Cascaded 3D U-Net with region-optimised Voxel Resolution for Abdominal Organ Segmentation - FLARE21 Challenge

Price Jackson
Peter MacCallum Cancer Centre
Melbourne, Australia
price.jackson@petermac.org

Lachlan McIntosh
Peter MacCallum Cancer Centre
Melbourne, Australia
lachlan.mcintosh@petermac.org

Abstract

In this work, an adaptation on the multi-stage 3D U-Net segmentation strategy is presented in which each organ is contoured independently. In the first stage, all cases are resampled and padded to match a standard "whole body" physical extent which will include all voxel data regardless of scan length. Label data in the first stage are expanded by 40mm improve class imbalance against a large background volume with label prediction using low-memory 3-resolution 3D U-Net. The predicted output labels are used to compute centre-of-mass for bounding box cropping in the second segmentation stage. The physical extent of the segmentation stage is optimised based on statistical analysis of all training cases. The resolution for segmentation of each organ is optimised to achieve a voxel spacing of approximately 1.5mm in each axis (kidneys, pancreas, and spleen) and 2.2mm for liver. The segmentation model was able to achieve dice scores of 0.811, 0.947, 0.953, 0.968, and 0.976 for pancreas, left kidney, right kidney, spleen, and liver, respectively.

1. Introduction

The task of abdominal organ segmentation in CT imaging presents a number of challenges as the image features and physical extent can vary appreciably between different tissues-of-interest. While it is preferable from a processing standpoint to contour all organs in a single computational stage, the accuracy of output will necessarily improve as the task becomes constrained to specific operations rather than approached as a single global task. For the purpose of creating a lightweight 3D segmentation network, the resolution of the input data and depth of the convolutional network in terms of stages and feature space are implicated in peak memory footprint required to process the image data. To best address this limitation, we propose to contour each organ independently thereby reducing the feature space required for any single encoding-decoding operation. Additionally, we present a reliable, general-purpose method to perform cascaded localisation and resolution-optimised segmentation as a two-stage process.

2. Method

2.1. Preprocessing

The baseline method includes the following preprocessing steps:

- Cropping strategy: Cascaded method with localisation and segmentation stages. Region-specific crop extents given in Table 2.
- Resampling method for anisotropic data:
  For region localisation, all images are resampled to the same physical extent of 400x400x1500mm (96x96x144 resolution). Segmentation stage employs region-specific cropping extents and resolutions to achieve nearly cubic voxel aspects as illustrated in Table 3.
- Intensity normalization method:
  No intensity normalisation applied - Hounsfield units considered as standard scale of physical density.

2.2. Proposed Method

- Network architecture details: 3D U-Net with standard 2x convolution + ReLu activation blocks at each resolution. 3x resolution levels for organ localisation and 4x levels for segmentation. Feature depths are reduced to consider memory footprint for each stage. Values are indicated in Figure 2.
- Loss function: Distance-weighted Dice (localiser) and Dice (final segmentation)
- Number of model parameters: 894,724 per region (localiser) and 12,938,562 per region (final segmentation)

Table 1. Data splits of FLARE2021.

Data Split	Center	Phase	# Num.
Training (361 cases)	The National Institutes of Health Clinical Center	portal venous phase	80
	Memorial Sloan Kettering Cancer Center	portal venous phase	281
Validation (50 cases)	Memorial Sloan Kettering Cancer Center	portal venous phase	5
	University of Minnesota	late arterial phase	25
	7 Medical Centers	various phases	20
Testing (100 cases)	Memorial Sloan Kettering Cancer Center	portal venous phase	5
	University of Minnesota	late arterial phase	25
	7 Medical Centers	various phases	20
	Nanjing University	various phases	50

- Number of flops: 1.54E11 (per localiser stage, 5x), 7.67E11 to 1.10E12 (per segmentation stage). Total: 5.28E12

2.3. Post-processing

Predicted labels are resampled to original CT resolution by linear interpolation and clipped above 0.5 value to achieve smooth contour edges at the native resolution. For each region, the largest contiguous label is kept in order to suppress the detection of small errant sub-regions.

2.4. Detailed Description of Methodology

Neural network architectures used for the task are comprised of modified 3D U-Net models (Tensorflow v2.4) with either 3 or 4 resolution levels for bounding box localisation or segmentation, respectively. Region localisation is performed by first resampling CT images and labels to a physical extent of 400x400x1500mm (x,y,z axes). CT images are augmented with random degrees of smoothing, sharpening, cropping, and rotation [1] and then localised in the centre of the extended "whole body" anatomical volume, the extents of which are illustrated in Figure 1. The dimensions of this physical space have been chosen as a standard general-purpose localisation pre-process which accounts for any clinical scan extent including PET/CT images ranging from vertex of skull to thighs. Though not specifically optimised to the FLARE21 task, the methodology has been shown to be reliable for a number of other clinical applications. Labelled organs in this stage are expanded by spherical convolution to increase the boundaries by 25mm in order to prevent a class imbalance with the extended background physical space. The symmetric expansion does not appreciably alter the computation of label centre-of-mass and provides consistently accurate region localisation.

Figure 1 - Illustration of whole-body physical extent pre-processing and label expansion for training localisation to pancreas centre-of-mass. True label is shown in red and prediction in blue with crosshairs representing label centre overlaid.

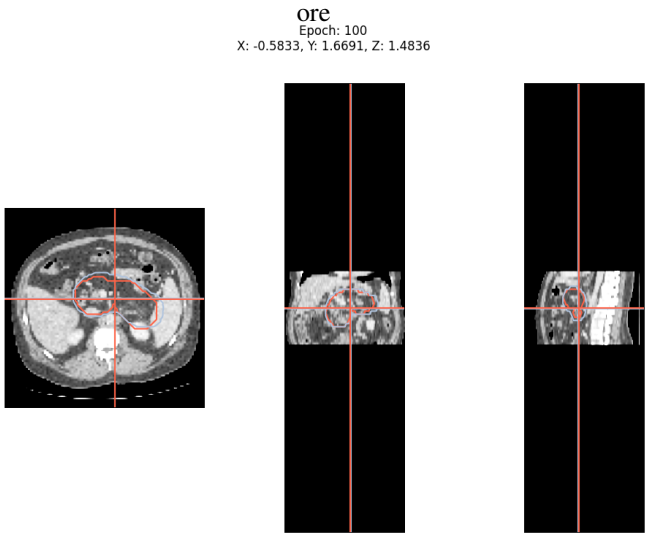

Figure 1. Pre-processing and model prediction for pancreas localisation. The input CT is expanded to a generic physical bounding box of 400x400x1500mm. Accuracy of centre-of-mass determination can be visually assessed by crosshairs and physical offsets in millimetres for this case are shown in header.

A secondary pre-processing consideration was to divide kidneys into separate right and left organs as this can be performed through automated means and is a sensible approach to reduce the required bounding box volume for each region. It should also offer greater accuracy by developing two task-specific models as opposed to a single kidney network that must balance feature learning for both structures. To split the labelled regions, we consider the patient midline to be approximated by the physical centre of mass in the lateral direction of the CT scan. This considers the density of the CT according to Hounsfield number and accounts for lateral set-up offsets of the patient on the reconstructed image field. Subsequently, in the left and right zones of the image the largest contiguous label with a volume greater than 20cc for each side is utilised for training with typical output illustrated in Figure 3.

Figure 2 - 3- and 4-resolution 3D U-Net models employed for region localisation and organ segmentation.

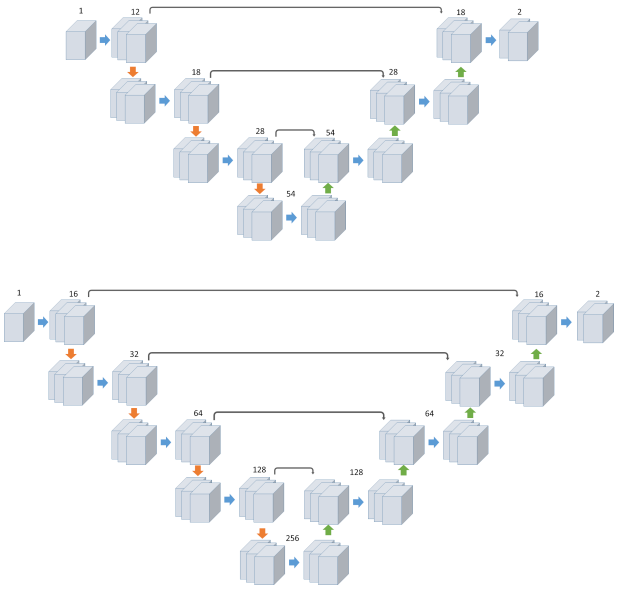

Figure 2. Architecture of 3D U-Net models used for bounding box localisation and region segmentation. Input image resolutions for organ localisation are 96x96x144 voxels (x,y,z) and optimised resolutions for segmentation of detected bounding box regions are table 3. Other than modifications of feature depth, a standard 3D U-net structure comprised of double convolution blocks at each resolution stage and a down or upsampling factor of 2x applied for each vertical arrow.

Figure 3 - Separation of right and left kidneys based on CT physical midline to improve renal segmentation precision.

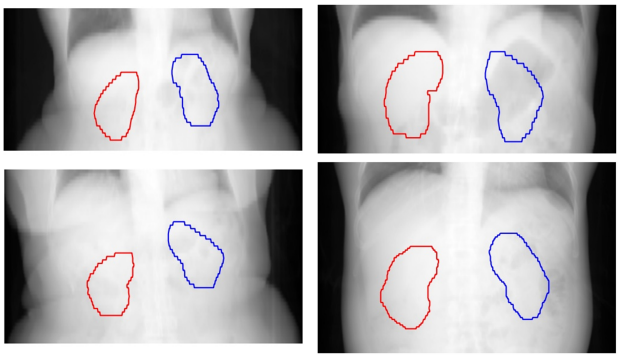

Figure 3. Automated pre-processing determination of right and left renal labels.

For both stages, training is performed by minimising Dice loss. In the localisation stage, additional weight is applied to voxels based on the distance from the label boundary in order to minimise the prediction of non-contiguous anatomy which could substantially offset the centre-of-mass determination. For region localisation, a lightweight 3-resolution version of the 3D U-Net (3 downsampling and skip blocks) was used with modest filter depths of 12, 18, 28, and 54 feature channels at each resolution [2]. For training the segmentation model, a 3D U-Net with 4 resolution downsampling and skip connection stages was designated. Feature depths at each resolution were set to 16, 32, 64, 128 and 256 channels.

To consider the memory usage in the finer image segmentation stage, the physical extent and resolution used for each organ was optimised independently. Bounding box extent was determined based on the population 90th percentile plus an additional 40mm padding in each axis. Based on the optimal extents, voxel spacing for each organ was chosen to achieve an approximate cubic resolution of 1.5mm. For liver, a slightly coarser resolution was used of 2.2mm to acount for its larger anatomical size. Using this convention, each segmentation model could be limited to an input footprint of 1.4-2.3M voxels. Model training considered 85 percent of available training cases and best model weights on the remaining validation data was determined for the first 100 epochs. The number of trainable parameters for localisation and segmentation models was 894,724 and 12,938,562, respectively

In FLARE21 competition scoring, test case inference was computed with CPU processing. Computation time and memory footprint in the final leaderboard may reflect this, however, inference can be designated with graphical processing with the appropriate system configuration.

3. Dataset and Evaluation Metrics

3.1. Dataset

- A short description of the dataset used:
  The dataset used of FLARE2021 is adapted from MSD [3] (Liver [4], Spleen, Pancreas), NIH Pancreas [5, 6, 7], KiTS [8, 9], and Nanjing University under the license permission. For more detail information of the dataset, please refer to the challenge website and [10].

- Details of training / validation / testing splits:
  The total number of cases is 511. An approximate 70%/10%/20% train/validation/testing split is employed resulting in 361 training cases, 50 validation cases, and 100 testing cases. The detail information is presented in Table 1.

3.2. Evaluation Metrics

- Dice Similarity Coefficient (DSC)
- Normalized Surface Distance (NSD)
- Running time
- Maximum used GPU memory (when the inference is stable)

4. Implementation Details

4.1. Environments and requirements

A description of the environment used for deployment of the method, including but not limited to the items illustrated in Table 4.

The environments and requirements of the baseline method is shown in Table 4.

4.2. Training protocols

Full description of the training protocols, including but not limited to the items illustrated in Table 5.

4.3. Testing protocols

Description of inference strategy to get the finial output on test dataset.

- Pre-processing steps of the network inputs:
  CT Resampled to "Whole Body" reference frame and 3D U-net inference used to localize to each organ. Secondary high resolution 3D U-net processed for each organ based on localised bounding box.
- Post-processing steps of the network outputs:
  Inferred labels upsampled to original CT resolution. Largest contiguous label kept for each organ.
- Patch-based strategy:
  Direct localisation of final crop bounding box by low-resolution region adaptation as cascaded method. No merging of multiple patches required.

5. Results

Based on segmentation of region-localised models with a separate network for each organ-of-interest, an average testing Dice accuracy in the range of 0.84-0.97 can be achieved. The average and median scores for testing data are summarised in Table 6. Training dice accuracy was comparable to the validation output for all organ models; a factor we attribute to a robust data augmentation scheme. We propose that the cascaded, organ-specific strategy is a reasonable approach to 3D image segmentation if constrained to a modest memory footprint. While this provides reasonable accuracy and smaller overall network weights, it does come at the cost of overall computation time as inference is required to run on each organ separately and may not be a practical approach for delineating a larger subset of anatomical regions.

Figure 4 - Model training accuracy for segmentation stage in pancreas

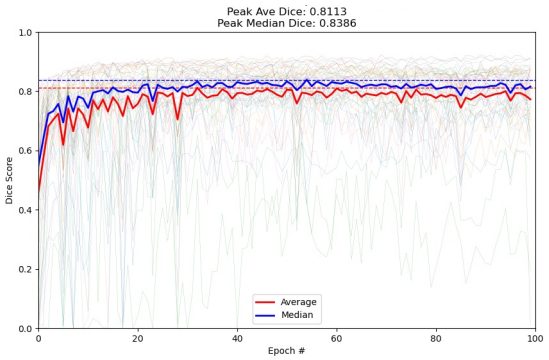

Figure 4. Recorded Dice score for pancreas testing cases during segmentation stage training. Scores for individual cases are indicated in fine lines with average(red) and median (blue) accuracy summarised by bold plots.

Figure 5 - Model training accuracy for segmentation stage other abdominal regions

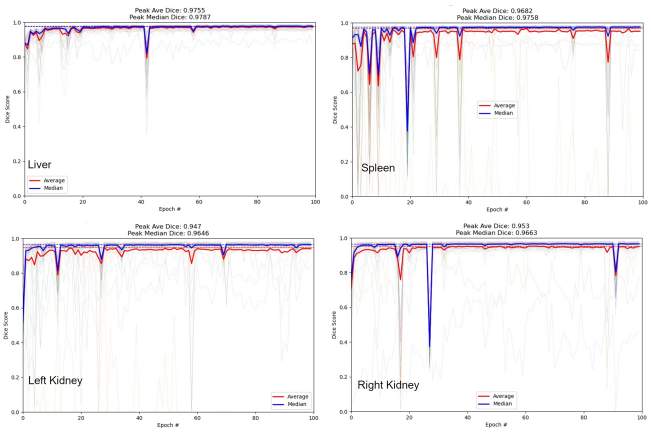

Figure 5. Testing Dice score accuracy for liver, spleen, left- and right kidneys based on 15 percent validation case split of all training data

5.1. Quantitative results on Test Set

Dice coefficient accuracy during training as scored on the 15% testing dataset are illustrated in 4 and 5. Individual cases are indicated by fine lines and group mean (red) and median (blue). It should be noted that scores are calculated

Table 2. Organ Bounding Box Offsets With Respect to Localiser Centres-of-mass

region	xmin	xmax	ymin	ymax	zmin	zmax	x_extent	y_extent	z_extent
liver	-122.4	192.1	-147.5	153.7	-157.5	117.7	314.5	301.2	275.3
lt_kidney	-82.9	81.0	-81.9	82.6	-102.8	100.8	163.9	164.5	203.6
pancreas	-108.3	135.5	-77.2	99.1	-97.2	88.5	243.8	176.3	185.6
rt_kidney	-79.0	81.1	-86.2	85.5	-99.5	99.2	160.2	171.7	198.8
spleen	-105.5	88.9	-113.6	94.3	-102.8	98.3	194.4	207.9	201.1

Table 3. Voxel Spacing for Organ Segmentation Models

region	x voxels	y voxels	z voxels	x spacing	y spacing	z spacing	Total Voxels (M)
liver	128	128	112	2.46	2.35	2.46	1.84
lt_kidney	112	112	128	1.46	1.47	1.59	1.61
pancreas	160	112	128	1.52	1.57	1.45	2.29
rt_kidney	112	112	128	1.43	1.53	1.55	1.38
spleen	128	128	128	1.52	1.62	1.57	2.10

Table 4. Environments and requirements.

Windows/Ubuntu version	Ubuntu 18.04.5 LTS
CPU	Intel(R) Core(TM) i9-7900X CPU@3.30GHz
RAM	64GB
GPU	Nvidia V100-32G
CUDA version	11.0
Programming language	Python3.6.9
Deep learning framework	Tensorflow 2.4
Specification of dependencies	
(Optional) code is publicly available at	github.com/jacksonmedphysics/FLARE21/

Table 5. Training protocols.

Data augmentation methods	Rotations, Translations, Gaussian noise, Gaussian blur, Sharpening, Brightness.
Initialization of the network	"glorot uniform" initialization
Patch sampling strategy	Cascaded: low resolution whole-volume localisation, with single segmentation bounding box region in 2nd stage
Batch size	1
Patch size	Localisation: 96×96×144, segmentation: organ-specific per Table 3
Total epochs stage	100 per organ
Optimizer	Adam
Initial learning rate	0.0001
Learning rate decay schedule	$\text{beta}_1 = 0.9, beta_2 = 0.999, epsilon = 1e-8, decay = 0.0$
Stopping criteria, and optimal model selection criteria	Stopping at maximum number of epochs, model with highest validation score saved.
Training time	localisation: 19 hours per organ, segmentation: 24-38 hours per organ
CO_2eq[1]	

on the post-cropped labels and the accuracy may be slightly lower when resampled to the original image resolution.

5.2. Quantitative results on External Validation

Results for the dice coefficient and normalised surface dice for the withheld validation set are presented in 7. The dice coefficients are consistently lower than the blind cases used to score training. Liver and spleen are comparable, however there is an appreciable reduction in accuracy for pancreas and kidney. It is noted that the standard deviation of scores in these regions is relatively high, indicative of very poor performance in a subset of cases.

5.3. Qualitative results

Figure 6 presents some challenging examples. It can be found that the baseline method can not segment the lesion-affected organs well. The first row of Figure 6 illustrates the accuracy of a reliable case. Our method displays unpredictable performance, however, in the presence of very high contrast opacity (abnormally high HU in liver and kidneys, Figure 6) perhaps due to irregular contrast phase or patient physiology in comparison to bulk training dataset.

6. Discussion and Conclusion

Pancreas proved to be the most challenging organ for segmentation, at least when assessed based on volumetric agreement by Dice score. This is to be expected in part because of it's relative surface area as an elongated structure in the abdomen. Additionally, the natural appearance may

Table 6. Quantitative results on Testing set.

region	Average Testing Dice Score	Median Testing Dice Score
liver	97.6	97.9
lt_kidney	94.7	96.5
pancreas	81.1	83.9
rt_kidney	95.3	96.6
spleen	96.8	97.6

Table 7. Quantitative results on validation set.

Organ	DSC (%)	NSD (%)
Liver	93.5±9.35	73.6±14.9
Kidney	78.7±25.5	71.0±22.6
Spleen	91.0±17.5	83.0±18.6
Pancreas	60.4±29.4	45.5±24.9

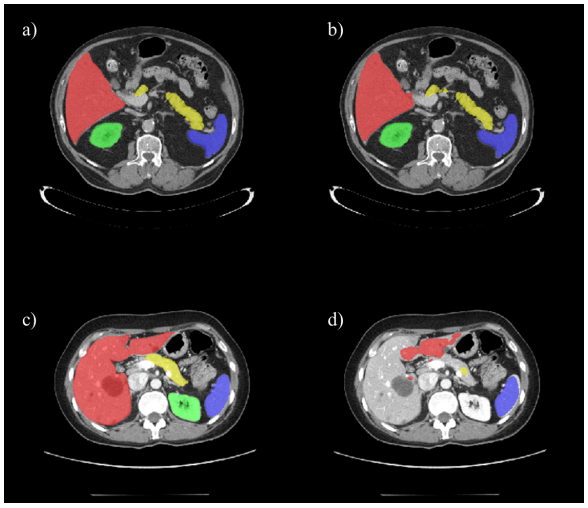

Figure 6. Challenging examples. First column is the ground truth and second column is the predicted results by our method. A well performing case is illustrated in the first row. A case with very high contrast uptake to liver and left kidney appears to have been problematic for the investigated segmentation model.

be more variable between subjects; particularly in contrast-enhanced CT where the degree of contrast uptake may be affected by an individual's physiology or the time between injection and scan initialisation.

In an effort to simplify the segmentation task at each stage, we also present an anatomically considered approach to separate renal labels according to an automated and straightforward methodology. Though it was beyond the scope of preparations for the current challenge submission, we plan to evaluate the relative improvement (or otherwise) on accuracy when using this approach compared to a single "whole renal" contouring model in the future.

The work presented provides a general purpose approach to 3D organ contouring. Region localisation is flexible to account for any physical scan extent. Class imbalance for very small regions during localisation may be overcome by automated label expansion which preserves the approximate object centroid. Additionally, we present a reasonable statistical approach to optimise bounding box size and second-stage cascaded segmentation resolution.

Acknowledgment

The authors of this paper declare that the segmentation method they implemented for participation in the FLARE challenge has not used any pre-trained models nor additional datasets other than those provided by the organizers.

Price Jackson is supported through a research fellowship with the Victorian Cancer Agency. We would also like to acknowledge the support of Jason Ellul and the Research Computing facility at Peter MacCallum Cancer Centre

References

[1] P. Jackson, N. Hardcastle, N. Dawe, T. Kron, M. S. Hofman, and R. J. Hicks, "Deep learning renal segmentation for fully automated radiation dose estimation in unsealed source therapy," *Frontiers in oncology*, vol. 8, p. 215, 2018. 2

[2] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: learning dense volumetric segmentation from sparse annotation," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2016, pp. 424–432. 3

[3] A. L. Simpson, M. Antonelli, S. Bakas, M. Bilello, K. Farahani, B. Van Ginneken, A. Kopp-Schneider, B. A. Landman, G. Litjens, B. Menze *et al.*, "A large annotated medical image dataset for the development and evaluation of segmentation algorithms," *arXiv preprint arXiv:1902.09063*, 2019. 3

[4] P. Bilic, P. F. Christ, E. Vorontsov, G. Chlebus, H. Chen, Q. Dou, C.-W. Fu, X. Han, P.-A. Heng, J. Hesser *et al.*, "The liver tumor segmentation benchmark (lits)," *arXiv preprint arXiv:1901.04056*, 2019. 3

[5] H. Roth, A. Farag, E. Turkbey, L. Lu, J. Liu, and R. Summers, "Data from pancreas-ct. the cancer imaging archive (2016)." 3

[6] H. R. Roth, L. Lu, A. Farag, H.-C. Shin, J. Liu, E. B. Turkbey, and R. M. Summers, "Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2015, pp. 556–564. 3

[7] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle *et al.*, "The

cancer imaging archive (tcia): maintaining and operating a public information repository," *Journal of digital imaging*, vol. 26, no. 6, pp. 1045–1057, 2013. 3

[8] N. Heller, F. Isensee, K. H. Maier-Hein, X. Hou, C. Xie, F. Li, Y. Nan, G. Mu, Z. Lin, M. Han *et al.*, "The state of the art in kidney and kidney tumor segmentation in contrast-enhanced ct imaging: Results of the kits19 challenge," *Medical Image Analysis*, vol. 67, p. 101821, 2021. 3

[9] N. Heller, S. McSweeney, M. T. Peterson, S. Peterson, J. Rickman, B. Stai, R. Tejpaul, M. Oestreich, P. Blake, J. Rosenberg *et al.*, "An international challenge to use artificial intelligence to define the state-of-the-art in kidney and kidney tumor segmentation in ct imaging." *American Society of Clinical Oncology*, vol. 38, no. 6, pp. 626–626, 2020. 3

[10] J. Ma, Y. Zhang, S. Gu, C. Zhu, C. Ge, Y. Zhang, X. An, C. Wang, Q. Wang, X. Liu, S. Cao, Q. Zhang, S. Liu, Y. Wang, Y. Li, J. He, and X. Yang, "Abdomenct-1k: Is abdominal organ segmentation a solved problem?" *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 3