

# XBoundNet++: Uncertainty-Aware Segmentation of Kidney Ablation Zones

Oren Arbel-Wood<sup>1</sup>, Maryam Rastegarpour<sup>1</sup>, Aaron Fenster<sup>1</sup>

<sup>1</sup> Robarts Research Institute, Western University, London, Canada  
oarbelwo@uwo.ca, mrasteg2@uwo.ca, afenster@uwo.ca

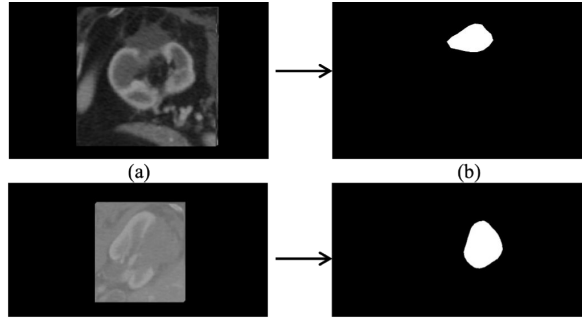
**Abstract.** Kidney ablation therapy is a minimally invasive procedure used to treat renal tumours. Evaluating treatment success for planning follow-up care relies on accurate kidney ablation zone (KAZ) segmentation in post-operative CT images. However, manual segmentation is time-consuming and prone to inter-observer variability and traditional segmentation is challenging as ground truth labels only provide a partial estimate of the area of interest. Segmenting the area of interest requires careful attention to the specific clinical needs of the resulting deep learning framework, including the addition of model interpretability and uncertainty estimation for further clinical review. We introduce a deep learning framework, XBoundNet++, that permits (1) precise segmentation of the boundary, (2) detailed attention maps for model layer-wise interpretability, and (3) model uncertainty estimation based on Bayesian Monte-Carlo dropouts and model ensembles. The model was trained and evaluated using a nested 5-fold cross-validation on a local dataset of 76 patients (with 912 CT 2D radial slices), collected at London Health Sciences Centre, which included manually annotated KAZs. Quantitative analysis showed that XBoundNet++ achieved promising segmentation results, including 88% precision, 83% recall, 84% DSC, 74% Jaccard, 6.89-pixel Mean Absolute Distance (MAD), -0.60-pixel Mean Signed Distance (MSD), and a 19.86-pixel Hausdorff distance (HD). Furthermore, heatmaps at each layer, probability and uncertainty maps, and uncertainty estimation at several thresholds indicate model trustworthiness, confidence, and justification for predictions. Our codebase can be found at <https://github.com/oarbelw/XBoundNetPlusPlus> and the dataset will be available upon request.

**Keywords:** Kidney Ablation, Segmentation, Interpretability, Uncertainty, Deep Learning, CT images

## 1 Introduction

Many segmentation methods perform well for known structures (e.g., kidney, liver in CT) [2, 17], as well as pathological structures (e.g., brain tumours) [18]. However, supervised learning requires the ground truth labels for objects of interest for training, but there are commonly contexts in which ground-truth labels are generally challenging to obtain. Furthermore, additional challenges are presented in these types of clin-

ical contexts as they typically consist of a limited number of cases. These contexts are not well studied in the literature and require particular care in terms of providing the clinician with model transparency and model uncertainty in order to trust the results and review the areas of relevance. This is crucial to enable trustworthy AI that aligns with established standards [14], helping with:



**Fig. 1.** Two sample patient images from the dataset, where (a) are raw images, and (b) clinically annotated images.

In this paper, we consider post-treatment delineation of the ablation zone in kidney CT images. Kidney cancer, or renal cell carcinoma, is one of the most prevalent urological malignancies worldwide. For patients unfit for surgical intervention, thermal ablation therapies like microwave or radiofrequency ablation offer a minimally invasive alternative. These procedures aim to destroy malignant cells by creating a “kidney ablation zone” (KAZ) that encapsulates the tumour and surrounding margin. Post-treatment assessment depends on accurately identifying the entirety of the KAZ in follow-up CT scans, which is a critical task for determining treatment success and guiding subsequent care [6].

The integration of uncertainty-aware deep learning techniques would be important in medical imaging domains involving ambiguous or low-contrast boundaries, such as post-ablation regions. However, to date, no deep learning models have been developed and published for this task.

Uncertainty estimation in deep learning models has been developed to help identify areas where predictions may be unreliable due to image ambiguity, label noise, or model uncertainty. A number of specific deep learning architectures were proposed to explicitly predict segmentation labels along with probabilistic outcomes, such as the Probabilistic U-Net [13] and PHISeg [1]. However, these models require architecture modification built specifically for particular kinds of uncertainty measures and are not easily adaptable to any context, for example, models that also embed attention modules within the framework. Other models were developed such as PULASKi [3], a model that explicitly models inherent ambiguity arising from expert disagreement, and the recent Stochastic Segmentation Network (SSNs) [19], which specifically captures the aleatoric uncertainty seen in medical images.

In their seminal paper, Kendall and Gal [12] introduced a framework that models both aleatoric and epistemic uncertainty using Bayesian deep learning (BDL)

techniques, which has become the foundation for uncertainty estimation in segmentation. BDL along with MC Dropout and deep ensembles permit a simple and effective mechanism for post hoc uncertainty estimates in several forms. The beauty of BDL models is that they can be used to estimate epistemic and aleatoric uncertainties through sampling techniques during inference for ANY network with dropout layers. As such, they are flexible and can be added to any network. They have been used in brain tumor segmentation [7, 15] and lesion segmentation [20]. In the context of kidney imaging, MC dropout has been used [25] in kidney tumors and cysts segmentation. While most segmentation studies [26] have focused on kidney tumors (e.g., the KiTS21 and KiTS23 datasets), segmentation of ablation zones, which present irregular, low contrast boundaries post-procedure, remains unexplored.

In this work, we introduce an *XBoundNet++*, an eXplainable Boundary-Aware modified ResU-Net++, a novel deep learning segmentation framework designed to provide clinicians with high-quality segmentation results, model transparency, interpretable tools, and uncertainty estimation using Bayesian Monte-Carlo (MC) dropout [8]. Our framework is aimed at shifting clinical practice from unclear binary masks to interpretable tools that explicitly provide confidence, uncertainty, probability, and transparency. We created an end-to-end pipeline to preprocess an image (as seen in **Fig. 1**), feed it into our model, generate segmentations of high quality that outperform other state-of-the-art models, provide comprehensive layer-wise transparency, and produce epistemic-uncertainty with probability maps.

## 2 Methods

### 2.1 XBoundNet++ Segmentation Network and Training

We propose XBoundNet++, an ensemble-based four-level modified U-Net [23] in **Fig. 2**, which introduces architectural elements that explicitly promote feature relevance, spatial focus, and post-hoc transparency. Our architecture integrates components from LeXNet++ [5], ResNet [9], attention mechanisms [21], Squeeze & Excitation (SE) [10], STEM [22], and several advanced architectures.

The Atrous Spatial Pooling Pyramid (ASPP) bridge, connecting the encoder and decoder, captured multi-scale context while maintaining dimensionality. It applied convolutions with dilations of 1, 6, 12, and 18 [4], performed a summation to merge the features, and applied a BN and ReLU activation.

Attention Gate blocks were introduced to selectively propagate relevant features during upsampling. They compute spatial attention maps via  $1 \times 1$  convolutions and ReLU-sigmoid activation, suppressing irrelevant activations and enhancing decoder focus on the ablation zone.

The network was optimized with Adam (learning rate  $10^{-4}$ , batch size 4) and a custom combined loss of Log-Dice ( $\alpha = 0.7$ ) and binary cross-entropy ( $\alpha = 0.3$ ), as Dice addresses class imbalance, while BCE improves per-pixel calibration, giving probabilistic outputs that can be further used for uncertainty estimation. Early stopping

(patience = 50) and Reduce-LR-on-Plateau (factor 0.1, patience = 15) were employed to prevent over-fitting and facilitate convergence. The final combined loss is:

$$\text{CombinedLoss}(y, \hat{y}) = \alpha * \text{LogDiceLoss}(y, \hat{y}) + (1 - \alpha) * \text{BCE}(y, \hat{y}) \quad (1)$$

where  $y$  is the ground truth,  $\hat{y}$  is the prediction, and  $\alpha$  is set to 0.7.

For each of the five patient-wise folds, we trained five instances of XBoundNet++ with differing seeds, yielding 25 independent models in total. Altering the seed affects weight initialization, alters the stochastic augmentation stream, and changes the sequence of dropout masks encountered during optimization. The resulting ensemble enhances predictive stability, generalizes the small dataset, and forms the basis for the uncertainty analysis described in the next section.

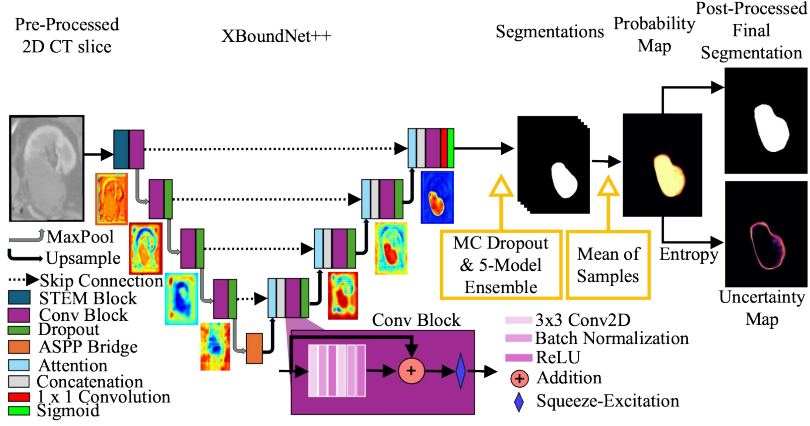


Fig. 2. Network pipeline and architecture with layer-wise activation maps.

## 2.2 Layer-wise Heatmap Generation

We propose a custom Gradient-weighted Class Activation Mapping (Grad-CAM) [24] method that helps visualize which regions of an image had the most influence on the model by analyzing gradient-weighted activations that serve as spatial attention maps. First, the image is processed by the model, a class prediction is made, and during back-propagation, the gradients of the prediction are computed for the feature map at the chosen convolutional layer. Finally, the results are passed through a ReLU activation (and upsampled if necessary) to produce the heatmap for any given convolutional layer in the network.

We then extract heatmaps from every convolutional block across the network to observe how feature abstraction evolves at different depths. This strategy allows us to visually trace the information flow and decision-making within the network, revealing where and how the network’s focus shifts, from low-level texture extraction to high-level semantic boundary recognition.

### 2.3 Model Inference and Uncertainty Estimation

During training, dropout mitigates over-fitting. During inference, the five seed-specific trained network models retained from each outer fold are evaluated with dropout kept active. For every unseen test slice, we generate 50 stochastic outputs using Bayesian MC dropout, yielding a collection of 250 predictions per slice. We then average this collection to create an ensemble-predictor, producing a probability map  $p$  for KAZ segmentation.

We use the probability map to generate an uncertainty map using normalized entropy,  $H$ , as a measure of uncertainty as portrayed below:

$$H = -[p \log(p) + (1 - p) \log(1 - p)]. \quad (2)$$

We pool the raw predictions of the validation slices and fit a one-dimensional logistic-regression calibrator. The fitted sigmoid is saved and applied to all test-set probabilities, producing a calibrated map. We then use a 0.4 threshold to binarize the prediction so that values are either 0 or 1. Next, we perform a morphological closing operation to seal small holes or gaps in the prediction if necessary. We also examined whether numerous disconnected components were present, in which case the largest foreground component is retained, and all other objects are suppressed. This didn't apply to instances where cysts are larger than the KAZ; in such a case, the second largest component is selected.

The cleaned final segmentation mask was then resized to the original CT image size ( $510 \times 788$  pixels) with Lanczos-4 interpolation and written to disk as an 8-bit BMP.

### 2.4 Evaluation Metrics

Standard pixel and distance-based metrics were used to assess both technical and clinical segmentation quality. These metrics included the Dice similarity coefficient (DSC), Precision, Recall, and Jaccard, which provide valuable quantitative insight on boundary overlap, precision, and quality of segmentation. Boundary accuracy was evaluated by the mean absolute distance (MAD), mean signed distance (MSD), and Hausdorff distance (HD), to quantify the comparative closeness and surface area.

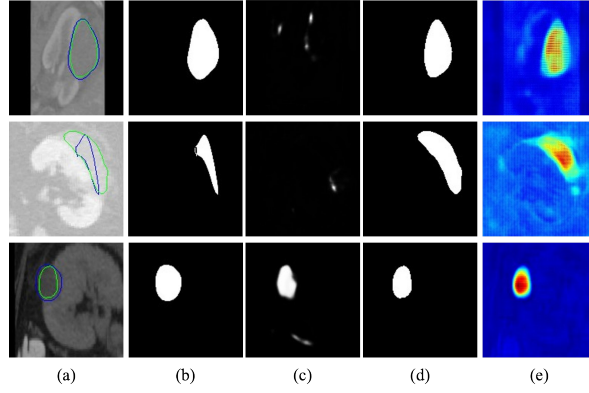
To validate the segmentation uncertainty estimations, we apply thresholds. This involved normalizing the entropy estimates per slice ranged from 0 and 100, and varying the thresholds ( $T = 25, 50, 75$ ) at different confidence levels as in [16]. Pixels exceeding the given threshold were labelled as uncertain and the remainder were cross-checked against the annotated mask to generate four disjoint classes: true positive (TP) (overlapping areas), false positive (FP) (over-prediction), false negative (FN) (under-prediction), and uncertain. As we lowered the uncertainty threshold, the FN and FP areas should have been filtered out while retaining the TP pixels. This validated that in areas where the model is confident, it is correct, while incorrect areas have high uncertainty. This permits clinical trust in areas of high model confidence. The entire spatial confidence map, along with the segmentation results, allows a framework for downstream clinical review.

### 3 Experiments and Results

#### 3.1 Patient Data, Preprocessing and Implementation Details

Our patient dataset was collected after approval by the Western University Research Ethics Board using a GE Lightspeed 64-slice CT scanner and included 76 patients' cases, each containing 12 axial CT slices obtained post-ablation. All the images were in DICOM format, grayscale, originally sized at  $510 \times 788$  pixels, and were accompanied by manually annotated binary masks of the KAZ, which were generated by an expert. The 3D CT images were resliced radially around an approximate vertical axis of the KAZ every  $15^\circ$  into 2D CT images. This transformation ensured that the zone appears more consistently across 2D image samples. Each slice was resized to  $256 \times 256$  pixels for computational feasibility to fit into the model, resulting in a dataset of 912 2D CT images, as shown in **Fig. 1**.

We normalized pixel intensities to a  $[0, 1]$  range, and split the dataset by patient into training (64%), validation (16%), and testing (20%) sets. This ensured that slices from the same patient did not appear in multiple subsets to avoid model bias. To enhance model reliability and reduce variance due to dataset partitioning, a nested 5-fold patient-wise cross-validation strategy was adopted. Each fold used a unique set of patients for training, validation, and testing, ensuring that no slices from a single patient were shared across splits.



**Fig. 3.** XBoundNet++ results for three image slices from three different patients, in each row a) Original image, showing **model prediction contour** and **clinical annotation**, b) Clinically annotated mask, c) LeXNET++ prediction, d) XBoundNet++ prediction, e) The prediction attention heatmap from the convolutional layer before the sigmoid is applied. Attention shows higher gradient activation in red and thus more involvement in the resulting prediction, as it is more confident in the centre and is less confident at the boundaries.

On-the-fly data augmentation was applied to expand the appearance diversity while preserving label fidelity to compensate for the relatively small dataset. Augmentations were executed in TensorFlow eager mode, so a new stochastic version of every training image was generated for each epoch without materializing augmented files on disk.

Each slice had a 30% chance of undergoing one or more spatial transformations: horizontal or vertical flip, translation of  $\pm 10\%$  of the image extent, rotation of  $\pm 20^\circ$ , or isotropic zoom between 0.9 and 1.1. Independently, there was a 30% chance of a photometric adjustment that scales contrast between 0.8 and 1.2.

### 3.2 Segmentation Results

**Fig. 3** shows the original image, mask, XBoundNet++ prediction, and the corresponding attention-based heatmap. These results show that the predictions generated by XBoundNet++ accurately align with the KAZ better than LeXNet++, as well as provide clarity on how strong the activations are that result in the arrival to the final prediction. This is evident in the first patient, where the KAZ, annotation mask, and model prediction all agree and cover the same area inside the kidney. While the second image may appear to be over-segmented, it is due to an incomplete annotation mask. The model correctly delineated the full ablation zone, outperforming the human annotation. The third prediction correctly under-segments, as the manual annotation extends beyond the actual KAZ and kidney region.

**Table 1.** Ablation analysis on different metrics in XBoundNet++, with the cumulative addition (+) of new components in descending order, highlighting the best result in grey.

Metrics	Precision	Recall	DSC	Jaccard	MAD (pixels)	MSD (pixels)	HD (pixels)
Models							
LeXNet++ Baseline	0.68 $\pm$ 0.33	0.54 $\pm$ 0.36	0.55 $\pm$ 0.34	0.45 $\pm$ 0.31	37.06 $\pm$ 57.94	27.44 $\pm$ 61.89	86.96 $\pm$ 119.89
+XBoundNet++	0.71 $\pm$ 0.20	0.68 $\pm$ 0.30	0.66 $\pm$ 0.27	0.54 $\pm$ 0.26	26.00 $\pm$ 41.83	17.59 $\pm$ 44.03	61.95 $\pm$ 66.28
+Augmentation	0.82 $\pm$ 0.18	0.80 $\pm$ 0.23	0.78 $\pm$ 0.19	0.76 $\pm$ 0.17	15.48 $\pm$ 26.10	8.19 $\pm$ 27.22	43.47 $\pm$ 55.10
+Post-Processing	0.81 $\pm$ 0.18	0.83 $\pm$ 0.19	0.81 $\pm$ 0.17	0.71 $\pm$ 0.18	12.47 $\pm$ 29.55	6.28 $\pm$ 30.76	28.29 $\pm$ 37.16
+CombinedLoss	0.84 $\pm$ 0.18	0.82 $\pm$ 0.19	0.82 $\pm$ 0.17	0.72 $\pm$ 0.17	10.54 $\pm$ 24.13	3.80 $\pm$ 25.39	24.83 $\pm$ 31.64
+Ensemble	<b>0.88<math>\pm</math>0.11</b>	<b>0.83<math>\pm</math>0.13</b>	<b>0.84<math>\pm</math>0.10</b>	<b>0.74<math>\pm</math>0.13</b>	<b>6.89<math>\pm</math>4.33</b>	<b>-0.60<math>\pm</math>5.90</b>	<b>19.86<math>\pm</math>12.40</b>

The results of the ablation study are shown in **Table 1** and were conducted to isolate the effect of each added XBoundNet++ component. Starting from the LeXNet++ [5] baseline, which lacks data augmentation, post-processing, and loss customization, we observe steady improvements across all metrics with each addition. XBoundNet++ alone improves DSC by 11%, recall by 14%, and HD by 25 pixels. Adding data augmentation further boosts DSC by 12%, Jaccard by 22%, and reduces MAD and MSD by over 9 pixels. Post-processing and the combined loss yield additional gains in boundary-related metrics, notably 3% DSC and a 15-pixel HD reduction. Finally, the ensemble improves all metrics, culminating in a 29% gain in DSC and Jaccard, 20% precision, and 67.1-pixel HD reduction compared to the baseline.

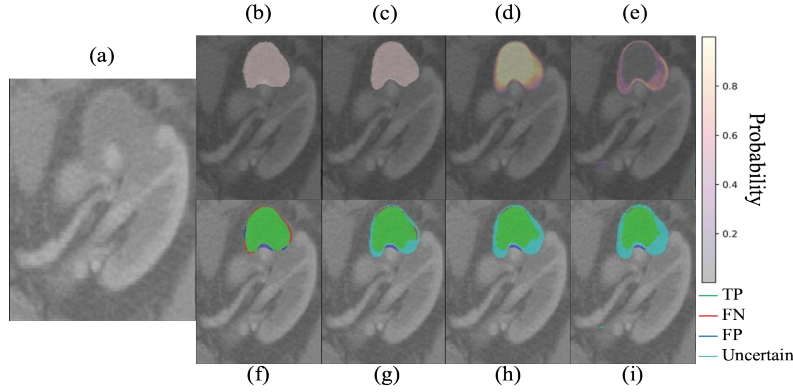
While these metrics demonstrate the quality of our proposed model, it is important to consider that there is no clear ground-truth in this application because KAZ boundaries are inherently ambiguous and the manually-drawn masks are subject to user variability. Thus, quantitative gains do not always capture the full clinical value (i.e., rows 2 and 3 in **Fig. 3**, where the model outperformed the annotation – based on post-hoc review).

### 3.3 Model Transparency

As visualized in **Fig. 2**, early convolutional layers tend to activate broadly across the ablation zone, while deeper layers increasingly emphasize peripheral boundary regions, particularly near ambiguous areas. The heatmaps clearly illustrate a transition from low-level texture detection in early layers to high-level semantic abstraction in deeper layers, confirming that the network progressively refines its attention toward clinically relevant boundaries. As a result, we can clearly track information flow and decision making, revealing the model’s focus, enabling trustworthy AI.

### 3.4 Uncertainty Analysis

**Fig. 4** illustrates a qualitative analysis of a given patients’ KAZ region and provides more insight for clinicians. The prediction doesn’t span over the healthy tissue at the bottom despite the manual annotation including it. The clinician can refer to the probability and uncertainty overlays to manually scrutinize areas with less confidence and higher uncertainty. The results show that decreasing the threshold leads to filtering out pixels of high uncertainty only.



**Fig. 4.** XBoundNet++ results, uncertainty, probability, and thresholding visualized over a patient’s CT slice. (a) CT original patient image slice, (b) Manually annotated mask, (c) XBoundNet++ predicted mask, (d) Probability map based on MC and ensembling, (e) Predicted entropy map from the probability map, (f) Uncertainty threshold = 100, (g) Uncertainty threshold = 75, (h) Uncertainty threshold = 50, (i) Uncertainty threshold = 25. It is desired that with more filtered out, more **False Positives** and **False Negatives** pixels are filtered out (marked **uncertain**), while **True Positive** pixels remain unfiltered.

## 4 Conclusions

In this work, we propose XBoundNet++, a novel deep learning segmentation framework that provides clinicians with several auxiliary interpretable and uncertainty tools to better equip them for clinically challenging contexts such as poor image contrast, no delineated boundary, or incomplete labels. The model excels at segmentation based on



several key metrics, provides in-depth transparency using Grad-CAM, and uncertainty estimation generated by Bayesian MC dropout and model ensembling. By offering transparency, spatial uncertainty, and probability overlays, XBoundNet++ enables more informed clinical review and supports safer, more trustworthy AI-assisted decision-making in interventional radiology.

The small dataset size, single-expert annotations, and use of 2D radial slices reflect common constraints in real-world clinical contexts. Rather than being limitations of the model, these challenges motivated our framework's design—tailored for clinically ambiguous labels and limited data. Standard models such as U-Net [23], while foundational in medical image segmentation, were not designed with uncertainty quantification or model transparency in mind. U-Net lacks mechanisms for epistemic or aleatoric uncertainty estimation, and offers no tools for layer-wise interpretability or confidence-guided clinical review—capabilities that are essential in interventional radiology. While adapting to 3D segmentation methods like nnU-Net [11] was not feasible in this setting as it requires large volumetric datasets and high-quality 3D annotations, future work will explore such extensions as larger, multi-center datasets and multi-rater annotations become available. We also aim to apply the model's uncertainty outputs to downstream tasks such as margin status, ablation volume, and residual tumour assessment.

**Acknowledgments.** The authors are grateful for funding from the Ontario Institute of Cancer Research (OICR) Grant RA#262 and the Canadian Institutes of Health Research (CIHR) – Grant FRN 154314.

**Disclosure of Interests.** The authors have no competing interests to declare that are relevant to the content of this article.

## References

1. Baumgartner, C.F., Koch, L.M., Tezcan, M., et al.: PHISeg: Capturing uncertainty in medical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 119–127. Springer, Cham (2019)
2. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In: MICCAI. pp. 424–432 (2016)
3. Chatterjee, S., Honchar, A., Seibold, M., et al.: PULASki: Learning inter-rater variability using statistical distances to improve probabilistic segmentation. *Med. Image Anal.* 85, 103623 (2025).
4. Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A. L.: DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* 40(4), 834–848 (2017)
5. Das, S., Khan, S.S., Sengupta, D., et al.: LeXNet++: Layer-wise eXplainable ResUNet++ framework for segmentation of colorectal polyp cancer images. *Neural Comput. Appl.* (2024). <https://doi.org/10.1007/s00521-024-10441-6>
6. Escudier, B., Porta, C., Schmidinger, M., et al.: Renal cell carcinoma: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann. Oncol.* 30(5), 706–720 (2019)

7. Fuchs, M., Gonzalez, C., Mukhopadhyay, A.: Practical uncertainty quantification for brain tumor segmentation. In: *Medical Imaging with Deep Learning (MIDL)* (2021)
8. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: representing model uncertainty in deep learning. In: *ICML*, pp. 1050–1059 (2016)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 770–778 (2016)
10. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: *IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 7132–7141 (2018)
11. Isensee, F., Jaeger, P.F., Kohl, S.A.A., et al.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods* 18, 203–211 (2021)
12. Kendall, A., Gal, Y.: What uncertainties do we need in Bayesian deep learning for computer vision? In: *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 20 (2017)
13. Kohl, S., Romera-Paredes, B., Meyer, C., et al.: A probabilistic U-Net for segmentation of ambiguous images. In: *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 31 (2018)
14. Lekadir, K., Frangi, A.F., Porras, A.R., Glocker, B., Cintas, C., Langlotz, C.P., et al.: FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *BMJ* 388:e081554 (2025)
15. Mehta, R., Paunovic, V., Arbel, T.: Propagating uncertainty across cascaded medical imaging tasks for improved deep learning inference. *IEEE Trans. Med. Imaging* 41(11), 3090–3102 (2022)
16. Mehta, R.: Integrating Bayesian deep learning uncertainties in medical image analysis. Ph.D. thesis, Dept. of Electrical & Computer Engineering, McGill University (2023)
17. Meine, H., Chlebus, G., Ghafoorian, M., Endo, I., Schenk, A.: Comparison of U-net based convolutional neural networks for liver segmentation in CT. *J. Intell. Fuzzy Syst.* 83, 71833–71862 (2024)
18. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., et al.: The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Trans. Med. Imaging* 34(10), 1993–2024 (2015)
19. Monteiro, M., Allken, V., Wang, B., Jacob, M.W.: Stochastic segmentation networks: Modelling spatially correlated aleatoric uncertainty. In: *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 12756–12767 (2020)
20. Nair, T., Chen, L., Yang, C., Precup, D.: Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation. *Med. Image Anal.* 59, 101557 (2020)
21. Oktay, O., Schlemper, J., Le Folgoc, L., et al.: Attention U-Net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999* (2018). <https://arxiv.org/abs/1804.03999>
22. Rastegarpour, M., Cool, D.W., Fenster, A.: Segmentation of kidney ablation zone using deep learning in CT images. In: *Proc. SPIE*, vol. 13406, San Diego, USA (2025)
23. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. In: *MICCAI*, pp. 234–241. Springer, Cham (2015)
24. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: Visual explanations from deep networks via gradient-based localization. *Int. J. Comput. Vis.* 128(2), 336–359 (2019)
25. Salahuddin, Z., Zhang, T., Wang, Y., Salama, M.S.: Leveraging uncertainty estimation for segmentation of kidney, kidney tumor and kidney cysts. In: *International Challenge on Kidney and Kidney Tumor Segmentation*, pp. 40–46. Springer, Cham (2023)
26. Sudre, C.H., Dalca, A., Baumgartner, C.F.: Uncertainty for safe utilization of machine learning in medical imaging. In: *MICCAI UNSURE Workshop* (2022)