Learned delineation of the gross tumor volume incorporating intra-observer variability

T Marin[1], C Ma[1], R M Lahoud[1], F Xing[1], P Wohlfahrt[1], M Moteabbed[1], J Woo[1], X Ma[1], K Grogg[1], Y-L E Chen[1], G El Fakhri[1]
[1]Massachusetts General Hospital, Harvard Medical School, Boston, USA

JULY 12–16 2020 VIRTUAL
JOINT AAPM | COMP MEETING
EASTERN TIME [GMT-4]

INTRODUCTION

Radiation therapy is a critical tool for cancer treatment

- Tumor delineation for treatment planning is a major bottleneck
- Contouring typically suffers from (inter/intra-reader) variability (Figure 1)
- Goal
  - Develop a deep learning-based method for delineation of the gross tumor volume (GTV) in computed tomography (CT) scans of patients with soft tissue sarcomas.
  - Utilize intra-reader variability to predict GTV as well as confidence maps.

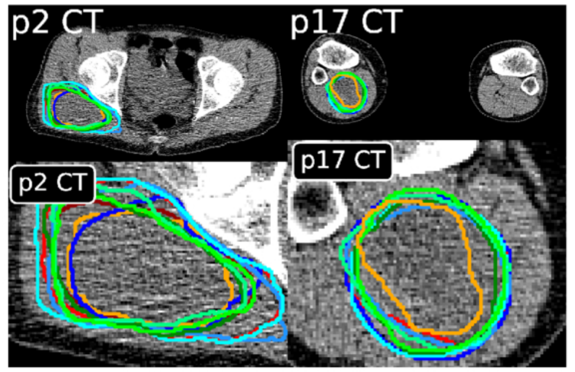

Figure 1. CT scans and corresponding gross tumor volume (GTV) delineated by three readers and multiple trials. The three blue contours, from the same reader, are used for training.

DATASET

- CT scans were acquired for 15 patients with soft tissue sarcomas chordomas
- Dataset was completed with publicly available datasets from The Cancer Imaging Archive [3] (McGill University, Canada) for a total of 45 patients
- GTV was specified three times by a clinician in three independent sessions
- The intersection (consensus), exclusive disjunction (partial agreement) and non-disjunction (healthy tissue) of the three GTV contours were used to classify CT voxels
- The input data is split into 2D slices for training

MODEL – DEEP NEURAL NETWORK

A deep convolutional neural network (Figure 2) based on the 2D U-Net architecture [1] with attention filters [2] in the expanding path was designed to learn the classification and thus the intra-observer variability in GTV delineation. The network was trained to minimize an ordinal cross-entropy loss [4].

Five datasets were excluded from the training and used only for validation.

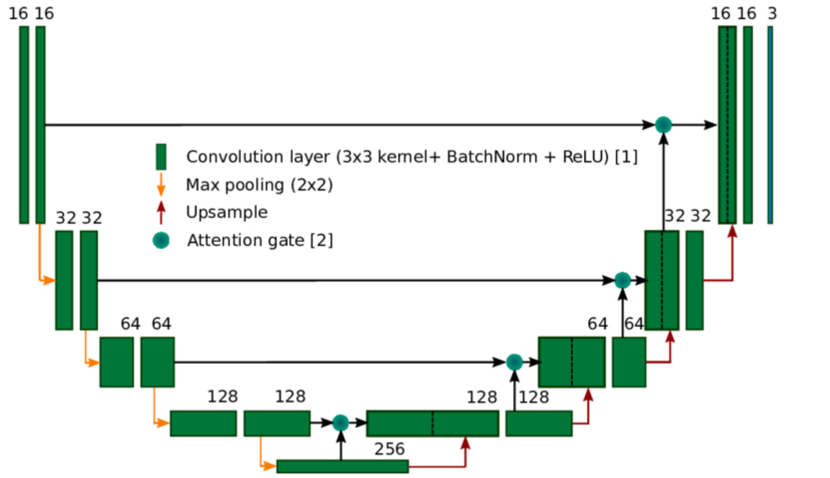

Figure 2. Network architecture. The contracting path performs successive convolutions and downsampling (max pooling) operations (the number of convolution units for each layer is indicated near the top of the block) producing feature maps at different scales. The expanding path concatenates upsampled feature maps (obtained using transposed convolution layers) with spatial information from the contracting path (skip connection) filtered by attention gates. The output layer encodes the three possible output classes (background, partial agreement, consensus).

RESULTS

Figure 3 shows CT slices with the reference confidence map calculated from the three trials along with the corresponding confidence maps predicted by the proposed neural network. The orange (resp. yellow) overlay corresponds to the consensus (resp. partial agreement) region. Visual analysis of the predicted masks suggest that:

- Most of the consensus region is correctly labeled
- A large part of the partial agreement region is also correcly labeled

Segmentation metrics were computed for both the consensus and the partial agreement contours. Besides traditional classification metrics (e.g. accuracy), the following metrics were computed:

- Dice coeffcient: overlap measured defined as: $\mathrm{Dice}(X,Y) = \dfrac{2\,|X \cap Y|}{|X| + |Y|}$
- Hausdorff distance: distance between contours, defined as:

$$H(X,Y) = \frac{1}{2}\left(h(X,Y) + h(Y,X)\right)$$

where $h(X,Y) = \max_{x \in X} \min_{y \in Y} \|x - y\|$

Results are shown in Table I. All metrics were computed in 3D for the consensus region, the partial agreement region and considering the input confidence map and prediction as continuous probability maps. The computed metrics demonstrate promising results (considering the relatively small dataset size)

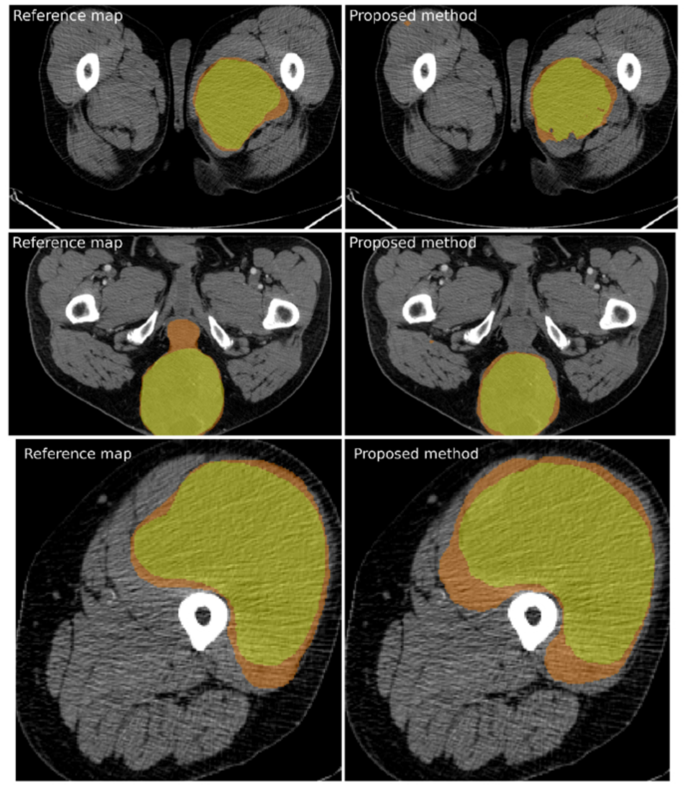

Figure 3. Examples of true (left) and predicted (right) GTV confidence maps. The yellow region corresponds to the consensus segmentation (i.e. pixels that were included by the reader for all three trials). The orange region indicates the partial agreement region. The predicted masks match the input confidence maps with reasonable accuracy.

Table I. Quantitative metrics measured in consensus and partial agreement regions. Metrics are calculated on 3D volumes, using the median over the evaluation dataset. Continuous metrics were obtained by using a fuzzy-logic extension of binary metrics.

	Consensus	Consensus + partial agreement	Continuous
Dice	0.79	0.81	0.80
Accuracy	0.90	0.91	0.92
Sensitivity	0.73	0.83	0.79
Specificity	0.95	0.94	0.95
Precision	0.82	0.79	0.80
Average Hausdorff distance (mm)	1.13	1.16	-

CONCLUSIONS

- The proposed deep convolutional neural network is able to accurately and reliably localize the GTV and integrate its intra-observer variability
- This can be a valuable tool to clinicians in GTV delineation thanks to its ability to accurately identify highly confident as well as uncertain tumor regions. For instance, it can be use to provide an initial GTV delineation

ACKNOWLEDGEMENTS

This work was supported in part by the National Institutes of Health under awards: T32EB013180, R01CA165221, R21EB021710 and P41EB022544.

REFERENCES

[1] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv*, Vol 1505; 2015.
[2] Oktay O, Schlemper J, Le Folgoc L, Lee M, Heinrich M, Misawa K, et al. Attention U-Net: Learning Where to Look for the Pancreas. *arXiv*, Vol 1804; 2018.
[3] Vallieres M, Freeman CR, Skamene SR, El Naqa I. A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities. *Physics in Medicine and Biology*. 2015;60(14):5471-96.
[4] Cheng J, Wang Z, Pollastri G. A neural network approach to ordinal regression. In: IEEE International Joint Conference on Neural Networks.; 2008. p. 1279-84.

CONTACT INFORMATION

Thibault Marin: tmarin@mgh.harvard.edu