

NEW Solution for Motion Synchronization QA



THE ONLY PLATFORM MOVING IN 7 DIMENSIONS*

We provide a realistic pre-treatment verification of the delivered treatment for Accuray Radixact® with Synchrony®.

In collaboration with Accuray, ScandiDos has developed a solution that improves the quality assurance (QA) of radiotherapy treatments of moving targets. The solution independently simulates the breathing motion of patients, therefore, adding a seventh dimension to the tumor movement simulation provided by the Delta4 HexaMotion.

*longitudinal, lateral, height, roll, tilt, time and breathing motion.

Learn more ►

Delta4family.com

Interactive contouring through contextual deep learning

Michael J. Trimpl^{a)}

Mirada Medical Ltd, Oxford, UK

Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, Oxford, UK
Oxford Institute for Radiation Oncology, University of Oxford, Oxford, UK

Djamal Boukerroui

Mirada Medical Ltd, Oxford, UK

Eleanor P. J. Stride

Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, Oxford, UK

Katherine A. Vallis

Oxford Institute for Radiation Oncology, University of Oxford, Oxford, UK

Mark J. Gooding

Mirada Medical Ltd, Oxford, UK

(Received 4 September 2020; revised 31 January 2021; accepted for publication 10 March 2021; published 3 May 2021)

Purpose: To investigate a deep learning approach that enables three-dimensional (3D) segmentation of an arbitrary structure of interest given a user provided two-dimensional (2D) contour for context. Such an approach could decrease delineation times and improve contouring consistency, particularly for anatomical structures for which no automatic segmentation tools exist.

Methods: A series of deep learning segmentation models using a Recurrent Residual U-Net with attention gates was trained with a successively expanding training set. Contextual information was provided to the models, using a previously contoured slice as an input, in addition to the slice to be contoured. In total, 6 models were developed, and 19 different anatomical structures were used for training and testing. Each of the models was evaluated for all 19 structures, even if they were excluded from the training set, in order to assess the model's ability to segment unseen structures of interest. Each model's performance was evaluated using the Dice similarity coefficient (DSC), Hausdorff distance, and relative added path length (APL).

Results: The segmentation performance for seen and unseen structures improved when the training set was expanded by addition of structures previously excluded from the training set. A model trained exclusively on heart structures achieved a DSC of 0.33, HD of 44 mm, and relative APL of 0.85 when segmenting the spleen, whereas a model trained on a diverse set of structures, but still excluding the spleen, achieved a DSC of 0.80, HD of 13 mm, and relative APL of 0.35. Iterative prediction performed better compared to direct prediction when considering unseen structures.

Conclusions: Training a contextual deep learning model on a diverse set of structures increases the segmentation performance for the structures in the training set, but importantly enables the model to generalize and make predictions even for unseen structures that were not represented in the training set. This shows that user-provided context can be incorporated into deep learning contouring to facilitate semi-automatic segmentation of CT images for any given structure. Such an approach can enable faster de-novo contouring in clinical practice. © 2021 Mirada Medical Ltd. *Medical Physics* published by Wiley Periodicals LLC on behalf of American Association of Physicists in Medicine. [https://doi.org/10.1002/mp.14852]

Key words: contouring, CT, deep learning, interactive, Radiotherapy, segmentation

1. INTRODUCTION

Many clinical procedures rely on accurate contouring of anatomical structures in medical images. For example, image segmentation is extensively used in radiotherapy planning to identify healthy and cancerous regions.¹ Accurate segmentation of both the tumor and the healthy tissue in the image, is essential to maximize the dose of radiation delivered to the tumor while minimizing the dose delivered to healthy tissues.

Manual contouring is a time-consuming process and is subject to significant intra- and interuser variability. As a result, semi-automatic and fully automatic image segmentation tools have been developed not only to reduce contouring time but also to improve the consistency of image segmentation. Machine learning (ML)-based approaches have been employed with great success to automatically contour medical images.^{2,3} When algorithm training is performed using a sufficiently large and representative dataset, ML-based

approaches have been shown to generate contours that are effectively indistinguishable from manually drawn contours.⁴ However, automatic image segmentation techniques fail when the postulated assumptions inherent to the ML model are violated.^{5,6} For example, the high variability in size, location and appearance of tumors makes it challenging to build a representative training set for multiple tasks, limiting the performance of ML methods.⁶ Human input is thus still currently required for the segmentation of structures whose appearance and location exhibit large variability.

To overcome the drawbacks of fully automatic image segmentation techniques, semi-automatic methods can be employed. Such approaches can improve segmentation accuracy compared to fully automatic techniques by integrating user inputs to guide the segmentation process. Compared to fully manual segmentation, interactive techniques have been shown to improve repeatability and consistency across multiple observers.^{7–10}

User inputs can be used in a ML model as additional prior information to the system. This has been shown to improve contouring results.^{11–13} For example, a User can draw a contour on one image slice in a CT or MR dataset that can then be propagated through the remaining slices if the context between the contour on the delineated slice and the remaining image slices can be established.^{11,12} As an alternative approach to incorporate user annotations, a three-dimensional (3D) U-Net for volumetric segmentation has been proposed that learns from sparsely annotated volumetric images.¹⁴ In this work, a network is trained using only a few manually annotated image slices by using a weighted loss function and special data augmentation. Previous work on semi-automatic methods has focused on the segmentation of a predefined anatomical structure or set of structures. For such methods, networks are trained and optimized to detect the features relevant to the specific structure(s). This means, however, that they are likely to fail when applied to unseen structures.^{5,6}

In the present study, a deep learning training strategy is evaluated to enable 3D segmentation of an arbitrary anatomical structure using a manual segmentation on a single slice as prior knowledge. The ability of this method to capture the context between the manual segmentation and the image slice and apply it to adjacent image slices regardless of the structure of interest was investigated. This study shows that interactive contouring is possible using contextual deep learning, given access to adequate contextual inputs and a diverse training set. Most importantly, it demonstrates that the contextual deep learning model is not limited to structures included in the training set but extends to previously unseen structures.

2. MATERIALS AND METHODS

2.A. Contextual deep learning

In this study, a deep learning model is proposed that makes predictions based on the provision of relevant contextual information, rather than being limited to a specific structure defined by the training set. Such a model is referred to as

a contextual deep learning model in this work. The training strategy, particularly the diversity of structures in the training set, has a great impact on the models ability to predict previously unseen structures.

For the contextual deep learning model to learn from context, three different inputs were provided that contain the information of the image slice to be contoured and the contextual information. These inputs are illustrated in Fig. 1(a). The first input is the slice to be contoured, here referred to as the target image slice. A previously contoured image slice provides the other two inputs to the model: the image slice itself, as well as the contour information in the form of a binary mask. The relationship between the image slice and its corresponding binary mask allows the model to identify what to segment on the target image slice. The binary mask does not label which structure it is, such as whether the structure is, for example, a heart, a tumor or any other structure. This information was deliberately omitted, to generalize the model to its context. In Supplemental Material 3, different sets of inputs to the models were studied. This analysis showed that all three inputs are essential for the network to learn based on context.

2.B. Neural network architecture

In this work, a modified U-Net was used.¹⁵ Specifically, a residual recurrent U-Net with attention gates.^{16,17} The architecture is shown in Fig. 1(b). Here, the network input was expanded from one to three channels. The additional channels provide prior information to the network, as described above. Attention gates were incorporated to connect the features from the contracting path with the expanding path. The attention gates for soft self-attention highlight salient image regions implicitly.^{18,19} In the last layer, a sigmoid activation was applied to get the probability map.

The network was trained using pixelwise binary cross entropy loss.²⁰ The Adam optimizer²¹ was used with an initial learning rate of 1×10^{-5} and a batch size of 4. Data augmentation was applied on-the-fly during training using a random affine matrix transformations consisting of translation (± 10 mm), rotation (± 10 degrees) and scaling ($\pm 5\%$). Each model presented in this study was trained from randomly initialized weights. Therefore, model performance may vary with initialization. To mitigate the variability and achieve the best possible performance for each model, eight models were initialized for training. After 5 epochs, the worst 50% of the models are discarded. This was repeated each time the number of epochs doubles until one model remains. Then the remaining model was trained for 20 more epochs.

2.C. Data

This study included CT images and delineations of 19 different structures of interest from various openly available datasets. An overview of all the used structures is shown in Table I.

The 3D CT volumes and corresponding contouring data were randomly split into a training, validation and test set

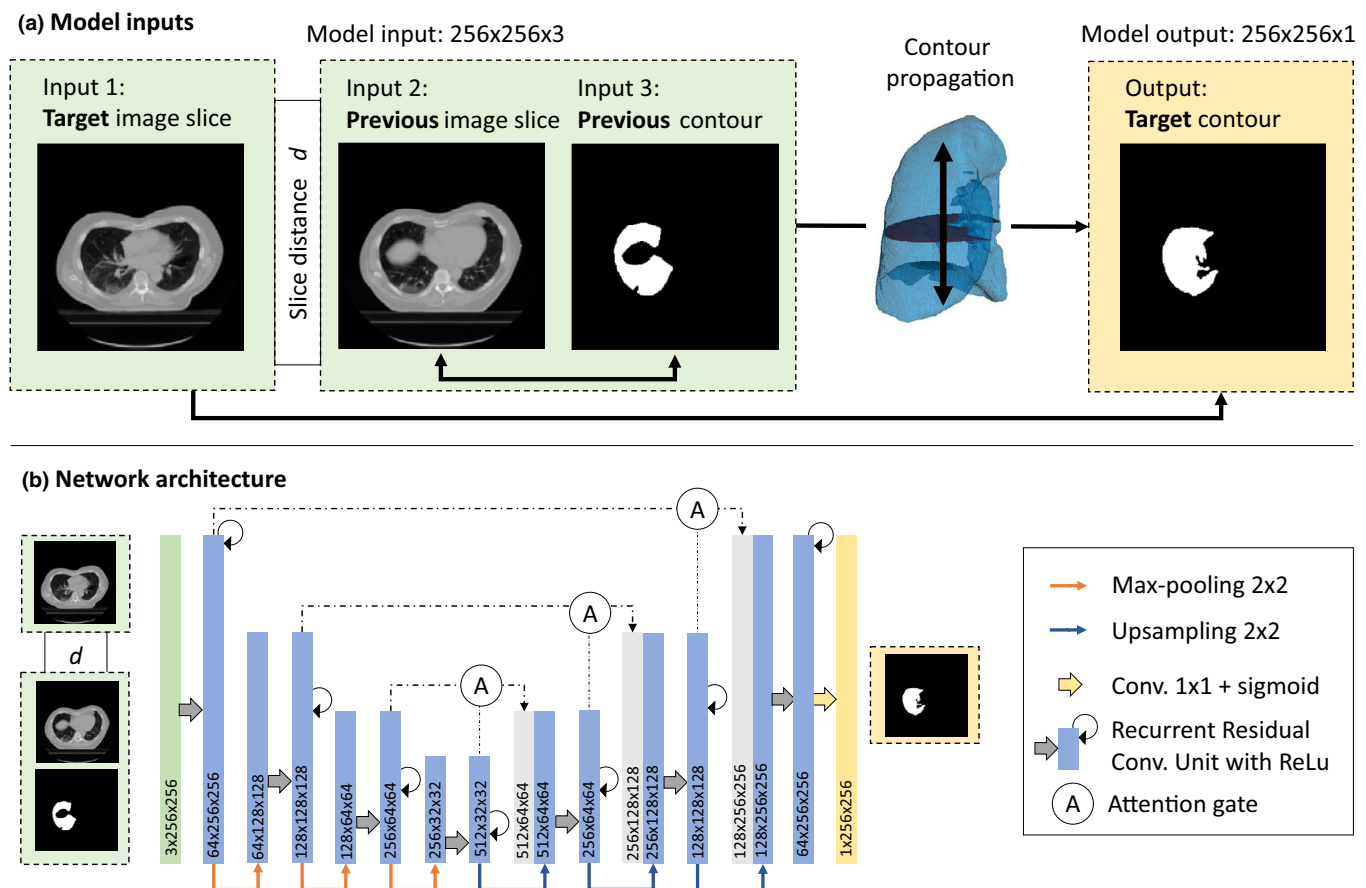


FIG. 1. (a) Inputs to the deep learning model. Target image slice and a contextual input which consists of an image slice and its corresponding binary mask at a slice distance d from the target image slice position. The predicted target contour corresponds to the target image slice (b) The network architecture is a Recurrent Residual U-Net variant using attention gates. The number of features F and layer dimension (width W and height H) is indicated in the network: $F \times W \times H$. [Color figure can be viewed at wileyonlinelibrary.com]

using a 60:20:20 split, respectively. In total, 2000 image slices per structure were randomly selected from the training data in order to limit the computational resources required and to create a balanced representation of each structure in the training set. Structures that were delineated on only very few patients or that extend over fewer than 10 slices were excluded from this study.

In the training set, the previously contoured and target slices were chosen carefully to support learning from context. The interslice distance between the contoured and target slice was particularly important. If the previously contoured image slice and the target image slice were adjacent image slices, there was insufficient change between the image slices for the model to learn the context and, in this case, the model learnt to copy the existing contour rather than generate a segmentation mask for the new slice. Target slices far away from the contoured slice are less relevant for the application of this model, as the target image slices corresponding too large interslice distances often lie outside of the structure to be segmented. The slice distance was sampled from a range of distances to balance between the two extremes and create a more diverse training set that was not biased towards a specific slice distance. It was important to sample the distance from a range of distances relative to the size of the structure

and not use one predefined set of distances for all structures. Thus, different structures with a variety of size (e.g., lungs, heart, tumor) were equally well represented. The previously contoured image slice must be chosen such that it lies within the structure, otherwise no contour exists from which context can be determined. However, the target image slice may lie outside the structure volume to allow the model to learn not only where to contour but also when to stop contouring.

2.D. Preprocessing

CT pixel intensities were clamped to the range from -1000 HU (air) to 2000 HU (cortical bone) to avoid artifacts in the high and low HU range. Intensities outside this range are typically irrelevant to segment biological tissue. The CT image slices of 512×512 pixels were rescaled to the network input size of 256×256 pixels. The rescaling was used to decrease computation time and reduce memory consumption.

2.E. Experimental details

A series of models was trained, for which the size of the training set was successively increased. Figure 2 shows the various models, together with the respective set of structures

TABLE I. Overview of number of slices used for training, validation, and testing for different anatomical structures, as well as the total number of contoured slices available in the respective dataset.

Set	Structure	Train.	Val.	Test.	Total	Ref.
A	Heart	1000	200	318	2096	[28]
	Heart	1000	200	404	3741	[29]
B	Esophagus	2000	200	802	5247	[28]
	Lung (left)	2000	200	805	5057	
	Lung (right)	2000	200	814	5305	
	Spinal cord	2000	200	1315	7907	
C	Lung tumor	2000	200	1219	10961	[29]
D	Pancreas	2000	200	987	8783	[31]
	Pancreas tumor	2000	200	273	2537	
E	Liver	2000	200	5052	27449	[31]
	Liver tumor	2000	200	2172	10972	
F	Neck (left)	2000	200	450	3136	[31]
	Neck (right)	2000	200	405	3099	
	Submandib. (right)	976	127	130	1233	
	Submandib. (left)	934	130	162	1266	
	Parotid (left)	1359	200	210	1806	
	Parotid (right)	1387	200	222	1832	
	Brain	2000	200	523	3454	
	Brain tumor	2000	200	260	2878	
G	Spleen	0	0	1363	1363	[31]

Note, that there are several other structures within the original datasets that were excluded due to their relatively small size or sparse annotations.

included in the training set. All models were evaluated on all structures. This permits the assessment of whether or not the ML model generalizes to unseen structures. The order in which the datasets are added to the training set was chosen arbitrarily. A combinatorial investigation would be prohibitively computationally expensive. A single GeForce GTX 1080 GPU was used for all experiments. The inference time is approximately 0.1 s per image slice.

2.F. Evaluation methods

2.F.1. Evaluation measures

The quality of the segmentation was quantified using the two-dimensional (2D) Dice similarity coefficient (DSC),²²

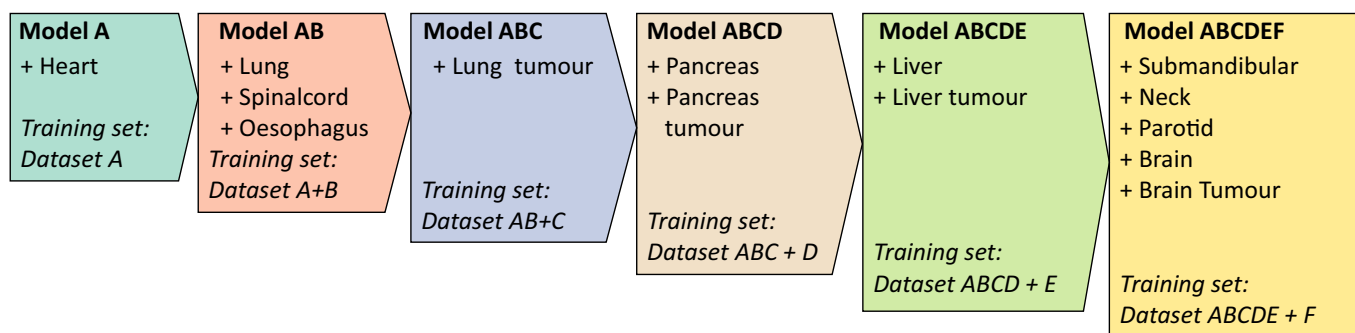


FIG. 2. Series of models trained on successively expanding training sets. Each new model uses the training set of the prior model and includes new structures previously excluded from the training set. [Color figure can be viewed at wileyonlinelibrary.com]

Hausdorff distance (HD),²³ and relative added path length (APL).²⁴

The DSC measures the overlap between two areas or two volumes, A and B . The DSC is defined as

$$\text{DSC}(A, B) = 2 \frac{A \cap B}{A \cup B} \quad (1)$$

The HD measures how far two subsets of a metric space are from each other. The HD $d_{\text{HD}}(A, B)$ is defined by

$$d_{\text{HD}}(A, B) = \max \left\{ \sup_{a \in A} \inf_{b \in B} d(a, b), \sup_{b \in B} \inf_{a \in A} d(a, b) \right\} \quad (2)$$

where d denotes the Euclidian distance.

The APL is the length of contour drawn when editing a segmentation. Because the absolute length of a contour varies between patients and different structures, the APL is reported relative to the ground truth contour length. A tolerance of 2 mm between predicted result and ground truth was used for APL. A low relative APL means that few edits are necessary, whereas at a relative APL of the full contour needs to be drawn by the user.²⁴

The DSC and HD are widely used to quantify the performance of segmentation tasks. However, these measures are not necessarily the best choice in a clinical context, where the contouring speed is more important. The time needed for an expert to review and adjust the predict contours shows better correlation with the APL than with DSC and HD.²⁴

To compute the performance measures, the masks that were generated were upsampled to the original 512×512 imaging resolution using bi-linear interpolation. It is important to evaluate the performance after upsampling because a clinician would review these contours at the original image resolution.

2.F.2. Contour prediction

In practice, it is expected that a user would select an image slice and delineate a structure on that slice. Then, the structure would be segmented by the model on the remaining image slices given the initial user input. Two different contour prediction methods were evaluated.

In the first approach, the initial input image slice (the central slice) was chosen as the contextual input for the

segmentation of all the remaining slices. This method is referred to as “direct prediction.” In the second approach, the expert contour at the central slice was used as the contextual input and the adjacent image slice as the target slice. The predicted target contour becomes the contextual input for the next slice. The process was repeated until all slices are contoured. This method is referred to as “iterative prediction.”

3. RESULTS

Results for all evaluation measures can be found in the Supplemental Materials, while the relative APL or DSC are reported here.

3.A. Impact of training set diversity on generalizability from context

Figure 3 shows the evolution of the model performance as new structures from an additional dataset are added. Here, results for iterative prediction are shown, averaging the slice-wise evaluation measures across all structures for each dataset.

Model A, with only the heart in the training set, showed the worst performance with a DSC of 0.86, 0.03, 0.24, 0.21, 0.40, 0.13, and 0.33 and a relative APL of 0.45, 0.96, 0.86, 0.88, 0.91, 0.95, and 0.85 for the test datasets A, B, C, D, E, F, G, respectively. The successive models included additional structures in the training set and showed a progressive overall improvement in performance compared to model A. The best performing model was the ABCDEF model with a DSC of 0.87, 0.77, 0.62, 0.49, 0.66, 0.69, and 0.80 and a relative APL of 0.44, 0.26, 0.60, 0.62, 0.56, 0.43, and 0.35 across the different datasets.

The boxplots in Fig. 4 show the relative APL for the heart, lung tumor, pancreatic tumor, liver, submandibular gland, and spleen. An equivalent figure for the rest of the structures and evaluation measures is provided in Supplemental Material 1. The most significant decrease in the relative APL for a given structure was observed when that structure was included in the training set of a model. However, observe that including additional structures in the training set also resulted in segmentation improvements to structures that were not included in the training set. For example, the relative APL for pancreas tumor, liver, submandibular gland and spleen decreased on adding the lung tumor in the training set (Model AB to Model ABC). The spleen was excluded from all models trained. Yet, the relative APL of the spleen decreases from 0.85 ± 0.22 for model A to 0.35 ± 0.21 for the final ABCDEF model.

3.B. Comparison of contour prediction approaches

In Fig. 5, the performance of model ABCDEF for direct and iterative prediction is compared at different interslice distances from the central slice. Figures for the remaining structures and evaluation measures can be found in Supplemental Material 2.

Predictions closer to the central slice had a lower relative APL, whereas the relative APL increased towards the superior and inferior most slice of each structure. The structures included in the training set [Figs. 5(a)–5(e)] showed similar performance between the direct and iterative prediction method. In contrast, the spleen [Fig. 5(f)] had a lower relative APL, when using iterative prediction as compared to direct prediction.

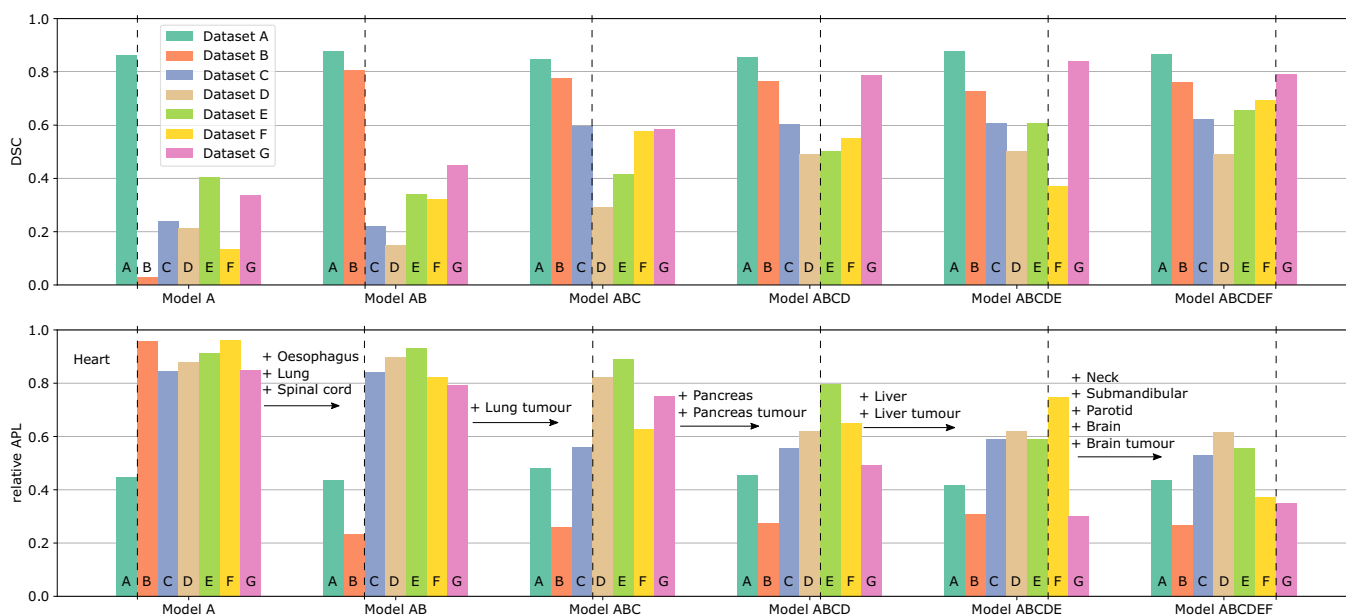


FIG. 3. Performance of models with successively increasing training set on different test datasets using iterative prediction. Datasets on the left of the dashed line indicate that this dataset was included in the training set of the respective model, whereas datasets to the right are excluded from the training set. [Color figure can be viewed at wileyonlinelibrary.com]

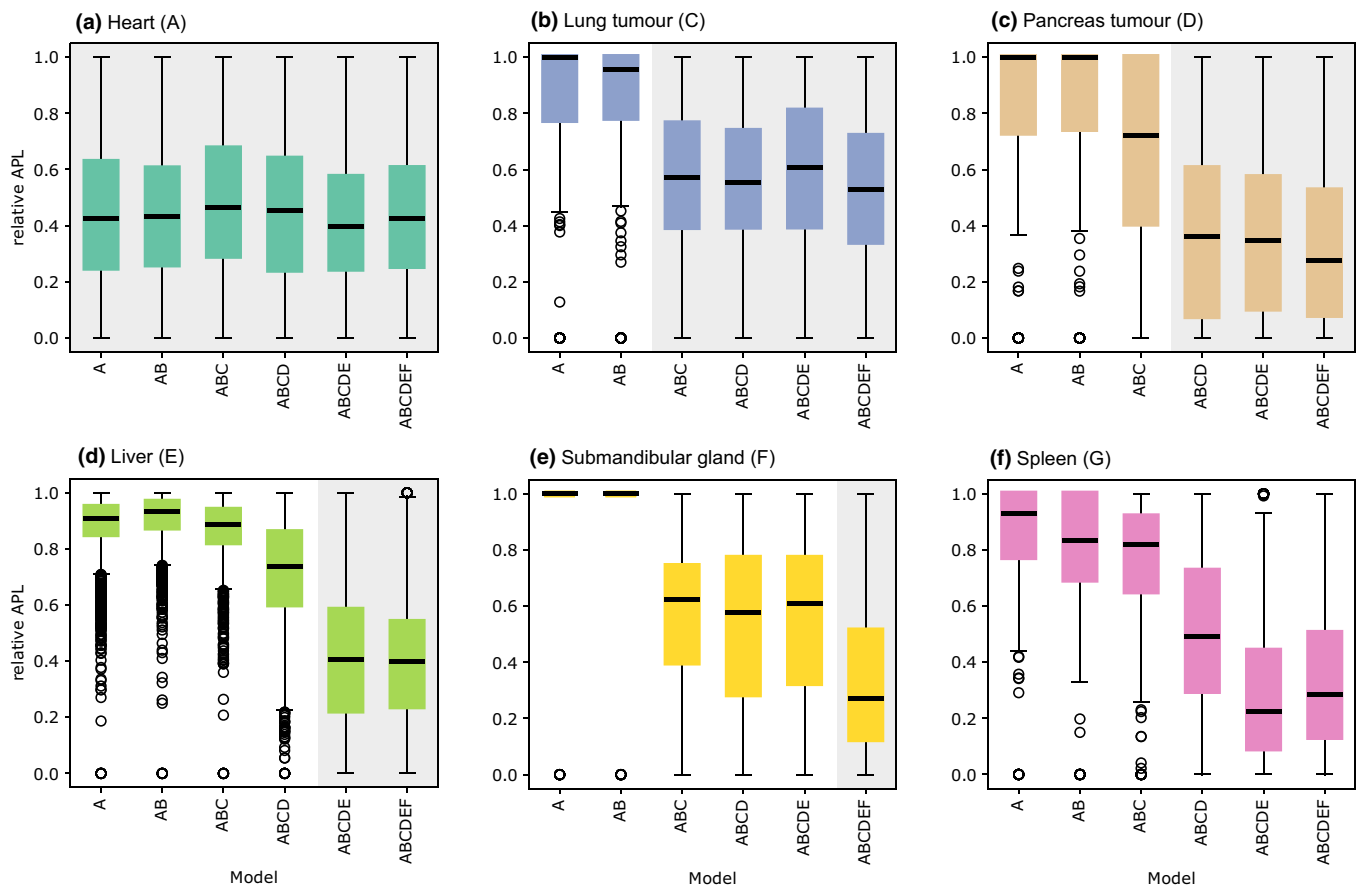


FIG. 4. Boxplots of the performance of different models on selected structures: (a) heart, (b) lung tumor, (c) pancreatic tumor, (d) liver, (e) submandibular gland, and (f) spleen. The gray-shaded background indicates the models that include the structure in their respective training set. [Color figure can be viewed at wileyonlinelibrary.com]

4. DISCUSSION

4.A. Impact of training set diversity on generalizability from context

The addition of training examples of structures would be expected to improve the performance of a model in delineating those structures. This can be observed in Figs. 3 and 4. However, with subsequent addition of structures to the training set, not only does the performance improve for the structures newly included in the training set, but also for the structures excluded from the training set, as seen in Fig. 4. For example, when adding lung tumors to the training set (going from model AB to model ABC), the performance improves for lung tumors, pancreatic tumors, liver, submandibular gland, and spleen. This improved performance cannot originate from learning any structure-specific features within the training set, but must be attributed to learning contextual features and making predictions based on the context from the provided input contour and image slice. For example, Model ABCDEF achieves a relative APL of 0.35 for the spleen, despite not having been trained on this structure before. This suggests that user interaction and thus time spent contouring can be decreased for de-novo contouring, when using such a model. Additionally, it is shown in Supplemental Material 4 that most structures can achieve a high

segmentation accuracy, even if that structure was excluded from training set and the training set consisted of all remaining structures. In contrast, a model trained on only a single structure fails to generalize as shown in Supplemental Material 5.

The ability to generalize to unseen structures distinguishes this approach from previous work, in which previously contoured image slices are used to assist in segmentation.^{11,12}

Certain structures appear to have a larger impact on the model's ability to generalize to unseen structures. For example, four new structures were added to the training set between model A and model AB, but the improvement in segmentation performance is almost exclusively seen in the newly added structures. In contrast, between model AB and ABC, the segmentation performance improved for all structures, although only lung tumors were added to the training set. The improvement in segmentation of unseen structures and thus, the ability to generalize, could potentially be attributable to the high diversity in the tumor appearance.

At the same time, the structures that have already been included in previous models show further improvement upon expanding the training set by new structures, see Figs. 4(b) and 4(c). This suggests that the quality of segmentation of an individual structure benefits from an increased diversity in the training set. This might be attributed to a refinement of

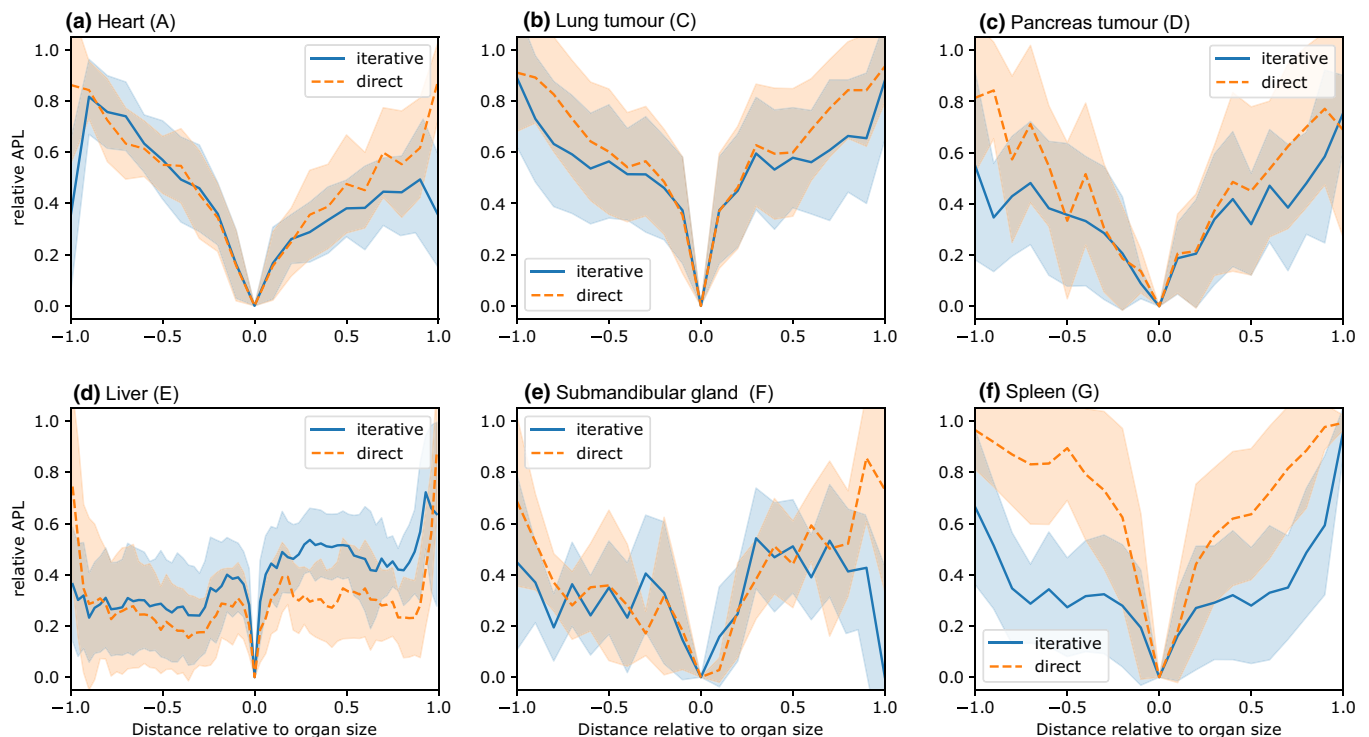


FIG. 5. Comparison of direct (orange-dashed line) and iterative (blue line) prediction using the model ABCDEF on different structures: (a) heart, (b) lung tumor, (c) pancreatic tumor, (d) liver, (e) submandibular gland, and (f) spleen. The shaded area indicates the standard deviation of the APL for different test samples. [Color figure can be viewed at wileyonlinelibrary.com]

the contextual features used in the segmentation over the shape-specific features of each structure. In this work, it has not been tested if the incremental changes between subsequent models are statistically significant due to the small and variable size of the test dataset for the different structures. This needs to be addressed in future work.

The interactive contouring approach proposed here achieves similar or higher DSC than fully automatic segmentation approaches reported for the heart,²⁵ lung,²⁵ spinal cord,²⁵ lung tumor²⁶ and pancreas tumor,²⁷ but lower performance for the esophagus,²⁵ and spleen.²⁷ This comparison is limited to studies in which fully automatic solutions were reported for the same dataset as used in this report. It is noted that the current approach benefits from input of an initial manual contour.

The contextual deep learning approach used here did not achieve a DSC comparable to the best performing fully-automatic models for unseen structures such as the spleen.²⁷ Note, however, that those fully automatic models are trained on a dataset of images that include the spleen, whereas the model proposed in this work, was trained on images that did not include the spleen. Therefore, this indicates that the current approach can be applied to unseen structures and may decrease delineation times for such structures for which no automatic solutions exist. Furthermore, contextual deep learning achieved high DSC values for pancreas and lung tumors compared to fully automated solutions, highlighting the benefits of this approach for segmentation of highly variable structures.

4.B. Comparison of contour prediction approaches

While direct prediction shows similar performance as iterative prediction for most structures, iterative prediction performed significantly better for an anatomical structure that was omitted from the training set, that is, the spleen.

In the direct prediction approach, the interslice information is not used and the slice distance varies. This may increase the error of the direct prediction approach. In the iterative approach, an initial error or error in a subsequent prediction can accumulate throughout the CT scan. The improved performance of iterative over direct prediction suggests that the cumulative segmentation error through iterative prediction is smaller than the error of direct prediction when compared over the same interslice distance. For direct prediction far from the initial input slice, there may be large differences between the target and input image slices. If these differences are too large, the model will fail to relate the two slices based on context and thus hinder accurate segmentation. In practice, if the error of either prediction approach becomes too large at a certain distance from the initial contoured slice, a user can adjust the predicted contour. These predicted contours in return can then be used to make a prediction.

The difference between iterative and direct prediction is less marked for structures included in the training set. This may be because the model still learns some features specific to the structures in the training set. These learnt features may account for the improved performance of iterative prediction.

Future work needs to study the impact of the proposed semi-automatic contouring tool on interobserver variability on the final segmentation. Furthermore, it is interesting to study the performance of the segmentation when compared to a reference segmentation from the same user or from different ones.

5. CONCLUSION

This work investigated the use of contextual information by a deep learning model using three input channels. The first channel was used for the image to be segmented. The second and third were used to provide context from an image slice and corresponding manual delineation of the structure to be contoured. The experiments demonstrated that a model trained using multiple distinct structures can accurately make predictions for structures within the training set and achieve performance similar or higher than fully automatic approaches, but more importantly it could generalize and predict structures excluded from the training set based on the provided context alone.

Two different prediction procedures (direct and iterative prediction) were introduced and compared. For unseen structures it was found that the error propagated using iterative prediction remains small when compared to the direct prediction approach.

The proposed contextual deep learning approach allows prior context to be incorporated into deep learning contouring to facilitate interactive contouring of CT imaging. Such an approach may enable faster de-novo contouring in clinical practice, where manual contouring is required and there is insufficient training data to train a specific model.

ACKNOWLEDGMENTS

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 766276. KAV acknowledges funding support from CRUK (A15935) and the CRUK Radnet Centre (A28736).

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in The Cancer Imaging Archive at <http://doi.org/10.7937/K9/TCIA.2017.3r3fvz08>,²⁸ <https://doi.org/10.7937/K9/TCIA.2015.PF0M9REI>,²⁹ <https://doi.org/10.7937/tcia.2019.8kap372n>,³⁰ and in the Medical Segmentation Decathlon at <http://medicaldecathlon.com/>.³¹

CONFLICT OF INTEREST

The authors have no conflicts to disclose

^{a)} Author to whom correspondence should be addressed. Electronic mail: michael.trimpl@mirada-medical.com.

REFERENCES

1. Ramkumar A, Dolz J, Kirisli HA, et al. User interaction in semi-automatic segmentation of organs at risk: a case study in radiotherapy. *J Digit Imaging* 2016;29:264-277.
2. Lustberg T, van Soest J, Gooding M, et al. Clinical evaluation of atlas and deep learning based automatic contouring for lung cancer. *Radiother Oncol* 2018;126:312-317.
3. Jarrett D, Stride E, Vallis K, Gooding MJ. Applications and limitations of machine learning in radiation oncology. *Br J Radiol* 2019;92:20190001.
4. Gooding M, Smith A, Peressutti D, et al. PV-0531: Multi-centre evaluation of atlas-based and deep learning contouring using a modified Turing Test. *Radiother Oncol* 2018;127:S282-S283.
5. Perone CS, Cohen-Adad J. Promises and limitations of deep learning for medical image segmentation. *J Med Art Int* 2019;2:1.
6. Lundervold AS, Lundervold A. An overview of deep learning in medical imaging focusing on MRI. *Zeitschrift für Medizinische Physik* 2019;29:102-127.
7. Olabariaga SD, Smeulders AWM. Interaction in the segmentation of medical images: A survey. *Med Image Anal* 2001;5:127-142.
8. Wang G, Li W, Zuluaga MA, et al. Interactive medical image segmentation using deep learning with image-specific fine tuning. *IEEE Trans Med Imaging* 2018;37:1562-1573.
9. Wang G, Zuluaga MA, Li W, et al. DeepIGeoS: A deep interactive geodesic framework for medical image segmentation. *IEEE Trans Pattern Anal Mach Intell* 2019;41:1559-1572.
10. Sakinis T, Milletari F, Roth H, et al. Interactive segmentation of medical images through fully convolutional neural networks. 2019, ArXiv. abs/1903.0.
11. Léger J, Brion E, Javaid U, Lee J, De Vleeschouwer C, Macq B. Contour Propagation in CT scans with Convolutional Neural Networks, in International Conference on Advanced Concepts for Intelligent Vision Systems ACIVS 2018: Advanced Concepts for Intelligent Vision Systems; 2018:380-391.
12. Zheng Q, Delingette H, Duchateau N, Ayache N. 3D consistent robust segmentation of cardiac images by deep learning with spatial propagation. *IEEE Trans Med Imaging* 2018;37:2137-2148.
13. Novikov A, Major D, Wimmer M, Lenis D, Bühler K, Deep sequential segmentation of organs in volumetric medical scans. *IEEE Trans Med Imaging* 2018;38:1207-1215.
14. Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation, in MICCAI, 2016:424-432.
15. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. *Med Image Comput Comput-Ass Int* 2015;9351:234-241.
16. Alom MZ, Hasan M, Yakopcic C, Taha TM, Asari VK. Recurrent Residual Convolutional Neural Network based on U-Net (R2U-Net) for Medical Image Segmentation. 2018, ArXiv. abs/1802.0.
17. Alom MZ, Yakopcic C, Taha TM, Asari V. Nuclei Segmentation with Recurrent Residual Convolutional Neural Networks based U-Net (R2U-Net), in IEEE National Aerospace and Electronics Conference; 2018:228-233.
18. Oktay O, Schlemper J, Le Folgoc L, et al. Attention U-Net: Learning Where to Look for the Pancreas, in Medical Imaging with Deep Learning. Amsterdam, 2018.
19. Schlemper J, Oktay O, Schaap M, Heinrich M, Kainz B, Glocker B, Rueckert D. Attention gated networks: Learning to leverage salient regions in medical images. *Med Image Anal* 2019;53:197-207.
20. Janocha K, Czarnecki WM. On loss functions for deep neural networks in classification. *Schedae Informaticae* 2016;25:49-59.
21. Kingma DP, Lei Ba J, ADAM: A method for stochastic optimization. ICLR. 2015.
22. Dice LR, Dice LR. Measures of the amount of ecologic association between species. *Ecology* 1945;26:297-302.
23. Huttenlocher DP, Klanderman GA, Rucklidge WJ. Comparing images using the Hausdorff distance. *IEEE Trans Pattern Anal Mach Int* 1993;15:850-863.
24. Vaassen F, Hazelaar C, Vaniqui A, et al. Evaluation of measures for assessing time-saving of automatic organ-at risk segmentation in radiotherapy. *Phys Imaging Radiat Oncol* 2020;13:1-6.

25. Feng X, Qing K, Tustison NJ, Meyer CH, Chen Q. Deep convolutional neural network for segmentation of thoracic organs at risk using cropped 3D images. *Med Phys* 2019;46:2169-2180.
26. Hossain S, Najeeb S, Shahriyar A, Abdullah ZR, Ariful Haque M. A Pipeline for Lung Tumor Detection and Segmentation from CT Scans Using Dilated Convolutional Neural Networks, in IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, Brighton; 2019:1348-1352.
27. Isensee F, Petersen J, Klein A, et al. nnU-Net: Self-adapting Framework for U-Net-Based Medical Image Segmentation, in Medical Segmentation Decathlon, Challenge 2018, 2018.
28. Yang J, Sharp G, Veeraraghavan H, et al. Data from Lung CT Segmentation Challenge; 2017.
29. Aerts HJWL, Wee L, Velazquez RE, et al. Data From NSCLC-Radiomics Lung1, Technical report. The Cancer Imaging Archive. 2019.
30. Wee L, Dekker A. Data from Head-Neck-Radiomics-HN1. Technical report: The Cancer Imaging Archive; 2019.
31. Simpson AL. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. Technical report. 2019.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Supplemental Material 1 and 2 list the experimental results for individual structures investigated along with additional evaluation metrics.

Supplemental Material 3 investigates why all 3 input channels to the contextual deep learning model are essential.

Supplemental Material 4 demonstrates that the contextual deep learning model can generalize for various unseen structures.

Supplemental Material 5 illustrates that a diverse training set is necessary to achieve a model that can generalize to unseen structures.