Semantic segmentation of computed tomography for radiotherapy with deep learning: Compensating insufficient annotation quality using contour augmentation

Umair Javaid[1,2,*], Damien Dasnoy[2], and John A. Lee[1,2]

[1]IREC/MIRO, UCLouvain, Brussels, Belgium
[2]ICTEAM, UCLouvain, Louvain-la-Neuve, Belgium

ABSTRACT

In radiotherapy treatment planning, manual annotation of organs-at-risk and target volumes is a difficult and time-consuming task, prone to intra and inter-observer variabilities. Deep learning networks (DLNs) are gaining worldwide attention to automate such annotative tasks because of their ability to capture data hierarchy. However, for better performance DLNs require large number of data samples whereas annotated medical data is scarce. To remedy this, data augmentation is used to increase the training data for DLNs that enables robust learning by incorporating spatial/translational invariance into the training phase. Importantly, performance of DLNs is highly dependent on the ground truth (GT) quality: if manual annotation is not accurate enough, the network cannot learn better than the annotated example. This highlights the need to compensate for possibly insufficient GT quality using augmentation, i.e., by providing more GTs per image, in order to improve performance of DLNs. In this work, small random alterations were applied to GT and each altered GT was considered as an additional annotation. **Contour augmentation** was used to train a dilated U-Net in *multiple GTs per image* setting, which was tested on a pelvic CT dataset acquired from 67 patients to segment bladder and rectum in a multi-class segmentation setting. By using contour augmentation (coupled with data augmentation), the network learnt better than with data augmentation only, as it was able to correct slightly offset contours in GT. The segmentation results produced were quantified using spatial overlap, distance-based and probabilistic measures. The Dice score for bladder and rectum are reported as 0.88 ± 0.19 and 0.89 ± 0.04, whereas the average symmetric surface distance are 0.22 ± 0.09 mm and 0.09 ± 0.05 mm, respectively.

Keywords: Radiotherapy, segmentation, contour augmentation, CT

1. INTRODUCTION

Radiation oncology treats cancer by irradiating tumors (target volumes, TVs) with either photon or proton beams. Cancerous cells are slightly more sensitive to irradiation than healthy ones and therefore daily targeted irradiation at low dose during several weeks maximizes the probability of local tumor control, while it minimizes undesired side effects. Medical images are used as a map to localize the TVs to be hit and the organs-at-risk (OARs) to be avoided. Computed tomography (CT) is an imaging modality obtained by aiming a narrow beam of x-rays at a patient using a motorized x-ray source to generate cross-sectional images/slices of the body. It allows identification and localization of basic anatomical structures, soft tissues, and possible TVs. Annotation of OARs and TVs on CT scans is a key step in radiotherapy treatment planning with a trade-off problem between delivering maximum dosage to the TVs and minimum dosage to OARs.

A huge number of CT scans must be analyzed by radiologists before any radiotherapy treatment planning. Therefore, manual and dense annotation of abundant CT scans is an arduous and time intensive task subject to intra and inter-observer variabilities, thereby highlighting the need for automated and accurate annotations for effective treatment planning. To address this problem, automated image segmentation using deep learning (DL) is currently being investigated. DL networks (DLNs) have been shown to achieve state-of-the-art performance

*umair.javaid@uclouvain.be

for classification, detection, and segmentation tasks.[1–3] However, in a supervised setting, performance of DLNs depends significantly on the quality of given ground truth (GT). For example in radiotherapy, annotating TVs is of prime concern and proper care is taken while drawing them. OARs are not always a priority and sometimes annotated with less care. When such rough annotation is used to train the DLNs, it limits their predictive capabilities in terms of localization of organ boundaries. The network learning is thus sub-optimal as it cannot improve over the unique, possibly offset contour that it is provided with and thus, the resulting segmentation map will be ambiguous in terms of localization of anatomical structures.

Given the data-demanding nature of DLNs, data augmentation is a common workaround to artificially add more data to the training sets. It enhances the training set by generating more instances from the available data which is representative of a specific task. Moreover, it also prevents overfitting. Augmentation can be considered as noise added to the data where the noise enables the network to learn more general features and useful invariances, like translations and rotations in images, instead of overfitting the training data. Therefore, data augmentation is useful for effective learning by DLNs where deformations are applied to the given input pair, i.e., image X and annotation Y collectively. However, learning by DLNs is still based on the given Y (GT) which can possibly be ill-defined.

In this study, we present a novel framework that can take into account the intra and inter-observer variabilities in terms of multi-organ annotations. We propose deformation of the annotation only, aiming for a better match between deformed annotation and salient features in the image. We augment the annotations defined in GT and produce several annotation variants or hypotheses per image, thus generating multiple GTs from the manually defined GT. Providing multiple GTs per image as opposed to a single GT, the DLN (in our case a *dilated* U-Net[4]) can take into account uncertainty (*intra-observer variability*) in GT. By incorporating multiple GT variations in its training phase, the network not only learns from one GT but from a sampled distribution of GTs. This can enhance its predictive capabilities especially if some sampled GTs better correlate with features in the input image than the original GT.

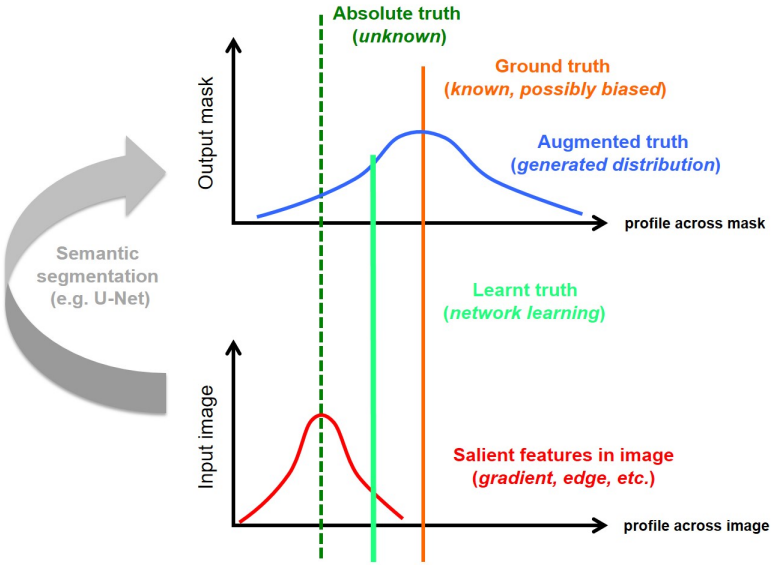

Figure 1. Principle of contour augmentation in a supervised setting, typically in an image-to-image task, like semantic segmentation. It is assumed that the absolute truth might be unknow in the annotation (output image or mask), with only a possibly biased ground truth that is available. Resampling the ground truth cannot mitigate a possible bias per se, unless some new samples better correlate with salient features in the input image, like gradients, edges, etc.

The motivation of contour augmentation is simply to learn an optimal distribution from the given sampled distribution of GTs as shown in Fig. 1. Given an input image, there exists a segmentation that is perfectly defined and accurately represents the anatomy. We term such segmentation as *absolute truth*, which may be unknown, due to inaccuracy in the annotation task. What we normally have at hand instead is the *ground*

truth, which is thus possibly slightly biased. Using contour augmentation, we sample the given *ground truth* to generate a distribution that is termed as *augmented truth*. This is used to train the DLN with the goal of learning an optimal distribution closer to the *absolute truth*. Thus, contour augmentation can equip the DLN with the flexibility to compensate for the induced bias in *ground truth*.

To the best of our knowledge, this is the first attempt to study DLN learning in a multiple GT per image setting. At the time of writing this work, probabilistic U-Net[5] (*work in progress*) was made available. The motivation is to learn a conditional density model over segmentations, conditioned on the image. The authors use probabilistic modeling for U-Net using a conditional variational auto-encoder to generate multiple segmentation hypotheses. They encode the possible segmentation variants in a low-dimensional latent space and take a random sample from this space to train the U-Net and produce a segmentation map. All the training is done in an iterative fashion. Unlike their approach, our proposed framework is simple and it is trained over all the GT variants simultaneously. Thus, we study the DLN learning given all annotation variants (*many vs single* GT training).

The objective of this work was to model the intra and inter-observer variabilities and study the behavior of our network with several GTs per image. Secondly, to observe if contour augmentation increases the network performance in terms of accurate localization of organ boundaries and good generalization of the DLN learning (given possibly ill-defined GT). We study the augmentation variants by training instances of the network with no augmentation, data augmentation, contour augmentation, and contour plus data augmentation.

The remainder of the paper is organized as follows: Section 2 outlines the dataset used and pre-processing steps. Section 3 introduces the different methods used to train and evaluate the networks. Results are given in Section 4 followed by discussion in section 5. Finally, Section 6 presents conclusion and future perspectives to this paper.

2. IMAGE DATA AND PRE-PROCESSING

2.1 Dataset

Data consists of pelvic CT scans acquired from 67 patients for another study and used here retrospectively. Each patient was referred for curative intent External Beam Radiation Therapy (EBRT) for prostate cancer at University Hospital UCLouvain Namur, Belgium. In the analyzed images, the OARs were defined according to the institutional guidelines and manually annotated by a single experienced radiation oncologist. The pelvic CT images and their corresponding OAR annotations, in particular bladder and rectum, were used for this study. All scans were acquired on an On-board imaging system, OBIWS (Varian Medical Systems, Palo Alto, California, USA).

2.2 Data Pre-processing

To obtain homogeneity across all patients' data, axial slices were selected from the 3D volume followed by geometric resampling in terms of uniform pixel spacing (0.94 mm) and slice thickness (2.5 mm). Image intensity values of all scans were truncated to the range of $[-1000, 3000]$ HU (Hounsfield Unit) to remove possible artifacts in the high HU values. Finally, slices were downsampled from 512×512 to 256×256 spatial resolution and rescaled between 0 and 1 with 8 bits per pixel.

3. METHODS

This paper aims to exploit the discriminative potential of DLNs for multi-organ segmentation subject to multiple GTs per image. We first briefly explain our network, *dilated* U-Net, and then present an overview of the proposed augmentation technique with different evaluation metrics used to assess network performance.

3.1 Dilated U-Net

U-Net[6] is a multi-scale convolutional network that consists of a *contracting* and an *expanding* path. The *contracting* phase extracts features from data and each step of the *contracting* path consists of two convolutions. The convolutions have 3×3 kernel and each convolution is followed by ReLU activation along with 2×2 max-pooling and strides of two in each direction of the spatial domain. In contrast, the *expanding* path upsamples the input using transposed convolutions, which is concatenated with features of the contracting path by *copy connections* (shown in yellow in Fig. 2). The upsampling layers preserve the maximum activation locations that were determined during the max-pooling step and re-use those locations to place the maximum activations back to their original spatial positions.

Fig. 2 shows the proposed U-Net architecture. Images of the size 256×256 were used as input whereas the network produces a segmentation mask as output. The output mask is a pixel-by-pixel one in which each pixel either belongs to one of the organs or the background. For this study, we had three classes: two organs, bladder and rectum, plus the background.

Dilated U-Net[4] is a variant of the standard U-Net in which standard convolutions are replaced with dilated ones. Dilated convolutions use sparse convolution kernel to enlarge the filters field-of-view in order to incorporate a larger context for dense feature extraction. This is done by deriving a larger but also sparser filter from the original filter and dilating it by inserting zeros to fetch context from farther away.[7] For a convolutional $k \times k$ size kernel, the size of the resulting dilated filter is $k_d \times k_d$, where $k_d = k + (k-1)(r-1)$ and r is the dilation rate. The dilation rate introduces spacing between the kernel values. Compared to normal convolution, dilated convolution enables network operation on a coarser scale.

Precise organ delineation requires knowledge of the spatial relationship between different organs. In the innermost part of U-Net, receptive field of the neurons is limited to a small subset of the image, thus limiting the information about relative organ positions. We therefore applied dilation (rate $r = 2$) to every second convolution before the max-pooling operation in the *contracting* path. Moreover, in the innermost bottleneck, we added a dilation block where we applied dilated convolutions of rates 2, 4, 8, and 16 to provide additional, broader context to the neurons.

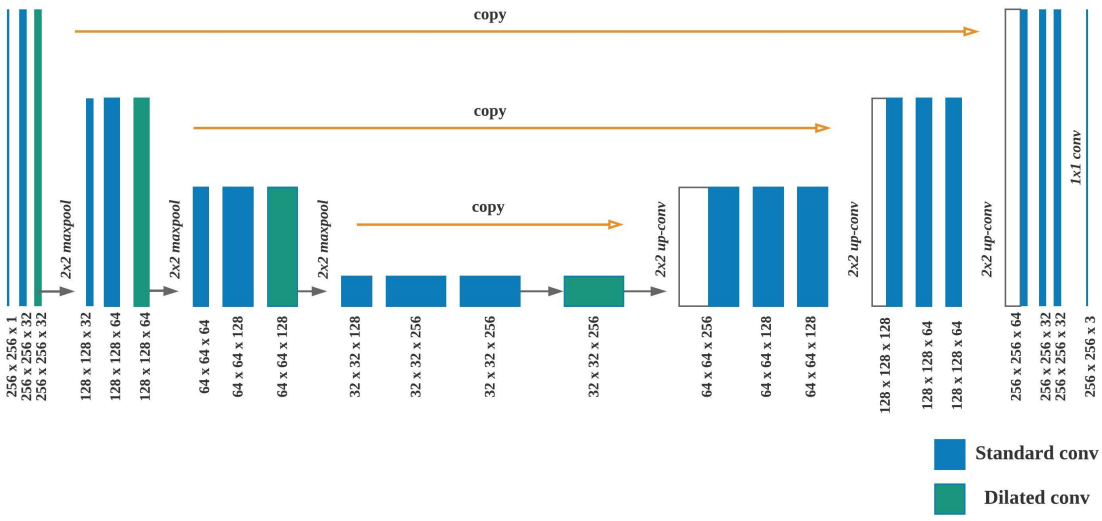

Figure 2. All convolutons are of 3×3 kernel size with ReLU as the activation function except the last layer that is 1×1 and has *softmax* activation where the output prediction involves three different classes.

3.2 Data Augmentation

Data augmentation techniques such as rotation, scaling, and translation were applied to the training images and masks followed by random local elastic deformations. This resulted in local stretch and compression in

images and masks.[8] During training, data augmentation was applied online (*augmenting every data sample stochastically*) which provided the network with new transformed examples, resulting in robust learning per epoch and potentially preventing overfitting. The new transformed examples are different versions of the same image. No flipping and/or mirroring was applied for practical reasons. Rotation angle of 10° with a random shift of 0.1 in total height and width range was used. Also, images were zoomed with a factor of 0.01 followed by elastic deformation.

3.3 Contour Augmentation

Ill-defined or rough annotation affects the predictive capabilities of DLNs. To bridge this limitation, we proposed contour augmentation to compensate for the insufficient annotation quality. Our contour augmentation block has three parameters, namely, the desired number of altered GTs N, alteration in millimeters m, and value of smoothing factor s. Given a single reference GT, GT_{R}, we applied random alteration within m ($m = 4$ mm) to its spatial dimensions to get altered GTs, GT_i where $i \in \{1, ..., N\}$ and $N = 5$. Gaussian smoothing (standard deviation for Gaussian kernel, $s = 2$) was performed to smoothen the GT_i to be qualitatively consistent with the original reference GT, GT_R. After smoothing, the GT_i were stored as additional GTs, augmenting our label set Y as $Y = \{GT_R, GT_1, ..., GT_i\}$. We checked the consistency of GT_i against reference GT_R by computing the standard deviation σ of pixel intensity distribution in the image under each GT and considered the GT_i as a good GT only when it was within the range of 2σ of GT_R. Several GTs per organ acquired by contour augmentation are given in Fig 3. For illustrative purposes, only 3 altered GTs along with the original one are shown.

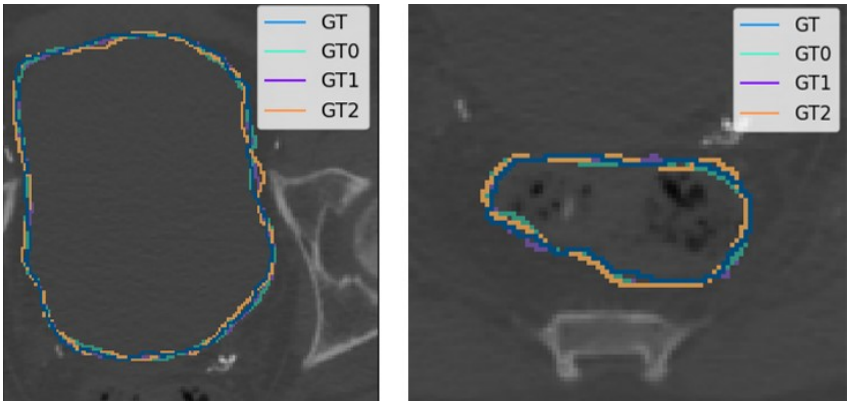

Figure 3. Several GTs per organ obtained by contour augmentation. *Left image* shows bladder where *right image* shows rectum. GT is the manually annotated GT whereas GT0, GT1 and GT2 correspond to altered GTs.

3.4 Quantitative Analysis and Evaluation

We use six metrics (three overlap, two contour-to-contour distance-based, and one probabilistic measure) to assess the similarity between predicted segmentation per organ and the manual organ segmentation performed by the clinical expert GT_R. The **Dice similarity coefficient** (DSC) measures the spatial overlap between two binary segmentation masks, seen as intersecting sets, typically the ground truth and a predicted segmentation. For two binary masks A and B, the DSC is the ratio of their intersection cardinality over their average cardinality given as:

$$\mathrm{DSC(A, B)} = \frac{2|A \cap B|}{|A| + |B|} \ , \tag{1}$$

where $|A|$ is the cardinality of set A that counts the number of nonzero pixels in mask A. In the context of binary classification, ***recall/sensitivity*** is a measure of how good a classifier is at detecting the positives whereas ***precision*** is a measure of how many of the positively classified samples of data are actually positive. They can

be formulated as:

$$\text{recall/sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \ , \tag{2}$$

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \ , \tag{3}$$

where TP, FP, and FN stand for true positive, false positive, and false negative.

The **95^{th} percentile Hausdorff distance**, HD_{95} is a measure of maximal distance (within the 95^{th} percentile, P_{95}) from a point in the first segmentation (*predicted*) to a nearest point in reference segmentation GT_R. Mathematically, it is given as:

$$\text{HD}_{95}(A, B) = \max\{P_{95} \min_{i \in A} \min_{j \in B} d(i, j), P_{95} \min_{j \in B} \min_{i \in A} d(i, j)\}, \tag{4}$$

where $d(i, j)$ is the Euclidean distance between the centers of nonzero pixels i and j in A and B, respectively. HD_{95} is used to measure segmentation accuracy and is expressed in millimeters (mm).

The **average symmetric surface distance** (ASSD)[9] is a distance-based measure used in the MICCAI 2007 scoring system. It gives the average of all distances from points on the boundary of predicted segmentation to the boundary of reference segmentation GT_R by computing the average symmetric surface distance between the surfaces, i.e.,

$$\text{ASSD}(A, B) = \frac{1}{|S(A)| + |S(B)|} \left(\sum_{p_A \in S(A)} d(p_A, S(B)) + \sum_{p_B \in S(B)} d(p_B, S(A)) \right) \ , \tag{5}$$

where $d(p_A, S(B))$ is the Euclidean distance from p_A, the points in surface A, to surface $S(B)$, and vice versa. ASSD is expressed in millimeters (mm).

Cohen's Kappa[10] is a probabilistic measure between observations. It indicates inter-annotator agreement by comparing an observed accuracy with an expected accuracy (*random chance*). Its value ranges from -1 (complete disagreeement) to 1 (full agreement). A null value means partial agreement by chance (*randomness*). It is given as:

$$\begin{gathered} P_o = \frac{1}{N} \sum_{j=1}^{k} f_{jj} \ , \\ r_i = \sum_{j=1}^{k} f_{ij}, \forall i \ , \text{ and } c_j = \sum_{i=1}^{k} f_{ij}, \forall j \ , \\ P_e = \frac{1}{N^2} \sum_{i=1}^{k} r_i c_i \ , \end{gathered} \tag{6}$$

where P_o is the observed proportional agreement, N is the number of cases, k is the number of categories in which a case can be rated, f_{ij} is the number of agreements for category i and j. r_i and c_j are the row and column totals for category i and j, and P_e is the expected proportion of agreement. The final measure of agreement, κ, is then given as

$$\kappa = \frac{P_o - P_e}{1 - P_e} \tag{7}$$

For a perfect segmentation, overlap measures should be 1 whereas the distance-based measures should be 0.

3.5 Network Training

A total of 1163 pairs of images and masks were considered as input. A 20% validation split was used on the input images and masks, resulting in 932 training and 231 validation pairs, respectively. To study the generalization of network learning, we evaluate it on several new patients. Standard pixel-wise unweighted cross-entropy was used as the loss function to train the network. It is given as:

$$L(y, \hat{y}) = -\sum_{nc} y_{nc} \log \hat{y}_{nc}, \tag{8}$$

where n refers to pixels, c represents the classes, and y_{nc} is the ground truth while $\hat{y}_{nc}$ is the predicted annotation.

Networks were trained from scratch for all the experiments reported in this paper. We performed *k-fold* cross-validation ($k = 5$) to tune the hyper-parameters. The split between train and test set for each fold of the *k-fold* cross-validation is shown in Fig. 4. The networks were trained using the Adam optimizer with learning rate LR = 0.001 and batch size equal to 32 for 3.96k iterations. The LR was monitored for 5 epochs and was reduced by a factor of 0.1 if no improvement in validation loss was observed. Early stopping with patience level of 15 epochs was also used to stop the training if no improvement in validation loss is observed. Batch normalization was used to train all networks.

Training was done using two Titan X (Pascal) GPUs and training the network with *no* augmentation, *contour* augmentation, *data* augmentation, and *contour* plus *data* augmentation took approximately 1 hour, 1.5 hours, 2 hours, and 5.5 hours, respectively. Network with *contour* plus *data* augmentation took longer than others due to processing of multiple GTs per image followed by data augmentation. In that case, computational complexity scales linearly with the number of contour resamplings. Keras[11] was used for all the experiments.

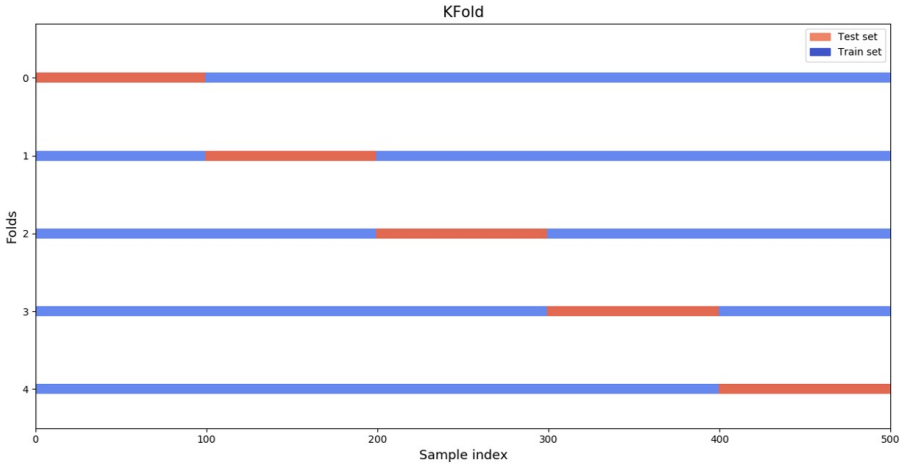

Figure 4. Illustration of the k-folds with train and test split.

4. RESULTS

We hereby report results of the experiments conducted to quantify the performance of contour augmentation in terms of multiple organ segmentation. All results given are obtained by training an instance of *dilated* U-Net for each model, i.e., no augmentation (NoAug), data augmentation (DataAug), contour augmentation (CtrAug), and contour plus data augmentation (CtrDataAug).

The predicted segmentations of each model against the GT segmentations evaluated by the above-mentioned quantitative measures are given in tables 1 to 4. Low recall signifies that the network is not able to identify the whole organ, thus missing some spatial details resulting in coarse segmentation. Low precision means that the network encompasses wrong details as part of the target organ, producing fragmented segmentation. We observe that NoAug and DataAug (tables 1 and 2) are unable to generalize to unseen test cases and perform

in an identical manner (comparable quantitative values). It can be seen from table 3 that CtrAug offers better quantitative values than NoAug and DataAug. CtrDataAug provides the best results (given in table 4) that indicate superior network learning by CtrDataAug compared to other models.

Table 1. Per organ quantitative measures computed for NoAug

Organs	Dice	Recall	Precision	ASSD (mm)	HD_{95} (mm)	CKappa
Bladder	0.80±0.26	0.78±0.25	0.88±0.26	0.30±0.12	2.76±1.27	0.80±0.26
Rectum	0.65±0.32	0.58±0.32	0.80±0.33	0.13±0.06	1.61±1.95	0.65±0.32
Collective	0.72±0.30	0.68±0.30	0.84±0.30	0.22±0.13	2.18±1.75	0.72±0.30

Table 2. Per organ quantitative measures computed for DataAug

Organs	Dice	Recall	Precision	ASSD (mm)	HD_{95} (mm)	CKappa
Bladder	0.83±0.22	0.82±0.19	0.88±0.24	0.30±0.17	2.68±1.43	0.83±0.22
Rectum	0.63±0.30	0.60±0.33	0.80±0.24	0.15±0.08	1.76±1.93	0.63±0.30
Collective	0.73±0.28	0.71±0.29	0.84±0.25	0.22±0.16	2.22±1.76	0.73±0.28

Table 3. Per organ quantitative measures computed for CtrAug

Organs	Dice	Recall	Precision	ASSD (mm)	HD_{95} (mm)	CKappa
Bladder	0.85±0.24	0.87±0.21	0.87±0.25	0.25±0.11	2.19±0.77	0.85±0.24
Rectum	0.77±0.16	0.68±0.21	0.95±0.06	0.12±0.07	1.24±1.42	0.77±0.16
Collective	0.81±0.21	0.78±0.23	0.91±0.18	0.19±0.11	1.71±1.23	0.81±0.21

Box plots (representative of median and quartiles) per model for DSC and ASSD are given in Figs. 5 and 6. It can be seen from Fig. 5 that CtrDataAug provides the best Dice score for both per organ and collective (organ 1 plus organ 2). Similarly, CtrDataAug provides the lowest ASSD in mm as shown in Fig. 6. We report mean ASSD with standard deviation of 0.22 ± 0.09 mm, 0.09 ± 0.05 mm and 0.15 ± 0.10 mm for bladder, rectum, and collective, respectively.

To check significance of each model, we applied Wilcoxon signed-rank test.[12] We found non-significant difference between collective Dice scores of NoAug and DataAug, and significant difference between collective Dice scores of CtrDataAug when compared against NoAug, DataAug, and CtrAug as given in table 5 and Fig. 7. Moreover, we found p-values of DataAug and CtrAug to be significantly different relative to their Dice scores indicating that network learning is different for the two augmentation types. All reported p-values were adjusted by *False Discovery Rate* (FDR)[13] method. CtrDataAug yields better results in terms of Dice scores compared to other models and is found to be the best performing model.

Finally, we illustrate a few predicted segmentations per organ for each model in Figs. 8 to 11. Each image is plotted against the GT segmentations where light colors correspond to GT and dark colors correspond to the predicted segmentations by the above-mentioned models. Fig. 8 is a very rare observation in our test case and it can be seen that each model tries to generalize the learning but is unable to scale to this observation. However, CtrDataAug is robust to such rare observations and scales well to this test case (*closest to the GT*). Another test case for bladder is given in Fig. 9. Interestingly, NoAug fails to generate any segmentation map for this test case. DataAug and CtrAug provide segmentation maps but suffer from low precision. CtrDataAug provides a segmentation map with higher precision and encompasses the target organ. Similar observations can be made for rectum (given in Figs. 10 and 11). Both qualitative and quantitative results indicate that CtrDataAug is the best performing model amongst the mentioned models that can better scale its learning to unseen test cases. Segmentation maps generated by CtrDataAug have both higher recall and precision compared to the other models (given in table 4).

Table 4. Per organ quantitative measures computed for CtrDataAug

Organs	Dice	Recall	Precision	ASSD (mm)	HD_{95} (mm)	CKappa
Bladder	0.88±0.19	0.88±0.18	0.91±0.19	0.22±0.09	1.95±0.87	0.88±0.19
Rectum	0.89±0.04	0.85±0.08	0.95±0.06	0.09±0.05	0.57±0.94	0.89±0.04
Collective	0.88±0.14	0.86±0.14	0.93±0.14	0.15±0.10	1.26±1.14	0.88±0.14

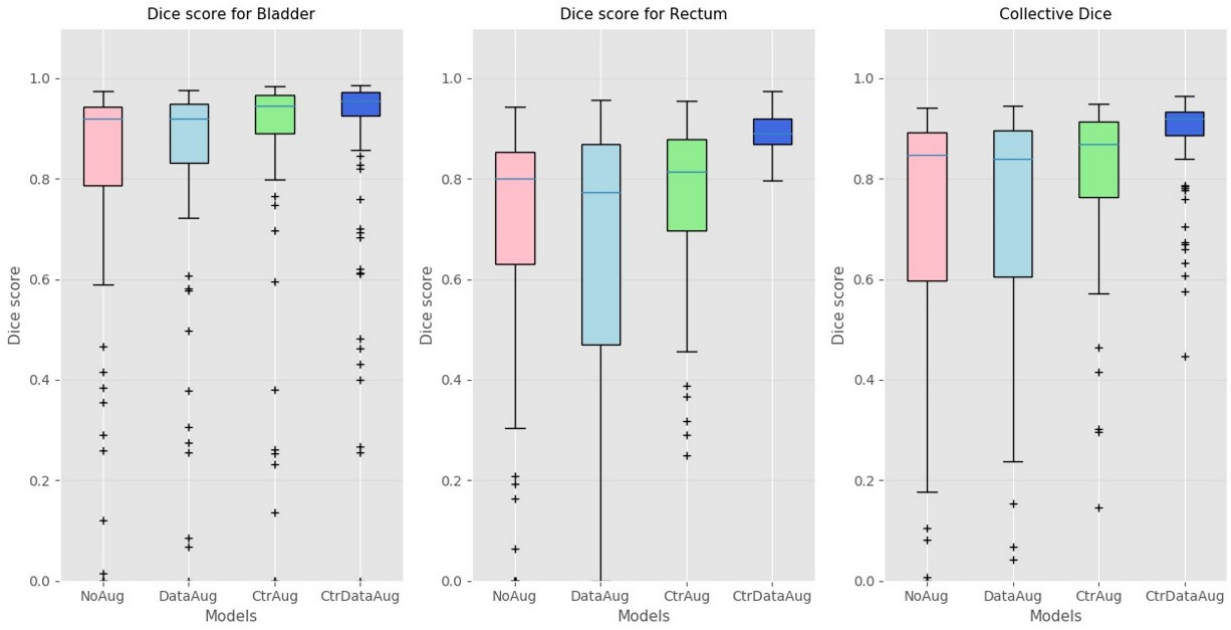

Figure 5. Boxplots of Dice score per model

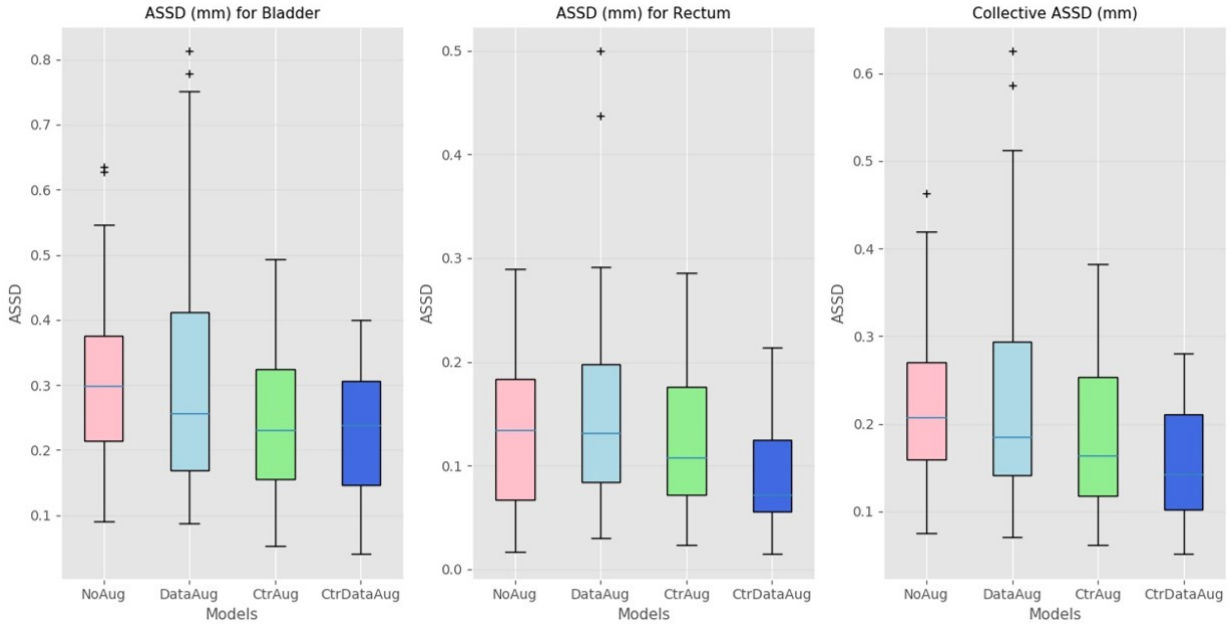

Figure 6. Boxplots of ASSD in mm per model

Table 5. Pairwise comparisons with Wilcoxon signed-rank test computed on Dice score (collective)

Models	NoAug	DataAug	CtrAug	CtrDataAug
NoAug	1.00	6.31×10^{-01}	7.46×10^{-10}	5.55×10^{-14}
DataAug	6.31×10^{-01}	1.00	3.35×10^{-08}	5.19×10^{-13}
CtrAug	7.46×10^{-10}	3.35×10^{-08}	1.00	2.40×10^{-10}
CtrDataAug	5.55×10^{-14}	5.19×10^{-13}	2.40×10^{-10}	1.00

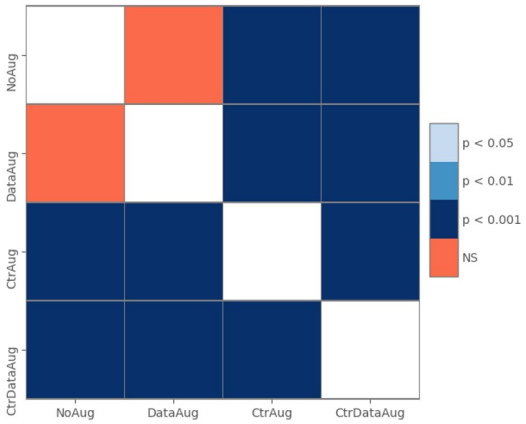

Figure 7. Pairwise comparisons with Wilcoxon signed-rank test computed on collective Dice score. It can be seen that NoAug and DataAug are non-significant (NS) in terms of their p-values.

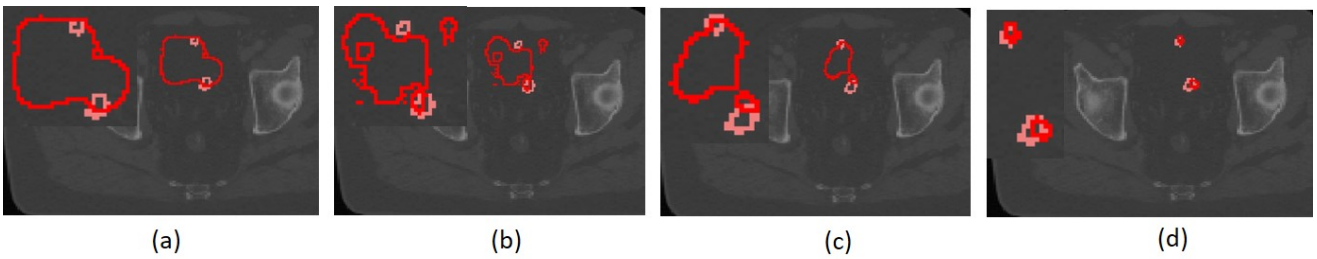

Figure 8. Per model segmentation of bladder: (a) NoAug, (b) DataAug, (c) CtrAug, and (d) CtrDataAug. Light color indicates manually annotated GT and dark color shows the model's predictions.

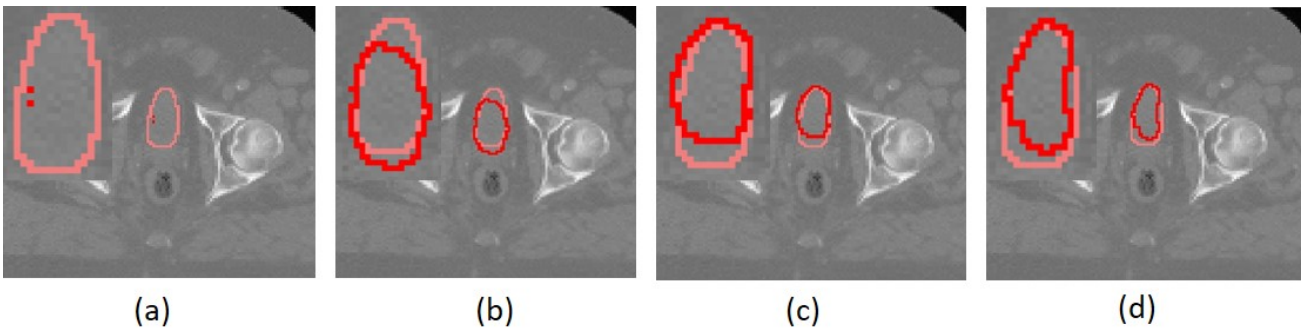

Figure 9. Per model segmentation of bladder: (a) NoAug, (b) DataAug, (c) CtrAug, and (d) CtrDataAug. Light color indicates manually annotated GT and dark color shows the model's predictions.

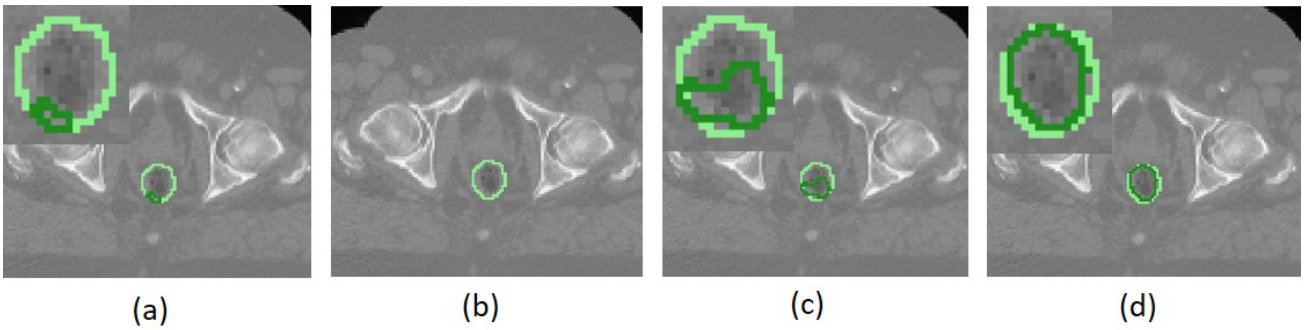

Figure 10. Per model segmentation of rectum: (a) NoAug, (b) DataAug, (c) CtrAug, and (d) CtrDataAug. Light color indicates manually annotated GT and dark color shows the model's predictions.

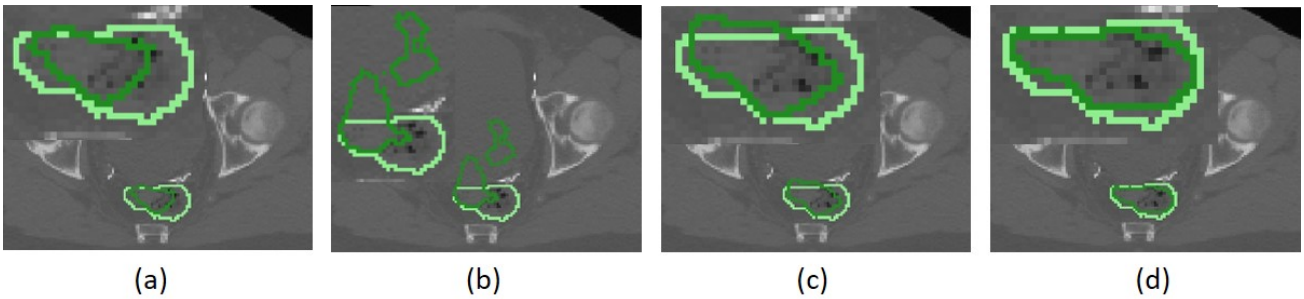

Figure 11. Per model segmentation of rectum: (a) NoAug, (b) DataAug, (c) CtrAug, and (d) CtrDataAug. Light color indicates manually annotated GT and dark color shows the model's predictions.

5. DISCUSSION

Medical practitioners usually perform manual organ delineation that is used as a ground truth for data analysis. There is a risk that the manual annotation might lack accuracy. For example, a physician might find some organs less relevant for the radiation treatment planning, therefore such organs might be roughly or not annotated at all. Interpolation of contours across slices also occurs frequently in manual delineation. Therefore, in order to benchmark the deep learning performance versus human-level accuracy, we need an accurately annotated ground truth that offers effective learning of the anatomical changes and hence, help the doctors to devise better and more precise treatment planning. Relying on retrospective images taken from clinical routine might be sub-optimal compared to images that could be prospectively collected and annotated specifically. While this has been feasible for atlases with only one or few images, the cost in time and money for deep learning where many examples are necessary, might turn out much too high.

Having a valid ground truth, we also need to have a standardized data augmentation technique. As annotated medical data is limited and deep learning networks are data intensive, data augmentation techniques are a common practice to increase the training datasets. It is a usual practice to augment the images but we should also explore the network learning by augmenting the contours. By doing so, we will potentially generate multiple contours per organ and hence, multiple ground truths which could be a surrogate for the actual inter/intra-observer variabilities.

The artificially introduced variations in annotations using contour augmentation do not convey additional information compared to the original, possibly ill-defined ground truth. However, it can be hypothesized that some of these variations in the segmentation masks match salient features in the underlying image, thereby providing more consistent pairs of images and masks than the available original.

6. CONCLUSION

In this paper, we addressed some of the shortcomings of possible insufficient annotation quality by augmenting the manual annotations of GTs. We call it contour augmentation. In a traditional setting, an input image X with its associated label Y is given to a network and data augmentation is performed on the input data pair (X, Y). This effectively increases the input data while the label always remains the same. If the label is poorly defined, then even with data augmentation, the network learning remains sub-optimal because it is conditioned on the poor label. On the contrary, in this work we propose to augment the labels using contour augmentation and thereby account for the annotation uncertainty. This enables the network to focus on the resampled label that is most consistent with the associated input.

We observed that by providing several GTs per image, the network can accomodate variability in terms of localizing the organ boundaries. This improves the network predictive capabilities by correcting the possibly offset or inaccurate annotations defined in the GT, hence bridging the limitations of ill-defined annotations. We compared the performance of a standard network using NoAug, DataAug, CtrAug, and CtrDataAug. Network trained with NoAug lacked the generalization to unseen test examples and performed the worst amongst the other models. Network trained with DataAug under-performed for some test cases whereas network trained with CtrAug performed better than NoAug and DataAug, indicating increased network learning given multiple GTs. However, when DataAug and CtrAug are combined into *CtrDataAug* and are applied together to the network, it outperforms other models both qualitatively and quantitatively. CtrDataAug was also found to be significantly different in terms of Dice scores (with p-value equal to 5.55×10^{-14}, 5.19×10^{-13}, and 2.40×10^{-10}) when compared against NoAug, DataAug, and CtrAug, respectively.

Network trained with DataAug lacked the ability to correct the offset or inaccurate annotations and was biased by the given GT whereas network trained using CtrAug was robust to ill-defined annotation. We found that by joining the two augmentation techniques, i.e., data and contour augmentation, we get the best performing model that leads to accurate organ localization as it can better cope with the possible bias introduced by annotation uncertainty.

Apart from the quantitative analyses, we also performed a qualitative analysis (*visual Turing test*[14]) on two medical annotation experts. Each model's predictions (segmentations) along with the corresponding GT were presented to them and they had to choose the best segmentations. Both of them mostly chose GT as a good segmentation whereas they selected predicted segmentation maps of CtrDataAug for some test cases. In short, their selection was limited to either GT or CtrDataAug. To our knowledge, this is the first attempt to explore the network learning in *multi-label per image* setting. Moreover, we believe contour augmentation (used in combination/union with data augmentation) can potentially incorporate the inter and intra-observer variabilities and thus, correct the possibly induced bias introduced during the manual annotation step.

Our proposed methodology is computationally inexpensive and can be generalized to any annotation/labeling task (*binary masks*) for supervised learning as the computational complexity is proportional to the GT resampling size where it scales linearly with the number of contour resamplings. We would further extend our proposed methodology for whole volume CT scans from different anatomical sites (*retrospective* and possibly *prospective*). We believe including prospectively collected CT scans can provide additional information to the network, which will enable it to learn more about the anatomical changes and generalize better on real-world cases.

7. ACKNOWLEDGMENT

Umair Javaid and Damien Dasnoy are funded by the Belgian Fonds de la Recherche Scientifique F.R.S.-FNRS, Télévie grants no. 7.4625.16 and 7.6509.17. John A. Lee is a Senior Research Associate with the F.R.S.- FNRS. We thank the University hospital (CHU Namur UCLouvain, Dr. Jean-François Daisne) for providing the data. We also thank the NVIDIA Corporation for providing Titan X (Pascal) GPUs.

REFERENCES

[1] Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., Van Der Laak, J. A., Van Ginneken, B., and Sánchez, C. I., "A survey on deep learning in medical image analysis," *Medical image analysis* **42**, 60–88 (2017).

[2] Bai, W., Sinclair, M., Tarroni, G., Oktay, O., Rajchl, M., Vaillant, G., Lee, A. M., Aung, N., Lukaschuk, E., Sanghvi, M. M., et al., "Human-level cmr image analysis with deep fully convolutional networks," *arXiv preprint arXiv:1710.09289* (2017).

[3] Liao, F., Liang, M., Li, Z., Hu, X., and Song, S., "Evaluate the malignancy of pulmonary nodules using the 3d deep leaky noisy-or network," *arXiv preprint arXiv:1711.08324* (2017).

[4] Javaid, U., Dasnoy, D., and Lee, J. A., "Multi-organ segmentation of chest ct images in radiation oncology: Comparison of standard and dilated unet," in [*International Conference on Advanced Concepts for Intelligent Vision Systems*], 188–199, Springer (2018).

[5] Kohl, S. A., Romera-Paredes, B., Meyer, C., De Fauw, J., Ledsam, J. R., Maier-Hein, K. H., Eslami, S., Rezende, D. J., and Ronneberger, O., "A probabilistic u-net for segmentation of ambiguous images," *arXiv preprint arXiv:1806.05034* (2018).

[6] Ronneberger, O., Fischer, P., and Brox, T., "U-net: Convolutional networks for biomedical image segmentation," in [*International Conference on Medical image computing and computer-assisted intervention*], 234–241, Springer (2015).

[7] Dutilleux, P., "An implementation of the "algorithme à trous" to compute the wavelet transform," in [*Wavelets*], 298–304, Springer (1990).

[8] Simard, P. Y., Steinkraus, D., Platt, J. C., et al., "Best practices for convolutional neural networks applied to visual document analysis.," in [*ICDAR*], **3**, 958–962 (2003).

[9] Heimann, T., Van Ginneken, B., Styner, M. A., Arzhaeva, Y., Aurich, V., Bauer, C., Beck, A., Becker, C., Beichel, R., Bekes, G., et al., "Comparison and evaluation of methods for liver segmentation from ct datasets," *IEEE transactions on medical imaging* **28**(8), 1251–1265 (2009).

[10] Cohen, J., "A coefficient of agreement for nominal scales," *Educational and psychological measurement* **20**(1), 37–46 (1960).

[11] Chollet, F. et al., "Keras," (2015).

[12] Wilcoxon, F., "Individual comparisons by ranking methods," *Biometrics bulletin* **1**(6), 80–83 (1945).

[13] Benjamini, Y. and Hochberg, Y., "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the royal statistical society. Series B (Methodological)* , 289–300 (1995).

[14] Geman, D., Geman, S., Hallonquist, N., and Younes, L., "Visual turing test for computer vision systems," *Proceedings of the National Academy of Sciences* , 201422953 (2015).