007. She is currently a Professor with the Department of Radiation Oncology, Fudan University Shanghai Cancer Center. Her research interests include new technology of radiotherapy and the application of image-guided radiotherapy.

HONGMEI YING received the B.S. and Ph.D. degrees from the Shanghai Medical College, Fudan University, Shanghai, China, in 1993 and 1998, respectively. She is currently a Chief Physician with the Department of Radiation Oncology, Fudan University Shanghai Cancer Center, Shanghai. Her research interests include clinical treatment of nasopharyngeal carcinoma, radiotherapy, and comprehensive treatment of head and neck tumors.

WEIGANG HU received the M.S. degree in biomedical engineering from Tsinghua University, Beijing, China, and the Ph.D. degree from Fudan University, Shanghai, China, in 2015. He is currently an Associate Research Fellow with the Department of Radiation Oncology, Fudan University Shanghai Cancer Center, Shanghai. His research interests include adaptive radiation therapy and auto planning.

CHAOSU HU received the B.S. degree from the Jiangxi Medical College, Nanchang University, Nanchang, China, in 1984, and the Ph.D. degree from the Shanghai Medical College, Fudan University, Shanghai, China, in 1992. He is currently a Professor with the Department of Radiation Oncology, Fudan University Shanghai Cancer Center, Shanghai. His research interests include radiotherapy of nasopharyngeal carcinoma, laryngocarcinoma, and cerebral tumor.