# Learning Clinically Acceptable Segmentation of Organs at Risk in Cervical Cancer Radiation Treatment from Clinically Available Annotations

**Author name(s) withheld** EMAIL(S) WITHHELD
*Address withheld*

**Editors:** Under Review for MIDL 2023

## Abstract

Deep learning models benefit from training with a large dataset (labeled or unlabeled). Following this motivation, we present an approach to learn a deep learning model for the automatic segmentation of Organs at Risk (OARs) in cervical cancer radiation treatment from a large clinically available dataset of Computed Tomography (CT) scans containing data inhomogeneity, label noise, and missing annotations. We employ simple heuristics for automatic data cleaning to minimize data inhomogeneity and label noise. Further, we develop a self-supervised training strategy utilizing a teacher-student setup, annotation imputation, and uncertainty-guided training to learn in presence of missing annotations. Our experimental results show that learning from a large dataset with our approach yields a significant improvement in the test performance despite missing annotations in the data. Further, the contours generated from the segmentation masks predicted by our best-performing model are found to be equally clinically acceptable as manually generated contours.

**Keywords:** organs at risk, segmentation, deep learning, missing annotations

## 1. Introduction

The planning for cervical cancer radiation treatment[1] requires manual contouring of the Organs at Risk (OARs) where the adverse effects of radiation must be minimized. Automatic segmentation of these OARs can save hours of manual work. In this paper, we focus on the automatic segmentation of four OARs in cervical cancer radiation treatment: bowel bag, bladder, hips, and rectum. A few studies have focused on developing deep learning based automatic OARs segmentation methods for cervical cancer radiation treatment (Liu et al., 2020b; Mohammadi et al., 2021; Liu et al., 2020a). All of these studies use a traditional setup for developing a deep learning model, which involves: (a) obtaining a fully annotated clinically available dataset, (b) splitting the data into training, validation, and testing, and (c) training a model and evaluating it on the test dataset. A major drawback in this setup is the limited size of the datasets used for training and testing. A small training dataset limits the possibility of a deep learning model capturing large variance in real-world data. Further, evaluation results from a small test dataset do not inform sufficiently in regard to the true test performance of a deep learning model. Although in the medical imaging domain, such a setup is understandable because of the underlying requirement of clinical

1. Radiation treatment for cancer involves giving high doses of radiation to the tumor to kill cancer cells.

© 2023 Anonymous.

expertise for annotating the data, it would be of interest to investigate if clinically available data can be leveraged to increase the size of the training and testing datasets.

Recent studies in the domains of natural image processing (Xie et al., 2020), and protein structure prediction (Jumper et al., 2021) have highlighted the use of unlabeled or partially labeled large datasets towards improving the performance of a deep learning model. In the case of OARs segmentation for cervical cancer radiation treatment, a large training dataset can be accumulated if the abdominal scans acquired for tumors other than cervical cancer are also included. However, all the OARs in cervical cancer radiation treatment may not be annotated in those scans. Furthermore, since the clinically available abdominal scans and annotations are retrospectively included, the acquisition protocols, contouring guidelines, and observers may be different giving rise to data inhomogeneity and label noise.

In this paper, we follow the motivation of harnessing the benefits of training on a large dataset. Therefore, we use the Computed Tomography (CT) scans and OARs contours delineated for clinical use during radiation treatment for tumors in the abdominal region to develop a deep learning model for segmentation of OARs in cervical cancer radiation treatment. The key contributions of our work are the following:

1. We propose heuristics based on the training dataset for automatic data cleaning to increase data homogeneity and reduce label noise.
2. We propose a self-supervised training strategy involving a teacher-student setup, annotation imputation, and uncertainty-guided training to develop a deep learning model with the use of clinically available data containing missing annotations.
3. We perform an ablation study to investigate the individual effect of different components of the proposed approach. Furthermore, we perform a clinical validation study to assess the clinical acceptability of contours generated from automatic segmentation masks predicted by our deep learning model.

## 2. Data

We retrospectively selected the CT scans of female patients who were treated in an academic hospital for a tumor in the abdominal region from 2009 to 2019. The CT scans with associated contours were received in anonymized form through a data transfer agreement. A total of 1170 scans from 1108 patients were received. These scans were used for training and validation. In addition, we used 105 CT scans from 95 cervical cancer patients who received radiation treatment as test data.

### 2.1. Preprocessing

The clinically available annotations of four OARs in cervical cancer radiation treatment (bowel bag, bladder, hips, and rectum) were extracted for all the CT scans (1170 from the training and validation dataset, and 105 from the test dataset) by using the following steps: (1) standardize different variations of organ labels (e.g., bowel, bowel bag, Bowel bag, bowel_bag, Bowel_bag were all considered bowel bag), (2) combine left and right hip annotations as a single organ, (3) remove voxels annotated as bladder or rectum from the bowel bag annotation to avoid ambiguous labeling in those voxels. Next, all the CT scans were resampled to 2.5mm $\times$ 2.5mm $\times$ 2.5mm voxel spacing. The Hounsfield units were

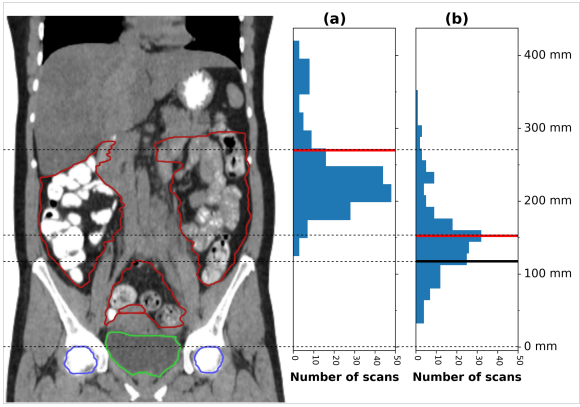



Figure 1: The histograms of the number of scans with respect to the distance from the most cranial point of the hip annotations to (a) the cranial border of the FOV of the scan, and (b) the cranial border of the bowel bag annotation. On the left, a representative CT scan and reconstructed contours in the coronal view are shown (red: bowel bag, green: bladder, blue: hips). Red lines in (a) and (b): thresholds to crop the FOV and delete the bowel bag annotations in the cranial direction. The black line in (b): threshold for discarding the scans, where the bowel bag annotations did not cover the pelvic region. Dashed lines: corresponding anatomy for each threshold.

converted to intensity values between 0 and 1 by windowing. In the training and validation dataset, the preprocessing resulted in a total of 186 scans that contained annotations for all the four OARs considered in this work (referred to as the fully annotated dataset, $\mathcal{D}_f$ in the rest of the paper). The remaining scans had missing annotations for at least one of the OARs (referred to as the partially annotated dataset, $\mathcal{D}_p$). The distribution of the available annotations in the training and validation dataset is described in Table 1.

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Bowel bag | ✓ | ✓ | | ✓ | ✓ | ✓ | | | ✓ | ✓ | | ✓ | | | |
| Bladder | ✓ | ✓ | ✓ | ✓ | | ✓ | ✓ | | | | ✓ | | ✓ | | |
| Hip | ✓ | ✓ | ✓ | | ✓ | | ✓ | ✓ | ✓ | | | | | ✓ | |
| Rectum | ✓ | | ✓ | ✓ | ✓ | | | ✓ | | ✓ | ✓ | | | | ✓ |
| **No. of scans** | 186 | 349 | 263 | 198 | 201 | 366 | 480 | 285 | 365 | 213 | 832 | 383 | 1103 | 504 | 865 |

Table 1: Distribution of annotations in the training and validation dataset. In each column, the No. (number) of scans for which the annotations were present for at least the OAR marked with ✓are reported .

### 2.2. Automatic Data Cleaning

Since the data was accumulated over 10 years and the scans belonging to patients who were treated for a tumor anywhere in the abdominal region were included, the data exhibited inhomogeneity in the transverse Field-of-View (FOV) of the scan (causing an increase in the number of background voxels and potentially less efficient training), and the cranial border of the bowel bag annotations (attributing to label noise).

To make the data more homogeneous such that the adverse effects of inefficient training and label noise could be avoided, we developed heuristics derived from the histograms of $\mathcal{D}_f$

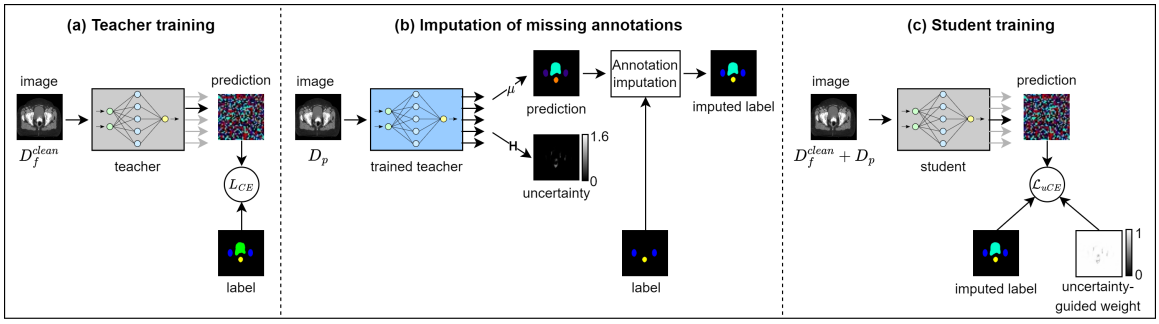



Figure 2: Schematic of the proposed self supervised training involving (a) training a K-head teacher model, (b) annotation imputation using predictions from trained teacher, and (c) training a student model with imputed labels and uncertainty-guided loss. $\mu$: mean, H: entropy, $L_{CE}$: cross-entropy loss, $\mathcal{L}_{uCE}$: uncertainty-guided loss.

(shown in Figure 1). We analyzed the histograms and decided on thresholds (marked in the figure) such that the histograms represented a unimodal distribution corresponding to the most frequently used scanning protocol and annotation style. Based on these thresholds, the scans were cropped in the cranial direction to remove the chest region. The bowel bag annotations in the abdominal region roughly above the level of the lumbar (L4) spinal segment were deleted. The scans that did not contain bowel bag annotations covering the entire pelvic region were discarded. These steps resulted in a decrease in the size of $\mathcal{D}_f$ from 186 to 134. The resulting dataset of 134 scans is referred to as $\mathcal{D}_f^{clean}$ in the rest of the paper.

## 3. Approach

We developed a self-supervised training approach utilizing a teacher-student setup. We trained a teacher model using the small, fully annotated dataset ($\mathcal{D}_f^{clean}$). The predictions from the trained teacher model were used to impute the remaining large dataset with missing annotations ($\mathcal{D}_p$). Then, a student was trained with the entire dataset ($\mathcal{D}_f^{clean} + \mathcal{D}_p$) containing the clinically available and imputed annotations. Alternative approaches include fine-tuning the teacher with the entire dataset containing imputed annotations (Zheng et al., 2020). We chose a teacher-student setup because existing literature suggests that training a student from scratch yields better results than fine-tuning the teacher (Xie et al., 2020). A schematic of our approach is illustrated in Figure 2.

### 3.1. Uncertainty Estimation

Epistemic uncertainty refers to the lack of knowledge in a model about the underlying data. Estimating epistemic uncertainty enables the estimation of the reliability of a model's prediction for a specific sample. We trained the teacher model to also estimate the epistemic uncertainty maps for each sample. For this purpose, we used a K-head neural network. At each iteration of training, a single head was selected randomly for backpropagation. We used the entropy of the mean prediction from K-heads as an estimate of epistemic uncertainty. We selected the K-head approach because it allows independence between

predictions from different heads with faster inference times as compared to the Monte-Carlo (MC) dropout approach (Gal and Ghahramani, 2016). Moreover, the memory overhead is not much compared to fully independent deep ensembles.

### 3.2. Uncertainty-Guided Training

We trained the student model with a cross-entropy loss weighted by an uncertainty-guided term ($e^{-u}$, where $u$ is uncertainty in the teacher's predictions) at each voxel. The weight $e^{-u}$ ensured a large weight on voxels where the uncertainty in the teacher's predictions was small and vice-versa. The loss at each voxel was defined as $\mathcal{L}_{uCE} = e^{-u} y \cdot log(\hat{y})$, where $y$ is the reference label, and $\hat{y}$ is the predicted probability. We set $u = 0$ at the voxels where annotations were clinically available. In this way, the student model could benefit from training with a large dataset while avoiding deterioration in performance due to uncertain label predictions from the teacher model.

### 3.3. Implementation Details

We used the U-Net architecture (Ronneberger et al., 2015) as a baseline neural network, which is well known for its performance on semantic segmentation tasks, especially in the case of medical images. The implementation[2] was done in Python by using the PyTorch library (Paszke et al., 2017) and the training was done on NVIDIA RTX2080 GPUs. Other hyperparameters were: network initialization = He norm; optimizer = Adam (Kingma and Ba, 2015); network initialization = Kaiming He (He et al., 2015); learning rate (LR) = $1e^{-3}$; weight decay = $1e^{-4}$; the number of training epochs = 500 for teacher models, 250 for student models; learning schedule = step LR with step size = $\frac{1}{3}\times$ total training steps; data augmentations = global brightness and contrast variations, random rotations; the number of heads (K) in teacher and student = 5.

## 4. Ablation Experiment

We conducted an ablation experiment to look into the effect of the different components mentioned in the above sections, on the test performance. Specifically, we compared the test performance of the following models: U-Net trained with the small, fully annotated dataset before automatic data cleaning ($\mathcal{D}_f$), U-Net trained with the small, fully annotated dataset after automatic data cleaning ($\mathcal{D}_f^{clean}$), K-head U-Net teacher model trained with $\mathcal{D}_f^{clean}$ (referred to as '*basic teacher*'), K-head U-Net student model trained with the large dataset ($\mathcal{D}_f^{clean} + \mathcal{D}_p$) with uncertainty-guided loss (referred to as '*basic student*').

Earlier studies suggest that training a noisy student (noise in the data or model) on a large dataset with imputed annotations yields a performance gain (Xie et al., 2020). To test this, we employed the following additional data augmentations as a part of the noise in the data: left-right flipping, masking an organ with a random intensity to simulate contrast, global elastic deformations, and elastic deformations centered in either bowel bag or bladder as additional augmentations. We compared the performance of three models: a teacher model trained with $\mathcal{D}_f^{clean}$ and additional augmentations (referred to as '*robust*

2. The source code will be made publicly available.

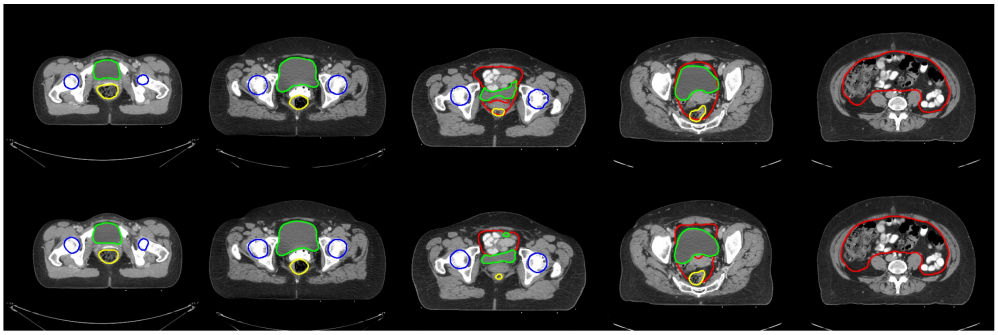
| **Method** | Dice | Surface Dice | HD |
|---|---|---|---|
| U-Net + $\mathcal{D}_f$ | .8347 (.0616) | .8023 (.0682) | 16.06 (9.07) |
| U-Net + $\mathcal{D}_f^{clean}$ | .8502 (.0592)$^*$ | .8200 (.0655)$^*$ | 12.44 (10.58)$^*$ |
| K-head teacher + $\mathcal{D}_f^{clean}$ (*basic teacher*) | .8536 (.0554)$^*$ | .8233 (.0618)$^*$ | 11.61 (7.94)$^*$ |
| K-head student + ($\mathcal{D}_f^{clean} + \mathcal{D}_p$) (*basic student*) | .8701 (.0462)$^{*\dagger}$ | .8464 (.0518)$^{*\dagger}$ | 10.64 (8.00)$^{*\dagger}$ |
| K-head teacher + $\mathcal{D}_f^{clean}$ + Aug. (*robust teacher*) | .8531 (.0525)$^*$ | .8230 (.0572)$^*$ | 11.57 (7.73)$^*$ |
| K-head student + ($\mathcal{D}_f^{clean}$+$\mathcal{D}_p$) + Aug. in student training (*basic teacher + robust student*) | .8711 (.0428)$^{*\dagger}$ | .8476 (.0485)$^{*\dagger}$ | 10.39 (6.68)$^{*\dagger}$ |
| K-head student + ($\mathcal{D}_f^{clean}$+$\mathcal{D}_p$) + Aug. in teacher and student training (*robust teacher + robust student*) | .8716 (.0419)$^{*\dagger}$ | .8482 (.0468)$^{*\dagger}$ | 9.92 (4.72)$^{*\dagger}$ |

Table 2: Mean (standard deviation) of mean test performance per scan of the best models obtained from 5-fold cross-validation. Aug.: additional augmentations, HD: Hausdorff distance in mm at 95 percentile. Surface Dice was computed at a tolerance of 2.5mm (voxel spacing). $^*$significant differences compared to U-Net + $\mathcal{D}_f$, $^\dagger$significant differences compared to U-Net + $\mathcal{D}_f^{clean}$.

*teacher*'), a student model trained with $\mathcal{D}_f^{clean} + \mathcal{D}_p$ and additional augmentation, and using the imputed annotations from *basic teacher* (referred to as '*basic teacher + robust student*'), and a student model trained with $\mathcal{D}_f^{clean} + \mathcal{D}_p$ and additional augmentation, and using the imputed annotations from *robust teacher* (referred to as '*robust teacher + robust student*').

The mean and standard deviations of the performance metrics on test data from the best models obtained after 5-fold cross-validation are reported in Table 2. The distributions of performance metrics for each method (N = 105 test scans $\times$ 5 models) were tested for normality using the Kolmogorov-Smirnov test. This was followed by a Friedman test for the main effect and Wilcoxon signed-rank test for post-hoc comparisons. A p-value less than 0.05 with adjustment for multiple comparisons was considered significant.

The automatic data cleaning had a significant impact on the test performance ($p = 5.96e^-18$, $p = 6.76e^-17$, $p = 2.18e^-29$ for Dice, Surface Dice (SD), and Hausdorff distance (HD), respectively), which was mainly due to better bowel bag segmentation. The automatic data cleaning increased the mean Dice coefficient of the bowel bag from 0.7947 to 0.8477 (performance metrics for all the OARs separately are provided in Appendix A). Furthermore, learning from a large dataset with the proposed teacher-student setup, annotation imputation, and uncertainty-guided training (*basic student*) provided a significant gain of 2.34% in mean Dice coefficient ($p = 4.51e^-38$), 3.22% in mean SD ($p = 1.21e^-35$), and 14.47% in mean HD ($p = 1.51e^-15$) as compared to learning from a small, fully annotated dataset (U-Net + $D_f^{clean}$). Adding noise to the data through additional augmentations provided a marginal gain in the mean performance of the student model. Intuitively, additional augmentations provided a considerable decrease in the standard deviations of performance metrics indicating robustness towards variations in the test data. A few representative examples from the results obtained by *basic teacher + robust student* are shown in Figure 3.



| | manual | automatic | manual | automatic | manual | automatic | manual | automatic | manual | automatic |
|---|---|---|---|---|---|---|---|---|---|---|
| bowel bag | - | - | - | - | 3 | 2 | 3 | 3 | 2 | 1 |
| bladder | 1 | 1 | 1 | 1 | 2 | 2 | 1 | 1 | - | - |
| hips | 2 | 2 | 2 | 2 | 1 | 1 | - | - | - | - |
| rectum | 2 | 1 | 1 | 1 | 1 | 3 | 1 | 1 | - | - |

Figure 3: Representative examples of OARs contours. *Top row*: clinically available contours (manual), *Bottom row*: contours generated from OARs segmentation masks predicted by our approach (automatic). Further, the clinical acceptability grades (smaller value indicates better quality) are reported for each OAR.

## 5. Clinical Acceptability Test

We conducted a validation study to assess the clinical acceptability of the automatically generated OARs segmentations. We used the best *basic teacher + robust student* model (based on cross-validation performance) from the first data-split, to predict segmentation masks for OARs in a subset of 4 scans, which were used to generate automatic contours. We showed both the clinically available contours and the automatically generated contours to a radiation oncologist (henceforth referred to as 'clinical expert'), without informing them about the method used to generate the contours. The contours were presented on 2D transverse slices spaced at a 10mm distance to make it similar to the clinical scenario where the contours are delineated on 2D transverse slices. The clinical expert optionally inspected the contours and scans in coronal and sagittal view also to ensure comprehensiveness. The clinical expert graded each contour for its clinical acceptability according to a 4-point Likert scale: 1 = acceptable as it is, 2 = acceptable but marginally deviating from exact anatomical definition (subjective to an observer), 3 = acceptable with minor corrections (e.g., only adapting a small part of the contour) because either a part of the organ was not delineated or a peripheral tissue was included in the contour, 4 = not acceptable because a correction involving both deletion, as well as delineation of an additional contour, was required.

The clinical acceptability grades for the automatically and manually generated contours for all the graded 2D transverse slices and OARs are shown in Figure 4. None of the contours were given grade 4 implying that all the contours were of clinically acceptable quality either as it is or with adaptations. Further, not all of the clinically available contours were graded as 1, representing inter-observer variation. A Chi-squared test of goodness of fit indicated that the histograms of clinical acceptability grades of the automatically generated contours were significantly different from the manually generated contours for the bowel bag ($\chi^2(1, N{=}58) = 11.402, p = 0.003$). However, as shown in the Figure 4, it was unclear which contours (automatically or manually generated) were better. The clinical acceptability grades for automatically and manually generated contours were not significantly different for the bladder ($\chi^2(1, N{=}27) = 2.667, p = 0.102$), and hips ($\chi^2(1, N{=}18) = 2.250, p = 0.134$).

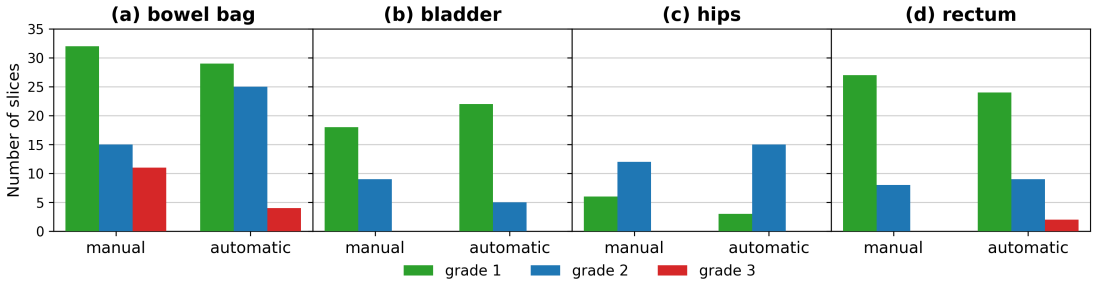



Figure 4: Comparison of clinical acceptability grades (smaller value indicates better quality) for clinically available contours (manual) and the contours generated from OARs segmentation masks predicted by our approach (automatic) for (a): bowel bag, (b): bladder, (c): hips, and (d) rectum.

For the rectum, the Chi-squared test statistics could not be obtained because the frequency counts corresponding to grade 3 were less than 5, however, it is apparent from the Figure 4 that the frequency counts in each category were similar for both the automatically and manually generated contours.

## 6. Discussion and Conclusions

We investigated the possibility of using a large clinically available dataset of the abdominal region to learn a deep learning model for the automatic segmentation of OARs in cervical cancer radiation treatment. To the best of our knowledge, this is one of the few works in the direction of utilizing a large clinically available dataset containing missing annotations for learning a deep learning model. Our experimental results show that learning from a large dataset using our proposed approach yields significant performance gain despite missing annotations in the data. The obtained segmentations from our deep learning model were of clinically acceptable quality, which is encouraging.

Limitations of our work include an ablation study involving only a single run (i.e., network initialization), and a lack of experiments with different semantic segmentation architectures. Both decisions were consciously taken to find sensible results despite the expensive nature of training deep neural networks on a large dataset. Interesting future directions are: 1) extending the current work to add the capability of segmenting more OARs in cervical cancer radiation treatment planning e.g., sigmoid and anal canal, and 2) performing a more extensive clinical acceptability study on a larger dataset.

In conclusion, we demonstrated that training a deep learning model without using curated and specifically annotated medical imaging data, but with the capability of predicting clinically acceptable segmentation is possible. Apart from saving clinicians' time, our proposed approach leads to faster development time because of using the readily available data and increased test performance because of the increased dataset size.

## Acknowledgments

Acknowledgments withheld.

## References

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pages 1050–1059, 2016.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In *IEEE International Conference on Computer Vision*, pages 1026–1034, 2015.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. URL http://arxiv.org/abs/1412.6980.

Zhikai Liu, Xia Liu, Hui Guan, Hongan Zhen, Yuliang Sun, Qi Chen, Yu Chen, Shaobin Wang, and Jie Qiu. Development and validation of a deep learning algorithm for auto-delineation of clinical target volume and organs at risk in cervical cancer radiotherapy. *Radiotherapy and Oncology*, 153:172–179, 2020a. ISSN 0167-8140. doi: https://doi.org/10.1016/j.radonc.2020.09.060. URL https://www.sciencedirect.com/science/article/pii/S0167814020308355. Physics Special Issue: ESTRO Physics Research Workshops on Science in Development.

Zhikai Liu, Xia Liu, Bin Xiao, Shaobin Wang, Zheng Miao, Yuliang Sun, and Fuquan Zhang. Segmentation of organs-at-risk in cervical cancer CT images with a convolutional neural network. *Physica Medica*, 69:184–191, 2020b. ISSN 1120-1797. doi: https://doi.org/10.1016/j.ejmp.2019.12.008. URL https://www.sciencedirect.com/science/article/pii/S1120179719305290.

Reza Mohammadi, Iman Shokatian, Mohammad Salehi, Hossein Arabi, Isaac Shiri, and Habib Zaidi. Deep learning-based auto-segmentation of organs at risk in high-dose rate brachytherapy of cervical cancer. *Radiotherapy and Oncology*, 159:231–240, 2021. ISSN 0167-8140. doi: https://doi.org/10.1016/j.radonc.2021.03.030. URL https://www.sciencedirect.com/science/article/pii/S0167814021061703.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in PyTorch. In *Advances in Neural Information Processing Systems-W*, 2017. URL https://github.com/pytorch/pytorch.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015. ISBN 978-3-319-24574-4.

Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves ImageNet classification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10687–10698, 2020.

Hao Zheng, Susan M Motch Perrine, M Kathleen Pitirri, Kazuhiko Kawasaki, Chaoli Wang, Joan T Richtsmeier, and Danny Z Chen. Cartilage segmentation in high-resolution 3D micro-CT images via uncertainty-guided self-training with very sparse annotation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 802–812. Springer, 2020.

## Appendix A. Performance Metrics for all OARs

| **Method** | Bowel bag | Bladder | Hips | Rectum |
|---|---|---|---|---|
| U-Net + $\mathcal{D}_f$ | .7947 (.1124) | .8925 (.1605) | .9157 (.0373) | .7358 (.1393) |
| U-Net + $\mathcal{D}_f^{clean}$ | .8477 (.0671) | .9024 (.1562) | .9191 (.0223) | .7315 (.1532) |
| K-head teacher + $\mathcal{D}_f^{clean}$ (*basic teacher*) | .8488 (.0621) | .9061 (.1459) | .9181 (.025) | .7415 (.1405) |
| K-head student + ($\mathcal{D}_f^{clean} + \mathcal{D}_p$) (*basic student*) | .8631 (.0559) | .9213 (.1143) | .9265 (.0214) | .7695 (.1263) |
| K-head teacher + $\mathcal{D}_f^{clean}$ + Aug. (*robust teacher*) | .8469 (.0701) | .9023 (.1543) | .9173 (.024) | .7458 (.127) |
| K-head student + ($\mathcal{D}_f^{clean}$+$\mathcal{D}_p$) + Aug. in student training (*basic teacher + robust student*) | .8586 (.0557) | .9208 (.1003) | .9262 (.0219) | .7786 (.1196) |
| K-head student + ($\mathcal{D}_f^{clean}$+$\mathcal{D}_p$) + Aug. in teacher and student training (*robust teacher + robust student*) | .8625 (.0554) | .9193 (.1058) | .9234 (.0222) | .781 (.1099) |

Table 3: Mean (standard deviation) of Dice coefficient of the best models obtained from 5-fold cross-validation. Aug.: additional augmentations.

| **Method** | Bowel bag | Bladder | Hips | Rectum |
|---|---|---|---|---|
| U-Net + $\mathcal{D}_f$ | .6155 (.1077) | .8824 (.1699) | .9662 (.0471) | .7451 (.1499) |
| U-Net + $\mathcal{D}_f^{clean}$ | .6637 (.0927) | .9007 (.1636) | .9703 (.0327) | .7452 (.1551) |
| K-head teacher + $\mathcal{D}_f^{clean}$ (*basic teacher*) | .6645 (.0868) | .9055 (.1561) | .9686 (.0384) | .7546 (.141) |
| K-head student + ($\mathcal{D}_f^{clean} + \mathcal{D}_p$) (*basic student*) | .6891 (.0857) | .9313 (.1197) | .9755 (.0318) | .7896 (.1292) |
| K-head teacher + $\mathcal{D}_f^{clean}$ + Aug. (*robust teacher*) | .6609 (.0884) | .9046 (.1625) | .968 (.0347) | .7586 (.1259) |
| K-head student + ($\mathcal{D}_f^{clean}$+$\mathcal{D}_p$) + Aug. in student training (*basic teacher + robust student*) | .6833 (.0861) | .9295 (.1043) | .9753 (.0323) | .8024 (.121) |
| K-head student + ($\mathcal{D}_f^{clean}$+$\mathcal{D}_p$) + Aug. in teacher and student training (*robust teacher + robust student*) | .6897 (.084) | .9274 (.1078) | .9729 (.0334) | .803 (.1137) |

Table 4: Mean (standard deviation) of Surface Dice computed at a tolerance of 2.5mm (voxel spacing) of the best models obtained from 5-fold cross-validation. Aug.: additional augmentations.

| **Method** | Bowel bag | Bladder | Hips | Rectum |
|---|---|---|---|---|
| U-Net + $\mathcal{D}_f$ | 35.27 (24.77) | 7.84 (10.51) | 4.16 (20.14) | 16.98 (11.86) |
| U-Net + $\mathcal{D}_f^{clean}$ | 19.34 (11.70) | 9.68 (37.38) | 2.93 (1.00) | 17.80 (13.26) |
| K-head teacher + $\mathcal{D}_f^{clean}$ (*basic teacher*) | 18.43 (9.97) | 7.96 (26.75) | 2.95 (1.03) | 17.10 (12.61) |
| K-head student + ($\mathcal{D}_f^{clean} + \mathcal{D}_p$) (*basic student*) | 17.26 (10.47) | 6.31 (22.70) | 2.87 (1.04) | 16.11 (18.35) |
| K-head teacher + $\mathcal{D}_f^{clean}$ + Aug. (*robust teacher*) | 18.43 (10.83) | 7.56 (22.30) | 2.95 (1.05) | 17.34 (17.86) |
| K-head student + ($\mathcal{D}_f^{clean}$+$\mathcal{D}_p$) + Aug. in student training (*basic teacher + robust student*) | 17.55 (10.23) | 5.57 (15.12) | 2.87 (1.05) | 15.58 (18.00) |
| K-head student + ($\mathcal{D}_f^{clean}$+$\mathcal{D}_p$) + Aug. in teacher and student training (*robust teacher + robust student*) | 17.10 (11.23) | 5.12 (7.22) | 2.90 (1.08) | 14.57 (10.63) |

Table 5: Mean (standard deviation) of Hausdorff distance at 95 percentile of the best models obtained from 5-fold cross-validation. Aug.: additional augmentations.