Quantifying Boundary Reliability in Endoscopic Ultrasound Pancreas Segmentation via Evidential Deep Learning

Hana Kim HKIM350@JH.EDU
Department of Electrical and Computer Engineering
Johns Hopkins University
Baltimore, Maryland, United States of America

Khalid Husain KHALIDAHUSAIN@JHU.EDU
Department of Gastroenterology
Johns Hopkins School of Medicine
Baltimore, Maryland, United States of America

Jianing Li JLI516@JH.EDU
Department of Gastroenterology
Johns Hopkins School of Medicine
Baltimore, Maryland, United States of America

Venkata Akshintala VAKSHIN1@JHMI.EDU
Department of Medicine, Division of Gastroenterology and Hepatology
Johns Hopkins School of Medicine
Baltimore, Maryland, United States of America

S. Swaroop Vedula SWAROOP@JHU.EDU
Malone Center for Engineering in Healthcare
Johns Hopkins University
Baltimore, Maryland, United States of America

Marcia Irene Canto MCANTO1@JHMI.EDU
Department of Medicine, Division of Gastroenterology and Hepatology
Johns Hopkins School of Medicine
Baltimore, Maryland, United States of America

Craig Jones CRAIGJ@JHU.EDU
Department of Computer Science
Johns Hopkins University
Baltimore, Maryland, United States of America

Abstract

Pancreatic cancer has a very high mortality rate, making its early detection essential for improving patient survival. Endoscopic Ultrasound (EUS) offers spatial resolution necessary for early-stage screening, but it is sensitive to artifacts, noise, and low contrast, which make automated image analysis challenging. We present an uncertainty-aware framework for EUS pancreas segmentation, benchmarking four U-Net variants across multi-institutional datasets. nnU-Net achieved the strongest performance with a Dice score of 0.811 ± 0.122 on our internal data. We introduce Inside Ground Truth Annulus Percentage (IGTP), a clinically-motivated metric revealing that models systematically fail to adhere to distal pancreatic margins where acoustic shadowing is severe. To localize these failures,

we trained an evidential nnU-Net that parameterizes pixel-wise class probabilities using a Dirichlet distribution and learns evidence through an expected cross-entropy objective with KL regularization. The learned Dirichlet parameters enable the estimation of aleatoric and epistemic uncertainty in a single forward pass. Quantitative analysis showed that higher epistemic uncertainty was associated with poorer segmentation and boundary adherence and provided a predictive signal to identify low-performing cases. Together, the proposed IGTP metric and trained evidential framework establish a trustworthy foundation for AI-assisted EUS screening.

Keywords: *endoscopic ultrasound pancreas, semantic segmentation, boundary evaluation, uncertainty quantification, clinical reliability*

1. Introduction

1.1. Motivation

Pancreatic cancer, especially pancreatic ductal adenocarcinoma (PDAC), remains one of the most lethal malignancies, characterized by a notoriously low 5-year survival rate of 13% Siegel et al. (2024); Park et al. (2021). The disease is largely asymptomatic in its early stages, meaning most patients are diagnosed at an advanced, unresectable stage, highlighting an urgent clinical need for earlier detection and follow-up of high-risk patients.

For early detection of pancreatic cancer, various medical imaging modalities have been used in multiple studies. The main imaging techniques used for pancreatic cancer diagnosis are Computed Tomography (CT), Magnetic Resonance Imaging (MRI), and Endoscopic Ultrasonography (EUS). Traditional modalities such as CT are the standard of care for initial staging and assessing vascular involvement, providing a highly standardized, global view of the abdomen Tummala et al. (2011). It offers comprehensive cross-sectional images of the pancreas, facilitating the evaluation of tumor size, location and involvement of adjacent structures Zhao et al. (2025). However, it can lack the soft-tissue contrast needed to identify very small, early-stage lesions.

MRI offers excellent soft-tissue contrast and is highly effective for characterizing cystic lesions and pancreatic ducts. The normal pancreas is hyperintense compared to surrounding tissue on T1-weighted images because of the large amount of acinar protein within the pancreatic parenchyma. On the other hand, the pancreas is quite variable in signal on T2-weighted images but tends to be relatively hypointense Raman et al. (2012). Despite being expensive, time-consuming, and susceptible to motion artifacts Zhao et al. (2025), MRI is better at characterizing cystic lesions of the pancreas and can provide some indirect radiological evidence to aid in diagnosis of pancreatic cancer. The choice of MRI or CT usually depends on whether local expertise is available and the clinician's comfort with one or the other imaging technique Tummala et al. (2011).

Transabdominal Ultrasound is a procedure used to examine the organs in the abdomen by positioning the ultrasound transducer firmly against the skin of the abdomen. While it allows clinicians to image the patient's region of interest in a non-invasive manner like CT or MRI, it faces challenges such as signal attenuation. The sound waves must travel 10-15cm to reach deep organs, but to survive this distance without being absorbed, the probe must emit low frequency. However, this lower frequency necessitates a resolution trade-off, making it difficult to observe small lesions or pancreatic tail tumors. To overcome these limitations, instead of placing the transducer probe on the skin, we combine flexible endoscopy and

high-frequency ultrasonography by inserting the endoscope through the mouth and into the stomach or duodenum, directly adjacent to the pancreas. This way, EUS provides an image with unparalleled, sub-millimeter spatial resolution.

Due to its high resolution, EUS is one of the primary modalities for identifying solid lesions at an early-stage, sub-centimeter and tracking subtle changes in high-risk screening cohorts. Despite its clinical value, EUS is inherently operator-dependent, requiring a steep learning curve. Interpretation is subjective, and the modality suffers from typical ultrasound artifacts, making automated, reliable image analysis highly valuable for standardizing care.

1.2. Background

Accurate delineation of the pancreas is a necessary prerequisite for advanced downstream tasks, for example, developing imaging-based biomarkers to enable early detection of pancreatic cancer, automated radiomics in texture analysis, longitudinal tumor volume tracking, and computer-aided diagnosis. However, accurate segmentation of the pancreas is extremely challenging due to the organ's small size, the complex variability it holds in shape, and the indistinct boundaries with its neighboring organs Antony et al. (2025). Manual segmentation is prohibitively time-consuming for routine clinical workflows, requiring an average of 3.37 ± 1.47 minutes per volume. Furthermore, it suffers from high inter-observer and intra-observer variability, with Dice scores dropping as low as 0.87 even among expert radiologists evaluating highly standardized CT scans Khasawneh et al. (2022).

Current segmentation methods in medical applications often face limitations when relying on object detection frameworks. While bounding boxes are computationally efficient for identifying general regions of interest, they often struggle with imprecise localization, especially when dealing with overlapping or small anatomical structures where a box may contain multiple objects or significant background noise Steele (2023).

While deep learning has shown promise in medical imaging with voxel-level semantic segmentation, achieving consistent results in EUS pancreas images remains a challenge. Unlike object detection, which localizes structures with coarse bounding boxes, semantic segmentation involves the process of assigning a specific class label to every individual pixel or voxel in an image. This allows for the precise delineation of complex anatomical boundaries, which is essential for quantifying organ volume and morphology. Specifically, the U-Net architecture and its variants have become the state-of-the-art for medical image segmentation due to their modular design and use of skip connections to preserve spatial context Azad et al. (2022). In the context of automated pancreas segmentation, recent advances have moved toward 3D U-Net configurations and attention mechanisms to better capture long-range dependencies and distinguish the organ from surrounding low-contrast tissues Suri et al. (2024).

However, the vast majority of automated pancreas segmentation literature focuses heavily on CT and MRI datasets. Robust segmentation frameworks tailored for EUS remain underexplored due to the lack of large, curated datasets and the unique complexities of ultrasound data Seo et al. (2022). EUS images are characterized by significant speckle noise, which reduces image contrast and obscures fine anatomical details to a much greater degree than MRI or CT Oktar et al. (2003). Furthermore, acoustic shadowing, which is caused by gas in the gastrointestinal tract or dense calcifications, can lead to complete signal dropout,

creating blind spots that degrade image quality and reliability Oktar et al. (2003); Dahiya et al. (2024).

These obstacles in EUS pancreas segmentation are aggravated by the anatomical variability that appears in the pancreas. The appearance varies based on the angle of the transducer and the specific region of the pancreas being imaged, which ranges from head, body, to tail. For instance, the pancreatic tail often presents the lowest detection and segmentation accuracy due to suboptimal acoustic access and interference from adjacent structures Dahiya et al. (2024); Jain et al. (2025). This is due to EUS being a highly operator-dependent modality. Therefore, standard segmentation models often fail to generalize across different scan planes, leading to overconfident but incorrect predictions where the model hallucinates boundaries in ambiguous, hypoechoic regions Seo et al. (2022); Jain et al. (2025).

Given these inherent ambiguities that lie in EUS images, there is a critical need for models that provide not only segmentation predictions for the region of interest, but also pixel-wise uncertainty estimates. Traditional methods for uncertainty quantification relied on Bayesian Neural Networks (BNN), which were often approximated by Monte Carlo Dropout (MC-Dropout) or Deep Ensembles Gal and Ghahramani (2016); Lakshminarayanan et al. (2017). Although effective, these methods are computationally expensive since they require multiple forward passes for a single inference. This limits their utility in real-time clinical applications like live EUS procedures.

Unlike Monte Carlo Dropout and Deep Ensembles, Evidential Deep Learning (EDL) does not require repeated stochastic inference. EDL provides an alternative framework for uncertainty quantification by training a network to predict evidence for each class rather than only fixed softmax probabilities Sensoy et al. (2018). There are two types of uncertainty, aleatoric uncertainty, which arises from the noise inherent in the data itself, such as speckle noise and acoustic shadowing. Next comes epistemic uncertainty, which quantifies its lack of confidence in its knowledge of making predictions. In EDL, placing a higher-order prior distribution over the model's likelihood function, it enables the estimation of both aleatoric and epistemic uncertainty in a single forward pass, providing a mathematically grounded score without the overhead of sampling-based methods.

To overcome these limitations in EUS-based pancreas analysis, namely the lack of specialized benchmarking, the ambiguity of anatomical boundaries, and the need for efficient uncertainty quantification, this work proposes a robust deep learning framework validated across multi-institutional datasets.

Contributions

1. **Benchmark Analysis on EUS Data:** We conduct an evaluation of the state-of-the-art nnU-Net framework against diverse U-Net variants using a unique, multi-visit internal site EUS dataset, capturing the temporal and anatomical complexity of high-risk screening.

2. **Multi-Institutional Validation:** We rigorously evaluate the cross-institutional generalizability of our proposed pipeline by using an independent external dataset, highlighting the critical need for cross-site reliability in medical AI.

3. **Introduction of the IGTP Metric:** We propose a novel, clinically-motivated metric, Inside Ground Truth Annulus Percentage (IGTP), which was designed specifically to quantify segmentation failures at the ambiguous lower border of the pancreas, where standard metrics like Dice often fail to capture subtle morphological inaccuracies.

4. **Evidentially Trained Uncertainty Quantification:** We train an EDL-enabled nnU-Net using a Dirichlet evidential objective to jointly learn segmentation predictions and pixel-wise evidence. The learned concentration parameters enable aleatoric and epistemic uncertainty estimation in a single forward pass, and we quantitatively evaluate their association with segmentation performance and failure detection.

Together, these contributions bridge the gap between high-performance deep learning and clinical utility in EUS. By providing both precise segmentations and a transparent measure of trustworthiness, this work offers a framework for standardizing the early detection of pancreatic malignancies in high-risk populations.

Generalizable Insights about Machine Learning in the Context of Healthcare

Our work provides several key insights into the deployment and evaluation of deep learning models within challenging clinical workflows:

- **Uncertainty Maps as a Defense Against Silent Failures:** In high-stakes medical imaging, a model's ability to "know when it doesn't know" is as vital as its predictive accuracy. We show that while standard architectures may yield overconfident, incorrect segmentations in noisy EUS environments, EDL provides a computationally efficient way to generate uncertainty maps. These maps serve as a visual audit trail, successfully highlighting ambiguous organ boundaries and hallucinated regions, thereby offering a blueprint for enhancing model reliability in other signal-starved modalities.

- **Bridging the Gap with Anatomically-Motivated Metrics:** Standard segmentation losses are often directionally and clinically agnostic. The introduction of the Inside Ground Truth Annulus Percentage (IGTP) metric illustrates that by designing custom evaluation frameworks, such as measuring specific boundary degradation at the lower organ border, researchers can pinpoint exact algorithmic failure modes that standard metrics overlook. This approach encourages a shift from optimizing for pixel-wise overlap toward optimizing for clinical utility and anatomical correctness.

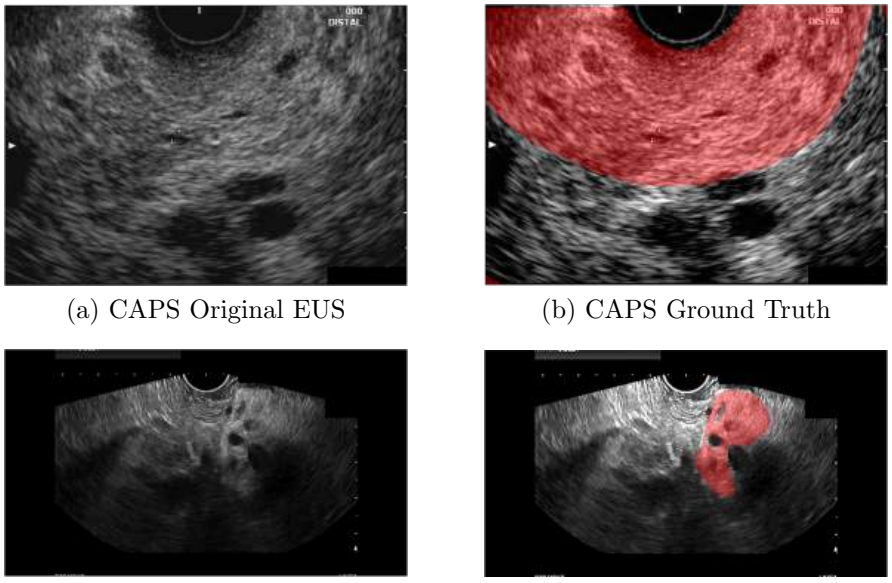
2. Data

2.1. CAPS Data (Internal Dataset)

The primary dataset for this study is derived from the Cancer of the Pancreas Screening (CAPS) cohort, a long-term, multi-center surveillance program based at Johns Hopkins Hospital. The program is specifically designed for the early detection of pancreatic neoplasia in high-risk individuals. Participants in the CAPS cohort exhibit significant genetic or familial predisposition to PDAC, with inclusion criteria targeting:

- individuals with at least two first-degree relatives with pancreatic cancer,
- carriers of specific germline mutations (e.g., *BRCA2*, *p16*, or *STK11*) regardless of family history.
- patients with a history of hereditary pancreatitis.

The data used in this work consist of archived static image captures (screenshots) obtained during longitudinal surveillance. All images were acquired using linear EUS transducers, providing sub-millimeter spatial resolution of the pancreatic parenchyma. Notably, the dataset contains multi-visit data, capturing anatomical changes in high-risk patients over multiple years.



(a) CAPS Original EUS (b) CAPS Ground Truth

(c) External Site Original EUS (d) External Site Ground Truth

Figure 1: Representative EUS scans and corresponding pancreatic segmentations (red) from the CAPS and the external site cohorts.

2.2. External Validation Dataset

To evaluate the cross-institutional generalizability of our models, we utilized a secondary dataset from Sibley Hospital. This dataset comprises patients who underwent EUS for standard indications, representing a distinct clinical environment and patient demographic from the primary CAPS cohort. Unlike the longitudinal nature of the CAPS data, the external dataset primarily consists of single-visit evaluations. Testing on this out-of-distribution data is critical for assessing the robustness of our segmentation frameworks against variations in institutional protocols, endosonographer technique, and hardware-specific image processing.

2.3. Manual Segmentation

The manual segmentation of anatomical structures, including the pancreas, pancreatic cysts, and ducts, followed a multi-stage pipeline to ensure high-fidelity ground truth labels. For both cohorts, we applied consistent inclusion and exclusion criteria. Images were excluded if they contained no visible pancreas, poor image quality, or duplicates. In cases where duplicate images differed only in annotations, one was excluded to avoid redundancy.

Initial segmentations were performed by trained student annotators using the Medical Imaging Interaction Toolkit (MITK). These preliminary labels then underwent a two-phase expert review process. In the first phase, a senior endoscopist performed a comprehensive quality control check, correcting structural boundaries, and providing specific feedback for refinement.

The final validation was conducted by an experienced endoscopist through interactive, real-time review sessions. During these sessions, segmentations were evaluated in the MITK environment with label overlays toggled to verify anatomical accuracy. The clinical expert provided precise feedback using on-screen drawing tools to resolve ambiguities, such as distinguishing between pancreatic ducts and cysts or adjusting the extension of the pancreatic tail. These expert annotations and session screenshots served as a definitive reference for the final corrective pass, resulting in the finalized dataset used for model training and evaluation.

2.4. Data Statistics

The distribution of patients, clinical visits, and images used for training and testing are summarized in Table 1.

Table 1: Summary of Dataset Distribution and Image Characteristics

Metric	**CAPS (Train)**	**CAPS (Test)**	**External**
Patients	54	20	41
Clinical Visits	187	46	45
Total Images	3,391	839	454
Mean Pixels (W × H)	615 × 412	592 × 406	1150 × 784

3. Methods

3.1. Segmentation Model

To address the challenges of high speckle noise and variable organ morphology in EUS, we evaluated several state-of-the-art deep learning architectures. Our primary framework is the self-configuring nnU-Net, which we benchmark against three distinct U-Net variants that represent convolutional and transformer-based medical segmentation.

3.1.1. nnU-Net

We utilized nnU-Net as our core segmentation engine Isensee et al. (2021). Unlike static architectures, nnU-Net is a self-configuring framework that automatically adapts preprocessing, such as resampling and normalization, network topology, and training hyperparameters based on the fingerprint of the specific dataset. For this study, the 2D configuration was utilized to process EUS frames due to our data's dimensionality. Given the high variability in EUS image acquisition in different centers, the ability of nnU-Net to systematically optimize the training pipeline ensures a robust baseline that has consistently outperformed manually tuned models in the challenges of pancreatic CT and MRI segmentation Antonelli et al. (2022).

3.1.2. Architectural Baselines

To thoroughly assess the performance of nnU-Net, we compared it with three alternative architectures that employ distinct strategies for modeling spatial context. U-Net, a foundational encoder-decoder architecture for medical imaging, serves as our primary baseline Ronneberger et al. (2015). It was developed to help provide the localization of the relatively small cross-sectional area of the pancreas in ultrasound frames. The next architecture we used for performance comparison is the U-Net++ Zhou et al. (2018). We selected U-Net++ to evaluate its ability to capture fine-grained features that help delineate the irregular and ambiguous boundaries characteristic of the pancreatic head and tail in EUS. Our last U-Net variant model, Swin U-Net, represents the shift toward transformer-based vision models Cao et al. (2021). We included this model to test whether global context is more effective than local convolutions at distinguishing the pancreas from similar-textured adjacent organs, such as the liver or spleen, in noisy ultrasound environments.

3.2. Experiments

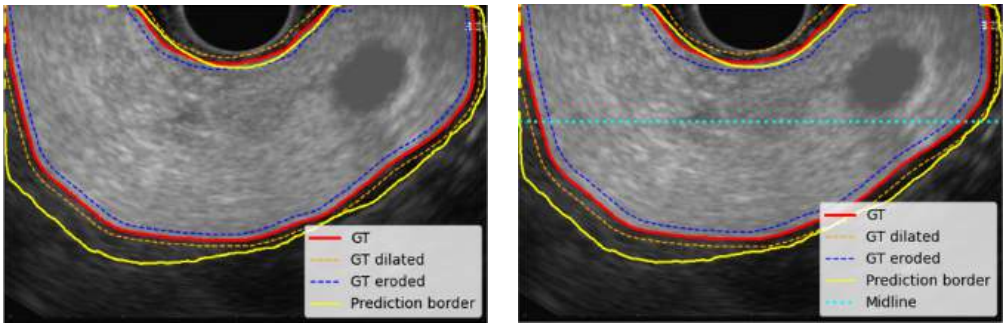
3.2.1. Cross-Validation Strategy

We conducted our experiment with a strict 5-fold cross-validation strategy on the CAPS cohort to ensure the robustness of the proposed models and mitigate overfitting on our EUS dataset. All folds were partitioned at the patient level rather than the image level to prevent data leakage. For the external site dataset, this cohort was specifically held-out as an independent external test set and was never used for training and validation, allowing us to assess cross-institutional generalizability. Together, this evaluation strategy guarantees that reported performance reflects the models' ability to generalize to unseen patients both within and across populations.

3.2.2. Evaluation Metrics

Given that EUS organ boundaries are often ambiguous due to speckle noise and artifacts, relying solely on volumetric overlap is insufficient for clinical validation. We evaluated model performance using a multi-faceted approach encompassing spatial overlap, boundary distance, and a novel clinical adherence metric.

For spatial overlap and distance measurement, we utilized the Dice Similarity Coefficient (DSC), Intersection over Union (IoU) to quantify the global overlap between the



(a) Full region analysis. (b) Upper and lower half region analysis.

Figure 2: Visual analysis of the IGTP metric on an EUS scan for the original view and the re-oriented view with the midline of the GT. The GT annulus is bounded by dilated (orange) and eroded (blue) lines. IGTP is calculated by the percentage of the intersection of the prediction boundary (yellow) with the GT annulus.

predicted and ground truth masks, and the 95th percentile Hausdorff Distance (HD95) of the predicted mask boundary and ground truth mask boundary to measure the maximum distance between the predicted and true boundaries, excluding the top 5% of outliers to provide a stable estimate of boundary error. Apart from these standard metrics used in medical images, we introduce a novel, clinically-motivated metric, Inside Ground Truth Annulus Percentage (IGTP). This new metric was designed to quantify the adherence of the predicted segmentation boundary to the ground truth (GT) border, evaluated within a defined spatial tolerance zone. It measures what percentage of the predicted boundary falls within a predefined pixel-space tolerance zone around the ground truth, and is calculated as the following:

$$IGTP = \frac{|\text{Pred}_{\text{border}} \cap \text{GT}_{\text{annulus}}|}{|\text{Pred}_{\text{border}}|} \times 100\%$$

A boundary tolerance zone, termed the GT annulus, is generated by dilating and eroding the manual ground truth mask using a square structuring element with a kernel size of 20 pixels, creating a symmetric buffer zone around the true GT boundary. A comprehensive sensitivity analysis validating the stability of this kernel size is detailed in Appendix A.2. Next, to evaluate the perimeter of the model's prediction mask, we isolate its 1-pixel thick contour. This is attained by extracting all foreground pixels within the model's binary prediction mask that are directly adjacent to the background. The IGTP is then calculated as the percentage of these predicted border pixels that spatially intersect with the GT annulus. This approach accommodates the inherent variability in manual annotation while still penalizing clinically significant deviations.

A higher IGTP indicates that the model's boundary adheres strictly to the ground-truth-centered tolerance region. Notably, a directional analysis of IGTP revealed a distinct failure mode: models yielded significantly lower IGTP scores on the lower border of the pancreas compared to the upper half. This observation aligns with the physics of EUS, where acoustic shadowing and tissue attenuation often obscure the distal margins of the

organ, making precise segmentation more challenging for both human annotators and AI models. We investigate this directional pattern quantitatively in Section 4.2 and validate it across the external dataset in Section 4.4.

3.2.3. Statistical Analysis

Statistical analyses were performed at the image level. For each metric, predictions from the different architectures were aligned by image and model performance was compared using two-sided paired Wilcoxon signed-rank tests. Significance was assessed using a Bonferroni-corrected threshold of $0.05/6 = 0.0083$, corresponding to the six pairwise comparisons among the four architectures. Effect sizes were reported as approximate standardized Wilcoxon effect sizes, $r = \frac{Z}{\sqrt{N}}$, where N denotes the number of paired images. Upper and lower-boundary IGTP values were compared within each architecture using two-sided paired Wilcoxon signed-rank tests, with Bonferroni correction across the four architecture-specific directional comparisons ($\alpha_{\text{corrected}} = 0.0125$). Epistemic uncertainty was evaluated at the image level using Spearman correlation and AUROC/AUPRC failure-detection analyses.

3.3. Evidential Deep Learning

Given the persistent boundary failures observed at the distal margins in both the CAPS and external site cohorts, standard softmax probabilities are insufficient, as they often yield overconfident predictions even in these ambiguous regions. To evaluate the reliability of our segmentation results, we employed a Dirichlet-based Evidential Deep Learning (EDL) framework to decompose the model's predictive variance into Aleatoric Uncertainty (AU) and Epistemic Uncertainty (EU). This framework allows us to identify where the model lacks sufficient evidence to make a clinical decision.

Following Sensoy et al. (2018), we place a Dirichlet prior distribution over the model's predicted class probabilities. Instead of directly predicting class probabilities $p = [p_1, p_2, ..., p_C]$ for C classes, the network predicts a concentration parameter $\alpha = [\alpha_1, \alpha_2, ..., \alpha_C]$, where $\alpha_k > 0$ represents the strength of evidence for class k.

$$\text{Dir}(p \mid \alpha) = \frac{\Gamma\left(\sum_k \alpha_k\right)}{\prod_k \Gamma(\alpha_k)} \prod_k p_k^{\alpha_k - 1}$$

The total evidence collected by the model is represented by the strength parameter $S = \sum_{k=1}^{C} \alpha_k$. The expected probability for any given class is therefore $\text{E}[p_k] = \frac{\alpha_k}{S}$.

During training, the concentration parameters were optimized directly using a Dirichlet evidential objective composed of an expected cross-entropy term and a KL-divergence regularizer toward a uniform Dirichlet prior. The expected cross-entropy encourages evidence for the correct class, whereas the KL term penalizes unsupported evidence. Therefore, the magnitude of evidence used for uncertainty estimation was learned during training. The complete loss formulation, regularization coefficient, and training schedule are provided in Appendix B. At inference, the learned concentration parameters were used to derive two distinct, quantifiable types of uncertainty, AU and EU.

AU represents inherent data noise that can't be eliminated by adding more data. In our EUS context, this includes acoustic shadowing, speckle noise, and inherently ambiguous

tissue borders. AU is calculated as the expected variance of the Dirichlet distribution over class probabilities.

$$AU = \sum_{k=1}^{C} \frac{p_k(1-p_k)}{S+1}$$

High AU indicates regions where the data is inherently noisy, regardless of model capacity. On the other hand, model uncertainty or a lack of evidence is represented using EU. A high EU indicates the model is guessing because it has not seen similar features in the training data. EU is inversely proportional to the total evidence S. Therefore, as S increases, EU decreases.

$$EU = \sum_{k=1}^{C} \frac{p_k(1-p_k)S}{(S+1)^2}$$

Unlike Monte Carlo Dropout and Deep Ensembles, this approach does not require repeated stochastic forward passes. The segmentation prediction and uncertainty maps are therefore generated together in a single deterministic forward pass. We subsequently evaluate whether elevated EU is associated with low DSC or IoU, high HD95, and poor IGTP boundary adherence.

4. Results and Analysis

4.1. Results on CAPS Data

The performance of the four benchmarked architectures on the CAPS internal dataset is summarized in Table 2 and visual representations can be found in Figure 4. Overall, nnU-Net demonstrated the most robust performance, achieving the highest scores in three of the four evaluation metrics, with a DSC of 0.811 ± 0.122 and an IoU of 0.698 ± 0.159. Notably, nnU-Net outperformed the other models for our IGTP metric (45.5%), suggesting that its self-configuring normalization and training pipeline are particularly effective at aligning predicted boundaries within the predefined ground-truth annulus. Complete pairwise test statistics, corrected p-values and effect sizes are reported in Appendix C.

Model	**DSC** (↑)	**IoU** (↑)	**HD95** (↓)	**IGTP** (↑)
nnU-Net	**0.811 ± 0.122**	**0.698 ± 0.159**	79.0 ± 45.6	**45.5 ± 19.0**
U-Net	0.794 ± 0.131	0.676 ± 0.164	**75.9 ± 42.7**	42.5 ± 18.7
U-Net++	0.790 ± 0.130	0.670 ± 0.161	80.0 ± 43.2	40.7 ± 17.9
Swin U-Net	0.778 ± 0.127	0.653 ± 0.159	81.7 ± 43.6	34.7 ± 16.6

Table 2: Model performance comparison across different U-Net variant models on the CAPS patient data.

Although nnU-Net achieved the strongest DSC, IoU, and IGTP results, U-Net achieved the lowest mean HD95 value (75.9 ± 42.7). This indicates that, on average, U-Net produced smaller 95th-percentile boundary-distance errors, despite its lower overlap and IGTP performance. However, the transformer-based Swin U-Net lagged across all categories, particularly in IGTP (34.7%), indicating that pure self-attention mechanisms may struggle more

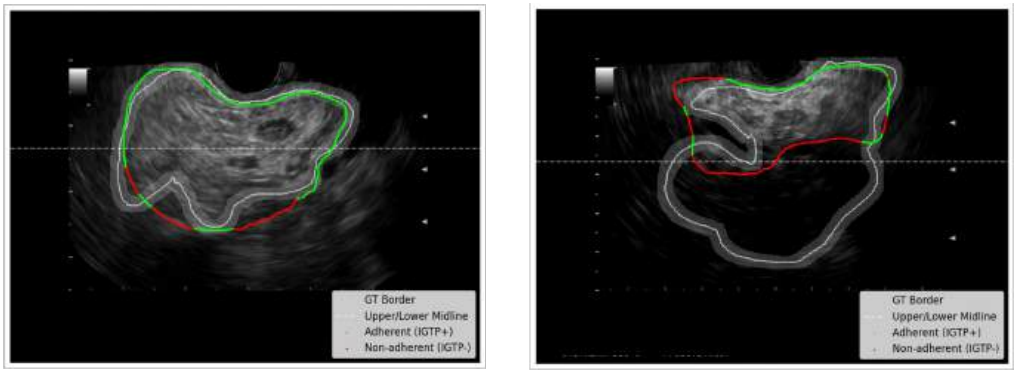
than convolutional architectures to delineate the fine-grained, speckle-heavy boundaries of the pancreas. These results validate the selection of nnU-Net as the primary framework for subsequent pancreatic analysis.

4.2. IGTP Analysis of Boundary Adherence on CAPS Data

Table 3 shows the performance of the directional decomposition of the IGTP metric to further investigate each model's failure modes. Across all architectures, a significant performance gap was observed between the upper and lower halves of the pancreas. For our best performing model, nnU-Net, the IGTP score decreased from 54.3% on the upper boundary to 35.6% on the lower boundary. This trend was even more pronounced in Swin U-Net's performance, which saw nearly a 50% relative reduction in boundary adherence (47.1% vs. 25.3%).

Model	IGTP Full (↑)	IGTP Upper Half (↑)	IGTP Lower Half (↑)
nnU-Net	**45.5 ± 19.0**	**54.3 ± 20.6**	**35.6 ± 28.1**
U-Net	42.5 ± 18.7	53.1 ± 21.8	32.1 ± 26.2
U-Net++	40.7 ± 17.9	51.8 ± 21.2	30.6 ± 24.8
Swin U-Net	34.7 ± 16.6	47.1 ± 20.9	25.3 ± 22.6

Table 3: IGTP comparison based on the upper and lower half of the pancreas across different U-Net variant models on the CAPS patient data.



(a) Representative Success (b) Shadowing Challenge

Figure 3: Directional IGTP heatmap analysis. (a) illustrates robust boundary adherence across both margins in a high-SNR frame. (b) demonstrates a characteristic failure mode where prominent vertical acoustic shadowing (center-left) obscures the distal boundary, causing significant non-adherence (red) along the lower pancreatic margin.

The spatial distribution of these adherence patterns is qualitatively visualized in Figure 3. While the model demonstrates precise boundary conformity in high-contrast regions (Fig. 3a), a distinct failure mode emerges when the organ boundary intersects with acoustic artifacts. These qualitative and quantitative findings provide strong empirical evidence for the impact of EUS-specific artifacts on deep learning models. The superior performance on

the upper boundary likely stems from its proximity to the duodenal wall, which provides a high-contrast acoustic interface.

Conversely, the lower boundary is frequently obscured by acoustic shadowing from intervening structures and signal attenuation in deeper tissues, as clearly seen in the non-adherent (red) clusters within the shadowed regions of Figure 3b. The high standard deviations observed in the lower-half metrics (28.1% for nnU-Net) further reflect the inconsistent visibility of these distal margins across different patient anatomies. This directional disparity highlights that while global overlap metrics may appear high, current models still struggle to reliably delineate fuzzy boundaries in low-SNR regions of the ultrasound frame.

4.3. Results on External Data

The evaluation on the external dataset, presented in Table 4, reveals significant domain shift. Spatial overlap metrics degraded substantially: the highest DSC dropped from 0.811 (internal) to 0.701 (external). Notably, the architecture ranking shifted, where standard U-Net achieved the highest DSC (0.701 ± 0.165) and IoU (0.562 ± 0.178), while nnU-Net dropped to 0.685. This inversion suggests that nnU-Net's self-configuring pipeline, though optimized for CAPS, does not generalize as robustly to out-of-distribution hardware and imaging protocols.

Model	**DSC** (↑)	**IoU** (↑)	**HD95** (↓)	**IGTP** (↑)
nnU-Net	0.685 ± 0.150	0.538 ± 0.158	214.447 ± 115.764	27.9 ± 18.3
U-Net	**0.701 ± 0.165**	**0.562 ± 0.178**	187.533 ± 104.044	**27.9 ± 15.3**
U-Net++	0.699 ± 0.159	0.559± 0.172	**183.506 ± 95.338**	26.6 ± 14.7
Swin U-Net	0.686 ± 0.151	0.539 ± 0.159	191.278 ± 102.942	22.5 ± 12.8

Table 4: Model performance comparison across different U-Net variant models on the external patient data.

All architectures showed lower mean performance on the external cohort, consistent with a cross-site domain shift. U-Net achieved the highest mean DSC and IoU, while U-Net and nnU-Net achieved equal mean full-boundary IGTP. The change in numerical rankings suggests sensitivity to cohort-specific acquisition characteristics. However, the modest number of external patients and high variability preclude broad conclusions regarding the intrinsic generalizability of individual architectures. The 5.4 percentage-point IGTP difference between the convolutional models and Swin U-Net is reported as a numerical difference. Its clinical significance has not yet been established.

4.4. IGTP Analysis of Boundary Adherence on External Data

The upper–lower disparity was reproduced in the external cohort, supporting the consistency of this directional failure pattern across institutions. However, the observational analysis can't isolate ultrasound physics from acquisition, anatomy, or annotation effects. Table 5 details the directional decomposition of IGTP to isolate where boundary failures manifest on external data. Critically, the directional vulnerability observed in internal CAPS data replicates on external data, confirming a modality-level phenomenon rather

than institutional artifact. All models show approximately 50% relative reduction in boundary adherence when comparing upper-half to lower-half regions. nnU-Net exemplifies this: upper-half IGTP of 40.0% plummets to 19.0% on the lower boundary, which is a −52.5% relative drop. Even the superior U-Net (27.9% full IGTP) shows near-identical directional degradation: 37.8% upper versus 18.9% lower.

Model	**IGTP Full** (↑)	**IGTP Upper Half** (↑)	**IGTP Lower Half** (↑)
nnU-Net	27.9 ± 18.3	**40.0 ± 20.9**	**19.0 ± 19.4**
U-Net	**27.9 ± 15.3**	37.8 ± 19.2	18.9 ± 19.1
U-Net++	26.6 ± 14.7	36.8 ± 19.2	17.2 ± 17.1
Swin U-Net	22.5 ± 12.8	31.1 ± 17.3	14.2 ± 13.7

Table 5: IGTP comparison based on the upper and lower half of the pancreas across different U-Net variant models on the external validation site patient data.

The upper–lower IGTP disparity was reproduced in the external cohort, suggesting that the directional pattern was not unique to the internal institution. Its localization to the deeper pancreatic margin is consistent with the known effects of attenuation and acoustic shadowing in EUS. However, the observational analysis cannot isolate imaging physics from acquisition technique, patient anatomy, equipment, or annotation variability. Together, these findings show that competitive global overlap can coexist with substantially poorer adherence along the lower pancreatic boundary.

4.5. Quantitative and Qualitative EDL Analysis on CAPS Data

To further understand where the segmentation model fails, we analyzed the uncertainty maps across high-performing and failure cases of the nnU-Net on our CAPS cohort. Figure 4 illustrates the spatial relationship between model uncertainty and boundary adherence. In high-performing image cases (Rows 1 & 2), where the DSC exceeds 0.95, the uncertainty maps show that the EU is minimal inside the pancreatic parenchyma and strictly localized to the inherently ambiguous interface at the organ boundary. This pattern indicates that the model has generalized to the internal texture of the pancreas and is confident in its volumetric representation, with uncertainty appropriately concentrated at the pancreas' boundary.

Conversely, the failure cases in Rows 3 & 4 reveal a distinct correlation between high uncertainty and poor boundary adherence. Elevated epistemic uncertainty extended beyond the predicted contour and into the pancreatic region, qualitatively co-localizing with regions of poor segmentation. These examples illustrate the spatial behavior of the uncertainty maps but do not independently establish calibration or clinical reliability. The quantitative associations and failure-detection analyses reported below provide the primary evidence for the relationship between uncertainty and segmentation quality.

We also quantitatively evaluated whether epistemic uncertainty was associated with segmentation quality among the cases with complete uncertainty outputs. The 95th-percentile epistemic uncertainty within the predicted pancreas was negatively associated with DSC ($\rho = -0.471$) and IGTP ($\rho = -0.207$) and positively associated with HD95 ($\rho = 0.335$).

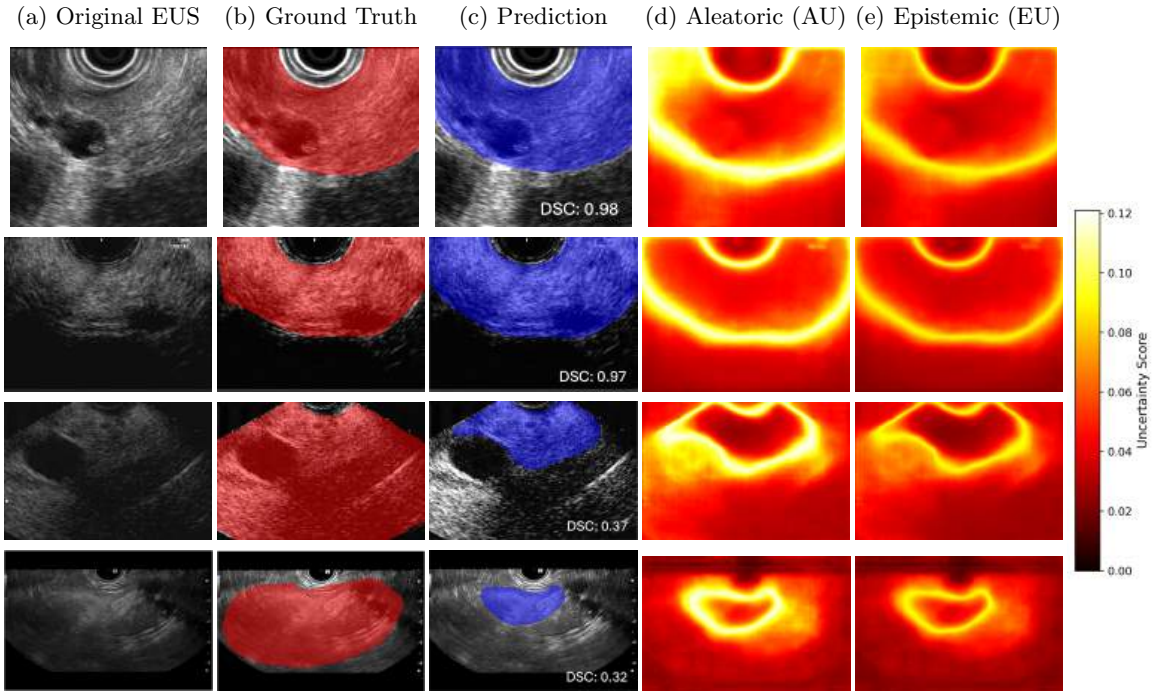


Figure 4: The top two rows display high-performing cases (DSC > 0.95) where the nnU-Net successfully captures the pancreas bulk with low epistemic uncertainty. The bottom two rows display failure cases where obscure lower boundaries result in high uncertainty (yellow regions) and corresponding poor boundary adherence.

Thus, images with higher epistemic uncertainty generally exhibited poorer overlap, larger boundary-distance errors, and lower boundary adherence.

For detecting bottom-quartile DSC failures, the 95th-percentile epistemic uncertainty within the predicted pancreas achieved an AUROC of 0.741 and an AUPRC of 0.568. Boundary-localized epistemic uncertainty achieved an AUROC of 0.715 and an AUPRC of 0.517. These results demonstrate that the learned uncertainty estimates contain quantitative failure-detection signal beyond the qualitative co-localization shown in Figure 4.

These results indicate that EU may support the prioritization of potentially low-performing segmentations for additional review. However, the observed discrimination is moderate, and no clinically validated uncertainty threshold is established in this study. Prospective calibration and evaluation of the sensitivity–workload tradeoff are required before uncertainty-based flags can be incorporated into clinical workflows.

5. Discussion

This study establishes a baseline for automated pancreas segmentation in EUS images. By evaluating nnU-Net and other U-Net variants on the multi-visit CAPS cohort and validating the models on an independent external dataset, we found that deep learning can successfully localize the pancreas, although generalization across institutions remains challenging. The

shift in architecture rankings suggests sensitivity to site-specific acquisition characteristics but does not by itself establish overfitting. Higher epistemic uncertainty was associated with poorer segmentation and provided moderate failure-detection signal, supporting its potential use for review prioritization. Prospective validation is required before uncertainty-based flags can be used in clinical workflows.

A central contribution of this work is the reconciliation of clinical requirements with computational evaluation. In routine EUS screening, endoscopists must identify pancreatic regions for cyst or mass detection while also evaluating organ contours and lesion margins for subtle morphological changes. From a machine learning perspective, global overlap metrics can conceal clinically relevant boundary failures. Models may achieve relatively high DSC while systematically misdelineating the lower pancreatic margin. On the external dataset, all four architectures achieved DSC values near or above 0.70 while exhibiting an approximately 50% relative reduction in IGTP from the upper to the lower boundary.

The lower pancreatic boundary showed significantly poorer adherence across architectures and cohorts. This pattern is consistent with the effects of acoustic attenuation and shadowing in deeper regions of EUS images, although the present analysis can't separate imaging physics from acquisition technique, patient anatomy, or equipment. Higher EU was associated with poorer segmentation and provided moderate discrimination of low-performing cases. Together, IGTP and EDL provide complementary information about segmentation reliability that is not captured by global overlap metrics alone.

While existing literature focuses heavily on the standardized environments of CT and MRI, our work highlights that EUS requires a fundamentally different evaluation paradigm. By shifting the focus from global overlap toward anatomically-motivated and uncertainty-guided frameworks, we provide a blueprint for developing trustworthy AI in signal-dependent medical modalities. Future research will focus on integrating these uncertainty-aware masks into real-time decision support systems, evolving from organ localization to automated anomaly detection and longitudinal characterization.

Limitations. The models were trained and evaluated using 2D static EUS frames rather than continuous video, and the datasets contained a modest number of independent patients despite a larger number of images. Because the statistical analyses were performed at the image level, correlations among images from the same patient were not explicitly modeled, which may overstate the effective amount of independent information. Future analyses should use patient-level aggregation, cluster-aware inference, or mixed-effects models to account for repeated observations. The selected IGTP tolerance was a methodological operating point and has not been clinically validated. In addition, the EDL model was not directly compared with MC Dropout, Deep Ensembles, or other uncertainty-estimation baselines in terms of calibration or failure-detection performance. Differences in external performance may also reflect variation in acquisition technique, equipment, patient anatomy, and annotation practices. Future work should evaluate temporal models, anatomical-station conditioning, alternative uncertainty methods, physics-informed training, and prospective integration into live endoscopy workflows.

Acknowledgments

This work was supported by the Lustgarten Foundation.

References

Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A. Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M. Summers, Bram van Ginneken, Michel Bilello, Patrick Bilic, Patrick F. Christ, Richard K. G. Do, Marc J. Gollub, Stephan H. Heckers, Henkjan Huisman, William R. Jarnagin, Maureen K. McHugo, Sandy Napel, Jennifer S. Golia Pernicka, Kawal Rhode, Catalina Tobon-Gomez, Eugene Vorontsov, James A. Meakin, Sebastien Ourselin, Manuel Wiesenfarth, Pablo Arbeláez, Byeonguk Bae, Sihong Chen, Laura Daza, Jianjiang Feng, Baochun He, Fabian Isensee, Yuanfeng Ji, Fucang Jia, Ildoo Kim, Klaus Maier-Hein, Dorit Merhof, Akshay Pai, Beomhee Park, Mathias Perslev, Ramin Rezaiifar, Oliver Rippel, Ignacio Sarasua, Wei Shen, Jaemin Son, Christian Wachinger, Liansheng Wang, Yan Wang, Yingda Xia, Daguang Xu, Zhanwei Xu, Yefeng Zheng, Amber L. Simpson, Lena Maier-Hein, and M. Jorge Cardoso. The medical segmentation decathlon. *Nature Communications*, 2022.

Ajith Antony, Sovanlal Mukherjee, Yan Bi, Eric A. Collisson, Madhu Nagaraj, Murlidhar Murlidhar, Michael B. Wallace, and Ajit H. Goenka. Ai-driven insights in pancreatic cancer imaging: from pre-diagnostic detection to prognostication. *Author Manuscript*, 50 (7), 2025.

Reza Azad, Ehsan Khodapanah Aghdam, Amelie Rauland, Yiwei Jia, Atlas Haddadi Avval, Afshin Bozorgpour, Sanaz Karimijafarbigloo, Joseph Paul Cohen, Ehsan Adeli, and Dorit Merhof. Medical image segmentation review: The success of u-net. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12), 2022.

Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. *arXiv preprint arXiv:2105.05537*, 2021.

Dushyant Singh Dahiya, Yash R. Shah, Hassam Ali, Saurabh Chandan, Manesh Kumar Gangwani, Andrew Canakis, Daryl Ramai, Umar Hayat, Bhanu Siva Mohan Pinnam, Amna Iqbal, Sheza Malik, Sahib Singh, Fouad Jaber, Saqr Alsakarneh, Islam Mohamed, Meer Akbar Ali, Mohammad Al-Haddad, and Sumant Inamdar. Basic principles and role of endoscopic ultrasound in diagnosis and differentiation of pancreatic cancer from other pancreatic lesions: A comprehensive review of endoscopic ultrasound for pancreatic cancer. *Journal of Clinical Medicine*, 13(9), 2024.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning, 2016.

Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2021.

Aryan Jain, Mayur Pabba, Aditya Jain, Sahib Singh, Hassam Ali, Rakesh Vinayek, Ganesh Aswath, Neil Sharma, Sumant Inamdar, and Antonio Facciorusso. Impact of artificial intelligence on pancreaticobiliary endoscopy. *Cancers*, 17(3), 2025.

Hala Khasawneh, Anurima Patra, Naveen Rajamohan, Garima Suman, Jason Klug, Shounak Majumder, Suresh T. Chari, Panagiotis Korfiatis, and Ajit Harishkumar Goenka. Volumetric pancreas segmentation on computed tomography: Accuracy and efficiency of a convolutional neural network versus manual segmentation in 3d slicer in the context of interreader variability of expert radiologists. *Journal of Computer Assisted Tomography*, 46(6), 2022.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles, 2017.

Suna Özhan Oktar, Cem Yücel, Hakan Özdemir, Aslı Ulutürk, and Sedat Işık. Comparison of conventional sonography, real-time compound sonography, tissue harmonic sonography, and tissue harmonic compound sonography of abdominal and pelvic lesions. *American Journal of Roentgenology*, 181(5), 2003.

William Park, Amit Chawla, and Eileen M O'Reilly. Pancreatic cancer: a review. *JAMA*, 326(9):851–862, 2021.

Siva P. Raman, Karen M. Horton, and Elliot K. Fishman. Multimodality imaging of pancreatic cancer—computed tomography, magnetic resonance imaging, and positron emission tomography. *The Cancer Journal*, 18(6), 2012.

Olaf Ronneberger, Philipp Fischer, and Thomas Bro. U-net: Convolutional networks for biomedical image segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241, 2015.

Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty, 2018.

Kangwon Seo, Jung-Hyun Lim, Jeongwung Seo, Leang Sim Nguon, Hongeun Yoon, Jin-Seok Park, and Suhyun Park. Semantic segmentation of pancreatic cancer in endoscopic ultrasound images using deep learning approach. *Cancers*, 14(20), 2022.

Rebecca L Siegel, Angela N Giaquinto, and Ahmedin Jemal. Cancer statistics, 2024. *CA: A cancer journal for clinicians*, 74(1):12–49, 2024.

Shannon-Morgan Steele. A unified semantic segmentation and object detection framework for synthetic aperture sonar imagery. *International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, 2023.

Abhinav Suri, Pritam Mukherjee, Perry J. Pickhardt, and Ronald M. Summers. A comparison of ct-based pancreatic segmentation deep learning models. *Academic Radiology*, 31 (11):4538–4547, 2024.

Pavan Tummala, Omer Junaidi, and Banke Agarwal. Imaging of pancreatic cancer: An overview. *Journal of Gastrointestinal Oncology*, 2(3):168–174, 2011.

Juan Zhao, Jiahuan Wang, Yuanlong Gu, Xiaoyi Huang, and Linyou Wang. Diagnostic methods for pancreatic cancer and their clinical applications (review). *Oncology Letters*, 30(1), 2025.

Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. *DLMIA*, 2018.

Appendix A: Inside the Ground Truth Annulus (IGTP) Details

A.1 IGTP Algorithm

For both evaluation and reproducibility, we provide the algorithmic implementation of our novel evaluation metric, IGTP. The IGTP metric evaluates the clinical reliability of a model's segmentation by measuring how much of the predicted mask's boundary falls within an acceptable tolerance zone of the ground truth (GT) mask. The algorithm takes the binary GT mask M_{GT}, the prediction mask M_{Pred}, and a kernel size K as inputs. For our experiments, K was set to 20 pixels. The pseudocode for this calculation is detailed in Algorithm 1.

Algorithm 1: Calculation of Inside Ground Truth Annulus Percentage (IGTP)

Input: Ground Truth Mask M_{GT}, Prediction Mask M_{Pred}, Kernel Size K
Output: Inside Ground Truth Annulus Percentage ($IGTP$)

`// Step 1: Define the acceptable tolerance zone (GT Annulus)`
$SE \leftarrow \text{SquareStructuringElement}(K)$;
$M_{GT\_dilated} \leftarrow \text{BinaryDilation}(M_{GT}, SE)$;
$M_{GT\_eroded} \leftarrow \text{BinaryErosion}(M_{GT}, SE)$;
$Annulus_{GT} \leftarrow M_{GT\_dilated} \wedge \neg M_{GT\_eroded}$;

`// Step 2: Isolate the 1-pixel wide prediction boundary`
$Border_{Pred} \leftarrow \text{FindInnerBoundaries}(M_{Pred})$;
$N_{total} \leftarrow \text{CountNonZero}(Border_{Pred})$;

if $N_{total} == 0$ **then**
  return NaN ; `// Handle empty predictions`
else
  `// Step 3: Calculate adherence within the annulus`
  $Pixels_{inside} \leftarrow Border_{Pred} \wedge Annulus_{GT}$;
  $N_{inside} \leftarrow \text{CountNonZero}(Pixels_{inside})$;
  $IGTP \leftarrow \left(\frac{N_{inside}}{N_{total}}\right) \times 100$;
  return $IGTP$;
end

A.2 Sensitivity Analysis of IGTP Kernel Size

To ensure that the IGTP metric is mathematically stable and not overly fitted to a specific hyperparameter, we conducted a sensitivity analysis across varying kernel sizes K. The kernel size directly dictates the thickness of the acceptable tolerance zone, or the annulus, surrounding the ground truth boundary.

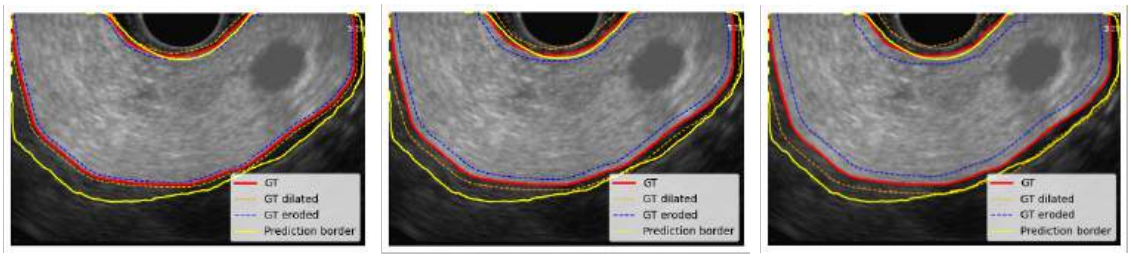
As shown in Table 6 and visualized in Figure 5, a smaller kernel ($K = 10$) severely penalizes minor boundary deviations, while a larger kernel ($K = 30$) provides a more forgiving clinical margin. The IGTP scores when using a kernel size of 30 start climbing up toward $60 - 65\%$ for the upper half boundaries. If the annulus is too wide, it starts

rewarding models that make relatively big hallucinations, eventually losing the ability to penalize models for being used in clinical settings.

Critically, the relative performance ranking of the evaluated architectures (nnU-Net > U-Net > U-Net++ > Swin U-Net) is perfectly preserved across all evaluated kernel sizes. We selected $K = 20$ as an intermediate methodological operating point between the stricter $K = 10$ and more permissive $K = 30$ settings. The consistent architecture ranking and upper-lower disparity across all three settings indicate that the principal conclusions are not dependent on one tolerance width. Clinical validation would be required before interpreting any kernel size as an acceptable clinical boundary margin.

Table 6: Sensitivity Analysis of IGTP Scores (Full, Upper, and Lower Half) Across Different Kernel Sizes (K)

Model	**Region**	$K = 10$	$K = 20$ **(Ours)**	$K = 30$
nnU-Net	**Full**	**26.9 ± 13.8**	**45.5 ± 19.0**	**57.3 ± 21.0**
	Upper	**34.5 ± 17.6**	**54.3 ± 20.6**	**65.1 ± 20.5**
	Lower	**18.2 ± 18.3**	**35.6 ± 28.1**	**48.6 ± 32.1**
U-Net	Full	24.3 ± 13.0	42.5 ± 18.7	53.8 ± 20.7
	Upper	32.7 ± 18.0	53.1 ± 21.8	64.1 ± 21.6
	Lower	16.2 ± 16.8	32.1 ± 26.2	43.6 ± 30.4
U-Net++	Full	23.3 ± 12.5	40.7 ± 17.9	51.9 ± 20.4
	Upper	31.7 ± 17.9	51.8 ± 21.2	62.8 ± 21.3
	Lower	15.4 ± 15.4	30.6 ± 24.8	42.2 ± 29.1
Swin U-Net	Full	19.1 ± 10.9	34.7 ± 16.6	46.2 ± 19.2
	Upper	27.5 ± 16.3	47.1 ± 20.9	59.6 ± 21.2
	Lower	12.4 ± 13.2	25.3 ± 22.6	36.1 ± 27.6



$K = 10$ $K = 20$ $K = 30$

Figure 5: Ground truth (GT) annulus annotations with varying kernel sizes ($K = 10, 20, 30$). The dilated (orange) and eroded (blue) boundaries show the range of the boundary tolerance zone in manual segmentation.

Relationship to Established Surface Metrics

IGTP is an asymmetric, precision-like boundary-adherence measure that quantifies the proportion of predicted contour pixels falling within a ground-truth-centered tolerance region. In contrast, normalized Surface Dice and symmetric surface-distance metrics evaluate correspondence in both directions and therefore also penalize portions of the reference surface that are not recovered by the prediction.

The primary distinction of IGTP is its directional decomposition into upper- and lower-boundary adherence. This decomposition is designed to localize anatomically asymmetric failure patterns that may be obscured by a single global symmetric surface score.

Appendix B: Evidential Training and Uncertainty Map Generation

B.1 Evidence Parameterization

For each pixel, the evidential segmentation head produces logits z_k, which are transformed into non-negative evidence using

$$e_k = \text{Softplus}(z_k).$$

Dirichlet concentration parameters are then defined as

$$\alpha_k = e_k + 1, S = \sum_{k=1}^{C} \alpha_k, \hat{p_k} = \frac{\alpha_k}{S}.$$

These concentration parameters are produced during both training and inference. They are not applied only as an inference-time transformation to a conventionally trained segmentation model.

B.2 Evidential Training Objective

The evidential network was optimized using an expected cross-entropy loss under the predicted Dirichlet distribution together with KL regularization:

$$\mathcal{L}_{\text{EDL}} = \mathcal{L}_{\text{ECE}} + \lambda \mathcal{L}_{\text{KL}},$$

where

$$\mathcal{L}_{\text{ECE}} = \frac{1}{|\Omega|} \sum_{i \in \Omega} \sum_{k=1}^{C} y_{ik} [\psi(S_i) - \psi(\alpha_{ik})],$$

and

$$\mathcal{L}_{\text{KL}} = \frac{1}{|\Omega|} \sum_{i \in \Omega} D_{\text{KL}} \left[\text{Dir}(\boldsymbol{\alpha}_i) \parallel \text{Dir}(\mathbf{1}) \right].$$

The KL coefficient was fixed at $\lambda = 10^{-3}$. Unlike formulations that suppress the ground-truth concentration before regularization, our implementation applies the KL term directly to the complete predicted concentration vector α_i.

Deep supervision was enabled by calculating the evidential loss at multiple decoder resolutions. Resolution-specific losses were weighted using normalized, exponentially decreasing weights, with the lowest-resolution output excluded from the total loss:

$$\mathcal{L}_{\text{total}} = \sum_{h=1}^{H} w_h \mathcal{L}_{\text{EDL}}^{(h)}.$$

Models were trained for 100 epochs using the AdamW optimizer with an initial learning rate of 3×10^{-4}, weight decay of 3×10^{-4}, and cosine-annealing learning-rate scheduling to a minimum learning rate of 3×10^{-5}. Training used 250 iterations per epoch, 50 validation iterations per epoch, and a batch size of 11.

B.3 Computational Timing

The calculation of the Dirichlet statistics and uncertainty maps added 14.24 ± 3.84 ms per image on an NVIDIA Tesla T4. This timing excludes preprocessing, file export, and disk I/O and therefore measures only the additional uncertainty-computation overhead. The present work evaluates offline or near-real-time feasibility rather than a prospectively deployed live endoscopy system.

B.4 EDL Analysis on External Data

To examine the cross-site behavior of the uncertainty maps, we applied the post-hoc EDL analysis of uncertainty maps generated by the evidentially trained model to our trained nnU-Net on the external dataset, representing a domain shift from our CAPS training cohort. Notably, the uncertainty patterns diverge markedly between the datasets. On CAPS data (Figure 4), high-performing cases exhibit well-localized, boundary-concentrated uncertainty with low interior AU. In contrast, on the external dataset as shown in Figure 6, AU is substantially more diffuse and widespread throughout the pancreatic region, manifesting as bright yellow distributions across the entire segmentation mask. This increased uncertainty dispersion reflects the domain shift and potentially greater inherent measurement uncertainty in the external cohort.

Across both datasets, elevated uncertainty qualitatively co-localized with regions of poor boundary adherence. This cross-site consistency supports further evaluation of uncertainty as a potential review-prioritization signal. However, the present retrospective study does not establish a clinically validated operating threshold or real-time endoscopy workflow.

Appendix C: Statistical and Uncertainty Results

This appendix provides the complete statistical results supporting the architecture comparison, directional boundary analysis, and uncertainty evaluation. All reported model comparisons were paired because each architecture was evaluated on the same set of images. Unless otherwise specified, two-sided Wilcoxon signed-rank tests were used because the metric distributions deviated significantly from normality.

For comparisons among the four segmentation architectures, six pairwise comparisons were possible for each metric. We therefore applied a Bonferroni-corrected significance

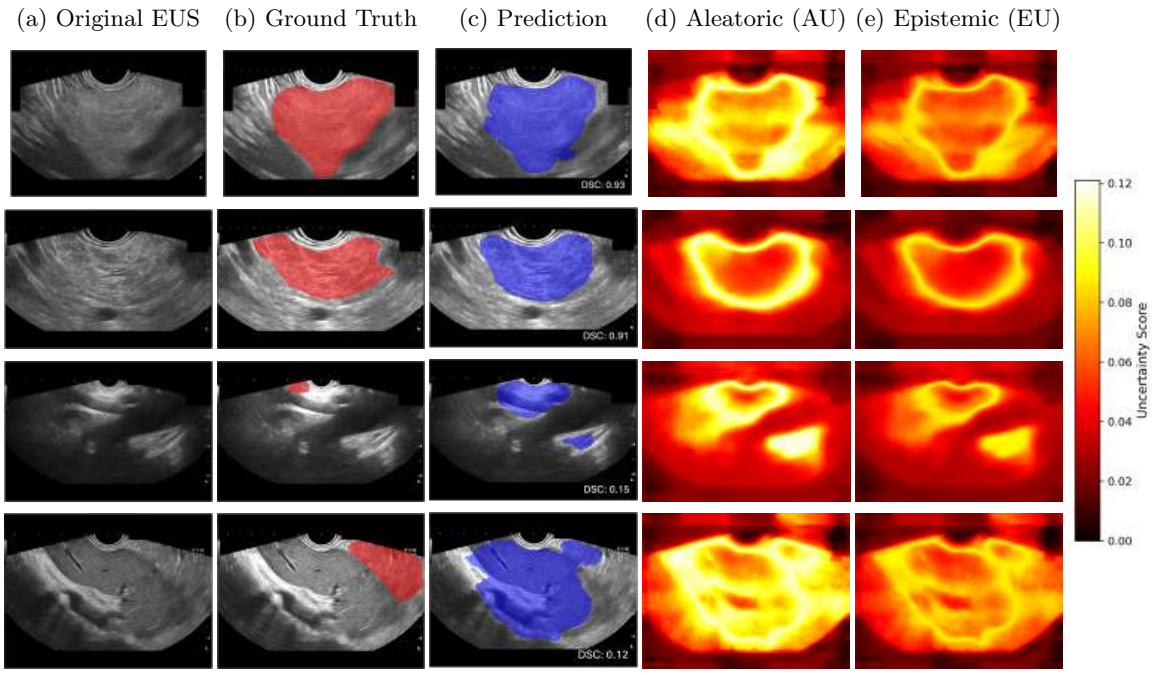


Figure 6: The top two rows display high-performing cases (DSC > 0.90), and the bottom two rows display failure cases where obscure lower boundaries result in high uncertainty (yellow regions) and corresponding poor boundary adherence.

threshold of

$$\alpha_{\text{corrected}} = \frac{0.05}{6} = 0.0083.$$

An approximate standardized Wilcoxon effect size was calculated as

$$r = \frac{Z}{\sqrt{N}}$$

where Z was reconstructed from the two-sided Wilcoxon p-value and signed according to the direction of the median paired difference. N denotes the number of paired images included in the comparison. Positive values indicate that the first model generally produced higher metric values, whereas negative values indicate that the first model generally produced lower metric values. For HD95, lower metric values indicate better performance.

C.1 Paired Architecture Comparisons

Table 7 reports paired Wilcoxon signed-rank comparisons between nnU-Net and each alternative architecture. All comparisons remained statistically significant after Bonferroni correction. nnU-Net achieved significantly higher DSC and IoU than the alternative architectures. U-Net achieved the lowest mean HD95. The paired HD95 results are reported in Table 7.

Table 7: Paired Wilcoxon signed-rank comparisons between nnU-Net and the alternative architectures. The reported p-values are unadjusted; statistical significance after Bonferroni correction was assessed using $\alpha_{\text{corrected}} = 0.0083$.

Metric	Comparison	nnU-Net mean	Comparator mean	Mean difference	p-value	Effect r	Bonferroni sig.
DSC	nnU-Net vs. U-Net	0.811	0.794	+0.017	1.26×10^{-12}	0.245	Yes
DSC	nnU-Net vs. U-Net++	0.811	0.790	+0.021	9.99×10^{-18}	0.296	Yes
DSC	nnU-Net vs. Swin U-Net	0.811	0.778	+0.033	8.84×10^{-30}	0.391	Yes
IoU	nnU-Net vs. U-Net	0.698	0.676	+0.022	2.66×10^{-13}	0.252	Yes
IoU	nnU-Net vs. U-Net++	0.698	0.670	+0.028	8.91×10^{-19}	0.305	Yes
IoU	nnU-Net vs. Swin U-Net	0.698	0.653	+0.045	9.77×10^{-31}	0.398	Yes
HD95	nnU-Net vs. U-Net	79.0	75.9	+3.1	3.20×10^{-11}	−0.229	Yes
HD95	nnU-Net vs. U-Net++	79.0	80.0	−1.0	2.70×10^{-17}	−0.292	Yes
HD95	nnU-Net vs. Swin U-Net	79.0	81.7	−2.7	5.72×10^{-16}	−0.279	Yes

The largest differences in overlap performance were observed between nnU-Net and Swin U-Net, with effect sizes of $r = 0.391$ for DSC and $r = 0.398$ for IoU. The corresponding HD95 difference was -2.7 in favor of nnU-Net.

C.2 IGTP Comparisons Between nnU-Net and Baseline Models

Table 8 reports paired comparisons between nnU-Net and each baseline architecture. nnU-Net had significantly higher full-boundary IGTP than all three baseline architectures. The largest difference was observed against Swin U-Net, with a mean difference of 10.81 percentage points.

For upper-boundary IGTP, the difference between nnU-Net and U-Net did not remain significant after correction for multiple comparisons ($p = 0.0436$). In contrast, nnU-Net retained significantly higher upper-boundary IGTP than U-Net++ and Swin U-Net. All lower-boundary comparisons remained significant after Bonferroni correction.

Table 8: Paired Wilcoxon signed-rank comparisons of IGTP between nnU-Net and the alternative architectures. Statistical significance was assessed using $\alpha_{\text{corrected}} = 0.0083$.

Region	Comparison	nnU-Net mean	Comparator mean	Mean difference	p-value	Bonferroni sig.
Full	nnU-Net vs. U-Net	45.5	42.5	+3.0	4.02×10^{-8}	Yes
Full	nnU-Net vs. U-Net++	45.5	40.7	+4.8	1.16×10^{-17}	Yes
Full	nnU-Net vs. Swin U-Net	45.5	34.7	+10.8	3.94×10^{-55}	Yes
Upper	nnU-Net vs. U-Net	54.3	53.1	+1.2	0.0436	No
Upper	nnU-Net vs. U-Net++	54.3	51.8	+2.4	9.23×10^{-6}	Yes
Upper	nnU-Net vs. Swin U-Net	54.3	47.1	+7.2	3.63×10^{-21}	Yes
Lower	nnU-Net vs. U-Net	35.6	32.1	+3.5	8.59×10^{-5}	Yes
Lower	nnU-Net vs. U-Net++	35.6	30.6	+5.0	1.59×10^{-7}	Yes
Lower	nnU-Net vs. Swin U-Net	35.6	25.3	+10.2	1.05×10^{-21}	Yes

Note. Mean differences are expressed in IGTP percentage points and are calculated as nnU-Net minus the comparison model.

C.3 Paired Upper and Lower-Boundary IGTP Analysis

To test whether the observed lower-boundary degradation was statistically significant, upper and lower-boundary IGTP values were compared within each architecture. Four directional comparisons were performed, producing a Bonferroni-corrected significance threshold of

$$\alpha_{\text{directional}} = \frac{0.05}{4} = 0.0125. \tag{1}$$

As shown in Table 9, upper-boundary IGTP was significantly higher than lower-boundary IGTP for every architecture. The relative decrease ranged from 34.6% for nnU-Net to 46.4% for Swin U-Net.

Table 9: Paired upper- versus lower-boundary IGTP comparisons. Positive mean gaps indicate superior adherence at the upper pancreatic boundary.

Model	Upper mean ± SD	Lower mean ± SD	Mean gap	Relative drop	p-value	Effect r	Bonferroni sig.
nnU-Net	54.3 ± 20.6	35.6 ± 28.1	18.8	34.6%	2.71×10^{-53}	0.532	Yes
U-Net	53.1 ± 21.8	32.1 ± 26.2	21.1	39.7%	2.18×10^{-63}	0.583	Yes
U-Net++	51.8 ± 21.2	30.6 ± 24.8	21.2	40.9%	5.11×10^{-67}	0.599	Yes
Swin U-Net	47.1 ± 20.9	25.3 ± 22.6	21.8	46.4%	3.35×10^{-72}	0.621	Yes

Note. The mean gap is calculated as upper-boundary IGTP minus lower-boundary IGTP. IGTP values and mean gaps are reported in percentage points.

The directional differences were substantially larger than most between-architecture differences. This finding indicates that boundary location and image quality exerted a stronger influence on adherence than incremental differences among the evaluated architectures.

C.4 Quantitative Uncertainty Evaluation

Quantitative uncertainty analysis was conducted among the images with complete segmentation and uncertainty outputs. The primary image-level uncertainty score was the 95th percentile of epistemic uncertainty within the predicted pancreatic region. Spearman rank correlation was used because the uncertainty and segmentation metrics were not assumed to be normally distributed.

Six correlation tests were performed. The corresponding Bonferroni threshold was

$$\alpha_{\text{uncertainty}} = \frac{0.05}{6} = 0.0083. \tag{2}$$

Epistemic uncertainty showed its strongest association with DSC and IoU ($\rho = -0.471$ for both metrics). Higher uncertainty was also associated with greater HD95 error ($\rho = 0.335$) and poorer full- and upper-boundary IGTP. The association with lower-boundary IGTP was weaker ($\rho = -0.112$) and did not remain significant after correction across the six correlation analyses.

C.5 Uncertainty-Based Failure Detection

To assess whether epistemic uncertainty could identify segmentation failures, failure cases were defined as images within the bottom quartile of DSC performance. The 95th-percentile

Table 10: Spearman associations between the 95th-percentile epistemic uncertainty within the predicted pancreas and segmentation-quality measurements.

Uncertainty summary	Quality metric	Spearman ρ	p-value	Bonferroni sig.
P95 EU within prediction	DSC	-0.471	1.35×10^{-31}	Yes
P95 EU within prediction	IoU	-0.471	1.35×10^{-31}	Yes
P95 EU within prediction	HD95	$+0.335$	8.48×10^{-16}	Yes
P95 EU within prediction	Full IGTP	-0.207	9.74×10^{-7}	Yes
P95 EU within prediction	Upper IGTP	-0.240	1.24×10^{-8}	Yes
P95 EU within prediction	Lower IGTP	-0.112	0.00864	No

Note. EU: epistemic uncertainty. P95: 95th percentile. Negative correlations with DSC, IoU, and IGTP indicate that greater uncertainty was associated with poorer segmentation quality. A positive correlation with HD95 indicates that greater uncertainty was associated with larger boundary-distance error.

epistemic uncertainty was then used as a continuous failure-detection score. Both uncertainty measured throughout the predicted pancreas and uncertainty localized to the predicted boundary were evaluated.

Table 11: Uncertainty-based detection of bottom-quartile DSC failures.

Uncertainty score	Failure definition	AUROC	AUPRC
P95 EU within predicted pancreas	Bottom-quartile DSC	0.741	0.568
P95 EU along predicted boundary	Bottom-quartile DSC	0.715	0.517

Note. AUROC: area under the receiver operating characteristic curve. AUPRC: area under the precision–recall curve. EU: epistemic uncertainty.

The 95th-percentile epistemic uncertainty within the predicted pancreas provided the strongest failure-detection performance, with an AUROC of 0.741 and an AUPRC of 0.568. Boundary-localized uncertainty also contained meaningful failure-detection information, achieving an AUROC of 0.715 and an AUPRC of 0.517. These results demonstrate that the evidentially learned uncertainty estimates provide quantitative information about segmentation reliability beyond qualitative spatial co-localization.

C.6 Summary of Statistical Evidence

The statistical analyses support three principal findings. First, nnU-Net achieved significantly higher DSC and IoU than all three alternative architectures. HD95 also differed significantly between nnU-Net and each alternative architecture. However, U-Net achieved the lowest mean HD95, while nnU-Net achieved lower mean HD95 than U-Net++ and Swin U-Net. Second, all four architectures exhibited a substantial and statistically significant reduction in boundary adherence from the upper to the lower pancreatic margin. Third, EU was meaningfully associated with segmentation quality and provided moderate discrimination of bottom-quartile DSC failures. Together, these findings support the use of uncertainty-aware and anatomically localized evaluation for identifying clinically important segmentation failures that may not be apparent from global overlap metrics alone.