Accelerating 2D abdominal organ segmentation with active learning

Yu, Xin, Tang, Yucheng, Yang, Qi, Lee, Ho Hin, Bao, Shunxing, et al.

Xin Yu, Yucheng Tang, Qi Yang, Ho Hin Lee, Shunxing Bao, Ann Zenobia Moore, Luigi Ferrucci, Bennett A. Landman, "Accelerating 2D abdominal organ segmentation with active learning," Proc. SPIE 12032, Medical Imaging 2022: Image Processing, 120323F (4 April 2022); doi: 10.1117/12.2611595

SPIE.

Event: SPIE Medical Imaging, 2022, San Diego, California, United States

Accelerating 2D Abdominal Organ Segmentation with Active Learning

Xin Yu[a], Yucheng Tang[b], Qi Yang[a], Ho Hin Lee[a], Shunxing Bao[b], Ann Zenobia Moore[c], Luigi Ferrucci[c], Bennett A. Landman[a,b]

[a]Computer Science, Vanderbilt University, Nashville, TN;
[b]Electrical and Computer Engineering, Vanderbilt University, Nashville, TN;
[c]National Institute on Aging, Baltimore, MD

Abstract

Abdominal computed tomography CT imaging enables assessment of body habitus and organ health. Quantification of these health factors necessitates semantic segmentation of key structures. Deep learning efforts have shown remarkable success in automating segmentation of abdominal CT, but these methods largely rely on 3D volumes. Current approaches are not applicable when single slice imaging is used to minimize radiation dose. For 2D abdominal organ segmentation, lack of 3D context and variety in acquired image levels are major challenges. Deep learning approaches for 2D abdominal organ segmentation benefit by adding more images with manual annotation, but annotation is resource intensive to acquire given the large quantity and the requirement of expertise. Herein, we designed a gradient based active learning annotation framework by meta-parameterizing and optimizing the exemplars to dynamically select the 'hard cases' to achieve better results with fewer annotated slices to reduce the annotation effort. With the Baltimore Longitudinal Study on Aging (BLSA) cohort, we evaluated the performance with starting from 286 subjects and added 50 more subjects iteratively to 586 subjects in total. We compared the amount of data required to add to achieve the same Dice score between using our proposed method and the random selection in terms of Dice. When achieving 0.97 of the maximum Dice, the random selection needed 4.4 times more data compared with our active learning framework. The proposed framework maximizes the efficacy of manual efforts and accelerates learning.

Keywords: Multi-organ segmentation, Annotation, Active learning, 2D slices

1. INTRODUCTION

Segmentation of the key organs in abdominal computed tomography CT imaging provides valuable information to assess the body habitus and organ health [1]. Many attempts have been made to develop automatic multi-organ segmentation approaches. This remains a unresolved research field [2] and there are some major challenges. For example, the organs have low intensity contrast with other soft tissues which results in burry boundaries. Furthermore, the organ structures are morphological complex since the organ sizes vary substantially within and between subjects [3]. With the success of deep learning in sematic segmentation tasks, many deep learning based automatic multi-organ segmentation approaches have been proposed [4,5,6]. However, these methods largely rely on 3D volumes which provide rich contextual information and relatively large amounts of sample slices. These approaches are not applicable when single slice imaging is used to minimize dose. Lack of 3D context, variety of acquired image levels between subjects, and lack of images slices add more challenges for automatic segmentation algorithms and even for manual annotation on 2D single slices. For the massive amount of data deep learning based approach require, adding more data into training benefits the model performance which makes manual annotation a necessity.

For the task of 2D single slices segmentation, the image level of the CT imaging varies, which makes the abdominal structure different. As shown in Figure 1, some subjects have clear abdominal structure while others have ambiguous ones. Deep learning based approaches might perform well on those that have clear structure and fail on those that do not. The failed samples are referred as ‘hard cases’. Adding more ‘hard cases’ into the training set can facilitate the generalizability and discrimination of the model. In a traditional approach, however, annotation was conducted in order or randomly, as shown in Figure 2 (a). This can end up selecting the slices that always have the same image level, which is less helpful than the ‘hard cases’, which will in turn increase the annotation burden. To reduce annotation efforts and accelerate the

Medical Imaging 2022: Image Processing, edited by Olivier Colliot, Ivana Išgum, Proc. of SPIE Vol. 12032, 120323F · © 2022 SPIE
1605-7422 · doi: 10.1117/12.2611595

segmentation process, we proposed an active learning framework to actively and quantitively select the slices that are most beneficial to the training process, as shown in Figure 2 (b). The active learning model in the proposed framework provided a suggested annotation ranking of the unannotated dataset by given them different weight. Higher weight indicates less similarity between the samples with the training set. The top n slices were selected to conduct manual annotation.

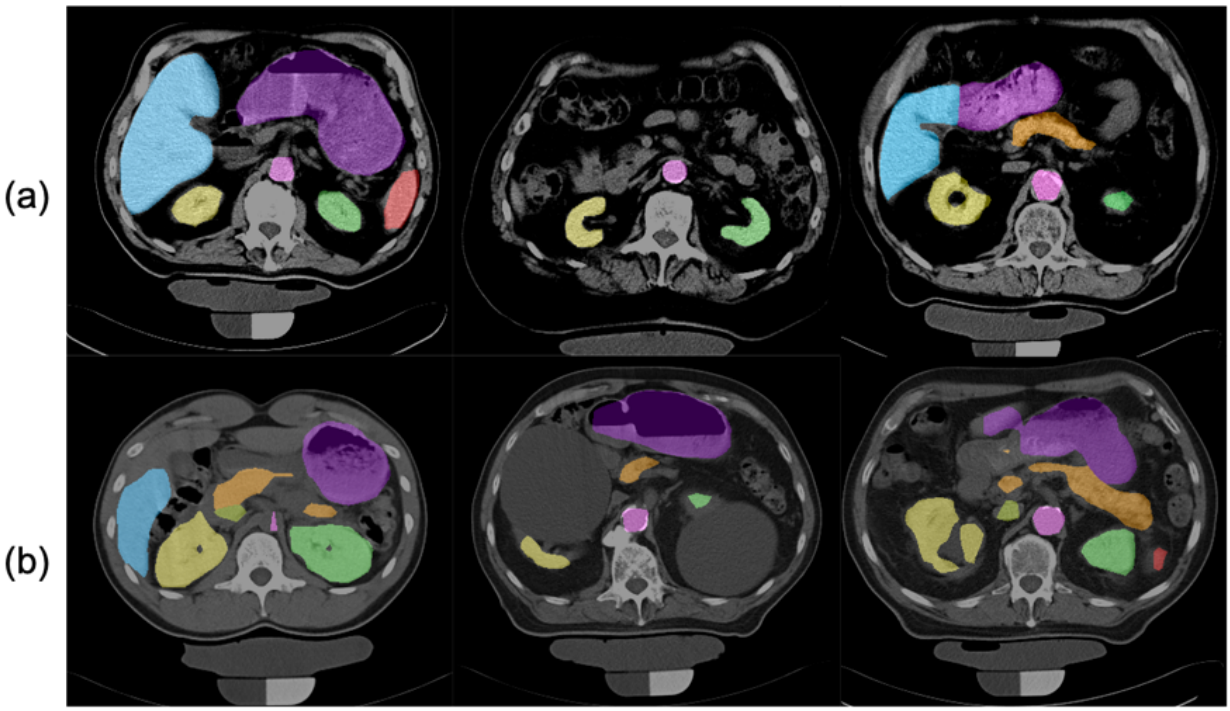

Figure 1. Abdominal segmentation results on 2D single slices of abdominal CT images. On some slices whose image level has clear abdominal structure, shown in (a), the deep learning model is able to achieve a high segmentation accuracy while they fail on some other slices that have ambiguous structure, shown in (b). To increase the model generalizability, we seek to add more 'hard case' slices illustrated in (b) into the training set.

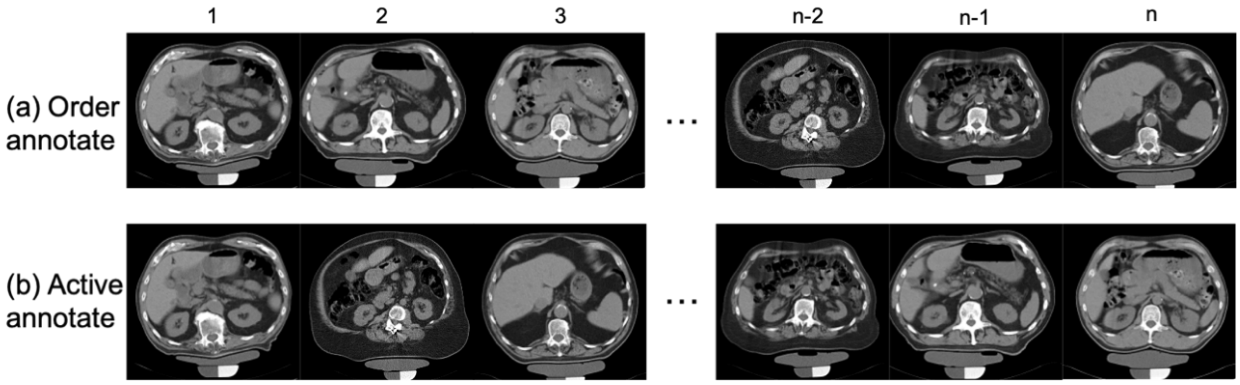

Figure 2. Unannotated 2D single slices of abdominal CT images. In a traditional approach, slices will be annotated in order or randomly as shown in (a) and can end up annotating the slices that are in the same image level which have similar abdominal structures (1,2,3 in (a)) and leave out the slices in the different image levels unannotated (n-2, n-1, n in (a)). We propose to actively choose the slices that have different abdominal structures to be annotated first, as shown in (b).

We conducted experiments on the Baltimore Longitudinal Study on Aging (BLSA) cohort. The method was evaluated on a total of 586 patients which starts from 286 patients and iterate for 5 times. The results were compared with random selection approach in terms of Dice coefficient. To achieve the same Dice coefficient, our active learning framework required substantively less data than the random selection approach.

2. METHODS

We designed an active learning framework that has three steps: a supervised segmentation model training step, an active sample selection step, and a manual annotation step. In the first step, we conducted 5-fold validation on training a segmentation model using our annotated dataset. In the second step, an active learning model was trained using the annotated dataset, and then applied on the unannotated dataset for testing. This step created a suggested annotation list for the unannotated dataset. In last step, we selected top n patients slices to be annotated, n=50 in our case, and added those into the annotated dataset. These three steps can be conducted iteratively by updating the annotated dataset and unannotated dataset each time. Figure 3 shows the overall pipeline of our proposed framework.

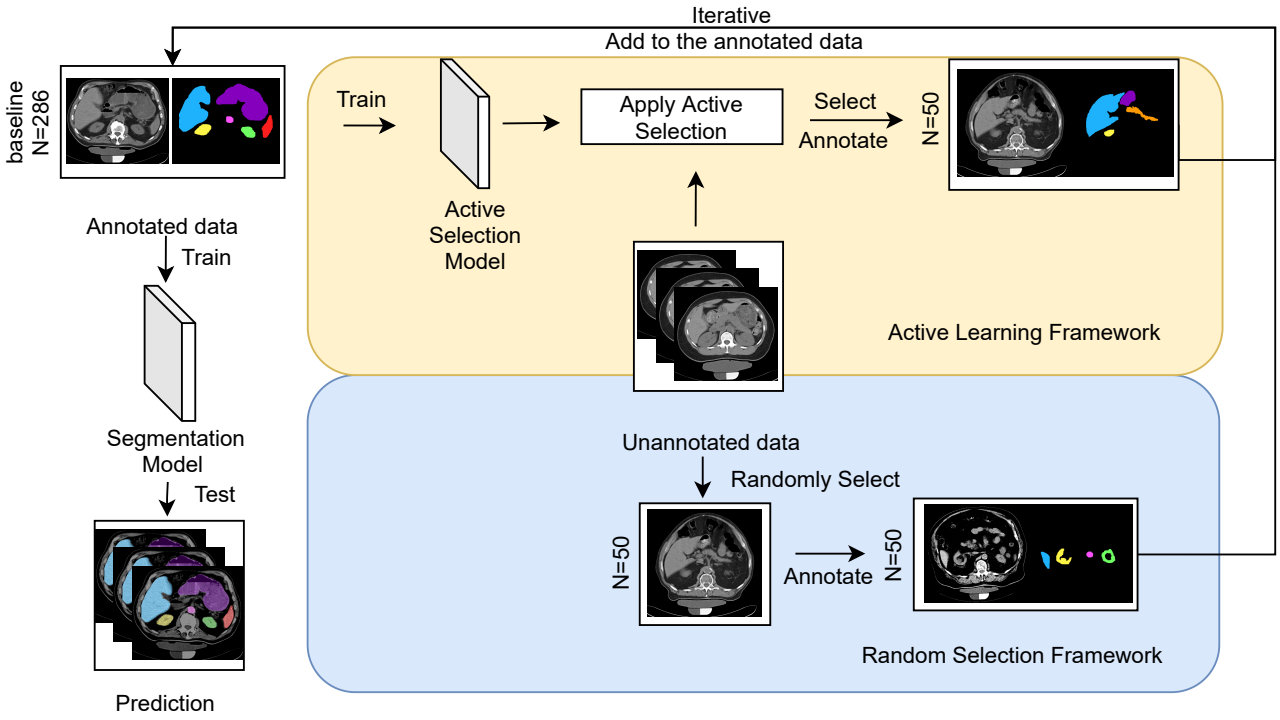

Figure 3. The pipeline of our proposed active learning framework and the comparison random selection framework. We started with a baseline segmentation model trained on 286 patient slices. In the active learning framework, we trained the active selection model with the annotated dataset. The unannotated dataset was applied to the trained active selection model and ranked each slice. The top 50 patient slices were selected to conduct the manual annotation, and then they were added to the annotated data. A new segmentation model was trained using the updated annotated dataset. This is the end of a round of active data selection. We iterated for 5 rounds by updating the unannotated dataset and annotated dataset in each round. In the random selection framework, the difference is that 50 patient slices were picked randomly from the unannotated dataset in each round to conduct the manual annotation.

2.1 Supervised segmentation model training

We adopted Deeplab-v3[7] with ResNet101[8] as our backbone. To make it more suitable for limited cases, we used the pre-trained weight from the ImageNet [9] to initialize our model. The total patients' slices were divided into 5 folds, and hence, 5 different segmentation models were trained.

2.2 Active data selection

Our active learning model was adapted from the method described in [10, 11, 12]. During training, the annotated data was passing to the duplicated segmentation network to reweight the training. For each sample, if the weight is high, which means the samples do not share a high degree of similarity with the other samples. The samples passed through the first network, the initial weight for each sample were the same. The new parameters were applied on the second network and the samples were used to finetune it. The weight was applied here during the gradient decent. The new parameters were used to update both networks and saved in the memory for next iteration. Based on the network performance on each sample, the weight gradient was calculated and used to update the weight for each sample. The weights were used to reweight the loss and further the gradient decent of the first network. During testing, the unannotated data was fed to the network to get the weight value for each slice in the unannotated data. The weight was sorted from large to small value. The larger the weight value, indicating the lower similarity between the training set and the slices. For this reason, we selected the top n patient slice to conduct the manual annotation and added into the training set, which is the n most representative patients.

2.3 Active selection pipeline iteration

The top n patient slices were selected to conduct manual annotation and added into the annotated dataset. Segmentation models were trained on the updated annotated dataset. This is the end of a round of active learning data selection. To iterate, we used the updated annotated dataset to train a new active learning model and applied it on the updated unannotated dataset, which gives us a new rank for the new unannotated dataset. Based on the new rank, another n patients were manually annotated and then added to the annotated dataset. A new segmentation model was trained based on the updated annotated dataset. This process was iterated until there were no images remaining in the unannotated dataset.

2.4 Random selection pipeline

For comparison, we designed a baseline random selection framework to validate the impact of the active selection model. In this framework, we randomly selected without replacement of the same number of samples for annotation from the unannotated data and added them into the annotated data as it in the active learning framework. A new segmentation model was trained using the updated annotated dataset. During iteration, only the annotated dataset and the unannotated dataset need to be updated and the process continue until the unannotated dataset was empty. Although the overall patients being used in active learning framework and random selection framework are the same, during each iteration, the annotated dataset and unannotated dataset are independent and different for these two frameworks except for the beginning of R1.

3. EXPERIMENTS AND RESULTS

3.1 Datasets and Implementations

This work is based on a subset of the 2D single slices of abdominal CT images of the Baltimore Longitudinal Study on Aging cohort. All the data has been approved by the Institutional Review Board (IRB) and accessed in de-identified form. A total of 586 patients were involved in study. We started with 286 patient slices and iterated for 5 times. For each iteration, 50 patients were added into the training dataset. We conducted 5-fold validation on all the experiments. For the baseline model with 286 patients, the patients were randomly divided into 5 folds, with 57/58 patients in each fold. During each experiment, 4 of the 5 folds were used as training dataset, and the other fold was used as testing. In each iteration, 50 new patients were randomly divided into 5-fold and added to the previous 5 folds. However, to make a fair comparison of the result, we keep the test set the same as the baseline model for both active learning framework and random selection framework. All the data have an image size of 512x512, and we processed them using the soft tissue CT window range [-125, 275] HU as in [6]. The intensity of the image slices was rescaled to [0, 255] and then the slices were randomly flipped with 0.5 probability and resized with range [0.5, 2.0] during online data augmentation. We used cross entropy as our loss function, and it was optimized by the stochastic gradient decent with a learning rate of 0.02 to train for 200 epochs.

3.2 Experimental Results

We evaluated the methods on 9 abdominal organs: spleen, right kidney, left kidney, gallbladder, liver, stomach, aorta, and pancreas in terms of Dice coefficient. Figure 4 (a) shows the average Dice coefficient in each round. When the patient

slices number reach 586, both active and random selection method have the same training dataset, so those patients were randomly divided into 5-fold to conduct the cross validation as maximum Dice coefficients it can achieved when includes all the patients into training process, $Dice_{max} = 0.784$. Comparing active and random selection, by adding 50 patient slices in R1, the Dice coefficient of the active selection framework is 15.8% higher than the random selection framework (0.771 versus 0.759). At R2, by adding 100 patients in total, the Dice coefficient of the active selection is 22.4% higher than the random selection (0.777 versus 0.760). Figure 4 (b) shows the amount of data required to be added to achieve the same Dice coefficient. When achieving 0.97 of the maximum Dice, the random selection needed 4.4 times more data compared with our active learning framework. The Dice coefficient difference starting to decrease gradually in the last few rounds mainly because the training dataset was becoming more and more similar. Figure 5 demonstrates the segmentation results changes on one patient slices from R1 to R5 using both active learning framework and random selection framework.

4. DISCUSSION AND CONCLUSION

Herein, we proposed an iteratively active learning framework to help dynamically select the 'hard cases' for manual annotation. We designed a random selection framework for comparison. The experiments iterated for 5 times by conducting the manual annotation on the selection patient slices and added to the annotated dataset. Our proposed framework required substantively less data to achieve the same Dice coefficient score compared with a random selection framework. This framework has the advantage to reduce the annotation efforts and accelerate the model training process.

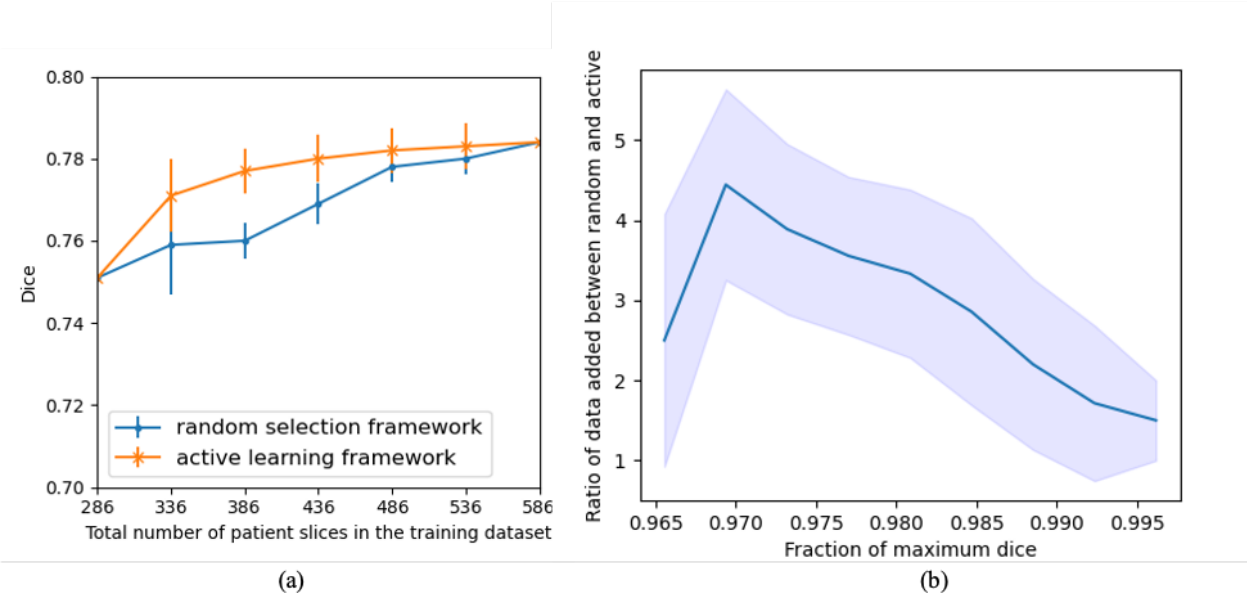

Figure 4. Comparison of our proposed active learning framework and random selection framework. We started with 286 patients in the baseline model and end with 586 patients. We conducted 5-fold validation on the framework iteratively for 5 time by adding 50 patient slices in each round. The error bar in (a) and confidence interval in (b) shows the standard error of the mean.

5. ACKNOWLEDEMENT

This research is supported by NSF CAREER 1452485 and the National Institutes of Health (NIH) under award numbers R01EB017230, R01EB006136, R01NS09529, T32EB001628, 5UL1TR002243-04, 1R01MH121620-01, and

T32GM007347; by ViSE/VICTR VR3029; and by the National Center for Research Resources, Grant UL1RR024975-01, and is now at the National Center for Advancing Translational Sciences, Grant 2UL1TR000445-06. This project was also supported by the National Science Foundation under award numbers 1452485 and 2040462. This research was conducted with the support from the Intramural Research Program of the National Institute on Aging of the NIH. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. This study was in part using the resources of the Advanced Computing Center for Research and Education (ACCRE) at Vanderbilt University, Nashville, TN. The identified datasets used for the analysis described were obtained from the Research Derivative (RD), database of clinical and related data.

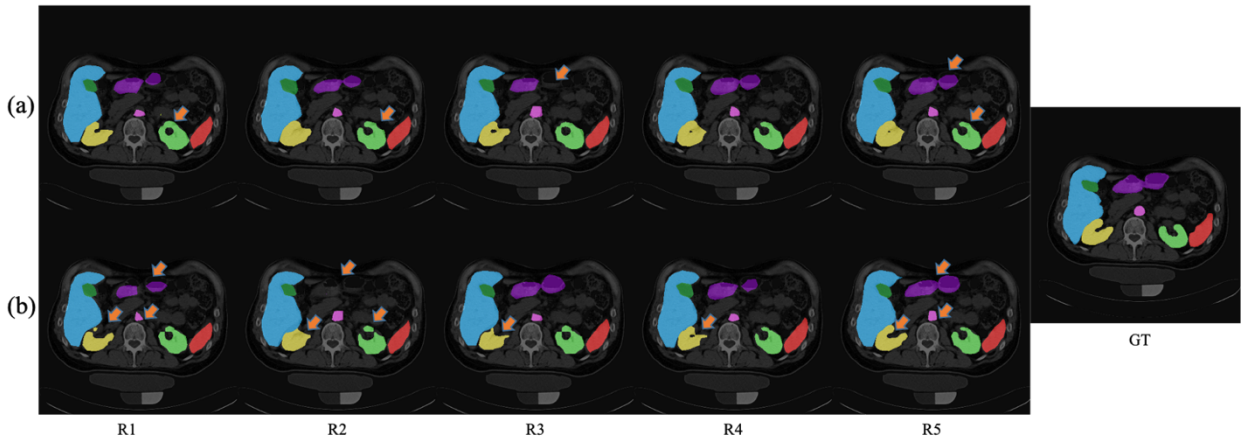

Figure 5. Visualization of a sample subject among 5 rounds. R1, R2, R3, R4 and R5 represents the 5 rounds, respectively. (a) shows the segmentation results from active learning framework. (b) shows segmentation results from random selection framework. GT shows the ground truth. Changes are highlighted with orange arrows.

6. REFERENCES

[1] E. M. Geraghty, J. M. Boone, J. P. McGahan, and K. Jain, "Normal organ volume assessment from abdominal CT," *Abdom Imaging*, vol. 29, no. 4, Jul. 2004, doi: 10.1007/s00261-003-0139-2.

[2] A. E. Kavur *et al.*, "CHAOS Challenge -- Combined (CT-MR) Healthy Abdominal Organ Segmentation," *Medical Image Analysis*, vol. 69, p. 101950, Apr. 2021, doi: 10.1016/j.media.2020.101950.

[3] Y. Wang, Y. Zhou, W. Shen, S. Park, E. K. Fishman, and A. L. Yuille, "Abdominal multi-organ segmentation with organ-attention networks and statistical fusion," *Medical Image Analysis*, vol. 55, pp. 88–102, Jul. 2019, doi: 10.1016/j.media.2019.04.005.

[4] H. H. Lee, Y. Tang, S. Bao, R. G. Abramson, Y. Huo, and B. A. Landman, "RAP-Net: Coarse-to-Fine Multi-Organ Segmentation with Single Random Anatomical Prior," *arXiv:2012.12425 [cs, eess]*, Dec. 2020, Accessed: Aug. 17, 2021. [Online]. Available: http://arxiv.org/abs/2012.12425

[5] Y. Tang *et al.*, "High-resolution 3D Abdominal Segmentation with Random Patch Network Fusion," p. 21.

[6] Y. Zhou *et al.*, "Prior-aware Neural Network for Partially-Supervised Multi-Organ Segmentation," *arXiv:1904.06346 [cs]*, Aug. 2019, Accessed: Aug. 11, 2021. [Online]. Available: http://arxiv.org/abs/1904.06346

[7] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking Atrous Convolution for Semantic Image Segmentation," *arXiv:1706.05587 [cs]*, Dec. 2017, Accessed: Aug. 11, 2021. [Online]. Available: http://arxiv.org/abs/1706.05587

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *arXiv:1512.03385 [cs]*, Dec. 2015, Accessed: Aug. 11, 2021. [Online]. Available: http://arxiv.org/abs/1512.03385

[9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2009, pp. 248–255. doi: 10.1109/CVPR.2009.5206848.

[10] M. Ren, W. Zeng, B. Yang, and R. Urtasun, “Learning to Reweight Examples for Robust Deep Learning,” *arXiv:1803.09050 [cs, stat]*, May 2019, Accessed: Apr. 18, 2021. [Online]. Available: http://arxiv.org/abs/1803.09050
[11] Z. Hu, B. Tan, R. Salakhutdinov, T. Mitchell, and E. P. Xing, “Learning Data Manipulation for Augmentation and Weighting,” *arXiv:1910.12795 [cs, stat]*, Oct. 2019, Accessed: Oct. 28, 2021. [Online]. Available: http://arxiv.org/abs/1910.12795
[12] T. Xu *et al.*, “Deep Mouse: An End-to-End Auto-Context Refinement Framework for Brain Ventricle amp; Body Segmentation in Embryonic Mice Ultrasound Volumes,” in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, Apr. 2020, pp. 122–126. doi: 10.1109/ISBI45749.2020.9098387.