FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Uncertainty-based Interactive 3D Medical Image Segmentation

Ana Filipa Macedo de Sousa

Mestrado em Bioengenharia

Supervisor: Prof. Dr. João Pedrosa

Second Supervisor: Prof. Dr. Helena Williams

July 16, 2025

Uncertainty-based Interactive 3D Medical Image Segmentation

Ana Filipa Macedo de Sousa

Mestrado em Bioengenharia

July 16, 2025

Abstract

Deep learning has significantly advanced medical image segmentation by enabling the automatic delineation of complex anatomical structures. Nonetheless, the clinical deployment of such models remains constrained by challenges in interpretability, robustness, and limited mechanisms for effective user interaction. This dissertation addresses these limitations by leveraging uncertainty quantification to support interactive 3D medical image segmentation, with a focus on the segmentation of the pericardium in computed tomography (CT) and the external anal sphincter muscle in transperineal ultrasound (TPUS).

A baseline segmentation model based on the 3D nnU-Net architecture was implemented, incorporating a calibration component into the loss function to enhance the reliability of the model's probabilistic outputs. Based on this, Monte Carlo Dropout was employed to estimate epistemic uncertainty at both voxel and structure levels. For structure-wise uncertainty, a metric based on volume-level entropy was introduced, while for voxel-wise uncertainty, a refined pipeline was proposed to filter and enhance uncertainty maps for interpretability.

Experimental results showed a correlation between predicted uncertainty and segmentation error. In the CT dataset, Spearman correlation coefficients reached 0.88 for voxel-wise and 0.65 for structure-wise uncertainty, while for TPUS the values were 0.61 and 0.59, respectively. These uncertainty estimates were used to guide simulated user corrections. Uncertainty-based selection strategies consistently outperformed random baselines and closely approximated ideal (Dice-based) strategies, important in real-world settings where ground truth is unavailable. Structure-wise uncertainty showed promising potential to prioritize image review. Specifically, to reduce the proportion of poor segmentations (defined as images with a Dice score below the inter-observer Dice) to under 5%, the number of images requiring review was reduced by 56% for CT and 74% for TPUS using uncertainty-based selection, compared to random selection. Voxel-wise uncertainty also demonstrated promising results in targeting erroneous regions within volumes, shown in patch correction experiments, and outperformed random region selection in guiding localized corrections. A hybrid approach combining both levels further improved correction efficiency.

An interactive segmentation framework, BISNet, was also implemented, incorporating user-provided boundary clicks. This framework was only applied to the TPUS data, where it demonstrated effective segmentation refinement through user interaction, but due to limitations in the training process, it was not possible to implement BISNet for the CT segmentation task. A clinical experiment with BISNet on TPUS data confirmed the utility of structure-wise uncertainty to priortize image review. Furthermore, the interactive correction process was positively evaluated by the participating clinician, who considered it intuitive and effective for refining segmentations.

Overall, this work demonstrates that integrating uncertainty estimation into interactive segmentation frameworks enables more targeted clinician intervention, reduces manual workload, and has the potential to enhance the reliability, transparency, and clinical applicability of Artificial

Intelligence-based medical image analysis systems.

Keywords: Medical image segmentation, Uncertainty quantification, Interactive segmentation, Monte Carlo Dropout, nnU-Net, 3D imaging, Model calibration, User-guided correction, Deep Learning

Resumo

A Aprendizagem Profunda tem contribuído significativamente para o avanço da segmentação de imagens médicas, ao permitir a delineação automática de estruturas anatómicas complexas. No entanto, a aplicação clínica destes modelos continua limitada por desafios relacionados com a interpretabilidade, robustez e ausência de mecanismos eficazes de interação com o utilizador. Esta dissertação aborda estas limitações através da utilização de quantificação de incerteza para apoiar a segmentação interativa de imagens médicas 3D, com foco na segmentação do pericárdio em Tomografia Computorizada (TC) e do esfíncter anal externo em Ecografia Transperineal (ET).

Foi implementado um modelo de segmentação base, utilizando a arquitetura da 3D nnU-Net e incorporando uma componente de calibração na função de perda, com o objetivo de melhorar a fiabilidade das previsões probabilísticas do modelo. Neste seguimento, foi utilizado Monte Carlo Dropout para estimar incerteza epistémica, tanto a nível de cada voxel individual como ao nível global da imagem. Para a incerteza global, foi introduzida uma métrica baseada na entropia do volume da estrutura, enquanto para a incerteza ao nível do voxel foi proposta uma pipeline para filtrar e melhorar os mapas de incerteza, facilitando a sua interpretabilidade.

Os resultados experimentais demonstraram uma correlação entre a incerteza e o erro de segmentação. No conjunto de dados de TC, os coeficientes de correlação de Spearman atingiram 0.88 para a incerteza a nível do voxel e 0.65 para a incerteza global, enquanto nos dados de ET os valores foram 0.61 e 0.59, respetivamente. Estas estimativas de incerteza foram utilizadas para orientar correções simuladas por parte do utilizador. As estratégias de seleção baseadas na incerteza superaram consistentemente os métodos aleatórios e aproximaram-se das estratégias ideais baseadas no Dice, o que é particularmente relevante em contextos reais onde não existe ground truth. A incerteza ao nível da estrutura revelou potencial para priorizar a revisão de imagens. Em concreto, para reduzir a proporção de segmentações de fraca qualidade (definidas como imagens com um valor de Dice inferior ao Dice entre observadores) para menos de 5%, o número de imagens que necessitam de ser revistas foi reduzido em 56% para TC e 74% para ET, utilizando seleção baseada na incerteza, em comparação com seleção aleatória. A incerteza ao nível do voxel também demonstrou resultados promissores na identificação de regiões com erros, que é demonstrado através de experiências de correção por patches, e superou a seleção aleatória de regiões para correções localizadas. Uma abordagem híbrida, que combina ambos os níveis, melhorou ainda mais a eficiência das correcções.

Foi também implementado uma framework de segmentação interativa, BISNet, que incorpora cliques do utilizador nas fronteiras das estruturas. Esta framework foi aplicada apenas aos dados ET, onde demonstrou eficácia no refinamento da segmentação através da interação do utilizador. Devido a limitações no processo de treino, não foi possível implementá-lo na tarefa de segmentação em TC. Uma experiência clínica com a BISNet, realizada sobre com os dados de ET, confirmou o potencial da incerteza global na priorização da revisão de imagens. Além disso, o processo de correção interativa foi avaliado de forma positiva pelo clínico participante, que o considerou intuitivo e eficaz para refinar as segmentações.

No seu conjunto, este trabalho demonstra que a integração da quantificação de incerteza em frameworks de segmentação interativa permite uma intervenção mais direcionada por parte do clínico, reduz o esforço manual necessário e tem o potencial para melhorar a fiabilidade, transparência e aplicabilidade clínica de sistemas de análise de imagem médica baseados em inteligência artificial.

Palavras-chave: Segmentação de imagem médica, Quantificação de incerteza, Segmentação interativa, Monte Carlo Dropout, nnU-Net, Imagem 3D, Calibração de modelos, Correção guiada por utilizador, Aprendizagem profunda

UN Sustainable Development Goals

The Sustainable Development Goals (SDGs), established by the United Nations in 2015, represent a comprehensive global framework aimed at promoting peace, prosperity, and sustainability by 2030 [1]. Comprising 17 interconnected goals, the SDGs cover areas of critical importance to humanity and the planet, including, among others, health, education, gender equality, innovation, social equity, and climate action. In the realm of engineering, these goals serve as a strategic reference to ensure that scientific and technological progress is pursued in an ethical, inclusive, and environmentally responsible manner.

This dissertation is situated within this broader paradigm of global transformation. The development and implementation of advanced Artificial Intelligence (AI) methodologies in the medical image analysis field not only aim to improve diagnostic accuracy and simplify clinical workflows but also contribute to the advancement of more accessible, efficient, and equitable healthcare systems. Such contributions are intrinsically aligned with the fundamental principles of universal health coverage, socially responsible technological innovation, and the development of resilient and inclusive infrastructure.

Therefore, a reflection on the relationship between this research and the SDGs reveals substantial direct and indirect contributions, particularly in the context of the following goals:

SDG 3 Ensure healthy lives and promote well-being for all at all ages.

SDG 9 Build resilient infrastructure, promote inclusive and sustainable industrialization, and foster innovation.

SDG 16 Promote peaceful and inclusive societies for sustainable development, provide access to justice for all, and build effective, accountable, and inclusive institutions at all levels.

Through an analysis of specific targets and associated performance indicators, it is possible to outline the ways in which this research can support the achievement of these SDGs. This alignment underscores the dissertation's commitment to sustainability, equity, and social responsibility. By pursuing scientific research guided by clearly defined goals and based on real-world challenges, the work reaffirms the fundamental role of biomedical AI image segmentation in engineering as a catalyst for positive transformation. Furthermore, the adoption of trustworthy AI-based frameworks, user-centered interaction, and robust uncertainty analysis increases not only the technical impact of the research but also its potential to provide sustainable, scalable, and ethically sound solutions in the field of healthcare.

SGD	Target	Contribution	Performance Indicators and Metrics
3	3.7	**Ensure universal access to sexual and reproductive health-care services.** The project contributes by enabling segmentation of the EAS in TPUS, aiding diagnosis of pelvic floor dysfunctions, especially important in maternal care and postnatal recovery, thus enhancing the accessibility and individualization of patient care.	**UN Indicator:** 3.7.1 – Proportion of women of reproductive age who have access to reproductive health services. **Project Metrics:** – Number of TPUS scans segmented using the AI tool. – Clinical time saved per EAS case. – Number of detected external anal sphincter injuries assisted by AI.
	3.8	**Achieve universal health coverage, including access to quality essential healthcare services.** The system improves access to diagnostic imaging through efficient and interactive AI models. Uncertainty estimation and interactivity allow for targeted human correction in out-of-distribution cases, improving inclusiveness and diagnostic fairness across diverse patient populations, while also reducing effort and speeding up decision-making, which is especially valuable in underserved regions with limited medical resources.	**UN Indicator:** 3.8.1 – Coverage of essential health services. **Project Metrics:** – Number of healthcare institutions adopting the tool. – Number of patients who benefit from the tool. – Performance across demographic groups.
9	9.5	**Enhance scientific research and upgrade technological capabilities.** The research fosters scientific advancement and technological innovation within the field of biomedical engineering, contributing to the research ecosystem in AI for healthcare.	**UN Indicator:** 9.5.1 – Research and development expenditure as a proportion of GDP. **Project Metrics:** – Number of AI models developed and evaluated. – Peer-reviewed publications and presentations. – Cross-disciplinary research collaborations established.
16	16.6	**Develop effective, accountable, and transparent institutions.** The project aligns with regulatory frameworks, promoting trustworthy AI through uncertainty quantification, explainability, and human-in-the-loop systems.	**UN Indicator:** 16.6.2 – Proportion of population satisfied with transparency of institutions. **Project Metrics:** – Number of AI outputs reviewed with human-in-the-loop interface. – Degree of compliance with transparency standards (e.g. MDR traceability checklist).

Acknowledgements

Em primeiro lugar, quero expressar a minha mais profunda gratidão aos meus pais, por serem um apoio constante e incansável, fundamentais em cada passo deste percurso, bem como em todos os aspetos da minha vida. Agradeço também por todos os esforços que fizeram para que eu pudesse chegar aqui. Foram, desde sempre, um exemplo de dedicação e resiliência para mim, que levo sempre comigo. Este agradecimento estende-se também a toda a minha família, uma base sólida e essencial ao longo de todo o caminho. Um agradecimento especial ao meu namorado, Tomás, pelo apoio incondicional, pela ajuda incansável, pela paciência nos momentos mais stressantes e por ter sido um suporte essencial sempre presente ao longo de todo este percurso.

Aos meus amigos e colegas, obrigada por cada momento vivido, dentro e fora da faculdade. Obrigada por tornarem a FEUP uma verdadeira casa e por ajudarem a tornar este percurso mais leve, mais alegre e inesquecível.

Quero ainda agradecer aos meus orientadores, João Pedrosa e Helena Williams, pelo acompanhamento atento, pela orientação dedicada, por tanto que me ensinaram e por todo o apoio prestado, que foram essenciais para a realização deste trabalho.

Por fim, um agradecimento aos meus colegas de laboratório no UZ Leuven, pelas valiosas contribuições e pelo espírito de colaboração, que enriqueceram de forma significativa esta investigação.

Ana Filipa Macedo de Sousa

"Man would not have attained the possible
unless time and again he had reached out for the impossible. "

Max Weber

Contents

1 Introduction **1**
- 1.1 Background and Motivation . . . 1
- 1.2 Objectives . . . 2

2 Bibliographic Review **4**
- 2.1 Segmentation . . . 4
- 2.2 Uncertainty Quantification . . . 8
  - 2.2.1 Uncertainty Quantification Methods . . . 9
  - 2.2.2 Structure-level Uncertainty . . . 16
- 2.3 Interactive Segmentation . . . 17
  - 2.3.1 Types of Interaction . . . 18
  - 2.3.2 Encoding User Interaction: Guidance Signal . . . 20
- 2.4 Uncertainty-guided Interactive Segmentation . . . 21

3 Methods **24**
- 3.1 Data . . . 24
  - 3.1.1 Cardiac CT Data . . . 24
  - 3.1.2 TPUS Data . . . 25
- 3.2 Baseline Segmentation Model . . . 25
  - 3.2.1 Model Architecture . . . 26
  - 3.2.2 Loss Function and Optimization . . . 27
  - 3.2.3 Model Selection Strategy . . . 29
- 3.3 Interactive Segmentation Model . . . 29
  - 3.3.1 Model Architecture . . . 30
  - 3.3.2 Boundary click encoding . . . 30
  - 3.3.3 Training Process . . . 31
- 3.4 Evaluation Metrics . . . 34
- 3.5 Uncertainty Quantification . . . 35
  - 3.5.1 MC Dropout Implementation . . . 35
  - 3.5.2 Uncertainty Metrics . . . 36
- 3.6 Experiments . . . 38
  - 3.6.1 Experiments on Baseline Model . . . 38

4 Baseline Model Results **40**
- 4.1 Segmentation Performance . . . 40
- 4.2 Model Calibration . . . 42
- 4.3 Tuning of Uncertainty Estimation Parameters . . . 45
  - 4.3.1 Structure-wise Correlation . . . 46

4.3.2 Voxel-wise Correlation 47
4.3.3 Analysis of Tuning Curves and Selection of Parameters 47

5 Leveraging Uncertainty on the Baseline Model **52**
5.1 Leveraging Structure-wise Uncertainty Estimates 52
5.1.1 Experiments with simulated perfect corrections 52
5.1.2 Experiments on CT Dataset using observer variability 55
5.2 Leveraging Voxel-wise Uncertainty Estimates 57
5.2.1 Patch Correction Experiments 59
5.3 Integrated Use of Structure-wise and Voxel-wise Uncertainty Estimates 65

6 Interactive Model Results **70**
6.1 BISNet Application on the TPUS Dataset 70
6.1.1 Segmentation Performance 70
6.1.2 Tuning of Uncertainty Estimation Parameters 71
6.2 BISNet Application on the CT Dataset 73

7 Integration of Uncertainty into an Interactive Segmentation Framework **75**
7.1 Clinical Experiment 75

8 Conclusion **78**
8.1 Contributions and Results 78
8.2 Limitations and Future Work 80
8.3 Final Remarks 81

References **82**

List of Figures

2.1 Basic structure of a CNN. The network consists of a feature extraction stage using convolution and pooling layers, followed by a classification stage using fully connected layers to produce the final output. 5

2.2 Overview of a FCN. The input image is passed through convolutional layers to produce pixel-wise class predictions. 5

2.3 A representative example of the original U-Net architecture. The network comprises an encoder path (on the left) and a decoder path (on the right). Blue boxes indicate multi-channel feature maps, with the number of channels shown above each box and spatial dimensions labeled at the bottom left corner. White boxes depict feature maps transferred via skip connections. Arrows illustrate the various operations performed throughout the network. 6

2.4 Illustration of TransU-Net architecture, a hybrid model combining a CNN encoder, Transformer for global context, and a U-Net decoder for precise segmentation. . . 7

2.5 Illustration highlighting the distinction between aleatoric and epistemic uncertainties. The dots on the plot represent observed data points. Aleatoric uncertainty accounts for the randomness or noise naturally present in the data, whereas epistemic uncertainty arises from limited knowledge or insufficient data, reflecting uncertainty that could be reduced with more information. 9

2.6 Illustration of aleatoric and epistemic uncertainties. Left: Data (aleatoric) uncertainty arises from overlapping or noisy regions between training data classes, where inherent ambiguity exists. Right: Epistemic (model) uncertainty appears in areas with limited or no training data, such as near class boundaries or with out-of-distribution inputs, where different models may disagree. 10

2.7 Comparison of model calibration. On the left, the model demonstrates over- and underconfidence in the alignment of predicted and observed frequencies, indicating imperfect calibration. On the right, the model is perfectly calibrated, with a perfect alignment between the predicted and observed frequencies. 11

2.8 Illustration of MC Dropout for uncertainty estimation in CNNs. An input image is passed through a convolutional neural network with dropout layers active during inference. By performing multiple stochastic forward passes (T steps), the model generates a set of output predictions, each with slight variations due to different dropout masks. These predictions are then aggregated to compute an uncertainty map, using statistical measures such as variance, entropy, or mutual information. 13

2.9 Illustration of a deep ensemble approach for uncertainty quantification in image analysis. Multiple models are independently trained and used to generate prediction distributions for each input image. These predictions are aggregated to estimate uncertainty through statistical measures such as variance, entropy, or mutual information, resulting in an uncertainty map highlighting regions of low model confidence. 14
2.10 Overview of TTA for uncertainty quantification. During inference, the input image undergoes multiple augmentations, each of which is passed through the same trained model. The ensemble of predictions is then aggregated, and statistical measures are computed to generate an uncertainty map. 15
2.11 Illustration of interaction types. The green region represents the ground truth, while the red contour indicates the current segmentation, showing both under- and over-segmented areas. Left: Interaction via region clicks. Positive clicks are shown in blue, and negative clicks in yellow. Center: Interaction via scribbles. Positive scribbles in blue, and negative in yellow. Right: Interaction via boundary points. 20
2.12 Illustration of click-based guidance signals. 21
2.13 Illustration of scribble-based guidance signals. 21

3.1 Internal outline of BISNet. In automatic segmentation mode, the guidance channel is empty. In interactive segmentation mode, the guidance channel contains boundary clicks based on the previous segmentation. 30

4.1 Examples of pericardium segmentation in CT slices. In each row, the manual segmentation (GT) and the model's automatic segmentation are shown, with the segmentation highlighted in green. The first row illustrates a case with above-average performance (Dice = 95.23%, ASSD = 1.47 mm), while the second row shows a below-average result (Dice = 77.84%, ASSD = 8.24 mm). 41
4.2 Examples of pericardium segmentation in TPUS slices. In each row, the manual segmentation (GT) and the model's automatic segmentation are shown, with the segmentation highlighted in green. The first row illustrates a case with above-average performance (Dice = 81.03%, ASSD = 1.65 mm), while the second row shows a below-average result (Dice = 64.22%, ASSD = 2.43 mm). 42
4.3 Calibration heatmaps and reliability curves for CT models trained without (left) and with (right) a calibration factor, evaluated on the training set (top) and test set (bottom). In each plot, the x-axis represents the predicted probability that a voxel belongs to the foreground class, and the y-axis shows the empirical frequency with which voxels at that probability level actually belong to the foreground. The red curve represents the average fraction of correct predictions made at each confidence level across the whole dataset. The white dashed line indicates perfect calibration, where predicted probabilities match the observed accuracy. The background heatmap reflects the density of predictions in each bin. The average ACE is reported in the bottom-right corner of each plot. 43

4.4 Calibration heatmaps and reliability curves for TPUS models trained without (left) and with (right) a calibration factor, evaluated on the training set (top) and test set (bottom). In each plot, the x-axis represents the predicted probability that a voxel belongs to the foreground class, and the y-axis shows the empirical frequency with which voxels at that probability level actually belong to the foreground. The red curve represents the average fraction of correct predictions made at each confidence level across the whole dataset. The white dashed line indicates perfect calibration, where predicted probabilities match the observed accuracy. The background heatmap reflects the density of predictions in each bin. The average ACE is reported in the bottom-right corner of each plot. 44
4.5 Variation in Dice score across different MC dropout rates and numbers of MC samples for (a) CT and (b) TPUS segmentation models. 46
4.6 Correlation between uncertainty and segmentation error across different MC dropout rates and numbers of MC samples for the CT segmentation model. Each curve shows how the Spearman correlation varies with uncertainty estimation settings, capturing the relationship between predicted uncertainty and actual segmentation error at the structure and voxel levels. 48
4.7 Uncertainty–performance relationships for the CT segmentation model using the selected MC dropout configuration (dropout rate 0.25, 20 MC samples). (a) Structure-wise correlation between Dice score and volume-normalized summed entropy. Each point represents a volume and the red dashed line shows a linear regression fit. (b) Voxel-wise uncertainty–error profile computed over boundary voxels. Voxels are grouped into uncertainty bins (x-axis), with corresponding misclassification rates (y-axis). The red dashed line indicates the mean misclassification rate across bins. The color scale reflects the occurrence across images, indicating how many times each uncertainty–error pairing appears throughout the dataset. 49
4.8 Correlation between uncertainty and segmentation error across different MC dropout rates and numbers of MC samples for the TPUS segmentation model. Each curve shows how the Spearman correlation varies with uncertainty estimation settings, capturing the relationship between predicted uncertainty and actual segmentation error at the structure and voxel levels. 50
4.9 Uncertainty–performance relationships for the TPUS segmentation model using the selected MC dropout configuration (dropout rate 0.1, 20 MC samples). (a) Structure-wise correlation between Dice score and volume-normalized summed entropy. Each point represents a volume and the red dashed line shows a linear regression fit. (b) Voxel-wise uncertainty–error profile computed over boundary voxels. Voxels are grouped into uncertainty bins (x-axis), with corresponding misclassification rates (y-axis). The red dashed line indicates the mean misclassification rate across bins. The color scale reflects the occurrence across images, indicating how many times each uncertainty–error pairing appears throughout the dataset. 51

5.1 Structure-wise uncertainty experiments results for CT data. (a) Mean Dice score as a function of the number of reviewed CT images. (b) Percentage of poor segmentations not yet reviewed as images are successively corrected, as well as the corresponding percentage in the dataset. In both plots, the purple line represents uncertainty-based selection, the orange line shows random selection (over 50 runs), and the green line represents the ideal Dice-based selection strategy. The shaded region around the random selection curve denotes the standard deviation across the 50 runs. 54

5.2 Structure-wise uncertainty experiments results for TPUS data. (a) Mean Dice score as a function of the number of reviewed TPUS images. (b) Percentage of poor segmentations not yet reviewed as images are successively corrected, as well as the corresponding percentage in the dataset. In both plots, the purple line represents uncertainty-based selection, the orange line shows random selection (over 50 runs), and the green line represents the ideal Dice-based selection strategy. The shaded region around the random selection curve denotes the standard deviation across the 50 runs. 54

5.3 Results of structure-wise uncertainty-guided review on CT data, showing the evolution of mean Dice score as a function of the number of corrected images. (a) Corrections modeled using intra-observer Dice scores, reflecting consistent single-expert review. (b) Corrections modeled using inter-observer Dice scores, simulating multi-expert variability. In both plots, the purple line represents uncertainty-based selection, the orange line shows the mean of 50 runs of random selection (with shaded area indicating standard deviation), and the green line corresponds to the ideal Dice-based oracle selection. The red dashed line shows the original (uncorrected) mean Dice score, while the blue and cyan dashed lines indicate intra- and inter-observer mean Dice scores, respectively. 57

5.4 Visualization of the uncertainty refinement process for two CT segmentation examples (top and bottom rows). Each row shows: (1) the original image; (2) ground truth (green) and predicted (red) segmentation boundaries; (3) raw voxel-wise uncertainty map; (4) uncertainty map after applying a 3×3×3 uniform averaging filter, reducing noise and emphasizing coherent uncertain regions; and (5) final uncertainty map after radial filtering applied only to boundary voxels, suppressing non-boundary uncertainty and enhancing alignment with actual missegmentation areas. The final map serves as input for the correction-guided experiments. 58

5.5 Visualization of the uncertainty refinement process for two TPUS segmentation examples (top and bottom rows). Each row shows: (1) the original image; (2) ground truth (green) and predicted (red) segmentation boundaries; (3) raw voxel-wise uncertainty map; (4) uncertainty map after applying a 3×3×3 uniform averaging filter, reducing noise and emphasizing coherent uncertain regions; and (5) final uncertainty map after radial filtering applied only to boundary voxels, suppressing non-boundary uncertainty and enhancing alignment with actual missegmentation areas. The final map serves as input for the correction-guided experiments. 59

5.6 Illustration of radial entropy smoothing applied to a 3D entropy map. The image shows a 2D slice through the CoM of the segmented structure. The green dot marks the center of mass, the blue dot indicates one example boundary voxel, and the white dashed line represents the radial direction from the center to this boundary point. Red dots denote the sampled points along this radial line, used to compute a local average. The left panel provides a zoomed-in view of the boxed region in the main image. This figure illustrates the process for a single boundary voxel, but the smoothing is applied to all boundary voxels across the segmentation. 60

5.7 Example from the CT dataset illustrating the correction process. (Left) Uncertainty map with the most uncertain boundary point marked in red. (Middle) Current segmentation overlaid with the ground truth boundary in green. The most uncertain point is highlighted in red, and the yellow square shows the 2D slice of the cubic patch centered on that point used for correction. (Right) Updated segmentation after pasting the ground truth within the patch, simulating the targeted correction. 61

5.8 Mean Dice score progression as a function of the number of simulated patch corrections. Three voxel selection strategies are compared: uncertainty-based, random, and distance-based. At each correction step, all images receive the same number of patch corrections. The shaded regions represent ±1 standard deviation across the dataset. 62

5.9 Example from the CT dataset illustrating the adaptive patch size correction. (Left) Uncertainty map with the most uncertain boundary point marked in red. (Middle) Current segmentation overlaid with the ground truth boundary in green. The most uncertain point is highlighted in red. Two patches are shown: the yellow square represents the patch defined by twice the distance to the ground truth boundary, and the blue square indicates the patch based on the estimated average radius of the structure. Since the blue patch is bigger, it is used for the correction. (Right) Updated segmentation after pasting the ground truth within the radius-based patch, simulating the targeted correction. 64

5.10 Mean Dice score progression as a function of the number of simulated corrections across two segmentation models using an adaptive minimum patch size. (a) Results for the CT model. (b) Results for the TPUS model. Three voxel selection strategies are compared: uncertainty-based, random, and distance-based. At each correction step, all images receive the same number of patch corrections. The shaded regions represent ±1 standard deviation across the dataset. 65

5.11 Comparison for the CT dataset between correction strategies using voxel-wise uncertainty alone and a hybrid strategy combining voxel- and structure-wise uncertainty. The bottom x-axis shows the total number of cumulative corrections applied. The top x-axis maps this total to the number of corrections per image in the voxel-wise-only strategy (since all images are corrected at each step in that method). This dual-axis design enables a fair comparison by aligning both curves according to the same total correction effort. (a) Evolution of mean Dice score (solid lines) as a function of the total number of patch corrections applied across the dataset. The shaded areas represent the full range (min to max) of Dice scores across all images. (b) Evolution of the percentage of poor segmentations (i.e., with Dice bellow the inter-observer Dice score) as a function of the total number of patch corrections applied across the dataset. 66

5.12 Comparison for the TPUS dataset between correction strategies using voxel-wise uncertainty alone and a hybrid strategy combining voxel- and structure-wise uncertainty. The bottom x-axis shows the total number of cumulative corrections applied. The top x-axis maps this total to the number of corrections per image in the voxel-wise-only strategy (since all images are corrected at each step in that method). This dual-axis design enables a fair comparison by aligning both curves according to the same total correction effort. (a) Evolution of mean Dice score (solid lines) as a function of the total number of patch corrections applied across the dataset. The shaded areas represent the full range (min to max) of Dice scores across all images. (b) Evolution of the percentage of poor segmentations (i.e., with Dice bellow the inter-observer Dice score) as a function of the total number of patch corrections applied across the dataset. 67
5.13 Evolution of the mean Dice score for images that were initially poor (below inter-observer Dice, shown in red) and good (above inter-observer Dice, shown in green), under the two correction strategies. Solid lines represent the proposed hybrid approach, while dashed lines correspond to the voxel-wise-only baseline. 68

6.1 Exemplification of the BISNet interactive segmentation process on a single case. The model first generates an automatic segmentation (a), which is then refined (c) based on user-provided boundary click (b). 71
6.2 Variation in Dice score across different MC dropout rates and numbers of MC samples using the BISNet on TPUS data. 71
6.3 Correlation between uncertainty and segmentation error across different MC dropout rates and numbers of MC samples using the BISNet in TPUS data. Each curve shows how the Spearman correlation varies with uncertainty estimation settings, capturing the relationship between predicted uncertainty and actual segmentation error at the structure and voxel levels. 72
6.4 Uncertainty–performance relationships for the BISNet using the selected MC dropout configuration (dropout rate 0.1, 15 MC samples). (a) Structure-wise correlation between Dice score and volume-normalized summed entropy. Each point represents a volume and the red dashed line shows a linear regression fit. (b) Voxel-wise uncertainty–error profile computed over boundary voxels. Voxels are grouped into uncertainty bins (x-axis), with corresponding misclassification rates (y-axis). The red dashed line indicates the mean misclassification rate across bins. The color scale reflects the occurrence across images, indicating how many times each uncertainty–error pairing appears throughout the dataset. 73

7.1 Evolution of the mean Dice score across different correction strategies as a function of the number of updates applied. The curves compare uncertainty-based ordering (blue), Dice-based ordering (green), and random ordering averaged over 50 runs (orange). The shaded region represents ±1 standard deviation for the random baseline. 76
7.2 Distribution of voxel-wise entropy values for boundary voxels with displacement above and bellow the mean correction distance. 77

List of Tables

4.1 Segmentation results and observer agreement metrics for the CT and TPUS models 40
4.2 Segmentation results for the CT and TPUS segmentation models trained with and without calibration parameter 45

6.1 Automatic segmentation performance of BISNet compared to the baseline model on the TPUS dataset. 70

Abbreviations and Symbols

ACE	Average Calibration Error
AI	Artificial Intelligence
ASSD	Average Symmetric Surface Distance
BBB	Bayes by Backprop
BNN	Bayesian Neural Network
BISNet	Boundary Interactive Segmentation Network
CHVNGE	Centro Hospitalar de Vila Nova de Gaia e Espinho
CNN	Convolutional Neural Network
CoM	Center of Mass
CT	Computed Tomography
EAS	External Anal Sphincter
ELBO	Evidence Lower Bound
FCN	Fully Convolutional Network
FDA	Food and Drug Administration
GT	Ground Truth
IoU	Intersection over Union
KL	Kullback-Leibler
MC	Monte Carlo
MDR	Medical Device Regulation
NN	Neural Network
OSIC	Open Source Imaging Consortium
ReLU	Rectified Linear Unit
SAM	Segment Anything Model
TPUS	Transperineal Ultrasound
TTA	Test-Time Augmentation
VI	Variational Inference
ViT	Vision Transformer

Chapter 1

Introduction

1.1 Background and Motivation

In recent years, artificial intelligence (AI) has emerged as one of the most discussed and rapidly evolving fields [2, 3], with the potential to transform numerous aspects of healthcare [4, 5]. As healthcare systems continue to evolve, the integration of AI technologies offers promising opportunities to enhance the quality of patient care, improving diagnosis accuracy [6] and streamlining clinical workflows [7]. With convolutional neural networks (CNNs), deep learning has become a powerful AI subfield in medical image analysis and disease diagnosis, covering a variety of tasks [8, 9, 10, 11, 12, 13].

Image segmentation, in the context of medical imaging, involves the precise delineation of anatomical structures, which may facilitate the automated identification of pathological regions. This capability is fundamental to numerous clinical applications, including surgical planning [14], accurate diagnosis [15, 9, 16], and monitoring of disease progression over time [17]. CNNs have achieved state-of-the-art performance in many medical image segmentation tasks and with a variety of different architectures [18, 19, 20, 21, 22]. This achievement is particularly impressive in the medical field, as it presents inherent challenges, including the difficulty in collecting large datasets with pixel-wise annotations for training, the ambiguity and variability of the ground truth, and the high level of skill required for manual segmentation. Additionally, automatic segmentation faces post-deployment data shifts, where differences in image acquisition protocols, biological variability, and pathologies among patients affect model performance [23].

Despite these challenges, CNNs have shown promise for automatic image segmentation and hold value in developing efficient image processing pipelines for research and clinical applications. However, it is important to recognize that fully automated procedures introduce additional responsibilities, since incorrect results can lead to misguided medical decisions and unnecessary interventions, which also reduces clinical acceptability [24]. Moreover, even when trained on extensive datasets and capable of making predictions with high accuracy, these systems may still exhibit some uncertainty in their predictions, especially when faced with low-quality or out-of-distribution

images, which are often difficult to analyse and may contain significant variability [25]. This variability can arise from differences in imaging protocols between vendors and countries, as well as variations in patient populations, including age, ethnicity, and underlying pathologies.

To safely integrate AI algorithms into clinical workflows, legal and ethical considerations must be addressed. Regulatory bodies such as the Food and Drug Administration (FDA) [26] in the United States of America and the Medical Device Regulation (MDR) [27] in the European Union play an important role in shaping the standards for AI technologies in healthcare. The FDA requires a high level of explanation for the development of algorithms that inform medical treatment, including traceability and transparency. Similarly, in 2021, the European Commission published the AI Act to further regulate AI technologies in Europe [28]. This proposal emphasizes safety, classifying all AI-based medical devices as high-risk and underscoring the importance of transparency and trustworthy AI. Both frameworks highlight the need for robust risk management to ensure reliable system performance.

Fostering interaction between AI systems and healthcare professionals, through a human-in-the-loop approach, becomes essential under these regulatory frameworks. Allowing clinicians to intervene or modify algorithm outputs can lead to more reliable results and enhance trust and acceptance of AI technologies in clinical settings.

Nevertheless, it is important to acknowledge that the incorporation of this interactivity alone would not be efficient. Manual review of every image, particularly in 3D, can be time-consuming and unfeasible for practitioners, as they would need to inspect each image to determine if any corrections are necessary. Therefore, it is important to develop methods that can automatically detect problematic segmentations or specific regions within them and guide clinicians through the correction process.

A promising strategy to support this process is uncertainty estimation, which involves quantifying the model's confidence in its predictions. This uncertainty quantification is a key technique of trustworthy AI, advocating for the integration of principles such as explainability, robustness, and accountability into the design and development of AI systems, as outlined in the guidelines [29]. However, this necessity highlights a significant limitation of current deep learning algorithms - they do not inherently provide uncertainty estimation for their segmentation results.

1.2 Objectives

This work is driven by the objective of enhancing the accuracy, reliability, and usability of 3D medical image segmentation in clinical workflows, by addressing the limitations of current automated methods through uncertainty-aware and interactive approaches.

The focus is specifically on 3D images, a domain that is both clinically significant and inherently more challenging. This high dimensionality and the complexity of volumetric data make manual correction especially difficult and time-consuming, highlighting the need for more effective tools.

To this end, the primary goals and methodological steps delineated for this project are as follows:

- **Develop robust baseline segmentation models.** Adapt and optimize state-of-the-art deep learning architectures for accurate segmentation of 3D medical imaging datasets, establishing reliable reference models for subsequent uncertainty analysis and interactive segmentation.
- **Design and implement a comprehensive uncertainty quantification pipeline.** Create methodologies to measure uncertainty at both voxel and structural levels, enabling the assessment of confidence in segmentation outputs.
- **Evaluate uncertainty metrics and their usability in guiding interaction.** Analyse how uncertainty correlates with actual segmentation inaccuracies and conduct experiments to assess the potential of uncertainty-guided intervention to improve segmentation correction efficiency.
- **Develop interactive segmentation frameworks.** Implement segmentation interactive frameworks that allow clinicians to manually refine segmentation results.
- **Incorporate uncertainty estimation within the interactive segmentation framework.** Leverage uncertainty information to assist user interaction to optimize manual intervention and reduce user burden.

These methods were applied and validated in two specific applications - 3D transperineal ultrasound (TPUS) imaging for external anal sphincter (EAS) segmentation and cardiac computed tomography (CT) for pericardial segmentation.

Chapter 2

Bibliographic Review

This chapter introduces and explains the key concepts addressed throughout this work, providing the necessary theoretical background for the methods and analyses that follow. It also presents an overview of previous and current research relevant to this work, with a focus on image segmentation, uncertainty estimation, and interactive segmentation methods. Fundamental approaches, recent advancements, and ongoing challenges in these areas are reviewed to establish a foundation for the methodologies explored in subsequent chapters. By analyzing existing literature, this review highlights the gaps and opportunities that motivate the proposed contributions and justify the methods employed.

2.1 Segmentation

Over the years, several segmentation methods have been proposed, evolving from low-level approaches to more complex and sophisticated techniques. Traditional methods, such as thresholding [30], region-growing [31] and active contours [32], rely on predefined heuristics to delineate structures of interest. These approaches do not include a learning component, instead using rules to narrow down search regions, initialize segment boundaries, or guide more complex strategies. While effective in controlled scenarios, traditional methods suffer from several limitations, such as high sensitivity to noise, poor generalization across datasets, difficulty handling subtle transitions between regions, strong dependence on initialization quality, and difficulty segmenting complex or irregularly shaped structures [33, 34].

With the rise of deep learning in recent years, CNNs have revolutionized the field of image analysis due to their ability to automatically learn hierarchical features from images through convolutional operations. Originally introduced in the 1990s, CNNs consist of convolutional layers that automatically extract spatial features from raw images, followed by pooling layers and, typically, fully connected layers for classification, which act as classifiers by mapping learned features to output categories [35]. An illustration is presented in 2.1. They gained widespread popularity after the introduction of AlexNet in 2012 [36], which demonstrated the power of deep CNNs trained on large-scale datasets. AlexNet introduced innovations such as ReLU activations,

dropout regularization, and GPU acceleration, which significantly boosted performance and scalability. However, the fully connected layers fix the input size and discard spatial resolution, which is unsuitable for pixel or voxel-level prediction.

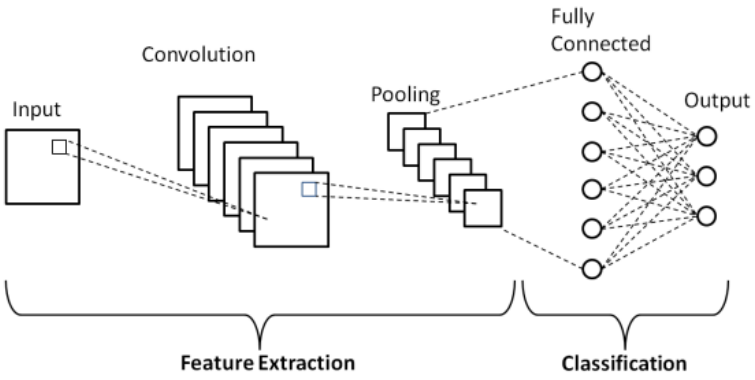


Figure 2.1: Basic structure of a CNN. The network consists of a feature extraction stage using convolution and pooling layers, followed by a classification stage using fully connected layers to produce the final output. Image from [37].

Building upon CNNs, Fully Convolutional Networks (FCNs) replaced the final fully connected layers with convolutional layers, allowing the network to produce dense, pixel or voxel-wise predictions for input images of arbitrary size [18]. FCNs introduced upsampling mechanisms to restore spatial resolution, enabling end-to-end training for semantic segmentation tasks. An overview of a FCN is presented in 2.2. However, early FCNs lacked mechanisms to capture global context, which limited their ability to segment objects with similar local features. To address this, methods like ParseNet proposed enhancing FCN features by integrating global context, for instance by computing global average features and fusing them with local activations [38].

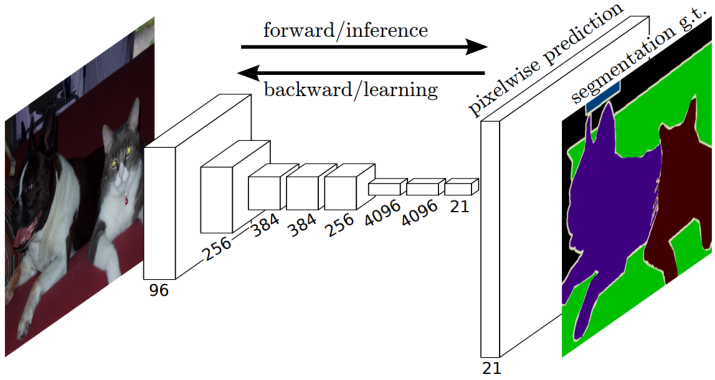


Figure 2.2: Overview of a FCN. The input image is passed through convolutional layers to produce pixel-wise class predictions. Image from [18].

A further breakthrough came with U-Net [19], an encoder-decoder architecture based on FCNs, specifically designed for biomedical image segmentation. U-Net adopts a symmetric encoder-decoder structure, forming a characteristic "U" shape. The encoder path utilizes convolution and

max-pooling layers to extract features within an image while reducing spatial dimensions. In contrast, the decoder path uses convolution and up-convolution (transposed convolution) operations to progressively upsample the feature map, effectively combining learned features and restoring spatial resolution to generate the final segmentation output. Skip connections connect encoder layer outputs to corresponding decoder layers, allowing the model to reuse high-resolution features lost during downsampling. This helps the decoder reconstruct more accurate and detailed segmentation maps by combining fine-grained spatial information with deeper semantic context. Figure 2.3 illustrates a typical example of the U-Net architecture, showing the encoder-decoder structure and the use of skip connections.

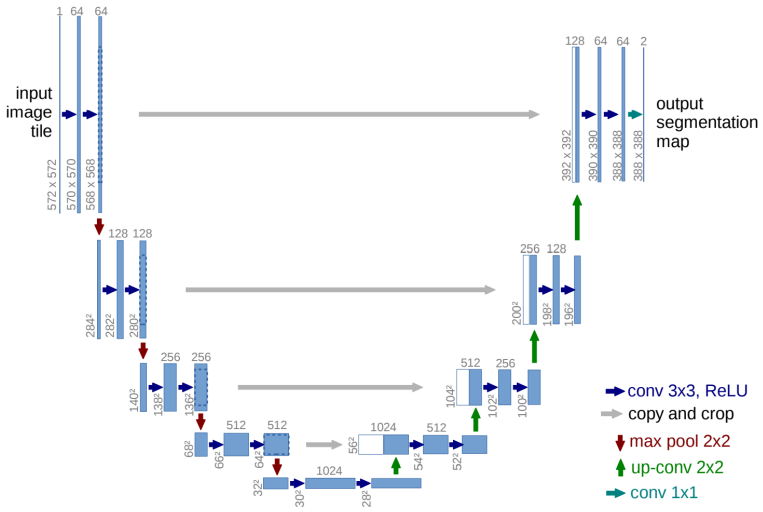


Figure 2.3: A representative example of the original U-Net architecture as proposed by [19]. The network comprises an encoder path (on the left) and a decoder path (on the right). Blue boxes indicate multi-channel feature maps, with the number of channels shown above each box and spatial dimensions labeled at the bottom left corner. White boxes depict feature maps transferred via skip connections. Arrows illustrate the various operations performed throughout the network. Image from [19].

To extend segmentation to 3D volumetric data, especially in medical imaging, architectures like 3D U-Net [39] and V-Net [40] were developed. 3D U-Net adapts U-Net by replacing all 2D operations with their 3D counterparts, enabling context aggregation across depth slices. Similarly, V-Net is a 3D segmentation model that includes residual connections, which consists of connecting the inputs and outputs from each stage to facilitate the learning of residual functions. This allows gradients to flow directly through the shortcut path, improving optimization.

Studies employing U-Net architectures have significantly improved the precision and efficiency of segmentation tasks in medical imaging. Currently, some of the best-performing segmentation networks, such as nnU-Net [41] and U-Net++ [42] are based on the U-Net framework.

The U-Net++ was designed to address the semantic gap between encoder and decoder features. This way, it inserts convolutional blocks along the skip paths, progressively refining encoder features before passing them to the decoder. This bridges the semantic gap and improves feature

fusion.

The nnU-Net has set a new benchmark by addressing the fact that designing and tuning U-Net-based pipelines can be complex, often requiring extensive parameter optimization and expert-driven decisions. Therefore, it consists of a self-adapting framework that automatically adjusts its architecture and data processing strategies to suit the input data. By leveraging dataset characteristics, nnU-Net applies heuristic rules to guide the automatic configuration of the segmentation pipeline. In general, nnU-Net allows for state-of-the-art results with minimal tuning and demonstrates strong generalization across diverse domains [41]. Its robust and adaptive design has established it as a new standard in medical image segmentation.

More recently, the field has embraced transformer-based models, such as Vision Transformers (ViTs). Transformers, originally developed for natural language processing, leverage self-attention to capture global context. ViTs adapt this idea for vision tasks by splitting an image into patches and learning relationships between them. Transformers lack the spatial hierarchy CNNs naturally build. Therefore, for dense prediction tasks like segmentation, ViTs are often hybridized with CNNs. An example of this is TransU-Net [43], which is a hybrid architecture that combines the strengths of CNNs and transformers for medical image segmentation. It uses a CNN encoder to extract low-level spatial features and reduce image dimensionality, which are then flattened into a sequence of tokens and passed through a ViT to model long-range dependencies and global context. After the transformer, the feature sequence is reshaped back into a spatial representation and fed into a U-Net-like decoder, which upsamples and combines features via skip connections from the CNN encoder. This model often outperforms other CNN-based methods in complex segmentation tasks by capturing global context more effectively. An overview of the TransU-Net architecture is presented in 2.4.

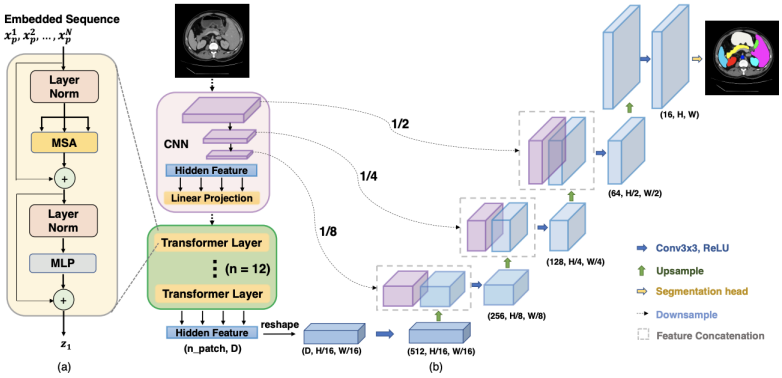


Figure 2.4: Illustration of TransUNet architecture, a hybrid model combining a CNN encoder, Transformer for global context, and a U-Net decoder for precise segmentation. Image from [43].

A current frontier in image segmentation is marked by the emergence of foundation models, which are large-scale, pre-trained architectures trained on diverse datasets using self-supervised or weakly supervised learning. These models, such as SAM (Segment Anything Model) [44], are designed to generalize across a wide range of tasks and domains through prompt-based interactions

rather than retraining. Building on this paradigm, MedSAM adapts SAM for the medical domain by fine-tuning it for a variety of medical imaging modalities [45]. Unlike traditional models that require task-specific labeled datasets, MedSAM supports zero-shot and few-shot segmentation, enabling high-quality delineation of anatomical structures even in previously unseen data. These models can adapt to new domains and tasks without retraining, making them powerful tools for both medical and general-purpose segmentation applications. However, besides offering a strong generalization, they may lack task-specific precision and rely heavily on prompt quality for accurate segmentation.

2.2 Uncertainty Quantification

Uncertainty quantification refers to the process of predicting confidence, or lack of it, associated with neural network predictions, thereby assisting in the identification of reliable outputs suitable for clinical use and indicating when extra caution is necessary [46]. This is particularly crucial in high-stakes domains such as clinical applications, where understanding how much trust to place in a model's output can guide decision-making.

Uncertainty can be interpreted and computed at different levels, depending on the task and the required granularity. The motivation for using uncertainty in deep learning stems from the need to improve model interpretability, allowing users to assess the reliability of the model's predictions. By capturing uncertainty at various levels, such as voxel-wise, volume-wise, or slice-wise, uncertainty helps ensure that predictions are not just seen as binary outcomes but as estimates that come with associated confidence. This comprehensive approach ultimately aids in the more accurate and reliable application of machine learning in critical fields like medical imaging.

For tasks like segmentation, voxel-wise uncertainty is particularly insightful and is typically obtained directly by applying uncertainty estimation methods. Uncertainty at this level provides a fine-grained view of confidence at the level of individual voxels. This degree of detail allows clinicians to pinpoint specific regions where the model's predictions may be less trustworthy, ultimately supporting more informed interpretation and decision-making. Additionally, volume-wise uncertainty can also be derived, which offers an overall estimate for an entire scan, making it useful for flagging cases where the model's predictions may be unreliable as a whole. This can be further refined through slice-wise uncertainty, which highlights the most uncertain slices within a volume, offering more localized insights.

One significant challenge in uncertainty estimation is the lack of Ground Truth (GT) for uncertainty in many medical scenarios. Unlike segmentation tasks where a clear label is available, uncertainty itself cannot always be easily quantified or verified. This makes it difficult to accurately assess how well the model's uncertainty predictions align with the actual uncertainty and complicates the evaluation of different uncertainty estimation methods. In addition, uncertainty estimation in deep models can be computationally demanding, especially for large and complex models [47].

It is important to understand that there are two major types of predictive uncertainties for deep CNNs - aleatoric uncertainty and epistemic uncertainty [46]. Aleatoric uncertainty, also known as data uncertainty, captures the inherent noise and randomness in the data itself. This type of uncertainty arises from factors such as acquisition noise, motion artifacts, biological variability, or ambiguous annotations, resulting in a level of unpredictability that cannot be reduced even with additional data. In contrast, epistemic uncertainty, also referred to as model uncertainty, reflects uncertainty in the model parameters, capturing the lack of knowledge about the true data-generating process. Therefore, it indicates the degree of uncertainty of the algorithm itself regarding its predictions, which often stems from insufficient training data or limitations in the model architecture itself. Unlike aleatoric uncertainty, epistemic uncertainty can be reduced by collecting more data. However, this is challenging, as obtaining more samples and performing manual annotation is a long and resource-intensive process. Figure 2.5 and Figure 2.6 present illustrations that highlight the distinction between the two types of uncertainty.

Aleatoric uncertainty can be modeled as a function of the input data, often by training the model to predict a distribution over possible outcomes, resulting in probabilistic outputs. In contrast, estimating epistemic uncertainty involves learning a probability distribution over the model's parameters, capturing uncertainty about the model itself [46]. By employing uncertainty estimation, high predicted uncertainty values could potentially flag incorrectly segmented regions, guiding user interactions and necessary refinements. If effective, this approach may enhance the decision-making process, help mitigate risks, and support the appropriate involvement of medical professionals in critical cases, paving the way for a more reliable and effective integration of AI technologies into healthcare.

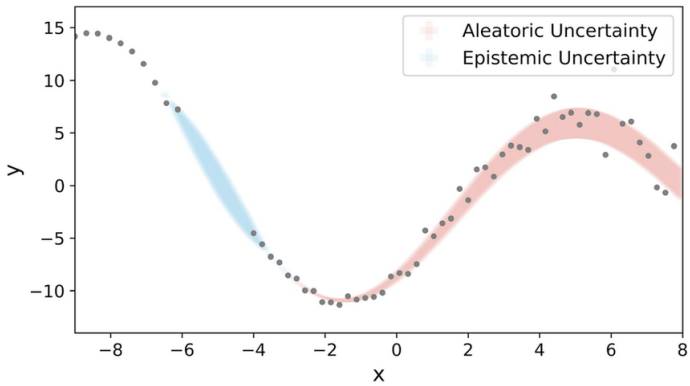


Figure 2.5: Illustration of aleatoric and epistemic uncertainties. The dots on the plot represent observed data points. Aleatoric uncertainty accounts for the randomness or noise naturally present in the data, whereas epistemic uncertainty arises from limited knowledge or insufficient data, reflecting uncertainty that could be reduced with more information. Image from [48].

2.2.1 Uncertainty Quantification Methods

Several methods have been proposed in the literature to estimate uncertainty in deep learning models. In this section, key frameworks are explored, including Softmax-based approaches, Bayesian

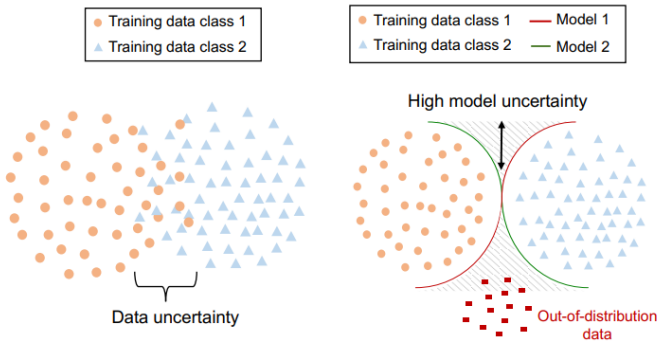


Figure 2.6: Illustration of aleatoric and epistemic uncertainties. Left: Data (aleatoric) uncertainty arises from overlapping or noisy regions between training data classes, where inherent ambiguity exists. Right: Epistemic (model) uncertainty appears in areas with limited or no training data, such as near class boundaries or with out-of-distribution inputs, where different models may disagree. Image from [47].

methods, Dropout-based techniques, Ensemble methods, Test-Time Augmentation (TTA) and Learned Uncertainty frameworks.

It is important to note that most of these frameworks do not produce uncertainty estimates directly. Instead, they generate a set or distribution of predictions from which uncertainty metrics, such as pixel or voxel-wise variance, entropy, or mutual information, can be derived [49]. These metrics enable the construction of uncertainty maps that visualize the spatial distribution of model confidence and help assess the reliability of predictions.

2.2.1.1 Softmax probabilities

An immediate and intuitive method is to use the softmax probabilities produced by a neural network. This approach relies on the assumption that the model is perfectly calibrated.

Calibration in Deep Learning refers to the alignment between a model's predicted probabilities and the actual observed frequencies of those predictions. A well-calibrated model produces probability estimates that accurately reflect real-world likelihoods [50]. For example, in a perfectly calibrated segmentation model, if 100 voxels are predicted with a confidence of 0.8, approximately 80 of them should truly belong to the target class, which makes model predictions more reliable and trustworthy. If a model is not well calibrated, it may be overconfident (assigning high probabilities to incorrect predictions) or underconfident (assigning low probabilities to correct predictions). This behavior is exemplified in Figure 2.7, where the relationship between predicted and observed frequencies is illustrated for a perfectly and not perfectly calibrated model.

However, modern neural networks tend to be miscalibrated. To address this mismatch, various probability calibration methods have been proposed. These approaches aim to transform raw softmax outputs into more faithful estimates of predictive uncertainty. Among the most commonly used techniques is temperature scaling [50], a post-hoc calibration method where the logits are

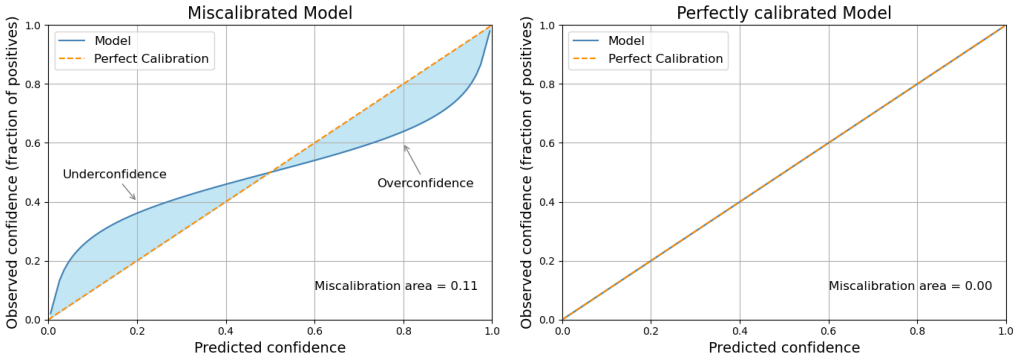


Figure 2.7: Comparison of model calibration. On the left, the model demonstrates over- and underconfidence in the alignment of predicted and observed frequencies, indicating imperfect calibration. On the right, the model is perfectly calibrated, with a perfect alignment between the predicted and observed frequencies.

divided by a learned scalar temperature parameter before applying softmax, adjusting the confidence of the output distribution. In addition, more sophisticated strategies involve integrating calibration objectives directly into the training loss to produce models that are inherently better calibrated [51].

Due to its simplicity, softmax-based uncertainty estimation has been explored and often serves as a baseline for evaluating more complex uncertainty estimation techniques [49]. For example, in [52] the Maximum Softmax Probability is proposed as a simple uncertainty score where lower maximum probabilities signal higher uncertainty, while in [53] the entropy of the softmax distribution is used as a baseline method to quantify prediction confidence. These methods only capture uncertainty in the model's predictions, not in the model's parameters. As such, they primarily reflect aleatoric uncertainty.

Given that most deep learning models are not inherently well-calibrated, accurate uncertainty estimation methods are essential to complement raw softmax probabilities.

2.2.1.2 Bayesian Neural Networks methods

In contrast, a more complex approach to estimating uncertainty is offered by Bayesian Neural Networks (BNNs). In Bayesian Deep Learning, each model weight is represented by a probability distribution rather than a fixed value [54]. Therefore, during training, the model attempts to learn the posterior distribution over weights given the training data, denoted as $p(W|X,Y)$, where W represents the network weights given data X and labels Y. Then during inference, the predictions are obtained by marginalizing over this posterior distribution.

However, computing the exact posterior is intractable for deep neural networks due to the high dimensionality of the parameter space [54]. To make Bayesian inference tractable in this context, Variational Inference (VI) has been widely adopted. VI approximates the true posterior with a parameterized variational distribution $q(W)$, typically chosen for its computational convenience [55]. The parameters of this distribution $q(W)$ are approximated by minimizing the

Kullback-Leibler (KL) divergence between $q(W)$ and the true posterior. This is equivalent to minimizing the variational free energy, also known as the Evidence Lower Bound (ELBO) [56]:

$$q^*(W) = \arg\min_{q(W)} KL(q(W) \parallel p(W \mid X, Y)) \tag{2.1}$$

Crucially, this minimization can be performed via gradient descent, similar to standard neural network training, a method commonly referred to as Bayes by Backprop (BBB) [56]. This reformulation transforms Bayesian learning into an optimization problem, making it more compatible with existing deep learning frameworks. Then, after training, predictions can be generated by sampling weights from the learned distributions, allowing multiple stochastic forward passes through the model to quantify predictive uncertainty.

Considering that it places a distribution on the model's weights, the BNN is used for epistemic uncertainty estimation. However, it introduces considerable computational overhead during both training and inference, and often requires architectural and training modifications to accommodate weight distributions [49].

2.2.1.3 Monte Carlo Dropout methods

An alternative approach to estimating uncertainty in a deep learning model is the Monte Carlo (MC) Dropout, which provides a practical and computationally efficient way to estimate model uncertainty without requiring a fully Bayesian network.

In [57], it is showed that training a network with dropout can approximate Bayesian inference in deep learning models. Dropout is a regularization technique that involves randomly dropping activations during training. In their work, based on these insights, the authors proposed the MC Dropout, where dropout is kept active during inference and the model is run multiple times with different dropout masks. This results in a distribution of outputs, allowing the model to estimate a predictive distribution, which allows the network to approximate the uncertainty typically captured by BNNs [57]. An illustration of the MC Dropout pipeline is presented in Figure 2.8.

In MC Dropout, $q(W)$ represents a distribution where weights are randomly dropped out based on Bernoulli random variables, aligning with the dropout mechanism. This can be interpreted as choosing the approximate distribution to be a mixture of two Gaussians with minimal variances, where one Gaussian is centered at zero and the other at the learned weight. By using dropout also at testing, it is possible to sample from the posterior distribution of the model.

MC Dropout has the advantage of being implemented within standard neural networks with minimal modification, allowing any dropout-trained model to approximate Bayesian behavior. This straightforward implementation has contributed to its rapid adoption and increasing popularity in the field. In fact, initial experiments with MC Dropout [57, 46] demonstrated that averaging the predictions from multiple dropout samples improves classification accuracy and yields better-calibrated uncertainty estimates.

Building on this idea, the authors in [58] developed a BNN with integrated MC Dropout to provide epistemic uncertainty feedback in image segmentation. Their findings showed that the

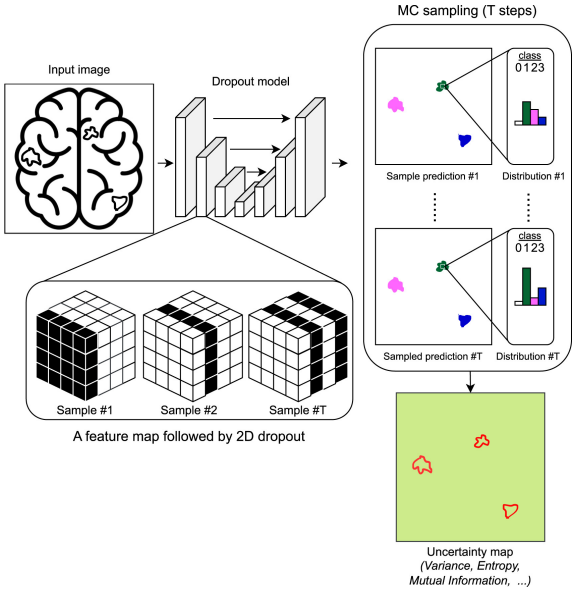


Figure 2.8: Illustration of MC Dropout for uncertainty estimation in CNNs. An input image is passed through a convolutional neural network with dropout layers active during inference. By performing multiple stochastic forward passes (T steps), the model generates a set of output predictions, each with slight variations due to different dropout masks. These predictions are then aggregated to compute an uncertainty map, using statistical measures such as variance, entropy, or mutual information. Image from [49].

estimated uncertainty inversely correlated with the model's performance, which means that regions with higher uncertainty often aligned with segmentation errors. This highlights MC Dropout's potential in practical applications where identifying uncertain or unreliable predictions.

Several studies have proposed improvements and variations to the standard MC Dropout technique. For instance, in [53] the Concrete Dropout is explored, a variant where the dropout rate at each layer is learned as part of the optimization process [59]. However, no significant improvement over standard dropout was obtained in this work. Other stochastic regularization techniques have also been proposed as alternatives, such as Monte Carlo DropConnect [60], which randomly drops weights instead of activations, and Monte Carlo Batch Normalization [61], which incorporates uncertainty into batch normalization parameters. In [62], the impact of both dropout type (Bernoulli and Gaussian) and dropout probability was evaluated. While the type of dropout had minimal influence on performance, the dropout rate was found to be a critical hyperparameter, strongly affecting both segmentation quality and uncertainty estimates.

2.2.1.4 Ensembling methods

Another widely adopted approach for uncertainty estimation is the use of Deep Ensembles [63]. In this method, multiple neural networks are trained independently with different random initializations, allowing each model to converge to a different local optimum. This diversity in model predictions enables the ensemble to estimate epistemic uncertainty by capturing variability across models.

While Deep Ensembles have shown strong empirical performance, training multiple networks from scratch is computationally expensive, especially for large or complex models. Despite this drawback, ensembles have been studied in medical imaging applications, with some works [64, 65] showing superior predictive performance and better-calibrated uncertainty compared to methods like MC Dropout.

To mitigate the computational overhead, recent research has focused on developing efficient ensembling strategies that preserve the benefits of standard ensembles. One such strategy involves multi-output architectures, where a single model produces a diverse set of predictions through multiple branches or heads, effectively simulating an ensemble in one forward pass [66].

Moreover, some studies propose combining ensembles with MC Dropout, resulting in Ensemble MC methods [67, 68] , which involve using several different MC Dropout models in parallel. This hybrid approach leverages both model-level and weight-level stochasticity to enhance uncertainty estimation. The motivation behind this is to integrate the benefits of both techniques, the diversity of model parameters from ensembling and the internal stochasticity of dropout during inference.

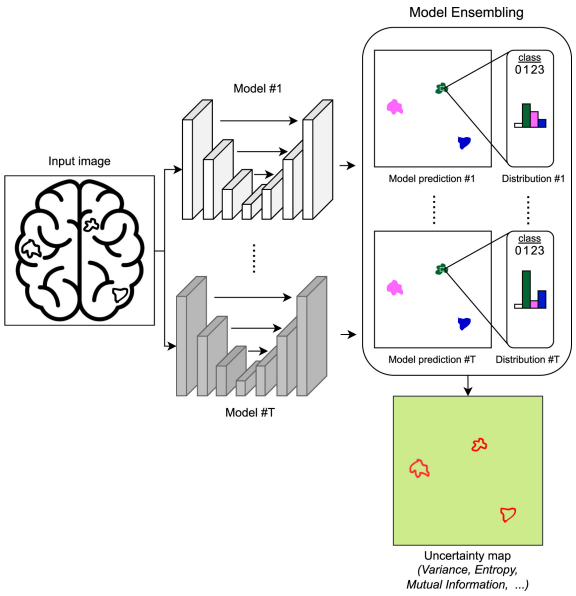


Figure 2.9: Illustration of a deep ensemble approach for uncertainty quantification in image analysis. Multiple models are independently trained and used to generate prediction distributions for each input image. These predictions are aggregated to estimate uncertainty through statistical measures such as variance, entropy, or mutual information, resulting in an uncertainty map highlighting regions of low model confidence. Image from [49].

2.2.1.5 Test Time Augmentation

Test-Time Augmentation (TTA) is another technique used for uncertainty estimation, particularly to assess aleatoric uncertainty [69]. The idea is straightforward; at inference time, multiple augmented versions of the input image are generated using standard data augmentation techniques,

such as spatial and intensity transformations. The model then predicts each augmented version independently and the variability in these predictions forms the basis for estimating uncertainty.

One of TTA's major advantages is that it is completely model-agnostic. It does not require any changes to the model architecture or training procedure, making it compatible with any pre-trained model. This makes TTA especially attractive in clinical settings, where retraining models are often infeasible. Numerous TTA-based strategies have been proposed for medical image analysis, demonstrating its potential for practical deployment [70, 69].

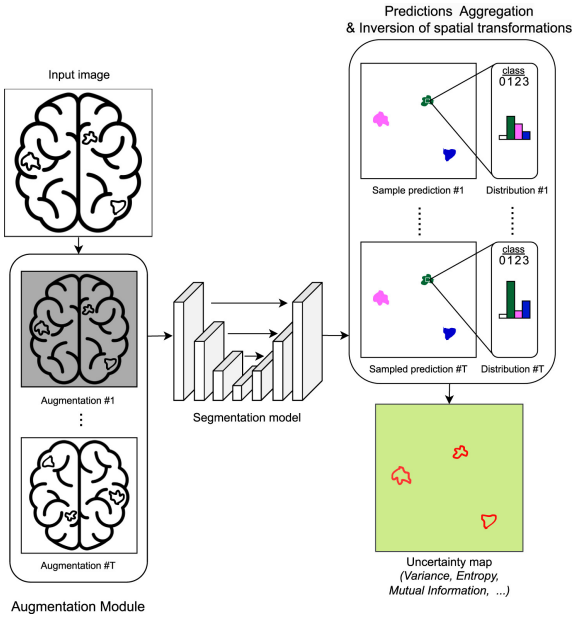


Figure 2.10: Overview of TTA for uncertainty quantification. During inference, the input image undergoes multiple augmentations, each of which is passed through the same trained model. The ensemble of predictions is then aggregated, and statistical measures are computed to generate an uncertainty map. Image from [49].

2.2.1.6 Learned Uncertainty frameworks

Beyond sampling-based approaches, Learned Uncertainty frameworks offer a different idea. Instead of generating prediction distributions at inference time, the model itself learns to predict uncertainty directly during training [71, 72].

One common strategy in medical image segmentation is to leverage inter-rater variability as ground truth for uncertainty [73, 74]. In this setup, multiple expert annotations are treated as a distribution, and a model is trained using supervised learning to reproduce this variability. While this approach allows for learning aleatoric uncertainty directly from the data, it is limited to datasets that include multiple annotations per image, which is often not the case in practice.

To overcome this limitation, most approaches have developed strategies to jointly learn the segmentation and uncertainty without requiring explicit uncertainty labels. One influential method assumes that the network output logits are corrupted by Gaussian noise with mean zero and a learnable variance [75]. Here, a higher predicted variance corresponds to higher aleatoric uncertainty.

Therefore, the model has two output heads, one for the predicted logits and one for the predicted noise variance. The learning process uses a modified cross-entropy loss that incorporates the predicted variance, allowing the model to learn both the segmentation and its associated uncertainty simultaneously. This framework was recently extended to skewed Gaussian distributions to better model asymmetric uncertainty [76]. Instead of assuming that the uncertainty around a prediction is symmetrically distributed, the model learns an additional skewness parameter that captures directional bias in the uncertainty, for example, when one side of a contour is consistently more ambiguous than the other.

Another direction consists of learning uncertainty directly from model error signals [77]. In this case, the model uses its own prediction errors, computed as the difference between predicted labels and ground truth labels during training, as a proxy for uncertainty. These error maps are treated as pseudo ground truth uncertainty maps. An additional output head is trained to predict these error-derived maps, often through an uncertainty-augmented loss function. This strategy enables the model to learn spatially aligned uncertainty estimates that correlate with its own weaknesses, even without external uncertainty labels.

2.2.2 Structure-level Uncertainty

Although previous methods focus on estimating voxel-wise uncertainty, a frequent and practical concern is whether the overall prediction can be trusted. To address this, structure-level uncertainty scores can be computed to give users a global sense of the model's confidence for a given input.

2.2.2.1 Aggregation of Voxel-Wise Uncertainty

One intuitive strategy involves aggregating voxel-wise uncertainty estimates - derived from methods like MC dropout, Deep Ensembles or TTA - into a single structure-level score, often using the mean. This is motivated by the intuition that an overall higher voxel-wise uncertainty should be an indicator of poor segmentation performance [53]. Recognizing that voxel uncertainty tends to be higher at the class boundaries, in [53] it is proposed to reduce the influence of the contour voxels during aggregation to obtain more accurate structural uncertainty estimates. For this effect, three different weightings are tested, specifically masking out voxels at the segmentation boundary, penalizing uncertainties close to the boundary and up-weighting uncertainties distant from the segmentation boundary, and normalizing uncertainty with the segmentation volume. Among these, distance-based weighting was found to most accurately reflect segmentation reliability, reducing the misleading influence of high boundary uncertainty and improving correlation with true performance.

2.2.2.2 Distributional Agreement Across Predictions

Moving beyond voxel-level fusion, other approaches leverage the set of plausible segmentation masks generated by uncertainty estimation methods. A pioneering study in this area [78] used MC dropout to produce multiple segmentation samples and measured their disagreement as an

indicator of uncertainty. With this propose, several proxy metrics were presented, such as the Coefficient of Variation of segmented volumes, which considers how much the segmented volume varies across predictions, as well as the Dice Score and Intersection over Union (IoU) agreement among samples. These metrics showed strong correlation with the true Dice score. This concept was further extended using various uncertainty estimation methods, alongside new proxy metrics like Average Symmetric Surface Distance (ASSD) Within Samples [65] and the Predictive Dice Coefficient [79].

2.2.2.3 Supervised Uncertainty Prediction

To enhance structure-level uncertainty estimation, some studies also used these metrics as features in supervised learning models trained to directly regress the expected segmentation quality, offering a data-driven path to reliable volume-level uncertainty estimation. For example, some works have leveraged linear regression models to estimate the expected Dice score based on uncertainty outputs from MC dropout–based segmentation models [80, 81].

2.3 Interactive Segmentation

Interactivity in segmentation, as defined in this work, refers to a process where the user can correct a segmentation generated by an automated model. This involves the user providing input to adjust the initial segmentation produced by the model. Based on this input, the model generates a new segmentation that incorporates the correction, ensuring that the updated result aligns more closely with the user's expectations. Therefore, this process combines the benefits of automatic segmentation with the precision and reliability of manual segmentation. However, it is important to note that this is the definition used in this context. In other works, interactive segmentation is also applied to different concepts, such as using input to initially guide the model with a spatial cue, but not necessarily allowing further input to refine or alter the segmentation. This type of interaction, though valuable, is not the focus of the current discussion.

Interactive segmentation is a topic of increasing study [82], driven by the aim of making this interaction as easy and fast as possible while maintaining high-quality segmentation results.

These systems involve several key aspects that are essential for their effective implementation, particularly in clinical settings. Firstly, the interaction must be user-friendly, allowing clinicians to engage with the system easily and intuitively, thereby minimizing training time. Additionally, the system's response to user input must be fast, ensuring that corrections are incorporated with minimal delay, which is essential to maintain workflow efficiency. Once the user provides input, the segmentation must be responsive, meaning the system should correctly interpret the corrections and update the segmentation accordingly. These updates should make sense biologically and remain consistent with the rest of the segmentation to avoid introducing irregular and unrealistic shapes. Finally, the incorporation of interactivity must not significantly affect the accuracy of the initial automatic segmentation. Otherwise, a large number of corrections would be required, which

could decrease both the system's overall effectiveness and user satisfaction, making the interactive process more frustrating.

There are various ways for users to interact with the segmentation model. Each type of input requires the model to interpret and act on it differently, demonstrating the importance of robust design in developing interactive segmentation systems.

2.3.1 Types of Interaction

In this subsection, the most popular frameworks for interactive segmentation are reviewed and grouped according to the type of interaction they support. Figure 2.11 exemplifies the different forms of user interaction described in this subsection.

2.3.1.1 Region Clicks

In this approach, the user provides positive or negative clicks on the image to guide the segmentation model. A positive click indicates that the clicked region should be included in the segmented object, while a negative click signals that the region should be excluded.

A key challenge in region-click-based interaction lies in the fact that segmentation errors often span entire regions of pixels rather than isolated points. Therefore, it is essential that the system can interpret a single click (or a minimal number of clicks) as a cue to correct a larger mis-segmented area. However, region clicks are often spatially limited in influence, typically affecting only the segmentation output within a local neighborhood around the point of interaction [83].

To address this limitation, some works have proposed architectures that are explicitly designed to integrate multiple region clicks and generalize their intent beyond their immediate vicinity. In [84], an FCN is trained to accept a variable number of positive and negative clicks as input, alongside the image, and to generate segmentations that accurately reflect the provided annotations. The model is trained to respond appropriately to the distribution and polarity of the clicks across the image. Furthermore, in [85], the authors propose a click attention mechanism aimed at overcoming the local impact limitation of region clicks. Their approach analyzes the similarity between the click locations and the broader image context, allowing the model to better infer the user's intended segmentation. This enhances the model's ability to propagate user intent across larger regions, improving the interactivity and reducing the effort required for segmentation refinement.

2.3.1.2 Scribbles

Scribbles are an intuitive form of user interaction commonly employed in interactive segmentation models. They involve drawing simple lines or strokes over the image to indicate regions that require correction. Scribbles can be positive, pointing out undersegmented areas, or negative, highlighting oversegmented regions.

This type of interaction offers significant flexibility, allowing users to guide the segmentation process with minimal effort. Unlike pixel-wise annotations or precise boundary delineations,

scribbles do not require high precision and are thus more user-friendly. However, the simplicity and sparsity of the scribbles also introduce challenges. Their coarse nature can lead to ambiguity in interpretation by the model, especially when scribbles do not clearly indicate the desired refinement. Since scribbles do not necessarily cover key structural features, the model may struggle to infer the correct corrections. Additionally, user inconsistency is a common issue, because scribble patterns can vary significantly across individuals, affecting model generalization and responsiveness [86].

In [87], an interactive segmentation method is proposed that uses two CNNs, the AutoCNN for generating the initial segmentation and InterCNN for refining it. InterCNN takes three inputs - the original image, the initial segmentation, and the user-provided scribbles — and outputs a corrected segmentation. In another work [86], ZScribbleNet is introduced as a unified framework for scribble-supervised segmentation. The authors first analyze the design of efficient scribble annotations, aiming to maximize the supervisory signal without increasing user effort. This includes strategies to optimize both the placement and the informational content of the scribbles, enabling robust model training at a minimal annotation cost, which is a novel contribution in the field. Subsequently, ZScribbleNet incorporates spatial priors to enhance the correction process. These priors encode feature similarity between pixels and help rectify the shape of predicted structures, enforcing anatomically and structurally consistent segmentations.

2.3.1.3 Boundary Clicks

Boundary clicks are user interactions that indicate specific points through which the segmentation boundary should pass. This approach provides a more intuitive and precise mechanism for humans to correct segmentation results, as it directly communicates the intended contour rather than relying on region-based cues. Although the model is still responsible for interpreting the click information and reason about how to modify the contour accordingly to correct the segmentation, boundary clicks offer a more explicit and structured signal than region-based inputs like interior/exterior clicks or scribbles, providing clear information on where the boundary should be and reducing ambiguity.

However, this method is sensitive to the accuracy of the clicks, since slight deviations from the actual boundary can mislead the model. Another limitation is its reduced effectiveness in removing unwanted segmentation components. In single-component cases, this can often be addressed with simple post-processing (e.g., selecting the largest connected component), but in multi-component scenarios, additional user input may be necessary.

In [88], an interactive framework is proposed in which users refine segmentation boundaries using boundary clicks, integrated into a B-spline Explicit Active Surfaces framework [89]. Here, the contour is modeled as a B-spline curve, and refined by minimizing an energy function that enforces consistency with the CNN output, local image features, and the user-provided boundary points (when provided). This ensures that the final delineation is both accurate and anatomically plausible, while allowing real-time interaction.

In another work [90], the authors introduce iW-Net, a two-stage deep network for automatic and interactive lung nodule segmentation. In the interactive phase, the model requires two boundary points, specifically the end-points of a diameter stroke across the nodule. These points are used to compute a physics-inspired weight map, which guides the network both through a feature input and as part of the loss function. Unlike traditional boundary click methods, this interaction uses the geometric relationship between two opposing points to guide refinement, rather than treating each point independently.

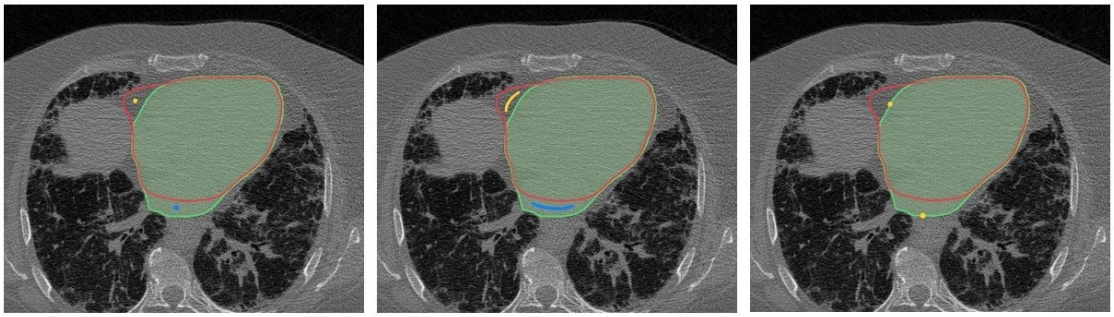


Figure 2.11: Illustration of interaction types. The green region represents the ground truth, while the red contour indicates the current segmentation, showing both under- and over-segmented areas. Left: Interaction via region clicks. Positive clicks are shown in blue, and negative clicks in yellow. Center: Interaction via scribbles. Positive scribbles in blue, and negative in yellow. Right: Interaction via boundary points.

2.3.2 Encoding User Interaction: Guidance Signal

For a model to effectively process user interaction information, it is crucial to represent that interaction in a form the model can interpret. The main challenge lies in converting human input—such as clicks or scribbles—into a structured format suitable for neural networks. This is typically done by encoding the inputs into tensors that share the same spatial dimensions as the input image.

Several encoding strategies have been proposed in the literature. In [91, 92, 25], a disk-based encoding is used, where each click is independently represented by assigning a value of one within a fixed-radius disk centered around the click. This results in binary guidance maps that emphasize the immediate neighborhood of each interaction point.

To create a smoother and more spatially informative representation, a Gaussian encoding, also known as Gaussian Heatmap, is used in [93, 94, 95, 96]. This approach applies a Gaussian filter centered at each click, producing soft-edged blobs whose values decay exponentially with distance from the click location. This leads to a more continuous representation of user intent, which can help the model to generalize better across regions.

An alternative approach, utilized in [97], is the Euclidean distance transform. Here, the guidance channel consists of a tensor where each voxel contains the minimum Euclidean distance to the closest click. This encoding provides a global spatial relationship between the clicks and the rest of the image.

Recognizing the limitations of purely geometric encodings, [98] argues that effective encodings should also incorporate image context. Based on this principle, the geodesic distance transform is introduced and used in several works [99, 45]. This transform computes the shortest geodesic path between each voxel and the nearest click [100], taking into account intensity differences along the path. As a result, the distance reflects not only spatial proximity but also appearance similarity. An extension of this method, the exponential geodesic encoding, is presented in [98], which applies an exponential decay to the geodesic distance values, further emphasizing nearby, contextually similar regions. Figure 2.12 provides a visual comparison of various click-based guidance transforms.

As for scribble-based interactions, they can be interpreted as a dense set of clicks, enabling the use of similar guidance encodings. Gaussian heatmaps, Euclidean distance transforms, and geodesic distance maps can all be adapted for scribbles, as illustrated in Figure 2.13.

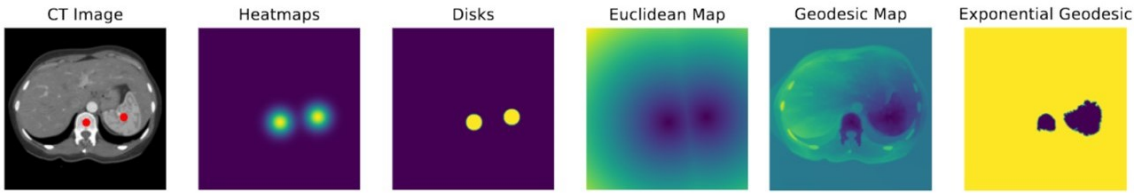


Figure 2.12: Illustration of click-based guidance signals. Image from [82]

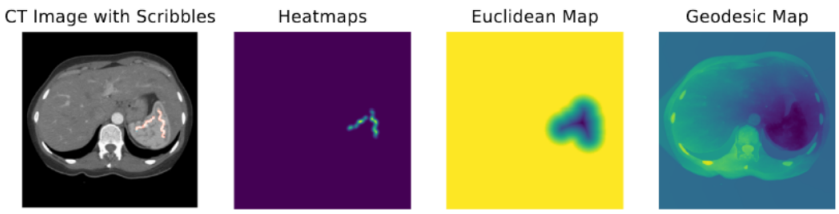


Figure 2.13: Illustration of scribble-based guidance signals. Image from [82]

2.4 Uncertainty-guided Interactive Segmentation

As discussed in the previous section, many recent segmentation frameworks adopt a hybrid approach where an initial automatic result is followed by user interaction to refine errors. While this semi-automatic strategy can improve performance, it still demands significant clinician involvement—particularly in the case of 3D volumetric data, making the process time-consuming and possibly inefficient.

An important step forward in this area is the use of intelligent cues to guide user interaction more effectively. In this context, leveraging uncertainty information from the initial segmentation has shown promise [101]. Most works on uncertainty estimation primarily focus on analyzing how different frameworks affect uncertainty quantification and the subsequent impact on segmentation

accuracy. These studies also explore correlations between uncertainty and mis-segmentation, but they rarely integrate these insights into interactive correction frameworks.

Some efforts have been made to bridge this gap by exploring the potential of uncertainty-aware tools within clinical workflows. For example, uncertainty has been used to flag entire images or specific slices that may require correction and, in some cases, to localize problematic regions. However, these approaches are still in their early stages, and the field remains largely unexplored.

In [65], the authors categorize segmentations as either "good" or "poor" based on a predefined quality criterion, where a segmentation is considered poor if its ASSD exceeds the inter-observer ASSD. To explore uncertainty estimation, they employ MC Dropout, Deep Ensembles, and BBB. A global structure-wise uncertainty metric is introduced, calculated as the variability of ASSD within the sampled predictions. Their results show that prioritizing images for review based on this uncertainty metric significantly reduces the number of poor-quality segmentations. Specifically, to reduce the proportion of poor segmentations in the dataset to 5%, only 31% to 48% of images needed to be reviewed across different datasets, compared to around 80% required with random selection, demonstrating the utility of uncertainty as a triaging tool.

Similarly, [102] investigates the use of uncertainty to train a classifier capable of detecting out-of-distribution images for review. A correlation between uncertainty estimates and Dice score was demonstrated, confirming the reliability of uncertainty as a quality indicator. Furthermore, the method uses localized uncertainty to flag the most uncertain subsegments within each image selected for manual inspection. While this represents a promising strategy for targeted review, the effectiveness of subsegment selection is only evaluated qualitatively through visual examples, with no quantitative assessment or integration into an interactive framework.

A more targeted application is presented in [103], which proposes a combined quality control and error correction pipeline for vessel segmentation in retinal fundus images. Instead of recommending full-image reannotations, the method selects specific candidate patches for correction based on uncertainty estimates. Corrections are simulated by pasting GT labels in these regions. Results show that uncertainty-guided patch selection improves performance over random selection. Moreover, the work also extends a Dice score estimator to predict the expected improvement of the segmentation following a correction. This allows annotators to make informed decisions on how many local patches need reannotation to meet quality standards. Hence, the study shows that, by adaptively determining the number of patches to be corrected per image, a more efficient resource and annotation effort allocation is achieved. While not fully interactive, this framework demonstrates the value of uncertainty in guiding efficient manual review.

Finally, [104] represents a step towards fully integrating uncertainty into an interactive segmentation framework. Their method selects the M most uncertain slices (with M being a fixed parameter) and allows user input in the form of scribbles on these slices. These local annotations are then propagated to yield volumetric corrections. Results indicate that uncertainty-based slice selection improves the refinement efficiency by approximately 30%, enabling high accuracy with fast runtimes. This work exemplifies how uncertainty-driven prioritization can be effectively coupled with interactive refinement to enhance large-scale medical image segmentation.

In summary, while the integration of uncertainty information into interactive segmentation is still emerging, early efforts suggest it can significantly streamline clinician input, reduce annotation burden, and improve segmentation quality. However, more work is needed to develop truly interactive systems that dynamically adapt to uncertainty and provide intuitive guidance for clinical users.

Chapter 3

Methods

This section presents and explains the methodologies used throughout this work, covering the model training process and the key characteristics of the selected architecture. It also details the approach to uncertainty estimation, with a focus on MC dropout to evaluate uncertainty at different levels. Additionally, the interactive segmentation model is discussed, highlighting its functionality and key features.

It is important to note that in the first part of this work, experiments were conducted using a baseline automatic segmentation model. These experiments focused on developing the uncertainty estimation framework, analysing uncertainty behavior, and assessing the potential usability of uncertainty information. In the second part, an interactive segmentation model was trained and additional experiments were performed to investigate how uncertainty could be leveraged to guide real user interactions during the segmentation process.

3.1 Data

In this study, two different modalities were utilized – EAS segmentation in TPUS and pericardium segmentation in cardiac CT.

3.1.1 Cardiac CT Data

For the cardiac CT analysis, both a publicly available dataset and a private dataset were employed. The publicly available dataset was sourced from the Open Source Imaging Consortium (OSIC), which was originally provided as part of a Kaggle challenge, while the private dataset was acquired from the Centro Hospitalar de Vila Nova de Gaia e Espinho (CHVNGE).

Segmenting the pericardium in CT scans is essential for various clinical and research applications, and can be used as a preprocessing step for extraction of different clinical parameters [105]. By accurately delineating the pericardial boundary, different anatomical and physiological measurements can be derived, aiding in disease assessment and risk stratification. One significant application of pericardial segmentation is in the quantification of cardiac fat.

However, manual segmentation is labor-intensive, time-consuming, and subject to observer variability, making it impractical for large-scale studies or routine clinical use [106, 107]. These limitations have motivated the development of automatic pericardial segmentation techniques using AI and machine learning algorithms.

To ensure data consistency, a preprocessing step was performed in which volumes were resampled when necessary to obtain a uniform slice spacing of 3 mm. Importantly, the OSIC dataset was used exclusively in the training phase, while the validation and test sets consisted solely of images from the CHVNGE hospital dataset. This split was intentional, as the models are ultimately intended for deployment on this clinical data. Specifically, 886 images were used for training, and 189 images from the CHVNGE dataset were allocated to each of the validation and test sets, respectively.

3.1.2 TPUS Data

For the TPUS analysis, a private dataset acquired at the UZ Leuven hospital was utilized.

TPUS has emerged as a promising imaging modality in the assessment of EAS injuries [108]. However, ultrasound imaging is inherently susceptible to artifacts, noise, acoustic shadows, and often exhibits low contrast between soft tissues, along with a limited field of view compared to other imaging techniques. A prominent issue is speckle noise, which results from the inhomogeneous nature of biological tissues. These challenges complicate image analysis and demand significant skill and experience to navigate the ultrasound volume and achieve accurate diagnoses. Consequently, automatic segmentation becomes highly relevant, but also particularly challenging to develop and train under these conditions [109].

The dataset consisted of 190 images in total, divided into 132 for training, 29 for validation, and 29 for testing, corresponding to approximately 70%, 15%, and 15% of the data, respectively.

3.2 Baseline Segmentation Model

The nnU-Net was selected for this study due to its demonstrated ability to perform robustly across a wide range of medical image segmentation tasks without the need for manual hyperparameter tuning or architectural adjustments. nnU-Net adapts its structure based on the characteristics of the dataset, ensuring that it performs optimally for different segmentation challenges. This makes it an ideal choice when working with medical images, where dataset characteristics can vary significantly, such as in terms of resolution, image size, and the complexity of the structures being segmented. The architecture of nnU-Net is inherently flexible and capable of adjusting key components like the number of layers, kernel sizes, and strides to accommodate the specific needs of the data at hand. The implementation of nnU-Net was carried out using the DynUNet architecture from MONAI [110].

3.2.1 Model Architecture

The nnU-Net architecture is a self-adapting framework built on the U-Net design, consisting of a typical encoder-decoder structure, which is specified for each application in the following subsections.

Furthermore, in both cases the architecture incorporates skip connections between corresponding downsampling and upsampling layers, enabling the model to retain low-level information from the encoder path, which is crucial for maintaining spatial accuracy in the final segmentation output.

Moreover, the models incorporate deep supervision, which involves the inclusion of intermediate supervision heads at different levels of the network. These heads offer additional loss terms during training, guiding the model to learn both low-level and high-level features. Deep supervision aids in faster convergence and better generalization, as the model is effectively trained using multi-level feature representations.

It is important to note that the architectural characteristics of the models, such as the number of downsampling blocks, convolutional layers, and kernel sizes, were not manually defined but were dynamically adapted to the data by the nnU-Net framework.

CT Segmentation Model

The CT segmentation model begins with an input block composed of two 3D convolutional layers with kernel sizes of 3×3×3 and unit strides (1×1×1), preserving the input dimensions while expanding the feature space. The encoder path then consists of four downsampling blocks, each containing two convolutional layers. In the first two blocks, the initial convolution uses a stride of 2×2×1, effectively reducing spatial resolution in the height and width dimensions while keeping the depth unchanged. The remaining two blocks apply a stride of 2×2×2, performing full 3D downsampling. The second convolution in each block consistently uses a stride of 1×1×1 to preserve spatial dimensions while refining the feature maps and increasing the number of channels. This hierarchical downsampling structure allows the network to gradually extract more abstract and semantically rich features. A bottleneck layer follows the encoder, further compressing the spatial dimensions using a 3D convolution with a 2×2×2 stride.

The decoder mirrors the previous operations with five upsampling blocks, restoring spatial resolution progressively. Each block starts with a transposed convolution to upsample the feature maps, followed by additional convolutions that refine the output and integrate information from the encoder through skip connections. A final upsampling block ensures full restoration of the input resolution. The output layer consists of a 1×1×1 convolution that reduces the feature maps to two channels corresponding to the segmentation classes.

Throughout the network, instance normalization and Leaky Rectified Linear Unit (ReLU) activations are applied, and a dropout rate of 0.3 is used for regularization.

TPUS Segmentation Model

The TPUS segmentation model adopts a similar design philosophy but with some structural simplifications. The input block also contains two 3D convolutional layers with 3×3×3 kernels and unit strides. The encoder includes three downsampling blocks, each consisting of two convolutional layers. The first convolution in each block reduces spatial resolution using a stride of 2×2×2, thereby downsampling the feature maps across all three spatial dimensions, while the second convolution maintains resolution with a stride of 1×1×1 and increases the depth of the feature representation. This is followed by a bottleneck layer that applies a 3D convolution with a 2×2×2 stride.

The decoder path comprises four upsampling blocks that symmetrically mirror the encoder, matching the downsampling steps to restore spatial dimensions accurately. These are followed by standard 3D convolutions to refine features, with skip connections linking blocks to its encoder counterpart to enhance spatial detail. The final layer uses a 1×1×1 convolution to reduce the feature maps to two output channels representing the segmentation classes.

As with the CT model, instance normalization and LeakyReLU activations are used throughout. A lighter dropout rate of 0.1 is applied for regularization, reflecting the structural simplicity of the model. A smaller dropout rate was used in the TPUS segmentation model compared to the CT model due to the substantially smaller size of the ultrasound training dataset and the nature of ultrasound images, which are generally lower in resolution and exhibit higher variability and noise. To prevent underfitting and allow the network to retain more relevant features during training, a lighter regularization strategy was adopted.

3.2.2 Loss Function and Optimization

For this work, a composite loss function combining Dice loss, Cross-Entropy loss, and a calibration parameter was adopted to effectively guide the training of the segmentation network.

The Dice loss directly optimizes the overlap between the predicted segmentation and the ground truth by measuring the similarity between the two. This is beneficial in handling class imbalance, a common challenge in medical datasets, where usually there is a superior proportion of background. Cross-Entropy loss, on the other hand, measures the voxel-wise classification accuracy by penalizing incorrect class predictions, encouraging the model to produce probabilistically accurate outputs. The combination of Dice and Cross-Entropy losses allows the model to benefit from both global region-level overlap optimization and local voxel-wise accuracy.

To improve the model's calibration, the approach proposed in [111] was followed, which introduces an additional loss term – ACE (Average Calibration Error) – to the standard loss function.

The ACE term is designed to minimize the gap between the model's predicted probabilities and the true frequencies of class occurrences. This helps align the predicted probability distribution with the true data distribution, ensuring that the model not only achieves accurate segmentation but also provides reliable probability estimates, which are critical in applications requiring uncertainty quantification.

To compute the ACE parameter, it is necessary to first discretize the continuous predicted probabilities of each of the C classes into M bins. 20 bins were used, following the approach used

in the original study. For each image, and for each class c, the observed frequency o_m^c in the mth bin is compared to the expected probability e_m^c within the same bin. An ideally calibrated model is one where the gap $|o_m^c - e_m^c|$ is zero for all bins and all classes, meaning the predicted probabilities match the true observed frequencies perfectly. The ACE loss is computed as:

$$L_{ACE} = \frac{1}{CM} \sum_{C=1}^{C} \sum_{m=1}^{M} |o_m^c - e_m^c| \tag{3.1}$$

To integrate this calibration into the model, the ACE term was added to the existing loss function, which already included the Cross-Entropy and Dice components. By combining these loss terms, the model is encouraged to minimize the calibration error while maintaining high segmentation accuracy.

This integrated loss function can then be expressed as a weighted sum of the original loss terms and the ACE loss, as presented in 3.2, where $\hat{y}$ is the predicted output, y is the ground truth segmentation, and each term corresponds to a different aspect of prediction quality.

$$L_{Dice,CE,ACE}(\hat{y}, y) = \frac{1}{3} L_{Dice}(\hat{y}, y) + \frac{1}{3} L_{CE}(\hat{y}, y) + \frac{1}{3} L_{ACE}(\hat{y}, y) \tag{3.2}$$

To further enhance learning, deep supervision was applied. This way, predictions from intermediate decoder layers are also supervised, not just the final output. This approach ensures that the model is not solely dependent on the final output for learning, but instead is guided at multiple levels throughout its depth. By exposing the network to supervisory signals from both early and late stages of the encoder-decoder path, deep supervision promotes the learning of meaningful hierarchical features, encouraging earlier layers to contribute more effectively to the segmentation task, rather than relying entirely on the deepest representations. Furthermore, the additional gradient signals from these intermediate outputs help alleviate the vanishing gradient problem, facilitating the training of deeper architectures.

These intermediate losses are then combined in a weighted fashion to form the final total loss, presented in 3.3, where N is the number of supervised outputs, $\hat{y}_i$ is the prediction from the i^{th} decoder layer, and w_i is a weight assigned to each supervision level, decreasing exponentially with depth. This strategy ensures stronger supervision at shallower layers, accelerating convergence and improving gradient flow throughout the network. In this work, two levels of supervision were used, based on coding practices for similar segmentation tasks [112].

$$L_{total} = \sum_{i=0}^{N-1} w_i \cdot L_{Dice,CE,ACE}(\hat{y}_i, y), \quad \text{where} \quad w_i = 0.5^i \tag{3.3}$$

The Adam optimizer was used for training, which adaptively adjusts learning rates for each parameter based on first and second-order moment estimates of gradients, promoting stable and efficient convergence.

3.2.3 Model Selection Strategy

For each of the two datasets, the final model used for evaluation was selected based on the highest Dice score achieved on the respective validation set. Dice Score calculation is presented in Equation 3.4, where A represents the set of predicted positive voxels and B the set of ground truth positive voxels.

$$Dice = \frac{2|A \cap B|}{|A| + |B|} \tag{3.4}$$

To prevent overfitting and reduce unnecessary training time, an early stopping criterion of 100 epochs was employed. This means that training was halted if no improvement in the validation Dice score was observed for 100 consecutive epochs.

3.3 Interactive Segmentation Model

This study employs the Boundary Interactive Segmentation Network (BISNet), an interactive 3D segmentation framework designed to segment anatomical structures in volumetric data [112]. BISNet supports both an initial fully automatic segmentation and subsequent interactive refinement through user-provided boundary clicks. Note that the BISNet was originally developed for the segmentation of the EAS in TPUS data. In this study, the same hyperparameters as reported in the original BISNet work are used.

These boundary clicks are interpreted as 3D spatial points within the image volume that indicate where the boundary of the segmented structure should lie, reflecting the user's intent to adjust the segmentation in a specific location. As such, the network modifies the initial segmentation in response to each interaction until the desired quality is reached.

The model operates in two modes. In its automatic mode, BISNet generates an initial segmentation solely based on the input volume. In its interactive mode, the segmentation is refined based on additional input received via the guidance channel, which encodes the user's boundary clicks. These interactions can be performed either all at once, by providing multiple clicks at once, or iteratively. A schematic of the model architecture is shown in Figure 3.1.

The model distinguishes between two types of user corrections:

- **Outward clicks**, which are placed outside the current segmented region to indicate that the boundary should expand and pass through the clicked point, correcting an under-segmentation.
- **Inward clicks**, which are placed inside the segmented region to signal that the boundary should contract inward through the click, correcting an over-segmentation.

Through this iterative process, BISNet integrates clinician feedback directly into the segmentation pipeline, enabling high-precision, user-guided corrections in 3D space.

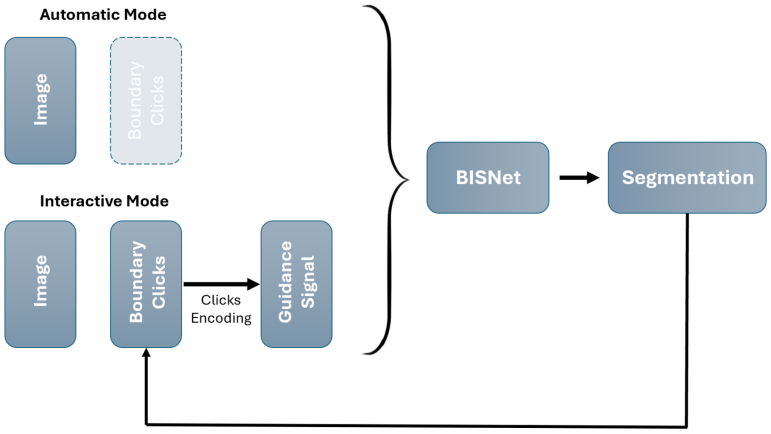


Figure 3.1: Internal outline of BISNet. In automatic segmentation mode, the guidance channel is empty. In interactive segmentation mode, the guidance channel contains boundary clicks based on the previous segmentation.

3.3.1 Model Architecture

The architecture used for interactive segmentation models remains consistent with that of the baseline models described in Section 3.2.1. Specifically, all models were configured using the nnU-Net framework, which automatically adapts the U-Net architecture based on the characteristics of the input data. Since the same datasets were used for both the baseline and interactive models within each modality, the nnU-Net generated identical architectures for both use cases.

3.3.2 Boundary click encoding

To enable BISNet to process user interactions, a second input channel is introduced to encode the provided boundary clicks. These clicks, which are initially given as explicit 3D coordinates in the image space, must be converted into a volumetric representation interpretable by the network. For this purpose, an inverse Euclidean distance encoding was employed.

This encoding transforms a set of boundary points $P = \{(x_i, y_i, z_i)\}_i$ into a 3D tensor $T[u, v, w]$, where each voxel value is inversely proportional to the Euclidean distance to the closest boundary point. Formally, the encoding is defined as:

$$T[u,v,w] = \left(1.1 + \alpha \min_{(x,y,z)\in P} D_E((x,y,z),(u,v,w))\right)^{-1} \tag{3.5}$$

Here, $D_E(\cdot)$ denotes the standard Euclidean distance, and α is a scaling factor that controls how sharply the values decay with distance. A higher value of α results in more localized influence around each boundary click. The coordinates (u, v, w) and (x, y, z) represent points within the 3D volume along the sagittal, coronal, and transverse axes. In accordance with the original BISNet proposal, the scaling factor was set to $\alpha = 1$, providing a linear balanced decay rate of the signal.

Note that the constant 1.1 in the denominator prevents division by zero and ensures numerical stability by limiting the maximum encoded value.

In the case of multiple clicks, only the distance to the nearest boundary point is considered for each voxel, as enforced by the minimum operator. This encoding ensures that voxels closer to user-specified points receive stronger guidance signals, steering the network to adjust the segmentation boundary toward the intended locations.

3.3.3 Training Process

Since boundary clicks are an implicit form of interaction, the network must learn to associate these with the intended refinement. To enable this, the model requires repeated exposure to boundary clicks during training. However, it is infeasible to collect such interactions manually. Consequently, boundary clicks were generated through a simulation procedure that mimics the behavior of a clinical expert.

Synthetic clicks are generated based on the ground truth and the current prediction of the network. Each voxel in the image space is assigned a probability of being selected as a boundary click. Voxels near incorrectly segmented boundaries, either over- or under-segmented, are more likely to be chosen, under the assumption that real users would focus their corrections near visible segmentation errors.

Formally, a 3D probability density function $D(x,y,z)$ is defined over the image space:

$$D(x,y,z) = G_\xi * \left(e^{\lambda E_p(x,y,z)} - 1\right) \tag{3.6}$$

Here, $E_p(x,y,z)$ represents the prediction error, calculated as the Euclidean distance from each boundary voxel in the ground truth to its nearest boundary voxel in the predicted segmentation. A boundary voxel is defined as any non-background voxel that neighbors at least one background voxel. Importantly, $E_p(x,y,z)$ is non-zero only at incorrectly segmented boundary regions, effectively ignoring already correct areas.

To give more weight to highly incorrect areas, the error map $E_p(x,y,z)$ is exponentiated using base e. To ensure that perfectly segmented points ($E_p = 0$) are never sampled, one is subtracted from the result, since $e^{\lambda \cdot 0} - 1 = 1 - 1 = 0$. This ensures that only regions with segmentation errors contribute to the click sampling distribution, which closely resembles real-world correction behavior.

The parameter λ determines the influence of the prediction error on the click sampling distribution. Higher values of λ result in stronger focus on high-error regions, while lower values lead to more uniform sampling across the boundary. For all experiments, a value of $\lambda = 0.1$ was used, providing a balance between focusing on errors and maintaining coverage across the segmentation boundary.

The Gaussian smoothing kernel G_ξ, with standard deviation ξ, is convolved with the error-based weighting to simulate inter- and intra-observer variability. For the experiments in this study,

$\xi = 1.0$ was used to introduce a mild variability while keeping the clicks close to the actual boundary.

To train BISNet to handle varying numbers of user interactions, a random number of clicks, between n_{min} and n_{max}, are generated for each training volume. A minimum of $n_{min} = 1$ ensures that each interactive example contains at least one click. To maintain the network's ability to perform automatic segmentation, a subset of training samples is processed without any interaction. This is controlled by a probability parameter p_g, which determines the likelihood of simulating guidance for a given training sample.

3.3.3.1 Dropout Mechanism

To ensure that the network performs well in both automatic and interactive segmentation modes, it must be trained under both conditions. However, there exists a trade-off, as excessive exposure to interaction during training may cause the model to become overly reliant on guidance, potentially degrading the quality of its initial automatic segmentations. Conversely, if the model is not sufficiently exposed to interaction, it may fail to learn the desired response to boundary clicks.

To address this, a dropout mechanism was implemented for the guidance signal. During each training iteration t, there is a probability $p(t)$ that the simulated boundary clicks are dropped. When this occurs, the guidance channel is set to zero, and the network operates in automatic segmentation mode. This probabilistic dropout provides finer control over the exposure of the network to interaction.

In this work, an Exponential Decay Dropout scheme was used, where the probability of dropping the guidance signal evolves over time according to the formula:

$$p(t) = \min\left(1, b^{a(t-c)}\right) \tag{3.7}$$

Here, the parameters a and b determine the rate of decay. Specifically, larger values of a or smaller values of b result in a faster reduction of the dropout probability. The parameter c defines a cold-start period during which all guidance is dropped (i.e., $p(t) = 1$ for $t < c$). This allows the network to initially learn a coherent automatic segmentation without the influence of guidance signals. This training dynamic encourages a balance between high-quality automatic performance and responsiveness to user guidance.

3.3.3.2 Loss Function

The primary loss used to train BISNet was the composite loss $L_{Dice,CE,ACE}$, combining Dice loss, cross-entropy loss, and the ACE loss for calibration, applied with deep supervision across decoder layers. This encourages accurate segmentation and stable training by supervising multiple output levels in the network.

However, this formulation alone does not explicitly encourage the model to respond to user-provided boundary clicks. While it can implicitly learn this behavior, it is not directly penalized

for ignoring the clicks. To address this, an additional loss term was introduced to focus learning on regions around the boundary interactions.

To guide the model's response toward the regions near the boundary clicks, a spatial weighting mechanism is introduced in this additional loss term. For every boundary click, a 3D Gaussian distribution is centered at the click location. These Gaussians are summed together to form a smooth weight map over the image volume. As a result, each voxel is assigned a weight depending on its distance to the nearest click, so voxels closer to clicks receive higher weights, and those further away receive lower or zero weight. This weight map is then used to compute a localized version of the Dice and cross-entropy losses, placing more importance on accurately segmenting areas influenced by the boundary interactions.

Formally, the Weighted Dice and Cross-Entropy Loss is defined as:

$$L_{WeightedDice,CE}(\hat{y}, y, P) = \sum_{v \in \Omega} T_\sigma[v] \left(\frac{\hat{y} \cdot y}{\hat{y} + y} + y \cdot \log \hat{y} \right) \tag{3.8}$$

where $\hat{y}$ is the predicted segmentation, y is the ground truth, P is the set of boundary clicks, T_σ is the Gaussian encoding of the guidance points P and Ω is the set of all voxels in the volume.

To ensure the model also retains global segmentation quality, the guidance-weighted loss is combined with the base loss function. The total loss used during training is therefore:

$$L_{combined} = K_1 \cdot L_{Dice,CE,ACE}(\hat{y}, y) + K_2 \cdot L_{WeightedDice,CE}(\hat{y}, y, P) \tag{3.9}$$

where K_1 and K_2 are the weighting factors that balance the influence of the guidance-aware term relative to the standard loss. It was used a weight of one for both components.

This formulation ensures that BISNet learns to produce accurate global segmentations while being explicitly encouraged to respond correctly to user-provided guidance during interaction.

Then deep supervision was applied in the same manner as described for the baseline model in Section 3.2.2. Specifically, loss terms were computed not only at the final output layer but also at multiple intermediate decoder stages of the network. These intermediate losses were then aggregated in a weighted fashion to form the total loss. This strategy helps guide the learning process at different levels of feature abstraction and improve gradient flow during training.

3.3.3.3 Model Selection Strategy

The final model was selected based on its performance on the validation set, using a custom evaluation metric that balances segmentation quality with responsiveness to interactive input.

To determine the most desirable training checkpoint, the following metric was used:

$$\Delta GD(S_{init}, S, P) \cdot Dice(S_{init}, L) \tag{3.10}$$

where S_{init} denotes the initial automatic segmentation produced by the model before any user interaction, S represents the segmentation result after incorporating interactive user guidance, P is

the set of guidance points provided to correct or refine the segmentation, and L denotes the ground truth segmentation labels.

This metric combines two key components:

- Dice coefficient of the initial (automatic) segmentation S_{init} against the ground truth labels L, assessing the baseline quality of the model's predictions.
- Fractional decrease in guidance distance $\Delta GD(S_{init}, S, P)$, which quantifies how effectively the model responds to interactive boundary corrections. This responsiveness is crucial for interactive use.

Guidance distance measures the average distance from a set of guidance points P to the nearest point on the segmentation boundary S, and is formally defined as:

$$GD(S,P) = \frac{1}{n} \sum_{p \in P} \min_{s \in S} d(p,s) \tag{3.11}$$

where $d(p,s)$ is the spatial distance between a guidance point p and the closest point s on the segmentation boundary.

The fractional improvement from interaction is then computed as:

$$\Delta GD(S_{init}, S, P) = \frac{GD(S_{init}, P) - GD(S, P)}{GD(S_{init}, P)} \tag{3.12}$$

This normalization ensures that models with higher initial segmentation quality are not unfairly penalized. For example, models starting with accurate predictions will naturally have smaller absolute guidance distances, and this fractional form allows for a fair comparison of responsiveness across different segmentation qualities.

Multiplying ΔGD with the initial Dice score emphasizes a balanced objective - a model that performs well both automatically and interactively. Models with poor automatic segmentations or weak responses to guidance will score lower, while the best models will show strong results in both components.

Similarly to the baseline models, an early stopping criterion of 100 epochs was employed.

3.4 Evaluation Metrics

The final segmentation performance was quantitatively assessed using multiple complementary metrics to evaluate both the baseline model and the interactive refinement process.

The Dice Score (Equation 3.4) and the Average Symmetric Surface Distance (ASSD) were used to measure the accuracy of the automatic segmentations. While the Dice Score captures volumetric overlap between the predicted segmentation and ground truth, the ASSD measures the average distance between the boundaries of the two masks, providing a more surface-sensitive evaluation. The ASSD is formally defined as:

$$ASSD(A,B) = \frac{1}{|S_A| + |S_B|} \left(\sum_{a \in S_A} \min_{b \in S_B} \|a - b\| + \sum_{b \in S_B} \min_{a \in S_A} \|b - a\| \right) \tag{3.13}$$

Here, S_A and S_B denote the sets of surface points for the predicted and ground truth segmentations, respectively, and $\|\cdot\|$ represents the Euclidean distance. The notation $|S_A|$ and $|S_B|$ represent the number of points in each surface set. This metric computes the average of the shortest distances from every boundary point in one mask to the closest boundary point in the other mask, ensuring a symmetric, bidirectional comparison of boundary proximity.

To assess the model's responsiveness to user input in the interactive segmentation setting, the fractional decrease in guidance distance, denoted $\Delta GD(S_{init}, S, P)$ (Equation 3.12), was measured. Together, these metrics provide a comprehensive evaluation of both the baseline segmentation quality and the effectiveness of interactive refinements.

3.5 Uncertainty Quantification

In this section, the techniques and metrics used to estimate uncertainty are described. MC Dropout was employed as a practical method for uncertainty estimation, leveraging its Bayesian interpretation to approximate posterior distributions. By applying dropout at inference, MC Dropout allows the quantification of uncertainty at both voxel and structure levels. The following subsections detail the implementation of MC Dropout and its role in uncertainty estimation.

3.5.1 MC Dropout Implementation

The method employed for uncertainty estimation in this study is based on MC Dropout, described in Section 2.2.1.3. MC Dropout is one of the most widely used and established techniques for capturing uncertainty in neural networks, chosen here for its simplicity and ease of integration since it requires no modifications to the training procedure.

During inference, MC dropout generates multiple predictions for the same input by applying different dropout masks across multiple forward passes. This approach allows the model to produce a distribution of predictions for one input image, rather than a single deterministic output.

To optimize performance, the dropout rate and the number of forward passes were carefully tuned and details of this tuning process are provided in a later section.

The final segmentation output was obtained by aggregating the predictions from all sampled forward passes, assigning to each voxel the label that appears most frequently across these predictions. Therefore, this approach serves a dual purpose - it enables effective uncertainty quantification by measuring variability across stochastic predictions, and it defines the segmentation output through the consolidation of multiple stochastic predictions into a single result.

3.5.2 Uncertainty Metrics

Using the previously described MC Dropout method, multiple segmentation predictions were generated for each input image. These predictions form a distribution that serves as the foundation for estimating uncertainty. The variability among these predictions quantifies the degree of uncertainty in the segmentation process. Uncertainty was computed at both the voxel level and the structure level to provide a comprehensive assessment.

3.5.2.1 Voxel-wise Uncertainty

In segmentation tasks, assessing uncertainty at the voxel level is particularly important. Given an input image X, multiple forward passes generate, for each voxel i, a set of predicted probabilities for a target class (e.g., foreground class), forming a distribution P^i, exemplified in Equation 3.14, where p_n^i denotes the predicted probability of the foreground class at voxel i from the nth pass, and N is the total number of stochastic forward passes.

$$P^i = \{p_1^i, p_2^i, \ldots, p_N^i\}, \quad p_n^i \in [0,1] \tag{3.14}$$

This set represents the variability in the model's probabilistic output at that voxel across multiple passes. Therefore, the mean predicted probability $\bar{p}^i$ is computed as:

$$\bar{p}^i = \frac{1}{N}\sum_{n=1}^{N} p_n^i \tag{3.15}$$

Voxel-wise uncertainty is then quantified using binary entropy, defined as:

$$H(\bar{p}^i) = -\bar{p}^i \log(\bar{p}^i + \varepsilon) - (1-\bar{p}^i)\log(1-\bar{p}^i + \varepsilon) \tag{3.16}$$

where ε is a small constant added for numerical stability. This measure reflects the model's epistemic uncertainty at each voxel based on the variability in softmax outputs across multiple stochastic passes, since high entropy corresponds to greater disagreement among the predictions for that voxel.

Uncertainty Map Post-Processing

The voxel-wise entropy computation described above results in an entropy map, where each voxel in the volume is assigned an uncertainty value. However, these raw entropy maps often exhibit noise across the volume, with elevated uncertainty frequently observed not only in regions of genuine ambiguity but also in background areas or isolated voxels.

To derive a more informative and robust uncertainty map, two filtering steps were applied.

First, to mitigate the influence of isolated voxels exhibiting spurious high uncertainty and to emphasize coherent regions of uncertainty, a $3\times3\times3$ averaging filter was applied to the raw entropy maps. This local smoothing suppresses random fluctuations while preserving broader

regions of consistently elevated uncertainty, which are more likely to reflect true segmentation ambiguities.

Despite this initial filtering, the boundary regions still exhibited considerable variability, as not all boundary voxels with high uncertainty correspond to actual segmentation errors. To further refine the uncertainty signal, a directional smoothing strategy based on radial profiling was implemented. The center of mass (CoM) of the segmented structure was first computed. Then, for each boundary voxel, six neighboring voxels were sampled along the radial direction defined by the vector connecting the CoM to the boundary point: three in the outward direction and three in the inward direction. The uncertainty values from these six neighbors, along with the boundary voxel itself, were averaged and assigned to the corresponding boundary voxel. This approach effectively smooths the uncertainty along directions orthogonal to the surface, better capturing over- and under-segmentation patterns that typically manifest as locally thickened regions of uncertainty.

The resulting map, produced through these two filtering stages, provided a refined boundary-focused uncertainty representation, better suited for downstream analysis and interaction guidance.

3.5.2.2 Structure-wise Uncertainty

At the structure level, the objective was to compute a single scalar value that reflects the overall uncertainty associated with each segmented structure.

For each segmentation prediction, the corresponding voxel-wise entropy map was generated based on the voxel-wise uncertainty estimation. However, as is common in uncertainty estimation for segmentation tasks, these voxel-wise entropy maps often contain a considerable amount of noise, particularly in background regions or distant from the object boundaries. In order to mitigate the influence of noise from irrelevant background regions while still preserving meaningful uncertainty information along structure boundaries, the entropy values were summed within the segmented structure and within an additional margin of three voxels surrounding its boundary. This approach allows for the inclusion of boundary-related uncertainty, which often reflects genuine segmentation ambiguity, while reducing the contribution of spurious background uncertainty.

Following this aggregation, three variants of the summed entropy were considered to account for differences in the size of the segmented structures:

- **Raw Summed Entropy:** the total sum of entropy values within the segmented structure and its surrounding margin.
- **Volume-Normalized Summed Entropy:** the raw summed entropy divided by the volume of the segmented structure.
- **Surface-Normalized Summed Entropy:** the raw summed entropy divided by the surface area of the segmented structure.

3.6 Experiments

A series of experiments were conducted to evaluate the applicability and effectiveness of the proposed uncertainty estimation methods in the context of medical image segmentation. Both non-interactive and interactive scenarios were investigated, with experiments designed to assess how uncertainty estimates can inform quality control, guide user interventions, and improve segmentation outcomes.

3.6.1 Experiments on Baseline Model

The first set of experiments aimed to assess the reliability and potential of uncertainty estimates to indicate missegmentations and guide correction processes, using the baseline model as the starting point. At this stage, no interactive tool was incorporated, and all corrections were simulated.

Initially, structure-wise uncertainty was employed to prioritize images for review. Images were ranked based on their global uncertainty estimates, and experiments evaluated both the improvement in Dice score as increasingly uncertain images were corrected, and the percentage of poor-quality segmentations remaining after correction. Corrections were simulated under two scenarios: (i) a perfect correction scenario, where corrected segmentations were assumed to achieve a Dice score of 1.0, and (ii) a scenario incorporating intra- and inter-observer variability, where corrections were simulated using manual annotations from multiple clinicians on a smaller subset of the dataset.

In addition, voxel-wise uncertainty analysis was performed through patch correction experiments. Here, voxel-wise uncertainty maps were utilized to identify localized regions for correction. Corrections were simulated by replacing the selected uncertain regions with the corresponding ground truth, and the resulting improvements in Dice score were quantified.

Finally, combined experiments were conducted, where structure-wise uncertainty guided the selection of images for review, followed by voxel-wise uncertainty to determine localized corrections within the selected images. Throughout all these experiments, interaction processes were fully simulated in the absence of an integrated interactive correction tool.

3.6.1.1 Experiments on Interactive Model

Subsequent experiments were conducted using the interactive model, which incorporates boundary clicks provided by the user as additional input for segmentation refinement. A clinical experiment was performed using the TPUS dataset, in which a clinician employed the trained BISNet model to iteratively refine the segmentations through boundary-based interactions.

In this context, structure-wise uncertainty was further explored as a mechanism to guide and prioritize images for correction, assisting in the identification of cases likely to benefit from user interaction. However, voxel-wise uncertainty was not utilized to directly guide user interactions during this stage. Nonetheless, the relationship between the corrected regions and the corresponding pre-correction uncertainty estimates was investigated, in order to assess the extent to which the

uncertainty maps were able to predict the locations of segmentation errors subsequently addressed through interaction.

Chapter 4

Baseline Model Results

In this chapter, the results of the baseline segmentation models are presented. These models, trained without any form of user interaction, served as a reference point for evaluating the effectiveness of subsequent interactive approaches. The model is assessed in terms of both segmentation performance and calibration quality, providing a comprehensive understanding of how well the model performs and how reliable its predictions are.

The evaluation is conducted in two distinct settings - Cardiac CT and TPUS. For each application, a separate model was trained and optimized to account for differences in data modality, anatomy, and imaging characteristics. Therefore, results for both cases are examined.

4.1 Segmentation Performance

The obtained Dice and ASSD scores for each application are summarized in Table 4.1, offering a baseline against which the impact of uncertainty estimation and interactive refinement will be measured in later sections. The table also includes reference inter-observer and intra-observer Dice scores and ASSD derived from manual segmentation experiments. For the CT data, these reference values were obtained from two clinicians segmenting the same 30 images, with one of them segmenting the images twice to assess intra-observer variability. For the TPUS data, the values were computed from a similar setup involving two clinicians and 5 volumes, although the small sample size may limit the robustness of this estimate. However, note that for the TPUS data, the ASSD values were not provided.

Table 4.1: Segmentation results and observer agreement metrics for the CT and TPUS models

Metric	**CT**	**TPUS**
Dice Score (%)	93.54 ± 3.41	72.85 ± 13.86
ASSD (mm)	2.81 ± 3.23	2.28 ± 1.05
Inter-observer Dice Score(%)	92.45 ± 1.98	68.35 ± 3.54
Inter-observer ASSD (mm)	2.54 ± 0.91	-
Intra-observer Dice (%)	95.35 ± 1.31	78.75 ± 7.16
Intra-observer ASSD (mm)	1.39 ± 0.55	-

The segmentation models trained for CT and TPUS data each show performance consistent with what is expected for their respective imaging modalities.

The CT results indicate reliable segmentation performance, with high overlap with ground truth. This aligns with the high anatomical contrast, spatial resolution, and relatively low inter-observer variability typically associated with CT imaging. The Dice score of 93.54% is consistent with values reported in the state-of-the-art literature for similar segmentation tasks [34], and also also exceeds the inter-observer reference score, indicating strong concordance with expert annotations. In contrast, the ASSD is slightly higher than the inter-observer benchmark, suggesting modest deviations in boundary placement despite the overall high segmentation accuracy.

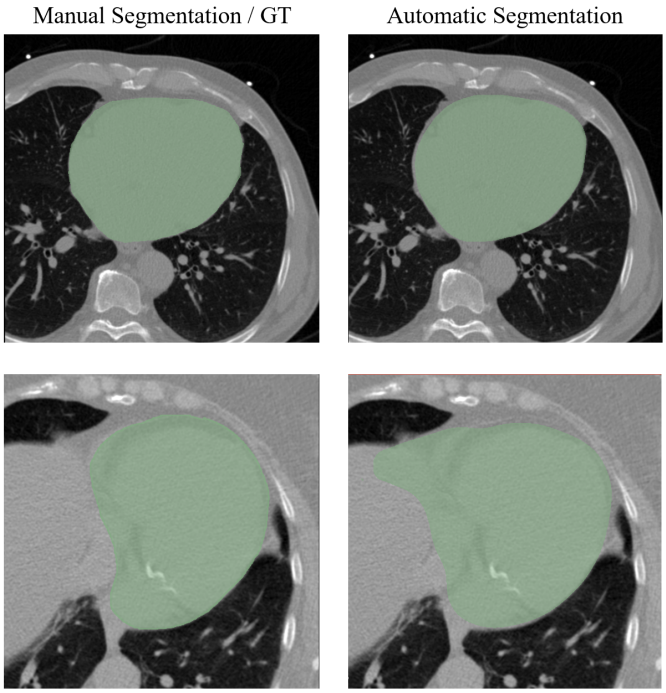


Figure 4.1: Examples of pericardium segmentation in CT slices. In each row, the manual segmentation (GT) and the model's automatic segmentation are shown, with the segmentation highlighted in green. The first row illustrates a case with above-average performance (Dice = 95.23%, ASSD = 1.47 mm), while the second row shows a below-average result (Dice = 77.84%, ASSD = 8.24 mm).

In contrast, the TPUS model results reflect greater segmentation variability and less consistent region overlap, as shown by a lower mean Dice score and a larger standard deviation. These outcome are in line with the known challenges of TPUS imaging, such as acoustic shadowing, speckle noise, and variable probe positioning, which often obscure anatomical boundaries and introduce ambiguity [109]. Even for expert annotators, segmentation of TPUS images tends to show higher inter-observer variability. Notably, the model's Dice score surpasses the inter-observer reference value. Although the ASSD is lower for TPUS than for CT, this likely reflects the smaller size of the EAS compared to the pericardium. The small size of the EAS makes even minor boundary deviations more impactful in terms of Dice, yet simultaneously amplifies the significance of small

absolute distances captured by ASSD. Moreover, the training dataset for TPUS was significantly smaller (132 images) compared to the CT dataset (886 images), likely limiting the ability for the model to generalize and hence contributes to the increased standard deviation observed in Dice scores.

Figures 4.1 and Figure 4.2 each show two examples of the segmentation of the pericardium and EAS, respectively, compared to the manual segmentation serving as the GT.

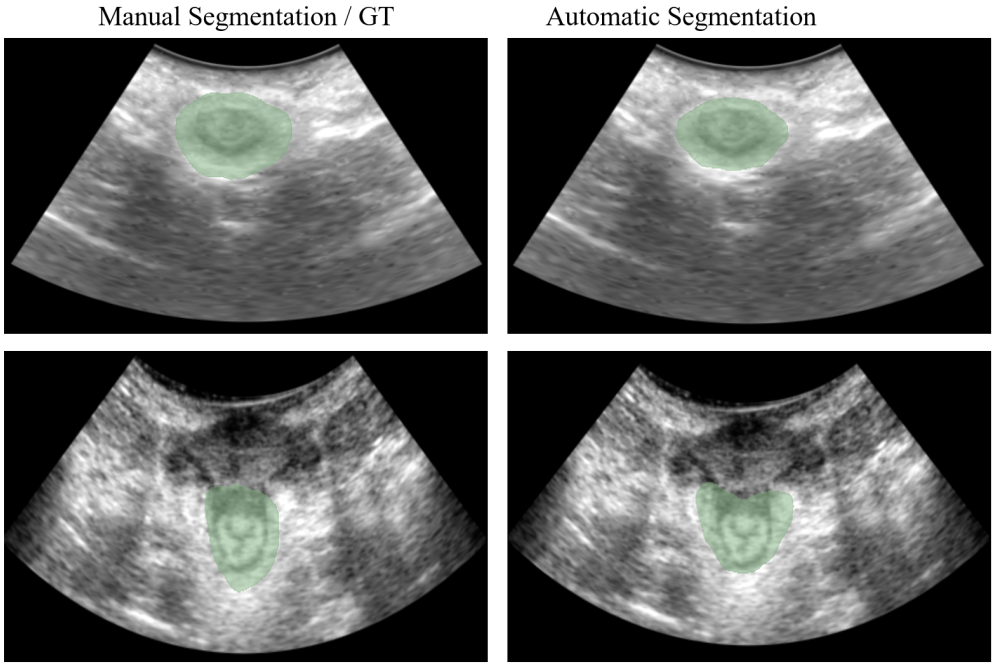


Figure 4.2: Examples of pericardium segmentation in TPUS slices. In each row, the manual segmentation (GT) and the model's automatic segmentation are shown, with the segmentation highlighted in green. The first row illustrates a case with above-average performance (Dice = 81.03%, ASSD = 1.65 mm), while the second row shows a below-average result (Dice = 64.22%, ASSD = 2.43 mm).

4.2 Model Calibration

In this section, the calibration quality of the baseline segmentation models is evaluated by comparing models trained with and without the calibration factor incorporated into the loss function. Calibration curves and the average ACE metric are used to assess how well predicted probabilities align with the true likelihood of correctness. A well-calibrated model will produce curves that closely follow the identity line, indicating that the predicted confidence levels match the actual likelihood of correct predictions.

Figures 4.3 and 4.4 present the calibration histograms for the cardiac CT and TPUS ultrasound datasets, respectively. For each dataset, separate plots are shown for the training and test sets, with and without the application of the calibration parameter during training. Rather than traditional

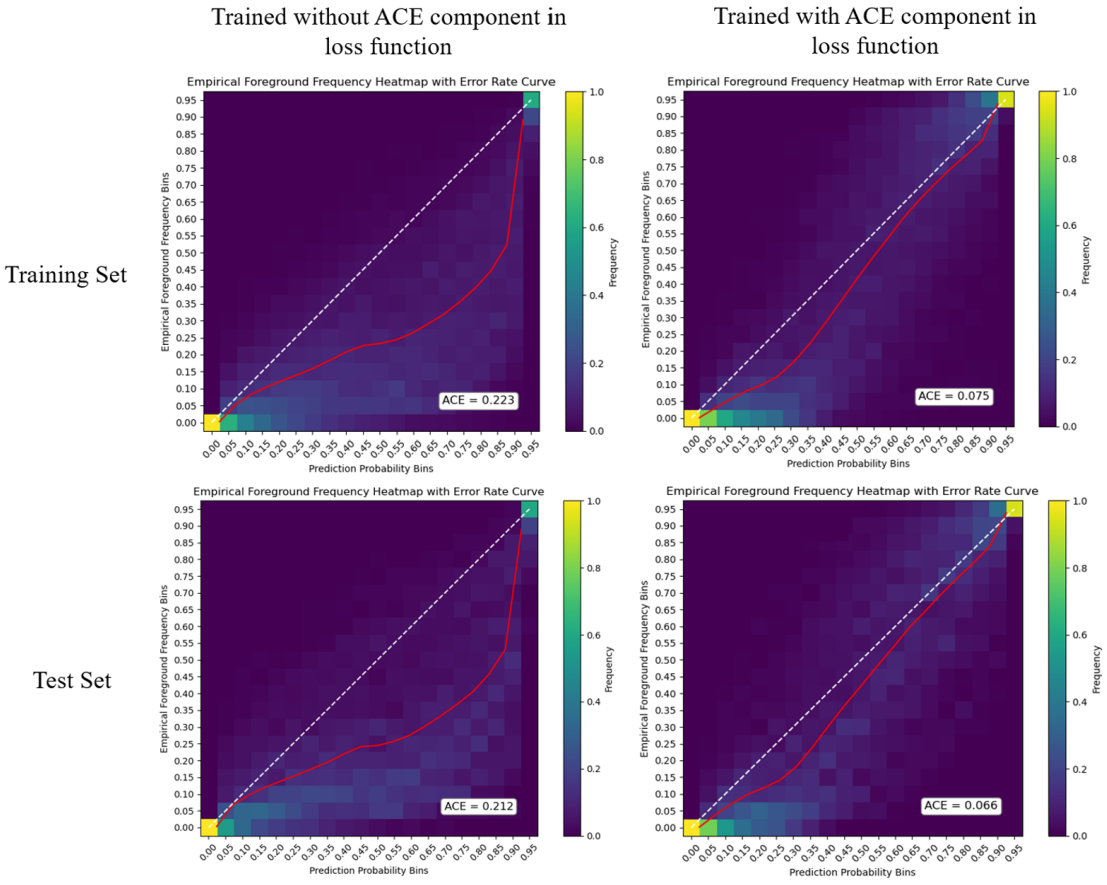


Figure 4.3: Calibration heatmaps and reliability curves for CT models trained without (left) and with (right) a calibration factor, evaluated on the training set (top) and test set (bottom). In each plot, the x-axis represents the predicted probability that a voxel belongs to the foreground class, and the y-axis shows the empirical frequency with which voxels at that probability level actually belong to the foreground. The red curve represents the average fraction of correct predictions made at each confidence level across the whole dataset. The white dashed line indicates perfect calibration, where predicted probabilities match the observed accuracy. The background heatmap reflects the density of predictions in each bin. The average ACE is reported in the bottom-right corner of each plot.

calibration curves, these figures display calibration histograms, providing an informative visualization of calibration across an entire dataset. The training set curves provide insight into whether the model is able to learn a calibrated mapping during optimization, while the test set curves reveal how well this calibration generalizes to unseen data.

In Table 4.2, the mean Dice score and ASSD values are reported for models trained with and without the calibration component, highlighting the impact of calibration on segmentation accuracy.

When trained without the calibration factor, the CT segmentation model exhibits poor calibration. This is evident from the reliability histograms, which show that the model does not produce well-calibrated confidence estimates. The behavior is consistent across both the training and test datasets, suggesting that the lack of calibration is a general issue of this model. However, when

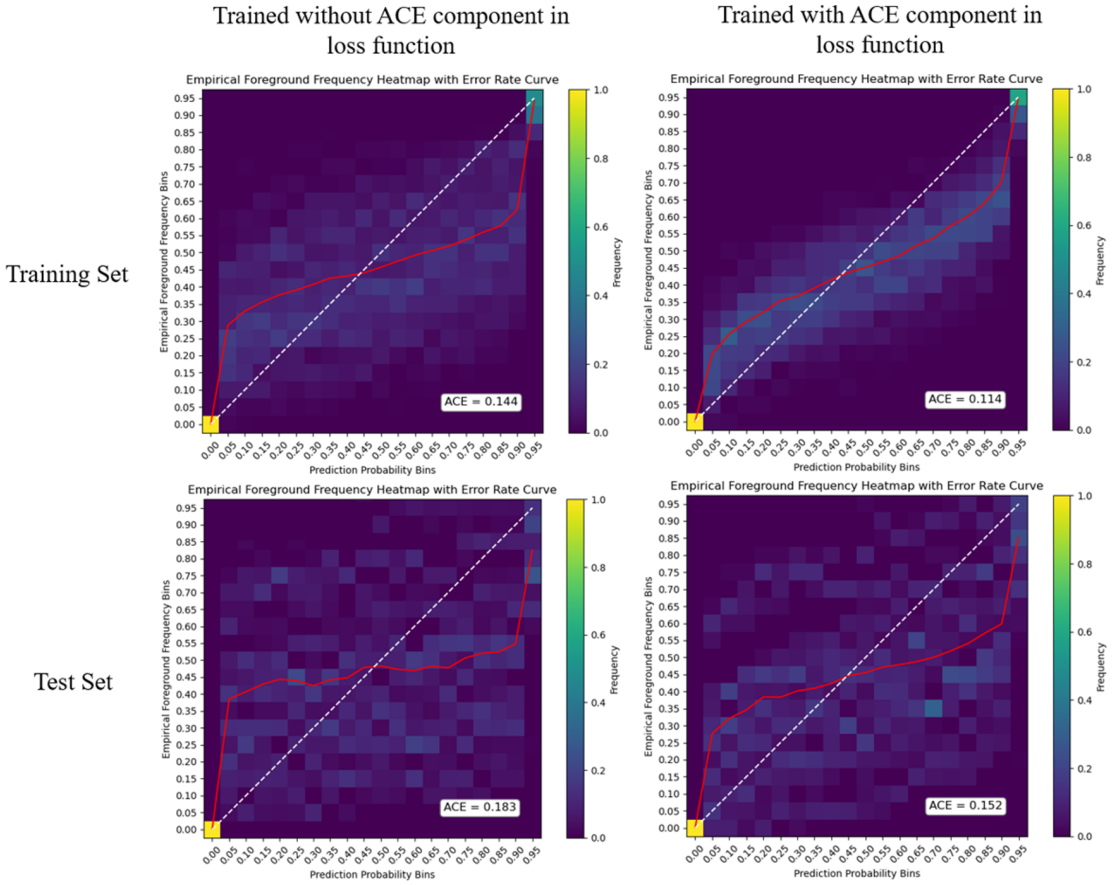


Figure 4.4: Calibration heatmaps and reliability curves for TPUS models trained without (left) and with (right) a calibration factor, evaluated on the training set (top) and test set (bottom). In each plot, the x-axis represents the predicted probability that a voxel belongs to the foreground class, and the y-axis shows the empirical frequency with which voxels at that probability level actually belong to the foreground. The red curve represents the average fraction of correct predictions made at each confidence level across the whole dataset. The white dashed line indicates perfect calibration, where predicted probabilities match the observed accuracy. The background heatmap reflects the density of predictions in each bin. The average ACE is reported in the bottom-right corner of each plot.

calibration is incorporated during training, there is a clear improvement. The histograms of predicted probabilities become much cleaner and more structured, indicating a better distribution of prediction confidence. As result, the average ACE becomes significantly lower, and the calibration curve aligns more closely with the ideal diagonal line. Importantly, the learned calibration generalizes well to the test set, where a similarly clean histogram and low ACE are observed. In terms of segmentation performance, the CT model clearly benefits from calibration, as the mean Dice score and ASSD shows a noticeable increase and decrease respectively. In addition, the standard deviations for both metrics are reduced, indicating more consistent and stable performance across samples. This suggests that calibration not only helps with confidence estimates but also contributes meaningfully to segmentation quality.

Regarding the TPUS segmentation model, from the training set results, it is evident that the

model trained without calibration exhibits poor calibration, being under-confident for low probabilities and overconfident for high ones. When trained with calibration, there is an improvement - the average ACE decreases from 0.144 to 0.114, and the mean error curve aligns more closely with the ideal calibration line. In addition, the histogram becomes cleaner, with reduced spread, indicating a more structured distribution of prediction probabilities. However, the calibration is still sub-optimal, with residual miscalibration patterns. On the test set, calibration generalization is clearly more challenging, as reflected in noisier histograms. Nevertheless, there is a slight improvement in average ACE from 0.183 without calibration to 0.152 with calibration, demonstrating that the calibration loss helps the model produce better-calibrated outputs on unseen data. In terms of segmentation accuracy, the model trained with calibration also shows a modest improvement, as the mean Dice score improves and the ASSD decreases. Moreover, the standard deviations for both metrics decrease as well, indicating more consistent and stable segmentation results. While these improvements are not large, they indicate that incorporating calibration contributes positively to both the confidence estimates and segmentation quality.

Table 4.2: Segmentation results for the CT and TPUS segmentation models trained with and without calibration parameter

Dataset	Training Method	Dice (%)	ASSD (mm)
CT	Without Calibration	91.71 ± 5.88	3.77 ± 6.21
	With Calibration	93.54 ± 3.41	2.81 ± 3.23
TPUS	Without Calibration	71.17 ± 15.72	2.41 ± 1.39
	With Calibration	72.85 ± 13.86	2.28 ± 1.05

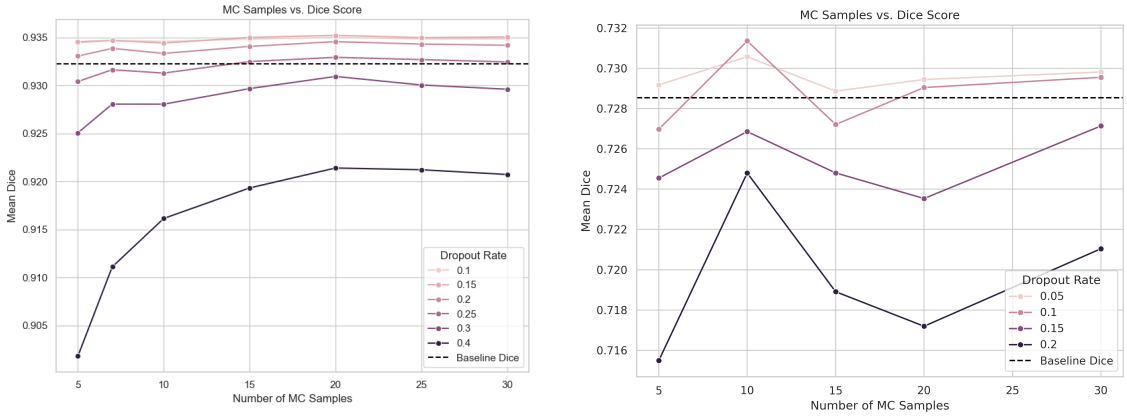
4.3 Tuning of Uncertainty Estimation Parameters

As explained in Section 3.5, the method adopted for uncertainty quantification in this work was the MC Dropout. Therefore, there are two key hyperparameters that are required, the number of MC samples (i.e., the number of forward passes) and the dropout probability applied during inference.

In uncertainty-aware segmentation, choosing the right combination of dropout rate and number of samples involves balancing three key factors - segmentation performance, uncertainty estimation quality, and computational cost. Higher segmentation performance is typically desired, but it must be weighed against the ability of uncertainty metrics, both voxel and volume-wise, to meaningfully correlate with prediction errors. On one hand, voxel-wise metrics provide fine-grained insights into uncertain regions, useful for assessing model confidence at the voxel level. On the other, structure-aware metrics capture broader spatial uncertainty, allowing for better understanding of confidence at the region or organ level. However, these benefits come with computational trade-offs, as increasing the number of MC samples is reported in the literature to improve both performance and uncertainty reliability but introduces higher inference time and computational cost, which may not always be acceptable in real-world pipelines. Similarly, lower dropout rates

may yield slightly higher Dice scores, but often reduce the variance necessary for reliable uncertainty quantification.

To evaluate the impact of these parameters, a two-stage analysis was performed. First, the effect of different combinations of MC dropout rate and sample count on segmentation performance, measured by the Dice score, was assessed. The results for both datasets are presented in Figure 4.5a and Figure 4.5b. These results indicate how sensitive the overall segmentation accuracy is to the uncertainty estimation configuration.



(a) CT segmentation model

(b) TPUS segmentation model

Figure 4.5: Variation in Dice score across different MC dropout rates and numbers of MC samples for (a) CT and (b) TPUS segmentation models.

Following this, the relationship between uncertainty and segmentation error was analyzed to assess the quality of the uncertainty estimates. For this, Spearman correlation was computed between the Dice scores and both voxel and structure-wise metrics.

4.3.1 Structure-wise Correlation

As described in the Methods section (Section 3.5.2.2), three structure-wise uncertainty metrics were evaluated:

- The raw summed voxel-wise entropy within each segmented structure.
- The entropy normalized by the structure's volume (volume-normalized entropy).
- The entropy normalized by the structure's surface area (surface area-normalized entropy).

To assess the relationship between uncertainty and segmentation quality at the structure level, the absolute Spearman correlation between each of these uncertainty metrics and the Dice error was computed across various configurations of MC dropout rate and number of samples. This analysis allowed to evaluate how well each uncertainty metric aligned with actual segmentation performance under different uncertainty estimation settings.

Among the three, the volume-normalized entropy consistently demonstrated stronger and more stable correlations with segmentation error across different modalities and configurations. In contrast, the raw summed entropy and surface area-normalized entropy showed less consistent and less meaningful correlations. Therefore, the volume-normalized metric was selected as the primary structure-wise uncertainty measure for further analyses in this study.

4.3.2 Voxel-wise Correlation

For the voxel-wise uncertainty evaluation, the analysis focused specifically on boundary voxels, as these regions typically exhibit the highest segmentation uncertainty and error rates. For each image in the dataset, boundary voxels were identified and their predicted uncertainty values were grouped into discrete intervals (bins). Within each uncertainty interval, the empirical misclassification probability (i.e., the proportion of voxels within that interval that were incorrectly segmented) was calculated. This process enabled the construction of uncertainty-error profiles, which reveal whether higher uncertainty values correspond to an increased likelihood of segmentation errors.

By aggregating the uncertainty and misclassification data across all images, the Spearman correlation between voxel-level uncertainty and error rate at boundary voxels was computed for different configurations of MC dropout rates and numbers of MC samples. This correlation served as an indicator of the quality and reliability of the voxel-wise uncertainty metric under various uncertainty estimation settings.

Figures 4.6 and 4.8 show the resulting correlations across all evaluated settings for the CT and TPUS datasets, respectively.

4.3.3 Analysis of Tuning Curves and Selection of Parameters

4.3.3.1 CT segmentation model

As shown in Figure 4.5a, the Dice score initially improves with an increasing number of MC samples, particularly for dropout rates above 0.2. For lower dropout values (0.1 to 0.2), performance is relatively stable and consistently surpasses the baseline Dice score, even with fewer samples. Notably, dropout values of 0.25 at 20 samples also result in performance above the baseline. Beyond 20 samples, improvements tend to plateau or even slightly decrease for all dropout settings.

In terms of uncertainty correlation, Figure 4.6 highlights two metrics of focus - voxel-wise correlation and volume-normalized entropy. Voxel-wise correlation remains relatively high across all dropout rates and numbers of samples, indicating consistent alignment between voxel-level uncertainty and segmentation errors. This suggests that voxel-wise uncertainty is a robust metric.

On the other hand, volume-normalized entropy shows a clearer trend with dropout - its correlation with performance increases with higher dropout values. However, its dependency on the number of MC samples is not systematic, with no strong pattern of improvement or degradation observed.

Based on these findings, the combination of dropout rate 0.25 and 20 MC samples was selected as the optimal setting. This choice balances several factors:

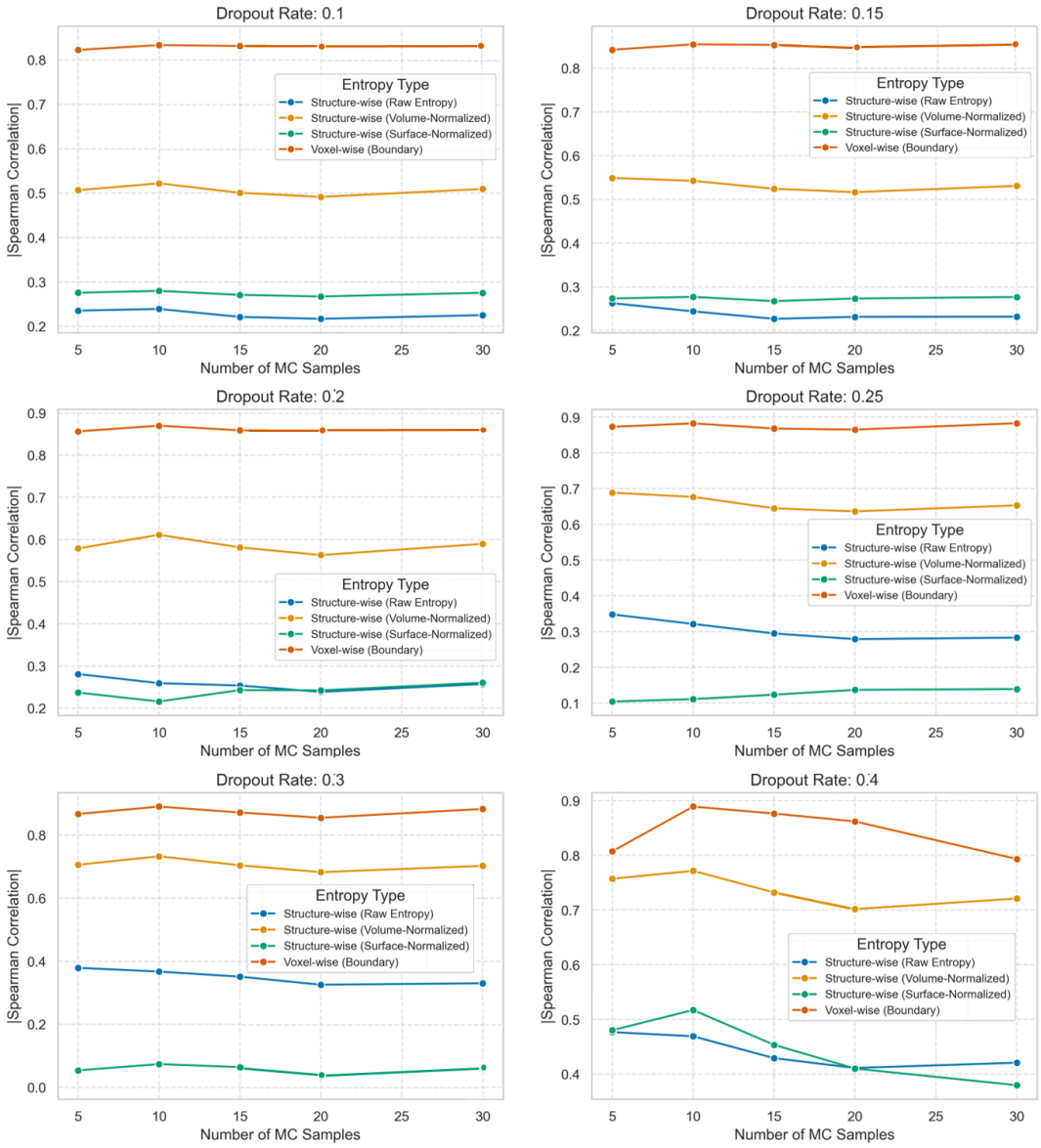


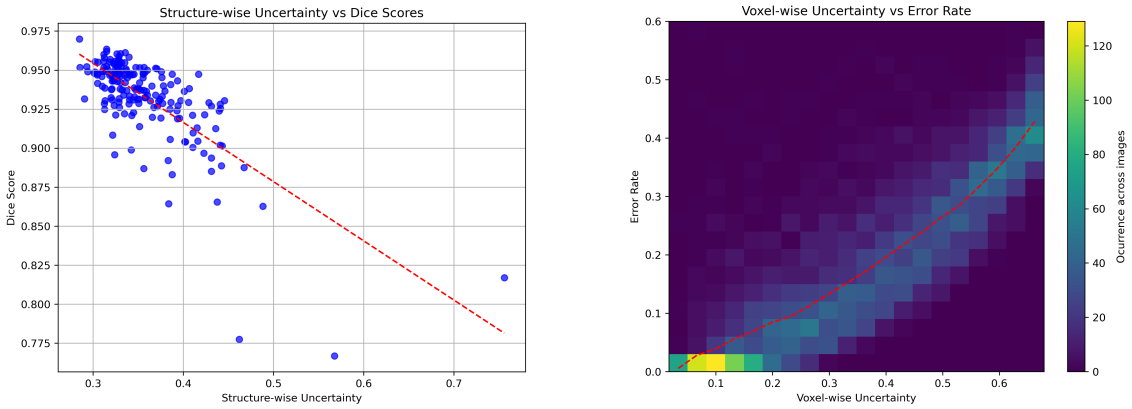
Figure 4.6: Correlation between uncertainty and segmentation error across different MC dropout rates and numbers of MC samples for the CT segmentation model. Each curve shows how the Spearman correlation varies with uncertainty estimation settings, capturing the relationship between predicted uncertainty and actual segmentation error at the structure and voxel levels.

- It yields a Dice score above the baseline, unlike higher dropout configurations that may lead to performance degradation.
- Voxel-wise correlation is among the highest at this setting, suggesting strong uncertainty–error alignment at the voxel level.
- Volume-normalized correlation, while not as high as in higher dropout conditions, still demonstrates good alignment.

- Computation-wise, using 20 samples is efficient and avoids unnecessary overhead, given that no significant Dice improvements were seen beyond this point.

This setting therefore offers a strong compromise between segmentation accuracy, uncertainty interpretability, and computational cost, and is used as the reference for subsequent analysis.

Figure 4.7 presents a visual summary of the results that underlie these correlation values for the selected setting of dropout rate 0.25 and 20 MC samples. Figure 4.7a shows the structure-wise relationship between segmentation accuracy and uncertainty, while Figure 4.7b illustrates the voxel-level uncertainty–error alignment. Together, they highlight how uncertainty metrics reflect segmentation quality under the chosen parameters. These graphs validate the strong correlation between uncertainty and performance, both at the volume and voxel levels. There is a clear negative correlation between structure-wise uncertainty and Dice scores, indicating that higher uncertainty corresponds to lower segmentation accuracy at the structure level. Similarly, a positive relationship is observed between voxel-wise boundary uncertainty and error rate. However, it is worth noting that even at the highest levels of voxel-wise uncertainty, the mean error rate remains around 0.45, suggesting that a substantial proportion of these uncertain voxels are still correctly segmented.



(a) (b)

Figure 4.7: Uncertainty–performance relationships for the CT segmentation model using the selected MC dropout configuration (dropout rate 0.25, 20 MC samples). (a) Structure-wise correlation between Dice score and volume-normalized summed entropy. Each point represents a volume and the red dashed line shows a linear regression fit. (b) Voxel-wise uncertainty–error profile computed over boundary voxels. Voxels are grouped into uncertainty bins (x-axis), with corresponding misclassification rates (y-axis). The red dashed line indicates the mean misclassification rate across bins. The color scale reflects the occurrence across images, indicating how many times each uncertainty–error pairing appears throughout the dataset.

4.3.3.2 TPUS segmentation model

Regarding the TPUS dataset, a dedicated tuning process was also conducted. Since the model was originally trained with a dropout rate of 0.1, tuning was performed using a reduced range of

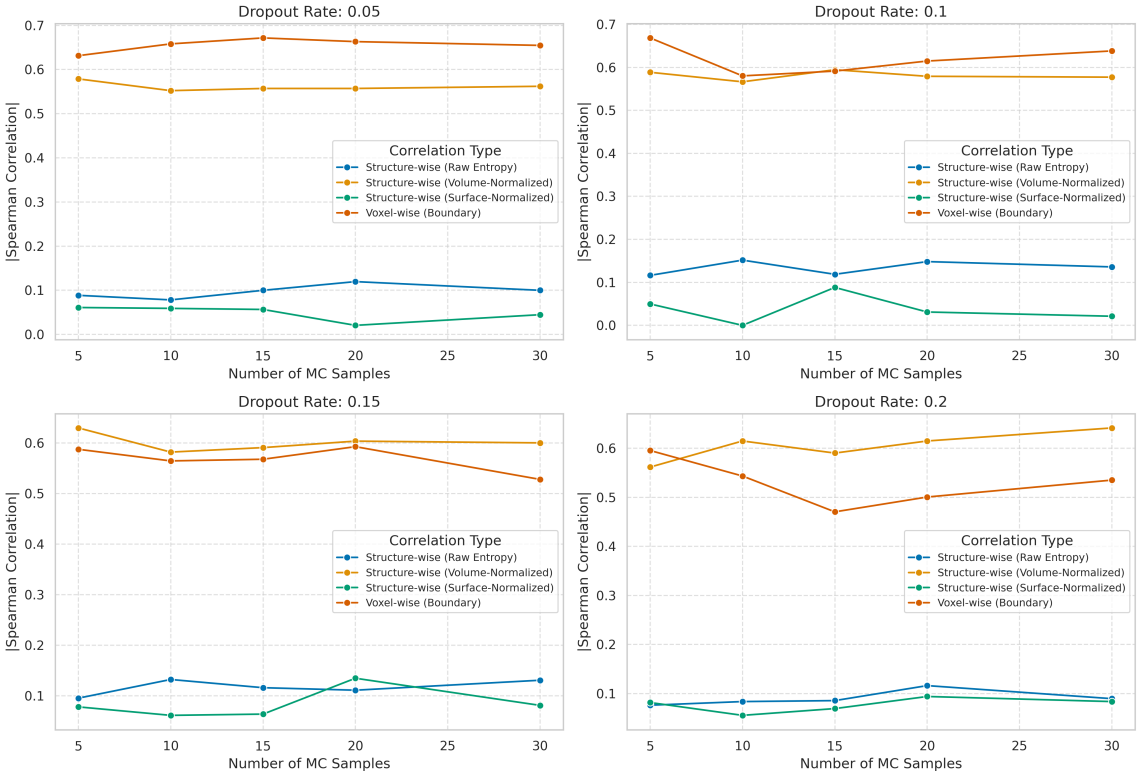


Figure 4.8: Correlation between uncertainty and segmentation error across different MC dropout rates and numbers of MC samples for the TPUS segmentation model. Each curve shows how the Spearman correlation varies with uncertainty estimation settings, capturing the relationship between predicted uncertainty and actual segmentation error at the structure and voxel levels.

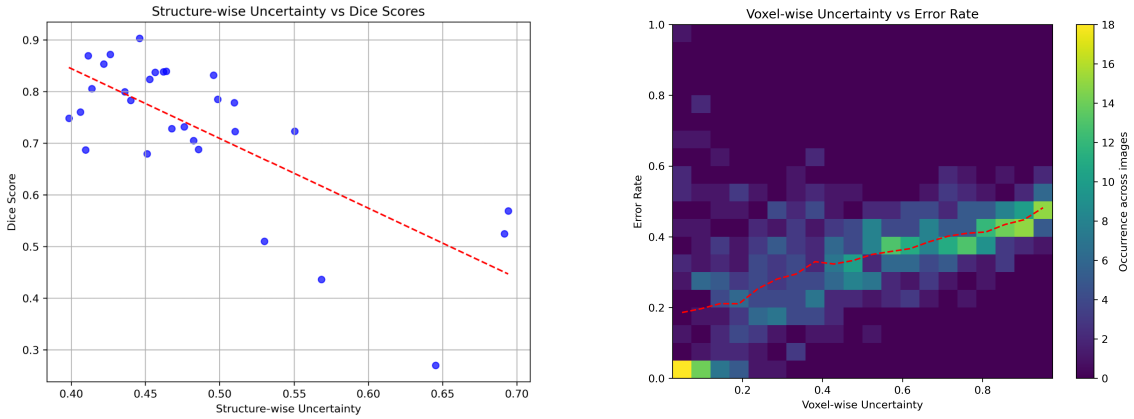
dropout values, focusing on smaller rates to reflect this constraint.

When analyzing the uncertainty–performance correlation curves, it is evident that overall correlations are generally lower in this scenario compared to the previous experiment. At a dropout of 0.05, the structure-wise correlation is relatively high, but voxel-wise correlation remains low, limiting its usefulness for voxel-level interpretability. Conversely, at dropout values of 0.15 and above, structure-wise correlation drops notably, reducing the effectiveness of uncertainty estimation at the volume level. Among the tested settings, dropout = 0.1 offers the best trade-off, maintaining a reasonable correlation for both structure-wise and voxel-wise metrics. Within this dropout setting, the use of 20 MC samples stands out as the most favorable, showing a consistent improvement in uncertainty correlation compared to lower sample counts.

In terms of Dice performance, this configuration of dropout 0.1 and 20 MC samples leads to a slight decrease compared to the baseline Dice, whereas using 15 samples yields a modest increase. However, the performance drop at dropout 0.1 with 20 samples is minimal and deemed acceptable given the stronger and more balanced uncertainty correlations achieved.

Therefore, the configuration of 20 MC samples with dropout 0.1 is adopted for the TPUS model, as it offers a practical compromise between segmentation accuracy and uncertainty interpretability.

Figure 4.9 presents a visual summary of the results that underlie these correlation values for the selected setting of dropout rate 0.1 and 20 MC samples. Figure 4.9a shows the structure-wise relationship between segmentation accuracy and uncertainty, while Figure 4.9b illustrates the voxel-level uncertainty–error alignment. Together, they highlight how uncertainty metrics reflect segmentation quality under the chosen parameters. In this scenario, worse results were expected compared to the CT results, based on the correlation curves analysed. Still, a negative trend can be observed in the structure-wise analysis, indicating that higher uncertainty continues to correlate with lower Dice scores. In the voxel-wise plot, the mean curve maintains an overall ascending tendency, suggesting that uncertainty still reflects local error to some extent. However, the histogram is noticeably noisier, the trend is less steep, and there is a broader spread of high error rates at lower uncertainty values. This indicates a weaker but still present alignment between uncertainty and segmentation error. It highlights that while uncertainty remains a useful indicator, its reliability in identifying segmentation errors may diminish in lower-performing models.



(a) (b)

Figure 4.9: Uncertainty–performance relationships for the TPUS segmentation model using the selected MC dropout configuration (dropout rate 0.1, 20 MC samples). (a) Structure-wise correlation between Dice score and volume-normalized summed entropy. Each point represents a volume and the red dashed line shows a linear regression fit. (b) Voxel-wise uncertainty–error profile computed over boundary voxels. Voxels are grouped into uncertainty bins (x-axis), with corresponding misclassification rates (y-axis). The red dashed line indicates the mean misclassification rate across bins. The color scale reflects the occurrence across images, indicating how many times each uncertainty–error pairing appears throughout the dataset.

Chapter 5

Leveraging Uncertainty on the Baseline Model

This chapter presents the results of uncertainty quantification applied to the baseline segmentation models and explores its practical utility through a series of targeted experiments[1].

To provide a comprehensive analysis, both structure-wise and voxel-wise uncertainty metrics are employed. These complementary perspectives allow for assessment of uncertainty at different granularities, structure-level and voxel-level. The results are reported for both CT and TPUS datasets.

By conducting these experiments on the baseline model, the aim is to gain a better understanding of the strengths and limitations of uncertainty estimation methods before integrating them into more advanced interactive models.

5.1 Leveraging Structure-wise Uncertainty Estimates

To evaluate the practical relevance of structure-wise uncertainty estimates, two analyses were conducted, both based on the premise that these estimates can be used to prioritize cases for manual review or further inspection in clinical workflows.

5.1.1 Experiments with simulated perfect corrections

In this experiments, the utility of structure-wise uncertainty for guiding manual review was assessed by simulating a scenario in which images are progressively reviewed and corrected. Images were selected successively according to their structure-wise uncertainty, beginning with the most uncertain. Each reviewed image was assumed to be perfectly corrected, and its Dice score was consequently set to 1.0. At each step, the mean corrected Dice score across the dataset was

[1] Parts of the findings presented in this chapter were submitted as an abstract to the IUGA/EUGA Joint Meeting, titled *"Automated External Anal Sphincter Analysis: Leveraging AI Uncertainty for Future Clinical Adoption"*, and were selected as an Unmoderated E-poster Presentation.

computed, enabling visualization of the expected performance improvement as a function of the number of reviewed cases.

To contextualize the effectiveness of this uncertainty-driven selection, two reference strategies were included:

- **Random Selection:** Images were reviewed in a random order. This simulation was repeated 50 times, and the mean trajectory was reported to provide a stochastic baseline.
- **Ideal Dice-based Selection:** An idealized scenario in which images are reviewed in ascending order of Dice score, corresponding to the actual worst-performing segmentations being corrected first.

This analysis aims to determine whether structure-wise uncertainty provides a meaningful criterion for prioritizing review and improving overall segmentation performance.

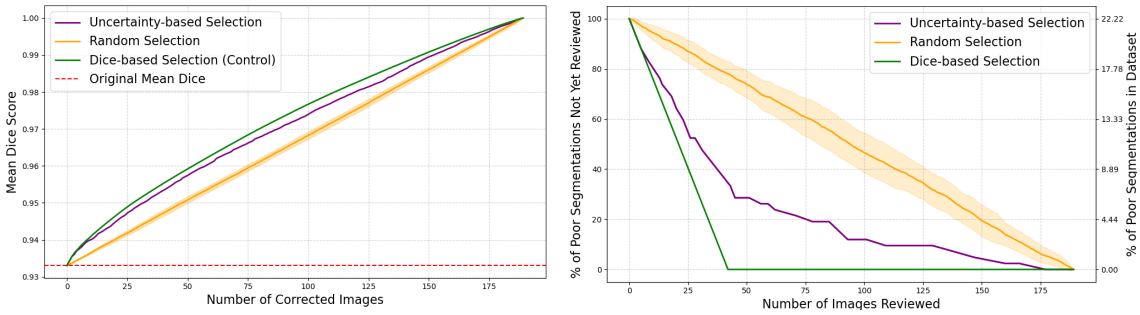
The results for CT data are shown in Figure 5.1a, while the corresponding results for TPUS data appear in Figure 5.2a.

A complementary analysis was performed to assess how effectively structure-wise uncertainty can identify low-quality segmentations. In this case, the percentage of poor segmentations remaining in the dataset was tracked as cases were reviewed successively according to the same three previous strategies - uncertainty-based, random, and Dice-based.

Poor segmentations were defined as those with Dice scores below the inter-observer Dice, chosen to reflect the expected variability between human annotators for each modality, considered a marker for clinical acceptability. At each step, as more cases were assumed to be reviewed and corrected, the proportion of poor segmentations was recalculated. The comparison between strategies enables evaluation of the utility of uncertainty estimates in reducing the number of problematic cases early in the review process.

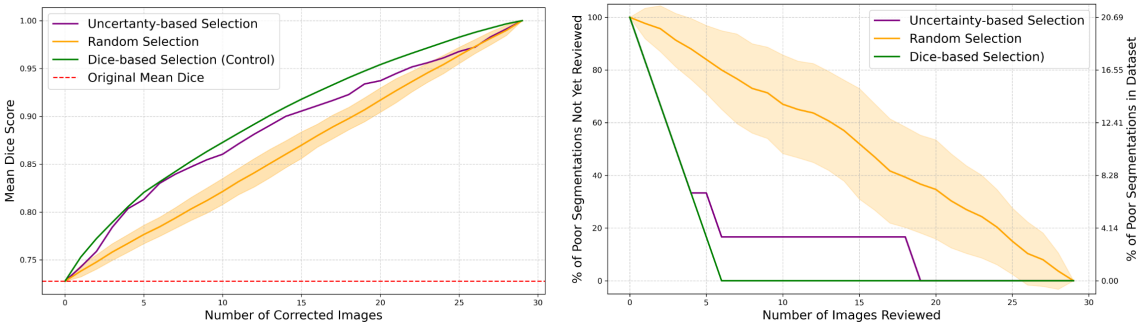
The results of this experiment for CT data are shown in Figure 5.1b, while the corresponding results for TPUS data appear in Figure 5.2b.

The results demonstrate that structure-wise uncertainty estimates are effective for guiding image review and correction in CT data. As shown in Figure 5.1a, selecting images for review based on uncertainty leads to a faster and more consistent improvement in mean Dice score compared to random selection. While the idealized Dice-based selection strategy unsurprisingly achieves the steepest performance gains, by correcting the worst segmentations first, the uncertainty-based approach closely approximates this trajectory, particularly in the early stages of review. This suggests that the structure-wise uncertainty metric serves as a strong proxy for segmentation quality, which is a critical advantage in practice, given that the Dice score cannot be computed on unseen data. Complementary findings in Figure 5.1b further support this conclusion, revealing that uncertainty-based selection is significantly more efficient at reducing the proportion of poor segmentations than random review. Specifically, to reduce the percentage of poor segmentations in the dataset to 5%, only 63 images need to be reviewed using the uncertainty-guided approach, compared to approximately 142 images required with random selection. This corresponds to an approximate 56% reduction in the number of reviews. These results underscore the potential of



(a) (b)

Figure 5.1: Structure-wise uncertainty experiments results for CT data. (a) Mean Dice score as a function of the number of reviewed CT images. (b) Percentage of poor segmentations not yet reviewed as images are successively corrected, as well as the corresponding percentage in the dataset. In both plots, the purple line represents uncertainty-based selection, the orange line shows random selection (over 50 runs), and the green line represents the ideal Dice-based selection strategy. The shaded region around the random selection curve denotes the standard deviation across the 50 runs.



(a) (b)

Figure 5.2: Structure-wise uncertainty experiments results for TPUS data. (a) Mean Dice score as a function of the number of reviewed TPUS images. (b) Percentage of poor segmentations not yet reviewed as images are successively corrected, as well as the corresponding percentage in the dataset. In both plots, the purple line represents uncertainty-based selection, the orange line shows random selection (over 50 runs), and the green line represents the ideal Dice-based selection strategy. The shaded region around the random selection curve denotes the standard deviation across the 50 runs.

uncertainty-guided review strategies to improve segmentation quality more effectively and efficiently.

The results for TPUS data, presented in Figure 5.2, also demonstrate that structure-wise uncertainty estimates offer benefit for guiding image review. As seen in Figure 5.2a, uncertainty-based selection yields a performance gain in the early stages of review, outperforming random selection in terms of mean Dice score improvement. However, this advantage diminishes as more images are corrected, and the trajectory begins to converge toward that of random selection. This suggests that while uncertainty is useful for identifying the most problematic cases early on, its discriminative power becomes less effective once the most uncertain (and likely poorest) segmentations have

been addressed.

As shown in Figure 5.2b, uncertainty-based selection demonstrates high initial effectiveness, closely mirroring the performance of the Dice-based strategy and rapidly targeting poor segmentations. This highlights the potential of uncertainty as a reliable indicator for prioritizing problematic cases. Following this initial sharp decline, the curve exhibits a temporary plateau, indicating a short period during which no additional poor segmentations are identified. However, it is important to note that at this stage, the proportion of poor segmentations in the dataset has already been reduced to approximately 4%. Despite this plateau, the uncertainty-guided approach ultimately achieves the complete elimination of poor segmentations by the 19^{th} reviewed image, outperforming the random selection strategy, which displays slower and more variable progress. Furthermore, to reduce the percentage of poor segmentations in the dataset to below 5%, only 6 images need to be reviewed using uncertainty-based selection, compared to approximately 23 images with random selection. This corresponds to a 74% reduction in the number of reviews.

It is important to note that these results for the TPUS are influenced by the much smaller dataset used in the experiment experiment (only 29 images), which limits the robustness of the comparison, especially for the random strategy, which exhibits high variance due to the small sample size. This may partially explain the closer alignment between random and uncertainty-based performance in later review stages.

Overall, these findings indicate that structure-wise uncertainty estimates are valuable for prioritizing early review efforts in both CT and TPUS images.

5.1.2 Experiments on CT Dataset using observer variability

In the previous uncertainty-guided review experiments, segmentation corrections following manual review were simulated by artificially setting the corrected Dice score to 1.0. While this approach helps measure an upper bound on potential improvement, it is inherently idealized and unrealistic, since it assumes that a human reviewer will always perfectly correct the segmentation to match the ground truth, which does not reflect actual clinical variability.

To simulate a more practical and clinically relevant review interaction under realistic annotation variability in a medical workflow with human-in-the-loop corrections, the uncertainty-guided review experiments were repeated on a subset of 30 volumes drawn from the CT test set. For each of these volumes, two additional expert segmentations were available - one from a different clinician and another from the same clinician who originally annotated the volume. This setup allowed the computation of both inter-observer Dice scores, which measure agreement between different experts, and intra-observer Dice scores, which quantify the consistency of a single expert across multiple annotations of the same image. The same experiment could not be performed on the TPUS data, as additional segmentations from other clinicians were not available for the corresponding test volumes.

The same experimental framework was applied in this approach, with one key modification. Instead of assuming perfect correction (setting corrected images' Dice scores to 1.0), the Dice score of each corrected volume was replaced with either the inter-observer Dice or intra-observer

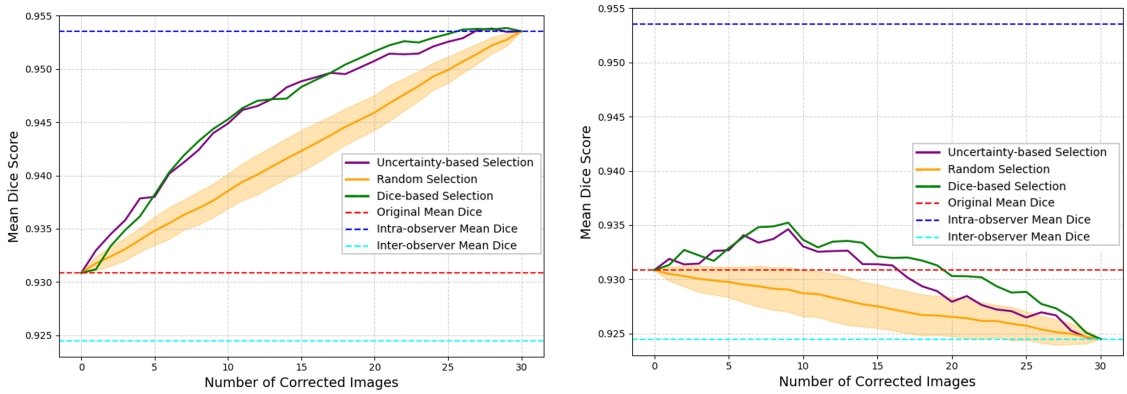
Dice score for that volume, depending on the experimental condition. This simulates a more realistic outcome of manual review, reflecting the variability among clinicians rather than an idealized perfect segmentation.

By using intra- and inter-observer Dice values instead, this setup reflects plausible upper and lower bounds for performance gains in clinical workflows. The intra-observer case represents a low-variance setting, where corrections are highly reliable and closely aligned with training data, possibly because the model was trained on segmentations by the same expert, enhancing its compatibility. In contrast, the inter-observer setting models real-world conditions where corrections may come from different clinicians with varying annotation styles. This simulates clinical workflows involving second opinions, crowdsourcing, or large-scale collaborative reviews.

In the intra-observer case (Figure 5.3a), where corrections mimic those made by the same expert who created the original ground truth, the corrected mean Dice consistently increases across all strategies, eventually approaching the intra-observer mean Dice of 95.35 ± 1.31%. The uncertainty-based selection performs similarly to the ideal Dice-based strategy and clearly outperforms random selection, particularly in early steps. Interestingly, uncertainty-based selection even slightly surpasses Dice-based selection in the initial phase, suggesting that images with lower Dice may not always yield the most significant performance gains when corrected.

In contrast, when corrections are guided by inter-observer Dice scores (Figure 5.3b), the trend is quite different. Here, corrections are based on segmentations from a different expert, and the mean inter-observer Dice is lower at 92.45 ± 1.98%, even below the model's baseline performance, causing the corrected mean Dice to eventually decline, even in the Dice-based correction, as more segmentations are replaced with more variable annotations, especially after the initial corrections. At the beginning, both uncertainty-based and Dice-based selection strategies generally lead to an increase in mean Dice and outperform random selection, reaching a peak around the correction of 9 images. This suggests that the most problematic predictions are being corrected early on. However, even in this initial phase, the improvement is not steady, the curves fluctuates, with occasional decreases in performance, indicating that the variability introduced by inter-observer disagreement leads to an unstable correction process. As corrections accumulate, the uncertainty-guided curve later approaches the random selection curve, highlighting the difficulty of maintaining improvement in the presence of annotation noise. It is also important to note that this decline is relative to the original ground truth, as corrections made by a different expert are penalized when they deviate from the initial annotations, even if clinically valid.

Overall, the results remain positive across both scenarios. As expected, the intra-observer setting shows visually better and more stable improvements, reflecting the high agreement between annotations made by the same clinician, who also annotated the ground truth. In contrast, the inter-observer setting exhibits more irregular behavior and overall lower performance, as corrections are based on annotations from a different clinician. Nonetheless, inter-observer Dice represents the standard level of acceptable agreement between experts in clinical practice, and therefore serves as a meaningful benchmark. Notably, uncertainty-based selection consistently shows promising results in both settings, particularly in the early stages of correction, supporting its use as an



(a) Intra-Observer Corrections

(b) Inter-Observer Corrections

Figure 5.3: Results of structure-wise uncertainty-guided review on CT data, showing the evolution of mean Dice score as a function of the number of corrected images. (a) Corrections modeled using intra-observer Dice scores, reflecting consistent single-expert review. (b) Corrections modeled using inter-observer Dice scores, simulating multi-expert variability. In both plots, the purple line represents uncertainty-based selection, the orange line shows the mean of 50 runs of random selection (with shaded area indicating standard deviation), and the green line corresponds to the ideal Dice-based oracle selection. The red dashed line shows the original (uncorrected) mean Dice score, while the blue and cyan dashed lines indicate intra- and inter-observer mean Dice scores, respectively.

effective strategy for prioritizing review and improving segmentation reliability in practice, while also emphasizing the need to account for annotation variability when designing and evaluating correction pipelines.

5.2 Leveraging Voxel-wise Uncertainty Estimates

As described in Section 3.5.2.1, a two-step filtering pipeline was applied to the voxel-wise uncertainty maps to enhance the identification of potential segmentation errors. This process is visually summarized in Figure 5.4 for the CT data and in Figure 5.5 for the TPUS data. Each figure presents two representative examples, helping to clarify the motivation behind each step by enabling a clearer comparison of how uncertainty is progressively refined throughout the pipeline.

From these figures, it is evident that the initial raw uncertainty maps exhibit widespread noise across the volume, with noticeably higher concentrations of uncertainty around the segmentation boundaries. This pattern, consistently observed across samples, supports the focus on boundary voxels as indicators of potential segmentation errors, since uncertainty tends to be naturally higher and more informative in these regions.

Although focusing on boundary voxels effectively reduces the search space, variability within these regions remains high. As shown in the figures, many boundary voxels display elevated uncertainty, but not all of them correspond to actual segmentation errors.

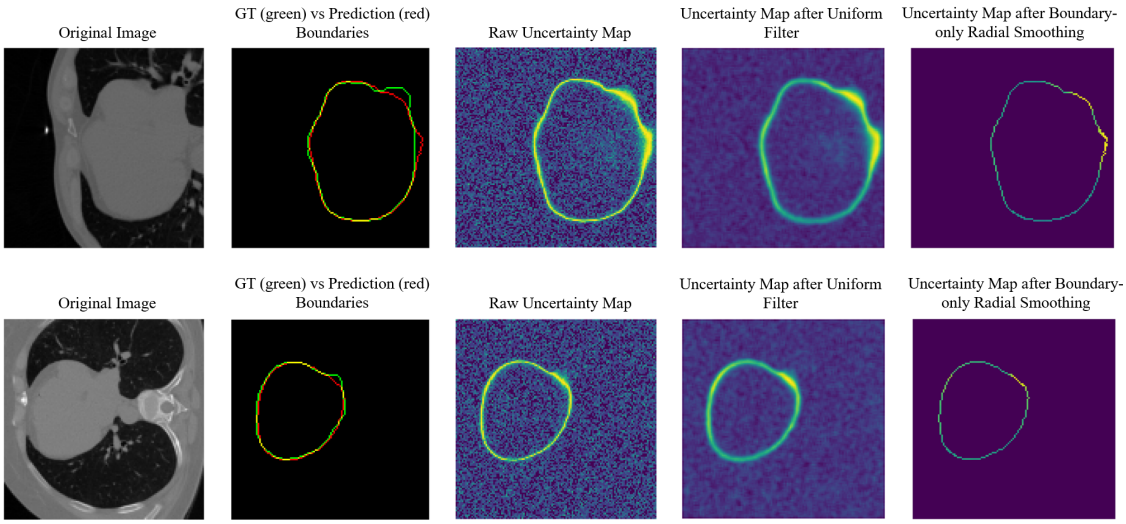


Figure 5.4: Visualization of the uncertainty refinement process for two CT segmentation examples (top and bottom rows). Each row shows: (1) the original image; (2) ground truth (green) and predicted (red) segmentation boundaries; (3) raw voxel-wise uncertainty map; (4) uncertainty map after applying a 3×3×3 uniform averaging filter, reducing noise and emphasizing coherent uncertain regions; and (5) final uncertainty map after radial filtering applied only to boundary voxels, suppressing non-boundary uncertainty and enhancing alignment with actual missegmentation areas. The final map serves as input for the correction-guided experiments.

The application of the uniform filter, visible in the fourth panel of each figure, reduces this variability by suppressing isolated high-uncertainty voxels and highlighting broader, more coherent regions. This improves visual interpretability and better emphasizes zones likely to correspond to true segmentation ambiguities. However, the uncertainty still tends to spread across surface voxels.

Another key insight from this fitered map was that missegmented regions tended to manifest as broader or thicker zones of high uncertainty after filtering, compared to general thin, well-defined boundaries in correctly segmented areas. To address this, the radial filtering step, shown in the final panel, further refines the signal by smoothing uncertainty along directions orthogonal to the boundary. This approach aligns with the typical manifestation of over- and under-segmentation and helps emphasize thicker, uncertain zones that are more likely to represent missegmented areas. Radial filtering involved defining a direction from the CoM to each boundary point and computing an average uncertainty from the boundary point and six neighboring points along this direction, which was then assigned to the boundary voxel. A schematic of this method is shown in Figure 5.6, illustrating the radial filtering process.

By comparing the final uncertainty maps with the ground truth and initial segmentation differences, it becomes clear that the refined maps localize problematic regions more accurately. They suppress spurious noise and highlight coherent areas of interest, forming a more reliable basis for the correction-guided strategies discussed in the remainder of this chapter.

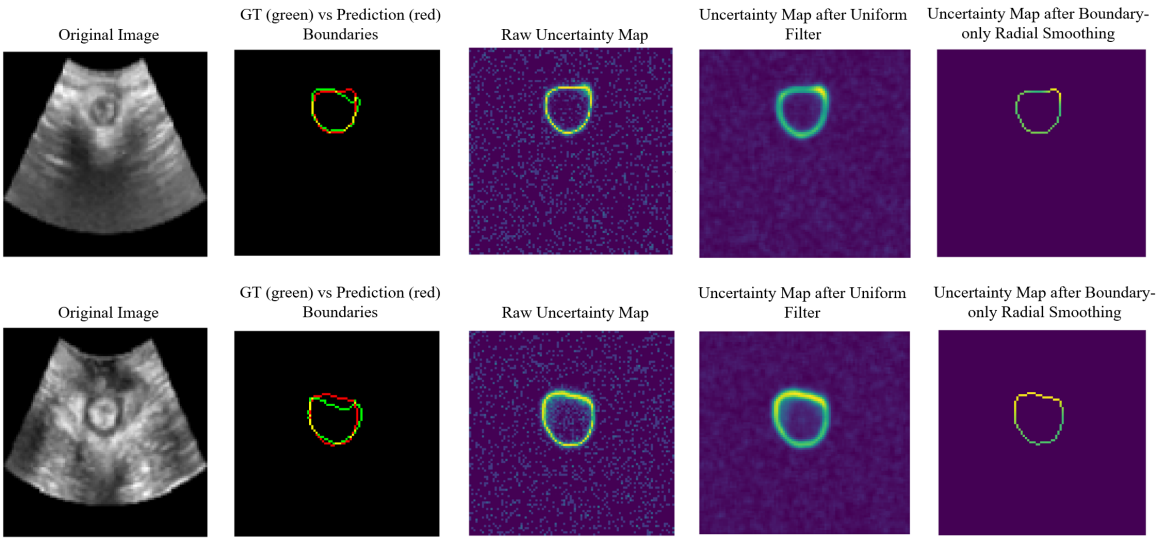


Figure 5.5: Visualization of the uncertainty refinement process for two TPUS segmentation examples (top and bottom rows). Each row shows: (1) the original image; (2) ground truth (green) and predicted (red) segmentation boundaries; (3) raw voxel-wise uncertainty map; (4) uncertainty map after applying a 3×3×3 uniform averaging filter, reducing noise and emphasizing coherent uncertain regions; and (5) final uncertainty map after radial filtering applied only to boundary voxels, suppressing non-boundary uncertainty and enhancing alignment with actual missegmentation areas. The final map serves as input for the correction-guided experiments.

Notably, in the first example of Figure 5.4, the regions of high uncertainty appear very prominent and distinctly highlighted, whereas in the second example of the same figure the highlighting is more subtle, though still noticeable. This variation reflects the natural differences in segmentation difficulty and prediction behavior across cases — some segmentations exhibit more pronounced uncertainty patterns, while others are more borderline or ambiguous. The filtering scheme was intentionally designed to be robust to such variability, ensuring that it can consistently provide meaningful and usable uncertainty information, even in cases where uncertainty patterns are less pronounced.

5.2.1 Patch Correction Experiments

Building upon the voxel-wise uncertainty maps introduced in the previous section, experiments were designed with the goal of evaluating the practical applicability of these maps in guiding corrective actions. The objective was to explore whether areas identified as highly uncertain could effectively indicate regions in need of correction.

To this end, a simplified correction simulation was designed. No interactive tool or model was involved at this stage, as this experiment was purely exploratory and aimed at evaluating the potential of uncertainty-driven correction without introducing human or algorithmic bias.

In practice, patches were defined around voxels exhibiting the highest uncertainty. Within each of these identified regions, the corresponding predicted segmentation was substituted with

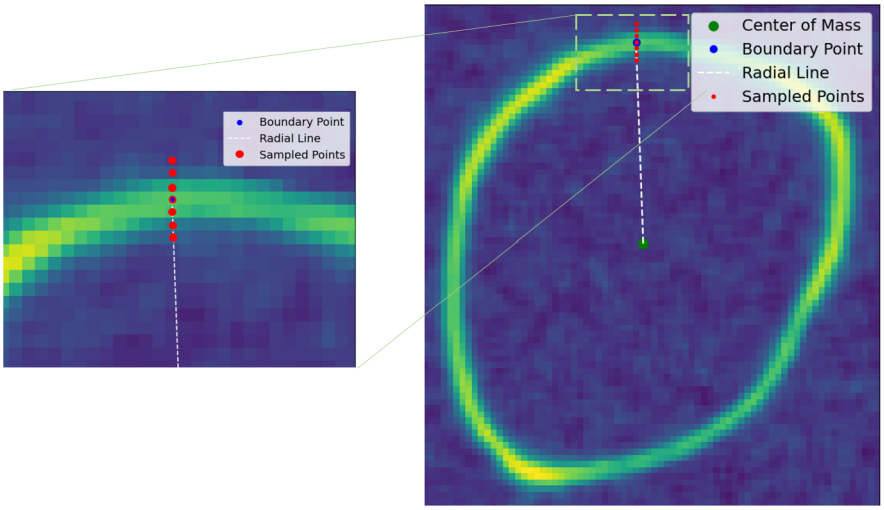


Figure 5.6: Illustration of radial entropy smoothing applied to a 3D entropy map. The image shows a 2D slice through the CoM of the segmented structure. The green dot marks the center of mass, the blue dot indicates one example boundary voxel, and the white dashed line represents the radial direction from the center to this boundary point. Red dots denote the sampled points along this radial line, used to compute a local average. The left panel provides a zoomed-in view of the boxed region in the main image. This figure illustrates the process for a single boundary voxel, but the smoothing is applied to all boundary voxels across the segmentation.

the ground truth label. This process, referred to as patch correction, was applied systematically to simulate the effect of addressing uncertainty-highlighted regions.

It is important to note that this approach has inherent limitations. Most notably, the corrections were not informed by human interaction knowledge and patch-level corrections applied independently across the volume may not preserve the anatomical coherence or structural continuity required for realistic segmentation. While individual patches may be corrected accurately in isolation, their collective integration into the full volume can result in fragmented or biologically implausible forms.

Despite these limitations, this method served as an initial approximation to assess whether uncertainty estimates could meaningfully guide localized corrections. The insights gained here lay the groundwork for methods that integrate uncertainty with more sophisticated, context-aware correction mechanisms.

The first experiment within this framework followed a structured methodology. Starting from the filtered uncertainty maps, the most uncertain boundary point in each image was identified. At that location, the radial distance to the ground truth boundary was computed, and a cubic patch was defined with dimensions twice that distance, ensuring sufficient coverage of the surrounding context. The ground truth segmentation was then inserted into the patch region, simulating a targeted correction. This process was repeated iteratively. After each correction, the next most uncertain point was selected, subject to a constraint that ensured it did not fall within any previously corrected patch. While some overlap between patches was allowed, this constraint guaranteed that

newly selected points had not already been addressed. Through this method, corrections were simulated incrementally, with each image receiving a fixed number of corrections per iteration (e.g., one correction per image, two corrections per image, and so on). At each stage, the mean Dice score was computed across the dataset, enabling a quantitative analysis of performance improvement as a function of the number of corrections.

Figure 5.7 provides an example from the CT dataset, illustrating the selection of the most uncertain point on the uncertainty map, the creation of a correction patch based on the radial distance to the ground truth boundary, and the insertion of the ground truth segmentation within that patch.

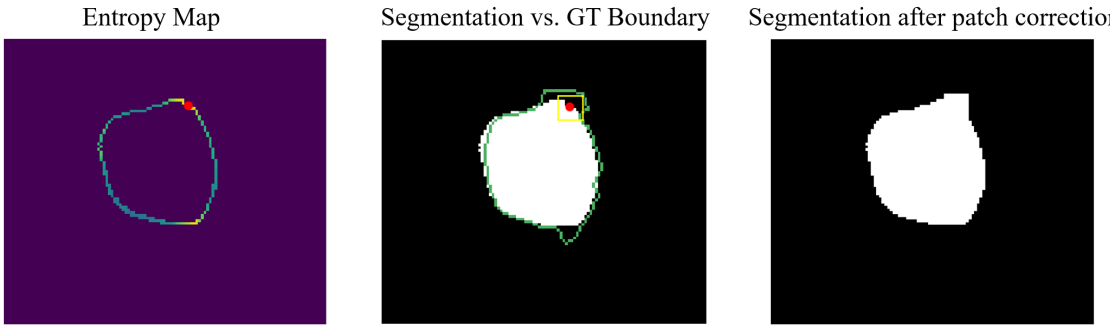


Figure 5.7: Example from the CT dataset illustrating the correction process. (Left) Uncertainty map with the most uncertain boundary point marked in red. (Middle) Current segmentation overlaid with the ground truth boundary in green. The most uncertain point is highlighted in red, and the yellow square shows the 2D slice of the cubic patch centered on that point used for correction. (Right) Updated segmentation after pasting the ground truth within the patch, simulating the targeted correction.

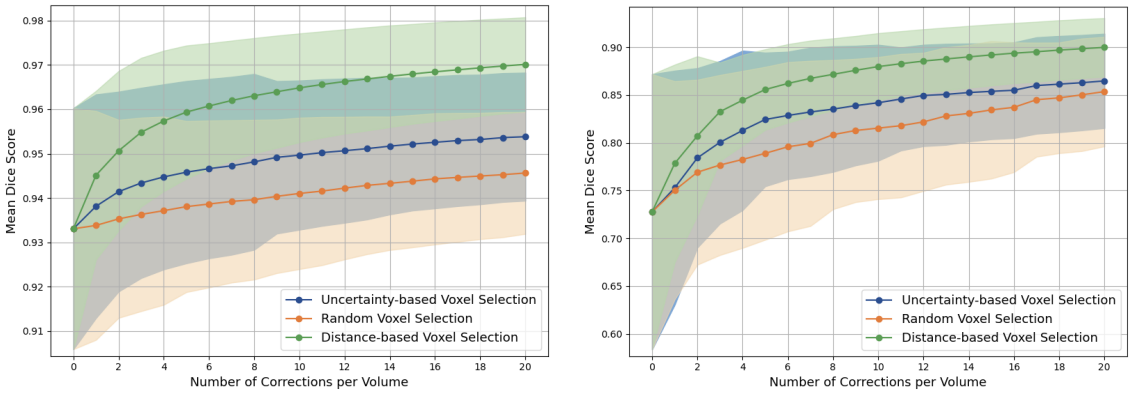
To assess the effectiveness of uncertainty-driven correction, this approach was compared against two alternative baselines:

- Random Patch Selection: Points were randomly selected as patch centers, while maintaining the same patch size definition and non-overlapping constraint. This baseline provided a reference for measuring whether uncertainty-based selection offered an actual advantage over unstructured correction.
- Max-Distance Patch Selection: Points with the greatest radial distance from the predicted boundary to the ground truth were selected as correction centers, reflecting the best-case selection strategy within this framework.

While any correction using ground truth data will naturally enhance the Dice score, the focus here was on the relative efficiency and prioritization capabilities of uncertainty estimates.

Figure 5.8 presents the results of this first experiment for the CT and TPUS segmentation models. The plots compare the mean Dice score as a function of the number of simulated corrections, using the three selection strategies - uncertainty-based, random, and distance-based. While

all methods show performance improvements with increasing corrections, the distance-based approach leads to the most rapid gains, followed by uncertainty-guided and random selection.



(a) Results for the CT model

(b) Results for the TPUS model

Figure 5.8: Mean Dice score progression as a function of the number of simulated patch corrections. Three voxel selection strategies are compared: uncertainty-based, random, and distance-based. At each correction step, all images receive the same number of patch corrections. The shaded regions represent ±1 standard deviation across the dataset.

In both cases, a clear separation can be observed between the uncertainty-guided and random correction curves. Although the improvement offered by uncertainty guidance is noticeable, it remains relatively modest and not strongly differentiated from the random baseline. In contrast, the curve corresponding to the distance-based correction approach consistently shows significantly higher Dice scores across all correction levels. This outcome aligns with expectations to some extent, as the distance-based strategy is designed to favor large, high-impact corrections. By always selecting the point with the greatest radial deviation from the ground truth boundary, this method ensures that the correction patch is centered over the most severe segmentation error. Since the patch size is determined as twice the measured distance at that point, the patches used in this strategy are, by definition, the largest possible corrections in this experiment at each step. As a result, the patches are more likely to cover entire missegmented regions in fewer steps.

For the uncertainty-guided approach, the observed improvements over random selection suggest that uncertainty is indeed informative. The improvement offered by uncertainty guidance, while modest, is still a meaningful indicator of its effectiveness compared to the random baseline. However, the gains are comparatively limited and the gap to the upper-bound curve is substantial, which may be attributed to two key limitations. First, regions of high uncertainty may occasionally not correspond precisely to regions of segmentation error, meaning that some corrections may be wasted on areas that are not truly missegmented, yielding little to no improvement. Second, even when uncertain regions do overlap with errors, the specific voxel selected as most uncertain may not lie at the point of greatest deviation from the ground truth. As a result, the corresponding patch may be too small to correct the full extent of the error, leading to the need for multiple successive corrections to resolve a single missegmented area.

Overall, the patch size definition used in this initial experiment, based on boundary distance, was a logical way to tie correction size to the severity of local error, but it does not fully reflect how a human expert would respond. In a real-world scenario, if a clinician were informed that a particular voxel was uncertain, they would not limit their correction to that voxel alone or to a narrow neighborhood determined by a distance metric. Instead, they would likely assess the broader context around the uncertain point and manually correct the entire missegmented region in one step. Thus, restricting the correction size purely based on local boundary distance underestimates the scope of real clinical interventions. This mismatch between patch size and the actual extent of the error is illustrated in Figure 5.7, where the missegmented region is detected but not fully corrected, making the process less efficient and more fragmented. In contrast, a human expert, when presented with the same uncertain point, would likely inspect the broader context around it and proceed to correct the entire missegmented region in a single step.

To better approximate this behavior, an additional constraint was introduced into the patch size definition - a minimum patch size based on the typical size of the segmented structure. Given that both target anatomies in this work are approximately spherical, the structure's average radius was estimated by first computing its CoM, then measuring the radial extent in multiple directions, and finally averaging these values. This adaptive minimum radius was then used as a lower bound for patch size. Specifically, for each uncertain point, the radial distance to the ground truth boundary was still computed, but if twice that distance fell below the structure's estimated radius, the radius value was used instead. This adjustment ensured that correction patches would be large enough to reflect the kind of region-wide correction a clinician would make in response to uncertainty, while maintaining adaptiveness to anatomical scale. It also avoided overly aggressive corrections in small structures that could result from using a fixed global minimum patch size.

Figure 5.9 illustrates this concept using the same volume from Figure 5.7, but showing both the patch defined by twice the distance to the ground truth boundary and the larger patch defined by the estimated radius. In this case, since the radius-based patch is larger, it is selected over the distance-based patch for simulating the correction in the pipeline. As demonstrated in the refined segmentation, this approach allows correction of the entire region associated with the missegmentation identified by the uncertain point, better reflecting a clinician's region-wide correction.

The results of this refined experiment, incorporating the adaptive minimum patch size constraint, are presented in Figure 5.10 for both the CT and TPUS datasets, following the same format as the previous set of results.

The results demonstrate a clear and consistent improvement across all three voxel selection strategies, for the two applications, when an adaptive minimum patch size is incorporated into the correction process, as this modification, which ensures patches are not smaller than the estimated radius of the target anatomy, leads to more substantial corrections. This is evident from the steeper improvement in Dice scores across the correction range for all methods, reflecting the greater impact of each correction.

Most notably, the uncertainty-guided strategy shows a significant improvement compared with the previous experiment, with its curve now closely tracking the distance-based (ideal) curve, es-

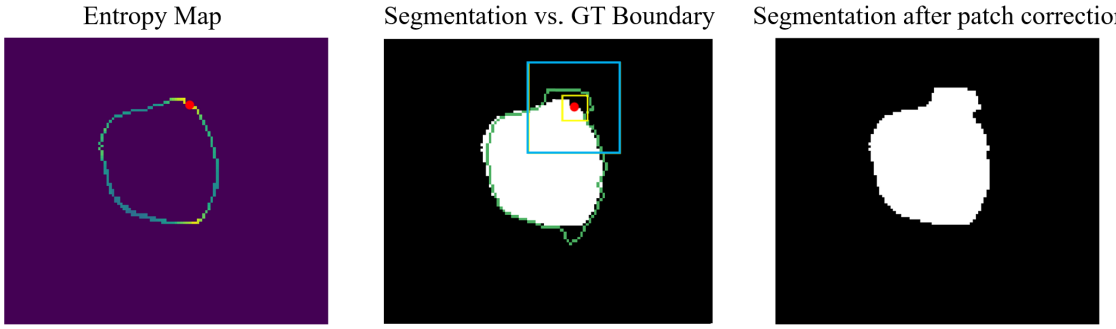
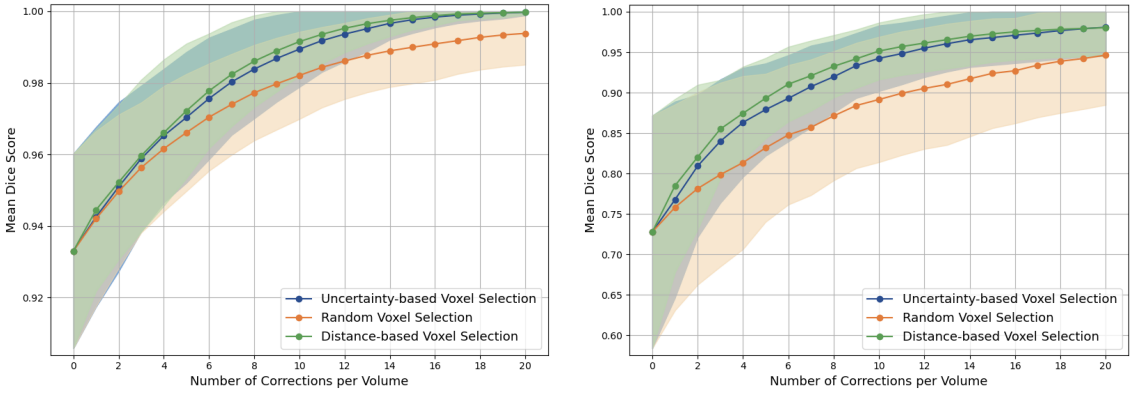


Figure 5.9: Example from the CT dataset illustrating the adaptive patch size correction. (Left) Uncertainty map with the most uncertain boundary point marked in red. (Middle) Current segmentation overlaid with the ground truth boundary in green. The most uncertain point is highlighted in red. Two patches are shown: the yellow square represents the patch defined by twice the distance to the ground truth boundary, and the blue square indicates the patch based on the estimated average radius of the structure. Since the blue patch is bigger, it is used for the correction. (Right) Updated segmentation after pasting the ground truth within the radius-based patch, simulating the targeted correction.

pecially as the number of corrections increases. This is a particularly encouraging result, as it demonstrates that uncertainty estimates are highly informative and can effectively approximate optimal correction strategies, even in the absence of direct ground truth knowledge. In the CT dataset (Figure 5.10a), for instance, both uncertainty and distance-based methods converge to nearly perfect Dice scores after approximately 16 corrections per volume, with minimal standard deviation, indicating robust and consistent performance. For the TPUS (Figure 5.10b), the uncertainty-guided curve also converges to the distance-based curve, with Dice scores improving markedly and approaching values close to one, while a reduction in standard deviation is also observed throughout the correction steps. Although the final Dice score and variability are not as optimal as in the CT case, reflecting the lower baseline performance of the TPUS model, the overall trend remains consistent, highlighting the effectiveness of guided correction even in more challenging scenarios.

In contrast, the random strategy, while also benefiting from the larger patch sizes and showing a significant improvement compared to the previous experiment, remains clearly inferior to the other curves. Additionally, it exhibits a noticeably higher standard deviation throughout the correction process, especially in later correction stages. This increased variability reflects less reliable performance and underlines the importance of guided correction, and the effectiveness of uncertainty as a prioritization signal. While the adaptive patch size might be considered aggressive, given the significant correction impact even for random selection, it was applied uniformly across all strategies, preserving the fairness of the comparison.

Overall, these findings reinforce the utility of voxel-wise uncertainty estimates as a guidance mechanism for interactive segmentation refinement, offering better generalization and stability than unguided approaches.



(a) Results for the CT model.

(b) Results for the TPUS model.

Figure 5.10: Mean Dice score progression as a function of the number of simulated corrections across two segmentation models using an adaptive minimum patch size. (a) Results for the CT model. (b) Results for the TPUS model. Three voxel selection strategies are compared: uncertainty-based, random, and distance-based. At each correction step, all images receive the same number of patch corrections. The shaded regions represent ±1 standard deviation across the dataset.

5.3 Integrated Use of Structure-wise and Voxel-wise Uncertainty Estimates

Building upon the previous experiments, this section explores a hybrid approach that integrates both structure-wise and voxel-wise uncertainty estimates to guide correction strategies more effectively. While the earlier structure-wise analysis used per-image uncertainty scores to rank and select the most uncertain segmentations for review, each selected image was treated as a whole, implicitly assuming that the entire structure required correction, without any spatial guidance regarding which regions were actually problematic. In contrast, the voxel-wise correction framework introduced spatial specificity by identifying and correcting small patches around the most uncertain voxels. However, it lacked image-level prioritization, as all images in the dataset were treated equally, receiving the same number of patch corrections at each step, regardless of their overall segmentation quality or need for correction.

The first experiment within this combined framework followed a sequential, uncertainty-driven correction strategy. The process was as follows:

1. All images in the dataset were ranked based on their structure-wise uncertainty scores. This ranking provided a global prioritization of the most uncertain segmentations.

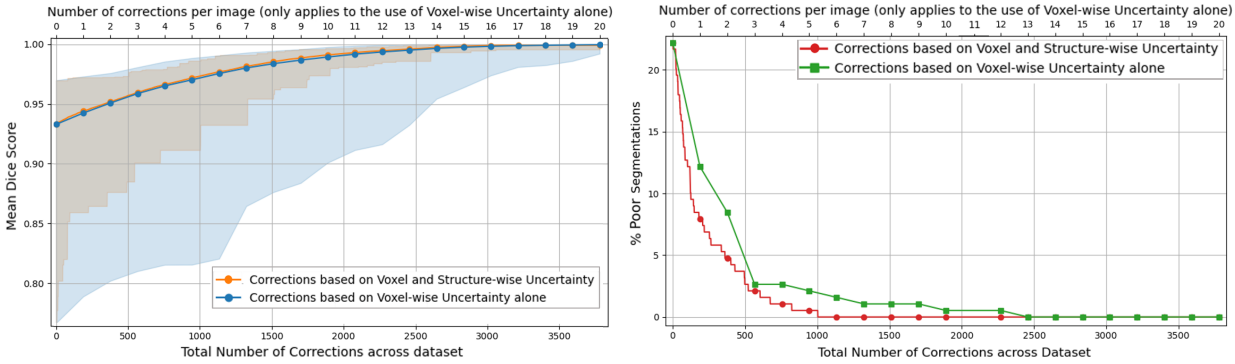
2. At each iteration, the image with the highest current structure-wise uncertainty was selected.

3. Within the selected image, the most uncertain voxel was identified based on the voxel-wise uncertainty map. A patch correction was then simulated, centered on this voxel. As in the previous experiment, the patch size was set to the maximum of twice the radial distance

from the predicted boundary to the ground truth boundary at that voxel, and the estimated radius of the segmented structure.

4. After inserting the patch:
    - The entropy values within the corrected patch were set to zero, assuming those voxels were now confidently correct.
    - The corresponding portion of the predicted segmentation within the patch was excluded from the volume normalization term.
    - The updated structure-wise uncertainty was recomputed by summing the remaining entropy and normalizing by the remaining volume. This dynamic updating step aimed to reflect the reduction in uncertainty caused by each correction
5. This updated uncertainty enabled re-prioritization of images. The process then proceeded iteratively: at each step, the image with the highest structure-wise uncertainty was selected, whether an uncorrected image or one already partially corrected, and a new patch was applied to its most uncertain region.

In Figure 5.11 and Figure 5.12 the proposed hybrid strategy, which combines structure-wise and voxel-wise uncertainty, is compared with the previous method that relies solely on voxel-wise uncertainty for patch corrections, applied to the CT and TPUS datasets, respectively.



(a) (b)

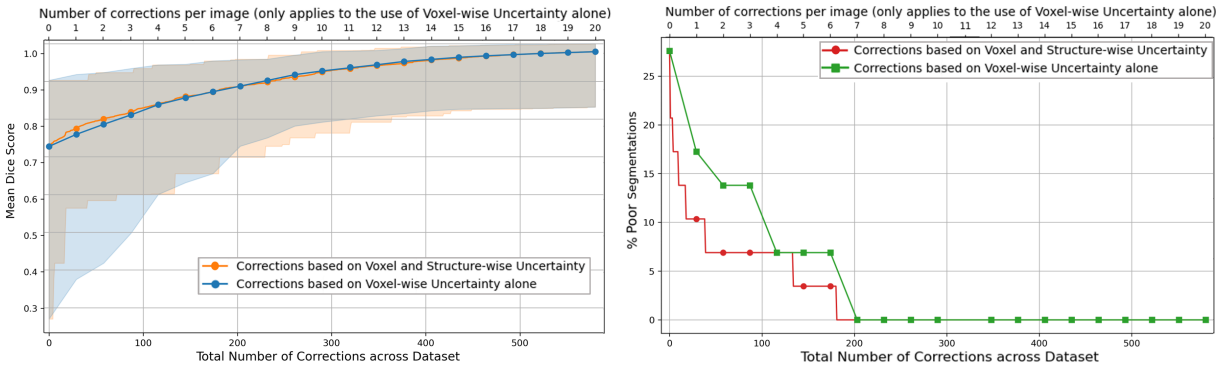
Figure 5.11: Comparison for the CT dataset between correction strategies using voxel-wise uncertainty alone and a hybrid strategy combining voxel- and structure-wise uncertainty. The bottom x-axis shows the total number of cumulative corrections applied. The top x-axis maps this total to the number of corrections per image in the voxel-wise-only strategy (since all images are corrected at each step in that method). This dual-axis design enables a fair comparison by aligning both curves according to the same total correction effort. (a) Evolution of mean Dice score (solid lines) as a function of the total number of patch corrections applied across the dataset. The shaded areas represent the full range (min to max) of Dice scores across all images. (b) Evolution of the percentage of poor segmentations (i.e., with Dice bellow the inter-observer Dice score) as a function of the total number of patch corrections applied across the dataset.

The key distinction between the two strategies lies in how images are selected for correction - in the voxel-wise-only approach, every image in the dataset receives a patch correction at each

iteration, since there is no prioritization across images. In contrast, the hybrid approach selects and corrects only one image per iteration, based on the current structure-wise uncertainty ranking. As a result, the first step in the voxel-wise-only strategy applies as many corrections as there are images in the dataset, while the hybrid method applies just one correction per step. This alignment is critical for fair comparison between the two strategies.

From Figure 5.11a, it is possible to observe that for the CT dataset both methods achieve similar trends in mean Dice score, with the hybrid approach performing slightly better overall. However, a notable difference is seen in the minimum Dice score - with the hybrid method, the minimum Dice increases much more rapidly in the early stages, indicating that the worst-performing segmentations are being corrected earlier. In contrast, the voxel-wise-only approach requires significantly more corrections to reach the same improvement in the minimum Dice score.

For the TPUS dataset, as shown in Figure 5.12a, the hybrid approach clearly outperforms the voxel-wise-only strategy during the initial phase (up to around 100 corrections). In this early stage, both the mean and minimum Dice scores improve more rapidly, indicating effective prioritization of problematic segmentations. However, as the number of corrections increases, the advantage of the hybrid approach diminishes and eventually both strategies converge to similar performance levels. One possible reason for this behavior is that the CT dataset has many more images (29 in TPUS vs. 189 in CT) and receives more total corrections, so the impact of prioritization lasts longer and has a bigger effect overall.



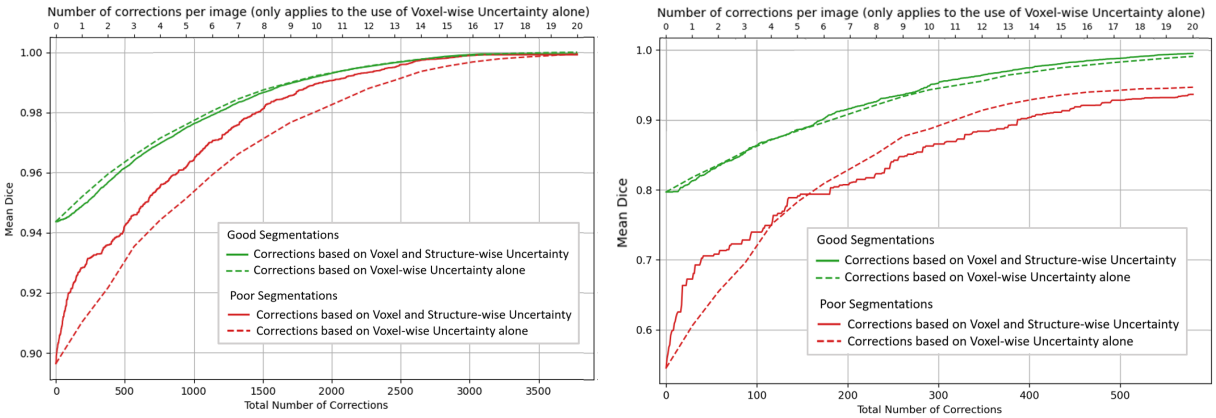
(a) (b)

Figure 5.12: Comparison for the TPUS dataset between correction strategies using voxel-wise uncertainty alone and a hybrid strategy combining voxel- and structure-wise uncertainty. The bottom x-axis shows the total number of cumulative corrections applied. The top x-axis maps this total to the number of corrections per image in the voxel-wise-only strategy (since all images are corrected at each step in that method). This dual-axis design enables a fair comparison by aligning both curves according to the same total correction effort. (a) Evolution of mean Dice score (solid lines) as a function of the total number of patch corrections applied across the dataset. The shaded areas represent the full range (min to max) of Dice scores across all images. (b) Evolution of the percentage of poor segmentations (i.e., with Dice bellow the inter-observer Dice score) as a function of the total number of patch corrections applied across the dataset.

This advantage is further reflected in the percentage of poor segmentations. Figure 5.11b shows that, at every correction step for the CT dataset, the hybrid approach results in a lower

percentage of poor segmentations compared to the voxel-wise-only method. This difference is particularly pronounced in the early stages, where the hybrid strategy rapidly eliminates many of the worst-performing cases. Notably, the hybrid approach brings the percentage of poor segmentations down to 0% after only 1000 corrections, while the voxel-wise-only method requires more than twice as many corrections (2457) to achieve the same outcome. In Figure 5.12b, an improvement achieved by the hybrid method on the TPUS dataset is also evident, which is particularly pronounced during the early stages, as was already expected from the previous analysis.

To further understand how these improvements are distributed across the segmentations, Figure 5.13 breaks down the evolution of the mean Dice score separately for images that were initially poor and good, for both strategies and datasets. As before, this classification is based on a threshold derived from the inter-observer Dice score.



(a) Results for the CT model.

(b) Results for the TPUS model.

Figure 5.13: Evolution of the mean Dice score for images that were initially poor (below inter-observer Dice, shown in red) and good (above inter-observer Dice, shown in green), under the two correction strategies. Solid lines represent the proposed hybrid approach, while dashed lines correspond to the voxel-wise-only baseline.

For the CT dataset (Figure 5.13a), the hybrid approach leads to a markedly higher improvement in the Dice score of poor segmentations, particularly in the early correction steps. This is reflected in the steep initial rise of the curve, clearly outperforming the voxel-wise-only strategy. In contrast, for the initially good segmentations, the hybrid strategy shows a slower and more modest improvement compared to the voxel-wise-only method in the early stages. This behavior aligns with the strategy's prioritization mechanism - the hybrid method focuses first on the worst-performing cases, resulting in rapid gains for poor segmentations while initially applying fewer corrections to already well-performing ones.

For the TPUS dataset (Figure 5.13b), a similar trend is observed initially. The hybrid method shows a clear advantage in improving poor segmentations during the early correction stages. However, after approximately 150 corrections, the performance on poor segmentations plateaus and eventually falls behind that of the voxel-wise-only approach. At the same point, the hybrid method

begins to outperform voxel-wise-only in the group of good segmentations. This turning point suggests that the most uncertain and poorly performing cases have already been addressed, and the hybrid method begins allocating corrections to images that were initially well-segmented but still contained residual uncertainty. As a result, the advantage in correcting poor segmentations diminishes, while the benefit shifts toward refining good ones. In contrast, the voxel-wise-only strategy, by applying corrections uniformly across all images, continues to improve both groups more steadily over time.

Chapter 6

Interactive Model Results

This chapter presents the implementation and evaluation of the BISNet model, previously introduced in Section 3.3. BISNet was designed to be applied independently to both datasets under study, with separate training for each. However, it was only successfully implemented and evaluated for the TPUS dataset. The challenges that prevented its application to the CT dataset are discussed later in this chapter.

For the TPUS dataset, this chapter presents the model's performance in automatic mode and describes the tuning process for uncertainty estimation, addressing both structure-wise and voxel-wise outputs.

6.1 BISNet Application on the TPUS Dataset

6.1.1 Segmentation Performance

Table 6.1 presents the mean Dice score and ASSD of the automatic segmentation produced by BISNet, as well as the corresponding ΔGD value obtained through simulated interaction, using the same simulation strategy employed during training.

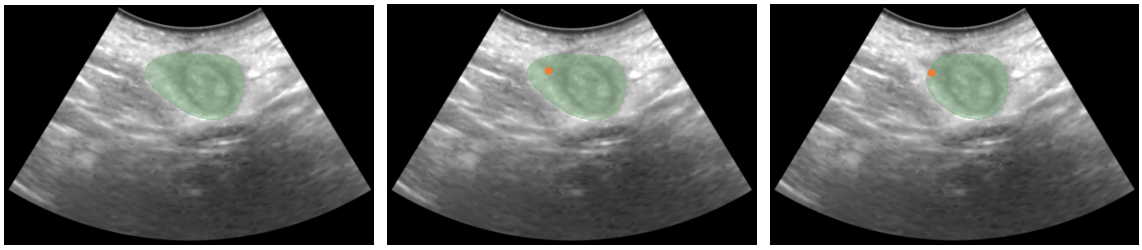
For comparison, the table also includes the Dice and ASSD metrics obtained from the baseline model.

Table 6.1: Automatic segmentation performance of BISNet compared to the baseline model on the TPUS dataset.

Model	**Dice Score (%)**	**ASSD (mm)**	**Δ GD (%)**
Baseline	72.85 ± 13.86	2.28 ± 1.05	—
BISNet	70.81 ± 12.53	2.47 ± 1.06	77.63 ± 3.24

The results show a slight decrease in Dice score and an increase in ASSD for the BISNet model compared to the baseline, indicating a reduction in the quality of the initial automatic segmentation. However, this drop in performance is within the range reported in related work involving BISNet, where the model prioritizes a balance between interactive responsiveness and performance over fully optimized initial predictions [112]. Importantly, the value of ΔGD demonstrates

responsiveness to boundary clicks, suggesting that BISNet effectively corrects segmentation errors when user guidance is provided. This supports its suitability for interactive medical image segmentation scenarios where refinement based on expert input is expected. The interactive segmentation process performed by BISNet is illustrated in Figure 6.1.



(a) Initial automatic segmentation (b) User-provided boundary clicks (c) Refined segmentation

Figure 6.1: Exemplification of the BISNet interactive segmentation process on a single case. The model first generates an automatic segmentation (a), which is then refined (c) based on user-provided boundary click (b).

6.1.2 Tuning of Uncertainty Estimation Parameters

As with the baseline model, a tuning process was conducted for the interactive model to identify an appropriate configuration of dropout rate and number of MC samples for uncertainty estimation, analysing their influence on both segmentation performance (Figure 6.2) and the correlation between uncertainty and segmentation quality (Figure 6.3).

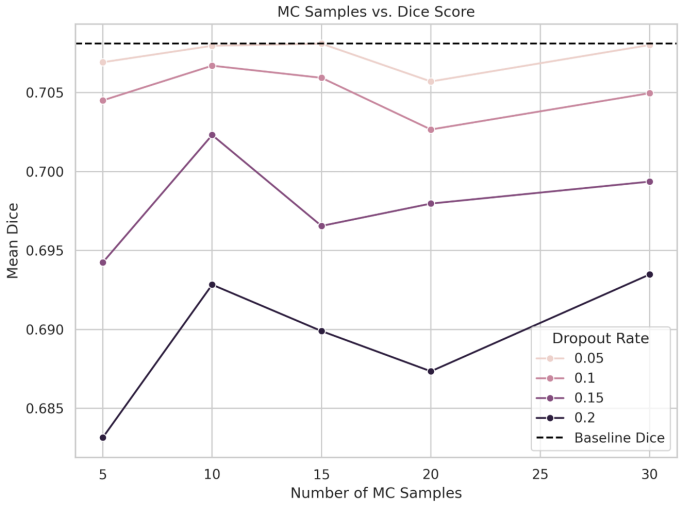


Figure 6.2: Variation in Dice score across different MC dropout rates and numbers of MC samples using the BISNet on TPUS data.

Among the tested configurations, a dropout rate of 0.1 and 15 MC samples was selected. While this setting results in a slight decrease in mean Dice score compared to the baseline model, the reduction is considered acceptable in light of the improved uncertainty estimation. Specifically,

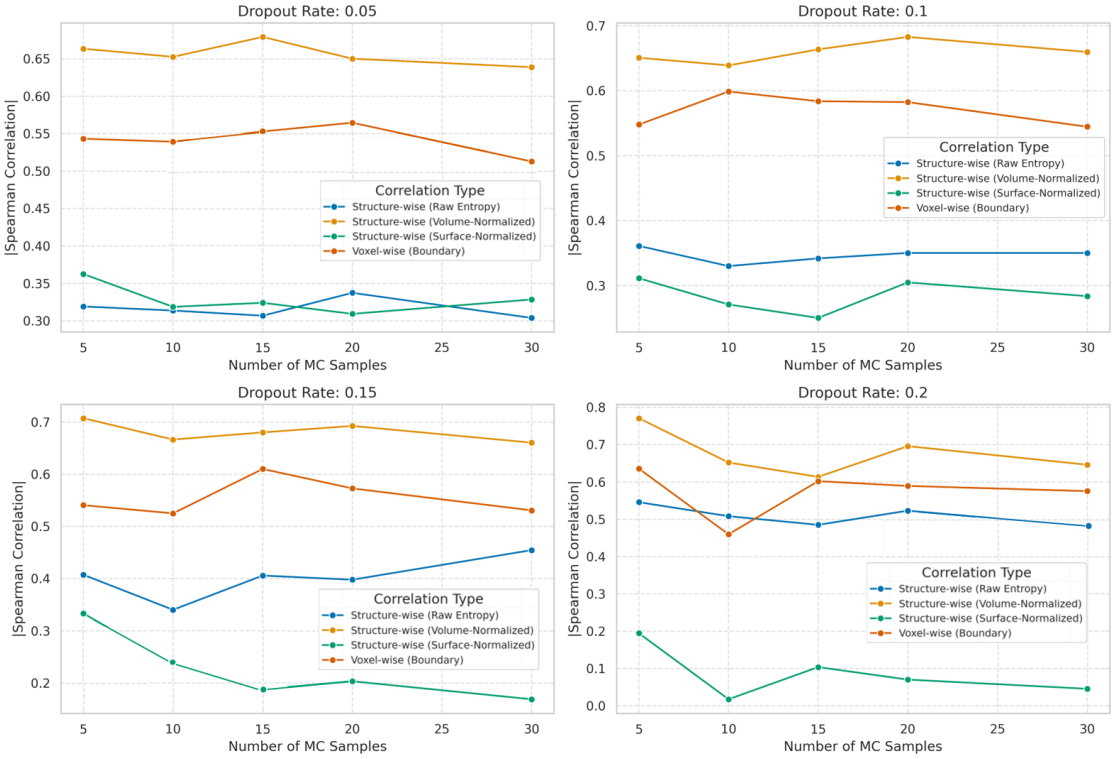


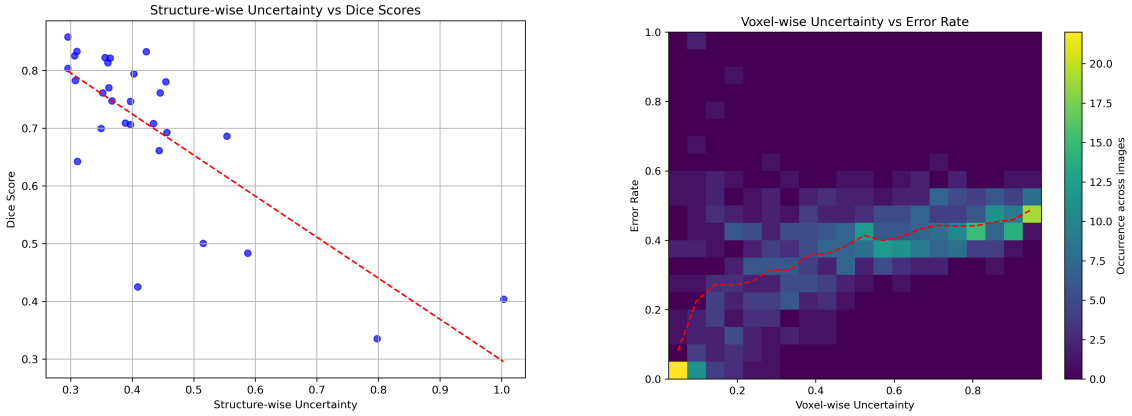
Figure 6.3: Correlation between uncertainty and segmentation error across different MC dropout rates and numbers of MC samples using the BISNet in TPUS data. Each curve shows how the Spearman correlation varies with uncertainty estimation settings, capturing the relationship between predicted uncertainty and actual segmentation error at the structure and voxel levels.

this configuration yields consistent Spearman correlations between entropy-based uncertainty estimates and segmentation performance, most notably for voxel-wise boundary and structure-wise volume-normalized measures. Consistent with findings from the baseline model tuning, raw entropy and surface area-normalized metrics yielded weak and unstable correlations with segmentation quality, highlighting volume-normalized entropy as a more robust and informative global uncertainty metric. Although using 20 MC samples with the same dropout rate of 0.1 produced stronger correlations, the associated drop in Dice score was more pronounced.

Compared to a lower dropout rate (e.g., 0.05), which exhibits slightly higher Dice scores but weaker uncertainty correlations, and higher dropout rates (e.g., 0.15 and 0.2), which result in decreased segmentation performance and less stable uncertainty estimates, the chosen configuration offers a favorable balance between predictive accuracy, uncertainty informativeness, and computational cost.

Figure 6.4 provides a visual summary of the outcomes obtained for a dropout rate of 0.1 and 15 MC samples. Figure 6.4a depicts the structure-wise relationship between segmentation accuracy and predictive uncertainty, while Figure 6.4b illustrates the alignment between uncertainty and segmentation error at the voxel level. Collectively, these figures demonstrate the extent to which uncertainty estimates capture segmentation quality under the specified experimental conditions.

The observed patterns closely mirror those of the baseline model, indicating that the overall behavior of uncertainty remains consistent within this framework. Specifically, the structure-wise analysis reveals a negative correlation, whereby increased uncertainty is associated with reduced Dice scores. At the voxel level, although the histogram some noise, the mean trend nonetheless displays an overall increasing trajectory. This suggests that voxel-wise uncertainty retains a degree of alignment with local segmentation errors.



(a) (b)

Figure 6.4: Uncertainty–performance relationships for the BISNet using the selected MC dropout configuration (dropout rate 0.1, 15 MC samples). (a) Structure-wise correlation between Dice score and volume-normalized summed entropy. Each point represents a volume and the red dashed line shows a linear regression fit. (b) Voxel-wise uncertainty–error profile computed over boundary voxels. Voxels are grouped into uncertainty bins (x-axis), with corresponding misclassification rates (y-axis). The red dashed line indicates the mean misclassification rate across bins. The color scale reflects the occurrence across images, indicating how many times each uncertainty–error pairing appears throughout the dataset.

6.2 BISNet Application on the CT Dataset

The BISNet model was also explored for application to the CT dataset.

It is important to note that BISNet was initially developed for the TPUS application. Although it was expected that BISNet could be adapted to different imaging modalities and segmentation tasks, its application to the CT dataset was not successful. Specifically, the model failed to respond effectively to interactive guidance during training and inference.

This lack of responsiveness may be mainly due to the loss function dynamics. The model optimizes a combined loss consisting of a global Dice and Cross-Entropy term and a Weighted Dice and Cross-entropy term focused around boundary clicks. The global loss covers the entire image and typically dominates the training signal because the CT baseline model already achieves a very high initial segmentation accuracy (mean Dice score of 93.54%). Consequently, the weighted loss component, which only influences a small region near the clicks, produces comparatively small

gradients and fails to significantly influence model updates. As a result, the model tends to rely heavily on the initial automatic segmentation and ignores the guidance from user clicks.

Several adjustments were attempted to address this issue:

- Increasing the weight of the guidance-specific loss component to 1.5, in order to strengthen the influence of user-provided guidance during learning.
- Disabling dropout on the guidance signal to ensure that the guidance input was always applied during training.
- Increasing the λ parameter in Equation 3.6. A higher λ places more emphasis on sampling points from areas with higher segmentation error. The objective was to increase the likelihood that guidance points fall near problematic regions.

Despite these efforts, no configuration was able to achieve a ΔGD greater than 1% on the validation set, indicating the model remained largely unresponsive to interactive clicks.

This lack of responsiveness can be mainly explained by the reasons outlined above. This outcome suggests that further refinement may be needed to better balance the learning signal in cases where the model generalizes well or converges quickly. As currently designed, BISNet appears to be more suitable for use cases where multiple rounds of interaction are required to achieve a satisfactory initial segmentation—allowing the model to become more aware of both the segmentation uncertainty and the corrective input provided through user interaction. Therefore, the application of BISNet to the CT dataset was not possible within the scope of this work.

Chapter 7

Integration of Uncertainty into an Interactive Segmentation Framework

This chapter explores the integration of uncertainty estimation into an interactive segmentation framework, extending the previous analyses performed on automatic, non-interactive models.

For the TPUS dataset, experiments were conducted using the BISNet. A clinical experiment was performed to assess the behavior of the model during user-guided corrections, and the impact of integrating structure-wise uncertainty into the interactive process was also evaluated.

Regarding the CT dataset, BISNet could not be successfully implemented. As a result, this dataset is not further addressed in the present chapter.

7.1 Clinical Experiment

A clinical experiment was conducted on the TPUS dataset in collaboration with a clinician. The BISNet model was integrated into the 3D Slicer software, enabling interactive segmentation through manual input in the form of boundary clicks. This setup allowed the clinician to test the network in a realistic setting by iteratively reviewing segmentations and providing corrections.

A total of 20 volumes were used in this experiment. For each case, the initial automatic segmentation produced by the model was presented to the user, who was able to navigate through the image in axial, sagittal, and coronal planes and add any number of corrective clicks. This setup enabled the assessment of the practical functionality of the interactive framework and its responsiveness to user input.

In addition to evaluating the interactive framework itself, the experiment was extended to incorporate structure-wise uncertainty. After the clinician completed the corrections, both the initial and final Dice scores were recorded for each volume. These values were then used to simulate different correction orderings across the dataset - by ascending initial Dice score (worst-performing segmentations corrected first), by descending global uncertainty (prioritizing the most uncertain cases), and at random. In these simulations, a correction was modeled by replacing the initial Dice with the improved Dice value obtained from the clinical session. The evolution of

the overall segmentation performance across the dataset was then analyzed under each correction strategy. The goal of this analysis was to assess whether prioritizing corrections based on initial uncertainty could lead to more efficient improvements in overall segmentation quality.

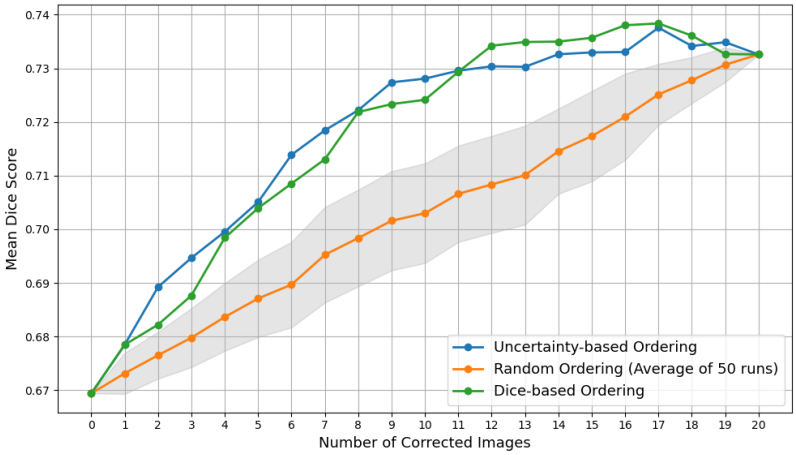


Figure 7.1: Evolution of the mean Dice score across different correction strategies as a function of the number of updates applied. The curves compare uncertainty-based ordering (blue), Dice-based ordering (green), and random ordering averaged over 50 runs (orange). The shaded region represents ±1 standard deviation for the random baseline.

The results of the clinical experiment are shown in Figure 7.1 and indicate that an informed, prioritized ordering of clinician corrections substantially accelerates the enhancement of segmentation quality compared to a random correction sequence. Notably, if the review process were to be halted at any intermediate stage, employing uncertainty-based prioritization would yield a higher Dice score at that point than would be achieved through random ordering. As illustrated in the figure, both uncertainty-guided and Dice-based correction strategies significantly outperform the random baseline. Moreover, uncertainty-based ranking demonstrates performance comparable to, and occasionally exceeding, the idealized Dice-based ordering.

These findings underscore the practical value of structure-wise uncertainty as an effective metric for identifying error-prone segmentations in the absence of ground truth annotations. This enables more efficient and targeted allocation of expert correction efforts, thereby optimizing clinical workflows. By the twentieth correction iteration, all strategies converge to a similar final Dice score, confirming that their ultimate segmentation accuracy is equivalent. However, the efficiency gains associated with uncertainty-based ranking—requiring fewer corrective interactions to achieve a given performance threshold—highlight its potential clinical utility in reducing expert burden while maintaining segmentation quality.

Although voxel-wise uncertainty was not utilized to guide user interactions during the clinical experiment, a post hoc analysis was performed to examine its relationship with the corrections provided by the clinician. For each of the 20 volumes, both the initial segmentation (produced automatically by the model) and the final segmentation (after user correction) were available, enabling a direct voxel-level comparison.

The analysis concentrated on the boundary voxels of the initial segmentation. For each of these voxels, the minimum Euclidean distance to the nearest boundary voxel in the corrected segmentation was computed. This distance served as a proxy for the magnitude of correction applied to a given voxel - smaller distances indicated minor or no changes, while larger distances reflected regions where the segmentation underwent more substantial modification.

To investigate whether voxel-wise uncertainty was predictive of correction magnitude, the entropy values at the initial boundary voxels were compared against the computed distances. Among voxels with non-zero displacement, the mean displacement distance was calculated to estimate the average magnitude of correction and serve as a threshold separating voxels into two groups: (i) those with displacement values below the mean, and (ii) those with values above the mean.

A statistical analysis was conducted to compare the entropy values between the two groups using Student's t-test. The test yielded a p-value of 1.23×10^{-299}, indicating a highly significant statistical difference. This result suggests that voxels requiring larger corrections tend to exhibit higher uncertainty. These findings support the potential utility of voxel-wise entropy as an indicator of segmentation error. The distribution of entropy values for the two groups is illustrated in Figure 7.2 using a violin plot, which visually highlights the clear separation between corrected regions with low and high displacement.

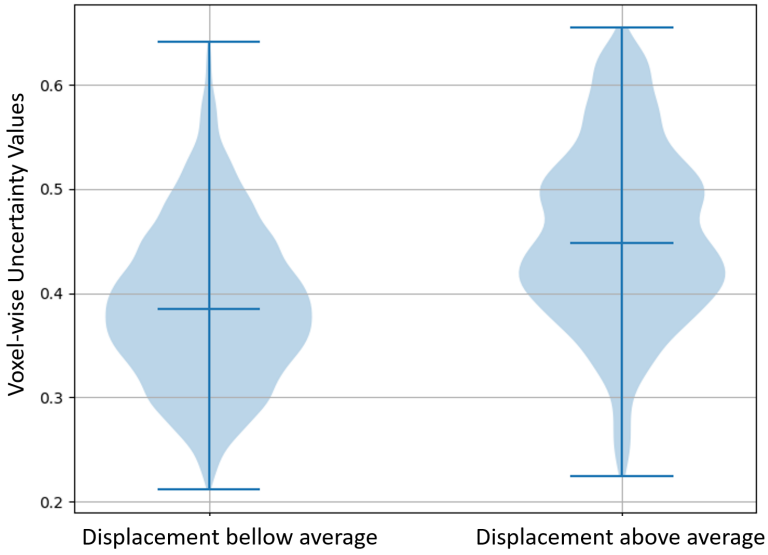


Figure 7.2: Distribution of voxel-wise entropy values for boundary voxels with displacement above and bellow the mean correction distance.

Chapter 8

Conclusion

This dissertation set out to investigate the role of uncertainty in 3D medical image segmentation, with the goal of improving both the efficiency of user interaction and the overall reliability of AI-assisted clinical workflows. By integrating uncertainty estimation into automated and interactive segmentation models, the work aimed to explore whether predictive confidence could be leveraged to guide clinicians during segmentation refinement and prioritization, ultimately supporting the development of safer and more trustworthy AI systems in healthcare. The findings presented in this dissertation strongly support that hypothesis.

8.1 Contributions and Results

The work began with the implementation of baseline deep learning models for two imaging modalities - pericardium segmentation in CT and EAS segmentation in TPUS, which served as reference points for estimating uncertainty. Uncertainty estimation pipelines were then developed using MC Dropout, producing both structure-wise and voxel-wise metrics.

Among the structure-wise uncertainty metrics evaluated, the most effective was the entropy summed across the segmentation mask and normalized by its volume. This approach substantially outperformed the other tested metrics — namely, unnormalized entropy and entropy normalized by surface area — in terms of correlation with segmentation performance, providing a more reliable measure of structure-level uncertainty. For voxel-wise uncertainty, the proposed entropy filtering and smoothing pipeline proved to be effective, enhancing the spatial coherence of uncertainty maps and enabling more precise localization of low-confidence regions within the segmentation. These enhancements facilitated more precise localization of regions requiring user attention and laid the foundation for targeted, uncertainty-guided corrections.

A strong inverse correlation was found between structure-wise uncertainty and segmentation accuracy, measured via Dice score, on both datasets. This confirmed that uncertainty could reliably serve as a proxy for identifying poor-quality segmentations, even in the absence of ground truth. Building on this, the study demonstrated that structure-wise uncertainty could be used to prioritize cases for review. Simulations involving both ideal corrections and realistic inter- and

intra-observer variability showed that uncertainty-based correction ordering significantly accelerated performance improvements when compared to random or uniform selection. In early correction stages, the uncertainty-guided strategy approached the effectiveness of a hypothetical oracle strategy that always reviews the worst segmentations first, an advantage made possible without requiring access to ground truth.

Voxel-wise uncertainty also demonstrated utility in localizing segmentation errors near the boundaries, which was investigated through patch-based simulations. These experiments, which tested both a more restrictive patch selection based on proximity to ground truth error and a more permissive strategy incorporating a minimum contextual region, consistently showed that uncertainty-guided corrections outperformed random selection. In particular, the broader patch approach, designed to better reflect clinical correction behavior (since users tend to address the surrounding context when correcting errors, rather than focusing on a single voxel), resulted in corrections that closely approximated the improvements achieved by an ideal strategy based on ground truth–guided corrections, which is especially important for real-world deployment scenarios where detailed ground truth labels are unavailable. This suggests that integrating a broader spatial context, informed by uncertainty, is more aligned with human correction patterns and ultimately more effective in practice.

Further performance gains were achieved by combining structure-wise and voxel-wise uncertainty within a hybrid correction strategy. In this approach, structure-wise uncertainty was used to prioritize which cases to review, while voxel-wise uncertainty guided the selection of regions within those cases for localized correction. This method proved especially effective for improving the lowest-performing segmentations in the dataset, significantly raising their Dice scores with a relatively small number of user interactions. By enabling a more efficient and focused allocation of correction effort, the combined use of structure- and voxel-level uncertainty represents a promising framework for scalable and clinician-in-the-loop segmentation systems.

Finally, to support real-time interaction, the BISNet was implemented and trained using the TPUS data. While its fully automatic performance was slightly lower than the baseline model, it showed responsiveness to user input, validating its potential for use in human-in-the-loop workflows. The integration of uncertainty into this interactive framework culminated in a clinical experiment in which an expert corrected TPUS segmentations using BISNet. The results demonstrated that uncertainty-based correction ordering performed comparably to Dice-based ordering in early correction stages. This confirmed the value of uncertainty as a stand-in for ground truth in prioritizing user attention, an essential feature for clinical viability. Furthermore, a post hoc analysis revealed that regions most affected by expert corrections tended to align with areas of higher voxel-wise uncertainty, reinforcing the relevance of the uncertainty maps.

Together, these findings show that the goals initially proposed in this dissertation were largely accomplished, as the integration of predictive uncertainty into segmentation workflows has been shown to improve the efficiency of human interaction.

8.2 Limitations and Future Work

While the framework successfully addressed the core research objectives and yielded encouraging results, some limitations remain, offering important directions for future research.

While BISNet performed well on TPUS data, it could not be successfully adapted to the CT dataset. The model exhibited poor responsiveness to user guidance in this context, which may be attributed to the interplay between the training objectives and data characteristics. In particular, the loss formulation appeared to prioritize global segmentation accuracy over localized adjustments, thereby diminishing the influence of user-provided corrections. This limitation suggests a need for alternative training strategies or model configurations that more effectively balance automatic prediction with responsiveness to interactive input. Therefore, future work building on this project should aim to adapt and validate an interactive framework on the CT dataset. In the longer term, extending the framework to support generalization across other imaging modalities and segmentation tasks would represent a significant step toward developing clinically robust and modality-agnostic interactive segmentation systems.

An important direction for future work lies in the integration of voxel-wise uncertainty into the interactive process. Although voxel-wise entropy maps were retrospectively analyzed and found to correlate with regions most frequently corrected by the clinician using BISNet, suggesting that higher uncertainty tended to align with areas requiring refinement, this information was not integrated into the real-time interactive segmentation process. Specifically, voxel-wise uncertainty was not incorporated into the BISNet model to guide user-provided boundary clicks or to inform the correction workflow during live interaction. As such, the practical application of voxel-wise uncertainty remains unexplored within the interactive framework developed in this dissertation. Instead, the potential of voxel-wise uncertainty was evaluated through patch correction simulations conducted on the baseline (non-interactive) model. The patch correction simulations, while informative and yielding promising results, also assumed a simplified model of user behavior. Although care was taken to simulate clinically relevant behavior through adaptive patch sizing and correction logic, these scenarios may not fully reflect the nuanced decisions clinicians make during annotation. Future work should explore how voxel-wise uncertainty can be directly integrated into real-time interaction, potentially by highlighting high-uncertainty regions within the interface or using entropy-driven mechanisms to guide boundary click placement. A prospective user study evaluating the effect of this integration on interaction efficiency, segmentation accuracy, and user workload would be a critical step toward translating these findings into practical, clinician-facing tools.

Another important area for future exploration involves the establishment of reliable, quantitatively defined thresholds for both structure-wise and voxel-wise uncertainty scores. Such thresholds could play a critical role in shaping the behavior of interactive segmentation systems, for example, by serving as triggers for initiating user intervention in cases of low confidence, as well as stopping criteria to indicate when a segmentation is already sufficiently reliable. Even if not used as rigid decision points, such thresholds could at least be presented as recommendations to

the user, providing supportive guidance without overriding clinical judgment. However, the effective deployment of uncertainty thresholds requires careful statistical calibration. This involves not only defining appropriate threshold values, but also validating their robustness across a diverse range of datasets. It will be important to assess the sensitivity and specificity of these thresholds in detecting segmentation failures. Additionally, thresholds should be adaptable to task-specific requirements and clinical tolerances for error, which may vary depending on the context and intended use of the segmentation output.

Finally, broader clinical evaluation is essential. The clinical study in this work involved one expert and a relatively limited dataset. Future studies should involve multiple clinicians with varying levels of experience and should evaluate performance across diverse datasets. In parallel, usability studies are needed to assess how clinicians interpret and act upon uncertainty information, and how such guidance affects cognitive load, confidence, and trust in the system.

8.3 Final Remarks

This dissertation contributes meaningfully to the field of interactive and trustworthy AI for medical imaging. It demonstrates that predictive uncertainty, far from being a limitation, can serve as a powerful tool to guide user attention, inform corrective actions, and ultimately enable more efficient and reliable segmentation workflows.

Rather than pursuing full automation, this work embraces the complexity of clinical environments and supports a collaborative, human-in-the-loop approach. By developing tools that incorporate predictive uncertainty to assist, rather than replace, clinical expertise, it presents a sustainable and realistic pathway for integrating AI into routine practice.

In closing, this work shows that uncertainty is not simply a marker of doubt, but a rich and actionable signal. When accurately estimated and effectively presented, it can inform where to act, when to trust, and how to collaborate with AI systems. This shifts the narrative of automation in medicine from one of replacement to one of partnership, pointing toward systems that are not only intelligent, but also self-aware, adaptive, and aligned with human values and needs.

References

[1] UNESCO. Engineering for sustainable development: Delivering on the sustainable development goals. Technical report, United Nations Educational, Scientific and Cultural Organization, 2021. Accessed: 2025-06-10.

[2] Giorgio Buttazzo. Rise of artificial general intelligence: risks and opportunities. *Frontiers in Artificial Intelligence*, 6, 8 2023.

[3] Chinimilli Venkata Rama Padmaja and Sadasivuni Lakshminarayana. The rise of ai: a comprehensive research review. *IAES International Journal of Artificial Intelligence (IJ-AI)*, 13:2226, 6 2024.

[4] Luís Pinto-Coelho. How artificial intelligence is shaping medical imaging technology: A survey of innovations and applications. *Bioengineering (Basel, Switzerland)*, 10, 12 2023.

[5] Shuroug A. Alowais, Sahar S. Alghamdi, Nada Alsuhebany, Tariq Alqahtani, Abdulrahman I. Alshaya, Sumaya N. Almohareb, Atheer Aldairem, Mohammed Alrashed, Khalid Bin Saleh, Hisham A. Badreldin, Majed S. Al Yami, Shmeylan Al Harbi, and Abdulkareem M. Albekairy. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. *BMC Medical Education*, 23:689, 9 2023.

[6] Yogesh Kumar, Apeksha Koul, Ruchi Singla, and Muhammad Fazal Ijaz. Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda. *Journal of ambient intelligence and humanized computing*, 14:8459–8486, 2023.

[7] Daniel J. Blezek, Lonny Olson-Williams, Andrew Missert, and Pangiotis Korfiatis. Ai integration in the clinical workflow. *Journal of Digital Imaging*, 34:1435–1446, 12 2021.

[8] Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542:115–118, 2 2017.

[9] Jeffrey De Fauw, Joseph R. Ledsam, Bernardino Romera-Paredes, Stanislav Nikolov, Nenad Tomasev, Sam Blackwell, Harry Askham, Xavier Glorot, Brendan O'Donoghue, Daniel Visentin, George van den Driessche, Balaji Lakshminarayanan, Clemens Meyer, Faith Mackinder, Simon Bouton, Kareem Ayoub, Reena Chopra, Dominic King, Alan Karthikesalingam, Cían O. Hughes, Rosalind Raine, Julian Hughes, Dawn A. Sim, Catherine Egan, Adnan Tufail, Hugh Montgomery, Demis Hassabis, Geraint Rees, Trevor Back, Peng T. Khaw, Mustafa Suleyman, Julien Cornebise, Pearse A. Keane, and Olaf Ronneberger. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature Medicine*, 24:1342–1350, 9 2018.

[10] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.

[11] Nicolas Coudray, Paolo Santiago Ocampo, Theodore Sakellaropoulos, Navneet Narula, Matija Snuderl, David Fenyö, Andre L. Moreira, Narges Razavian, and Aristotelis Tsirigos. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nature Medicine*, 24:1559–1567, 10 2018.

[12] Yiming Ding, Jae Ho Sohn, Michael G. Kawczynski, Hari Trivedi, Roy Harnish, Nathaniel W. Jenkins, Dmytro Lituiev, Timothy P. Copeland, Mariam S. Aboian, Carina Mari Aparici, Spencer C. Behr, Robert R. Flavell, Shih-Ying Huang, Kelly A. Zalocusky, Lorenzo Nardo, Youngho Seo, Randall A. Hawkins, Miguel Hernandez Pampaloni, Dexter Hadley, and Benjamin L. Franc. A deep learning model to predict a diagnosis of alzheimer disease by using 18 f-fdg pet of the brain. *Radiology*, 290:456–464, 2 2019.

[13] Hyunkwang Lee, Sehyo Yune, Mohammad Mansouri, Myeongchan Kim, Shahein H. Tajmir, Claude E. Guerrier, Sarah A. Ebert, Stuart R. Pomerantz, Javier M. Romero, Shahmir Kamalian, Ramon G. Gonzalez, Michael H. Lev, and Synho Do. An explainable deep-learning algorithm for the detection of acute intracranial haemorrhage from small datasets. *Nature Biomedical Engineering*, 3:173–182, 12 2018.

[14] Rui-Qi Li, Xiao-Liang Xie, Xiao-Hu Zhou, Shi-Qi Liu, Zhen-Liang Ni, Yan-Jie Zhou, Gui-Bin Bian, and Zeng-Guang Hou. Real-time multi-guidewire endpoint localization in fluoroscopy images. *IEEE Transactions on Medical Imaging*, 40:2002–2014, 8 2021.

[15] Wenjia Bai, Matthew Sinclair, Giacomo Tarroni, Ozan Oktay, Martin Rajchl, Ghislain Vaillant, Aaron M. Lee, Nay Aung, Elena Lukaschuk, Mihir M. Sanghvi, Filip Zemrak, Kenneth Fung, Jose Miguel Paiva, Valentina Carapella, Young Jin Kim, Hideaki Suzuki, Bernhard Kainz, Paul M. Matthews, Steffen E. Petersen, Stefan K. Piechnik, Stefan Neubauer, Ben Glocker, and Daniel Rueckert. Automated cardiovascular magnetic resonance image analysis with fully convolutional networks. *Journal of Cardiovascular Magnetic Resonance*, 20:65, 2 2018.

[16] Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, Gerard Sanroma, Sandy Napel, Steffen Petersen, Georgios Tziritas, Elias Grinias, Mahendra Khened, Varghese Alex Kollerathu, Ganapathy Krishnamurthi, Marc-Michel Rohé, Xavier Pennec, Maxime Sermesant, Fabian Isensee, Paul Jäger, Klaus H. Maier-Hein, Peter M. Full, Ivo Wolf, Sandy Engelhardt, Christian F. Baumgartner, Lisa M. Koch, Jelmer M. Wolterink, Ivana Išgum, Yeonggul Jang, Yoonmi Hong, Jay Patravali, Shubham Jain, Olivier Humbert, and Pierre-Marc Jodoin. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: Is the problem solved? *IEEE Transactions on Medical Imaging*, 37:2514–2525, 11 2018.

[17] Philipp Kickingereder, Fabian Isensee, Irada Tursunova, Jens Petersen, Ulf Neuberger, David Bonekamp, Gianluca Brugnara, Marianne Schell, Tobias Kessler, Martha Foltyn, Inga Harting, Felix Sahm, Marcel Prager, Martha Nowosielski, Antje Wick, Marco Nolden, Alexander Radbruch, Jürgen Debus, Heinz-Peter Schlemmer, Sabine Heiland, Michael Platten, Andreas von Deimling, Martin J van den Bent, Thierry Gorlia, Wolfgang Wick,

Martin Bendszus, and Klaus H Maier-Hein. Automated quantitative tumour response assessment of mri in neuro-oncology with artificial neural networks: a multicentre, retrospective study. *The Lancet Oncology*, 20:728–740, 5 2019.

[18] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[19] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[20] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

[21] Ran Gu, Guotai Wang, Tao Song, Rui Huang, Michael Aertsen, Jan Deprest, Sébastien Ourselin, Tom Vercauteren, and Shaoting Zhang. Ca-net: Comprehensive attention convolutional neural networks for explainable medical image segmentation. *IEEE transactions on medical imaging*, 40(2):699–711, 2020.

[22] Jiacheng Ruan, Jincheng Li, and Suncheng Xiang. Vm-unet: Vision mamba unet for medical image segmentation. *arXiv preprint arXiv:2402.02491*, 2024.

[23] Robert Challen, Joshua Denny, Martin Pitt, Luke Gompels, Tom Edwards, and Krasimira Tsaneva-Atanasova. Artificial intelligence, bias and clinical safety. *BMJ quality safety*, 28:231–237, 3 2019.

[24] GEORGE MALIHA, SARA GERKE, I. GLENN COHEN, and RAVI B. PARIKH. Artificial intelligence and liability in medicine: Balancing safety and innovation. *The Milbank Quarterly*, 99:629–647, 9 2021.

[25] Qin Liu, Zhenlin Xu, Yining Jiao, and Marc Niethammer. isegformer: interactive segmentation via transformers with application to 3d knee mr images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 464–474. Springer, 2022.

[26] US Food, Drug Administration, et al. Artificial intelligence and machine learning in software as a medical device. *US Food & Drug Administration: Silver Spring, MD, USA*, 2021.

[27] European Union. Regulation (eu) 2017/745 of the european parliament and of the council on medical devices, 2017. Accessed: 2025-04-26.

[28] European Commission. Proposal for a regulation laying down harmonized rules on artificial intelligence (artificial intelligence act), 2021. Accessed: 2025-04-26.

[29] High-Level Expert Group on Artificial Intelligence. Ethics guidelines for trustworthy ai, 2019. Accessed: 2025-04-26.

[30] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9:62–66, 1 1979.

[31] R. Adams and L. Bischof. Seeded region growing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16:641–647, 6 1994.

[32] Michael Kass, Andrew Witkin, and Demetri Terzopoulos. Snakes: Active contour models. *International Journal of Computer Vision*, 1:321–331, 1 1988.

[33] Rafael C. Gonzalez and Richard E. Woods. *Digital Image Processing*. Pearson Education, 4th edition, 2018.

[34] Marin Benčević, Irena Galić, Marija Habijan, and Aleksandra Pižurica. Recent progress in epicardial and pericardial adipose tissue segmentation and quantification based on deep learning: A systematic review. *Applied Sciences*, 12:5217, 5 2022.

[35] Yann LeCun, Bernhard Boser, John Denker, Donnie Henderson, R. Howard, Wayne Hubbard, and Lawrence Jackel. Handwritten digit recognition with a back-propagation network. In D. Touretzky, editor, *Advances in Neural Information Processing Systems*, volume 2. Morgan-Kaufmann, 1989.

[36] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[37] Reymond Mesuga and Brian James Bayanay. A deep transfer learning approach on identifying glitch wave-form in gravitational wave data. *arXiv preprint arXiv:2107.01863*, 2021.

[38] Wei Liu, Andrew Rabinovich, and Alexander C Berg. Parsenet: Looking wider to see better. *arXiv preprint arXiv:1506.04579*, 2015.

[39] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention*, pages 424–432. Springer, 2016.

[40] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016.

[41] Fabian Isensee, Jens Petersen, Andre Klein, David Zimmerer, Paul F Jaeger, Simon Kohl, Jakob Wasserthal, Gregor Koehler, Tobias Norajitra, Sebastian Wirkert, et al. nnu-net: Self-adapting framework for u-net-based medical image segmentation. *arXiv preprint arXiv:1809.10486*, 2018.

[42] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *International workshop on deep learning in medical image analysis*, pages 3–11. Springer, 2018.

[43] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

[44] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023.

[45] Chaofan Ma, Qisen Xu, Xiangfeng Wang, Bo Jin, Xiaoyun Zhang, Yanfeng Wang, and Ya Zhang. Boundary-aware supervoxel-level iteratively refined interactive 3d image segmentation with multi-agent reinforcement learning. *IEEE Transactions on Medical Imaging*, 40(10):2563–2574, 2020.

[46] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 5580–5590, Red Hook, NY, USA, 2017. Curran Associates Inc.

[47] Ke Zou, Zhihao Chen, Xuedong Yuan, Xiaojing Shen, Meng Wang, and Huazhu Fu. A review of uncertainty estimation and its application in medical imaging. *Meta-Radiology*, 1(1):100003, 2023.

[48] Chu-I Yang and Yi-Pei Li. Explainable uncertainty quantifications for deep learning-based molecular property prediction. *Journal of Cheminformatics*, 15:13, 2 2023.

[49] Benjamin Lambert, Florence Forbes, Senan Doyle, Harmonie Dehaene, and Michel Dojat. Trustworthy clinical ai solutions: A unified review of uncertainty quantification in deep learning models for medical image analysis. *Artificial Intelligence in Medicine*, 150:102830, 4 2024.

[50] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.

[51] Balamurali Murugesan, Bingyuan Liu, Adrian Galdran, Ismail Ben Ayed, and Jose Dolz. Calibrating segmentation networks with margin-based label smoothing. *Medical Image Analysis*, 87:102826, 2023.

[52] Terrance Devries and Graham W. Taylor. Leveraging uncertainty estimates for predicting segmentation quality. *ArXiv*, abs/1807.00502, 2018.

[53] Alain Jungo, Fabian Balsiger, and Mauricio Reyes. Analyzing the quality and challenges of uncertainty estimations for brain tumor segmentation. *Frontiers in Neuroscience*, 14, 4 2020.

[54] Kumar Shridhar, Felix Laumann, and Marcus Liwicki. A comprehensive guide to bayesian convolutional neural network with variational inference. *arXiv preprint arXiv:1901.02731*, 2019.

[55] Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

[56] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1613–1622, Lille, France, 07–09 Jul 2015. PMLR.

[57] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1050–1059, New York, New York, USA, 20–22 Jun 2016. PMLR.

[58] Jose Ignacio Orlando, Philipp Seebock, Hrvoje Bogunovic, Sophie Klimscha, Christoph Grechenig, Sebastian Waldstein, Bianca S. Gerendas, and Ursula Schmidt-Erfurth. U2-net: A bayesian u-net model with epistemic uncertainty feedback for photoreceptor layer segmentation in pathological oct scans. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 1441–1445. IEEE, 4 2019.

[59] Yarin Gal, Jiri Hron, and Alex Kendall. Concrete dropout. *Advances in neural information processing systems*, 30, 2017.

[60] Li Wan, Matthew Zeiler, Sixin Zhang, Yann LeCun, and Rob Fergus. Regularization of neural networks using dropconnect. In *30th International Conference on Machine Learning, ICML 2013*, pages 2095–2103. International Machine Learning Society (IMLS), 2013. 30th International Conference on Machine Learning, ICML 2013 ; Conference date: 16-06-2013 Through 21-06-2013.

[61] Mattias Teye, Hossein Azizpour, and Kevin Smith. Bayesian uncertainty estimation for batch normalized deep networks. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4907–4916. PMLR, 10–15 Jul 2018.

[62] Robin Camarasa, Daniel Bos, Jeroen Hendrikse, Paul Nederkoorn, Eline Kooi, Aad Van Der Lugt, and Marleen De Bruijne. Quantitative comparison of monte-carlo dropout uncertainty measures for multi-class segmentation. In *International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging*, pages 32–41. Springer, 2020.

[63] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

[64] Alireza Mehrtash, William M Wells, Clare M Tempany, Purang Abolmaesumi, and Tina Kapur. Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. *IEEE transactions on medical imaging*, 39(12):3868–3878, 2020.

[65] Matthew Ng, Fumin Guo, Labonny Biswas, Steffen E Petersen, Stefan K Piechnik, Stefan Neubauer, and Graham Wright. Estimating uncertainty in neural networks for cardiac mri segmentation: a benchmark study. *IEEE Transactions on Biomedical Engineering*, 70(6):1955–1966, 2022.

[66] Moritz Fuchs, Camila González, and Anirban Mukhopadhyay. Practical uncertainty quantification for brain tumor segmentation. In Ender Konukoglu, Bjoern Menze, Archana Venkataraman, Christian Baumgartner, Qi Dou, and Shadi Albarqouni, editors, *Proceedings of The 5th International Conference on Medical Imaging with Deep Learning*, volume 172 of *Proceedings of Machine Learning Research*, pages 407–422. PMLR, 06–08 Jul 2022.

[67] Moloud Abdar, Soorena Salari, Sina Qahremani, Hak-Keung Lam, Fakhri Karray, Sadiq Hussain, Abbas Khosravi, U. Rajendra Acharya, Vladimir Makarenkov, and Saeid Nahavandi. Uncertaintyfusenet: Robust uncertainty-aware hierarchical feature fusion model with ensemble monte carlo dropout for covid-19 detection. *Information Fusion*, 90:364–381, 2 2023.

[68] Angelos Filos, Sebastian Farquhar, Aidan N Gomez, Tim GJ Rudner, Zachary Kenton, Lewis Smith, Milad Alizadeh, Arnoud De Kroon, and Yarin Gal. A systematic comparison of bayesian deep learning robustness in diabetic retinopathy tasks. *arXiv preprint arXiv:1912.10481*, 2019.

[69] Guotai Wang, Wenqi Li, Michael Aertsen, Jan Deprest, Sébastien Ourselin, and Tom Vercauteren. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing*, 338:34–45, 4 2019.

[70] Murat Seckin Ayhan and Philipp Berens. Test-time data augmentation for estimation of heteroscedastic aleatoric uncertainty in deep neural networks. In *Medical Imaging with Deep Learning*, 2018.

[71] Yanwu Yang, Xutao Guo, Yiwei Pan, Pengcheng Shi, Haiyan Lv, and Ting Ma. Uncertainty quantification in medical image segmentation with multi-decoder u-net. In *International MICCAI brainlesion workshop*, pages 570–577. Springer, 2021.

[72] Wei Ji, Wenting Chen, Shuang Yu, Kai Ma, Li Cheng, Linlin Shen, and Yefeng Zheng. Uncertainty quantification for medical image segmentation using dynamic label factor allocation among multiple raters. In *MICCAI on QUBIQ workshop*, volume 2, 2020.

[73] Sabri Can Cetindag, Mert Yergin, Deniz Alis, and Ilkay Oksuz. Meta-learning for medical image segmentation uncertainty quantification. In *International MICCAI brainlesion workshop*, pages 578–584. Springer, 2021.

[74] Shi Hu, Daniel Worrall, Stefan Knegt, Bas Veeling, Henkjan Huisman, and Max Welling. Supervised uncertainty quantification for segmentation with multiple annotations. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 137–145. Springer, 2019.

[75] Tanya Nair, Doina Precup, Douglas L Arnold, and Tal Arbel. Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation. *Medical image analysis*, 59:101557, 2020.

[76] Thierry Judge, Olivier Bernard, Woo-Jin Cho Kim, Alberto Gomez, Agisilaos Chartsias, and Pierre-Marc Jodoin. Asymmetric contour uncertainty estimation for medical image segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 210–220. Springer, 2023.

[77] Richard McKinley, Micheal Rebsamen, Katrin Dätwyler, Raphael Meier, Piotr Radojewski, and Roland Wiest. Uncertainty-driven refinement of tumor-core segmentation using 3d-to-2d networks with label uncertainty. In *International MICCAI brainlesion workshop*, pages 401–411. Springer, 2020.

[78] Abhijit Guha Roy, Sailesh Conjeti, Nassir Navab, and Christian Wachinger. Bayesian quicknat: Model uncertainty in deep whole-brain segmentation for structure-wise quality control. *NeuroImage*, 195:11–22, 7 2019.

[79] Yuta Hiasa, Yoshito Otake, Masaki Takao, Takeshi Ogawa, Nobuhiko Sugano, and Yoshinobu Sato. Automated muscle segmentation from clinical ct using bayesian u-net for personalized musculoskeletal modeling. *IEEE transactions on medical imaging*, 39(4):1030–1040, 2019.

[80] Evan Hann, Ricardo A Gonzales, Iulia A Popescu, Qiang Zhang, Vanessa M Ferreira, and Stefan K Piechnik. Ensemble of deep convolutional neural networks with monte carlo dropout sampling for automated image segmentation quality control and robust deep learning using small datasets. In *Annual conference on medical image understanding and analysis*, pages 280–293. Springer, 2021.

[81] Audrey Xie, Elhoucine Elfatimi, Sambuddha Ghosal, and Pratik Shah. Deep learning with uncertainty quantification for predicting the segmentation dice coefficient of prostate cancer biopsy images. In *2024 International Conference on Machine Learning and Applications (ICMLA)*, pages 1158–1163. IEEE, 2024.

[82] Zdravko Marinov, Paul F. Jäger, Jan Egger, Jens Kleesiek, and Rainer Stiefelhagen. Deep interactive segmentation of medical images: A systematic review and taxonomy. *arXiv preprint arXiv:2311.13964*, November 2023.

[83] Soumajit Majumder and Angela Yao. Content-aware multi-level guidance for interactive instance segmentation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11594–11603. IEEE, 6 2019.

[84] Tomas Sakinis, Fausto Milletari, Holger Roth, Panagiotis Korfiatis, Petro Kostandy, Kenneth Philbrick, Zeynettin Akkus, Ziyue Xu, Daguang Xu, and Bradley J Erickson. Interactive segmentation of medical images through fully convolutional neural networks. *arXiv preprint arXiv:1903.08205*, 2019.

[85] Long Xu, Shanghong Li, Yongquan Chen, Junkang Chen, Rui Huang, and Feng Wu. Clickattention: Click region similarity guided interactive segmentation. *arXiv preprint arXiv:2408.06021*, 2024.

[86] Ke Zhang and Xiahai Zhuang. Zscribbleseg: Zen and the art of scribble supervised medical image segmentation. *arXiv preprint arXiv:2301.04882*, 2023.

[87] Gustav Bredell, Christine Tanner, and Ender Konukoglu. Iterative interaction training for segmentation editing networks. In *International workshop on machine learning in medical imaging*, pages 363–370. Springer, 2018.

[88] Helena Williams, João Pedrosa, Laura Cattani, Susanne Housmans, Tom Vercauteren, Jan Deprest, and Jan D'hooge. Interactive segmentation via deep learning and b-spline explicit active surfaces. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 315–325. Springer, 2021.

[89] D. Barbosa, T. Dietenbeck, J. Schaerer, J. D'hooge, D. Friboulet, and O. Bernard. B-spline explicit active surfaces: An efficient framework for real-time 3-d region-based segmentation. *IEEE Transactions on Image Processing*, 21:241–251, 1 2012.

[90] Guilherme Aresta, Colin Jacobs, Teresa Araújo, António Cunha, Isabel Ramos, Bram van Ginneken, and Aurélio Campilho. iw-net: an automatic and minimalistic interactive lung nodule segmentation deep network. *Scientific Reports*, 9:11591, 8 2019.

[91] Rajshree Daulatabad, Roberto Vega, Jacob L. Jaremko, Jeevesh Kapur, Abhilash R. Hareendranathan, and Kumaradeven Punithakumar. Integrating user-input into deep convolutional neural networks for thyroid nodule segmentation. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, pages 2637–2640. IEEE, 11 2021.

[92] Rodrigo Benenson, Stefan Popov, and Vittorio Ferrari. Large-scale interactive object segmentation with human annotators. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[93] Holger R. Roth, Dong Yang, Ziyue Xu, Xiaosong Wang, and Daguang Xu. Going to extremes: Weakly supervised medical image segmentation. *Machine Learning and Knowledge Extraction*, 3:507–524, 6 2021.

[94] Andres Diaz-Pinto, Pritesh Mehta, Sachidanand Alle, Muhammad Asad, Richard Brown, Vishwesh Nath, Alvin Ihsani, Michela Antonelli, Daniel Palkovics, Csaba Pinter, et al. Deepedit: Deep editable learning for interactive segmentation of 3d medical images. In *MICCAI Workshop on Data Augmentation, Labelling, and Imperfections*, pages 11–21. Springer, 2022.

[95] Ragavie Pirabaharan and Naimul Khan. Interactive segmentation using u-net with weight map and dynamic user interactions. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 4754–4757. IEEE, 2022.

[96] Ti Bai, Anjali Balagopal, Michael Dohopolski, Howard E. Morgan, Rafe McBeth, Jun Tan, Mu-Han Lin, David J. Sher, Dan Nguyen, and Steve Jiang. A proof-of-concept study of artificial intelligence–assisted contour editing. *Radiology: Artificial Intelligence*, 4, 9 2022.

[97] Lei Bi, Michael Fulham, and Jinman Kim. Hyper-fusion network for semi-automatic segmentation of skin lesions. *Medical Image Analysis*, 76:102334, 2 2022.

[98] Xiangde Luo, Guotai Wang, Tao Song, Jingyang Zhang, Michael Aertsen, Jan Deprest, Sebastien Ourselin, Tom Vercauteren, and Shaoting Zhang. Mideepseg: Minimally interactive segmentation of unseen objects from medical images using deep learning. *Medical image analysis*, 72:102102, 2021.

[99] Xuan Liao, Wenhao Li, Qisen Xu, Xiangfeng Wang, Bo Jin, Xiaoyun Zhang, Yanfeng Wang, and Ya Zhang. Iteratively-refined interactive 3d medical image segmentation with multi-agent reinforcement learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9394–9402, 2020.

[100] Antonio Criminisi, Toby Sharp, and Andrew Blake. Geos: Geodesic image segmentation. In *European Conference on Computer Vision*, pages 99–112. Springer, 2008.

[101] Andrew Top, Ghassan Hamarneh, and Rafeef Abugharbieh. Active learning for interactive 3d image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 603–610. Springer, 2011.

[102] Dilek Mirgun Yalcinkaya, Khalid Youssef, Bobby Heydari, Luis Zamudio, Rohan Dharmakumar, and Behzad Sharif. Deep learning-based segmentation and uncertainty assessment for automated analysis of myocardial perfusion mri datasets using patch-level training and advanced data augmentation. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, pages 4072–4078. IEEE, 11 2021.

[103] Patrick Köhler, Jeremiah Fadugba, Philipp Berens, and Lisa M. Koch. Efficiently correcting patch-based segmentation errors to control image-level performance in retinal images. In Ninon Burgos, Caroline Petitjean, Maria Vakalopoulou, Stergios Christodoulidis, Pierrick

Coupe, Hervé Delingette, Carole Lartizien, and Diana Mateus, editors, *Proceedings of The 7nd International Conference on Medical Imaging with Deep Learning*, volume 250 of *Proceedings of Machine Learning Research*, pages 841–856. PMLR, 03–05 Jul 2024.

[104] Guotai Wang, Michael Aertsen, Jan Deprest, Sébastien Ourselin, Tom Vercauteren, and Shaoting Zhang. Uncertainty-guided efficient interactive refinement of fetal brain segmentation from stacks of mri slices. In *International conference on medical image computing and computer-assisted intervention*, pages 279–288. Springer, 2020.

[105] Rúben Baeza, Carolina Santos, Fábio Nunes, Jennifer Mancio, Ricardo Fontes-Carvalho, Miguel Coimbra, Francesco Renna, and João Pedrosa. A generalization study of automatic pericardial segmentation in computed tomography images. In *International Conference on Wireless Mobile Communication and Healthcare*, pages 157–167. Springer, 2022.

[106] Mohamed Marwan and Stephan Achenbach. Quantification of epicardial fat by computed tomography: Why, when and how? *Journal of Cardiovascular Computed Tomography*, 7:3–10, 1 2013.

[107] Carmelo Militello, Leonardo Rundo, Patrizia Toia, Vincenzo Conti, Giorgio Russo, Clarissa Filorizzo, Erica Maffei, Filippo Cademartiri, Ludovico La Grutta, Massimo Midiri, and Salvatore Vitabile. A semi-automatic approach for epicardial adipose tissue segmentation and quantification on cardiac ct scans. *Computers in Biology and Medicine*, 114:103424, 11 2019.

[108] Annika Taithongchai, Isabelle M.A. van Gruting, Ingrid Volløyhaug, Linda P. Arendsen, Abdul H. Sultan, and Ranee Thakar. Comparing the diagnostic accuracy of 3 ultrasound modalities for diagnosing obstetric anal sphincter injuries. *American Journal of Obstetrics and Gynecology*, 221:134.e1–134.e9, 8 2019.

[109] Suraj Bhute, Subhamoy Mandal, and Debashree Guha. Speckle noise reduction in ultrasound images using denoising auto-encoder with skip connection. *arXiv preprint arXiv:2403.02750*, 2024.

[110] Project MONAI. Dynamic unet pipeline - monai tutorials. https://github.com/Project-MONAI/tutorials/tree/main/modules/dynunet_pipeline, 2025. Accessed: 2025-06-19.

[111] Theodore Barfoot, Luis C Garcia Peraza Herrera, Ben Glocker, and Tom Vercauteren. Average calibration error: A differentiable loss for improved reliability in image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 139–149. Springer, 2024.

[112] F. Ramakers, M. Asad, T. Vercauteren, J. Deprest, and H. Williams. BISNet: Boundary Interactive Segmentation Network: Enhancing Automatic Segmentation with Precise Boundary Identification. In *IEEE International Ultrasonics Symposium (IUS)*. IEEE, 2023.