

Contour interpolation by deep learning approach

Chenxi Zhao,^{a,*} Ye Duan,^a and Deshan Yang^b

^aUniversity of Missouri, Electrical Engineering and Computer Science Department,
Columbia, Missouri, United States

^bDuke University, Department of Radiation Oncology, Durham,
North Carolina, United States

Abstract

Purpose: Contour interpolation is an important tool for expediting manual segmentation of anatomical structures. The process allows users to manually contour on discontinuous slices and then automatically fill in the gaps, therefore saving time and efforts. The most used conventional shape-based interpolation (SBI) algorithm, which operates on shape information, often performs suboptimally near the superior and inferior borders of organs and for the gastrointestinal structures. In this study, we present a generic deep learning solution to improve the robustness and accuracy for contour interpolation, especially for these historically difficult cases.

Approach: A generic deep contour interpolation model was developed and trained using 16,796 publicly available cases from 5 different data libraries, covering 15 organs. The network inputs were a $128 \times 128 \times 5$ image patch and the two-dimensional contour masks for the top and bottom slices of the patch. The outputs were the organ masks for the three middle slices. The performance was evaluated on both dice scores and distance-to-agreement (DTA) values.

Results: The deep contour interpolation model achieved a dice score of 0.95 ± 0.05 and a mean DTA value of 1.09 ± 2.30 mm, averaged on 3167 testing cases of all 15 organs. In a comparison, the results by the conventional SBI method were 0.94 ± 0.08 and 1.50 ± 3.63 mm, respectively. For the difficult cases, the dice score and DTA value were 0.91 ± 0.09 and 1.68 ± 2.28 mm by the deep interpolator, compared with 0.86 ± 0.13 and 3.43 ± 5.89 mm by SBI. The t-test results confirmed that the performance improvements were statistically significant ($p < 0.05$) for all cases in dice scores and for small organs and difficult cases in DTA values. Ablation studies were also performed.

Conclusions: A deep learning method was developed to enhance the process of contour interpolation. It could be useful for expediting the tasks of manual segmentation of organs and structures in the medical images.

© 2022 Society of Photo-Optical Instrumentation Engineers (SPIE) [DOI: [10.1117/1.JMI.9.6.064003](https://doi.org/10.1117/1.JMI.9.6.064003)]

Keywords: medical imaging segmentation; deep learning; contour interpolation.

Paper 22002GRR received Jan. 1, 2022; accepted for publication Dec. 2, 2022; published online Dec. 21, 2022.

1 Introduction

Complete manual delineation of organs and other structures in medical images is labor intensive and often very time consuming, especially for time sensitive situations, e.g., online plan adaptation.¹ Computer assistance is very desirable, and autosegmentation is an important potential answer.² The new deep learning autosegmentation models have shown remarkable performance improvements in recent years;³ however, their robustness and accuracy are still inadequate for difficult cases. In addition, many organs or structures are still not supported [e.g., the gastrointestinal (GI) organs and tumor targets]. Because of the current challenges in robustness, accuracy, and breadth for autosegmentation models, most organ segmentations are still performed manually, and the results autosegmentation tools are manually evaluated and

*Address all correspondence to Chenxi Zhao, cz3d6@mail.missouri.edu

corrected before the results are used to support diagnosis and treatments in clinical settings, for example, in radiation therapy.⁴

Contour interpolation is an important tool for expediting the manual segmentation process by allowing users to manually contour on discontinuous two-dimensional (2D) slices, for example, on one of every three slices, and then automatically fill in the gaps. In the era of auto-segmentation using deep-learning models, contour interpolation is still playing an important role in (1) manually generating ground truth labels and (2) aiding manual segmentation of new or difficult anatomical structures that are not yet supported by deep learning models, for example, tumor targets and the small and large intestines. Conventional methods for image slice interpolation can be roughly grouped in three categories—contour-based, intensity-based, and shape-based methods.⁵ The contour-based interpolation methods are often used in surface reconstruction, which takes a set of binary images representing cross-sectional boundaries of an object.⁶ In intensity-based interpolation, the interpolation results are computed from the intensity values of the input images directly and on the organ contour masks for the contour interpolation cases. Linear interpolation is typically used.⁷ The shape-based interpolation (SBI) methods are explained in Fig. 1. The SBI methods are simple and versatile, thus suitable and commonly used for medical structure interpolation. The basic SBI method has mainly three steps:⁸ (1) compute the 2D distance maps for the manual contours on the top and the bottom slices, separately. (2) The distance map for an in-between slice is computed by linearly interpolating the top and bottom distance map based on the distance from the intermediate slice to the top and bottom slices. (3) The interpolated contour on the intermediate slice is computed by thresholding the interpolated distance map at 0.

The effects of saving the manual contouring effort depend on the accuracy and robustness of the contour interpolation method. The basic SBI method is simple and universally applicable to any contour interpolation case because it uses nothing but the shapes of manual contours. Although it works well for most situations, it has difficulties when the two to-be-interpolated manual contours differ dramatically, especially near the superior and inferior borders of organs and for the gastrointestinal structures. An example of such a difficult case is shown in Fig. 2. In the authors' opinion, a major reason for such difficulties is that only the distance to the manually delineated contours is considered in the interpolation process. Other potentially useful information, e.g., the image intensity similarity, image contrast, structure shape, are not utilized.

Multiple methods have been proposed to improve the contour interpolation performance. Lutufu et al.⁹ proposed a method in 1992 to combine shape distance and image gray level for interpolation (CSGI). However, the CSGI method was not robust. It incorporated the image intensity information into the interpolation process, but the incorporation was neither accurate nor smooth. This algorithm also requires an extensive manual configuration; therefore, it is not universal by default. Albu et al.¹⁰ proposed a morphology-based interpolation (MBI) method in

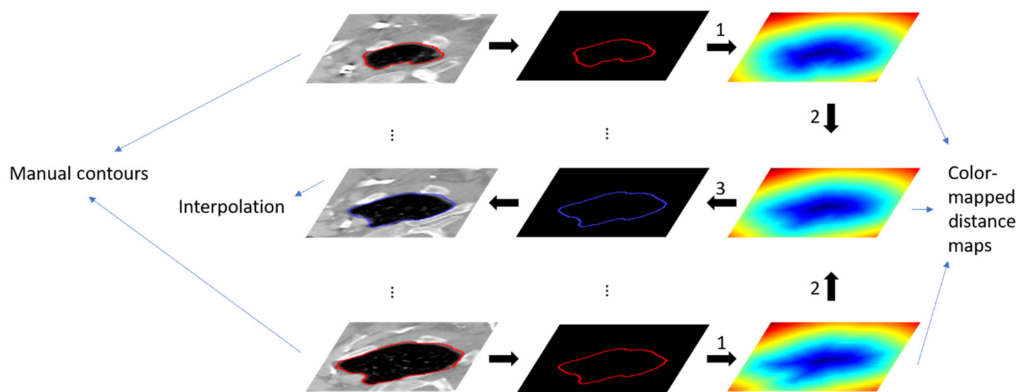


Fig. 1 Illustration of the SBI method. The manual contours are in red on the top and bottom slices. The interpolation results (the interpolated color map and the interpolated contour by thresholding the interpolated distance map at 0) are shown in the middle row. The figures on the right are color-mapped distance maps that take the value of zero (green) on the contours, positive values (green to red) outside the contours, and negative values (blue to green) inside the contours.

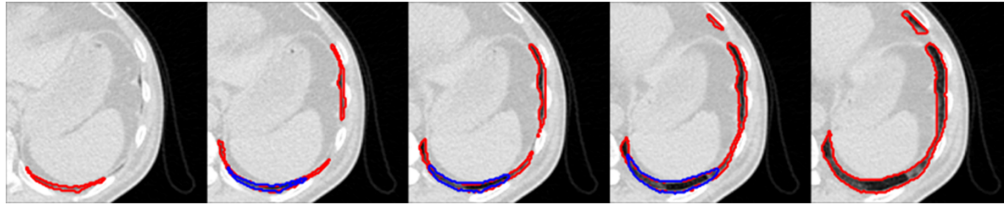


Fig. 2 An example of unsatisfactory results by the basic SBI method for slices near the bottom of the left lung. The ground truth manual contours are shown in red, and the interpolation results are shown in blue on the three middle slices. The interpolation results were generated using SBI to interpolate the manual contours on the first and last slices.

2008 to generate smooth interpolation between interslice containing one or more regions. This method was later extended to N -dimension by Zukic et al.¹¹ in 2016 and implemented in the insight toolkit. Liao et al. improved the MBI method in 2011 by adding image intensity-based classification. The new method, called morphology-based interpolation with local intensity information (MBILII),⁵ as demonstrated, outperformed MBI, modified cubic spline method,¹² and CSGI. Combining the binary weighted averaging for contour-based interpolation and random forest for intensity-based classification, another automatic method improved the accuracy of the interpolation more than morphology.¹³ Also employed in tumor segmentation, interpolation based on radial basis functions (RBF) distance map was used in a three-dimensional (3D) reconstruction in prostate cancer studies.¹⁴ To the authors' understanding, these more complicated methods that use image gray-level information solely or combined with shape distance information only work relatively well if the gray level of the segmented structure is uniform and clearly separated from the gray level of the background. This is often not the case in medical images in which adjacent organs have very similar voxel values. The basic SBI method often performs better for abdominal organs, e.g., liver and spleen, that have minimal gray-level difference from the surrounding abdominal tissues and organs and for GI organs that do not have consistent gray level.

Deep convolutional neural network (DCNN) methods have been successfully used for image segmentation and have shown greatly improved performances over the conventional image segmentation methods.¹⁵ To the authors' knowledge, no DCNN method has been proposed for contour interpolation. In this study, we propose a new DCNN-based contour interpolation method. The contour interpolation schemes (the slice distance between the two manual contours, the general applicability to any organs) were optimally selected. The DCNN model was designed to provide a balanced performance between model prediction accuracy and robustness, be universal applicability to any organs, and be computational efficiency for supporting real-time user interactive contouring. Both shape distance information and image intensity information are utilized together in the new method to achieve the improved contour interpolation performance, especially for the cases that are usually difficult for the conventional methods.

2 Materials and Methods

2.1 Datasets

A total of 1174 datasets acquired from 5 different data libraries were used in this study. Each dataset contains the computerized tomography (CT) image and manual contours of one or multiple organs of the 15 organs covered by this study. The CT images are in $512 \times 512 \times S$, where S is the number of CT axial slices. The S ranges from 100 to 400. The slice thickness of different CTs was from 0.7 to 2 mm.

Table 1 provides the details about contoured organs on datasets from different libraries. Fifteen different organs, from the thorax to the pelvis, were covered by the 1174 datasets from 5 different data libraries. Specifically, the non-small cell lung cancer (NSCLC)-radiomics¹⁶ data library has contours of lungs, spinal cord, esophagus, and heart. The beyond the cranial vault (BTCV)¹⁷ data library has contours of spleen, kidneys, gallbladder, esophagus, liver, stomach, aorta, inferior vena cava, veins, pancreas, and duodenum. The liver tumor segmentation (LiTS)¹⁸

Table 1 The list of publicly available benchmark data libraries used in this study.

Sl. No	Data libraries	Contoured organs	Number of datasets	Year
1	NSCLC-radiomics ¹⁶	Lung, esophagus, spinal cord, and heart	422	2019
2	BTCV ¹⁷	Spleen, right kidney, left kidney, gallbladder, esophagus, liver, stomach, aorta, inferior vena cava, veins, pancreas, and duodenum	450	2019
3	LITS ¹⁸	Liver	130	2017
4	Pancreas-CT ¹⁹	Spleen, left kidney, gallbladder, esophagus, liver, stomach, pancreas, and duodenum	82	2016
5	Medical segmentation decathlon ²⁰	Colon	190	2019

data library has liver contours. The pancreas-CT¹⁹ data library has spleen, left kidney, gallbladder, esophagus, liver, stomach, pancreas, and duodenum contours. Medical segmentation decathlon²⁰ data library contains colon contours.

Overall, the collection of CT and organ contour datasets from multiple data libraries was diverse and comprehensive. It covered clinical CT images from a wide range of settings and situations, including various image quality levels, image noise levels, in-plane pixel sizes, slice thicknesses, and ranges of human anatomy. The diversity of the training datasets was useful for ensuring the generality of the trained deep interpolator models. In the future work we plan to apply additional image modalities such as magnetic resonance imaging (MRI). Contours of additional organs in other regions of body such as head and neck, could also be adopted in further improvement.

The ground truths were contoured manually by experts and are available for most of the slices. If one slice happened to be not contoured, it was removed in the preprocessing procedure.

2.2 Preparation of Patches

Following the model design, the following data preprocessing procedure was applied to convert the full volumes of CT image and the structure mask to $128 \times 128 \times 5$ patches to be ready for model training and evaluation. For each organ in each dataset, (1) a 5-slice image was selected for preprocessing only if the corresponding ground truth was available for all of the continuous slices. Both the CT image and structure mask were cropped based on the 3D volume of the organ. The kidneys and lungs were separated into left and right parts. In each slice, a bounding square was utilized to crop the organs in the original image according to the sizes of the organs with an additional 10-pixel axial margin and a 2-pixel superior-inferior margin. (2) The cropping was applied in the axial view. Each axial slice was resampled to 128×128 . There was no other preprocessing procedure in the sagittal view. (3) The CT image intensity values are in the range of 0 to 4191, and the voxel intensity to water is normalized at 1000. The raw voxel values of masks were 0 or 1. To keep the mask image value intensity consistent with the water voxel intensity in the CT images, the masks were multiplied by 1000 element wisely. (4) The three middle slices were duplicated so that five slices (the top and bottom slices with manual contours, and three middle slices) were converted to eight slices before the data were fed to the DCNN model. This step was needed because the model required the patch size to be on an order of 2 in each dimension to be computationally efficient. Figure 3 is an illustration of the preprocessing procedure.

After preprocessing, 19,963 patches were generated from the 1174 original full volume datasets. 840 patches (5%) were selected randomly for validation during model training. 3167 patches of the randomly selected original (15%) patches were separated for testing. The different patches of the same patient's same organ were not used for both training and validation.

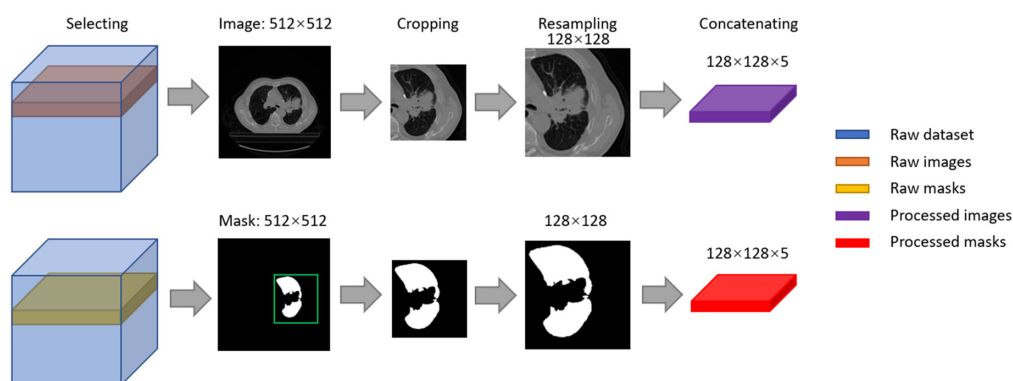


Fig. 3 An illustration of the preprocessing procedure. A five continuous slice raw patch was selected from the dataset and preprocessed. In each slice, the mask was cropped with the assistance of the bounding square (green), and the image was cropped in the same position. Next, it was sampled to 128×128 . After all five slices were processed, they were stacked together as $128 \times 128 \times 5$ and used as input.

2.3 DCNN Model Design and Considerations

We designed our deep interpolator model by adapting a three-layer 3D U-Net model.²¹ Figure 4 shows the procedure of our contour interpolation method. The network has two inputs and a single output. The first input is a $128 \times 128 \times 5$ 3D patch of the CT image with five continuous axial slices. The second input is the corresponding $128 \times 128 \times 5$ patch of the manual contour mask, in which only the top and the bottom slices have the ground truth organ contours, and the middle three slices are empty. These two contours on the top and bottom slices serve as the reference contours for the interpolation process to produce the contours for the middle three slices, as explained in Fig. 1. The output of the model is the predicted organ mask for the middle three slices. Compared with the conventional SBI method, which only uses the shape information, both the image and the shapes (i.e., the manual contours) are provided to the network model, thus providing the possibility of the improved interpolation performance.

The deep CNN model was designed under the following considerations.

1. Three middle slices were assumed between the two slices with manual contours. The slice distance between the two to-be-interpolated manual contours is probably the most important factor to affect the efficiency of using contour interpolation to expedite the manual contouring. The slice distance also directly affects the performance of any given contour interpolation algorithm.

According to our preliminary results, the selection of the three middle slices was an optimal choice. It ensured that the two to-be-interpolated manual contours were not too far away; otherwise, the interpolation accuracy would be compromised. It also ensured

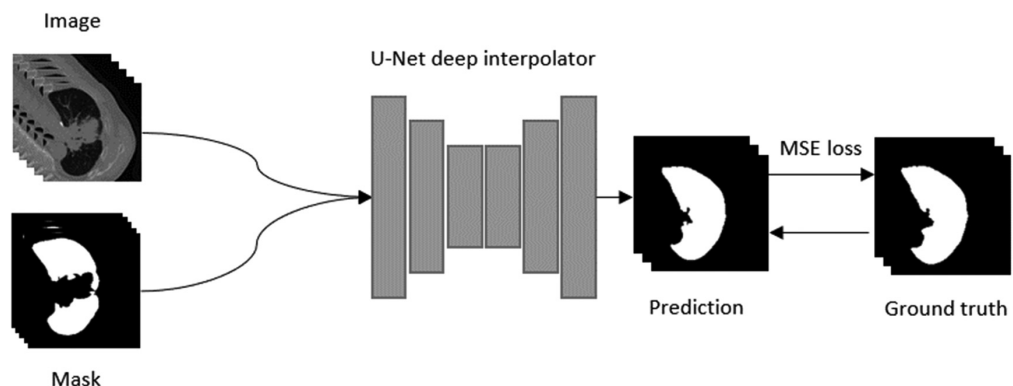


Fig. 4 An overview of our contour interpolating procedure.

- that the manual contours were not too close; otherwise, the effectiveness of saving the manual contour efforts using contour interpolation would be reduced.
2. The axial patch size 128×128 was chosen to support various organ sizes and to ensure both the pixel resolution and the overall speed of the computation. An organ usually occupies only a portion of the entire CT slice of 512×512 pixels. The computation should be focused on the organ contours and not on voxels far outside the organ contours. It would be very inefficient to include the entire CT slice into the computation regardless of the organ size. In addition, the DCNN model was designed to be universal for all organs so that there would be a single trained DCNN model to be easily managed and readily applied to any contour interpolation cases. However, the size of different organs (listed in Table 1) varies significantly. For example, the liver and the lung are much larger than the spinal cord on the axial slice. A medium patch size of 128×128 was therefore chosen to support the dramatic organ size differences. For all organs with different sizes on the axial slice, the image and contour mask were resampled to fit in the 128×128 patch size. The details were explained above in the data preprocessing section. To be explained in the discussion, an alternative option is to train and manage multiple different DCNN models for supporting different axial patch sizes.
 3. The DCNN model, based on the 3D U-Net architecture, was intentionally designed to be shallow (only three stages) and narrow (only eight channels in the first encoder and the convolution filter size = 3) to ensure fast computation. The final computation speed is important because the contour interpolation should be completed instantaneously to support the interactive manual contouring task in near real time. We chose the U-Net architecture due to its reported efficiency and performance in the medical image segmentation literature. The normalization layers were all removed because the data were already normalized.
 4. After the DCNN model is trained for predicting the contours on the three middle slices, a different number of middle slices (1, 2, or >3) will still be supported when the trained model is applied for contour interpolation. To do so, the middle slices are duplicated or subsampled to make up six middle slices. For example, if there is only a single middle slice, it will be duplicated six times. If there are two middle slices, they will be duplicated three times each. If there are more than three middle slices, contour interpolation will be carried out in multiple runs and each run will only cover three middle slices. For example, five middle slices can be covered by two runs, with the slices 1, 3, and 5 covered by the first run, and the slices 1, 2, and 4 covered by the second run.
 5. The transverse colon were very difficult to interpolate because it is on only a few axial slices instead of going through many axial slices in the superior–inferior direction, whereas the deep interpolator is designed to interpolate between slices in the superior–inferior direction. Therefore, we used 3-slice patches, instead of the regular 5-slice patches, for the transverse colon. To make a $128 \times 128 \times 3$ patch into a $128 \times 128 \times 8$ patch, the top and bottom slices were duplicated two times and the single middle slice was copied four times. The model output was still in the format of $128 \times 128 \times 3$, but the middle slice of the output was kept as the final single-slice interpolation result.

2.4 Model Implementation, Training, and Performance Evaluation

The model was implemented using MATLAB 2020b. The models were trained on a laptop computer equipped with a GTX 1650 GPU. The Adam optimizer was used to train the 3D U-Net for 20 epochs with the following hyperparameters: mini batch size = 16, initial learning rate = 0.001, and learn rate schedule = piecewise.

We evaluated the model prediction performance on the preprocessed 3167 testing patches. The dice scores and the distance-to-agreement (DTA) values were computed referring to the known ground truth contours for each case. The dice coefficient is a measure of overlap between the predicted contours and the ground truth manual contour. The DTA measures the per voxel distance between the predicted contour and the ground truth contour.²² The average values of the maximum, the mean, and the 95th percentile of per-case DTA values were computed.

As a comparison, contour interpolations were also performed using three conventional methods: SBI,⁸ MBI,¹⁰ and MBILII.⁵ Unfortunately, the implementation for these methods is not publicly available; therefore, we implemented these methods ourselves based on the original paper. Dice scores and DTA values were computed for the results and compared. A student t-test was conducted to test the statistical significance between the results of our deep interpolator and the SBI method. Because our implementation for MBI and MBILII may not be optimal, the comprehensive comparison did not include them, and it will be our future work.

3 Result

Figure 5 shows the interpolation results for 12 selected organs. The deep contour interpolator worked well visually for all organs, regardless of the organ and the organ size, as one can see the closeness between the interpolated contours in green color and the blue ground truth 3D surfaces. Inferior vena cava and veins were not included in Fig. 5 because they had a similar shape and scan appearance to aorta. Colon was not included because the number of slices was not enough for rendering 3D surfaces and the result was suboptimal.

The computed dice scores are listed in Table 2. Esophagus and spinal cord were counted together. Aorta, inferior vena cava, and veins were counted together. Our deep interpolator performed better in each group with higher dice scores and a smaller standard deviation than the basic SBI method.

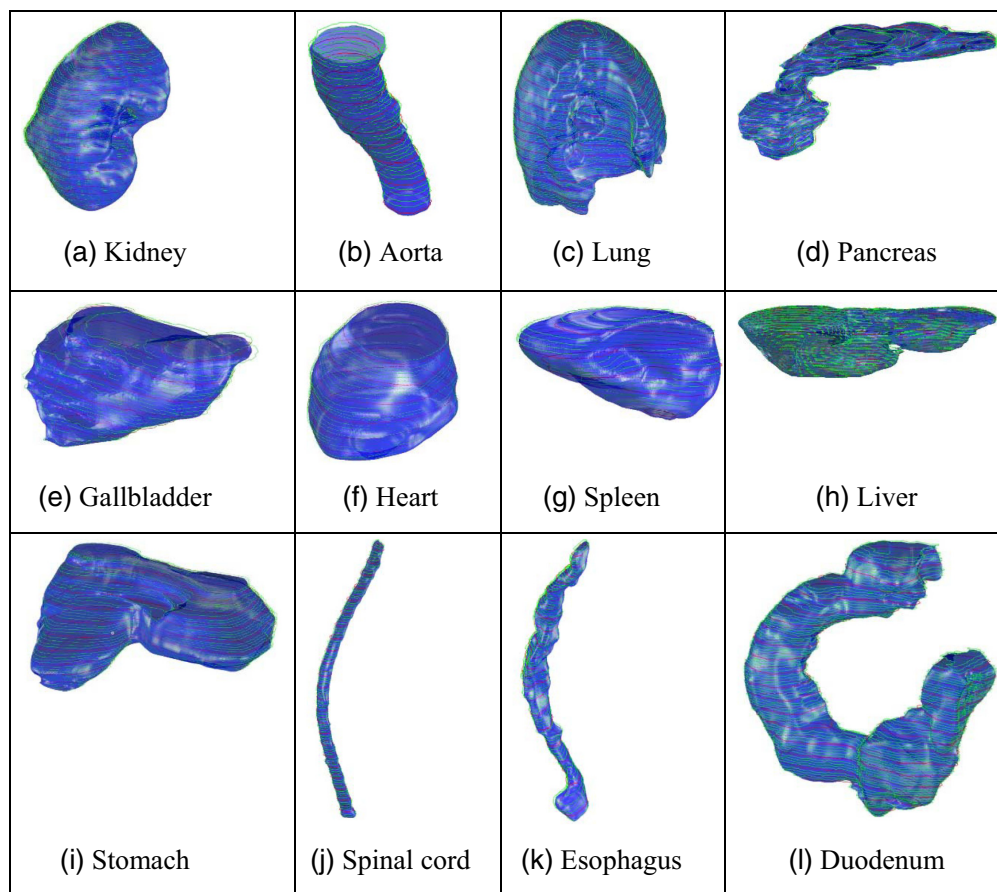


Fig. 5 Examples of contour interpolation results of 12 selected organs: (a) kidney, (b) aorta, (c) lung, (d) pancreas, (e) gallbladder, (f) heart, (g) spleen, (h) liver, (i) stomach, (j) spinal cord, (k) esophagus, and (l) duodenum. In each plot, the blue surface is the ground truth of the whole organ. The red curves are the ground truth contours on the 2D slices. The green curves are the interpolated contours.

Table 2 Comparison of dice scores by the conventional SBI method and the proposed deep contour interpolator.

Organs	3D U-Net deep contour interpolator	SBI method	Dice score differences	t-Test p -values
Lung	0.96 ± 0.07	0.93 ± 0.13	0.03 ± 0.10	$5.70e^{-10}$
<i>E</i> and <i>Sa</i>	0.94 ± 0.04	0.93 ± 0.05	0.01 ± 0.02	$4.62e^{-39}$
Heart	0.94 ± 0.06	0.93 ± 0.07	0.01 ± 0.03	$3.66e^{-05}$
Duodenum	0.95 ± 0.05	0.94 ± 0.07	0.01 ± 0.03	$5.88e^{-04}$
Stomach	0.97 ± 0.03	0.96 ± 0.05	0.01 ± 0.02	$5.46e^{-07}$
Spleen	0.98 ± 0.02	0.96 ± 0.05	0.01 ± 0.04	$3.02e^{-09}$
<i>A</i> , <i>I</i> , and <i>Vb</i>	0.91 ± 0.09	0.88 ± 0.15	0.03 ± 0.07	$4.66e^{-04}$
Liver	0.98 ± 0.02	0.98 ± 0.03	0.00 ± 0.01	$9.72e^{-17}$
Kidney	0.97 ± 0.03	0.96 ± 0.06	0.01 ± 0.04	$3.95e^{-08}$
Gallbladder	0.94 ± 0.05	0.92 ± 0.11	0.02 ± 0.06	0.01
Pancreas	0.88 ± 0.05	0.86 ± 0.08	0.02 ± 0.04	$2.16e^{-08}$
Colon	0.81 ± 0.10	0.66 ± 0.15	0.15 ± 0.14	$1.03e^{-05}$
Difficult cases	0.91 ± 0.09	0.86 ± 0.13	0.05 ± 0.09	$3.83e^{-15}$
All	0.95 ± 0.05	0.94 ± 0.08	0.01 ± 0.05	$1.71e^{-60}$

^aEsophagus and spinal cord.

^bAorta, inferior vena cava, and veins.

We checked the SBI method result and noticed that the performance was always not satisfying when the shapes of the contoured organs were not regular, especially in the superior and inferior parts. Comparing with the regular shape and suboptimal performance of the SBI method, we selected 286 difficult cases from relatively bigger organs including lung, liver, kidney, pancreas, stomach, gallbladder, and duodenum. The results for the difficult cases are listed in a separate row. The t-test for every group confirmed that the difference between the two result groups was statistically significant, with all p -values being <0.05 . This dice score comparison suggests that our deep interpolator was more accurate and robust. A few examples for such difficult cases, as shown in Fig. 6, suggest that our deep interpolator performed visually better than the basic SBI method, as the results by our deep interpolator are visually closer to the ground truth manual contours.

We computed both 2D DTA (on the axial slice) and 3D DTA values. 3D DTA values were computed among the three slices, and the 2D DTA values were computed in each individual slice. Both are clinically relevant evaluation metrics. The DTA results are shown in Tables 3 and 4. In Tables 3 and 4, smaller DTA values represent the better results of the method. We computed the difference values between DTA from the deep interpolator and SBI method in each group. Our deep interpolator overperformed the SBI method as all of the difference values are negative.

The t-test results, as shown in Table 5, indicate that our deep interpolator performed significantly better overall, for most DTA measurements and for the difficult cases, with p -values < 0.05 . However, the performance differences were not statistically significant for a few easy organs, e.g., heart and liver, with p -values > 0.05 .

The 3D views of lung and stomach are shown in Fig. 7; they demonstrate that our deep interpolator performed well in the regions that are usually difficult for the basic SBI method. The difficult regions were the inferior portions of the organs in these two cases, as indicated by arrows. The deep contour interpolator uses both image intensity and organ contour information.

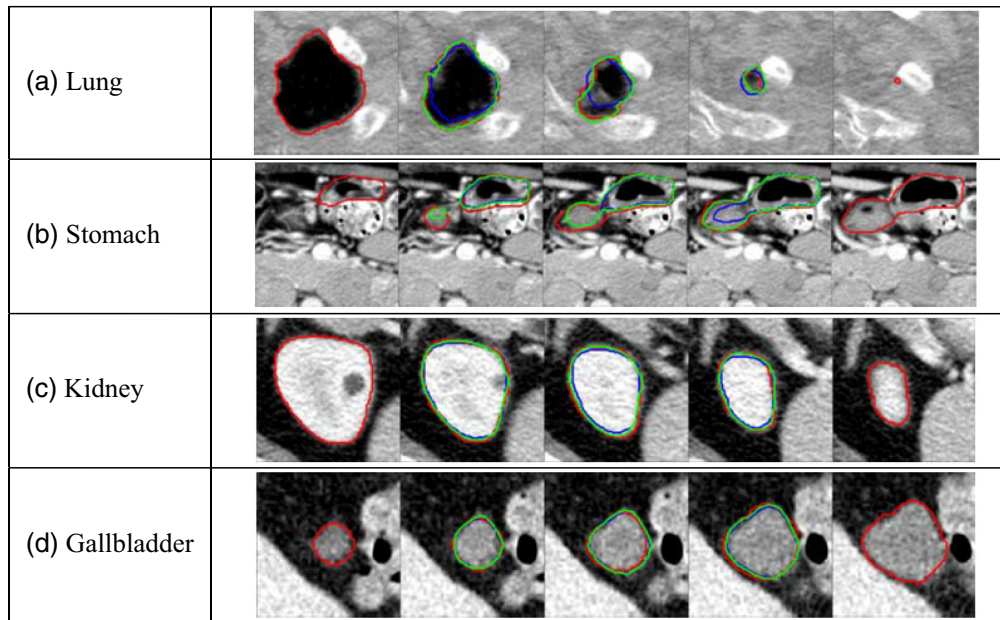


Fig. 6 Contour interpolation examples of a few difficult cases: (a) lung, (b) stomach, (c) kidney, and (d) gallbladder. The ground truth manual contours are in red. The SBI interpolation results are in blue. The deep interpolator results are in green.

Table 3 Mean DTA values and the differences (mm).

Name	DI ^a (3D)	DI (2D)	SBI ^b (3D)	SBI (2D)	Diff ^c (3D)	Diff (2D)
Lung	0.78 ± 1.98	0.55 ± 1.92	1.88 ± 4.25	1.07 ± 2.29	-1.10 ± 3.78	-0.52 ± 1.99
<i>E</i> and <i>S</i>	1.19 ± 0.61	0.87 ± 0.62	1.30 ± 0.66	0.96 ± 0.67	-0.11 ± 0.29	-0.09 ± 0.38
Heart	1.75 ± 1.46	1.07 ± 0.81	2.03 ± 1.74	1.22 ± 1.20	-0.28 ± 0.82	-0.15 ± 0.82
Duodenum	0.80 ± 1.05	0.53 ± 0.96	1.17 ± 2.38	0.77 ± 2.01	-0.37 ± 2.12	-0.24 ± 1.74
Stomach	0.79 ± 0.57	0.53 ± 0.50	0.97 ± 1.19	0.63 ± 0.88	-0.01 ± 0.95	-0.09 ± 0.64
Spleen	0.59 ± 0.45	0.39 ± 0.36	0.83 ± 0.88	0.53 ± 0.70	-0.24 ± 0.57	-0.14 ± 0.45
<i>A</i> , <i>I</i> , and <i>V</i>	2.12 ± 2.65	1.44 ± 1.89	2.64 ± 3.53	1.85 ± 2.90	-0.51 ± 1.94	-0.41 ± 1.81
Liver	0.51 ± 0.48	0.31 ± 0.35	0.64 ± 0.92	0.36 ± 0.63	-0.13 ± 0.64	-0.05 ± 0.51
Kidney	0.81 ± 0.59	0.52 ± 0.54	1.23 ± 2.85	0.73 ± 1.38	-0.45 ± 2.75	-0.21 ± 1.23
Gallbladder	1.19 ± 0.54	0.82 ± 0.66	1.43 ± 1.25	0.99 ± 1.38	-0.24 ± 0.86	-0.17 ± 0.93
Pancreas	1.87 ± 1.02	1.28 ± 0.85	2.42 ± 3.20	1.60 ± 2.14	-0.54 ± 3.09	-0.31 ± 2.00
Colon	—	15.67 ± 16.00	—	23.93 ± 20.72	—	-8.26 ± 19.86
Difficult cases	1.68 ± 2.28	1.31 ± 2.23	3.43 ± 5.89	2.12 ± 2.82	-1.75 ± 5.67	-0.82 ± 2.58
All	1.09 ± 2.30	0.79 ± 2.25	1.50 ± 3.63	1.05 ± 3.22	-0.41 ± 2.76	-0.26 ± 2.29

^aDeep interpolator.

^bSBI.

^cDifferences.

Table 4 The 95-percentile DTA values and the differences (mm).

Name	DI (3D)	DI (2D)	SBI (3D)	SBI (2D)	Diff (3D)	Diff (2D)
Lung	3.41 ± 10.13	2.87 ± 10.08	7.42 ± 18.05	5.59 ± 16.08	-4.01 ± 14.74	-2.72 ± 13.05
<i>E</i> and <i>S</i>	3.12 ± 2.63	3.02 ± 2.69	3.35 ± 2.51	3.13 ± 2.49	-0.24 ± 1.33	-0.11 ± 1.61
Heart	5.72 ± 7.21	4.20 ± 4.24	6.22 ± 7.75	5.09 ± 6.46	-0.50 ± 5.03	-0.89 ± 5.07
Duodenum	3.13 ± 8.59	2.44 ± 6.76	5.16 ± 13.52	4.27 ± 12.27	-2.04 ± 11.66	-1.83 ± 11.30
Stomach	2.32 ± 1.80	2.11 ± 1.98	3.02 ± 6.14	2.70 ± 5.71	-0.70 ± 5.93	-0.60 ± 5.47
Spleen	1.74 ± 1.16	1.54 ± 1.15	2.24 ± 2.79	1.97 ± 2.71	-0.50 ± 2.27	-0.43 ± 2.25
<i>A</i> , <i>I</i> , and <i>V</i>	5.92 ± 10.40	5.84 ± 10.57	7.50 ± 13.54	6.47 ± 11.47	-0.62 ± 9.24	-1.21 ± 8.19
Liver	1.86 ± 2.78	1.50 ± 2.32	2.32 ± 5.27	1.80 ± 4.65	-0.46 ± 4.20	-0.30 ± 4.22
Kidney	2.61 ± 4.49	2.19 ± 4.70	4.18 ± 9.66	3.42 ± 9.18	-1.57 ± 8.46	-1.22 ± 7.97
Gallbladder	3.03 ± 1.45	2.90 ± 2.08	4.07 ± 6.05	3.77 ± 6.09	-1.04 ± 5.69	-0.87 ± 5.70
Pancreas	5.10 ± 5.01n	4.62 ± 2.89	7.42 ± 12.23	6.14 ± 9.15	-2.31 ± 11.48	-1.52 ± 8.86
Colon	—	32.21 ± 24.17	—	41.25 ± 25.83	—	-9.05 ± 25.99
Difficult cases	6.49 ± 13.94	6.10 ± 14.09	12.92 ± 23.87	11.71 ± 22.37	-6.43 ± 20.28	-5.61 ± 18.82
All	3.24 ± 6.45	2.90 ± 6.15	4.53 ± 10.32	3.85 ± 9.35	-1.29 ± 8.16	-0.95 ± 7.51

Table 5 The results of the t-test on the DTA values.

Name	Max (3D)	Mean (3D)	p95% (3D)	Max (2D)	Mean (2D)	p95% (2D)
Lung	2.62e ⁻⁰⁹	5.37e ⁻⁰⁸	3.18e ⁻⁰⁷	2.40e ⁻⁰⁸	8.67e ⁻⁰⁷	8.25e ⁻⁰⁵
<i>E</i> and <i>S</i>	9.13e ⁻⁰⁸	1.05e ⁻²⁴	1.58e ⁻⁰⁶	0.004	2.38e ⁻¹¹	0.06
Heart	0.75	1e ⁻³	0.34	0.69	0.07	0.09
Duodenum	0.08	0.01	0.01	0.04	0.05	0.02
Stomach	0.16	0.003	0.06	0.19	0.02	0.08
Spleen	3.14e ⁻⁰⁴	5.69e ⁻¹⁰	8.52e ⁻⁰⁴	2.12e ⁻⁰⁴	3.97e ⁻⁰⁶	0.004
<i>A</i> , <i>I</i> , and <i>V</i>	0.08	0.01	0.07	0.58	0.03	0.51
Liver	0.09	3.51e ⁻⁰⁶	0.01	0.43	0.01	0.08
Kidney	3.52e ⁻⁰⁴	0.006	0.001	5.27e ⁻⁰⁴	0.003	0.007
Gallbladder	0.09	0.01	0.07	0.12	0.07	0.14
Pancreas	0.001	0.02	0.01	0.002	0.03	0.02
Colon	—	—	—	0.08	0.03	0.07
Difficult cases	1.38e ⁻⁰⁸	3.54e ⁻⁰⁷	1.78e ⁻⁰⁷	1.32e ⁻⁰⁷	2.02e ⁻⁰⁷	8.59e ⁻⁰⁷
All	1.15e ⁻¹⁹	8.94e ⁻¹⁷	1.17e ⁻¹⁸	1.71e ⁻¹⁵	2.70e ⁻¹⁰	1.27e ⁻¹²

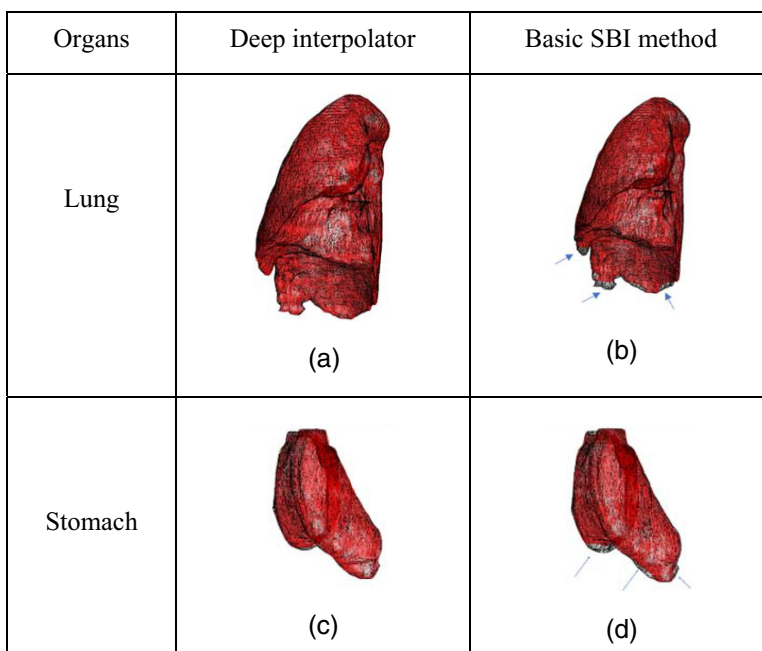


Fig. 7 (a)–(d) Comparison of the interpolation results by the deep interpolator and by SBI. The red surfaces are the interpolation results. The black meshes are the ground truth.

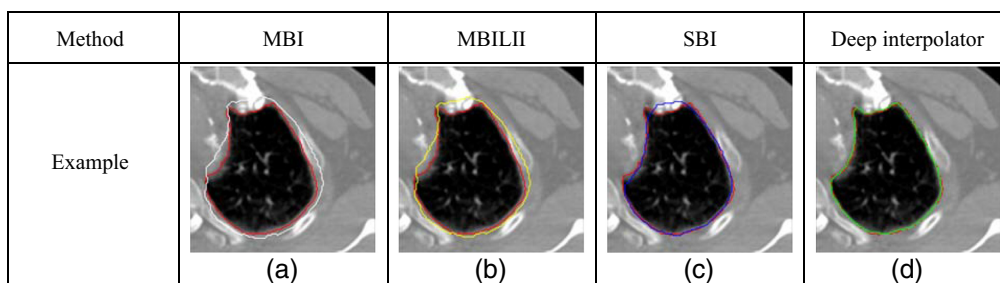


Fig. 8 Comparison of the interpolation results on a lung case by (a) MBI, (b) MBILII, (c) SBI, and (d) deep interpolator. The ground truth manual contours are in red. The MBI results are in white. The MBILII results are in yellow. The SBI results are in blue. The deep interpolator results are in green.

We compared the contour results with the current shape-based and complex methods, including SBI, MBI, and MBILII. An example from a lung case is shown in Fig. 8. The dice values and 3D DTA values on difficult cases are listed in Table 6. MBI and MBILII performed well, but the interpolated contours were not as close to the ground truth as the results by our deep contour interpolator. MBI is a shape-based method, and MBILII is based on MBI combined with the image intensity. The results for MBILII are slightly better than MBI.

Table 6 The dice scores and DTA (mm) values of MBI, MBILII, SBI, and deep method.

Method	Dice scores	Mean (DTA)	p95% (DTA)
MBI	0.83	2.22 ± 1.68	6.92 ± 8.92
MBILII	0.83	2.15 ± 1.55	6.90 ± 8.64
SBI	0.86	3.43 ± 5.89	12.92 ± 23.87
Deep interpolator	0.91	1.68 ± 2.28	6.49 ± 13.94

Table 7 Ablation studies on the deep model of varying depth, number of output channels for the first encoder stage, and training epochs.

Changes	Dice (difficult)	Mean (DTA)
Depth = 1	0.89 ± 0.09	1.72 ± 1.42
Depth = 2	0.91 ± 0.08	1.83 ± 2.24
Epoch = 10	0.91 ± 0.08	1.71 ± 2.13
NumFirstEncoder = 4	0.90 ± 0.09	1.70 ± 2.10
Default	0.91 ± 0.09	1.68 ± 2.28

3.1 Ablation Studies

We adopted a light-weight network design for the deep interpolator. Because there are no other additional blocks or layers, some simple ablation studies were conducted to investigate the impact of the most important design parameters. We compared the dice scores and mean of DTA values on difficult cases with different network depths, training epochs, and numbers of output channels for the first encoder stage. Table 7 summarizes the results. For the network depth settings, we tested one- and two-layer designs against the default three layers. The comparison results show that the model prediction accuracy was not sensitive to the network depth. We confirmed that more training epochs contribute to a better results. We also tested a simpler 4-channel output for the first encoder stage and demonstrated that the more output channels for the first encoder stage contribute to a better result. Due to the limitation of our GPU hardware, we did not try deeper or wider networks.

3.2 Generality and Slice Thickness Limitation Tests

To prove the generality of the trained deep interpolator model, we tested the model on three new organs (brainstem, parotid gland, and submandibular) from PDDCA²³ and HNSCC-3DCT-RT.²⁴ These three new organs are in the head-neck region, and the deep interpolator did not see any CT images or organs from the head-neck region in model training.

The computed dices scores are listed in Table 8. They show that the results from our deep interpolator are as good as or better than SBI. Figure 9 shows a comparison between the results of SBI and our deep interpolator. The smaller distance between deep interpolator results and ground truth indicates a better performance.

We tested the performance of our deep interpolator to interpolate between two 7-slice patches apart for different organs. The two examples in Fig. 10 show the comparison between 5-slice cases and 7-slice cases. The first row contains the comparisons on one relatively regularly shape from lung. The performances were in the same level. The second row contains some suboptimal performances on an irregular shape from stomach. The distance between the prediction and ground truth is smaller in the 5-slice case than the 7-slice cases in the area pointed to by blue circle. It can be observed that the performance can still be robust when intermediate slices are fed into the model with the increased slice thickness. However, when the differences between the

Table 8 Comparison of dice scores on new organs.

Organs	Deep interpolator	SBI	Differences	t-test p -value
Brainstem	0.92 ± 0.03	0.92 ± 0.03	0.00 ± 0.00	0.09
Parotid gland	0.90 ± 0.06	0.88 ± 0.09	0.02 ± 0.05	$7.10e^{-14}$
Submandibular	0.86 ± 0.06	0.81 ± 0.09	0.05 ± 0.07	$3.49e^{-18}$

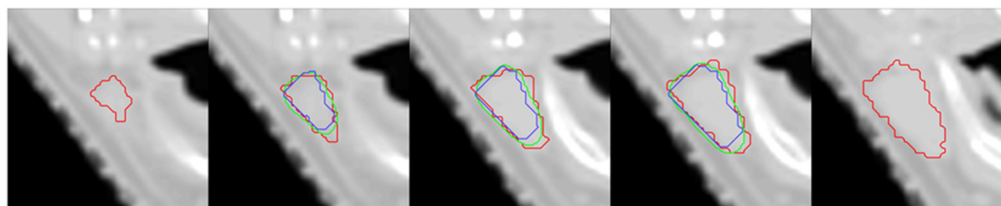


Fig. 9 Comparison of performance on newly added organs between different methods (submandibular). The ground truth manual contours are in red. The SBI interpolation results are in blue. The deep interpolator results are in green.

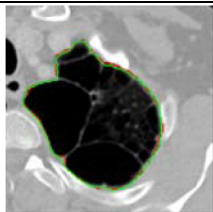
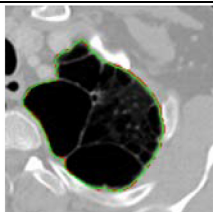
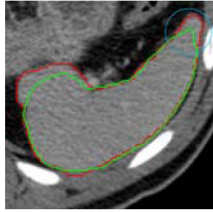
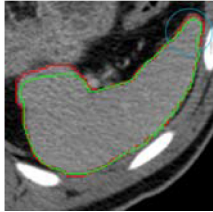
Thickness	7-slice	5-slice
Lung	 <p>(a)</p>	 <p>(b)</p>
Stomach	 <p>(c)</p>	 <p>(d)</p>

Fig. 10 (a)–(d) Comparison of results from different slice thickness. The ground truth contours are in red. The deep interpolator results are in green.

reference slices are dramatic, an increasing distance between the reference slices could lead to suboptimal interpolation results. Model robustness over larger slice distances might warrant a further investigation.

4 Discussions

The results confirmed that the proposed deep interpolator was significantly more accurate and robust than the conventional shape-based interpolator method, especially for the difficult cases in which the shape changes between the adjacent slices were more dramatic. On the difficult transverse colon cases, our deep interpolator trained with only a few training cases performed much better than the conventional SBI method. The deep contour interpolator also outperformed the MBI and MBILII methods on dice score and DTA measurements. We decided not to compare our results with the CSGI method because the algorithm requires too many user-configurable parameters per case to work properly.

There are many directions to further improve our deep interpolator model. More training data will be very useful, especially for difficult cases. However, datasets for the GI organs, e.g., colon and small intestines, are very difficult to obtain because they are not commonly contoured in the clinic and the contour datasets are rarely publicly shared due to low contour qualities. The data augmentation methods, for example, rotation and noise addition, could be applied to further improve the model robustness. A shallow and narrow 3D U-Net architecture was used in the current study to ensure the computation performance and the general robustness. More advanced network architectures, e.g., generative adversarial network²⁵ and the attention network,²⁶ can still be explored in further work. It is also possible to use the distance map as the second input to our

deep interpolator to provide richer information than the current structure mask. The network can also be trained to output the distance map instead of the organ mask to allow for more post-processing options. The prediction of our current small U-Net model is very fast. The mean processing time for each $128 \times 128 \times 5$ patch is <0.05 s on a GTX1650 GPU, compared with SBI (<0.01), MBI (0.2), and MBILII (0.6).

This deep interpolator model could be further extended in future work to include more image modalities and more organs. Only CT images and 15 organs were covered so far. Other organs can be supported once the organ contour datasets are available. In the authors' opinion, MRI data can also be added and be supported using a single universal model, but the image intensity normalization step might require additional adjustments.

5 Conclusion

A deep learning model was developed in this study to perform 2D contour interpolation on CT images. The deep model was trained and evaluated on 15 organs and significantly outperformed the basic shape-based contour interpolation method. It could be useful to expedite the tasks of manual segmentation of organs and structures in medical images.

Disclosures

No conflicts of interest.

Acknowledgments

This study was supported by the National Institute of Biomedical Imaging and Bioengineering (NIBIB) (Grant No. R03-EB028427).

References

1. J. Lamb et al., "Online adaptive radiation therapy: implementation of a new process of care," *Cureus* **9**(8), e1618 (2017).
2. M. H. Hesamian et al., "Deep learning techniques for medical image segmentation: achievements and challenges," *J. Digital Imaging* **32**(4), 582–596 (2019).
3. Y. Lei et al., "Deep learning in multi-organ segmentation," arXiv:2001.10619 (2020).
4. F. Vaassen et al., "Evaluation of measures for assessing time-saving of automatic organ-at-risk segmentation in radiotherapy," *Phys. Imaging Radiat. Oncol.* **13**, 1–6 (2020).
5. X. Liao, D. Reutens, and Z. Yang, "Morphology-based interslice interpolation using local intensity information for segmentation," in *4th Int. Conf. Biomed. Eng. and Inf. (BMEI)*, IEEE (2011).
6. K. Jha, "Construction of branching surface from 2-D contours," *Int. J. CAD/CAM* **8**(1), 21–28 (2009).
7. A. Baghaie and Z. Yu, "An optimization method for slice interpolation of medical images," arXiv:1402.0936 (2014).
8. G. J. Grevera and J. K. Udupa, "Shape-based interpolation of multidimensional grey-level images," *IEEE Trans. Med. Imaging* **15**(6), 881–892 (1996).
9. R. A. Lutufo, G. T. Herman, and J. K. Udupa, "Combining shape-based and gray-level interpolations," *Proc. SPIE* **1808**, 1–10 (1992).
10. A. B. Albu, T. Beugeling, and D. Laurendeau, "A morphology-based approach for interslice interpolation of anatomical slices from volumetric images," *IEEE Trans. Biomed. Eng.* **55**(8), 2022–2038 (2008).
11. D. Zukic et al., "ND morphological contour interpolation," *Insight J.* 1–8 (2016).
12. G. T. Herman, C. A. Bucholtz, and J. Zheng, "Shape-based interpolation using modified cubic splines," in *Proc. Annu. Int. Conf. IEEE Eng. in Med. and Biol. Soc., Volume 13: 1991*, IEEE (2005).

13. S. Ravikumar et al., "Facilitating manual segmentation of 3D datasets using contour and intensity guided interpolation," in *IEEE 16th Int. Symp. Biomed. Imaging (ISBI 2019)*, IEEE (2019).
14. R. R. Wildeboer et al., "Three-dimensional histopathological reconstruction as a reliable ground truth for prostate cancer studies," *Biomed. Phys. Eng. Express* **3**(3), 035014 (2017).
15. Q. Wu et al., "A multi-stage DCNN method for liver tumor segmentation," in *IEEE 3rd Int. Conf. Safe Prod. and Inf. (IICSPI)*, IEEE (2020).
16. H. J. W. L. Aerts, "Data from NSCLC-radiomics," *Cancer Imaging Arch.* (2019).
17. B. Landman, "MICCAI multi-atlas labeling beyond the cranial vault – workshop and challenge," (2015).
18. P. Bilic et al., "The liver tumor segmentation benchmark (LiTS)," *Med. Image Anal.* **84**, 102680 (2023).
19. H. R. Roth et al., "Data from pancreas-CT," *Cancer Imaging Archive* (2016).
20. A. L. Simpson et al., "A large annotated medical image dataset for the development and evaluation of segmentation algorithms," arXiv:1902.09063 (2019).
21. O. Ronneberger, P. Fischer, and T. Brox, "U-net: convolutional networks for biomedical image segmentation," *Lect. Notes Comput. Sci.* **9351**, 234–241 (2015).
22. D. Thomson, C. Boylan, and T. Liptrot, "Evaluation of an automatic segmentation algorithm for definition of head and neck organs at risk," *Radiat. Oncol.* **9**, 173 (2014).
23. P. F. Raudaschl et al., "Evaluation of segmentation methods on head and neck CT: auto-segmentation challenge 2015," *Med. Phys.* **44**(5), 2020–2036 (2017).
24. T. Bejarano, M. De Ornelas-Couto, and I. B. Mihaylov, "Head-and-neck squamous cell carcinoma patients with CT taken during pre-treatment, mid-treatment, and post-treatment Dataset," *Cancer Imaging Arch.* (2018).
25. I. Goodfellow et al., "Generative adversarial networks," *Commun. ACM* **63**(11), 139–144 (2020).
26. A. Vaswani et al., "Attention is all you need," *Adv. Neural Inf. Process. Syst.* **30** (2017).

Chenxi Zhao received his BS degree in computer science and technology from Jilin University in 2018. Currently, he is pursuing a PhD in computer science at the EECS Department of the University of Missouri under Dr. Ye Duan's supervision. His research interests include medical imaging, computer vision, deep learning, and computer graphics.

Ye Duan received his BA degree in mathematics from Peking University in 1991, his MS degree in mathematics from Utah State University in 1996, and his MS degree and PhD in computer science from the State University of New York at Stony Brook, in 1998 and 2003, respectively. From September 2003 to August 2009, he was an assistant professor of computer science at the University of Missouri, Columbia, Missouri, United States, where he is an associate professor of computer science. His research interests include computer graphics, computer vision, machine learning, and biomedical imaging.

Deshan Yang received his PhD in biomedical engineering from the University of Wisconsin–Madison. Currently, he is a professor at the Department of Radiation Oncology, Duke University. His research interests include medical imaging (registration, segmentation, reconstruction, motion management, and image guidance) for radiation oncology applications, machine learning and deep learning, adaptive radiotherapy, image guidance, treatment planning automation, quality assurance, cardiac radiosurgery, health information technologies, and clinical applications for radiation oncology and medical physics.