Improving Error Detection in Deep Learning based Radiotherapy Autocontouring using Bayesian Uncertainty

Prerak Mody[1], Nicolas F. Chaves-de-Plaza[3], Klaus Hildebrandt[3], and Marius Staring[1,2]

[1] Department of Radiology, Leiden University Medical Centre, Leiden, The Netherlands
{P.P.Mody, M.Staring}@lumc.nl
[2] Department of Radiation Oncology, Leiden University Medical Centre, Leiden, The Netherlands
[3] Computer Graphics and Visualization Lab, TU Delft, Delft, The Netherlands

Abstract. Bayesian Neural Nets (BNN) are increasingly used for robust organ auto-contouring. Uncertainty heatmaps extracted from BNNs have been shown to correspond to inaccurate regions. To help speed up the mandatory quality assessment (QA) of contours in radiotherapy, these heatmaps could be used as stimuli to direct visual attention of clinicians to potential inaccuracies. In practice, this is non-trivial to achieve since many accurate regions also exhibit uncertainty. To influence the output uncertainty of a BNN, we propose a modified accuracy-versus-uncertainty (AvU) metric as an additional objective during model training that penalizes both accurate regions exhibiting uncertainty as well as inaccurate regions exhibiting certainty. For evaluation, we use an uncertainty-ROC curve that can help differentiate between Bayesian models by comparing the probability of uncertainty in inaccurate versus accurate regions. We train and evaluate a FlipOut BNN model on the MICCAI2015 Head and Neck Segmentation challenge dataset and on the DeepMind-TCIA dataset, and observed an increase in the AUC of uncertainty-ROC curves by 5.6% and 5.9%, respectively, when using the AvU objective. The AvU objective primarily reduced false positives regions (uncertain and accurate), drawing less visual attention to these regions, thereby potentially improving the speed of error detection.

Keywords: Radiotherapy · Quality Assessment · Organs-at-Risk · Bayesian Uncertainty · Deep learning · AvU Loss · Uncertainty-ROC · FlipOut

1 Introduction

Radiotherapy for cancer treatment requires one to acquire diagnostic scans like CT and contour the boundaries of tumors and organs-at-risk (OARs). This process is time-consuming, prone to human error as well as inter-and intra-annotator disagreement [5,27]. For head-and-neck CT scans, these issues are further exacerbated due to the large OAR count (~ 35) and a lack of soft-tissue contrast.

Although deep learning has made great leaps in auto-contouring of tumors and OARs [26], the predicted contours still need to be manually verified before treatment. In this paper we investigate the use of Bayesian uncertainty heatmaps to help speed up this quality assessment (QA) by directing visual attention of clinicians to regions potentially containing contouring errors. Specifically, to enable faster error detection, we improve upon literature by incentivizing deep Bayesian models to produce uncertainty heatmaps only in inaccurate and not in accurate regions.

There exists a large body of work on uncertainty estimation for medical image segmentation. Some show that uncertainty heatmaps correspond to erroneous regions [24,23] indicating their potential to be used during autocontouring QA. Others investigate loss functions (c.f. Dice vs cross-entropy) [15,24], uncertainty metrics (e.g entropy, standard deviation) [20,6,23] or the use of uncertainty for error detection [23], contour refinement [25,2] and training data sampling [12]. The aforementioned works design for epistemic (or model) uncertainty which represents the variation in outputs, given an input. Conversely, other works design for aleatoric (or data) uncertainty [18,4,11] for e.g. inter- and intra-annotator disagreement, a phenomenon common in medical image segmentation. To the extent of our knowledge, Bayesian approaches to medical image segmentation have only explored the direct use of uncertainty and have not attempted to influence its nature. Our approach instead trains Bayesian segmentation models to produce both accurate contours along with uncertainty present only in inaccurate regions. This can potentially speed up autocontouring QA by ensuring that clinicians are not distracted by the uncertainty in accurate regions.

To ensure that accurate regions are certain and inaccurate regions are uncertain, we are inspired by [13] and their use of the Accuracy-vs-Uncertainty (AvU) metric in image classification tasks. Specifically, the AvU metric measures the ratio of the sum of accurate and certain (n_{ac}) and inaccurate and uncertain (n_{iu}) voxels to the total number of voxels (N). Our contribution is to use this metric as a loss term to improve the clinical utility of uncertainty heatmaps by providing a higher signal-to-noise ratio for error detection in a medical image segmentation context. In addition, we propose a loss term to specifically reduce uncertainty in accurate regions, as these regions are largest and may play a major role in influencing the visual attention of clinicians during autocontouring QA. Unlike [13], we maximize AvU across a range of uncertainty thresholds and evaluate our approach using the uncertainty-ROC metric [16].

2 Method

2.1 Dataset and Neural Architecture

We used two public datasets of CT scans of the head-and-neck region which were annotated with 9 organs-at-Risk (OARs). The MICCAI 2015-Head and Neck Segmentation Challenge dataset [22] contains 33 training, 5 validation and 10 test samples. The DeepMindTCIA dataset [21], containing 15 patients, is used as an independent test set.

Building upon literature, we use OrganNet2.5D [7], a non-Bayesian model as our base neural architecture. It follows the encoder-decoder design and uses a combination of both 2D-only and 3D-only convolutions. Its middle layers use dilated convolutions to obtain a sufficient receptive field since it only performs two down-sampling steps to maintain resolution for the smaller optic organs. Inspired by earlier work [1], we add Bayesian layers in the middle of our network.

To perform training for Bayesian models, a prior on the weights is assumed and updated to a posterior using the available dataset. For inference, we perform Monte-Carlo sampling of the posterior weights to estimate the output distribution as follows:

$$p(y|x, D) = \mathbb{E}_{W \sim p(W|D)}[p(y|x, W)]. \tag{1}$$

Here, x is an input mapped to an output y, $p(y|x, D)$ is the output distribution and $p(W|D)$ is the posterior [3] used to sample model weights W. We use the FlipOut technique [28], a form of variational inference in Bayesian neural nets that enables GPU-efficient sampling of weights from the posterior. Here, the prior is assumed to be a Gaussian factorizable across the Bayesian layers which is initialized with zero mean and identity covariance. FlipOut with a Gaussian prior was chosen over methods like Dropout [9] or DropConnect [16] since they use a Bernoulli distribution that may not be as representative of the neural net weight distribution when compared to a Gaussian. Unlike OrganNet2.5D, we use a lower count of filters in our model to be able to efficiently perform training and inference on its Bayesian version.

2.2 Losses

Segmentation Loss: In 3D segmentation, for each OAR class ($c \in C$) the model produces 3D probability maps (P_c) where each voxel (i) is represented by a probability vector summing to 1. We use the Cross Entropy (L_{CE}) loss and the Dice loss (L_{Dice}) on each P_c to learn organ geometry as also done in [17].

Accuracy-vs-Uncertainty (AvU) Loss: After prediction, each voxel has two properties – accuracy and uncertainty. Uncertainty is calculated on the output distribution $p(y|x, D)$. We chose predictive entropy, a commonly used uncertainty statistic, as it is capable of capturing both epistemic and aleatoric uncertainty [8]. Here, entropy represents the average amount of ambiguity present in the OAR probability vector P_c^i of each voxel i in the probability map P_c and is calculated as shown in [17]. In this work, we use the normalized entropy calculated by dividing the entropy by $\ln |C|$. Each voxel then belongs to one of four categories: n_{ac}, n_{au}, n_{ic} and n_{iu}, where n stands for the number of voxels, a for accurate, i for inaccurate, c for certain and u for uncertain. For QA, it is desirable to have a high n_{ac} when compared to n_{au} to prevent clinicians from spending time investigating accurate regions as well as high n_{iu} when compared to n_{ic} to prevent omission of errors. This requirement leads to the formulation

of the AvU metric [19]:

$$\text{AvU} = \frac{n_{\text{ac}} + n_{\text{iu}}}{n_{\text{ac}} + n_{\text{au}} + n_{\text{ic}} + n_{\text{iu}}}, \tag{2}$$

with a range between [0,1]. To maximize AvU, we follow [13] and minimize the negative logarithm of the AvU metric which uses a differentiable version of n_{ac}, n_{au}, n_{ic} and n_{iu}. This loss term is minimal when all accurate voxels are certain and inaccurate voxels are uncertain, i.e. $n_{\text{au}} = n_{\text{ic}} = 0$. While [13] applies the AvU loss to each image in a classification task, for organ segmentation we apply it on a dilated region around the ground truth and predicted mask since the background usually has low error as well as low uncertainty. In addition, rather than using a fixed uncertainty threshold calculated by the average uncertainty on a held-out validation set, we instead propose penalizing the AvU metric across a range of uncertainty thresholds ($t \in T$) and average their AvU loss values:

$$L_{\text{AvU}} = \frac{1}{T} \sum_{t=1}^{T} \ln \left(1 + \frac{n_{\text{au}}^t + n_{\text{ic}}^t}{n_{\text{ac}}^t + n_{\text{iu}}^t} \right). \tag{3}$$

***p*(*u*|*a*) Loss:** In practice, inaccuracies are usually present along the contour and not in the core. To avoid unnecessary visual inspection, it is desirable to have low uncertainty in the core of such organs. Thus, we investigate an additional loss on the probability of uncertainty in accurate regions, $p(u|a)$:

$$L_{\text{p(u|a)}} = \frac{1}{T} \sum_{t=1}^{T} \ln \left(1 + \frac{n_{\text{au}}^t}{n_{\text{ac}}^t + n_{\text{au}}^t} \right). \tag{4}$$

This loss is at its minimum when n_{ac} is 0. Thus, the final model objective is:

$$L = L_{\text{CE}} + L_{\text{Dice}} + \alpha \cdot L_{\text{AvU}} + \beta \cdot L_{\text{p(u|a)}}. \tag{5}$$

2.3 Evaluation

As a first measure of evaluation, we evaluate AvU for each uncertainty threshold t in the full range of normalized entropy ($0 \leq t \leq 1$). Then a single Area-under-the-curve (AUC) score is computed for each model using the AvU scores across a range of uncertainty thresholds. However, the AvU score compresses information of all voxels in a single value and does not allow to evaluate uncertainty separately in accurate and inaccurate regions. For faster radiotherapy contour QA, we need high probability of uncertainty in inaccurate regions – $p(u|i)$, and low probability of uncertainty in accurate regions – $p(u|a)$. Let us plot $p(u|i)$ on the y-axis and $p(u|a)$ on the x-axis of a graph and also define n_{iu} as True Positive (TP), n_{au} as False Positive (FP), n_{ac} as True Negative (TN) and n_{au} as False Negative (FN). This makes $p(u|i)$ the True Positive Rate and $p(u|a)$ the False Positive Rate, essentially giving us the commonly used Receiver Operating Characteristic (ROC) curve. We dub this measure as the uncertainty-ROC curve [16].

Calculation of the AUC would provide us with insight into whether the additional AvU loss has been useful. In this work, the $p(u|i)$ and $p(u|a)$ values plotted on the graph are the average across all test samples calculated using a discretized set of uncertainty values. Instead of uncertainty-ROC, [13] uses the Uncertainty Calibration Error (UCE) [14] metric to evaluate the effect of the AvU loss. This metric motivated by model trustworthiness metric Expected Calibration Error (ECE) [10], requires a normalized uncertainty value of $x \in [0, 1]$ to also provide an error percentage of the same value. We believe that this approach of treating the scalar value of uncertainty as a proxy for error percentage is not the correct approach. Finally, we also report ECE where a high score would indicate that the model, on average, produces high confidence probability estimates (i.e. low entropy), even for inaccurate predictions.

Due to inter-observer variation common in radiotherapy contouring [5], we consider voxels with "tolerable" errors within the inaccuracy map as accurate since they do not require clinical intervention [23]. Two morphological operations on the inaccuracy map i.e. erosion (to remove) followed by dilation (to repair partially eroded error regions) help us output an error map consisting only of segmentation failures that require clinical intervention.

3 Experiments and Results

3.1 Experimental Details

For our data, we ensure homogeneity by resampling all CT volumes to a resolution of (0.8, 0.8, 2.5) mm. As is commonly done in radiotherapy anatomical contouring, we trim the Hounsfield units of the CT scan from -125 to +225 for improved contrast of soft tissues. Finally, during training, random 3D patches of size 140 x 140 x 40 were extracted and augmented with 3D translations, 3D rotations, 3D elastic deformation and Gaussian noise.

We compare the original FlipOut model to the one with AvU loss (FlipOut-A) and the one with the AvU and $p(u|a)$ loss (FlipOut-AP), by training them on the MICCAI2015 training set and evaluating on the MICCAI2015 and DeepMindTCIA test sets. Additionally, the FlipOut-A(t1), FlipOut-A(t2) and FlipOut-A models use the following thresholds (Eq. (3)): the median, the mean [13] and the 25^{th} to 75^{th} percentile range in steps of 0.05%, of the uncertainty values in the MICCAI2015 validation set.

All our models are trained for 1000 epochs, with the FlipOut-A and Flipout-AP models having their first 50 epochs dedicated to segmentation losses alone, so that they can learn the geometry of individual organs-at-risk prior to tuning their uncertainty. They are further trained till a 1000 epochs when the KL-divergence between the posterior $p(W|D)$ and prior $p(W)$ has stabilized. The loss balancing terms $\alpha = 100$ and $\beta = 100$ (Eq. (5)) are experimentally determined from the training set $\{1, 10, 100\}$ such that the volumetric and surface measures of the newer model are either better or equivalent to the base model. For training we use a fixed learning rate of 10^{-3} with the

Table 1: Comparing Bayesian models for the MICCAI2015 (MIC2015) and DeepMindTCIA (DMTCIA) datasets. † and ‡ represents statistical significant difference ($p < 0.05$) when compared to FlipOut and FlipOut-A respectively. The mean and standard deviation are calculated across all patients.

Dataset	Model	HD95(mm)	ECE (x10^{-2})	AvU (x10^{-1})	unc-ROC (x10^{-1})
MIC2015	FlipOut	3.3±0.4	7.3±0.6	8.5±0.1	6.5±0.2
	FlipOut-A(t1)	3.1±0.3	6.5±0.7†‡	8.7±0.5†	6.7±0.6†‡
	FlipOut-A(t2)	3.2±0.4	6.4±0.2†‡	8.7±0.6†	6.7±1.0†‡
	FlipOut-A	3.1±0.4	5.8±0.6†	8.9±0.1†	6.9±0.2†
	FlipOut-AP	3.2±0.3	5.8±1.0†	9.0±0.1†	6.9±0.3†
DMTCIA	FlipOut	4.4±1.4	8.2±1.1	8.1±0.2	6.1±0.3
	FlipOut-A(t1)	4.5±0.9	8.3±0.5	8.2±0.9†	6.4±0.7†
	FlipOut-A(t2)	4.4±1.2	8.1±0.5	8.3±0.5†	6.4±0.2†
	FlipOut-A	4.4±1.3	8.2±0.7	8.5±0.3†	6.5±0.3†
	FlipOut-AP	4.5±1.4	8.2±0.9	8.6±0.2†	6.6±0.1†

Adam optimizer. For output distribution estimation, we perform 5 and 30 Monte Carlo sampling steps during training and inference respectively. During training and evaluation, we identify tolerated errors by using a kernel size corresponding to (2.4,2.4,2.5) mm. Code is implemented using Tensorflow version 2.4 on an Nvidia V100 (16 GB). Code to reproduce results is available on https://github.com/prerakmody/hansegmentation-uncertainty-errordetection.

3.2 Results and Analysis

Table 1 shows 95$^{\text{th}}$ percentile of Hausdorff Distance (HD95), Expected Calibration Error (ECE) [10], and AUC (Area-under-the-curve) scores for the AvU and uncertainty-ROC (unc-ROC) curves along with their statistical significances, using a Wilcoxon signed-rank test. There are no significant differences between the HD95 scores. For the internal MICCAI2015 dataset, the results show that using even a single threshold in the AvU loss causes a significant decrease in the ECE score, thus leading to more trustworthy probability estimates while a range of thresholds further improves calibration performance. This is in line with the results observed in [13], but the same does not hold for the external dataset. We also observe that upon using the AvU (Eq. (3)) and $p(u|a)$ (Eq. (4)) loss, AUC scores for the AvU and uncertainty-ROC curves have significant improvements in the uncertainty outputs over FlipOut (Figure 1). Compared to FlipOut-A(t1) and FlipOut-A(t2), FlipOut-A has significantly better unc-ROC scores for the MICCAI dataset, but not for the DeepMindTCIA dataset. Finally, compared to FlipOut-A, FlipOut-AP increases the AvU score slightly, at similar uncertainty-ROC scores.

In Figure 2, the rows depict CT slices with varying "qualities" of uncertainty maps. The first row shows the mandible bone with both the models having near-perfect contour predictions. Despite this, the FlipOut model still exhibits

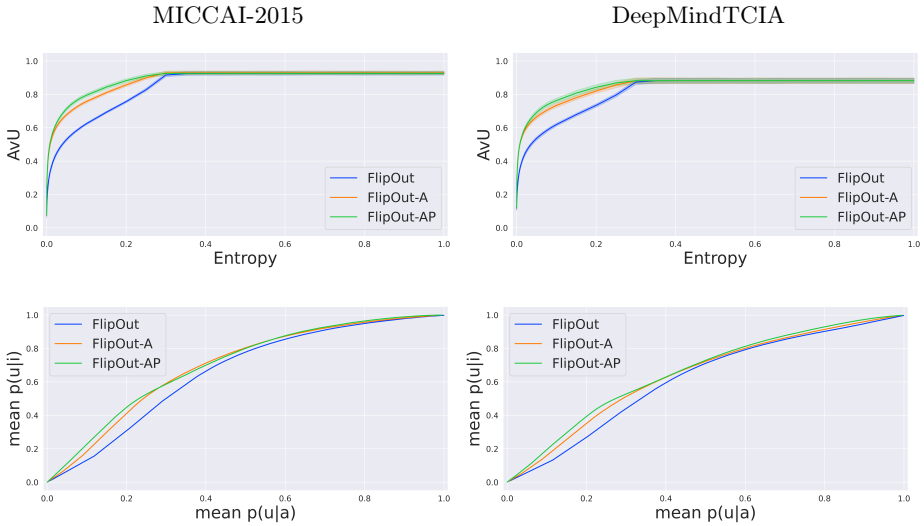

Fig. 1: The top row shows the AvU scores while the bottom row shows the uncertainty-ROC curve for the MICCAI2015 (left) and DeepMindTCIA (right) datasets respectively.

uncertainty in the core of the mandible bone even though there are no contrast issues. Conversely, the other models show uncertainty only on the contour of the prediction since uncertainty in accurate voxels has been suppressed leading to less visual attention in those regions. The second row shows the left parotid gland with all models having acceptable contours with minor errors (and associated uncertainty) in the medial lobe. The differences in uncertainty lies in the white blob (i.e. a vein) in the core of the gland as well as its lateral boundaries. While the FlipOut model shows uncertainty on both, the FlipOut-A model shows uncertainty only on the white blob and the FlipOut-AP model does not show uncertainty in both due to the use of the $p(u|a)$ loss. In the third row, we show the top-most slice of the brainstem with the FlipOut and FlipOut-AP model showing uncertainty along the predicted contours and the FlipOut-A model showing uncertainty in the core region. Thus, although there may not be significant differences in the uncertainty-ROC metric between the FlipOut-A and Flipout-AP models (Figure 1), the Flipout-AP may still be useful in certain scenarios. Note here that the slight contour differences are simply a result of the Monte-Carlo sampling as the overall models have similar geometric scores. In the fourth row, we again show the left parotid gland, but with larger errors on the medial lobe. In all cases, the models do not show any uncertainty in the erroneous region despite textural differences in the same. Such errors in contouring and omission in uncertainty may be attributed to the small size of our training dataset and such false negatives (in context of uncertainty) can be a potential blocker to adoption in the clinic.

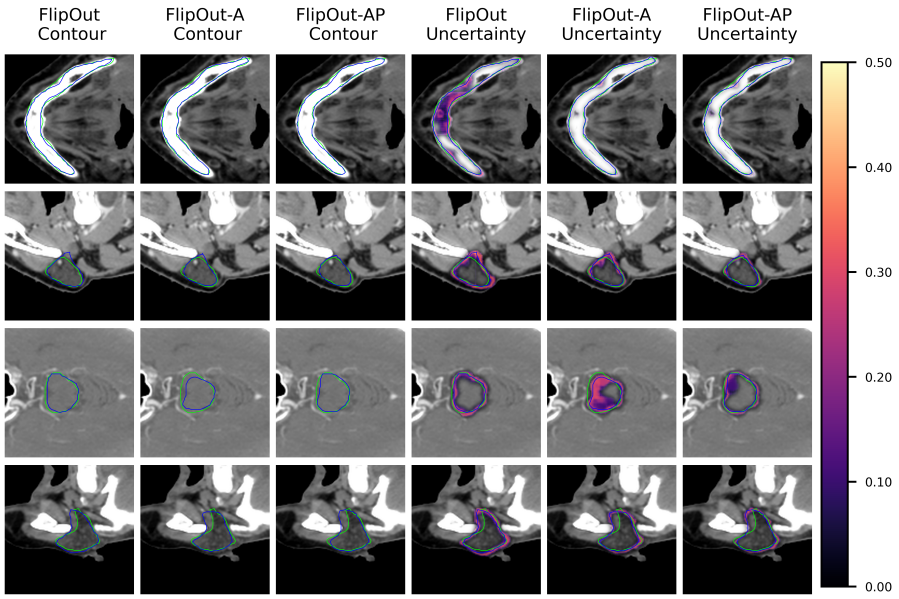

Fig. 2: Comparing the entropy heatmaps for the FlipOut, FlipOut-A and FlipOut-AP models. Here, we see the clinical (green) and predicted (blue) contours (columns 1-3) and the entropy heatmaps (columns 4-6).

4 Conclusion

This work investigates the use of the Accuracy-vs-Uncertainty (AvU) and $p(u|a)$ metrics as an additional objective in deep Bayesian modelling for improving error detection using uncertainty. Specifically, the proposed losses potentially enables faster error detection in radiotherapy contouring by motivating the model to produce accurate voxels which are certain and inaccurate voxels which are uncertain. This can assist clinicians during the mandatory quality assessment (QA) of autocontouring algorithms prior to radiation dosage calculation. We evaluate the effect of the AvU loss by using the uncertainty-ROC curve which shows that we improve the correlation of uncertainty with contouring inaccuracies. Our modified AvU loss does not require a manual choice of the uncertainty threshold, and improves the uncertainty-ROC metric on both an internal and external test dataset. We also explore an uncertainty-based loss specifically designed for the more abundant accurate regions, but found that although we observe visual differences, it does not lead to a significantly improved uncertainty-ROC. Future work will consider exploring the effects of a larger dataset and the utility of uncertainty from our proposed models by conducting trials with clinicians.

Acknowledgements The research for this work was funded by Varian, a Siemens Healthineers Company, through the HollandPTC-Varian Consortium (grant id 2019022) and partly financed by the Surcharge for Top Consortia for Knowledge and Innovation (TKIs) from the Ministry of Economic Affairs and Climate, The Netherlands.

References

1. Alex Kendall, V.B., Cipolla, R.: Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. In: Proceedings of the British Machine Vision Conference (BMVC). BMVA Press (2017)
2. Arega, T.W., Bricq, S., Meriaudeau, F.: Leveraging uncertainty estimates to improve segmentation performance in cardiac mr. In: Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Perinatal Imaging, Placental and Preterm Image Analysis, pp. 24–33. Springer (2021)
3. Blundell, C., Cornebise, J., Kavukcuoglu, K., Wierstra, D.: Weight uncertainty in neural network. In: International conference on machine learning. PMLR (2015)
4. Bragman, F.J., Tanno, R., Eaton-Rosen, Z., Li, W., Hawkes, D.J., Ourselin, S., Alexander, D.C., McClelland, J.R., Cardoso, M.J.: Uncertainty in multitask learning: joint representations for probabilistic mr-only radiotherapy planning. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 3–11. Springer (2018)
5. Brouwer, C.L., Steenbakkers, R.J., van den Heuvel, E., Duppen, J.C., Navran, A., Bijl, H.P., Chouvalova, O., Burlage, F.R., Meertens, H., Langendijk, J.A., et al.: 3D variation in delineation of head and neck organs at risk. Radiation Oncology **7**(1), 1–10 (2012)
6. Camarasa, R., Bos, D., Hendrikse, J., Nederkoorn, P., Kooi, E., Lugt, A.v.d., Bruijne, M.d.: Quantitative comparison of monte-carlo dropout uncertainty measures for multi-class segmentation. In: Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Graphs in Biomedical Image Analysis, pp. 32–41. Springer (2020)
7. Chen, Z., Li, C., He, J., Ye, J., Song, D., Wang, S., Gu, L., Qiao, Y.: A novel hybrid convolutional neural network for accurate organ segmentation in 3d head and neck ct images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 569–578. Springer (2021)
8. Gal, Y.: Uncertainty in deep learning. PhD Thesis, University of Cambridge (2016)
9. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: International conference on machine learning. pp. 1050–1059. PMLR (2016)
10. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International Conference on Machine Learning. pp. 1321–1330. PMLR (2017)
11. Islam, M., Glocker, B.: Spatially varying label smoothing: Capturing uncertainty from expert annotations. In: International Conference on Information Processing in Medical Imaging. pp. 677–688. Springer (2021)
12. Iwamoto, S., Raytchev, B., Tamaki, T., Kaneda, K.: Improving the reliability of semantic segmentation of medical images by uncertainty modeling with bayesian deep networks and curriculum learning. In: Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Perinatal Imaging, Placental and Preterm Image Analysis. Springer (2021)

13. Krishnan, R., Tickoo, O.: Improving model calibration with accuracy versus uncertainty optimization. Advances in Neural Information Processing Systems (2020)
14. Laves, M.H., Ihler, S., Kortmann, K.P., Ortmaier, T.: Well-calibrated model uncertainty with temperature scaling for dropout variational inference. arXiv preprint arXiv:1909.13550 (2019)
15. Mehrtash, A., Wells, W.M., Tempany, C.M., Abolmaesumi, P., Kapur, T.: Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. IEEE transactions on medical imaging **39** (2020)
16. Mobiny, A., Yuan, P., Moulik, S.K., Garg, N., Wu, C.C., Van Nguyen, H.: Dropconnect is effective in modeling uncertainty of bayesian deep networks. Scientific reports **11**(1), 1–14 (2021)
17. Mody, P.P., de Plaza, N.C., Hildebrandt, K., van Egmond, R., de Ridder, H., Staring, M.: Comparing Bayesian models for organ contouring in head and neck radiotherapy. In: Medical Imaging 2022: Image Processing. International Society for Optics and Photonics, SPIE (2022)
18. Monteiro, M., Le Folgoc, L., Coelho de Castro, D., Pawlowski, N., Marques, B., Kamnitsas, K., van der Wilk, M., Glocker, B.: Stochastic segmentation networks: Modelling spatially correlated aleatoric uncertainty. Advances in Neural Information Processing Systems **33**, 12756–12767 (2020)
19. Mukhoti, J., Gal, Y.: Evaluating bayesian deep learning methods for semantic segmentation. CoRR **abs/1811.12709** (2018)
20. Nair, T., Precup, D., Arnold, D.L., Arbel, T.: Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation. Medical image analysis **59**, 101557 (2020)
21. Nikolov, S., Blackwell, S., Zverovitch, A., Mendes, R., Livne, M., De Fauw, J., Patel, Y., Meyer, C., Askham, H., Romera-Paredes, B., et al.: Clinically applicable segmentation of head and neck anatomy for radiotherapy: deep learning algorithm development and validation study. Journal of Medical Internet Research **23**(7), e26151 (2021)
22. Raudaschl, P.F., Zaffino, P., Sharp, G.C., Spadea, M.F., Chen, A., Dawant, B.M., Albrecht, T., Gass, T., Langguth, C., Lüthi, M., et al.: Evaluation of segmentation methods on head and neck CT: Auto-segmentation challenge 2015. Medical physics **44**(5), 2020–2036 (2017)
23. Sander, J., de Vos, B.D., Išgum, I.: Automatic segmentation with detection of local segmentation failures in cardiac MRI. Scientific Reports **10** (2020)
24. Sander, J., de Vos, B.D., Wolterink, J.M., Išgum, I.: Towards increased trustworthiness of deep learning segmentation methods on cardiac MRI. In: Medical imaging 2019: image Processing. vol. 10949, pp. 324–330. SPIE (2019)
25. Soberanis-Mukul, R.D., Navab, N., Albarqouni, S.: Uncertainty-based graph convolutional networks for organ segmentation refinement. In: Medical Imaging with Deep Learning. pp. 755–769. PMLR (2020)
26. Van Dijk, L.V., Van den Bosch, L., Aljabar, P., Peressutti, D., Both, S., Steenbakkers, R.J., Langendijk, J.A., Gooding, M.J., Brouwer, C.L.: Improving automatic delineation for head and neck organs at risk by deep learning contouring. Radiotherapy and Oncology **142**, 115–123 (2020)
27. van der Veen, J., Gulyban, A., Nuyts, S.: Interobserver variability in delineation of target volumes in head and neck cancer. Radiotherapy and Oncology **137** (2019)
28. Wen, Y., Vicol, P., Ba, J., Tran, D., Grosse, R.: Flipout: Efficient pseudo- independent weight perturbations on mini-batches. In: Proceedings of the 6th International Conference on Learning Representations (2018)