Second-Course Esophageal Gross Tumor Volume Segmentation in CT with Prior Anatomical and Radiotherapy Information

Yihua Sun[1], Hee Guan Khor[1], Sijuan Huang[2], Qi Chen[3], Shaobin Wang[1,3], Xin Yang[2(✉)], and Hongen Liao[1(✉)]

[1] Department of Biomedical Engineering, School of Medicine, Tsinghua University, Beijing, China
wsb20@mails.tsinghua.edu.cn, liao@tsinghua.edu.cn

[2] Sun Yat-Sen University Cancer Center, State Key Laboratory of Oncology in South China, Collaborative Innovation Center for Cancer Medicine, Guangdong Key Laboratory of Nasopharyngeal Carcinoma Diagnosis and Therapy, Guangzhou, China
yangxin@sysucc.org.cn

[3] MedMind Technology Co., Ltd., Beijing, China

Abstract. Esophageal cancer is a significant global health concern, and radiotherapy (RT) is a common treatment option. Accurate delineation of the gross tumor volume (GTV) is essential for optimal treatment outcomes. In clinical practice, patients may undergo a second round of RT to achieve complete tumor control when the first course of treatment fails to eradicate cancer completely. However, manual delineation is labor-intensive, and automatic segmentation of esophageal GTV is difficult due to the ambiguous boundary of the tumor. Detailed tumor information naturally exists in the previous stage, however the correlation between the first and second course RT is rarely explored. In this study, we first reveal the domain gap between the first and second course RT, and aim to improve the accuracy of GTV delineation in the second course RT by incorporating prior information from the first course. We propose a novel prior **A**natomy and **RT** information enhanced **S**econd-course **E**sophageal **G**TV segmentation network (**ARTSEG**). A region-preserving attention module (RAM) is designed to understand the long-range prior knowledge of the esophageal structure, while preserving the regional patterns. Sparsely labeled medical images for various isolated tasks necessitate efficient utilization of knowledge from relevant datasets and tasks. To achieve this, we train our network in an information-querying manner. ARTSEG incorporates various prior knowledge, including: 1) Tumor volume variation between first and second RT courses, 2) Cancer cell proliferation, and 3) Reliance of GTV on esophageal anatomy. Extensive quantitative and qualitative experiments validate our designs.

H. Liao and X. Yang are the co-corresponding authors.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43990-2_48.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14226, pp. 511–520, 2023.
https://doi.org/10.1007/978-3-031-43990-2_48

Keywords: Second course radiotherapy · Esophageal gross tumor volume · Data efficient learning · Prior anatomical information · Attention

1 Introduction

Esophageal cancer is a significant contributor to cancer-related deaths globally [3,15]. One effective treatment option is radiotherapy (RT), which utilizes high-energy radiation to target cancerous cells [4]. To ensure optimal treatment outcomes, both the cancerous region and the adjacent organ-at-risk (OAR) must be accurately delineated, to focus the high-energy radiation solely on the cancerous area while protecting the OARs from any harm. Gross tumor volume (GTV) represents the area of the tumor that can be identified with a high degree of certainty and is of paramount importance in clinical practice.

In the clinical setting, patients may undergo a second round of RT treatment to achieve complete tumor control when initial treatment fails to completely eradicate cancer [16]. However, the precise delineation of the GTV is labor-intensive, and is restricted to specialized hospitals with highly skilled RT experts. The automatic identification of the esophagus presents inherent challenges due to its elongated soft structure and ambiguous boundaries between it and adjacent organs [12]. Moreover, the automatic delineation of the GTV in the esophagus poses a significant difficulty, primarily attributable to the low contrast between the esophageal GTV and the neighboring tissue, as well as the limited datasets.

Recently, advances in deep learning [21] have promoted research in automatic esophageal GTV segmentation from computed tomography (CT) [18,19]. Since the task is challenging, Jin et al. [9,10] improve the segmentation accuracy by incorporating additional information from paired positron emission tomography (PET). Nevertheless, such approaches require several imaging modalities, which can be both costly and time-consuming, while disregarding any knowledge from previous treatment or anatomical understanding. Moreover, the correlation between the first and second courses of RT is rarely investigated, where detailed prior tumor information naturally exists in the previous RT planning.

In this paper, we present a comprehensive study on accurate GTV delineation for the second course RT. We proposed a novel prior **A**natomy and **RT** information enhanced **S**econd-course **E**sophageal **G**TV segmentation network (**ARTSEG**). A region-preserving attention module (RAM) is designed to effectively capture the long-range prior knowledge in the esophageal structure, while preserving regional tumor patterns. To the best of our knowledge, we are the first to reveal the domain gap between the first and second courses for GTV segmentation, and explicitly leverage prior information from the first course to improve GTV segmentation performance in the second course.

The medical images are labeled sparsely, which are isolated by different tasks [20]. Meanwhile, an ideal method for automatic esophageal GTV segmentation in the second course of RT should consider three key aspects: 1) Changes in tumor volume after the first course of RT, 2) The proliferation of cancerous cells from a tumor to neighboring healthy cells, and 3) The anatomical-dependent

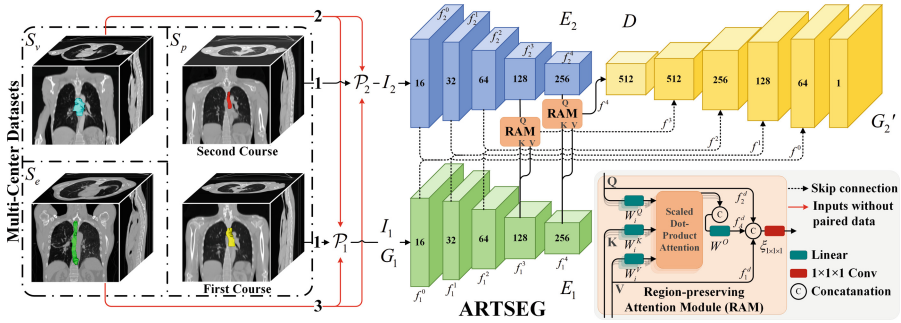

Fig. 1. Our training approach leverages multi-center datasets containing relevant annotations, that challenges the network to retrieve information from E_1 using the features from E_2. The decoder D utilizes the prior knowledge obtained from I_1 and G_1 to generate the mask prediction. Our training strategy leverages three datasets that introduce prior knowledge to the network of the following three key aspects: 1) Tumor volume variation, 2) Cancer cell proliferation, and 3) Reliance of GTV on esophageal anatomy.

nature of GTV on esophageal locations. To achieve this, we efficiently exploit knowledge from multi-center datasets that are not tailored for second-course GTV segmentation. Our training strategy does not specific to any tasks but challenges the network to retrieve information from another encoder with augmented inputs, which enables the network to learn from the above three aspects. Extensive quantitative and qualitative experiments validate our designs.

2 Network Architecture

In the first course of RT, a CT image denoted as I_1 is utilized to manually delineate the esophageal GTV, G_1. During the second course of RT, a CT image I_2 of the same patient is acquired. However, I_2 is not aligned with I_1 due to soft tissue movement and changes in tumor volume that occurred during the first course of treatment. Both images $I_{1/2}$ have the spatial shape of $H \times W \times D$.

Our objective is to predict the esophageal GTV G_2 of the second course. It would be advantageous to leverage insights from the first course, as it comprises comprehensive information pertaining to the tumor in its preceding phase. Therefore, the input to encoder E_1 consists of the concatenation of I_1 and G_1 to encode the prior information (features f_1^d) from the first course, while encoder E_2 embeds both low- and high-level features f_2^d of the local pattern of I_2 (Fig. 1),

$$f_1^d = E_1(I_1, G_1),\ f_2^d = E_2(I_2),\ d = 0, 1, 2, 3, 4 \tag{1}$$

where the spatial shape of $f_{1/2}^d$ is $\frac{H}{2^d} \times \frac{W}{2^d} \times \frac{D}{2^d}$, with 2^{d+4} channels.

Region-Preserving Attention Module. To effectively learn the prior knowledge in the elongated esophagus, we design a region-preserving attention module

(RAM), as shown in Fig. 1. The multi-head attention (MHA) [17] is employed to gather long-range informative values in f_1^d with f_2^d as queries and f_1^d as keys. The features $f_{1/2}^d$ are reshaped to $\frac{HWD}{2^{3d}} \times C$ before passed to the MHA, where C is the channel dimension. The attentive features f_A^d can be formulated as:

$$f_A^d = MHA(Q,\, K,\, V) = MHA(f_2^d,\, f_1^d,\, f_1^d),\, d = 3, 4. \tag{2}$$

Since MHA perturbs the positional information, we preserve the tumor local patterns by concatenating original features to the attentive features at the channel dimension, followed by a $1 \times 1 \times 1$ bottleneck convolution $\xi_{1\times1\times1}$ to squeeze the channel features (named as RAM), as shown in the following equations,

$$f^d = \begin{cases} Concat(f_1^d,\, f_2^d), & d = 0, 1, 2, \\ \xi_{1\times1\times1}\left(Concat(f_1^d,\, f_2^d,\, f_A^d)\right), & d = 3, 4, \end{cases} \tag{3}$$

where the lower-level features from both encoders are fused by concatenation. The decoder D generates a probabilistic prediction $G_2' = D(f^0, \cdots, f^4)$ with skip connections (Fig. 1). We utilize the 3D Dice [14] loss function, $\mathcal{L}_{DICE}(G_2', G_2)$.

3 Training Strategy

The network should learn from three aspects: 1) **Tumor volume variation**: the structural changes of the tumor from the first to the second course; 2) **Cancer cell proliferation**: The tumor in esophageal cancer tends to infiltrate into the adjacent tissue; 3) **Reliance of GTV on esophageal anatomy**: The anatomical dependency between esophageal GTV and the position of the esophagus.

Medical images are sparsely labeled which are isolated by different tasks [20], and are often inadequate. In this study, we use a paired first-second course GTV dataset S_p, an unpaired GTV dataset S_v, and a public esophagus dataset S_e.

In order to fully leverage both public and private datasets, the training objective should not be specific to any tasks. Here, we denote G_1/G_2 as prior/target annotations respectively, which are not limited only to the GTV areas. As shown in Fig. 1, our strategy is to challenge the network to retrieve information from augmented inputs in E_1 using the features from E_2, which can incorporate a wide range of datasets that are not tailored for second-course GTV segmentation.

3.1 Tumor Volume Variation

The differences in tumor volume between the first and second courses following an RT treatment can have a negative impact on the state-of-the-art (SOTA) learning-based techniques, which will be discussed in Sect. 4.2. To adequately monitor changes in tumor volume and integrate information from the initial course into the subsequent course, a paired first-second courses dataset $S_p = \{i_p^1, i_p^2, g_p^1; g_p^2\}$ is necessary for training. In S_p, i_p^1 and i_p^2 are the first and second course CT images, while g_p^1 and g_p^2 are the corresponding GTV annotations.

3.2 Cancer Cell Proliferation

The paired dataset S_p for the first and second courses is limited, whereas an unpaired GTV dataset $S_v = \{i_v; g_v\}$ can be easily obtained in a standard clinical workflow with a substantial amount. S_v lacks its counterpart for the second course, in which i_v/g_v are the CT image and the corresponding annotation for GTV. To address this, we apply two distinct randomized augmentations, $\mathcal{P}_1, \mathcal{P}_2$, to mimic the unregistered issue of the first and second course CT. The transformed data is feed into the encoders $E_{1/2}$ as shown in the following equations:

$$I_1, G_1, I_2, G_2 = \begin{cases} \mathcal{P}_1(i_p^1), \mathcal{P}_1(g_p^1), \mathcal{P}_2(i_p^2), \mathcal{P}_2(g_p^2), & \text{when } i_p^1, i_p^2, g_p^1, g_p^2 \in S_p, \\ \mathcal{P}_1(i_v), \mathcal{P}_1(g_v), \mathcal{P}_2(i_v), \mathcal{P}_2(g_v), & \text{when } i_v, g_v \in S_v, \\ \mathcal{P}_1(i_e), \mathcal{P}_1(g_e), \mathcal{P}_2(i_e), \mathcal{P}_2(g_e), & \text{when } i_e, g_e \in S_e. \end{cases} \tag{4}$$

The esophageal tumor can proliferate with varying morphologies into the surrounding tissues. Although not paired, S_v contains valuable information about the tumor. Challenging the network to query information within GTV will enhance the capacity to retrieve pertinent information for the tumor positions.

3.3 Reliance of GTV on Esophageal Anatomy

To make full use of the datasets of relevant tasks, we incorporate a public esophagus segmentation dataset, denoted as $S_e = \{i_e; g_e\}$, where i_e/g_e represent the CT images and corresponding annotations of the esophagus structure. By augmenting the data as described in Eq. (4), S_e challenges the network to extract information from the entire esophagus, which enhances the network's embedding space with anatomical prior knowledge of the esophagus. Similarly, data from the paired S_p is also augmented by $\mathcal{P}_{1/2}$ to increase the network's robustness.

In summary, our training strategy is not dataset-specific or target-specific, thus allowing the integration of prior knowledge from multi-center esophageal GTV-related datasets, which effectively improves the network's ability to retrieve information for the second course from the three key aspects stated in Sect. 3.

4 Experiments

4.1 Experimental Setup

Datasets. The paired first-second course dataset, S_p, is collected from Sun Yat-Sen University Cancer Center (Ethics Approval Number: B2023-107-01), comprising paired CT scans of 69 distinct patients from South China. We collected the GTV dataset S_v from MedMind Technology Co., Ltd., which has CT scans from 179 patients. For both S_p and S_v, physicians annotated the esophageal cancer GTV in each CT. The GTV volume statistics (cm^3, mean $\pm$ std.) in S_v is 40.60 $\pm$ 29.75, and is 83.70 $\pm$ 55.97/71.66 $\pm$ 49.36 for the first/second course RT in S_p respectively. Additionally, we collect S_e from SegTHOR [12], consisting of CT scans and esophagus annotations from 40 patients who did not

Table 1. The results suggest a domain gap between the first and second courses, which indicates increased difficulty in GTV segmentation for the second course. Asterisks indicate p-value < 0.05 for the performance gap between the first and second course.

Methods	First Course				Second Course			
	DSC (%) ↑		ASD (mm) ↓		DSC (%) ↑		ASD (mm) ↓	
	mean ± std.	med.	mean ± std.	med.	mean ± std.	med.	mean ± std.	med.
UNETR [7]	59.77 ± 20.24	62.90	10.57 ± 14.66	7.06	53.03 ± 17.62*	55.17	11.29 ± 11.44	8.42
Swin UNETR [6]	60.84 ± 19.74	64.07	10.29 ± 17.78	6.67	57.04 ± 20.16*	60.73	9.76 ± 15.43	6.21
DenseUnet [19]	63.95 ± 18.23	68.11	8.94 ± 13.82	6.04	55.35 ± 18.59*	58.54	9.84 ± 6.91*	8.54
3D U-Net [5]	66.73 ± 17.21	69.86	8.04 ± 16.83	4.19	57.50 ± 19.49*	62.62	9.14 ± 12.03*	6.09

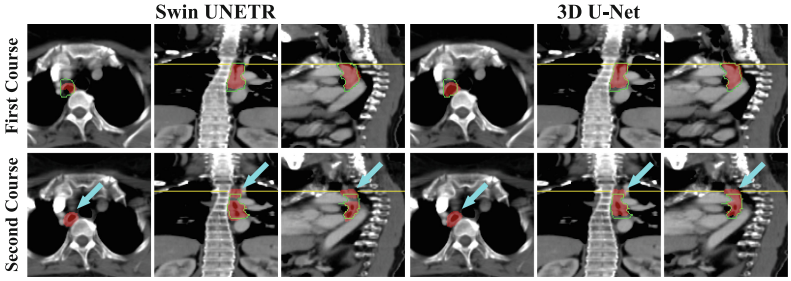

Fig. 2. The domain gap observed between the first and second courses RT. The blue arrows indicate that the methods tend to exhibit false delineation in the second course, suggesting a lack of consideration for tumor changes after the first course. (Yellow: Locations of the transverse planes; Green: GTV ground truth contours; Red: predictions.) (Color figure online)

have esophageal cancer. We randomly split S_p into training and test datasets at the patient-level. The training dataset includes S_v, S_e, and 41 patients from S_p (denoted as S_p^{train}), while the test dataset comprises 28 patients from S_p (denoted as S_p^{test}).

Implementation Details. The CT volumes from the first and second course in S_p are aligned based on the center of the lung mask [8]. The CT volumes are applied with a windowing of $[-100, 300]$ HU, and resampled to 128^3, with a voxel size of $1.2 \times 1.2 \times 3$ mm^3. The augmentations $\mathcal{P}_{1/2}$ involve a combination of random 3D resized cropping, flipping, rotation in the transverse plane, and Gaussian noise. We employ the Adam [11] optimizer with $(\beta_1, \beta_2, lr) = (0.9, 0.999, 0.001)$ for training for 500 epoches. The network is implemented using PyTorch [2] and MONAI [1], and detailed configurations are in the supplementary material. Experiments are performed on an NVIDIA RTX 3090 GPU with 24GB memory.

Performance Metrics. Dice score (DSC), averaged surface distance (ASD) and Hausdorff distance (HSD) are used as metrics for evaluation. The Wilcoxon signed-rank test is used to compare the performance of different methods.

4.2 Domain Gap Between the First and Second Course

As previously mentioned, the volume of the tumors changes after the first course of RT. To demonstrate the presence of a domain gap between the first and second courses, we train SOTA methods with datasets S_p^{train} and S_v, by feeding the data sequentially into the network. We then evaluate the models on S_p^{test}. The results presented in Table 1 indicate a performance gap between GTV segmentation in the first and second courses, with the latter being more challenging. Notably, the paired first-second course dataset S_p^{test} pertains to the same group of patients, thereby ensuring that any performance drop can be attributed solely to differences in courses of RT, rather than variations across different patients.

Figure 2 illustrates the reduction in the GTV area after the initial course of RT, where the transverse plane is taken from the same location relative to the vertebrae (yellow lines). The blue arrows indicate that the networks failed to track these changes and produced false predictions in the second course of RT. This suggests that deep learning-based approaches may not rely solely on the identification of malignant tissue patterns, as doctors do, but rather predict high-risk areas statistically. Therefore, for accurate second-course GTV segmentation, we need to explicitly propagate prior information from the first course using dual encoders in ARTSEG, and incorporate learning about tumor changes.

4.3 Evaluations of Second-Course GTV Segmentation Performance

Combination of Various Datasets. Table 2 presents the information gain derived from multi-center datasets using quantified metrics for segmentation performance. We first utilize a standard ARTSEG (w/o RAM) as an ablation network. When prior information from the first course is explicitly introduced using S_p, ARTSEG outperforms other baselines for GTV segmentation in the second course, which reaches a DSC of 66.73%. However, in Fig. 3, it can be observed that the model failed to accurately track the GTV area along the esophagus (orange arrows) due to the soft and elongated nature of the esophageal tissue, which deforms easily during CT scans performed at different times.

By subsequently incorporating S_e for structural esophagus prior knowledge, the DSC improved to 69.42%. Meanwhile, the esophageal tumor comprises two primary regions, the original part located in the esophagus and the extended part that has invaded the surrounding tissue. As shown in Fig. 3, identifying the tumor proliferation into the surrounding tissue without comprehensive knowledge of tumor morphology can be challenging (blue arrows). To address this, incorporating S_v to comprehensively learn the tumor morphology is required.

When S_v is incorporated for learning tumor proliferation, the DSC improved to 72.64%. We can observe from Case 2 in Fig. 3 that the network has a better understanding of the tumor proliferation with S_v, while it still fails to track the GTV area along the esophagus as pointed by the orange arrow. Therefore, S_v and S_e improve the network from two distinct aspects and are both valuable. Our proposed training strategy fully exploits the datasets S_p, S_v, and S_e, and

Table 2. Quantitative comparison of GTV segmentation performance in the second course. Our proposed ARTSEG+RAM achieved better overall performance, where asterisks indicate ARTSEG+RAM outperforms other methods with p-value < 0.05.

Methods	S_p^{train}	S_v	S_e	DSC (%) ↑		ASD(mm) ↓		HSD(mm) ↓		Inference
				mean ± std.	med.	mean ± std.	med.	mean ± std.	med.	speed(ms)
Swin UNETR§ [6]	✓	✓		57.04 ± 20.16*	60.73	9.76 ± 15.43*	6.21	58.73 ± 46.67*	51.00	59.25
3D U-Net§ [5]	✓	✓		57.50 ± 19.49*	62.62	9.14 ± 12.03*	6.09	63.54 ± 49.59*	48.29	4.01
ARTSEG w/o RAM	✓			66.73 ± 16.20*	71.85	4.45 ± 4.10*	3.10	38.22 ± 27.51*	31.22	9.05
	✓		✓	69.42 ± 11.05*	71.12	3.98 ± 3.03*	2.99	47.89 ± 50.82*	33.04	
	✓	✓		72.64 ± 13.53*	74.64	2.69 ± 1.69	2.23	21.71 ± 11.95	19.69	
ARTSEG w/o RAM	✓	✓	✓	74.54 ± 13.33	**76.84**	2.51 ± 1.85	1.83	27.00 ± 41.79	16.35	
ARTSEG+MHA [17]	✓	✓	✓	74.34 ± 13.27	76.25	2.49 ± 1.62	1.97	27.33 ± 33.25	15.99	12.24
ARTSEG+RAM	✓	✓	✓	**75.26 ± 12.24**	76.40	**2.39 ± 1.57**	**1.73**	**19.75 ± 11.83**	**15.54**	12.60

§Without explicit information from the first-course. (Best methods in Table 1)

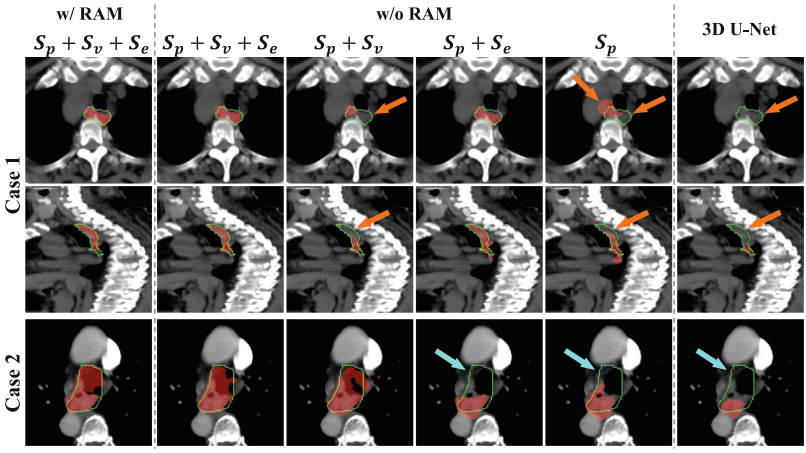

Fig. 3. The impact of different prior knowledge on esophageal tumor detection. Networks with inadequate knowledge of the esophagus may fail in identifying the tumor within the esophagus (orange arrows), whereas a limited understanding of tumor morphology can deteriorate the ability to detect the tumor in the adjacent area (blue arrows). Our proposed approach, encompassing comprehensive prior knowledge, shows superior performance. (Green: GTV ground truth contours; Red: predictions.) (Color figure online)

further improve the DSC to 74.54% by utilizing comprehensive knowledge of both the tumor morphology and esophageal structures.

Region-Preserving Attention Module. Although introducing the esophageal structural prior knowledge using S_e can improve the performance in DSC and ASD (Table 2), the increase in HSD (38.22 to 47.89 mm; 21.71 to 27.00 mm) indicates that there are outliers far from the ground truth boundaries. This may be attributed to the convolution that cannot effectively handle the long-range knowledge of the esophagus structure. The attention mechanism can effectively capture the long-range relationship as shown recently in [13].

However, there is no performance gain with MHA as shown in Table 2, and the HSD further increased to 27.33 mm. We attribute the drawback is due to the location-agnostic nature of the operations in MHA, where the local regional correlations are perturbed.

To tackle the aforementioned problem, we propose RAM which involves the concatenation of the original features with attention outputs, allowing for the preservation of convolution-generated regional tumor patterns while effectively comprehending long-range prior knowledge specific to the esophagus. Finally, our proposed ARTSEG with RAM achieves the best DSC/HSD of 75.26%/19.75 mm, and outperforms its ablations as well as other baselines, as shown in Table 2.

Limitations. For the method's generalizability, analysis of diverse imaging protocols and segmentation backbones are inadequate. Besides, ARTSEG requires more computational resources due to its dual-encoder and attention design.

5 Conclusion

In this paper, we reveal the domain gap between the first and second courses of RT for esophageal GTV segmentation. To improve the accuracy of GTV declination in the second course, we explicitly incorporated the naturally existing prior information from the first course. Besides, to efficiently leverage prior knowledge contained in various medical CT datasets, we train the network in an information-querying manner. We proposed RAM to capture long-range prior knowledge in the esophageal structure, while preserving the regional tumor patterns. Our proposed ARTSEG incorporates prior knowledge of the tumor volume variation, cancer cell proliferation, and reliance of GTV on esophageal anatomy, which enhances the GTV segmentation accuracy in the second course RT. Our future research includes accurate delineation for multiple targets in the second course and knowledge transferring through the time series of multiple courses.

Acknowledgments. Thanks to National Key Research and Development Program of China (2022YFC2405200), National Natural Science Foundation of China (82027807, U22A2051), Beijing Municipal Natural Science Foundation (7212202), Institute for Intelligent Healthcare, Tsinghua University (2022ZLB001), Tsinghua-Foshan Innovation Special Fund (2021THFS0104), Guangdong Esophageal Cancer Institute Science and Technology Program (Q202221, Q202214, M-202016).

References

1. Medical Open Network for Artificial Intelligence (MONAI). https://monai.io/
2. PyTorch. https://pytorch.org/
3. Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A., Jemal, A.: Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: Cancer J. Clin. **68**(6), 394–424 (2018)
4. Burnet, N.G., Thomas, S.J., Burton, K.E., Jefferies, S.J.: Defining the tumour and target volumes for radiotherapy. Cancer Imaging **4**(2), 153–161 (2004)

5. Falk, T., et al.: U-net: deep learning for cell counting, detection, and morphometry. Nat. Methods **16**(1), 67–70 (2019)
6. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin UNETR: swin transformers for semantic segmentation of brain tumors in MRI images. In: Crimi, A., Bakas, S. (eds.) BrainLes 2021. LNCS, vol. 12962, pp. 272–284. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-08999-2_22
7. Hatamizadeh, A., et al.: UNETR: transformers for 3D medical image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 574–584 (2022)
8. Hofmanninger, J., Prayer, F., Pan, J., Röhrich, S., Prosch, H., Langs, G.: Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem. Eur. Radiol. Exp. **4**(1), 1–13 (2020). https://doi.org/10.1186/s41747-020-00173-2
9. Jin, D., et al.: DeepTarget: gross tumor and clinical target volume segmentation in esophageal cancer radiotherapy. Med. Image Anal. **68**, 101909 (2021)
10. Jin, D., et al.: Accurate esophageal gross tumor volume segmentation in PET/CT using two-stream chained 3D deep network fusion. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11765, pp. 182–191. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32245-8_21
11. Kingma, D.P., Ba, J.L.: Adam: a method for stochastic optimization. In: International Conference on Learning Representations (2015)
12. Lambert, Z., Petitjean, C., Dubray, B., Kuan, S.: Segthor: segmentation of thoracic organs at risk in CT images. In: 2020 Tenth International Conference on Image Processing Theory, Tools and Applications (IPTA), pp. 1–6 (2020)
13. Li, J., Chen, J., Tang, Y., Wang, C., Landman, B.A., Zhou, S.K.: Transforming medical imaging with transformers? a comparative review of key properties, current progresses, and future perspectives. Med. Image Anal. **85**, 102762 (2023)
14. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV), pp. 565–571 (2016)
15. Pennathur, A., Gibson, M.K., Jobe, B.A., Luketich, J.D.: Oesophageal carcinoma. Lancet **381**(9864), 400–412 (2013)
16. Van Andel, J.G., et al.: Carcinoma of the esophagus: results of treatment. Ann. Surg. **190**(6), 684–689 (1979)
17. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
18. Yousefi, S., et al.: Esophageal tumor segmentation in CT images using a dilated dense attention Unet (DDAUnet). IEEE Access **9**, 99235–99248 (2021)
19. Yousefi, S., et al.: Esophageal gross tumor volume segmentation using a 3D convolutional neural network. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11073, pp. 343–351. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00937-3_40
20. Zhou, S.K., et al.: A review of deep learning in medical imaging: imaging traits, technology trends, case studies with progress highlights, and future promises. Proc. IEEE **109**(5), 820–838 (2021)
21. Zhou, S.K., Rueckert, D., Fichtinger, G.: Handbook of Medical Image Computing and Computer Assisted Intervention. Academic Press (2019)