Rel-UNet: Reliable Tumor Segmentation via Uncertainty Quantification in nnUNet

Seyed Sina Ziaee[1], **Katie Ovens**[1], **Farhad Maleki**[1]

[1] Department of Computer Science, University of Calgary, Canada

Abstract

Accurate and reliable tumor segmentation is crucial in medical imaging for enhancing cancer diagnosis, treatment planning, and monitoring. However, existing segmentation models often need more robust mechanisms for quantifying the uncertainty associated with their predictions, which is essential for supporting clinical decision-making. This study presents a novel approach for uncertainty quantification in kidney tumor segmentation using deep learning, specifically leveraging multiple local minima during training. Our method generates uncertainty maps without altering the original model architecture or requiring extensive computational resources. We evaluated our approach on the KiTS23 dataset, which effectively identified ambiguous regions and improved segmentation accuracy. The generated uncertainty maps provide critical insights into model confidence, ultimately enhancing the reliability of the segmentation and aiding in more accurate medical diagnoses. Our approach offers a valuable tool for integrating uncertainty quantification into clinical workflows, with potential applications across various medical image segmentation tasks.

Code — https://github.com/sinaziaee/sgdr_nnunet

Introduction

Medical imaging plays a crucial role in the diagnosis, treatment planning, and monitoring of cancer. Accurate and reliable tumor segmentation improves patient outcomes and enables precision medicine. Despite advancements in imaging technologies, accurate tumor segmentation remains challenging due to the heterogeneous nature of tumors, variations in imaging modalities, and the presence of noise and artifacts. These challenges can lead to significant variability in segmentation results, which impacts clinical decision-making and patient care. To ensure the reliability of automated segmentation models in clinical practice, it is crucial to predict the tumor boundaries and quantify the uncertainty associated with these predictions. Uncertainty quantification provides insights into model confidence, identifies potential errors, and helps in risk assessment, supporting more informed clinical decisions. While several deep learning-based methods have achieved promising results in tumor segmentation, they often need robust mechanisms for uncertainty quantification. Some of the main downsides of the existing approaches include providing overly conservative estimates, being computationally expensive, or changing the model architecture to calculate uncertainty (Heller et al. 2021; Yogananda et al. 2020; Roy et al. 2022; Dorta et al. 2018).

This paper presents a novel approach to estimating uncertainty in image segmentation with an application in kidney tumor segmentation. Our method leverages multiple checkpoints from the local minima obtained during the training process of the segmentation model to generate robust uncertainty maps in one training and inference process. This approach provides critical insights into model confidence and improves segmentation accuracy without changing the model architecture or using computationally expensive resources. We evaluated our approach on the KiTS23 dataset (Heller et al. 2023), demonstrating its effectiveness in identifying ambiguous regions and out-of-distribution data. The generated uncertainty maps enhance clinical decision-making by highlighting areas of low confidence, ultimately aiding in more accurate and reliable medical diagnoses.

The next section provides a summary of the existing literature before a discussion of the proposed methodology.

Literature Review

Various approaches have been introduced to estimate uncertainty in medical segmentation tasks and can be categorized into Deterministic Single Networks, Bayesian Neural Networks (BNNs), Ensemble method, and Test-time data augmentation approaches (Zou et al. 2023).

In Deterministic Single networks, uncertainty is inferred from the network's output, where it is assumed that the model is deterministic, and the uncertainty can be estimated based on one single forward pass. Holder and Shafique (2021) proposed an Uncertainty Distillation method for segmentation tasks. It utilizes a pre-trained teacher network and trains a smaller student network to mimic the teacher's output distribution, including the inherent uncertainty. While the student network is a Deterministic Single Network, the distillation process leverages the teacher's implicit uncertainty for improved confidence estimation. Franchi et al. (2021) introduced OVNNI, which addresses uncertainty in

deep learning classifications by employing two networks: ”One-vs-All” specialists, each trained to detect a single class, and an ”All-vs-All” network that distinguishes between all class pairs. Together, these networks collaborate during predictions, and high scores from both indicate confidence, while low scores signal uncertainty, particularly in identifying out-of-distribution data.

Bayesian Neural Networks (BNNs) address this limitation by treating the network's weights as probability distributions. This allows the network to learn a distribution of possible predictions instead of a deterministic prediction obtained with the point estimate values of the network parameters, inherently capturing the model's uncertainty. Salahuddin et al. (2023) proposed a modification to cascaded nnUNet framework Isensee et al. (2021) for segmenting kidneys, tumors, and cysts in CT scans. This approach leverages uncertainty estimation through Monte Carlo Dropout to identify potentially ambiguous structures and improve segmentation accuracy, particularly for tumors and cysts. Uncertainty maps highlight regions of lower model confidence, aiding in clinical decision-making. Baumgartner et al. (2019) proposed a method that leverages a hierarchical probabilistic model to capture the uncertainty in medical image segmentation. Unlike traditional deterministic models, PHISeg estimates a distribution over possible segmentations by learning a conditional variational autoencoder. This allows the model to produce a variety of plausible segmentations, reflecting both model and data uncertainty. This approach is particularly valuable in medical imaging, where understanding the confidence in predictions is crucial for clinical decision-making. Holder and Shafique (2021) introduced an efficient approach for posterior sampling of weight space to estimate Bayesian uncertainty. Their approach maintains the original segmentation model architecture without requiring a variational framework, thereby preserving the performance of nnUNet. Additionally, they boost the segmentation performance over the original nnU-Net via marginalizing multi-modal posterior models. They applied their method to the public ACDC and M&M datasets of cardiac MRI and demonstrated improved uncertainty estimation.

Ensemble methods leverage the power of multiple models where an ensemble of DNNs, potentially with different architectures or training data, is trained. The final prediction is often an aggregation of the individual predictions, leveraging the disagreement among the ensemble members. Causey et al. (2021) proposed to train multiple variations of the U-Net architecture and data augmentation, where each U-Net votes on kidney and tumor segmentation in new CT images, and their combined predictions yield the final segmentation.

Test-Time Data Augmentation (TTA) focuses on creating variations of the input data through techniques like random crops, rotations, or adding noise. The model's predictions for these augmented versions are used for uncertainty estimation, where high variability in the network's predictions for these variations suggests the model is sensitive to slight changes in the input and, hence, less certain about the prediction. Wang et al. (2019) explored the application of TTA with uncertainty estimation for deep learning-based medical image segmentation in a broader context. They compare TTA-based uncertainty to model-based uncertainty estimation (limitations of the model itself) and demonstrate the benefits of TTA for improving segmentation performance and reducing overconfident incorrect predictions.

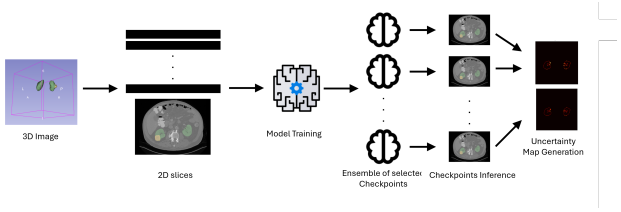

Figure 1: The project pipeline starts with converting 3D CT images into 2D axial images. Next, 2D images are fed into the nnUNet model for training, where inference is performed on the test set, and the uncertainty maps are generated after selecting multiple checkpoints from the trained model.

One of the significant downsides of the previous state-of-the-art uncertainty estimation approaches, particularly the Monte Carlo dropout and other Bayesian Neural Networks, is that they change the model architecture, which leads to decreasing the performance of the models that have been fine-tuned on specific tasks. Our approach for uncertainty quantification is independent of the model architecture, where we sample multiple checkpoints from the local minima in the training process. These checkpoints generate uncertainty maps that indicate the regions in which the model is uncertain about in the segmentation.

Methodology

The steps of our method are shown in Figure 1 and described in the following subsections:

Dataset

We evaluated our method on the Kidney and Kidney Tumor Segmentation Challenge (KiTS23) dataset (Heller et al. 2023), a recent and comprehensive dataset of cross-sectional medical images for kidney and kidney tumor segmentation. The dataset is composed of 458 cases (patients). Each case contains three different types of label masks: kidney, tumor, and cyst.

Pre-process

- **Dataset Split**: We split the dataset into 400 cases for training and validation, and the rest for testing. Using 5-fold cross-validation, 320 cases are used for training, and 80 cases are used for validation.
- **Resampling**: Different cases include different voxel spacing , and to ensure common voxel spacing and reduce variability in spatial resolutions, we perform resampling.
- **Intensity normalization**: Intensity values are normalized by subtracting the mean and dividing by the standard deviation, done on a per-case basis, normalizing each 3D volume independently to improve model training.

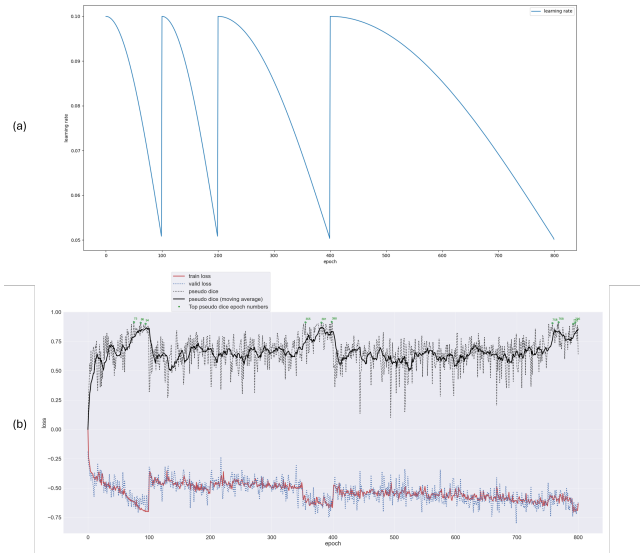

Figure 2: a) The learning rate value over the training progress with the SGDR learning rate scheduler. b) Training progress over 800 epochs and 4 cycles. Green numbers indicate epoch numbers, where the local minimum is happening (highest validation dice score) in the training progress. These are around epochs 100, 200, 400, and 800, where the top three checkpoints around each local minimum is selected for uncertainty map generation. Note that model training progress does not show any peak around epoch 200. Thus, no checkpoint is selected around this epoch.

- **Slicing**: For each 3D CT scan volume, we perform 2D axial slicing to generate 2D images. After inference and prediction of the model, they are stacked up on each other to reconstruct the images similar to the original 3D volumes.
- **Data Augmentation**: We apply several data augmentation techniques to increase the diversity of the training data, including rotations, scaling, mirroring, gamma correction, etc.

Segmentation model (Training and Inference)

Our methodology is built upon nnUNet (Isensee et al. 2021), a deep learning framework specifically designed for medical image segmentation tasks. It tackles the challenge of adapting segmentation pipelines to new datasets by automatically extracting key dataset characteristics and using them to configure a U-Net-based architecture with optimal parameters. This eliminates the need for manual configuration and improves generalizability to unseen datasets. We specifically used nnUNet ResEnc Large as the segmentation model. We performed a 5-fold cross-validation to have a better evaluation. The nnUNet framework was used for both training and inference of the model. The original nnUNet method uses polynomial learning rate decay, where the learning rate decreases to nearly zero in the last epochs of the training phase. However, we updated the learning rate scheduler, which is adapted using Stochastic Gradient Descent with Warm Restarts (SGDR) as proposed by Loshchilov and Hutter Loshchilov and Hutter (2016). This method allows the model to escape local minima by periodically restarting the learning rate to a higher value, creating multiple local minima before the restart happens in the training progress.

As shown in Figure 2, the model has three peaks based on the pseudo dice as a metric. Then, the learning rate scheduler modifies the learning rate. In this way, the last peak of the model has better performance in segmentation, and we sample posterior weights from the peaks of this training progress. In other words, we take the checkpoints from the peak points of the training process. Next, we use the ensemble of these checkpoints to generate segmentation and uncertainty maps. These peaks are the local minima. The goal is to use the model's local minima obtained by the previous peaks (epochs 100 and 400 in our case) and the optimal minima in the last peak to have an ensemble of models for uncertainty quantification. This approach is independent of the model architecture and can be used with any other model or framework for quantifying uncertainty. What makes it computationally efficient is that we use the ensemble of models but in the inference time and we do not train multiple models to have this ensemble. The learning rate formula used in this methodology is as follows:

$$T_i = T_0 \cdot \eta^i \tag{1}$$

$$\text{lr}(t) = \frac{lr_{min}}{2}\left(\cos\left(\frac{\pi \cdot \text{mod}(t, T_i)}{T_i}\right) + 1\right) + lr_{min} \tag{2}$$

In the given formula, SGDR adjusts the learning rate in a cyclical manner, where each cycle's length is determined by $T_i = T_0 \cdot \eta^i$, with T_0 as the initial cycle length and η as the factor by which the cycle length increases after each restart. Within each cycle, the learning rate $\text{lr}(t)$ follows a cosine annealing schedule, starting from a higher value and gradually decreasing to a lower bound, lr_{min}. Using the cosine function ensures a smooth decrease in the learning rate within each cycle, and the periodic warm restarts at the end of each cycle allow the model to escape local minima by resetting the learning rate and improving convergence while exploring different regions of the loss landscape. Figure 2 demonstrates the training progress.

Checkpoint Selection

Three checkpoints are selected from the peaks of each cycle, represented by green numbers as indicated in Figure 2. Predictions are made with each of the sampled checkpoints. The ensemble of these predictions creates uncertainty maps as models tend to predict different segmentations, especially around the borders of kidneys, tumors, and cysts.

Bayesian Inference

This step is done for four types of classes, which are background, kidney, tumor, and cyst. We generate four probability maps and estimate an average of individual checkpoint probability, from the test output y_{test}, given the test input x_{test} and dataset D using the following formulas:

$$p(y_{\text{test}}|x_{\text{test}}, D) \approx \frac{1}{n}\sum_{i=1}^{n} p(y_{\text{test}}|x_{\text{test}}, w_{t_i}), \text{for } w_{t_i} \in W \tag{3}$$

The ensemble prediction is approximated by averaging the predictions from multiple checkpoints, where n is the number of checkpoints. w_{t_i} are the model parameters at checkpoint t_i. The set of all model parameters w_{t_i} is defined as W.

Uncertainty Map Generation

In order to generate uncertainty maps, we compute class-specific entropy for kidneys, tumors, and cysts. The process is as follows:

Entropy Calculation We calculated the entropy of each pixel using the following formula:

$$H(y_{\text{test}}^{ij}) = -\sum_{k=1}^{c} p(y_{\text{test}}^{ij} = k|x_{\text{test}}, D) \log_2 p(y_{\text{test}}^{ij} = k|x_{\text{test}}, D) \tag{4}$$

In equation 4, $H(y_{\text{test}}^{ij})$ denotes the entropy of the predicted output, $p(y_{\text{test}}^{ij} = k|x_{\text{test}}, D)$ is the probability that the predicted output y_{test}^{ij} is equal to class k given the test input x_{test} and the dataset D. $\sum_{k=1}^{c}$ indicates that the summation is performed over all c classes, and c is the number of classes. The entropy is defined as the sum of the product of the probability of each class and the logarithm of that probability.

Normalization Due to the variation in segmentation areas across samples and classes, more than the regular mean approach for quantifying entropy is needed to achieve a single uncertainty score per image. We applied normalization by generating a dilated version of each image, subtracting it from the original image, and isolating only the segmentation contours. Dilation of the binary image effectively enlarges the segmented regions by expanding the boundaries. This dilation helps isolate the segmentation contours more effectively by emphasizing the borders between segmented regions and the background. This approach provided a normalized sum of entropy values for each image. The formulas for computing the normalized entropy are as follows:

$$p_{\text{class}}(y_{\text{test}}|x_{\text{test}}, D) \approx \frac{1}{n}\sum_{i=1}^{n} p_{\text{class}}(y_{\text{test}}|x_{\text{test}}, w_{ti}) \tag{5}$$

Equation 5 represents a process for calculating an uncertainty score from image segmentation outputs. It approximates the mean image of foreground probabilities for a given class, where the threshold is set to 0.5 to create a binary image I.

$$H_{\text{total}} = \sum_{i}\sum_{j} H(y_{\text{test}})_{ij} \tag{6}$$

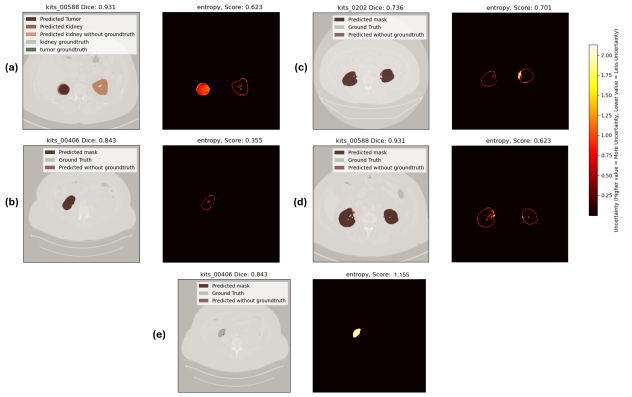

Figure 3: Some samples of the results to compare the segmentation performance with uncertainty maps. Lighter pixels usually indicate higher uncertainty and darker pixels indicate cases where the checkpoint models have the same inference results. In the labels of the images, "without ground truth" indicates pixels that the ground truth does not consider that part as a mask (kidney, tumor, or cyst), but the model has segmented that pixel. The dice scores and uncertainty scores are for the whole 3D images and not the 2D slices on the figure alone.

Equation 6 H_{total} is the sum of entropy values across all pixels in the test image, reflecting the total uncertainty.

$$I^* = I \oplus A; \tag{7}$$

Equation 7 describes I^* as the result of dilating the binary image I using a structuring element A. Dilation is a morphological operation used to grow or expand the boundaries of regions of foreground pixels (in our case, it is pixel values of 1, 2, and 3 for kidney, tumor, and cyst, respectively) in a binary image. The structuring element (or kernel) defines the neighborhood where the dilation operation is performed.

$$I_{\text{normalization}} = I - I^* \tag{8}$$

Equation 8 normalizes the binary image by subtracting the dilated image from the original binary image, yielding $I_{\text{normalization}}$.

$$\text{Uncertainty score} = \frac{H_{\text{total}}}{\sum_{i,j} I_{\text{normalization}}} \tag{9}$$

Equation 9 calculates the uncertainty score by dividing the total entropy H_{total} by the sum of the normalized image values $I_{\text{normalization}}$.

Finally, at this stage we have the entropy and inference result for each of the 2D images in our dataset, and by stacking up the 2D images on top of each other, we can regenerate uncertainty maps and segmentation masks similar to the original 3D images.

Results

With the help of the uncertainty maps, we improved the quality of the segmentation results. Figure 3 shows test data

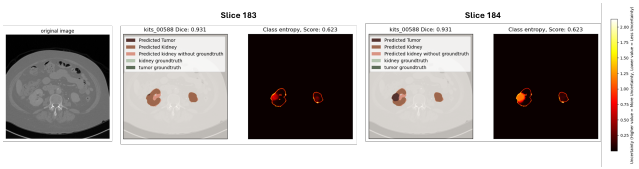

Figure 4: Slice 184 indicates a kidney slice where a tumor exists, and both the segmentation and the uncertainty map cover it. However, slice 183 (only one slice before 184) has no tumor in its ground truth, and the model's segmentation result shows a similar pattern. However, the uncertainty map demonstrates a tumor portion in this slice. So, while this slice has a tumor, neither the ground truth nor the segmentation map is showing it. As we have this evidence in exactly the next slice, we are sure that the uncertainty map is doing better in finding this slice as an unhealthy kidney with a tumor (Approved by a specialist physician).

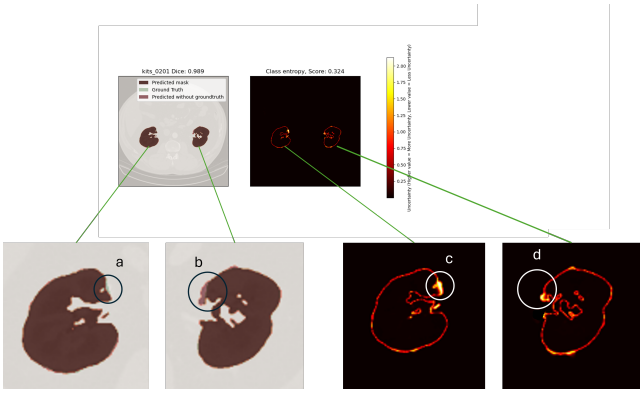

Figure 5: Circles a and c indicate the small portion of the kidney (green color) that the model could not predict and is highly uncertain about, which results are reflected in the uncertainty map. Circles b and d indicate portions of the kidney that are not available in the ground truth but are, in fact, a kidney portion. The segmentation model has found that part is the kidney, and the uncertainty is also high in this area (Approved by a specialist physician.

samples to compare the segmentation and uncertainty maps. Cases a, b, c, and d represent segmentations where the uncertainty map matches the predicted segmentation. In contrast, in some cases, the performance of the segmentation models is not good, but the uncertainty map has detected those parts as a potential kidney, tumor, or cyst portion. Case e indicates a case where the model segmentation has not detected the ground truth, the uncertainty is high and the uncertainty map has covered that area.

Figure 4 demonstrates a case where the segmentation model seems to match the ground truth, but both the ground truth and the segmentation model are wrong (case 183). This problem comes from the original annotation that the uncertainty map found, and it shows a tumor portion in the picture (in the left kidney).

In some cases, the ground truth is incomplete, and the segmentation model performs better than the ground truth. In Figure 5, in circle b, the ground truth is not considering the light red color as a kidney, but it is indeed a kidney portion, and we see the segmentation model has considered that portion as a kidney (Approved by a specialist physician). This phenomenon also happens in case c in Figure 3 on the right kidney. In addition, the model on circle a in Figure 5, has not detected a segment of the kidney, and the uncertainty in that segment is high in circle c.

Method	ECE (%)
5-fold	2.89
Phiseg	2.52
Monte Carlo	3.44
Deep Ensemble	2.70
HMC	1.36
Rel-UNet	**1.11**

Table 1: Comparison of Estimated Calibration Error (ECE) of our model with previous state-of-the-art uncertainty Quantification models on KITS23 dataset. Lower ECE indicates improved calibration performance. As shown, Rel-UNet, HMC (Zhao et al. 2022), and Phiseg (Baumgartner et al. 2019) are among the top performers with lower ECE values. 5-fold (Khalili et al. 2024), Deep Ensemble (Causey et al. 2021), and Monte Carlo (Monteiro et al. 2020) are next in line.

The comparison of the segmentation results of our method and previous approaches is available in supplementary materials in Table 3, and an analysis of the dice score with respect to uncertainty score is available in the supplementary materials section in Figure 7. The hyperparameters and the model's setup are available in the Experimental Setup section of the Supplementary Material section.

In summary, our proposed method consistently outperformed baseline models in uncertainty estimation, with a 15% reduction in Expected Calibration Error (ECE) and a comparable Dice score in tumor segmentation.

Discussion

In this section, we compared the result of the uncertainty maps generated by our model with previous state-of-the-art models in uncertainty quantification in two ways. First is the uncertainty maps that they generate, available in Figure 6, and second is the comparison of Expected Calibration Error (ECE) available in Table 1. In this figure, cases (a) and (d) indicate a successful uncertainty map generation, where all approaches generate similar patterns for segmenting the kidney and the tumor available in the picture. Case (b) indicates a healthy kidney and an unhealthy kidney with a tumor, and the blue circle shows the kidney portion of the unhealthy kidney. In this case, Phiseg and Monte Carlo do not show the healthy portion of the unhealthy kidney, whereas deep ensemble, HMC, and our approach show that portion in their uncertainty map. In case (c), Monte Carlo and Deep Ensemble cannot find the tumor's position in the image, whereas Phiseg, HMC, and approach do. Case (e) shows a relatively complex example where both tumor and cyst exist in the im-

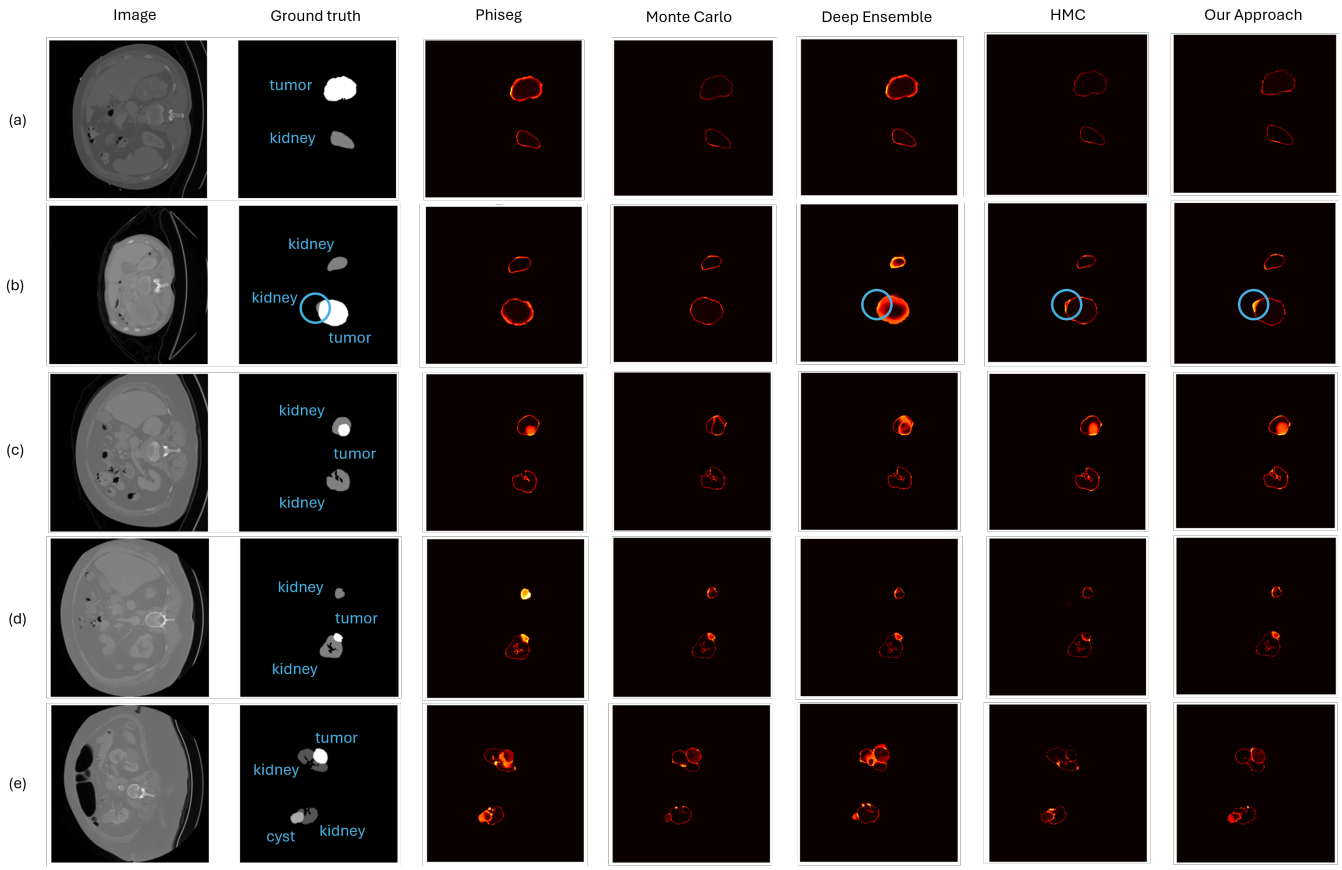

Figure 6: Comparison of our approach with previous state-of-the-art models in uncertainty estimation in uncertainty maps. Cases are kits 420 (slice 24), kits 178 (slice 30), kits 588 (slice 136), kits 401 (slice 70), and kits 403 (slice 38), respectively.

age, and previous approaches cannot fully locate the regions of the tumor, the cyst, and the healthy portions of the kidney accurately, while our method is locating these areas more precisely. See Figure 8 in the supplementary material section for the original version of this figure without any annotation.

To quantitatively compare the results of our model on the test dataset with previous state-of-the-art models, we used Expected Calibration Error (ECE) that is available in Table 1. ECE is a metric used to measure the difference between predicted probabilities and actual outcomes to assess the calibration of a model's confidence. It is computed by partitioning predictions into bins based on confidence scores and then calculating the weighted average of the absolute difference between accuracy and confidence within each bin. A lower ECE indicates that the model's predicted probabilities are well-calibrated to the true likelihood of segmentations.

Our approach to leveraging multiple local minima for uncertainty quantification presents a significant advancement over traditional methods, offering a computationally efficient solution without compromising segmentation accuracy. This could have broad implications for integrating uncertainty quantification into clinical workflows.

Conclusion

In conclusion, our study presents a novel approach to quantifying uncertainty in kidney tumor and cyst segmentation using deep learning by leveraging multiple local minima from the training process. We generate robust uncertainty maps that enhance segmentation accuracy and provide critical insights into model confidence. This approach not only improves the reliability of segmentation models but also has the potential to be integrated into clinical workflows, thereby aiding in accurate and reliable medical diagnoses.

While our study has provided valuable insights into the utilization of uncertainty maps to increase the level of trust and reliance in tumor and cyst segmentation models, several areas remain unexplored and present opportunities for further research. First, Rel-UNet starts uncertainty estimation only after the training process is finished, which prevents the model from using the uncertainty estimation while the model is being trained. Second, in order to increase the usability of our model, we will incorporate it into other datasets and tumor segmentation tasks. Third, we will integrate our framework into other image segmentation models to enhance the reliability of their methodology.

References

Baumgartner, C. F.; Tezcan, K. C.; Chaitanya, K.; Hötker, A. M.; Muehlematter, U. J.; Schawkat, K.; Becker, A. S.; Donati, O.; and Konukoglu, E. 2019. Phiseg: Capturing uncertainty in medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*, 119–127. Springer.

Causey, J.; Stubblefield, J.; Qualls, J.; Fowler, J.; Cai, L.; Walker, K.; Guan, Y.; and Huang, X. 2021. An ensemble of U-Net models for kidney tumor segmentation with CT images. *IEEE/ACM transactions on computational biology and bioinformatics*, 19(3): 1387–1392.

Dorta, G.; Vicente, S.; Agapito, L.; Campbell, N. D.; and Simpson, I. 2018. Structured uncertainty prediction networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5477–5485.

Franchi, G.; Bursuc, A.; Aldea, E.; Dubuisson, S.; and Bloch, I. 2021. One versus all for deep neural network for uncertainty (ovnni) quantification. *IEEE Access*, 10: 7300–7312.

Heller, N.; Isensee, F.; Maier-Hein, K. H.; Hou, X.; Xie, C.; Li, F.; Nan, Y.; Mu, G.; Lin, Z.; Han, M.; et al. 2021. The state of the art in kidney and kidney tumor segmentation in contrast-enhanced CT imaging: Results of the KiTS19 challenge. *Medical image analysis*, 67: 101821.

Heller, N.; Isensee, F.; Trofimova, D.; Tejpaul, R.; Zhao, Z.; Chen, H.; Wang, L.; Golts, A.; Khapun, D.; Shats, D.; Shoshan, Y.; Gilboa-Solomon, F.; George, Y.; Yang, X.; Zhang, J.; Zhang, J.; Xia, Y.; Wu, M.; Liu, Z.; Walczak, E.; McSweeney, S.; Vasdev, R.; Hornung, C.; Solaiman, R.; Schoephoerster, J.; Abernathy, B.; Wu, D.; Abdulkadir, S.; Byun, B.; Spriggs, J.; Struyk, G.; Austin, A.; Simpson, B.; Hagstrom, M.; Virnig, S.; French, J.; Venkatesh, N.; Chan, S.; Moore, K.; Jacobsen, A.; Austin, S.; Austin, M.; Regmi, S.; Papanikolopoulos, N.; and Weight, C. 2023. The KiTS21 Challenge: Automatic segmentation of kidneys, renal tumors, and renal cysts in corticomedullary-phase CT. arXiv:2307.01984.

Holder, C. J.; and Shafique, M. 2021. Efficient uncertainty estimation in semantic segmentation via distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3087–3094.

Isensee, F.; Jaeger, P. F.; Kohl, S. A.; Petersen, J.; and Maier-Hein, K. H. 2021. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2): 203–211.

Khalili, N.; Spronck, J.; Ciompi, F.; van der Laak, J.; and Litjens, G. 2024. Uncertainty-guided annotation enhances segmentation with the human-in-the-loop. *arXiv preprint arXiv:2404.07208*.

Liu, S.; and Han, B. 2023. Dynamic Resolution Network for Kidney Tumor Segmentation. In *International Challenge on Kidney and Kidney Tumor Segmentation*, 14–21. Springer.

Loshchilov, I.; and Hutter, F. 2016. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*.

Monteiro, M.; Le Folgoc, L.; Coelho de Castro, D.; Pawlowski, N.; Marques, B.; Kamnitsas, K.; van der Wilk, M.; and Glocker, B. 2020. Stochastic segmentation networks: Modelling spatially correlated aleatoric uncertainty. *Advances in neural information processing systems*, 33: 12756–12767.

Myronenko, A.; Yang, D.; He, Y.; and Xu, D. 2023. Automated 3D Segmentation of Kidneys and Tumors in MICCAI KiTS 2023 Challenge. In *International Challenge on Kidney and Kidney Tumor Segmentation*, 1–7. Springer.

Qian, L.; Luo, L.; Zhong, Y.; and Zhong, D. 2023. A Hybrid Network Based on nnU-Net and Swin Transformer for Kidney Tumor Segmentation. In *International Challenge on Kidney and Kidney Tumor Segmentation*, 30–39. Springer.

Roy, S.; Trapp, M.; Pilzer, A.; Kannala, J.; Sebe, N.; Ricci, E.; and Solin, A. 2022. Uncertainty-guided source-free domain adaptation. In *European conference on computer vision*, 537–555. Springer.

Salahuddin, Z.; Kuang, S.; Lambin, P.; and Woodruff, H. C. 2023. Leveraging Uncertainty Estimation for Segmentation of Kidney, Kidney Tumor and Kidney Cysts. In *International Challenge on Kidney and Kidney Tumor Segmentation*, 40–46. Springer.

Stoica, G.; Breaban, M.; and Barbu, V. 2023. Analyzing domain shift when using additional data for the MICCAI KiTS23 Challenge. In *International Challenge on Kidney and Kidney Tumor Segmentation*, 22–29. Springer.

Uhm, K.-H.; Cho, H.; Xu, Z.; Lim, S.; Jung, S.-W.; Hong, S.-H.; and Ko, S.-J. 2023. Exploring 3D U-Net Training Configurations and Post-processing Strategies for the MICCAI 2023 Kidney and Tumor Segmentation Challenge. In *International Challenge on Kidney and Kidney Tumor Segmentation*, 8–13. Springer.

Wang, G.; Li, W.; Aertsen, M.; Deprest, J.; Ourselin, S.; and Vercauteren, T. 2019. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing*, 338: 34–45.

Yogananda, C. G. B.; Shah, B. R.; Vejdani-Jahromi, M.; Nalawade, S. S.; Murugesan, G. K.; Yu, F. F.; Pinho, M. C.; Wagner, B. C.; Emblem, K. E.; Bjørnerud, A.; et al. 2020. A fully automated deep learning network for brain tumor segmentation. *Tomography*, 6(2): 186–193.

Zhao, Y.; Yang, C.; Schweidtmann, A.; and Tao, Q. 2022. Efficient Bayesian uncertainty estimation for nnU-Net. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 535–544. Springer.

Zou, K.; Chen, Z.; Yuan, X.; Shen, X.; Wang, M.; and Fu, H. 2023. A review of uncertainty estimation and its application in medical imaging. *Meta-Radiology*, 100003.

Supplementary Material

Experimental Setup

The training is performed using nnUNet ResEnc Large model, SGD as optimizer, SGDR as learning rate scheduler, initial learning rate with 0.1, T_0 with 100 (The number of epochs before the first restart), T_mult with 2 (The factor by which the cycle length (T_0) is multiplied after each restart), eta_min is 0.0001 (minimum learning rate). The number of epochs is 800. Thus, with the configuration above, the restart happens around epochs 100, 200, 400, and 800. The loss function is Dice Loss, and the batch size is 40.

Comparison with other experiments

We ran different setups of our method to compare the performance of segmentation and uncertainty maps under different variations. Table 2 shows the comparison of 5-fold, 3D, and Full train comparison. 5-fold is training five different models with five different folds of the dataset and create an ensemble of models for uncertainty generation. 3D is the exact configuration of our method but with the original 3D images. Full-train is training a single model and using the epochs at 60%, 70%, 80%, 90%, and the last epoch to create the ensemble of checkpoint models and generate uncertainty maps. Analysis of Table 2 is as follows:

- 5-fold approach has a higher dice score and lower entropy compared to the SGDR approach that we used for training, but it also has higher ECE. This means the model could be achieving a high Dice score by correctly predicting most of the correct segmentations, but when it makes mistakes, it does so with high confidence, as it is not well-calibrated.
- Full-train setup indicates low ECE, which means it does not show good calibration and has high entropy (uncertainty).
- 3D experiment with the same 800 training epochs shows a lower dice score, higher ECE, and higher entropy. Although 3D segmentation considers the voxels in its predictions, it does not necessarily perform better as we see in this experiment. The reason is that with the same architecture and the same number of epochs for training, it does not perform well, and the model requires a deeper architecture with more epochs to perform well. In addition, training in 3D architecture requires more training data than 2D training in the same architecture. With 3D, we have at most 458 cases and images in Kits23, but with 2D, we have more than 90000 images to train after the 2D slicing of the CT scan images.

Comparison in Segmentation

Table 3 indicates the segmentation result compared to the top 6 performers in the Kits23 challenge. Although the segmentation result of our model is not as good as the result in the top 6 methods, our aim is not to solely perform segmentation but to incorporate uncertainty to qualify the segmentation results more precisely.

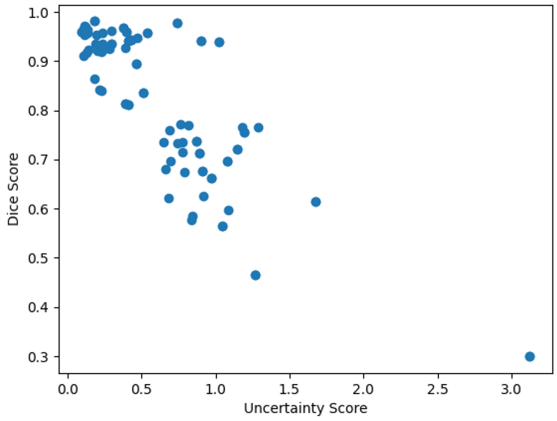

Figure 7: This figure indicates that lower uncertainty usually concludes higher segmentation performance. It indicates that segmentation performance (dice score) is more than 80% when the uncertainty score is less than 0.5.

Method	ECE (%)	Dice score (%)	Avg. Entropy
5-fold	3.02	81.1	0.52
3D	5.20	73.3	0.98
Full-train	3.37	80.5	0.71
Rel-UNet	**1.11**	**80.0**	**0.67**

Table 2: Comparison of our approach with 5-fold cross-validation model, the 3D architecture of our approach, and full-train in ECE and Dice score.

Uncertainty and Segmentation Performance

Figure 7 shows the segmentation performance and the uncertainty score calculated from the test dataset. With some approximation, the segmentation model performs better in areas with lower uncertainty scores. In cases where the uncertainty score is less than 0.5, the dice score (segmentation performance) is higher than 80%.

Method	Kidney Dice	Masses Dice	Tumor Dice
Andriy Myronenko et al. (Myronenko et al. 2023)	0.835	0.723	0.758
Kwang-Hyun Uhm et al. (Uhm et al. 2023)	0.820	0.712	0.738
Yasmeen George et al.	0.819	0.707	0.713
Shuolin Liu et al. (Liu and Han 2023)	0.805	0.706	0.697
George Stoica et al. (Stoica, Breaban, and Barbu 2023)	0.807	0.691	0.713
Lifei Qian et al. (Qian et al. 2023)	0.801	0.680	0.687
Rel-UNet	0.800	70.2	0.641

Table 3: Comparison of our SGDR segmentation model with top 6 performers in KITS23 challenge. The purpose of our model is not segmentation alone, and we incorporate uncertainty in our prediction in contrast to these top 6 performers.

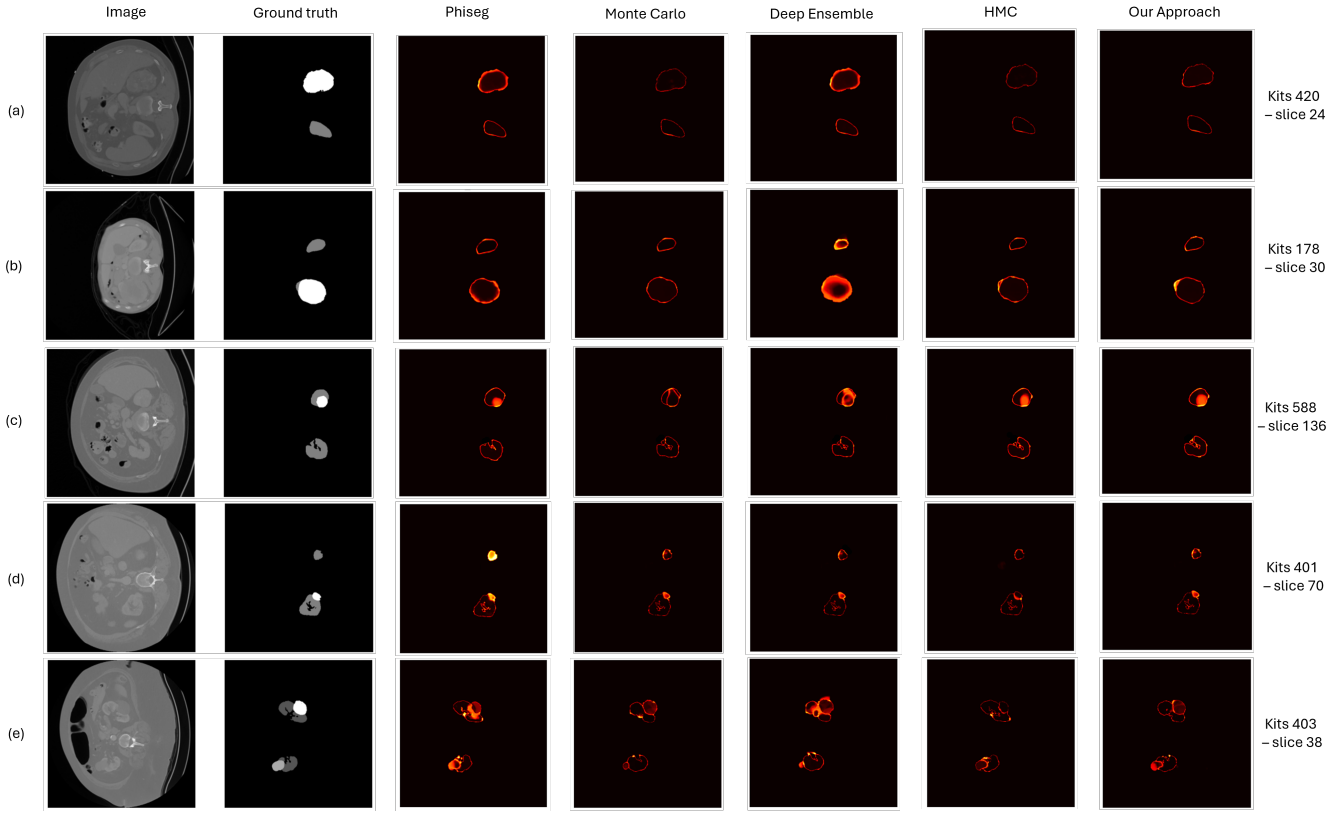

Figure 8: Comparison of our approach with previous state-of-the-art models in uncertainty estimation in uncertainty maps. Case (a) is kits 420 - slice 24, case (b) is kits 178 - slice 30, case (c) is kits 588 - slice 136, case (d) is kits 401 - slice 70, and case (e) is kits 403 - slice 38.