INTERNATIONAL JOURNAL OF
RADIATION ONCOLOGY • BIOLOGY • PHYSICS

www.redjournal.org

PHYSICS CONTRIBUTION

Comprehensive Evaluation of a Deep Learning Model for Automatic Organs at Risk Segmentation on Heterogeneous Computed Tomography Images for Abdominal Radiation Therapy

Wenjun Liao, MD,[*] Xiangde Luo, PhD,[†,‡] Yuan He, MD,[§] Ye Dong, MD,[‖] Churong Li, MD,[*] Kang Li, PhD,[¶] Shichuan Zhang, MD,[*] Shaoting Zhang, PhD,[†,‡] Guotai Wang, PhD,[†,‡] and Jianghong Xiao, PhD[#]

[*]*Department of Radiation Oncology, Radiation Oncology Key Laboratory of Sichuan Province, Sichuan Clinical Research Center for Cancer, Sichuan Cancer Hospital & Institute, Sichuan Cancer Center, Affiliated Cancer Hospital of University of Electronic Science and Technology of China, Chengdu, China;* [†]*School of Mechanical and Electrical Engineering, University of Electronic Science and Technology of China, Chengdu, China;* [‡]*Shanghai AI Laboratory, Shanghai, China;* [§]*Department of Radiation Oncology, The First Affiliated Hospital of USTC, Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei, Anhui, China;* [‖]*Department of NanFang PET Center, Nanfang Hospital, Southern Medical University, Guangzhou, China;* [¶]*West China Biomedical Big Data Center; and* [#]*Radiotherapy Physics & Technology Center, Department of Radiation Oncology, Cancer Center, West China Hospital, Sichuan University, Chengdu, China*

Received Nov 8, 2022; Accepted for publication May 18, 2023

Purpose: Our purpose was to develop a deep learning model (AbsegNet) that produces accurate contours of 16 organs at risk (OARs) for abdominal malignancies as an essential part of fully automated radiation treatment planning.

Methods and Materials: Three data sets with 544 computed tomography scans were retrospectively collected. Data set 1 was split into 300 training cases and 128 test cases (cohort 1) for AbsegNet. Data set 2, including cohort 2 (n = 24) and cohort 3 (n = 20), were used to validate AbsegNet externally. Data set 3, including cohort 4 (n = 40) and cohort 5 (n = 32), were used to clinically assess the accuracy of AbsegNet-generated contours. Each cohort was from a different center. The Dice similarity coefficient and 95th-percentile Hausdorff distance were calculated to evaluate the delineation quality for each OAR. Clinical accuracy evaluation was classified into 4 levels: no revision, minor revisions (0% < volumetric revision degrees [VRD] ≤ 10%), moderate revisions (10% ≤ VRD < 20%), and major revisions (VRD ≥20%).

Results: For all OARs, AbsegNet achieved a mean Dice similarity coefficient of 86.73%, 85.65%, and 88.04% in cohorts 1, 2, and 3, respectively, and a mean 95th-percentile Hausdorff distance of 8.92, 10.18, and 12.40 mm, respectively. The performance of AbsegNet outperformed SwinUNETR, DeepLabV3+, Attention-UNet, UNet, and 3D-UNet. When experts evaluated contours from cohorts 4 and 5, 4 OARs (liver, kidney_L, kidney_R, and spleen) of all patients were scored as having no

Corresponding author: Jianghong Xiao, PhD; E-mail: xiaojh@scu.edu.cn

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

Disclosures: The authors report no conflicts of interest in this work.

The code for neural network training can be obtained at https://github.com/Luoxd1996/AbsegNet. Additionally, we will release part of the carefully annotated data set with 16 organ annotations (150 volumes from the development cohort and 20 volumes from LiTS2017) to boost this research task. Other data generated and analyzed during this study can be obtained by contacting the corresponding author with reasonable requirements.

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.ijrobp.2023.05.034.

https://doi.org/10.1016/j.ijrobp.2023.05.034

revision, and over 87.5% of patients with contours of the stomach, esophagus, adrenals, or rectum were considered as having no or minor revisions. Only 15.0% of patients with colon and small bowel contours required major revisions.
Conclusions: We propose a novel deep-learning model to delineate OARs on diverse data sets. Most contours produced by AbsegNet are accurate and robust and are, therefore, clinically applicable and helpful to facilitate radiation therapy workflow.

Introduction

Radiation therapy is one of the most important local treatment modalities for abdominal malignancies, such as cervical, prostate, pancreatic, and hepatic cancers. In the process of radiation therapy administration, delineating abdominal organs at risk (OARs) on computed tomography (CT) images is an essential step. Radiation treatment planning requires accurately calculating radiation dose, especially for OARs close to gross tumor volumes. Hence, inaccurate delineation might lead to dose miscalculations and unexpected side effects.[1]

In past decades, radiation oncologists manually conducted OAR delineation with slice-by-slice CT images. It is labor-intensive and may take several hours per case. Additionally, manual delineation of OAR often leads to inter- and intraobserver variabilities that can influence treatment outcomes in certain cases.[2-5] Therefore, consistent and high-quality abdominal OAR delineation is greatly desired in clinical practice. In this context, if feasible, full autosegmentation of whole abdominal OARs by deep learning (DL) methods could be more advantageous.

Benefitting from the advantages of feature learning, DL-based automatic delineation has offered a promising solution to solve the problems of manual delineation.[6] In recent years, a wide range of DL-based models have been proposed to segment abdominal OARs, and huge progress has been made. For example, in previous studies, spleen and liver segmentation could obtain 96% and 95% accuracy, respectively, regarding their Dice similarity coefficients (DSC).[7,8] In a recent work on kidney and pancreas segmentation, 93% and 79% DSC were obtained, respectively.[9]

However, several limitations still exist in current automatic segmentation models for abdominal OARs. For instance, many abdominal data sets just contain single-institutional, single-scanner, or single-disease patients.[10] It is unclear whether the segmentation performance acquired on those data sets might generalize well on more heterogeneous data. Huge differences in organ morphologic structure, disease status, image appearance, and image quality obtained from various patients by different scanners could affect the accuracy of image segmentation.

The main focus of most studies is on the reporting of segmentation results in their own cohorts, but they fail to perform a comprehensive clinical assessment for automatically segmented contours, a critical step in medical image segmentation.[11] Van Dijk et al[12] reported that it took an average of 34 to 54 minutes to make the automatic contours suitable for clinical use. Other disadvantages should not be ignored. For example, some studies just segmented a few OARs, and in some studies, ground truth OARs used network predictions as initial results and were refined by experts, which might have rater bias.[10,13,14]

To address the aforementioned issues, in this paper, we retrospectively collected whole abdominal CT images of 544 patients with various tumors from 5 different centers. Then, we proposed a new DL model, named “AbsegNet,” that could delineate a comprehensive set of 16 abdominal OARs. The performance of AbsegNet was validated in 172 patients across 3 different cohorts, and the accuracy was compared with 5 previous state-of-the-art methods. Moreover, 2 experienced radiation therapy practitioners were invited to evaluate the accuracy of AbsegNet in another independent 2 cohorts with 72 patients.

Methods and Materials

Data summary

Three data sets with a total number of 544 patients in this study were used (Table 1). Data set 1, collected from the West China Hospital (WCH), contained 428 planning CT images, which were randomly split into 300 cases for training and 128 cases for internal testing (cohort 1), with a ratio of around 3:1. Data set 2 included cohort 2 (n = 24) collected from the Sichuan Cancer Hospital (SCH), and cohort 3 (n = 20) was collected from a public data set (LiTS2017).[15] These 2 cohorts were used to externally validate the performance of AbsegNet. Data set 3 consisted of cohort 4 (n = 40) collected from the Southern Medical University (SMU), and cohort 5 (n = 32) was from the Anhui Provincial Hospital (APH), which were used to clinically assess the accuracy of AbsegNet-generated contours. The flow chart of this study is illustrated in Fig 1.

Distributions of sex, age, tumor sites, and scanning parameters of these data sets are presented in Table 1. Patients with various tumors with CT images from different scanners using different slices were incorporated into this study. In addition, a comparison of the included data sets with several available public data sets is shown in Table E1. The included data sets mostly covered OARs for abdominal radiation therapy with ample sample sizes from multiple centers.

This retrospective study was approved by the Ethics Committee on Biomedical Research for these hospitals, and informed consent was waived.

Table 1 **Baseline characteristics and scanning parameters**

Characteristics	Data set 1 (n = 428)		Data set 2 for external testing (n = 44)		Data set 3 for clinical evaluation (n = 72)	
	Training cohort WCH (n = 300, %)	Internal testing cohort 1 WCH (n = 128, %)	Cohort 2 SCH (n = 24, %)	Cohort 3 Public[‡] (n = 20)	Cohort 4 SMU (n = 40, %)	Cohort 5 APH (n = 32, %)
Sex						
Male	182 (60.7)	76 (59.4)	10 (41.7)	NA	0	0
Female	118 (39.3)	52 (40.6)	14 (58.3)	NA	40 (100.0)	32 (100.0)
Age (median)	47 (17-75)	47 (20-72)	49 (36-72)	NA	55 (35-62)	53 (46- 70)
Tumor site						
Rectal cancer	143 (47.7)	60 (46.9)	5 (20.8)	NA	0	0
Prostate cancer	39 (13.0)	18 (14.1)	7 (29.2)	NA	0	0
Gynecologic[*]	34 (11.3)	11 (8.6)	12 (50.0)	NA	40 (100.0)	32 (100.0)
Bladder cancer	11 (3.6)	5 (3.9)	0	NA	0	0
Metastatic tumor	44 (14.7)	20 (15.6)	0	NA	0	0
Others[†]	29 (9.7)	14 (10.9)	0	NA	0	0
OAR types annotated	16	16	16	15[§]	16	16
Scanning parameters						
Total slice (median)	200 (123-368)	201 (145- 436)	168 (118- 248)	501 (276- 842)	176 (152-201)	172 (103-260)
Thickness (median, mm)	3.0 (0.60-0.98)	3.0 (2.5-3.0)	3.0 (3.0- 3.0)	1.0 (0.70- 1.5)	3.0 (3.0- 3.0)	3.0 (3.0- 5.0)
In plane spacing	0.98 (0.60-0.98)	0.98 (0.78- 1.27)	0.98 (0.95- 0.98)	0.74 (0.60 -0.90)	0.96 (0.90- 1.04)	1.04 (0.91- 1.17)
Manufacturer	Siemens	Siemens	Philips	NA	Philips	Philips

Abbreviations: WCH = West China Hospital; SCH = Sichuan Cancer Hospital; SMU = Southern Medical University; APH = Anhui Provincial Hospital; NA = not applicable; OAR = organ at risk.

[*] Included cervical cancer and endometrial cancer.

[†] Included liver cancer, pancreatic cancer, kidney cancer, sarcoma, and testicular cancer.

[‡] These patients were collected from a public data set, LiTS2017.[15]

[§] Considering the liver was manually previously labeled in LiTS2017, it was excluded, and the remaining 15 OARs were labeled in this study.

Ground truth contours

To ensure data uniformity, abdominal OARs in each case were manually delineated by a radiation oncologist from WCH with >8 years of experience treating abdominal malignancies. After that, another senior oncologist from the same hospital with >20 years of experience checked and revised these annotations carefully and, in cases of disagreement, produced consensus annotations. All CT scans of these data sets, except for LiTS2017,[15] were exhaustively labeled with 16 anatomic organs, including the liver, spleen, kidneys (left and right), stomach, gallbladder, esophagus, pancreas, adrenals, duodenum, colon, small bowel, rectum, bladder, and head of the femurs (left and right). Considering the liver was manually previously labeled in LiTS2017, it was excluded, and the remaining 15 OARs were labeled in this study. The OAR delineation principles were in accordance with relevant radiation guidelines.[16,17] An example CT scan and ground truth contour from data set 1 is shown in Fig E1.

All manually delineated contours were performed in the MIM 7.07 Software (Microsoft, Corp).[18,19]

Segmentation network construction

The AbsegNet framework is shown in the top right of Fig 1, and detailed architecture parameters are presented in Table E2. In this work, we present a new method to train accurate and robust segmentation networks by employing data augmentation and knowledge distillation, consisting of a teacher-student model. Considering that CT images come from different centers, patients, scanning protocols, tumors, and contrast types may cause data distribution shifts. As expected, our data sets had huge intensity distribution gaps (Fig. E2). These distribution shifts could lead to model collapse on unseen centers.[20] To boost the network's robustness on unseen data sets, we used a wide range of data augmentation strategies to generate different augmented images for network training online, such as intensity transformations (randomly using random noise, sharpening, histogram match, nonlinear transformation, and histogram equalization) and spatial transformations (randomly applying rotation, rescaling, elastic deformation). Furthermore, we combined data augmentation with a general knowledge distillation framework to train segmentation models.[21,22] Specifically, in the training stage, teacher (Θ) and student (Ψ) networks take augmented images ($T^1(i)$ and $T^2(i)$) as inputs and produce corresponding predictions ($\Theta T^1(i)$) and $\Psi(T^2(i))$). Here, T^1 is a random noise transformation, and T^2 is a random one in the intensity and spatial transformations sets. T^1 and T^2 can be considered weak and strong augmentations, respectively. Then, we encouraged the student to generate predictions based on the teacher via a knowledge distillation loss (L_{kd}) according to the following equation.

$$L_{kd} = L_{kl}\ \left(\Psi\left(T^2(i)\right),\ \Theta\left(T^1(i)\right)/t\right)$$

Where t is a temperature factor that controls the importance of the teacher's predictions and is set to 4, and Kl is

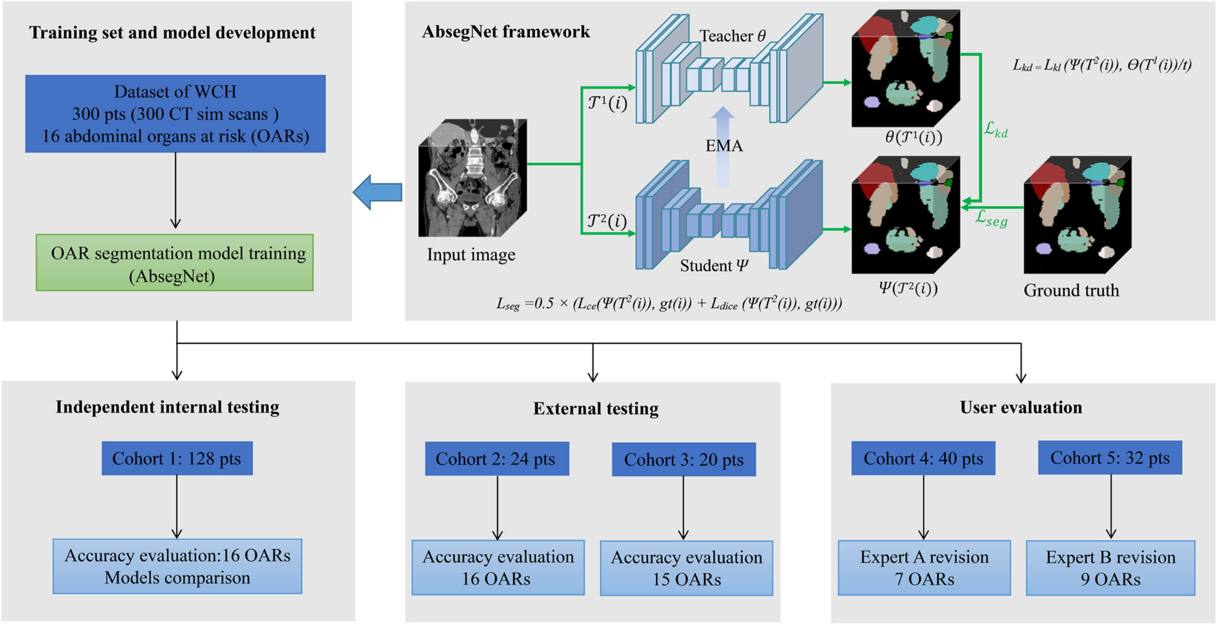

Fig. 1. The flow chart of this study. The top right of the figure is the AbsegNet framework, which consists of a teacher-student model. In the training stage, the teacher and the student take images with different augmentation strategies (the teacher with weak augmentation and the student with strong augmentation as inputs) and then employ the teacher's output to teach the student to be more robust. The student network updates parameter by minimizing loss functions, and the teacher parameter is updated as an exponential moving average of the student's parameter. In the testing stage, the teacher model was used to produce the final segmentation results. The mathematical definitions are presented in the section of segmentation network construction. *Abbreviations*: WCH = West China Hospital; CT = computed tomography; EMA = exponential moving average; OAR = organ at risk.

the Kullback-Leibler divergence function. At the same time, the student network is also supervised by the ground truth (*gt(i)*) via a combination loss as the following.

$$L_{seg} = 0.5 \times \left(L_{ce}\left(\Psi\left(T^2(i)\right), gt(i)\right) + L_{dice}\left(\Psi\left(T^2(i)\right), gt(i)\right)\right)$$

Where *ce* and *dice* represent the cross-entropy loss and Dice loss, respectively, and the total objective loss function is L_{total} = L_{seg} + *0.1*× L_{kd}. Afterward, the student network updates the parameter by minimizing L_{total}, and the teacher's parameter is updated as an exponential moving average of the student's parameter. Based on the proposed method, the segmentation network is encouraged to learn the anatomic context feature and ignore the intensity distribution to boost the generalization on unseen data sets. In the testing stage, the teacher model was used to produce final results following previous suggestions.[22] Different from previous works,[23-25] AbsegNet could be applied to unseen data sets without fine-tuning or retraining.

To confirm the usefulness of distillation learning in the segmentation of CT images from various centers, we first reported the results of the proposed method with and without knowledge distillation in cohorts 1, 2, and 3, which showed that using knowledge distillation could improve performance in the 3 cohorts (Table E3). The overall average DSC of AbsegNet with knowledge distillation was significantly higher than that of AbsegNet without knowledge distillation (cohort 1, 86.73% vs 85.34%; cohort 2, 85.65% vs 81.98%; cohort 3, 88.04% vs 84.65%; all *P* values < .05) (Table E3).

Preprocessing of images

In the preprocessing phase, all images were reformatted into a standard orientation of right-to-left, anterior-to-posterior, and inferior-to-superior, in the x, y, and z axes, respectively. Each image's intensities (Hounsfield units) were adjusted based on the gray-level histogram and cut-off intensities outside the 0.5 and 99.5 percentiles. Then, we resampled the images to the fixed resolution of 0.98 × 0.98 × 3.0 mm^3, which was the medium resolution of the training set. Finally, all images were normalized to zero mean and unit variance. In the postprocessing phase, the largest connected component selection and morphologic operation were used to refine the network's predictions and generate final results.

Implementation details

The proposed method was implemented by PyTorch on a Ubuntu20.04 desktop with 2 NVIDIA 3090 GPUs. A 3-dimensional (3D)-UNet[23] was used as a baseline model. The total epoch was set to 1000, and the batch size was 2. The input patch size was randomly cropped from the preprocessed image with a shape of 64 × 192 × 192. We employed a set of data augmentation strategies and knowledge distillation to train the network (detailed in Supplementary Materials). We used the stochastic gradient descent optimizer (weight decay = 10-4, momentum = 0.9) to update network parameters. The initial learning rate was 0.01 and adjusted by a poly learning rate strategy. In the testing stage, we used a sliding window strategy with a stride of 32 × 96 × 96 to produce final predictions.

Method comparison

In this work, AbsegNet was compared with 5 famous and widely used methods: (1) UNet, which presents a U-shape encoder-decoder network for biomedical image segmentation and achieves very promising results on many tasks[26]; (2) 3D-UNet, an extension of UNet from 2D space to 3D space for volumetric image segmentation[23]; (3) DeepLabV3 +, an encoder-decoder with an atrous separable convolution network for natural image semantic segmentation[24]; (4) Attention-UNet, which extends UNet attention gates to focus on target structures of varying shapes and sizes for better segmentation results[25]; and (5) SwinUNETR, a new combination framework that uses the merits of Swin Transformers and U-Shape networks for medical image segmentation and achieves encouraging performance.[27] To ensure consistency of comparisons, public implementations of the methods were used to directly train the network based on the same training data set and procedures (Table E4).

Assessment of AbsegNet-generated contours by experts

Two senior experts (A and B) with >15 years of radiation experience from SMU and APH, respectively, were invited to assess the accuracy of AbsegNet-generated contours from cohorts 4 and 5 (Fig. 1). Each expert was required to revise incorrect OAR segmentation when necessary. In the correction process, experts were blinded to the ground truth contours and encouraged to obey the same delineation guidelines described previously. Considering the heavy burden of annotating with 16 OARs, 7 representative OARs, including some solid and gastrointestinal organs, were assigned to expert A, and the other 9 OARs, including some of those organ types, were assigned to expert B. The kidneys (left and right), pancreas, duodenum, bladder, and femurs (left and right) were reviewed by expert A. The liver, spleen, stomach, gallbladder, esophagus, adrenals, colon, small bowel, and rectum were reviewed by expert B.

Next, AbsegNet-predicted contours were compared with their corresponding revised contours to calculate volumetric revision degrees (VRD), which were defined as the volume required to be edited divided by the volume of AbsegNet-generated contours multiplied by 100.[28] Accuracy was classified into 4 levels: no revision (VRD = 0%), minor revisions (0 < VRD ≤ 10%), moderate revisions (10% < VRD < 20%), and major revisions (VRD ≥ 20%).

Qualitative analysis was also an important part of our research. Similar to a previous work,[29] experts A and B were invited to subjectively evaluate each automatic contouring result together. At first, we randomly selected 5 cases from each cohort. Then, autosegmentations from these 25 patients were evaluated by the 2 experts. We used 3-grade criteria to estimate the degree of clinically acceptable: (1) completely acceptable (the prediction can be used in the treatment planning without any revision), (2) acceptable (the prediction needs a few refinements but has no obvious clinical effect without corrections), and (3) unacceptable (the prediction needs to be substantially revised before treatment planning or needs to be redelineated manually).

A study was also performed to compare the time spent by expert A in delineating OARs under 2 modes: with or without assistance from AbsegNet. In the first mode, the contours of 16 OARs produced by AbsegNet were provided to expert A, who would then examine the predictions and revise the incorrect ones when necessary. In the second mode, the OAR delineation was conducted completely manually. The contouring time for each patient includes the time spent verifying the results and revising the model's predictions. We randomly selected 10 cases from the 25 patients mentioned previously to conduct this experiment.

Evaluation metrics

To evaluate the performance of AbsegNet, the volumetric DSC and 95th percentile Hausdorff distance (HD95) were adopted, which are the most commonly used metrics in this field.[30] The DSC measures the volumetric overlap between 2 contours, and the HD measures the boundaries of 2 contours. Because the max HD is very sensitive to outliers, HD95, which measures the 95th percentile distance between 2 contours, is often used instead.[31]

Statistical analysis

Statistical analysis was performed using an SPSS software package (version 22.0; IBM SPSS, Inc). Numeric variables were denoted as mean ± SD and compared by paired *t* test when necessary. A 2-tailed *P* value < .05 was considered significant.

Results

Intraobserver variability examination

To determine the intraobserver variability, 10 CT images were randomly selected from data set 1 and recontoured by the same expert after an interval of 2 months. The second contours were compared with the first corresponding contours to calculate DSC and HD95. We found that the distribution of mean DSC (mDSC) for 12 out of 16 OARs was around 95%, and the mean HD95 (mHD95) across all OARs was <5 mm (Fig. E3). It was suggested that the intraobserver discrepancy was minor and the annotations were reliable.

Performance comparison in the internal testing cohort

The DSC and HD95 of 16 OARs obtained by 6 DL algorithms in cohort 1 are listed in Table 2. Regarding mDSC, AbsegNet performed best in 14 out of 16 OARs among 6 algorithms and achieved an mDSC >90% for 8 out of 16 OARs. Only 2 OARs (adrenals and duodenum) had an mDSC <80%. Considering all OARs as a whole, the mDSC was 86.73%, 84.39%, 82.11%, 84.67%, 84.66%, and 82.82% for AbsegNet, SwinUNETR, DeepLabV3+, Attention-UNet, UNet, and 3D-UNet, respectively. In terms of HD95, AbsegNet showed the best performance for 12 OARs. We also observed that 5 OARs produced by AbsegNet had an mHD95 <5 mm and 11 OARs <10 mm. The overall mHD95 of AbsegNet, SwinUNETR, DeepLabV3+, Attention-UNet, UNet, and 3D-UNet were 8.92, 12.50, 16.44, 13.12, 13.00, and 23.22 mm, respectively. Furthermore, compared with SwinUNETR (the best among all previous algorithms), a 29% reduction of mHD95 was observed for AbsegNet. Visualization of a randomly selected CT scan from cohort 1 is illustrated in Fig 2.

Table E5 summarizes previously reported delineation results for multiple abdominal OARs, with comparable accuracy observed in AbsegNet.

Performance comparison in external testing cohorts

Table 3 summarizes the DSC and HD95 in external cohort 2. Regarding DSC, AbsegNet was prone to produce more accurate contours than other algorithms, showing the best performance at 13 OARs. The overall mDSCs of AbsegNet, SwinUNETR, DeepLabV3+, Attention-UNet, UNet, and 3D-UNet were 85.65%, 79.27%, 78.51%, 79.73%, 81.14%, and 77.58%, respectively. Regarding HD95, the accuracy of 13 OARs produced by AbsegNet outperformed 5 previous algorithms. Furthermore, the advantage of AbsegNet over other algorithms is more obvious when evaluating the mHD95 across all OARs, with the value decreasing nearly 50% compared with SwinUNETR, DeeplabV3+, UNet, and 3D-UNet. Similarly, the visualization of a randomly selected patient from cohort 2 is shown in Fig 2.

Consistent results were obtained from the public CT images (Table E6). AbsegNet presented the best accuracy in 13 out of 15 OARs. Overall mDSC of AbsegNet, SwinUNETR, DeepLabV3+, Attention-UNet, UNet, and 3D-UNet was 88.04%, 79.87%, 83.03%, 81.75%, 82.27%, and 82.43%, respectively. An improvement of 5.01% in mDSC was observed when comparing AbsegNet with DeepLabV3+, the best among the 5 previous methods. For HD95, AbsegNet

Table 2 Accuracy comparison in cohort 1 (n = 128)

Variable		DSC (%)					HD95 (mm)					
OAR	AbsegNet	SwinUNETR	DeepLabV3+	Att	UNet	3D-UNet	AbsegNet	SwinUNETR	DeepLabV3+	Att	UNet	3D-UNet
Liver	96.40 ± 1.13[‖]	95.94 ± 1.58	95.00 ± 2.14	95.61 ± 2.06[*]	95.49 ± 1.44	95.25 ± 1.79	4.26 ± 5.66[‖]	4.63 ± 8.27	6.27 ± 9.60	5.04 ± 4.22	5.55 ± 4.20	5.58 ± 4.46
Spleen	95.13 ± 5.32[‖]	94.17 ± 6.31	93.81 ± 3.66	94.51 ± 5.39[*]	94.46 ± 3.45	93.77 ± 4.26	3.38 ± 7.71[‖]	4.07 ± 7.25	4.31 ± 5.92	3.61 ± 5.46	6.79 ± 19.15	3.98 ± 5.21
Kidney_L	95.60 ± 1.07[‖]	94.36 ± 2.93	94.31 ± 2.68	94.52 ± 6.29[†]	94.86 ± 2.93	94.16 ± 3.08	2.54 ± 0.72[‖]	8.79 ± 23.72[‡]	3.18 ± 1.84	4.09 ± 9.63	3.03 ± 1.82	3.78 ± 4.01
Kidney_R	95.60 ± 1.31[‖]	94.29 ± 2.76	94.11 ± 2.98	94.72 ± 3.60[‡]	94.97 ± 2.30	94.53 ± 2.32	2.41 ± 0.83[‖]	7.97 ± 23.01[‡]	3.59 ± 3.77	3.07 ± 2.43	3.63 ± 4.77	3.26 ± 3.26
Stomach	91.46 ± 4.08[‖]	89.11 ± 7.53	88.08 ± 8.00	88.95 ± 7.38[*]	89.41 ± 5.95	86.90 ± 9.38	9.87 ± 12.46[‖]	13.48 ± 20.1	18.21 ± 42.38	13.59 ± 19.91	14.30 ± 14.85	18.24 ± 38.04
Gallbladder[§]	80.74 ± 15.17[‖]	71.26 ± 25.41	74.42 ± 17.79	74.56 ± 20.29[*]	76.02 ± 17.09	71.79 ± 21.59	6.93 ± 9.56[‖]	13.37 ± 19.34[*]	10.86 ± 12.38	13.46 ± 18.53	10.22 ± 13.32	10.95 ± 15.33
Esophagus	81.36 ± 6.25[‖]	78.81 ± 6.33	73.24 ± 11.91	77.54 ± 8.79[*]	76.96 ± 8.41	73.00 ± 12.51	5.19 ± 4.46[‖]	5.26 ± 2.89	6.15 ± 3.97	31.91 ± 91.13	5.64 ± 3.16	7.28 ± 4.84
Pancreas	82.71 ± 7.54[‖]	80.46 ± 7.92	77.64 ± 8.33	79.82 ± 8.43[*]	79.82 ± 8.00	76.90 ± 9.63	7.64 ± 7.25[‖]	8.82 ± 9.66	10.56 ± 8.79	8.62 ± 7.72	9.99 ± 8.20	12.08 ± 22.83
Adrenals	71.34 ± 11.63[‖]	65.16 ± 16.36	49.30 ± 10.72	68.13 ± 12.99[*]	67.58 ± 12.06	67.71 ± 12.23	6.58 ± 5.28[‖]	22.4 ± 17.49[*]	65.22 ± 10.81	8.18 ± 9.59	8.66 ± 7.06	8.39 ± 8.66
Duodenum	67.77 ± 16.28	83.46 ± 8.99[‖]	61.21 ± 17.15	65.51 ± 17.04[†]	64.64 ± 17.77	62.45 ± 17.65	25.85 ± 35.60	16.78 ± 17.34[‖]	23.14 ± 16.43	24.31 ± 18.97	21.61 ± 16.50	23.65 ± 17.64
Colon	85.74 ± 9.51[‖]	84.42 ± 7.71	80.03 ± 10.94	82.91 ± 9.36[*]	81.82 ± 9.39	80.45 ± 10.21	14.14 ± 15.44	9.22 ± 8.55[‖]	19.34 ± 16.98	17.11 ± 16.49	25.65 ± 36.03	19.70 ± 17.20
Small bowel	86.42 ± 8.36[‖]	68.54 ± 11.88	82.21 ± 8.33	84.56 ± 8.09[*]	82.70 ± 8.66	82.00 ± 9.36	8.49 ± 9.43	7.28 ± 6.88[‖]	11.44 ± 8.70	9.66 ± 9.07	11.48 ± 9.04	11.10 ± 9.43
Rectum	80.06 ± 12.43	78.08 ± 12.24	78.19 ± 11.59	78.85 ± 12.35[†]	80.51 ± 10.12[‖]	76.92 ± 11.02	14.06 ± 12.36	14.12 ± 11.71	15.28 ± 15.79	14.40 ± 11.75	13.29 ± 10.46[‖]	17.18 ± 29.77
Bladder	93.14 ± 7.51[‖]	91.1 ± 9.85	90.33 ± 12.45	91.22 ± 11.15[*]	91.46 ± 11.82	90.48 ± 11.06	4.56 ± 7.80[‖]	14.87 ± 53.42[†]	25.97 ± 73.62	7.00 ± 13.48	19.96 ± 62.55	16.69 ± 53.41
Femur_L	91.87 ± 4.03[‖]	90.46 ± 9.36	90.84 ± 4.22	91.59 ± 4.49	91.86 ± 4.59	88.96 ± 5.60	15.50 ± 56.57[‖]	23.6 ± 85.71	25.06 ± 91.52	17.45± 70.20	21.28 ± 81.89	132.8 ± 171.6
Femur_R	92.40 ± 4.56[‖]	90.56 ± 9.2	91.00 ± 5.01	91.72 ± 4.37	92.02 ± 4.48	89.91 ± 5.70	11.32 ± 30.85[‖]	25.32 ± 88.26	25.38 ± 92.43	28.50 ± 95.49	25.99 ± 89.30	88.89 ± 150.7
Average	86.73[‖]	84.39	82.11	84.67[*]	84.66	82.82	8.92[‖]	12.50[*]	16.44	13.12	13.00	23.22

Data were denoted as mean ± SD. Bold numbers represent the best results. *P* values were obtained by comparing our method with the best one among the 5 previous methods according to the overall average DSC.

Abbreviations: Att = attention-UNet; DSC = Dice similarity coefficient; HD95 = 95th percentile Hausdorff distance; OAR = organ at risk.

[*] *P* < .001

[†] *P* < .05

[‡] *P* < .01

[§] Four patients underwent gallbladder resection, so the number of gallbladders was 124.

[‖] Best results.

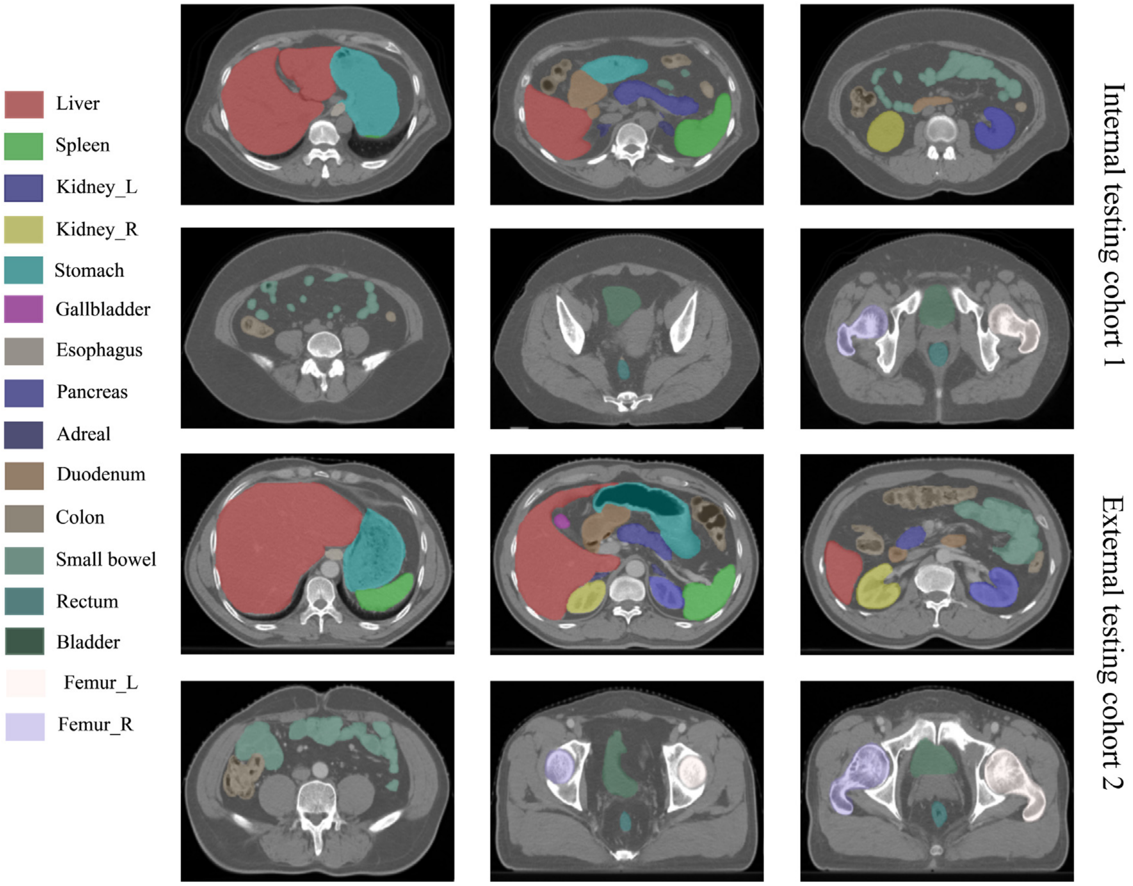

Fig. 2. Visualization of 2 randomly selected patients from internal testing cohort 1 and external testing cohort 2, respectively.

showed the most accuracy in 66.7% (10/15) of OARs. The whole mHD95s of AbsegNet, SwinUNETR, DeepLabV3+, Attention-UNet, UNet, and 3D-UNet were 12.40, 23.77, 22.06, 15.72, 24.28, and 25.46 mm, respectively.

Performance of AbsegNet in all cohorts

We examined the performance of AbsegNet in 2 data sets, including cohorts 1, 2, and 3. The mDSC and mHD95 of each OAR for all patients are shown in Fig E4. The mDSC of 8 OARs exceeded 90%, whereas only 2 OARs had an mDSC of <80% (adrenals and duodenum). For HD95, 9 out of 16 OARs had an mHD95 of <10 mm, whereas only 3 OARs were >15 mm (colon, duodenum, and femur_L).

Performance of AbsegNet in different types of organs

The 16 OARs were divided into 4 groups according to morphologic structures and size: solid organs (liver, spleen, kidneys, and pancreas), gastrointestinal organs (esophagus, stomach, duodenum, colon, small bowel, rectum, and bladder), bone tissues (femurs), and small organs (gallbladder and adrenals). Then, the performance of AbsegNet was examined in these 4 groups. The exact results are listed in Table E7. Of these cohorts, all mDSC were >90% for solid organs and bone tissues, all mDSC exceeded 80% in gastrointestinal organs, and all mDSC were >75% in small organs.

Clinical assessment of contours produced by AbsegNet

When using our 4-grading criteria to assess contour accuracy, all patients with AbsegNet-produced contours for kidneys, bladder, and femur_L were deemed satisfactory by the expert, with no or minor revisions in cohort 4 (Table 4). However, 25% (n = 10) and 10% (n = 4) of patients with AbsegNet-produced duodenum were considered to have moderate and major revisions, respectively (Table 4). Similarly, in cohort 5, AbsegNet-generated contours for the liver and spleen in all patients were not required to be revised, and over 87.5% of patients with contours of the stomach, esophagus, adrenals, and rectum were considered satisfactory, with no or minor revisions (Table 4). Only 2 (6.2%), 5 (15.6%), and 4 (12.5%) patients with autosegmentations of the gallbladder, colon, and small bowel, respectively, required major revisions (Table 4). A visual display of

Table 3 Accuracy comparison in cohort 2 (n = 24)

Variable	DSC (%)						HD95 (mm)					
OAR	AbsegNet	SwinUNETR	DeepLabV3+	Att	UNet	3D-UNet	AbsegNet	SwinUNETR	DeepLabV3+	Att	UNet	3D-UNet
Liver	96.75 ± 1.21[§]	90.52 ± 15.86	92.94 ± 5.93	92.71 ± 9.72	91.79 ± 10.77[*]	89.53 ± 18.85	4.14 ± 3.93[§]	15.32 ± 17.85	10.75 ± 11.19	11.16 ± 16.33[*]	12.14 ± 12.65	15.86 ± 19.16
Spleen	91.50 ± 11.27[§]	86.81 ± 14.85	89.78 ± 10.34	88.81 ± 13.25	90.04 ± 12.94	90.13 ± 7.52	12.92 ± 22.47[§]	33.54 ± 56.53	13.45 ± 20.29	13.98 ± 21.75	19.64 ± 36.07	13.41 ± 19.97
Kidney_L	94.91 ± 2.14[§]	89.81 ± 8.69	91.76 ± 12.41	92.94 ± 6.69	92.30 ± 12.22	91.91 ± 7.26	5.60 ± 12.44	32.27 ± 47.48	4.02 ± 3.66[§]	3.92 ± 3.17	4.76 ± 5.49	10.83 ± 20.26
Kidney_R	95.13 ± 1.15[§]	91.23 ± 6.43	92.78 ± 8.40	93.24 ± 5.11	93.43 ± 7.21	93.36 ± 4.13	2.65 ± 0.76[§]	31.25 ± 62.95	3.61 ± 2.22	6.67 ± 15.44	5.41 ± 10.90	4.67 ± 6.07
Stomach	86.45 ± 14.42[§]	67.62 ± 27.35	75.47 ± 22.71	67.11 ± 27.16	72.81 ± 25.16[†]	61.32 ± 29.86	12.43 ± 14.03[§]	31.54 ± 35.77	22.64 ± 18.22	27.98 ± 22.66[†]	25.31 ± 16.09	32.96 ± 28.19
Gallbladder	87.85 ± 6.36[§]	74.23 ± 28.67	79.13 ± 16.45	78.03 ± 20.53	77.43 ± 20.47[‡]	72.51 ± 23.77	6.93 ± 14.15[§]	23.42 ± 76.39	10.33 ± 16.78	10.28 ± 16.39	10.32 ± 14.95	11.02 ± 16.73
Esophagus	79.90 ± 4.56[§]	78.23 ± 5.82	71.24 ± 12.05	74.50 ± 8.97	75.58 ± 10.00[‡]	69.89 ± 13.44	5.35 ± 3.01[§]	16.1 ± 49.74	11.54 ± 12.89	45.34 ± 90.36[*]	10.68 ± 13.17	9.36 ± 6.70
Pancreas	81.32 ± 9.68[§]	75.51 ± 13.43	74.12 ± 14.14	73.40 ± 18.10	76.05 ± 14.89[*]	72.47 ± 16.54	7.91 ± 7.36[§]	17.58 ± 19.11	12.93 ± 8.15	15.05 ± 13.85[‡]	13.48 ± 10.86	19.05 ± 18.91
Adrenals	70.87 ± 11.66[§]	59.32 ± 19.3	46.11 ± 13.05	62.93 ± 17.22	68.00 ± 11.96[*]	66.26 ± 11.92	6.19 ± 3.56[§]	30.31 ± 17.8	65.06 ± 7.98	7.42 ± 4.52	6.87 ± 3.59	6.79 ± 4.06
Duodenum	63.40 ± 19.62	81.84 ± 8.39[§]	54.75 ± 20.29	58.03 ± 21.19	60.80 ± 19.43	56.88 ± 20.42	23.73 ± 13.52	28.76 ± 44.95	26.46 ± 18.91	34.23 ± 23.70[*]	23.02 ± 15.72[§]	42.34 ± 68.51
Colon	85.82 ± 5.64[§]	66.79 ± 18.77	78.91 ± 8.47	79.98 ± 9.06	81.86 ± 5.54[†]	78.33 ± 14.94	12.24 ± 8.17[§]	23.24 ± 16.16	18.76 ± 8.70	21.18 ± 12.74[†]	29.87 ± 39.42	19.41 ± 11.51
Small bowel	82.41 ± 9.50[§]	68.18 ± 10.31	63.80 ± 17.85	67.87 ± 16.39	68.73 ± 14.88[†]	63.92 ± 17.30	14.33 ± 13.73[§]	7.0 ± 4.23	20.52 ± 9.44	17.63 ± 9.05	26.74 ± 51.21	19.34 ± 9.92
Rectum	83.02 ± 6.42	82.06 ± 6.6	81.74 ± 7.77	81.58 ± 13.24	83.30 ± 6.48[§]	79.57 ± 15.72	10.53 ± 7.69[§]	15.66 ± 19.66	11.00 ± 10.08	10.91 ± 11.14	32.77 ± 105.80	11.29 ± 10.62
Bladder	87.74 ± 18.25[§]	81.48 ± 26.98	84.97 ± 17.67	82.71 ± 25.46	84.72 ± 19.75	85.08 ± 20.98	6.51 ± 8.07[§]	33.36 ± 120.12	25.86 ± 79.16	10.55 ± 14.56	41.72 ± 94.01	62.64 ± 144.60
Femur_L	93.14 ± 2.79[§]	89.09 ± 6.45	90.75 ± 4.11	91.01 ± 7.30	91.68 ± 19.75	82.15 ± 16.76	13.31 ± 46.50[§]	39.62 ± 53.69	37.26 ± 54.82	22.15 ± 63.90	28.69 ± 44.01	45.19 ± 45.11
Femur_R	90.22 ± 11.00	85.52 ± 17.93	87.76 ± 7.25	90.85 ± 10.68[§]	89.71 ± 6.21	87.95 ± 15.50	18.04 ± 43.91	48.34 ± 58.75	37.69 ± 52.65	11.02 ± 20.71[§]	30.62 ± 40.68	32.06 ± 48.48
Average	85.65[§]	79.27	78.51	79.73	81.14[†]	77.58	10.18[§]	26.71	20.62	16.84[†]	20.13	22.26

Data were denoted as mean ± SD. *P* values were obtained by comparing our method with the best one among the 5 previous methods according to the whole average DSC.

Abbreviations: Att = attention-UNet; DSC = Dice similarity coefficient; HD95 = 95th percentile Hausdorff distance; OAR = organs at risk.

[*] $P < .05$

[†] $P < .001$

[‡] $P < .01$

[§] Best results.

Table 4 Clinical accuracy evaluation for each OAR in cohorts 4 and 5

Cohort 4 (n = 40)	No revision (n, %)	Minor revision (n, %)	Moderate revision (n, %)	Major revision (n, %)
Kidney_L	40 (100.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Kidney_R	40 (100.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Pancreas	32 (80.0%)	6 (15.0%)	2 (5.0%)	0 (0.0%)
Duodenum	7 (17.5%)	19 (47.5%)	10 (25.0%)	4 (10.0%)
Bladder	38 (95.0%)	2 (5.0%)	0 (0.0%)	0 (0.0%)
Femur_L	39 (97.5%)	1 (2.5%)	0 (0.0%)	0 (0.0%)
Femur_R	35 (87.5%)	3 (7.5%)	1 (2.5%)	0 (0.0%)
Cohort 5 (n = 32)				
Liver	32 (100.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Spleen	32 (100.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Stomach	23 (71.9%)	6 (18.8%)	3 (9.3%)	0 (0.0%)
Gallbladder	21 (65.7%)	6 (18.8%)	3 (9.3%)	2 (6.2%)
Esophagus	27 (84.4%)	2 (6.3%)	3 (9.3%)	0 (0.0%)
Adrenals	13 (40.6%)	15 (46.9%)	4 (12.5%)	0 (0.0%)
Colon	16 (50.0%)	7 (21.9%)	4 (12.5%)	5 (15.6%)
Small bowel	23 (71.9%)	4 (12.5%)	1 (3.1%)	4 (12.5%)
Rectum	24 (75.0%)	4 (12.5%)	4 (12.5%)	0 (0.0%)

Abbreviation: OAR = organ at risk.

AbsegNet-generated OARs revised by experts is presented in Fig E5.

In addition, 2 experts together subjectively evaluated each OARs predicted by AbsegNet from the 25 patients. All patients with AbsegNet-generated liver, spleen, and kidney predictions were considered completely acceptable (Table E8). Only 2 (8.0%), 2 (8.0%), 5 (20.0%), 3 (12.0%), and 2 (8.0%) patients with autosegmentations of the stomach, adrenals, duodenum, colon, and small bowel, respectively, were considered as clinically unacceptable (Table E8).

We recorded the time spent by expert A to delineate 16 OARs in each of the 10 patients. With the assistance of AbsegNet, the expert spent, on average, 12.04 ± 2.93 minutes to delineate 1 patient. However, without assistance from AbsegNet, the delineation time was significantly increased to an average of 39.73 ± 3.38 minutes per case ($P < .001$) (Fig. E6).

Discussion

In this study, we aimed to develop a novel DL model to segment OARs for abdominal radiation therapy accurately and robustly. At first, large-scale and multicenter CT scans were collected, and a training cohort with high-quality annotations was used to train AbsegNet. Then, a comprehensive model performance evaluation was conducted across 5 different institutions, including a clinical assessment. These results showed that AbsegNet achieved state-of-the-art performance in seen subjects and generalized well to unseen subjects. Compared with 5 previous sophisticated DL methods, the accuracy of most OARs produced by AbsegNet was higher. And considering all OARs as a whole, AbsegNet demonstrated the best performance. When experts evaluated these autosegmentations, most OARs were considered satisfactory with no or minor revisions, suggesting that AbsegNet-generated contours were clinically acceptable.

In recent years, many fully automatic segmentation models have been proposed to delineate OARs in the abdomen.[8,9,32,33] Nevertheless, robust segmentation of OARs remains a challenge in real-world clinical scenarios. Most existing data sets using abdominal organ segmentation vary in size (from dozens to hundreds) and the number of annotations (single or several).[10] For instance, the BTCV (Beyond The Cranial Vault) provided only 50 CT scans covering 13 organs.[34] Although the AbdomenCT-1K offered over 10,000 CT scans, only 4 organs were included.[10] Besides, some data sets only included a single disease, such as all patients with gastric cancer.[33] Considering underrepresentation problems in those data sets, using those models to directly segment abdominal OARs for radiation therapy is not easy. Compared with previous data sets,[1,9,14,35,36] the data sets included in this study were more advantageous from the following aspects: (1) large-scale: data sets contained over 500 CT scans covering nearly all abdominal and pelvic OARs; (2) diverse and clinically relevant: the data sets used in this study were collected from real-world clinical settings; for example, these patients had various abdominal primary or metastatic tumors, were scanned by different scanners at different medical centers, and both contrast and non-contrast CT images were included; and (3) high-quality annotations: principles for OAR delineation in this study were in line with recommendations of the Radiation Therapy Oncology Group,[16,17] with small intraobserver variability, making them more suitable for radiation treatment planning. Parts of the carefully annotated data set of 16 organ annotations (150 volumes from the development cohort and 20 volumes from LiTS2017) will be released to boost this research task. In short, our data sets are more promising for developing a robust segmentation model for clinical application.

We propose combining data augmentation and knowledge distillation for deep network training to obtain accurate and robust segmentations. First, a series of data augmentation strategies could simulate more challenging scenarios to boost the robustness of networks. Furthermore, we used knowledge distillation to minimize the difference between the teacher and the student. In general, the student's input is more challenging than that of the teacher, and the teacher's output is more accurate than that of the student. We encouraged student outputs to be consistent with their teachers, reminding them to pay more attention to the common anatomic context rather than the variance of the intensity distribution. Based on these approaches, the AbsegNet can learn from 1 data center and generalize well to many unseen centers. Different from those previous works, which focused on improving network performance on a single data center,[23-25] the AbsegNet considered the domain shift between different centers and used data augmentation and knowledge distillation to boost the models' generalization. In addition, the proposed training strategy can improve performance by a large margin compared with the standard 3D-UNet model,[23] further suggesting the efficiency and effectiveness of the proposed approach.

After finishing construction and training, the performance of AbsegNet was first validated in cohort 1. It was demonstrated that good segmentation results were acquired for most OARs, with the exception of the duodenum and adrenals. Volumetric overlaps between AbsegNet-generated and ground truth contours were around 95% for solid organs (liver, spleen, and kidneys). And for all OARs, overall mDSC reached 87%, and overall mHD95 was lower than 10 mm. In contrast to previous DL methods, AbsegNet was prone to generate more accurate contours for most OARs, particularly for gastrointestinal organs that are difficult to delineate because of their huge and variable spatial information.

AbsegNet was tested on 2 completely unseen data acquired from 2 hospitals. In such heterogeneous data, AbsegNet still acquired comparable segmentation results, with the mDSC approaching that which was obtained in the internal cohort 1 (overall mDSC, cohorts 1, 2, and 3: 86.73% vs 85.65% vs 88.04%), and it outperformed 5 existing methods. On the contrary, the performance of previous algorithms was unstable, affected by data differences. As observed, the mDSC of Attention-UNet in cohort 1 was 84.67%. However, mDSC considerably reduced to 79.63% in cohort 3. Moreover, compared with previous methods, AbsegNet achieved smaller standard deviations for DSC and HD95 overall, suggesting AbsegNet is less affected by individual differences and confirming its robustness for delineating abdominal OARs.

Comparable accuracy was obtained in AbsegNet by contrasting with historical results for multiple abdominal OAR delineation. It should be noticed that our results were acquired on heterogeneous CT scans and patients, more closing to the clinical setting, whereas those[1,9] were acquired on more homogeneous data. As indicated, different scanners and CT phases on patients with heterogeneous lesions could lead to obvious variances in organ appearances, resulting in a degradation of model performance. Hence, results showed the powerful generalizability of AbsegNet.

To further confirm the accuracy of AbsegNet clinically, 2 independent experts were invited to review AbsegNet-generated OARs. AbsegNet-produced contours for solid organs (liver, kidneys, and spleen) of all patients did not need to be modified. Only 15.6%, 12.5%, 10%, and 6.2% of patients with autosegmentations of the colon, small bowel, duodenum, and gallbladder, respectively, required major revisions. For other OARs, only a small portion of patients needed minor or moderate revisions. In a subjective evaluation by the 2 experts together, only several patients with OARs, such as duodenum, adrenals, and colon, were unacceptable, whereas most patients with most OARs were considered totally acceptable or acceptable. Moreover, we showed that, with the aid of the model, delineating efficiency was substantially improved, saving the delineation time by as much as over 65%. These results indicated that most OARs produced by AbsegNet were clinically applicable and could be used for radiation therapy.

This study had several limitations. First, to improve the consistency of OAR delineation, only 1 experienced expert was invited to delineate these contours, which introduced potential subjective variation. Second, although most OARs obtained high accuracy, the performance of AbsegNet on the duodenum was unsatisfactory. The possible reason might be that the duodenum has the most complex anatomic structure, consisting of 4 parts: the superior, the descending, the horizontal, and the ascending. Hence, the organ volume and location can vary dramatically, posing a great challenge for DL models. Similar results were also observed in the study by Gibson et al,[14] with an mDSC of 63%. There is much room to improve the segmentation accuracy for this organ. In the future study, we are going to experiment with different loss functions to optimize our model, hoping to obtain better results on lower-performing structures. Third, when performing a clinical evaluation, 2 experts were required to review complementary OARs (7 OARs for expert A and 9 OARs for expert B) in 2 cohorts rather than each checking all OARs in each cohort because of the heavy workload and time requirements. This might have had a certain effect on the evaluative accuracy of AbsegNet. Alternatively, pelvic bone delineation was not involved in this study, though it is also an important OAR for abdominal radiation therapy.[37] We propose to incorporate it in future research.

Conclusions

In summary, we proposed a novel, fully automatic DL model to delineate whole abdominal OARs. Despite heterogeneous CT scans and individual differences, our findings showed that most OARs produced by AbsegNet were accurate and robust. It is clinically applicable and helpful to facilitate radiation treatment planning and workflow with ongoing efforts.

References

1. Chen X, Sun S, Bai N, et al. A deep learning-based auto-segmentation system for organs-at-risk on whole-body computed tomography images for radiation therapy. *Radiother Oncol* 2021;160:175-184.
2. Joskowicz L, Cohen D, Caplan N, et al. Inter-observer variability of manual contour delineation of structures in CT. *Eur Radiol* 2019;29:1391-1399.
3. Peng YL, Chen L, Shen GZ, et al. Interobserver variations in the delineation of target volumes and organs at risk and their impact on dose distribution in intensity-modulated radiation therapy for nasopharyngeal carcinoma. *Oral Oncol* 2018;82:1-7.
4. Spoelstra FO, Senan S, Le Péchoux C, et al. Variations in target volume definition for postoperative radiotherapy in stage III non-small-cell lung cancer: Analysis of an international contouring study. *Int J Radiat Oncol Biol Phys* 2010;76:1106-1113.
5. Vinod SK, Jameson MG, Min M, et al. Uncertainties in volume delineation in radiation oncology: A systematic review and recommendations for future studies. *Radiother Oncol* 2016;121:169-179.
6. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. *Med Image Anal* 2017;42:60-88.
7. Zhou SK, Greenspan H, Davatzikos C, et al. A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. *Proc IEEE* 2021;109:820-838.
8. Humpire-Mamani GE, Bukala J, Scholten ET, et al. Fully automatic volume measurement of the spleen at CT using deep learning. *Radiol Artif Intell* 2020;2 e190102.
9. Weston AD, Korfiatis P, Philbrick KA, et al. Complete abdomen and pelvis segmentation using U-net variant architecture. *Med Phys* 2020;47:5609-5618.
10. Ma J, Zhang Y, Gu S, et al. AbdomenCT-1K: Is abdominal organ segmentation a solved problem? *IEEE Trans Pattern Anal Mach Intell* 2022;44:6695-6714.
11. Cardenas CE, Beadle BM, Garden AS, et al. Generating high-quality lymph node clinical target volumes for head and neck cancer radiation therapy using a fully automated deep learning-based approach. *Int J Radiat Oncol Biol Phys* 2021;109:801-812.
12. van Dijk LV, Van den Bosch L, Aljabar P, et al. Improving automatic delineation for head and neck organs at risk by deep learning contouring. *Radiother Oncol* 2020;142:115-123.
13. Rister B, Yi D, Shivakumar K, et al. CT-ORG, a new dataset for multiple organ segmentation in computed tomography. *Sci Data* 2020;7:381.
14. Gibson E, Giganti F, Hu Y, et al. Automatic multi-organ segmentation on abdominal CT with dense V-networks. *IEEE Trans Med Imaging* 2018;37:1822-1834.
15. Antonelli M, Reinke A, Bakas S, et al. The medical segmentation decathlon. *Nat Commun* 2022;13:1-13.
16. Gay HA, Barthold HJ, O'Meara E, et al. Pelvic normal tissue contouring guidelines for radiation therapy: A Radiation Therapy Oncology Group consensus panel atlas. *Int J Radiat Oncol Biol Phys* 2012;83:e353-e362.
17. Jabbour SK, Hashem SA, Bosch W, et al. Upper abdominal normal organ contouring guidelines and atlas: A Radiation Therapy Oncology Group consensus. *Pract Radiat Oncol* 2014;4:82-89.
18. Pukala J, Johnson PB, Shah AP, et al. Benchmarking of five commercial deformable image registration algorithms for head and neck patients. *J Appl Clin Med Phys* 2016;17:25-40.
19. Nakajima Y, Kadoya N, Kanai T, et al. Evaluation of the effect of user-guided deformable image registration of thoracic images on registration accuracy among users. *Med Dosim* 2020;45:206-212.
20. Zhang L, Wang X, Yang D, et al. Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation. *IEEE Trans Med Imaging* 2020;39:2531-2540.
21. Gou J, Yu B, Maybank SJ, et al. Knowledge distillation: A survey. *Int J Comput Vis* 2021;129:1789-1819.
22. Tarvainen A, Valpola HJ. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Adv Neural Inf Process Syst* 2017;30:1-10.
23. Ballestar LM, Vilaplana V. *MRI Brain Tumor Segmentation and Uncertainty Estimation Using 3D-UNet Architectures. International MICCAI Brainlesion Workshop*. Switzerland: Springer; 2020:376-390.
24. Chen LC, Zhu Y, Papandreou G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Switzerland AG: Springer Nature; 2018:801-818.
25. Schlemper J, Oktay O, Schaap M, et al. Attention gated networks: Learning to leverage salient regions in medical images. *Med Image Anal* 2019;53:197-207.
26. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. 234-241.
27. Tang Y, Yang D, Li W, et al. Self-supervised pre-training of swin transformers for 3d medical image analysis. *Proc IEEE/CVF Conf Comput Vis Pattern Recognit* 2022;20730-20740.
28. Liang S, Tang F, Huang X, et al. Deep-learning-based detection and segmentation of organs at risk in nasopharyngeal carcinoma computed tomographic images for radiotherapy planning. *Eur Radiol* 2019;29:1961-1967.
29. Liu Z, Liu X, Guan H, et al. Development and validation of a deep learning algorithm for auto-delineation of clinical target volume and organs at risk in cervical cancer radiotherapy. *Radiother Oncol* 2020;153:172-179.
30. Vrtovec T, Močnik D, Strojan P, et al. Auto-segmentation of organs at risk for head and neck radiotherapy planning: From atlas-based to deep learning methods. *Med Phys* 2020;47:e929-e950.
31. Tang H, Chen X, Liu Y, et al. Clinically applicable deep learning framework for organs at risk delineation in CT images. *Nat Mach Intell* 2019;1:480-491.
32. Qayyum A, Lalande A, Meriaudeau F. Automatic segmentation of tumors and affected organs in the abdomen using a 3D hybrid model for computed tomography imaging. *Comput Biol Med* 2020;127 104097.
33. HR Roth, H Oda, Y Hayashi, et al., Hierarchical 3D fully convolutional networks for multi-organ segmentation, Computer Vision and Pattern Recognition (CVPR), 2017, Arxiv, *arXiv*: https://doi.org/10.48550/arXiv.1704.06382.
34. Landman B, Xu Z, Igelsias J, et al. Miccai multi-atlas labeling beyond the cranial vault—workshop and challenge, *Proc. MICCAI*

Multi-Atlas Labeling Beyond Cranial Vault—Workshop. Challenge 2015;12.

35. Kavur AE, Gezer NS, Barıș M, et al. CHAOS Challenge - combined (CT-MR) healthy abdominal organ segmentation. *Med Image Anal* 2021;69 101950.
36. Heller N, Isensee F, Maier-Hein KH, et al. The state of the art in kidney and kidney tumor segmentation in contrast-enhanced CT imaging: Results of the KiTS19 challenge. *Med Image Anal* 2021;67 101821.
37. Albuquerque K, Giangreco D, Morrison C, et al. Radiation-related predictors of hematologic toxicity after concurrent chemoradiation for cervical cancer and implications for bone marrow-sparing pelvic IMRT. *Int J Radiat Oncol Biol Phys* 2011; 79:1043-1047.