Deep learning-assisted multiple organ segmentation from whole-body CT images

Yazdan Salimi1*, Isaac Shiri1*, Zahra Mansouri1 and Habib Zaidi1,2,3,4†

1 Division of Nuclear Medicine and Molecular Imaging, Geneva University Hospital, CH-1211, Geneva, Switzerland
2 Geneva University Neurocenter, Geneva University, Geneva, Switzerland
3 Department of Nuclear Medicine and Molecular Imaging, University of Groningen, University Medical Center Groningen, Groningen, Netherlands
4 Department of Nuclear Medicine, University of Southern Denmark, Odense, Denmark

*Yazdan Salmi and Isaac Shiri contributed equally to this work

†Corresponding Author:
Habib Zaidi, Ph.D
Geneva University Hospital
Division of Nuclear Medicine and Molecular Imaging
CH-1211 Geneva, Switzerland
Tel: +41 22 372 7258
Fax: +41 22 372 7169
email: habib.zaidi@hcuge.ch

Short running title: Deep learning-assisted multiple organ segmentation

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.
Abstract

**Background:** Automated organ segmentation from computed tomography (CT) images facilitates a number of clinical applications, including clinical diagnosis, monitoring of treatment response, quantification, radiation therapy treatment planning, and radiation dosimetry.

**Purpose:** To develop a novel deep learning framework to generate multi-organ masks from CT images for 23 different body organs.

**Methods:** A dataset consisting of 3106 CT images (649,398 axial 2D CT slices, 13,640 images/segment pairs) and ground-truth manual segmentation from various online available databases were collected. After cropping them to body contour, they were resized, normalized and used to train separate models for 23 organs. Data were split to train (80%) and test (20%) covering all the databases. A Res-UNET model was trained to generate segmentation masks from the input normalized CT images. The model output was converted back to the original dimensions and compared with ground-truth segmentation masks in terms of Dice and Jaccard coefficients. The information about organ positions was implemented during post-processing by providing six anchor organ segmentations as input. Our model was compared with the online available “TotalSegmentator” model through testing our model on their test datasets and their model on our test datasets.

**Results:** The average Dice coefficient before and after post-processing was 84.28% and 83.26% respectively. The average Jaccard index was 76.17 and 70.60 before and after post-processing respectively. Dice coefficients over 90% were achieved for the liver, heart, bones, kidneys, spleen, femur heads, lungs, aorta, eyes, and brain segmentation masks. Post-processing improved the performance in only nine organs. Our model on the TotalSegmentator dataset was better than their models on our dataset in five organs out of 15 common organs and achieved almost similar performance for two organs.

**Conclusions:** The availability of a fast and reliable multi-organ segmentation tool leverages implementation in clinical setting. In this study, we developed deep learning models to segment multiple body organs and compared the performance of our models with different algorithms. Our model was trained on images presenting with large variability emanating from different databases producing acceptable results even in cases with unusual anatomies and pathologies, such as splenomegaly. We recommend using these algorithms for organs providing good performance. One of the main merits of our proposed models is their lightweight nature with an average inference time of 1.67 seconds per case per organ for a total-body CT image, which facilitates their implementation on standard computers.

**Keywords:** Segmentation, Organs at Risk, Computed Tomography, Deep Learning, Computational Models.
Introduction

Segmentation of healthy organs from Computed Tomography (CT) images is critical and beneficial in a number of applications, including the generation of anthropomorphic computational models, delimitation of organs at risk in radiation therapy (RT) treatment planning (1-4), and other kinds of computer-assisted applications, such as pathologic detection (5, 6), prognosis and outcome prediction (7, 8), image quantification (9, 10), and radiation dosimetry calculations (11-13). The manual slice-by-slice segmentation of organs can be labor-intensive and time-consuming, in addition to the high inter- and intra-observer variability reported for segmentation of healthy organs and malignant lesions (14, 15). Since the emergence of machine learning and deep learning (DL) algorithms in medical imaging research, especially medical image segmentation, a number of studies focused on automatic segmentation of structures from CT images and other imaging modalities (16-18). Most published studies attempted to improve segmentation accuracy (commonly quantified by the Dice coefficient), robustness and generalizability on new unseen dataset acquired with different imaging settings on disparate patient characteristics and including a large number of organs (19-21). Newly developed neural network architectures, loss functions, and image processing algorithms contributed to the improvement of the performance of image segmentation models. Yet, the number of datasets and their diversity remains the bottleneck for successful implementation of deep learning-based algorithms (22). Most studies conveyed the performance of the developed models on a test set excluded from the training set, thus reaching very high Dice coefficients as reported in few challenges held on multiple organ segmentations (23). Yet, the majority of these studies didn’t investigate models’ performance on unseen external datasets. Xu et al. (24) focused on the occurrence of outliers during image segmentation and how to solve this problem. Recent studies addressed the limitations and benefits of DL-based organ segmentation in real-life clinical scenarios (14, 25). The comparison of the results achieved by different techniques using private/local databases is not straightforward given that the used datasets are not publicly available. Besides, it’s well established that acquisition, scanner, and demographic parameters can affect the performance of a model on external unseen datasets from other centers (14, 26). Ma et al. (19) described the low performance of segmentation models trained and inferenced on different databases for abdominal organs segmentation task. In this context, a segmentation model trained on a dataset presenting with a large variability and tested on an unseen dataset may be beneficial in estimating the performance in real clinical scenarios.

In this study, we aimed to develop a deep neural network to segment multiple healthy organs (28 organs) from total-body CT images targeting improvement of the accuracy and generalizability compared to previously developed models. We also compared the performance of our models with existing methods and considered the effect of post-processing algorithms to take into account organ-specific anatomical information during the segmentation process.

Materials and Methods

Patient population

This study included 3106 CT images (649,398 axial 2D CT slices, 13,640 3D image/segment pairs) collected from multiple online available datasets (27-32). A total of 300 pediatric cases with 18.9 ± 4.13 cm effective diameter and 2806 adult cases with 27.53 ± 5.35 cm effective diameter as defined by the AAPM #204 Report (33) were included. The average age was 6.32 ± 4.34 years for pediatric patients and 66.98 ± 9.84 years for adult patients. It should be noted that the age, gender, and acquisition parameters were available only in a limited number of datasets, the rest were either anonymized or in NIFTI format without additional information. The number of slices and patient size characteristics were summarized in supplementary Table 1. The data were split into training and test...
set for each organ according to the number of cases from each database to ensure the test and train
data use cases from each available database, i.e., the training (80%) and testing (20%) data for each
organ include cases from all databases.

Figure 1 depicts the number of CT images used for training each model for each organ segmentation.
The number of training and test datasets are summarized in Table 1. The detailed number of training
cases from each database is provided in supplementary Table 2. The masks (segmentations) of the 23
different organs were used to train separate segmentation models. The summed gastrointestinal
segment (GIs) was defined by adding distinct segmentations of the duodenum, small intestine, and
colon together to define a single organ.
Figure 1. Number of 3D CT images extracted from the clinical studies included in all databases used for training and testing of the models. The upper (color) bar depicts the test, whereas the bottom (gray) bar depicts the training numbers. UB: Urinary Bladder, SC: Spinal Cord, GB: Gall Bladder, AG: Adrenal Gland, IVC: Inferior Vena Cava, SI: Small Intestine, GIs: Gastrointestinal.

Preprocessing and Network architecture

The external body contour was extracted from axial CT images through image processing algorithms developed and used in previous studies (34, 35). The CT images were cropped to a bounding box (BB) including the body contour in the lateral and AP directions to remove the background area. The images were cropped to 30 slices in the superior direction and 30 slices in the inferior direction according to the BB covering organ segmentation in the Z-axis (cranio-caudal direction). The model was trained in 2D fashion, meaning that the input to the network consisted of 2D axial images with the output being the corresponding segmentation masks. A Res-UNET neural network architecture used in a previous PET segmentation study (36) was employed in this work (Figure 2). The 2D images and masks were resized to 304 (right to left) × 224 (anterior posterior) pixel dimensions. The image intensities were clipped between -70 HU and +170 HU and normalized between zero and one and then discretized to 240 intensity values.

The cropped information was stored in the image header and used later to reverse the cropped model output segmentation to the original image dimensions. Figure 3 summarizes the steps performed to train the network.
Figure 2. The Res-UNET neural network architecture adopted for multi-organ segmentation from CT images. Conv2d: 2D convolution layer.

**Post-processing and prior knowledge implementation**

Organ-specific post-processing algorithms were used to take into account the anatomical locations of the organs. For all test datasets, segmentation masks of six anchor organs including the liver, spleen, lung, femoral head, bladder, and kidneys were generated by our trained models. These generated segmentation masks were used to perform organ-specific post-processing. For each organ, a specific algorithm was used to remove the segmented voxels outside of the BB delimitating each organ. For instance, for spleen post-processing, the post-processing function input was the liver, lung, and femoral head segmentation. The BB of the femur and lung were determined considering the known prior anatomical information that the liver is in the abdomen and always higher than the femoral head BB and lower than the lungs apex. We used the same strategy for other organs, such as the gall bladder (GB), and adrenal glands (AG) to remove unwanted (false positive) voxels from the network output by exploiting the fact that the gall bladder is at the inferior part of the liver and the adrenal gland are upper than kidneys. In the end, the network performance evaluation was compared with/without organ-specific post-processing.

**Evaluation metrics**

For each organ segmentation task, 20% of each database was randomly defined and used as the test dataset. The predicted segmentation masks were compared to the ground truth masks by measuring Dice and Jaccard coefficients.
Figure 3. The flowchart of training and inference steps followed in implementing the segmentation algorithm. From left to right: The upper axial slices show CT images, with the liver mask defined with the yellow contour, segmented body contour, and pre-processed (cropped, normalized, and resized) images. The lower segmentation shows the 3D visualization of the network output after post-processing.

Real-life evaluation on external datasets

To evaluate the performance of our model in real clinical scenarios on an external unseen dataset and compare our method to previously reported deep learning models, we tested our trained models on the online available TotalSegmentator dataset published recently by Wasserthal et al. (37) and then tested...
their trained model on the databases we used for testing. The dataset used in the above reference was
local for the involved centers and could be considered completely unseen data for our models. In
addition, they separated the test and train dataset to make the comparison more reproducible. They
have used state-of-the-art nnU-Net (38) network/training methodology and managed to won 9 out of
10 MICCAI 2020 (39) and AMOS (23) challenges. We compared the performance of our model and
models reported in the reference above for organs included in both studies (15 organs listed in
Table 2). Figure 4 shows the dataflow in this study. Overall, we performed three evaluations: a) our
model tested on our test dataset (23 organs), b) our model tested on the TotalSegmentator test dataset
(15 organs), and c) the TotalSegmentator model tested on our dataset (15 organs). TotalSegmentator
trained model was collected from GitHub on April 28, 2023.

Figure 4. The Dataflow adopted for external evaluation. Green, light blue, and dark blue lines show the
dataflow for strategies (a), (b), and (c), respectively.

Statistical analysis
We used the Wilcoxon rank t-test to evaluate the effect of post-processing on each organ.

Results
Evaluation on the test dataset
Table 1 summarizes the Dice and Jaccard image segmentation metrics for our model on test sets
separated from our data (strategy (a) mentioned in the Methods section). The highest Dice coefficients
were achieved for the lung, spleen, liver, and brain organs, while the lowest values were obtained for
the thymus, adrenal gland organs. Organ-specific post-processing increased the Dice coefficient in
seven organs by more than 0.15 absolute value and this increase was statistically significant.
However, it did not increase or even significantly decrease the Dice coefficient for the remaining
organs. These seven organs were the heart, spleen, UB, SC, aorta, GB, and thymus. Figure 5 depicts
examples of 3D visualization of segmentations of CT images corresponding to different subjects
shown from eight different perspectives. Supplementary figure 1 extends the examples shown in
Figure 5 for pediatric cases and cases with unusual anatomical variations such as splenomegaly. Figure
6 depicts the Dice coefficients in different organs with/without post-processing. The detailed results
of segmentation accuracy for each database included in the assessment are presented in supplementary
Table 2.
External comparison

Table 2 summarizes the Dice and Jaccard evaluation metrics for strategies (b) and (c). A relative difference larger than 2% for the Dice coefficient was considered as significant when comparing strategies (b) and (c). Our model outperformed TotalSegmentator for five organs, including the liver, pancreas, spleen, UB, and femur heads, while the outcome was the same for the lungs and kidneys. For the remaining organs listed in Table 2 (7 organs), the TotalSegmentator models outperformed our models. The brain mask was not valid due to the blurring generated on the face area for privacy preserving concerns on TotalSegmentator dataset. Organ-specific post-processing improved the segmentation accuracy, reflected by higher Dice coefficients for strategy (b) in seven other organs, including the liver, kidneys, spleen, UB, esophagus, femur heads, and GB. Table 3 compares our model’s performance in strategy (a) to recent studies reported in the literature for common organs.

Figure 5. 3D visualization of segmentation masks of different organs showing: the kidneys (dark green), femoral heads (lime), bones (yellow), liver (dark red), aorta (light red), spleen (cyan), heart (purple), stomach (light green), spinal cord (dark blue), urinary bladder (light blue), and the rectum (pink).
Figure 6. Box plots of Dice coefficients achieved for different organ segmentations before and after post-processing. The red, blue, and green reference lines depict 60%, 80%, and 90% Dice coefficients, respectively. UB: Urinary Bladder, SC: Spinal Cord, GB: Gall Bladder, AG: Adrenal Gland, IVC: Inferior Vena Cava, SI: Small Intestine, GIs: Gastrointestinal.

<table>
<thead>
<tr>
<th>Organ</th>
<th>Train#</th>
<th>Test#</th>
<th>Dice_network</th>
<th>Dice_post</th>
<th>Dice_gain_post</th>
<th>Dice_P-value</th>
<th>Jaccard_network</th>
<th>Jaccard_post</th>
<th>Jacc_gain_post</th>
<th>Jacc_P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Liver</td>
<td>908</td>
<td>227</td>
<td>96.93 ± 1.67</td>
<td>96.98 ± 1.62</td>
<td>0.06 ± 0.14</td>
<td>&lt;0.001</td>
<td>94.08 ± 3.01</td>
<td>94.19 ± 2.93</td>
<td>0.11 ± 0.25</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Heart</td>
<td>368</td>
<td>92</td>
<td>91.60 ± 6.52</td>
<td>91.95 ± 6.13</td>
<td>0.35 ± 0.87</td>
<td>&lt;0.001</td>
<td>85.08 ± 9.82</td>
<td>85.62 ± 9.38</td>
<td>0.54 ± 1.32</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Pancreas</td>
<td>984</td>
<td>246</td>
<td>77.52 ± 15.99</td>
<td>76.24 ± 17.02</td>
<td>-1.28 ± 5.07</td>
<td>0.563</td>
<td>65.76 ± 19.15</td>
<td>64.25 ± 19.78</td>
<td>-1.50 ± 5.89</td>
<td>0.662</td>
</tr>
<tr>
<td>Bones</td>
<td>300</td>
<td>75</td>
<td>94.08 ± 4.84</td>
<td>93.84 ± 4.75</td>
<td>-0.24 ± 0.32</td>
<td>&lt;0.001</td>
<td>89.17 ± 7.77</td>
<td>88.72 ± 7.60</td>
<td>-0.45 ± 0.56</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Kidneys</td>
<td>548</td>
<td>137</td>
<td>94.03 ± 4.80</td>
<td>94.13 ± 5.05</td>
<td>0.10 ± 0.79</td>
<td>&lt;0.001</td>
<td>89.07 ± 7.57</td>
<td>89.28 ± 7.89</td>
<td>0.21 ± 1.24</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Spleen</td>
<td>732</td>
<td>183</td>
<td>94.29 ± 9.88</td>
<td>94.68 ± 9.28</td>
<td>0.39 ± 3.26</td>
<td>&lt;0.001</td>
<td>90.28 ± 11.81</td>
<td>90.85 ± 11.16</td>
<td>0.57 ± 4.33</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>UB</td>
<td>464</td>
<td>116</td>
<td>83.41 ± 17.99</td>
<td>83.78 ± 17.97</td>
<td>0.37 ± 1.34</td>
<td>&lt;0.001</td>
<td>74.75 ± 21.24</td>
<td>75.24 ± 21.40</td>
<td>0.50 ± 1.84</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Esophagus</td>
<td>720</td>
<td>180</td>
<td>73.20 ± 10.90</td>
<td>72.35 ± 12.55</td>
<td>-0.84 ± 4.07</td>
<td>0.010</td>
<td>58.76 ± 12.27</td>
<td>58.01 ± 13.73</td>
<td>-0.75 ± 3.98</td>
<td>0.010</td>
</tr>
<tr>
<td>Femur Heads</td>
<td>220</td>
<td>55</td>
<td>95.65 ± 2.78</td>
<td>95.72 ± 2.79</td>
<td>0.07 ± 0.14</td>
<td>&lt;0.001</td>
<td>91.78 ± 4.77</td>
<td>91.92 ± 4.79</td>
<td>0.14 ± 0.24</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Lungs</td>
<td>1240</td>
<td>310</td>
<td>97.63 ± 1.18</td>
<td>96.56 ± 4.70</td>
<td>-1.07 ± 4.61</td>
<td>&lt;0.001</td>
<td>95.39 ± 2.18</td>
<td>93.42 ± 9.06</td>
<td>-1.97 ± 8.46</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>SC</td>
<td>588</td>
<td>147</td>
<td>89.69 ± 3.65</td>
<td>89.85 ± 3.70</td>
<td>0.17 ± 0.39</td>
<td>&lt;0.001</td>
<td>81.49 ± 5.63</td>
<td>81.76 ± 5.72</td>
<td>0.28 ± 0.64</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Aorta</td>
<td>184</td>
<td>46</td>
<td>92.99 ± 2.85</td>
<td>93.55 ± 2.22</td>
<td>0.56 ± 1.34</td>
<td>&lt;0.001</td>
<td>87.03 ± 4.77</td>
<td>87.95 ± 3.77</td>
<td>0.93 ± 2.18</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>GB</td>
<td>384</td>
<td>96</td>
<td>78.74 ± 16.92</td>
<td>79.32 ± 16.78</td>
<td>0.57 ± 5.35</td>
<td>0.095</td>
<td>67.60 ± 19.56</td>
<td>67.39 ± 21.00</td>
<td>-0.21 ± 9.09</td>
<td>0.298</td>
</tr>
<tr>
<td>AG</td>
<td>380</td>
<td>95</td>
<td>54.64 ± 17.40</td>
<td>51.62 ± 17.52</td>
<td>-3.02 ± 5.86</td>
<td>&lt;0.001</td>
<td>39.38 ± 15.21</td>
<td>30.14 ± 18.41</td>
<td>-9.24 ± 13.01</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>IVC</td>
<td>184</td>
<td>46</td>
<td>80.57 ± 11.30</td>
<td>62.59 ± 11.44</td>
<td>-17.98 ± 10.78</td>
<td>&lt;0.001</td>
<td>68.77 ± 14.33</td>
<td>40.97 ± 13.22</td>
<td>-27.80 ± 11.26</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Colon</td>
<td>224</td>
<td>56</td>
<td>70.61 ± 11.97</td>
<td>67.48 ± 15.25</td>
<td>-3.13 ± 7.05</td>
<td>0.167</td>
<td>55.76 ± 13.20</td>
<td>52.16 ± 16.96</td>
<td>-3.60 ± 8.15</td>
<td>0.133</td>
</tr>
<tr>
<td>Rectum</td>
<td>228</td>
<td>57</td>
<td>77.90 ± 13.31</td>
<td>77.70 ± 13.33</td>
<td>-0.20 ± 2.44</td>
<td>0.064</td>
<td>65.41 ± 15.13</td>
<td>65.14 ± 15.23</td>
<td>-0.26 ± 3.31</td>
<td>0.048</td>
</tr>
<tr>
<td>SI</td>
<td>228</td>
<td>57</td>
<td>70.79 ± 14.23</td>
<td>71.45 ± 14.52</td>
<td>0.66 ± 1.00</td>
<td>&lt;0.001</td>
<td>56.35 ± 14.63</td>
<td>56.67 ± 16.10</td>
<td>0.32 ± 4.34</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Stomach</td>
<td>408</td>
<td>102</td>
<td>86.22 ± 10.76</td>
<td>75.13 ± 12.17</td>
<td>-11.09 ± 11.46</td>
<td>&lt;0.001</td>
<td>77.12 ± 14.36</td>
<td>57.20 ± 18.80</td>
<td>-19.92 ± 18.09</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Thymus</td>
<td>108</td>
<td>27</td>
<td>64.11 ± 23.55</td>
<td>65.36 ± 23.57</td>
<td>1.25 ± 2.63</td>
<td>&lt;0.001</td>
<td>50.89 ± 22.67</td>
<td>52.26 ± 22.75</td>
<td>1.37 ± 3.01</td>
<td>&lt;0.002</td>
</tr>
<tr>
<td>GIs</td>
<td>220</td>
<td>55</td>
<td>88.88 ± 8.87</td>
<td>89.26 ± 8.22</td>
<td>0.38 ± 1.44</td>
<td>&lt;0.001</td>
<td>80.91 ± 11.79</td>
<td>81.40 ± 11.14</td>
<td>0.50 ± 1.66</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Eyes</td>
<td>404</td>
<td>101</td>
<td>91.13 ± 3.13</td>
<td>91.13 ± 3.15</td>
<td>0.00 ± 0.09</td>
<td>0.860</td>
<td>83.84 ± 5.09</td>
<td>25.75 ± 38.89</td>
<td>-58.09 ± 39.10</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Brain</td>
<td>276</td>
<td>69</td>
<td>97.39 ± 1.66</td>
<td>96.72 ± 5.26</td>
<td>-0.67 ± 4.48</td>
<td>0.909</td>
<td>94.95 ± 3.03</td>
<td>33.10 ± 44.58</td>
<td>-61.85 ± 44.10</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>
Table 2. Results of the external evaluation using Wasserthal et al. (37) algorithm in strategies (b) and (c). * The brain images were distorted and blurred in Wasserthal et al. dataset. UB: Urinary Bladder, SC: Spinal Cord, GB: Gall Bladder, AG: Adrenal Gland, IVC: Inferior Vena Cava, SI: Small Intestine, Gls: Gastrointestinal.

<table>
<thead>
<tr>
<th>Organ</th>
<th>Test #</th>
<th>Dice_summary_test</th>
<th>Jaccard_summary_test</th>
<th>Test #</th>
<th>Dice_network</th>
<th>Dice_post</th>
<th>Jaccard_network</th>
<th>Jaccard_post</th>
</tr>
</thead>
<tbody>
<tr>
<td>Liver</td>
<td>227</td>
<td>87.18 ± 27.58</td>
<td>83.87 ± 27.11</td>
<td>46</td>
<td>94.96 ± 3.57</td>
<td>95.02 ± 3.67</td>
<td>90.60 ± 6.11</td>
<td>87.40 ± 16.82</td>
</tr>
<tr>
<td>Pancreas</td>
<td>246</td>
<td>57.61 ± 35.31</td>
<td>48.04 ± 30.94</td>
<td>37</td>
<td>70.36 ± 19.36</td>
<td>68.60 ± 19.31</td>
<td>57.05 ± 19.35</td>
<td>29.73 ± 24.94</td>
</tr>
<tr>
<td>Kidneys</td>
<td>137</td>
<td>91.44 ± 6.34</td>
<td>84.76 ± 9.36</td>
<td>39</td>
<td>88.47 ± 7.60</td>
<td>90.41 ± 7.56</td>
<td>80.03 ± 10.75</td>
<td>79.44 ± 21.59</td>
</tr>
<tr>
<td>Spleen</td>
<td>183</td>
<td>90.38 ± 19.37</td>
<td>85.84 ± 19.51</td>
<td>45</td>
<td>93.13 ± 6.31</td>
<td>93.68 ± 4.97</td>
<td>87.73 ± 10.00</td>
<td>85.57 ± 15.39</td>
</tr>
<tr>
<td>UB</td>
<td>116</td>
<td>75.03 ± 26.66</td>
<td>65.63 ± 26.79</td>
<td>36</td>
<td>77.91 ± 18.06</td>
<td>80.17 ± 17.18</td>
<td>66.67 ± 20.29</td>
<td>65.25 ± 26.47</td>
</tr>
<tr>
<td>Esophagus</td>
<td>180</td>
<td>72.25 ± 13.46</td>
<td>58.05 ± 14.36</td>
<td>44</td>
<td>51.75 ± 14.69</td>
<td>53.42 ± 17.00</td>
<td>36.18 ± 13.22</td>
<td>24.86 ± 26.43</td>
</tr>
<tr>
<td>Femur Heads</td>
<td>55</td>
<td>81.15 ± 17.30</td>
<td>70.94 ± 18.91</td>
<td>30</td>
<td>79.24 ± 12.72</td>
<td>86.52 ± 10.44</td>
<td>67.30 ± 16.56</td>
<td>71.41 ± 25.32</td>
</tr>
<tr>
<td>Lungs</td>
<td>310</td>
<td>97.81 ± 1.10</td>
<td>95.74 ± 2.04</td>
<td>47</td>
<td>96.64 ± 2.33</td>
<td>96.33 ± 3.53</td>
<td>93.60 ± 4.23</td>
<td>85.06 ± 26.90</td>
</tr>
<tr>
<td>Aorta</td>
<td>46</td>
<td>90.98 ± 2.57</td>
<td>83.56 ± 4.26</td>
<td>46</td>
<td>82.01 ± 7.85</td>
<td>82.53 ± 10.37</td>
<td>70.20 ± 10.64</td>
<td>69.90 ± 16.12</td>
</tr>
<tr>
<td>GB</td>
<td>96</td>
<td>77.63 ± 17.72</td>
<td>66.07 ± 18.74</td>
<td>31</td>
<td>69.72 ± 25.72</td>
<td>72.59 ± 26.15</td>
<td>58.24 ± 25.09</td>
<td>57.86 ± 30.36</td>
</tr>
<tr>
<td>AG</td>
<td>95</td>
<td>55.46 ± 20.49</td>
<td>41.07 ± 19.40</td>
<td>36</td>
<td>44.12 ± 21.95</td>
<td>43.98 ± 22.10</td>
<td>30.72 ± 17.78</td>
<td>15.95 ± 16.57</td>
</tr>
<tr>
<td>Colon</td>
<td>56</td>
<td>65.55 ± 14.22</td>
<td>50.28 ± 14.83</td>
<td>42</td>
<td>46.83 ± 19.88</td>
<td>45.28 ± 18.82</td>
<td>32.59 ± 16.12</td>
<td>27.67 ± 17.78</td>
</tr>
<tr>
<td>SI</td>
<td>57</td>
<td>58.85 ± 16.82</td>
<td>43.63 ± 16.71</td>
<td>39</td>
<td>39.62 ± 17.72</td>
<td>41.10 ± 18.38</td>
<td>26.24 ± 14.34</td>
<td>25.79 ± 17.11</td>
</tr>
<tr>
<td>Stomach</td>
<td>102</td>
<td>88.59 ± 11.43</td>
<td>80.80 ± 12.86</td>
<td>46</td>
<td>59.42 ± 26.78</td>
<td>58.14 ± 25.73</td>
<td>46.95 ± 25.49</td>
<td>22.45 ± 17.98</td>
</tr>
<tr>
<td>Brain*</td>
<td>69</td>
<td>90.13 ± 17.80</td>
<td>85.29 ± 20.51</td>
<td>9</td>
<td>66.52 ± 37.45</td>
<td>65.51 ± 36.75</td>
<td>58.62 ± 35.63</td>
<td>35.59 ± 42.47</td>
</tr>
</tbody>
</table>
Table 3. Model performance reported in terms of Dice coefficient in strategy (a) compared to recent studies in the field. UB: Urinary Bladder, SC: Spinal Cord, GB: Gall Bladder, AG: Adrenal Gland, IVC: Inferior Vena Cava.

<table>
<thead>
<tr>
<th>Study</th>
<th>Liver</th>
<th>Heart</th>
<th>Pancreas</th>
<th>Kidneys</th>
<th>Spleen</th>
<th>UB</th>
<th>Esophagus</th>
<th>Femur Heads</th>
<th>Lungs</th>
<th>SC</th>
<th>Aorta</th>
<th>GB</th>
<th>AG</th>
<th>IVC</th>
<th>Colon</th>
<th>Rectum</th>
<th>Stomach</th>
<th>Eyes</th>
<th>Brain</th>
</tr>
</thead>
<tbody>
<tr>
<td>This study</td>
<td>96.98</td>
<td>91.95</td>
<td>77.52</td>
<td>94.13</td>
<td>94.68</td>
<td>83.78</td>
<td>73.20</td>
<td>95.72</td>
<td>97.63</td>
<td>89.85</td>
<td>93.55</td>
<td>79.32</td>
<td>54.62</td>
<td>80.57</td>
<td>70.60</td>
<td>77.90</td>
<td>86.22</td>
<td>91.13</td>
<td>97.39</td>
</tr>
<tr>
<td>Tang et al. (40)</td>
<td>96.60</td>
<td>76.00</td>
<td>93.80</td>
<td>96.30</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>92.30</td>
<td>82.60</td>
<td>73.60</td>
<td>85.30</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>85.70</td>
</tr>
<tr>
<td>Duan et al. (41)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Xiao et al. (42)</td>
<td>96.00</td>
<td>84.00</td>
<td>92.00</td>
<td>95.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>90.00</td>
<td>83.00</td>
<td>78.00</td>
<td>83.00</td>
<td>89.00</td>
<td>83.00</td>
<td>89.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Song et al. (43)</td>
<td>95.35</td>
<td>74.81</td>
<td>94.07</td>
<td>83.81</td>
<td>78.26</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>89.01</td>
</tr>
<tr>
<td>Wang et al. (44)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Crespi et al. (45)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>89.42</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shi et al. (1)</td>
<td>98.00</td>
<td>96.90</td>
<td>90.70</td>
<td>97.90</td>
<td>96.90</td>
<td>95.50(Male)</td>
<td>97.50</td>
<td>98.10</td>
<td>91.10</td>
<td>93.40</td>
<td>94.40</td>
<td>87.40</td>
<td>93.70</td>
<td>97.80</td>
<td>97.70</td>
<td>92.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F a et al. (19)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hao et al. (21)</td>
<td>95.47</td>
<td>74.52</td>
<td>95.49</td>
<td>96.15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>76.92</td>
<td>84.79</td>
<td>90.21</td>
<td>90.40</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Siciarz et al. (46)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chen et al. (47)</td>
<td>95.56</td>
<td>93.88</td>
<td>94.02</td>
<td>81.60</td>
<td>99.33</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>89.15</td>
</tr>
</tbody>
</table>

All rights reserved. No reuse allowed without permission.
Discussion

Automated organ segmentation is a critical step in a wide range of clinical applications, including personalized radiation dosimetry and quantification, and radiation treatment planning. The availability of a fast and reliable organ segmentation tool can facilitate the automation of these procedures and their adoption/deployment in clinical setting. In this work, we developed novel deep learning models to segment multiple organs from total-body CT images and compared the performance of our models with previous algorithms reported in the literature. Our model was trained on images presenting with high variability using large datasets, including adults, pediatrics and patients presenting with a wide range of pathologies and anatomic pathologies. The proposed model demonstrated an acceptable outcome even on cases with uncommon anatomies and pathologies, such as splenomegaly as shown in Error! Reference source not found. Besides, we used prior anatomical knowledge for some organs in the body for organ-specific post-processing to improve the outcome. This methodology enabled to successfully improve the results in nine organs by achieving higher Dice coefficients in strategy (a) and could help improving the Dice coefficient in more than five organs in strategy (b). We used the model output anchor organ segmentation as reference for prior knowledge implementation i.e., no manual segmentation or ground truth segmentation was needed in post-processing. The error in anchor organ masks was used as input for the post-processing function that can propagate to the post-processed mask and accumulated errors can be problematic. One possible explanation is that our decision algorithm to exclude false positive segmented voxels was not successful in a number of organs. We believe that using ground truth anchor organ segmentations for post-processing can improve the post-processing capabilities. According to the results achieved through post-processing, we suggest using these algorithms for organs achieving good performance (Table 2 and Table 3).

To evaluate the performance of the proposed model on real-world unseen external datasets, we tested our model on the online available dataset provided by Wasserthal et al. (37). Our model outperformed their model for five organs in strategies (b) and (c). For most small organs and gastrointestinal organs, our model’s performance was inferior to their model, which can be explained by the different 2D and 3D training strategies used in our and their models, respectively. They have used the nn-Unet (38) model which demands a high computational burden for both training and inference. Their images were resampled to 1.5 mm isotropic voxel dimension in a specific orientation. Conversely, we resampled the images again during cropping and resizing in strategy (b), while our test set in strategy (c), the images were in the original dimensions and the used preprocessing steps were similar to the training step in their model. As shown in

Table 2, the number of valid cases in the testing group was limited and the effect of statistical difference can be significant, while the number of our test dataset was larger. It should be mentioned that their brain images were blurred in the face region to preserve the privacy of subjects. This has affected the performance of our model in strategy (b). In addition, their initial model was trained and improved based on models trained on the same dataset used for training our models and then manually edited the segmentations. As such, our test dataset was not really unseen for their model in strategy (c), while in strategy (b) their local data were completely unseen for our model. The effect of post-processing improved the performance of image segmentation of five organs and there is still scope for improvement that can be explored in future studies by changing the number of anchor organs or providing manual edited segmentation as an aid to generate robust post-processing without initial anchor segmentation errors.
The comparison of our models with recent studies revealed that our model’s performance was better or at least comparable to algorithms reported in the literature for most organs, except for small and gastrointestinal organs (Table 3). One of the main merits of our proposed model is its lightweight requiring a small number of parameters (533 K). In addition, for each organ, we have a separate lightweight model, and the user can select a lower number of organs to be segmented to save time. We calculated the inference time on an NVIDIA RTX 4090 GPU where the average inference time for a total body CT was 1.67 seconds per case per organ. Besides, we tested inferring our model on an Intel Corei9 13900KF CPU and the inference time on the CPU was 14.5 seconds per study and per organ, which is a bearable and acceptable inference time for centers lacking access to dedicated GPUs.

We trained different deep learning models to segment 23 organs from total-body CT images which can be beneficial in various clinical tasks. We evaluated our models on an external dataset. The number of cases was limited to a few organs, and the segmentation criteria were different for each manual segmentation available from the online databases, inherently causing inter-observer variability, e.g., some databases provided whole segmented kidneys while others excluded pelvicalyceal systems. These differences may mislead our models and affect their performance.

Conclusion

We have developed a fully automated deep learning algorithm capable of generating accurate masks for multiple organs from CT images in an affordable computing time. After training these models on a diverse dataset comprising images from various databases, we compared the performance of our model with other algorithms on external datasets in real clinical scenarios. The proposed model exhibited remarkable capabilities even in cases involving uncommon anatomies and pathologies, such as splenomegaly.

Based on our analysis and results, we recommend using this algorithm especially for organs achieving excellent performance. One key advantage of our proposed models is their lightweight nature, enabling to run them efficiently on standard devices without access to dedicated GPUs in a bearable time. With an average GPU inference time of only 1.67 seconds per organ for a total-body CT image, they provide fast results and can be exploited in most routine tasks, even for verification in RT positioning. Overall, a reliable organ segmentation tool enables wider adoption by medical professionals in clinical setting.

Acknowledgments

This work was supported by the Euratom research and training programme 2019-2020 Sinfonia project under grant agreement No 945196.

Competing interests

The authors have no relevant financial or non-financial interests to disclose and the authors have no competing interests to declare that are relevant to the content of this article.
References


27. Patrick Blic1a PFCa, Eugene Vorontsov1e,1, Grzegorz Chlebus, Hao, Chenm, Qi Doum, Chi-Wing Fum, Xiao Hanp, Pheng-Ann Hengm, Jrgen Hesserq, Samuel, Kadourye, Tomasz Kopczyskiv, Miao Leo, Chunming Lio, Xiaoming Lim, Jana Lipkovaa, John Lowengrubn HM, Jan Hendrik Moltzr, Chris Pale,1, Marie Pirauda, Xiaojian Qim JQ,1, Markus Remplera, Karsten Rothq, Andrea Schenkr, Anjany, Sekuboyinaa PZ, Christian Hiulemeyera, Marcel Beetza, Florian Ettingera, Felix, et al. The Liver Tumor Segmentation Benchmark (LiTS). 2019.


