Ashim Dhor

+91 7086482909 ashimdhor2003@gmail.com LinkedIn ashimdhor.github.io

Education

Indian Institute of Science Education and Research Bhopal Madhya Pradesh, India
BS-MS in Data Science and Engineering, CPI: 6.85/10 *Dec. 2021 – Present*

Publications

Ashim Dhor, Abhirup Banerjee, Tanmay Basu. **TUNE++: Topology-Guided Uncertainty Estimation for Reliable 3D Medical Image Segmentation** *Accepted at the MIDL 2026 Conference.*

Ashim Dhor, Emily Das, Rasel Mondal, Sweta Azad, Vaishali Walke, Shakti Kumar Yadav, Abhirup Banerjee, Tanmay Basu. **HISTO-UNet: Histopathology Image Segmentation Using Topology-Aware UNet with Dual Uncertainty Quantification**. *IEEE International Symposium on Biomedical Imaging (ISBI)*, 2026 (accepted).

Ashim Dhor, Smily Bharadwaj. **Topo-GraT: Learning to Grow with Causal Graph Transformers** *Accepted at the 40th AAAI Conference on Artificial Intelligence (Student Abstract Track).*

Ashim Dhor.et al **MOSAIC: Morphologically Orchestrated Specialized Agents for Interpretable Cancer Diagnosis** *IEEE International Conference on Healthcare Informatics (ICHI), 2026 - Young Scholar Symposium* ***(accepted, Lightning Talk)***

Ashim Dhor, Emily Das. **An Uncertainty-Aware Deep Learning Model for Gland Segmentation in Prostate Histopathology Images** *Accepted at the AAAI 2026 Empowering Global South AI Community Activity.*

Innan, N., Sawaika, A., **Dhor, A.**, Dutta, S., Thota, S., Gokal, H., Patel, N., Khan, M. A., Theodonis, I., and Bennai, M. (2024). **Financial fraud detection using quantum graph neural networks.** *Quantum Machine Intelligence*

Research Experience

Master's Thesis, IISER Bhopal *July 2025 - Present*
Biomedical Data Science Lab, PI: Dr. Tanmay Basu

- Conducting joint research with clinicians at Jawaharlal Nehru Cancer Hospital and Research Centre, Bhopal, and All India Institute of Medical Sciences Bhopal on whole slide image (WSI) analysis to build a large, high-quality dataset for Head and Neck Cancer, with ongoing work on Breast, Lung, and Ovarian Cancer datasets.
- Developing an interpretable multi-agent learning framework for whole-slide histopathology, where concept-specialized vision models collaboratively analyze multi-resolution tissue patches, resolve diagnostic disagreement through iterative discussion, and explicitly flag uncertain cases for human review.

Bioinformatics Scientist Intern *December, 2025 - Present*
Elucidata Data Consulting Pvt. Ltd. *Hybrid*

- Aiming to develop an AI agentic system for flow cytometry data analysis that automates and streamlines the work clinicians reducing manual gating effort providing interpretable, evidence-backed results. Developing an end-to-end Python pipeline covering FCS file ingestion, QC gating, etc across multi-sample heterogeneous datasets.
- Building a supervised Random Forest classifier and an unsupervised contrastive learning encoder that learns biologically meaningful cell population embeddings without manual labels. Conducting systematic data quality characterisation of an 18-sample clinical dataset and generating publication-quality diagnostic visualisations.

BS Thesis, IISER Bhopal *Jan. - Apr. 2025*
Biomedical Data Science Lab, PI: Dr. Tanmay Basu

- Evaluated baseline models - UNet, LinkNet, and UNet++ for gland segmentation in prostate histopathology.
- Developed a dual uncertainty framework by combining epistemic uncertainty via Monte Carlo Dropout with a novel aleatoric uncertainty approach that injects Gaussian noise into model logits.
- Performed Uncertainty-Filtered Calibration Analysis, sanity checks using label/prediction flipping, and probabilistic evaluation. Found UNet++ with data uncertainty worked best.

Presentation , Report

Research Internship, QWorld *June - Sept. 2023*
Fraud Detection using Quantum Machine Learning

- Architected a Quantum Graph Neural Network (QGNN) that uses Topological Data Analysis (TDA) to process transactions as graphs and a Variational Quantum Circuit (VQC) to enhance feature learning.
- Established a performance baseline using a classical GraphSAGE model, which achieved 92.4% accuracy and a 0.77 Precision-Recall AUC on the fraud detection task.
- Demonstrated a quantum advantage by validating that the proposed QGNN significantly outperformed the classical benchmark, achieving a superior accuracy of 94.5% and a Precision-Recall AUC of 0.85.

Our team received the Best Project and Best Presentation Award, Presentation

Achievements

Selected for the Rising Star Award
IEEE International Conference on Healthcare Informatics (ICHI), 2026 - Young Scholar Symposium

Awarded the MICCAI Society Membership Grant 2026
Medical Image Computing and Computer Assisted Intervention Society

The 40th Annual AAAI Conference on Artificial Intelligence, 2026
*Student **Scholarship** Travel Grant Recipient*

13th ACM iKDD International Conference on Data Science, 2025
*Student **Scholarship** Travel Grant Recipient*

Selected for presentation as a 3MT Contest Finalist
Empowering Global South AAI - AAAI'26 Community Activity

MFF CARE Conference on Data Science & Artificial Intelligence, 2025
*Student **Scholarship** Travel Grant Recipient*

National Bio Entrepreneurship Competition (NBEC) 2025
Selected for the second round for HistoAI, an ideathon concept for an AI-powered diagnostic tool that quantifies uncertainty in cancer segmentation

Received Best Project and Best Presentation Prize
QIntern Internship

Position of Responsibility

The 7th Annual Conference on Health, Inference, and Learning
*Selected as a **Reviewer***

The 40th Annual AAAI Conference on Artificial Intelligence
*Selected as a **Mentor** for Undergraduate Consortium Program*

The AHLI Machine Learning for Health (ML4H) Symposium 2025
*Selected as a **Reviewer***

The 40th Annual AAAI Conference on Artificial Intelligence
*Selected as a **Reviewer***

Sports Council
Sports Secretary, IISER Bhopal — *Aug. 2023 – Aug. 2024*

Institute Career Development and Placement Council (ICDPC)
Student Placement Head, IISER Bhopal — *June 2022 - May 2023*

Technical Skills

Languages: Python, C, SQL, MATLAB, Qiskit, LaTeX
Core ML / DL: PyTorch, PyTorch Lightning, torchvision, timm, scikit-learn
Deep Learning & NLP: Hugging Face (Transformers, PEFT), NLTK, Gensim, spaCy, stanza
GenAI & Research: Multi-Agent Systems (AutoGen), LoRA/QLoRA, G-Eval
Computer Vision & Imaging: OpenCV, scikit-image, MONAI, OpenSlide
Compute & Ops: NVIDIA GPUs (H200/A100), Weights & Biases, Distributed Training

TUNE++: Topology-Guided Uncertainty Estimation for Reliable 3D Medical Image Segmentation

Ashim Dhor[1] ASHIMDHOR2003@GMAIL.COM
Abhirup Banerjee[2] ABHIRUP.BANERJEE@ENG.OX.AC.UK
Tanmay Basu[1] TANMAY@IISERB.AC.IN
[1] *Indian Institute of Science Education and Research Bhopal, India*
[2] *Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, UK*

Editors: Accepted for publication at MIDL 2026

Abstract

Deep learning models for medical image segmentation lack mechanisms to assess their own reliability, leading to two critical failures: they provide no uncertainty estimates to distinguish confident predictions from error-prone ones, and often produce anatomically implausible segmentations or incorrect connectivity that violate known structural constraints. We observe that uncertainty and topology are intrinsically linked and anatomically complex regions naturally exhibit higher prediction uncertainty, while uncertain predictions require stronger enforcement of structural constraints. Building on this insight, we propose TUNE++, a unified framework that jointly learns segmentation, uncertainty quantification, and topology preservation through a novel Topology-Uncertainty aware Paired Attention (TUPA) mechanism. Our method decomposes uncertainty into aleatoric and epistemic components while simultaneously enforcing anatomical correctness through persistent homology-based constraints. A key innovation is our topology-uncertainty alignment loss that minimizes the discrepancy between predicted total uncertainty and a topological complexity score computed from organ boundaries, multi-organ junction counts, and critical points extracted from persistence diagrams, teaching the model to be uncertain precisely where anatomical structure is geometrically complex. Our empirical results demonstrate that joint modeling of TUNE++ produced enhanced segmentation accuracy, well-calibrated uncertainty estimates that successfully identify errors, substantial reduction in topological violations, and learned confidence that correlates strongly with anatomical complexity. Our source code will be available at: https://github.com/AshimDhor/tune_plus_plus.

Keywords: Medical image segmentation, Uncertainty quantification, Topology, Homology.

1. Introduction

Deep learning models, including recent transformer-based architectures (Cao et al., 2022; Hatamizadeh et al., 2022; Tang et al., 2022), achieve strong segmentation accuracy but remain difficult to deploy clinically due to limited reliability assessment. Without uncertainty quantification, clinicians cannot distinguish high-confidence predictions from error-prone cases, while models frequently generate anatomically implausible segmentations – organs with holes or disconnected components – violating known structural constraints. Recent work has separately addressed these issues. Uncertainty quantification methods (Roy et al., 2019; Wang et al., 2019; Jungo and Reyes, 2019) estimate prediction confidence but ignore anatomical structures. Topology-preserving approaches (Hu et al., 2019; Shit et al., 2021; Clough et al., 2020) enforce structural correctness but provide no uncertainty estimates.

We observe that these aspects are fundamentally interconnected: topologically complex regions are inherently more difficult to segment and should exhibit elevated uncertainty, while uncertain predictions require stronger enforcement of topological constraints. A detailed literature review is provided in Appendix A.1

We introduce **TUNE++** (**T**opology and **UN**certainty-aware **E**fficient transformers), a unified framework that jointly learns segmentation, uncertainty quantification, and topology preservation. Our contribution is the **T**opology-**U**ncertainty aware **P**aired **A**ttention (TUPA) block, which extends efficient paired attention (EPA) (Shaker et al., 2024) with two innovations: a topology-aware attention branch that focuses on anatomically critical regions identified through persistent homology, and an uncertainty-guided adaptive fusion mechanism that dynamically weights spatial, channel, and topological features based on prediction confidence. We introduce a topology-uncertainty alignment loss that enforces correlation between uncertainty estimates and topological complexity, ensuring the model exhibits higher uncertainty at boundaries, junctions, and regions with complex anatomical structure. We evaluate TUNE++ on three benchmarks spanning different anatomical regions (Synapse, ACDC, BTCV), achieving state-of-the-art segmentation accuracy (mean DSC 89.4%), topological error reduction (72%, Betti 1.94→0.54), and superior uncertainty calibration (ECE 0.043), demonstrating that joint topology-uncertainty modeling produces synergistic improvements over approaches addressing these objectives in isolation.

2. Methods

Medical image segmentation faces two linked challenges: models produce anatomically implausible outputs (disconnected organs, spurious holes) because they lack structural constraints, and they provide no confidence estimates to distinguish reliable from error-prone predictions. We observe that anatomically complex regions-boundaries, junctions, irregular structures-are precisely where models should express higher uncertainty, while uncertain predictions should trigger stronger structural enforcement. We propose TUNE++ (**T**opology and **UN**certainty-aware **E**fficient transformers), a unified framework that jointly learns segmentation, uncertainty quantification, and topology preservation by explicitly coupling these objectives. Given a 3D medical image $\mathbf{x} \in \mathbb{R}^{H\times W\times D}$ (height H, width W, depth D), our model learns a mapping that produces three outputs: (1) segmentation mask $\mathbf{y} \in \{0,1\}^{H\times W\times D\times C}$ for C anatomical classes, (2) uncertainty estimates decomposed into aleatoric uncertainty $\boldsymbol{\sigma}_a^2 \in \mathbb{R}_+^{H\times W\times D\times C}$ (data uncertainty) and epistemic uncertainty $\boldsymbol{\sigma}_e^2 \in \mathbb{R}_+^{H\times W\times D\times C}$ (model uncertainty), and (3) topological descriptors encoding structural correctness through persistence diagrams. We formalize this as:

$$\theta^* = \arg\min_\theta \mathbb{E}_{(\mathbf{x},\mathbf{y}^*)} \left[\mathcal{L}_{\text{seg}} + \lambda_1 \mathcal{L}_{\text{topo}} + \lambda_2 \mathcal{L}_{\text{unc}} + \lambda_3 \mathcal{L}_{\text{calib}} + \lambda_4 \mathcal{L}_{\text{hier}}\right] \tag{1}$$

where $\mathcal{L}_{\text{seg}}$ ensures accurate segmentation, $\mathcal{L}_{\text{topo}}$ enforces anatomical structure, $\mathcal{L}_{\text{unc}}$ learns uncertainty decomposition, $\mathcal{L}_{\text{calib}}$ ensures well-calibrated confidence, and $\mathcal{L}_{\text{hier}}$ maintains multi-scale consistency. Our key innovation is the explicit coupling between topology and uncertainty through the TUPA attention mechanism, detailed in Section 2.2.

2.1. Architecture Overview

TUNE++ adopts a four-stage hierarchical architecture operating at spatial resolutions $\{H/4, H/8, H/16, H/32\}$ with progressively increasing channel dimensions $\{C_1, 2C_1, 4C_1, 8C_1\}$ where $C_1 = 32$ is the base channel count. This multi-scale design serves two purposes: coarse resolutions (H/32, H/16) capture global organ-level context and inter-organ relationships, while fine resolutions (H/8, H/4) preserve boundary details critical for accurate delineation. Figure 1 illustrates the complete architecture.

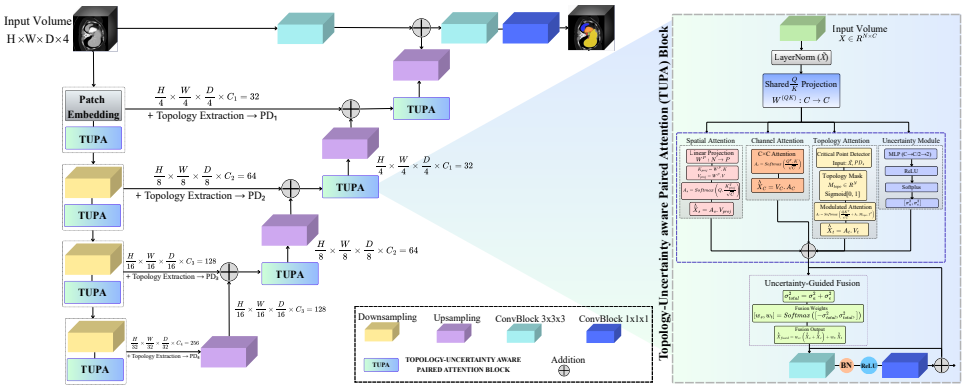

Figure 1: TUNE++ architecture. The encoder extracts multi-scale features while computing topological descriptors (persistence diagrams) at each stage. TUPA blocks integrate spatial, channel, and topology-aware attention. The decoder produces segmentation and uncertainty estimates through specialized heads.

2.1.1. Input Processing and Patch Embedding

Following Vision Transformer (ViT) design (Dosovitskiy et al., 2021) , we divide the input volume into non-overlapping 3D patches of size $(P_h, P_w, P_d) = (4, 4, 2)$. The asymmetric patch dimensions accommodate typical medical image anisotropy where through-plane (axial) resolution is coarser than in-plane resolution. This yields $N = \frac{HWD}{32}$ tokens, each linearly projected to $C_1 = 32$ dimensions with learnable positional embeddings encoding spatial structure. This tokenization converts the volumetric input into a sequence suitable for transformer processing while preserving 3D spatial relationships.

2.1.2. Encoder Architecture

Each of the four encoder stages contains three components working together to extract increasingly abstract features while tracking their topological properties. First, stride-2 convolution with $3 \times 3 \times 3$ kernels performs downsampling, reducing spatial dimensions by half while doubling channel count to progressively capture higher-level features. Second, our novel TUPA block (detailed in Section 2.2) processes these features through three parallel branches: spatial attention, channel attention, and topology-aware attention, and then fuses them using uncertainty-guided weighting, allowing the model to adaptively balance data-driven and structure-driven information. Third, at each encoder stage s, we extract

topological features from the feature maps using persistent homology that analyzes multi-scale structure by treating feature activations as height functions and tracking when topological features like connected components, holes, and voids appear and disappear across thresholds. This produces a persistence diagram $\text{PD}_s = \{(b_i, d_i)\}$, where each (birth, death) coordinate pair represents a topological feature and its persistence $|d_i - b_i|$ quantifies how significant it is. Long-lived features with large persistence correspond to true anatomical structures, while short-lived features represent noise. This topological information then guides the TUPA attention mechanism to focus on structurally critical regions, creating a feedback loop between topology extraction and feature processing.

2.1.3. Decoder Architecture and Output Heads

The decoder mirrors the encoder with four stages, each containing: $2 \times 2 \times 2$ transposed convolutions with stride 2 for upsampling, doubling spatial dimensions while halving channels, skip connections from corresponding encoder stages concatenated before TUPA blocks (following UNet design (Ronneberger et al., 2015)), and TUPA blocks for feature refinement. The final decoder output at resolution $H/2 \times W/2 \times D/2$ with $C_1 = 32$ channels feeds into four parallel heads that produce our different outputs. The segmentation head applies $\text{Conv}_{3\times3\times3}$ followed by $\text{Conv}_{1\times1\times1}$, then uses bilinear upsampling to reach original resolution and applies softmax to produce $\mathbf{y} \in \mathbb{R}^{H\times W\times D\times C}$. For uncertainty quantification, we have two separate heads: the aleatoric uncertainty head uses an MLP with architecture $256 \to 128 \to C$ followed by softplus activation to produce $\boldsymbol{\sigma}_a^2 \in \mathbb{R}_+^{H\times W\times D\times C}$, where softplus ensures the variance remains positive and captures inherent data ambiguity. The epistemic uncertainty head uses the same MLP structure to generate an initialization $\boldsymbol{\sigma}_{e,\text{init}}^2$, with the final epistemic uncertainty computed via Monte Carlo dropout at inference (detailed in Section 2.4). Finally, the topology descriptor head computes a persistence diagram on the final segmentation $\mathbf{y}$ for topology loss computation.

2.2. Topology-Uncertainty Aware Paired Attention (TUPA)

Standard self-attention in transformers has quadratic complexity $\mathcal{O}(N^2C)$ where N is the number of tokens and C is the channel dimension, making it prohibitively expensive for 3D medical volumes. Moreover, standard attention treats all spatial locations equally, failing to account for (1) structural importance (organ boundaries and junctions require more careful modeling than homogeneous regions), and (2) prediction confidence (uncertain regions should leverage stronger structural priors). TUPA addresses both limitations by: (a) decomposing attention into efficient spatial, channel, and topology branches, reducing complexity to $\mathcal{O}(NPC)$ where $P \ll N$, and (b) dynamically fusing these branches using predicted uncertainty to adaptively balance data-driven and structure-driven features.

2.2.1. Shared Query-Key Projection

Given input features $\mathbf{X} \in \mathbb{R}^{N\times C}$ where $N = h \cdot w \cdot d$ is the number of spatial tokens at the current scale, we first apply layer normalization $\tilde{\mathbf{X}} = \text{LayerNorm}(\mathbf{X})$ for training stability. We then compute shared query and key projections:

$$\mathbf{Q}_{\text{shared}} = \mathbf{K}_{\text{shared}} = \mathbf{W}^{QK}\tilde{\mathbf{X}} \in \mathbb{R}^{N\times C} \tag{2}$$

where $\mathbf{W}^{QK} \in \mathbb{R}^{C\times C}$ is a learnable projection. Sharing queries and keys across attention branches reduces parameters while enabling consistent attention patterns, following efficient attention designs (Shaker et al., 2024).

2.2.2. Efficient Spatial and Channel Attention

To avoid quadratic $\mathcal{O}(N^2)$ complexity, we project keys and values to lower dimension $P \ll N$, reducing spatial attention complexity to $\mathcal{O}(NPC)$:

$$\mathbf{K}_{\text{proj}} = (\mathbf{W}^P \mathbf{K}_{\text{shared}}^T)^T \in \mathbb{R}^{P\times C}, \tag{3}$$

$$\mathbf{V}_{\text{proj}}^{\text{spatial}} = (\mathbf{W}^P (\mathbf{V}_{\text{spatial}})^T)^T \in \mathbb{R}^{P\times C}, \tag{4}$$

$$\hat{\mathbf{X}}_{\text{spatial}} = \text{Softmax}\left(\frac{\mathbf{Q}_{\text{shared}}\mathbf{K}_{\text{proj}}^T}{\sqrt{C}}\right)\mathbf{V}_{\text{proj}}^{\text{spatial}}, \tag{5}$$

where $\mathbf{V}_{\text{spatial}} = \mathbf{W}^{V_{\text{spatial}}}\tilde{\mathbf{X}} \in \mathbb{R}^{N\times C}$ is the spatial value projection and $\mathbf{W}^P \in \mathbb{R}^{P\times N}$ performs learned dimensionality reduction. In parallel, channel attention models feature interdependencies by attending across the channel dimension rather than spatial locations, computing a channel-wise attention matrix that recalibrates feature importance:

$$\mathbf{A}_c = \text{Softmax}\left(\frac{\mathbf{Q}_{\text{shared}}^T\mathbf{K}_{\text{shared}}}{\sqrt{C}}\right) \in \mathbb{R}^{C\times C}, \tag{6}$$

$$\hat{\mathbf{X}}_c = \mathbf{V}_c\mathbf{A}_c, \tag{7}$$

where $\mathbf{V}_c = \mathbf{W}^{V_c}\tilde{\mathbf{X}}$ is the channel value projection. These two branches capture both spatial relationships between locations and semantic relationships between feature channels, providing complementary representations that are later fused based on predicted uncertainty.

2.2.3. Topology-Aware Attention Branch

The key innovation of TUPA is incorporating structural knowledge through topological attention. We identify anatomically critical regions using a Critical Point Detector:

$$\mathbf{M}_{\text{topo}} = \text{CriticalPointDetector}(\tilde{\mathbf{X}}, \text{PD}_s) \in \mathbb{R}^N \tag{8}$$

where PD_s is the persistence diagram extracted at encoder stage s. The detector (implemented as a 3-layer CNN) takes as input both the current features $\tilde{\mathbf{X}}$ and the topological descriptor PD_s, producing a scalar importance score for each spatial location. High scores are assigned to organ boundaries (transitions between tissues), multi-organ junctions, and critical points identified from the persistence diagram (locations where topological features appear or disappear in the filtration). We then compute topology-modulated attention:

$$\mathbf{A}_t = \text{Softmax}\left(\frac{\mathbf{Q}_{\text{shared}}\mathbf{K}_{\text{shared}}^T}{\sqrt{C}} + \lambda_t\mathbf{M}_{\text{topo}}\mathbf{1}^T\right) \in \mathbb{R}^{N\times N}, \tag{9}$$

$$\hat{\mathbf{X}}_t = \mathbf{A}_t\mathbf{V}_t, \tag{10}$$

where $\mathbf{V}_t = \mathbf{W}^{V_t}\tilde{\mathbf{X}} \in \mathbb{R}^{N\times C}$ is the topology value projection, and $\lambda_t = 0.3$ controls how strongly topological structure biases attention (determined via grid search over $\{0.1, 0.2, 0.3, 0.4, 0.5\}$ on validation data). The broadcast operation $\mathbf{M}_{\text{topo}}\mathbf{1}^T$ upweights attention to structurally critical regions identified by persistent homology.

2.2.4. Uncertainty-Guided Adaptive Fusion

Parallel to the three attention branches, we estimate voxel-wise uncertainty to guide how these branches should be combined:

$$[\boldsymbol{\sigma}_a^2, \boldsymbol{\sigma}_{e,\text{init}}^2] = \text{MLP}(\tilde{\mathbf{X}}) \in \mathbb{R}^{N\times 2} \tag{11}$$

where σ_a^2 captures aleatoric uncertainty and $\sigma_{e,\text{init}}^2$ initializes epistemic uncertainty. We compute total uncertainty as $\boldsymbol{\sigma}_{\text{total}}^2 = \boldsymbol{\sigma}_a^2 + \boldsymbol{\sigma}_{e,\text{init}}^2$. The key insight is that uncertain regions should rely more heavily on topological structure, while confident regions can rely on data-driven spatial/channel attention. We implement this through adaptive fusion weights:

$$[w_s, w_t] = \text{Softmax}([-\boldsymbol{\sigma}_{\text{total}}^2, \boldsymbol{\sigma}_{\text{total}}^2]) \in \mathbb{R}^{N\times 2} \tag{12}$$

where $w_s, w_t \in \mathbb{R}^N$ are per-voxel weights for spatial/channel versus topology attention. When uncertainty is low ($\sigma_{\text{total}}^2 \approx 0$), the softmax produces balanced weights $w_s \approx w_t \approx 0.5$, allowing data-driven features to dominate. When uncertainty is high ($\sigma_{\text{total}}^2 \gg 0$), topology weight increases ($w_t \to 1, w_s \to 0$), enforcing structural constraints where the model is uncertain. The three attention outputs are fused as:

$$\hat{\mathbf{X}}_{\text{fused}} = w_s \odot (\hat{\mathbf{X}}_s + \hat{\mathbf{X}}_c) + w_t \odot \hat{\mathbf{X}}_t \tag{13}$$

where $\odot$ denotes element-wise multiplication. Finally, convolutional refinement processes fused features:

$$\mathbf{X}_{\text{out}} = \text{Conv}_{1\times1\times1}(\text{BN}(\text{ReLU}(\text{Conv}_{3\times3\times3}(\hat{\mathbf{X}}_{\text{fused}})))). \tag{14}$$

This completes the TUPA block, which is applied at each encoder and decoder stage, progressively refining features with topology-uncertainty aware attention.

2.3. Training Objectives

Training TUNE++ requires balancing five complementary objectives, each targeting a distinct aspect of reliable segmentation. The total training objective is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{seg}} + \lambda_1\mathcal{L}_{\text{topo}} + \lambda_2\mathcal{L}_{\text{unc}} + \lambda_3\mathcal{L}_{\text{calib}} + \lambda_4\mathcal{L}_{\text{hier}} \tag{15}$$

$\mathcal{L}_{\text{seg}}$ ensures segmentation accuracy (unweighted, coefficient 1.0), $\mathcal{L}_{\text{topo}}$ enforces topological correctness ($\lambda_1 = 0.3$), $\mathcal{L}_{\text{unc}}$ learns uncertainty decomposition and calibration ($\lambda_2 = 0.2$), $\mathcal{L}_{\text{calib}}$ ensures confidence calibration ($\lambda_3 = 0.1$, serving as a regularizer), and $\mathcal{L}_{\text{hier}}$ maintains multi-scale topology consistency ($\lambda_4 = 0.15$, also a regularizer). This weight hierarchy reflects our task priorities: topology provides the strongest structural constraints preventing anatomically impossible outputs, uncertainty enables reliability assessment critical for clinical deployment, and calibration together with hierarchical consistency serve as a regularizer refining model behavior. Detailed analysis is given in Appendix A.5.

Segmentation Loss combines Dice and cross-entropy to handle class imbalance while providing dense gradients. Complete formulations are provided in Appendix A.3.1.

$$\mathcal{L}_{\text{seg}} = \mathcal{L}_{\text{Dice}} + \mathcal{L}_{\text{CE}} \tag{16}$$

Topology Preservation Loss ensures anatomically plausible structures by enforcing correct connectivity, preventing spurious holes, and avoiding fragmented organs. We combine three complementary topological constraints:

$$\mathcal{L}_{\text{topo}} = \mathcal{L}_{\text{PH}} + 0.5\mathcal{L}_{\text{Betti}} + 0.3\mathcal{L}_{\text{critical}}. \tag{17}$$

where $\mathcal{L}_{\text{PH}}$ measures multi-scale structural similarity through persistent homology (capturing when topological features appear and disappear across scales), $\mathcal{L}_{\text{Betti}}$ penalizes incorrect topological invariants, and $\mathcal{L}_{\text{critical}}$ ensures critical topological locations are correctly positioned. Complete formulations are provided in Appendix A.3.2.

Uncertainty Loss: We decompose total uncertainty into aleatoric and epistemic components, then align them with topological complexity:

$$\mathcal{L}_{\text{unc}} = \mathcal{L}_{\text{aleatoric}} + \mathcal{L}_{\text{epistemic}} + 0.5\mathcal{L}_{\text{align}}. \tag{18}$$

Aleatoric Uncertainty Loss learns heteroscedastic noise following (Kendall and Gal, 2017):

$$\mathcal{L}_{\text{aleatoric}} = \frac{1}{N_{\text{vox}}C} \sum_{i=1}^{N_{\text{vox}}} \sum_{c=1}^{C} \left(\frac{\|p_{i,c} - g_{i,c}\|^2}{2\sigma_{a,i,c}^2} + \log(\sigma_{a,i,c}^2 + \epsilon) \right). \tag{19}$$

The first term penalizes prediction error normalized by predicted noise, while the second term prevents the model from trivially minimizing loss by predicting infinite uncertainty. Epistemic Uncertainty Loss enforces consistency across multiple stochastic forward passes:

$$\mathcal{L}_{\text{epistemic}} = \text{KL}(p_{\text{single}} \| p_{\text{MC}}) = \frac{1}{N_{\text{vox}}C} \sum_{i,c} p_{i,c}^{\text{single}} \log \frac{p_{i,c}^{\text{single}}}{p_{i,c}^{\text{MC}}} \tag{20}$$

where p_{single} is a single forward pass prediction, and p_{MC} is the mean over dropout samples. This KL divergence encourages the learned epistemic initialization to approximate the true variability captured by MC dropout. Topology-Uncertainty Alignment Loss creates synergy by correlating prediction uncertainty with topological complexity:

$$\mathcal{L}_{\text{align}} = \frac{1}{N_{\text{vox}}} \sum_{i=1}^{N_{\text{vox}}} \|\sigma_{\text{total},i}^2 - C_{\text{topo},i}\|_2^2 \tag{21}$$

where the topological complexity score combines three components:

$$C_{\text{topo},i} = w_b B_i + w_j J_i + w_a A_i \tag{22}$$

with $B_i \in \{0, 1\}$ indicating boundaries (organ edges), $J_i \in \mathbb{Z}_+$ counting organ junctions (where 3+ structures meet), and $A_i \in [0, 1]$ measuring topological anomalies from persistence diagrams. The weights $w_b = 1.0$, $w_j = 2.0$, $w_a = 3.0$ reflect increasing topological severity: boundaries are baseline features, junctions involve multiple organs requiring stronger enforcement, and anomalies represent severe violations warranting highest penalty.

Calibration Loss ensures predicted confidence matches actual correctness probability through Expected Calibration Error (ECE) and Brier Score (BS):

$$\mathcal{L}_{\text{calib}} = \text{ECE} + 0.5 \cdot \text{Brier} = \sum_{m=1}^{M} \frac{|B_m|}{N_{\text{vox}}} |\text{acc}(B_m) - \text{conf}(B_m)| + \frac{0.5}{N_{\text{vox}}C} \sum_{i,c} (p_{i,c} - g_{i,c})^2 \tag{23}$$

where predictions are discretized into $M = 10$ confidence bins B_m, $\text{acc}(B_m)$ is the empirical accuracy within bin m, and $\text{conf}(B_m)$ is the average predicted confidence.

Hierarchical Topology Consistency Loss: We expect topology to remain stable across scales-downsampling shouldn't fundamentally change whether an organ appears as one piece or fragments:

$$\mathcal{L}_{\text{hier}} = \sum_{s=2}^{4} W_2(\text{PD}_s, \text{Downsample}(\text{PD}_{s-1})) \tag{24}$$

where PD_s is the persistence diagram at encoder stage s, and $\text{Downsample}(\cdot)$ recomputes the diagram at the next coarser scale, preventing scale-dependent artifacts.

2.4. Inference Protocol

At inference, we perform $K = 25$ stochastic forward passes with dropout enabled (Monte Carlo dropout (Gal and Ghahramani, 2016)), each producing prediction $\mathbf{y}_k$ and aleatoric estimate $\boldsymbol{\sigma}^2_{a,k}$. The final prediction is the mean output $\mathbf{y}_{\text{final}} = \frac{1}{K}\sum_{k=1}^{K} \mathbf{y}_k$. Aleatoric uncertainty is estimated as $\boldsymbol{\sigma}^2_a = \frac{1}{K}\sum_{k=1}^{K} \boldsymbol{\sigma}^2_{a,k}$, epistemic uncertainty is computed from predictive variance $\boldsymbol{\sigma}^2_e = \frac{1}{K}\sum_{k=1}^{K}(\mathbf{y}_k - \mathbf{y}_{\text{final}})^2$, and total uncertainty combines both: $\boldsymbol{\sigma}^2_{\text{total}} = \boldsymbol{\sigma}^2_a + \boldsymbol{\sigma}^2_e$. This decomposition allows distinguishing irreducible data ambiguity from model confidence, critical for clinical decision-making.

3. Experiments and Results

3.1. Datasets and Implementation

We evaluated TUNE++ on three public benchmark datasets: Synapse multiorgan segmentation CT scans (Landman et al., 2015) containing 30 volumetric scans containing 8 abdominal organs with an 18/12 train/test split following (Zhou et al., 2023), Automatic Cardiac Diagnosis Challenge (ACDC) (Bernard et al., 2018) (100 subjects, 70/10/20 - train/validation/test split) follwoing (Zhou et al., 2023), and BTCV abdominal CT (Landman et al., 2015) (50 subjects, standard 30/20 - train/test split). Implementation details including training hyperparameters and augmentation strategies are provided in Appendix A.2. Experiments are conducted on a NVIDIA H100 (95GB) GPU. We evaluate performance across three metric groups: segmentation accuracy - Dice Similarity Coefficient (DSC), Normalized Surface Distance (NSD), 95th percentile Hausdorff Distance (HD95), uncertainty quality (ECE and Brier Score) and topological correctness (Betti Error). We introduce **Topology-Aware Uncertainty Score** (TAUS), measuring correlation between predicted uncertainty and topological complexity:

$$\text{TAUS} = \text{Pearson}(\sigma^2_{\text{total}}, C_{\text{topo}}) \tag{25}$$

where C_{topo} is topological complexity score. TAUS $\in$ [-1, 1], with higher values indicating better uncertainty-anatomy alignment. Full metric definitions are given in Appendix A.4.

3.2. Results

Table 1 summarizes the Synapse multi-organ benchmark results, where TUNE++ achieves 89.3% mean DSC, establishing a new state-of-the-art with statistically significant gains over all baselines ($p < 0.001$). Beyond average accuracy, it offers substantially improved boundary quality, evidenced by an 89.1% NSD and HD95 reduced from 7.5mm (UNETR++) to 6.2mm. Performance gains are notable in challenging organs such as pancreas (84.2% DSC) and gallbladder (73.8%), while accuracy remains high for large organs (e.g., spleen 96.5%), demonstrating robustness across scale and anatomy. A comparison with baseline models and qualitative visual results are shown in Figure 2.

Table 1: Synapse dataset comparison. Best results in **bold**, second best underlined.

Method	Spl	RK	LK	Gal	Liv	Sto	Aor	Pan	Mean DSC↑	NSD↑ (%)	HD95↓ (mm)	Betti Err↓
U-Net	86.7	68.6	77.8	69.7	93.4	75.6	89.1	54.0	76.9	83.2	39.7	2.87
nnUNet	90.5	86.2	86.6	70.2	96.8	86.8	92.0	83.4	86.6	84.5	10.6	1.89
UNETR	86.7	85.6	85.6	56.3	94.6	70.5	89.8	60.5	78.4	76.6	18.6	2.41
Swin-UNETR	95.4	86.3	87.0	66.5	95.7	77.0	91.1	68.8	83.5	80.9	10.6	1.76
nnFormer	90.5	86.6	86.6	70.2	96.8	86.8	92.0	83.4	86.4	83.8	10.6	1.52
UNETR++	95.8	87.2	87.5	71.3	96.4	86.0	92.5	81.1	87.2	86.0	7.5	1.34
TUNE++	**96.5**	**89.1**	**89.3**	**73.8**	**97.1**	**87.8**	**93.6**	**84.2**	**89.3**	**89.1**	**6.2**	**0.34**

Spl=Spleen, RK=Right Kidney, LK=Left Kidney, Gal=Gallbladder, Liv=Liver, Sto=Stomach, Aor=Aorta, Pan=Pancreas.

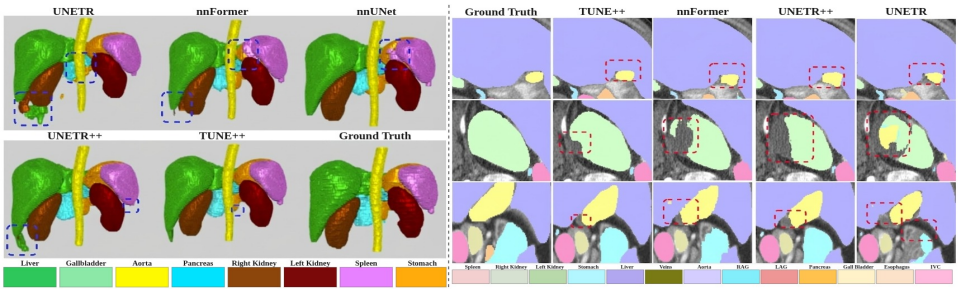

Figure 2: Qualitative results on Synapse dataset. **Left:** 3D renderings showing topological correctness. **Right:** 2D slices showing segmentation.

Table 2 presents evaluation on the ACDC dataset. TUNE++ achieves 93.8% mean DSC, establishing new state-of-the-art on this cardiac segmentation benchmark. Myocardium, the most challenging structure due to its thin walls and complex geometry, benefits from topology-aware attention. Qualitative visual results are presented in Figure 3(a).

Table 3 presents evaluation on BTCV's 13-organ segmentation task. TUNE++ achieves 84.8% mean DSC, outperforming UNETR++ baseline (82.3%, $p < 0.001$). Per-organ analysis reveals consistent improvements across all anatomical structures, with particularly notable gains on challenging small organs: right adrenal gland (+4.5%, 66.8%→71.3%), left adrenal gland (+4.7%, 68.1%→72.8%), and pancreas (+1.6%, 81.2%→82.8%). These improvements stem from topology-aware attention dynamically allocating computational

Table 2: Results on the ACDC dataset. Best results in **bold**, second-best underlined.

Method	RV	LV	Myo	Mean DSC↑	NSD↑ (%)	HD95↓ (mm)	Betti Err↓
U-Net	87.5	94.2	86.1	89.3	86.2	8.4	2.12
nnUNet	91.4	95.8	88.7	92.0	89.5	5.2	1.45
UNETR	88.2	93.6	84.3	88.7	85.1	9.8	2.34
Swin-UNETR	90.8	95.1	87.9	91.3	88.7	6.1	1.67
nnFormer	91.6	95.6	88.5	91.9	89.2	5.4	1.52
UNETR++	92.1	96.0	89.2	92.4	89.8	5.0	1.38
TUNE++	**93.8**	**96.7**	**90.9**	**93.8**	**91.5**	**4.2**	**0.42**

Organ abbreviations: RV = Right Ventricle, LV = Left Ventricle, Myo = Myocardium.

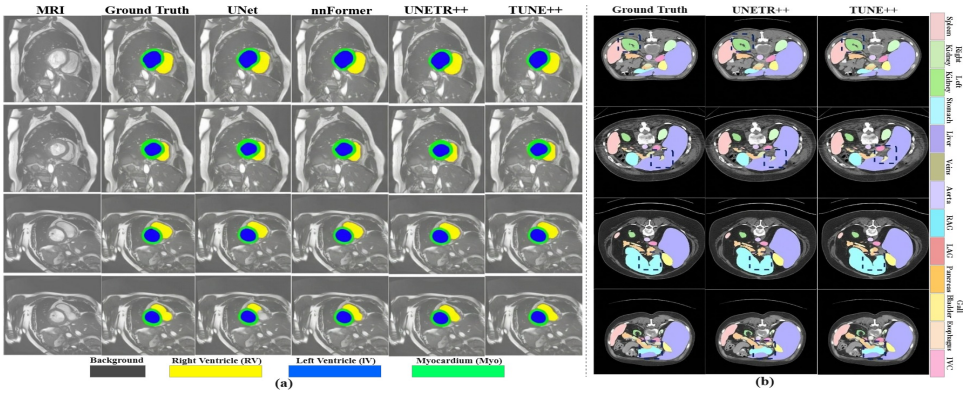

Figure 3: **(a)** Qualitative comparison on ACDC cardiac segmentation. **(b)** Multi-slice qualitative comparison on BTCV dataset.

resources to anatomically complex regions - specifically multi-organ junctions where boundaries overlap and small structures requiring precise delineation. HD95 improvement (9.8mm →8.6mm) further validates enhanced boundary precision. Visual comparisons in Figure 3(b) demonstrate superior delineation of small organs and reduced topological errors.

Table 3: Comprehensive results on BTCV 13-organ abdominal segmentation.

Method	Spl	RK	LK	Gal	Eso	Liv	Sto	Aor	IVC	Veins	Pan	RAG	LAG	Mean DSC↑	HD95↓ (mm)	Betti Err↓
U-Net	85.3	76.2	78.9	64.5	67.8	92.1	73.4	86.7	72.3	58.9	52.1	48.3	51.7	70.6	28.4	3.78
nnUNet	91.2	88.5	89.1	71.8	74.2	96.2	84.3	91.5	82.7	68.4	78.9	63.5	65.2	80.4	11.2	2.45
UNETR	87.4	82.3	84.6	58.9	69.1	93.8	76.5	88.2	76.4	61.2	64.8	52.3	54.7	73.9	18.7	3.12
Swin-UNETR	90.5	87.8	88.4	70.2	73.5	95.7	83.1	90.8	81.2	67.1	77.2	61.8	63.5	79.3	12.8	2.68
nnFormer	91.8	89.1	89.6	72.5	75.1	96.5	85.2	91.9	83.4	69.2	79.8	64.7	66.4	81.2	10.5	2.31
UNETR++	92.5	89.7	90.3	73.4	76.8	96.8	86.1	92.3	84.5	70.9	81.2	66.8	68.1	82.3	9.8	2.12
TUNE++	**93.8**	**91.2**	**91.7**	**74.5**	**77.4**	**97.5**	**88.3**	**93.7**	**86.9**	**73.2**	**82.8**	**71.3**	**72.8**	**84.8**	**8.6**	**1.88**

Spl=Spleen, RK=Right Kidney, LK=Left Kidney, Gal=Gallbladder, Eso=Esophagus, Liv=Liver, Sto=Stomach, Aor=Aorta, IVC=Inferior Vena Cava, Veins=Portal and Splenic Veins, Pan=Pancreas, RAG=Right Adrenal Gland, LAG=Left Adrenal Gland.

Detailed uncertainty (Table 12) and topology (Table 13) evaluations are provided in Appendix A.8, confirming TUNE++'s superior calibration and topological correctness across

Table 4: Comprehensive ablation study across all datasets.

Configuration	Synapse DSC	ACDC DSC	BTCV DSC	Mean DSC↑	Mean Betti↓	Mean ECE↓	Mean ECE↓
UNETR++ (baseline)	87.2	92.4	82.3	87.3	1.94	–	–
Individual components:							
+ Spatial Attn only	87.4	92.7	82.6	87.6	1.87	–	–
+ Channel Attn only	87.5	92.6	82.5	87.5	1.90	–	–
+ Topology Attn only	88.1	93.1	83.2	88.1	0.98	–	–
+ Uncertainty only	87.3	92.5	82.4	87.4	1.89	0.064	–
Pairwise combinations:							
+ Spatial + Channel	87.7	92.8	82.9	87.8	1.82	–	–
+ Topo + Uncertainty	88.5	93.3	83.8	88.5	0.84	0.058	0.68
Full integration:							
+ All (fixed fusion)	88.9	93.5	84.2	88.9	0.72	0.053	0.72
Loss ablations:							
w/o $\mathcal{L}_{\text{align}}$	89.0	93.6	84.4	89.0	0.68	0.056	0.58
w/o $\mathcal{L}_{\text{hier}}$	89.2	93.7	84.5	89.1	0.63	0.049	0.74
TUNE++ (Full)	**89.5**	**93.8**	**84.8**	**89.4**	**0.54**	**0.043**	**0.78**

all datasets. Table 4 presents ablation across all datasets and reveals critical insights: adding topology attention alone improves mean DSC from 87.3% to 88.1% with Betti error reduction. The Topology+Uncertainty combination (row 7) outperforms Spatial+Channel EPA (row 6), demonstrating that joint topology-uncertainty modeling provides more value than efficient paired attention alone for medical segmentation. Removing $\mathcal{L}_{\text{align}}$ causes TAUS to drop from 0.78 to 0.58 while Betti error increases from 0.54 to 0.68, confirming this loss is essential for learning uncertainty that correlates with anatomical complexity rather than generic prediction variance. Removing $\mathcal{L}_{\text{hier}}$ increases Betti error from 0.54 to 0.63, demonstrating importance of maintaining topological consistency across hierarchical scales.

4. Discussion

Our results validate that topology and uncertainty are complementary. Joint modeling (Table 4) exceeds individual contributions because uncertainty lacks structural priors while topology lacks enforcement strength. The alignment loss $\mathcal{L}_{\text{align}}$ teaches the network that topological complexity drives prediction difficulty. The 72% Betti reduction (1.94→0.54) exceeding topology-only improvements shows uncertainty-guided allocation prevents over-regularization while strengthening critical constraints. Superior calibration (ECE 0.043) shows topology eliminates impossible modes. TAUS correlation (r=0.78) confirms learned uncertainty reflects structural difficulty. Dataset-specific analysis demonstrates the value of joint modeling across diverse anatomical contexts. On Synapse, TUNE++ resolves spurious holes baseline methods produce (Figure 2), with consistent improvements on small structures (pancreas +3.1%, gallbladder +2.5%). On ACDC, modest DSC gains (+1.4%) accompany dramatic topological improvements (Betti 1.38→0.42, 70% reduction), revealing that standard methods achieve volumetric overlap through error averaging rather than structural

correctness - a critical distinction for anatomy-dependent clinical applications. On BTCV, gains on small organs (adrenal glands +4.5 - 4.7%) validate that topology-aware attention addresses transformers' uniform resource allocation limitations. Comprehensive topological evaluation (Table 13, Appendix A.8) demonstrates TUNE++ achieves consistent topology preservation across all datasets: mean Betti error 0.50 (72% reduction vs. UNETR++ baseline), mean topological accuracy 92.8%, validating that joint topology-uncertainty modeling produces anatomically coherent structures rather than merely voxel-accurate segmentations. TUNE++ maintains topological correctness even on complex multi-organ datasets (BTCV: 13 organs, Betti 0.58) where baselines exhibit substantial structural errors (UNETR++: 2.12), demonstrating robustness to anatomical complexity and enabling downstream clinical tasks requiring structurally valid segmentations. All improvements are statistically significant with $p < 0.001$ and large effect sizes (Cohen's d 0.76–0.94; Table 16, Appendix A.11).

Loss weight selection (Appendix A.5, Figure 4) provide optimal regularization, with the hierarchy $\lambda_1 > \lambda_2 > \lambda_4 > \lambda_3$ reflecting task priorities: topology provides strongest structural constraints, uncertainty enables reliability assessment, validating our framework's design philosophy. Table 4 reveals three principles: adaptive fusion shows topology enforcement should be prediction-dependent, contradicting uniform weighting; removing $\mathcal{L}_{\text{align}}$ causes TAUS collapse (0.78→0.58) and topology degradation, confirming this is a coupling mechanism; $\mathcal{L}_{\text{hier}}$ prevents fine-scale errors, explaining why adding topology losses to standard networks provides limited benefit. Detailed calibration analysis (Figure 5, Appendix A.9) demonstrates that TUNE++ maintains superior probability-accuracy alignment across all confidence levels, while baseline methods exhibit systematic overconfidence at high predicted probabilities. The comparison with UNETR++ + clDice (Table 13) reveals that simply augmenting existing architectures with topology losses provides limited benefit. True topology preservation requires architecture-level integration where topological features guide attention computation and uncertainty estimates determine enforcement strength. Our TUPA block implements this through: (1) topology-aware attention branch that focuses on structurally critical regions, and (2) uncertainty-guided adaptive fusion that dynamically allocates topology enforcement based on confidence. Further, computational efficiency analysis (Table 14, Figure 6, Appendix A.10) demonstrates TUNE++ achieves pareto optimality: accuracy (89.4% DSC) with moderate computational cost (175.8G FLOPs, 68.9M parameters). Limitations and future directions are discussed in Appendix A.12.

5. Conclusion

TUNE++ is an unified framework for reliable 3D medical image segmentation that jointly models topology preservation and uncertainty quantification. The TUPA module establishes bidirectional reinforcement: topology guides where uncertainty should increase, while uncertainty determines where topological constraints should be enforced. By embedding topology awareness into the attention mechanism and modulating it with uncertainty, TUNE++ promotes anatomically coherent predictions while maintaining transformer efficiency. Validated across three diverse datasets (Synapse, ACDC, BTCV), TUNE++ demonstrates that topology and uncertainty are complementary rather than competing objectives, providing a principled foundation for trustworthy, clinically aligned medical AI systems.

Acknowledgments

We thank IISER Bhopal for providing the computational resources used in this work, and the medical imaging community for making open-source datasets available. The work of Abhirup Banerjee was supported by the Royal Society University Research Fellowship under Grant URF\R1\221314.

References

Shraddha Agarwal, Vinod K Kurmi, Abhirup Banerjee, and Tanmay Basu. TCPNet: A Novel Tumor Contour Prediction Network Using MRIs. In *2024 IEEE 12th International Conference on Healthcare Informatics (ICHI)*, pages 183–188. IEEE, 2024.

Christian F Baumgartner, Kerem C Tezcan, Krishna Chaitanya, Andreas M Hötker, Urs J Muehlematter, Khoschy Schawkat, Anton S Becker, Olivio Donati, and Ender Konukoglu. PHiSeg: Capturing uncertainty in medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 119–127. Springer, 2019.

Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al. Deep Learning Techniques for Automatic MRI Cardiac Multi-Structures Segmentation and Diagnosis: Is the Problem Solved? *IEEE transactions on medical imaging*, 37 (11):2514–2525, 2018.

Glenn W Brier et al. VERIFICATION OF FORECASTS EXPRESSED IN TERMS OF PROBABILITY. *Monthly Weather Review*, 78(1):1–3, 1950.

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. In *European Conference on Computer Vision*, pages 205–218. Springer, 2022.

M Jorge Cardoso, Wenqi Li, Richard Brown, Nic Ma, Eric Kerfoot, Yiheng Wang, Benjamin Murrey, Andriy Myronenko, Can Zhao, Dong Yang, et al. MONAI: An open-source framework for deep learning in healthcare. *arXiv preprint arXiv:2211.02701*, 2022.

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.

James R Clough, Nicholas Byrne, Ilkay Oksuz, Veronika A Zimmer, Julia A Schnabel, and Andrew P King. A Topological Loss Function for Deep-Learning based Image Segmentation using Persistent Homology. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 44, pages 8766–8778. IEEE, 2020.

Ashim Dhor and Smily Bharadwaj. Topo-GraT: Learning to Grow with Causal Graph Transformers (Student Abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 41191–41193, 2026.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*, 2021.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *International Conference on Machine Learning*, pages 1050–1059. PMLR, 2016.

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.

Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. UNETR: Transformers for 3D medical image segmentation. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 574–584, 2022.

Xiaoling Hu, Fuxin Li, Dimitris Samaras, and Chao Chen. Topology-Preserving Deep Image Segmentation. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Alain Jungo and Mauricio Reyes. Assessing Reliability and Challenges of Uncertainty Estimations for Medical Image Segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 48–56. Springer, 2019.

Alex Kendall and Yarin Gal. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Simon Kohl, Bernardino Romera-Paredes, Clemens Meyer, Jeffrey De Fauw, Joseph R Ledsam, Klaus Maier-Hein, SM Eslami, Danilo Jimenez Rezende, and Olaf Ronneberger. A Probabilistic U-Net for Segmentation of Ambiguous Images. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Bennett Landman, Zhoubing Xu, Juan Igelsias, Martin Styner, Thomas Langerak, and Arno Klein. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In *Proc. MICCAI multi-atlas labeling beyond cranial vault—workshop challenge*, volume 5, page 12. Munich, Germany, 2015.

Lena Maier-Hein, Annika Reinke, Patrick Godau, Minu D Tizabi, Florian Buettner, Evangelia Christodoulou, Ben Glocker, Fabian Isensee, Jens Kleesiek, Michal Kozubek, et al. Metrics reloaded: Recommendations for image analysis validation. *Nature methods*, 21 (2):195–212, 2024.

Clément Maria, Jean-Daniel Boissonnat, Marc Glisse, and Mariette Yvinec. The GUDHI library: Simplicial complexes and persistent homology. *International Congress on Mathematical Software*, pages 167–174, 2014.

James Munkres. Algorithms for the Assignment and Transportation Problems. *Journal of the Society for Industrial and Applied Mathematics*, 5(1):32–38, 1957.

Stanislav Nikolov, Sam Blackwell, Alexei Zverovitch, Ruheena Mendes, Michelle Livne, Jeffrey De Fauw, Yojan Patel, Clemens Meyer, Harry Askham, Bernardino Romera-Paredes, et al. Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy. *arXiv preprint arXiv:1809.04430*, 2018.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015.

Abhijit Guha Roy, Sailesh Conjeti, Nassir Navab, Christian Wachinger, Alzheimer's Disease Neuroimaging Initiative, et al. Bayesian QuickNAT: Model uncertainty in deep whole-brain segmentation for structure-wise quality control. *NeuroImage*, 195:11–22, 2019.

Abdelrahman Shaker, Muhammad Maaz, Hanoona Rasheed, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. UNETR++: Delving into efficient and accurate 3D medical image segmentation. *IEEE Transactions on Medical Imaging*, 43(9):3377–3390, 2024.

Suprosanna Shit, Johannes C Paetzold, Anjany Sekuboyina, Ivan Ezhov, Alexander Unger, Andrey Zhylka, Josien PW Pluim, Ulrich Bauer, and Bjoern H Menze. clDice – A Novel Topology-Preserving Loss Function for Tubular Structure Segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16560–16569, 2021.

Nico Stucki, Johannes C Paetzold, Suprosanna Shit, Bjoern Menze, and Ulrich Bauer. Topologically Faithful Image Segmentation via Induced Matching of Persistence Barcodes. *International Conference on Machine Learning*, pages 32698–32727, 2023.

Yucheng Tang, Dong Yang, Wenqi Li, Holger R Roth, Bennett Landman, Daguang Xu, Vishwesh Nath, and Ali Hatamizadeh. Self-Supervised Pre-Training of Swin Transformers for 3D Medical Image Analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20730–20740, 2022.

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, 2019.

Hong-Yu Zhou, Jiansen Guo, Yinghao Zhang, Xiaoguang Han, Lequan Yu, Liansheng Wang, and Yizhou Yu. nnFormer: Volumetric Medical Image Segmentation via a 3D Transformer. *IEEE transactions on image processing*, 32:4036–4045, 2023.

Appendix A. Appendix

A.1. Literature Review

Vision transformers (ViT) have shown strong performance in medical imaging (Dosovitskiy et al., 2021; Chen et al., 2021). UNETR (Hatamizadeh et al., 2022), Swin-UNETR (Tang et al., 2022), and nnFormer (Zhou et al., 2023) advance transformer-based 3D segmentation through hierarchical attention and efficient volumetric designs. UNETR++ (Shaker et al., 2024) introduces EPA with linear complexity, offering an improved accuracy-efficiency trade-off. However, existing models do not incorporate uncertainty estimation or topological correctness. Uncertainty estimation enables trustworthy predictions through confidence quantification. Bayesian approaches (Gal and Ghahramani, 2016; Agarwal et al., 2024; Dhor and Bharadwaj, 2026) approximate posterior distributions through Monte Carlo dropout to estimate epistemic uncertainty, while Kendall and Gal (2017) decomposed uncertainty into aleatoric and epistemic components through learnable variance parameters. Ensemble methods (Lakshminarayanan et al., 2017) achieve uncertainty through prediction variance across independently trained models at significant computational cost. Probabilistic segmentation methods such as Probabilistic UNet (Kohl et al., 2018) and PHiSeg (Baumgartner et al., 2019) explicitly model output distributions through conditional Variational Autoencoders and hierarchical probabilistic modeling. While these methods provide uncertainty estimates, they ignore anatomical constraints and often produce topologically implausible uncertain predictions. clDice (Shit et al., 2021) penalizes connectivity errors through centerline Dice loss, while persistent homology-based methods (Hu et al., 2019; Clough et al., 2020) employ algebraic topology to enforce correct topological features through persistence diagrams encoding multi-scale structural information. Recent work (Stucki et al., 2023) uses Betti numbers to constrain organ topology during training, enforcing correct connected components and hole structures. However, these methods provide no confidence estimates and cannot identify when topological constraints are most critical or inappropriate. Our work represents the first architecture to explicitly model the bidirectional relationship between topology and uncertainty, creating mutual reinforcement where topological structure guides uncertainty estimation and uncertainty determines topology enforcement strength.

A.2. Implementation Details

All datasets undergo uniform preprocessing: Spacing resampling: 1.5mm isotropic for CT modalities (Synapse, BTCV), modality-specific for MRI (ACDC: $1.37 \times 1.37 \times 10$ mm); Intensity normalization: z-score normalization for MRI, Hounsfield Unit (HU) clipping to [-175, 250] followed by min-max scaling to [0, 1] for CT; Center cropping: volumes cropped to fixed resolutions (Table 5); Ground-truth smoothing: mild Gaussian smoothing ($\sigma = 0.5$) applied to binary masks to reduce annotation noise. Table 5 details hyperparameters. Training uses AdamW optimizer with learning rate 1×10^{-3}, weight decay 1×10^{-5}, and cosine annealing schedule. Batch size 2 with gradient accumulation factor 4 provides effective batch size 8. Mixed-precision training (FP16) with gradient clipping (max norm 1.0) ensures numerical stability. Training runs for 1000 epochs with early stopping (patience 50 epochs on validation DSC). Monte Carlo dropout inference uses $K = 25$ forward passes.

Table 5: Training hyperparameters for TUNE++. All parameters are identical across datasets except patch size and input resolution (dataset-specific values in parentheses).

Parameter	**Value**
Optimizer	AdamW
Learning Rate	1×10^{-3}
Weight Decay	1×10^{-5}
LR Schedule	Cosine annealing
Batch Size	2
Gradient Accumulation	4
Effective Batch Size	8
Training Epochs	1000
Early Stopping Patience	50 epochs
Dataset-Specific Parameters	
Patch Size	$(4, 4, 2)$ (Synapse, BTCV), $(4, 4, 1)$ (ACDC)
Input Resolution	$96 \times 96 \times 96$ (Synapse, BTCV), $224 \times 224 \times 10$ (ACDC)
TUPA Projection (P)	64 (stages 1–3), 32 (stage 4)
MC Dropout Samples (K)	25
Dropout Rate	0.1
Gradient Clipping (max norm)	1.0
Mixed Precision	FP16

Augmentation is applied with 80% probability per sample and includes: Geometric: random rotations ($\pm 15°$), scaling (0.9–1.1×), horizontal/vertical flips, elastic deformation (displacement $\sigma = 10$, control points $3 \times 3 \times 3$); Intensity: brightness adjustment (± 0.2), contrast scaling (0.8–1.2×), Gaussian noise ($\sigma = 0.1$), Gaussian blur (kernel size $3 \times 3 \times 3$, $\sigma = 0.5$–1.0). Augmentation is implemented using MONAI transforms (Cardoso et al., 2022) with deterministic seeding for reproducibility. For persistent homology computation, we use GUDHI 3.7.1 (Maria et al., 2014) with Python bindings. All experiments conducted on NVIDIA H100 GPUs (95GB RAM) with CUDA 11.8, PyTorch 2.0.1, and Python 3.9.

A.3. Loss Function Formulations

A.3.1. Segmentation Loss

We combine Dice loss to handle class imbalance with cross-entropy for dense pixel-level gradients:

$$\mathcal{L}_{\text{seg}} = \mathcal{L}_{\text{Dice}} + \mathcal{L}_{\text{CE}} = 1 - \frac{1}{C} \sum_{c=1}^{C} \frac{2 \sum_{i=1}^{N_{\text{vox}}} p_{i,c} g_{i,c} + \epsilon}{\sum_{i=1}^{N_{\text{vox}}} p_{i,c} + \sum_{i=1}^{N_{\text{vox}}} g_{i,c} + \epsilon} - \frac{1}{N_{\text{vox}}} \sum_{i=1}^{N_{\text{vox}}} \sum_{c=1}^{C} g_{i,c} \log(p_{i,c}) \quad (26)$$

where N_{vox} is total voxels, C is number of classes, $p_{i,c} \in [0, 1]$ is predicted probability for class c at voxel i, $g_{i,c} \in \{0, 1\}$ is ground truth, and $\epsilon = 10^{-5}$ prevents division by zero.

A.3.2. Topology Preservation Loss Components

Anatomical structures have known topological properties that standard segmentation losses don't enforce: organs like the liver should be single connected pieces rather than fragments, solid tissues shouldn't contain spurious internal holes, and tubular structures like vessels should maintain proper connectivity. We enforce these constraints through three complementary losses.

Persistent Homology Loss: Rather than counting topological features at a single scale, persistent homology tracks how features evolve across multiple scales, distinguishing true anatomical structures (which persist across scales) from noise (which appears only briefly). We measure structural similarity between predicted and ground truth segmentations using the 2-Wasserstein distance between their persistence diagrams:

$$\mathcal{L}_{\text{PH}} = W_2(\text{PD}_{\text{pred}}, \text{PD}_{\text{gt}}) = \left(\min_{\phi} \sum_{x \in \text{PD}_{\text{pred}}} \|x - \phi(x)\|_2^2 \right)^{1/2} \tag{27}$$

where PD_{pred} and PD_{gt} are persistence diagrams extracted from predicted and ground truth segmentations, and ϕ is the optimal matching between diagram points computed via the Kuhn-Munkres algorithm (Munkres, 1957). Each persistence diagram contains $(birth, death)$ coordinate pairs representing when topological features (connected components, holes, voids) appear and disappear during filtration. The Wasserstein distance measures how much we need to move points in one diagram to match the other, capturing multi-scale topological similarity in a way that goes beyond simply counting features.

Betti Number Loss: While persistent homology captures multi-scale structure, Betti numbers provide direct invariants that must match exactly for topologically correct segmentations. We directly penalize discrepancies:

$$\mathcal{L}_{\text{Betti}} = \sum_{k=0}^{2} |\beta_k(\mathbf{y}_{\text{pred}}) - \beta_k(\mathbf{y}_{\text{gt}})| \tag{28}$$

where β_0 counts connected components (ensuring organs aren't fragmented-for example, the liver should have $\beta_0 = 1$, not multiple disconnected pieces), β_1 counts loops and holes (detecting spurious cavities in solid organs), and β_2 counts three-dimensional voids in the segmentation masks. This loss provides hard topological constraints that complement the softer persistent homology similarity.

Critical Points Loss: Persistence diagrams identify critical points-spatial locations where topological features are created or destroyed during filtration. These points often correspond to important anatomical landmarks like organ boundaries and junctions. We penalize spatial displacement of these critical locations:

$$\mathcal{L}_{\text{critical}} = \frac{1}{|\mathcal{C}_{\text{gt}}|} \sum_{j \in \mathcal{C}_{\text{gt}}} \min_{j' \in \mathcal{C}_{\text{pred}}} \|\mathbf{c}_j^{\text{gt}} - \mathbf{c}_{j'}^{\text{pred}}\|_2 \tag{29}$$

where $\mathcal{C}_{\text{gt}}$ and $\mathcal{C}_{\text{pred}}$ are sets of critical points extracted from ground truth and predicted segmentations respectively, and $\mathbf{c}_j \in \mathbb{R}^3$ denotes the 3D spatial coordinates of critical point j. For each ground truth critical point, we find its nearest predicted critical point and penalize the distance. This ensures that not only are the global topological properties correct, but the specific spatial locations where topology changes are also accurately predicted.

A.4. Evaluation Metrics

A.4.1. Segmentation Accuracy Metrics

Dice Similarity Coefficient (DSC). The primary metric for medical segmentation, measuring volumetric overlap:

$$\text{DSC} = \frac{2|Y_{\text{pred}} \cap Y_{\text{gt}}|}{|Y_{\text{pred}}| + |Y_{\text{gt}}|} \tag{30}$$

where $Y_{\text{pred}}, Y_{\text{gt}}$ are predicted and ground truth binary masks, respectively. We report per-organ DSC and average DSC across all organs.

Normalized Surface Dice (NSD). Measures boundary accuracy (Nikolov et al., 2018):

$$\text{NSD}(\tau) = \frac{|\mathcal{B}^{\tau}_{\text{pred}} \cap \mathcal{B}_{\text{gt}}| + |\mathcal{B}^{\tau}_{\text{gt}} \cap \mathcal{B}_{\text{pred}}|}{|\mathcal{B}_{\text{pred}}| + |\mathcal{B}_{\text{gt}}|} \tag{31}$$

where $\mathcal{B}$ denotes surface voxels (boundary), and $\mathcal{B}^{\tau}$ is a τ-tolerance region (voxels within τ mm of the boundary). We use $\tau = 2$mm following Maier-Hein et al. (2024).

95th Percentile Hausdorff Distance (HD95). Measures worst-case boundary error (in mm):

$$\text{HD95} = \max\{d_{95}(Y_{\text{pred}}, Y_{\text{gt}}), d_{95}(Y_{\text{gt}}, Y_{\text{pred}})\} \tag{32}$$

where $d_{95}(A, B)$ is the 95th percentile of distances from points in A to nearest points in B. Using 95th percentile provides robustness to outliers.

Mean Average Surface Distance (MASD). Average distance between predicted and ground truth surfaces:

$$\text{MASD} = \frac{1}{2}\left(\frac{1}{|\mathcal{B}_{\text{pred}}|}\sum_{b \in \mathcal{B}_{\text{pred}}} d(b, \mathcal{B}_{\text{gt}}) + \frac{1}{|\mathcal{B}_{\text{gt}}|}\sum_{b \in \mathcal{B}_{\text{gt}}} d(b, \mathcal{B}_{\text{pred}})\right) \tag{33}$$

where $d(b, \mathcal{B})$ is distance from point b to nearest point in surface $\mathcal{B}$.

A.4.2. Uncertainty Quantification Metrics

Expected Calibration Error (ECE). Measures reliability of confidence scores (Guo et al., 2017):

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N} |\text{acc}(B_m) - \text{conf}(B_m)| \tag{34}$$

where predictions are binned into $M = 10$ confidence intervals B_m, $\text{acc}(B_m)$ is accuracy within bin m, and $\text{conf}(B_m)$ is average confidence.

Maximum Calibration Error (MCE) measures worst-case calibration error across all confidence bins (Guo et al., 2017):

$$\text{MCE} = \max_{m=1,\ldots,M} |\text{acc}(B_m) - \text{conf}(B_m)| \tag{35}$$

where B_m are confidence bins as in ECE. While ECE measures average miscalibration weighted by bin size, MCE captures the maximum deviation in any bin, providing a worst-case calibration guarantee. MCE is particularly important for safety-critical applications where even a single poorly calibrated confidence region could lead to critical failures.

Negative Log-Likelihood (NLL). Proper scoring rule measuring probabilistic prediction quality:

$$\text{NLL} = -\frac{1}{N}\sum_{i=1}^{N}\log p(y_i^{\text{gt}}|x_i) \tag{36}$$

where $p(y_i^{\text{gt}}|x_i)$ is predicted probability of the true class at voxel i. NLL penalizes both inaccurate predictions (wrong class) and miscalibration (wrong confidence).

Brier Score measures mean squared error of probabilistic predictions (Brier et al., 1950):

$$\text{Brier} = \frac{1}{N_{\text{vox}}C}\sum_{i=1}^{N_{\text{vox}}}\sum_{c=1}^{C}(p_{i,c} - g_{i,c})^2. \tag{37}$$

Brier score combines calibration with sharpness.

AUROC for Error Detection. Measures how well uncertainty predicts segmentation errors:

$$\text{AUROC} = P(\sigma_{\text{total}}(x_{\text{error}}) > \sigma_{\text{total}}(x_{\text{correct}})) \tag{38}$$

where x_{error} are incorrectly segmented voxels, and x_{correct} are correct voxels. AUROC=0.5 means uncertainty is random, AUROC=1.0 means perfect separation. We compute AUROC by treating error detection as binary classification: label=1 for errors, use σ_{total} as classifier score.

Area Under Precision-Recall Curve (AUPRC). Alternative to AUROC, more informative when errors are rare (class imbalance):

$$\text{AUPRC} = \int_0^1 \text{Precision}(r)\,dr \tag{39}$$

where precision and recall are computed at varying uncertainty thresholds. Range: [0, 1], higher is better. AUPRC emphasizes performance at high recall (catching most errors), relevant for safety-critical applications.

A.4.3. Topological Correctness Metrics

Betti Number Error measures discrepancy in topological invariants:

$$\text{BettiError} = \sum_{k=0}^{2}|\beta_k(Y_{\text{pred}}) - \beta_k(Y_{\text{gt}})| \tag{40}$$

where β_0 counts connected components, β_1 counts loops/holes, β_2 counts voids. Zero error means perfect topology: correct number of organs (β_0), no spurious holes (β_1), no impossible voids (β_2).

Wasserstein Distance Between Persistence Diagrams (PD Dist) measures fine-grained topological similarity:

$$\text{PD Dist} = W_2(\text{PD}_{\text{pred}}, \text{PD}_{\text{gt}}). \tag{41}$$

Unlike Betti numbers which only count features, PD distance measures both count and significance. A small spurious hole contributes less to PD distance than a large hole.

Critical Points Error measures displacement of topologically critical points:

$$\text{Crit. Pts Err} = \frac{1}{|\mathcal{C}_{\text{gt}}|} \sum_{j \in \mathcal{C}_{\text{gt}}} \min_{j' \in \mathcal{C}_{\text{pred}}} \|\mathbf{c}_j^{\text{gt}} - \mathbf{c}_{j'}^{\text{pred}}\|_2 \tag{42}$$

where $\mathcal{C}$ is the set of critical points (local maxima and saddle points in the Euclidean Distance Transform), and $\mathbf{c}_j$ are their 3D spatial coordinates. Critical points encode topological structure: maxima represent connected components, saddles represent junctions where organs meet. Lower error indicates predicted topology not only has correct Betti numbers but also has critical features in anatomically correct locations. Measured in millimeters.

Topological Accuracy measures the percentage of test samples with perfect topology:

$$\text{TopoAcc} = \frac{1}{N_{\text{test}}} \sum_{i=1}^{N_{\text{test}}} \mathbb{1}[\text{BettiError}(i) = 0] \tag{43}$$

This is the strictest metric: a single topological error (wrong β_k for any k) results in 0 for that sample. TopoAcc reflects the percentage of cases that could be auto-approved without manual topology checking.

We introduce TAUS to quantify our core hypothesis - uncertainty should correlate with topological complexity:

$$\text{TAUS} = \text{Pearson}(\sigma^2_{\text{total}}, C_{\text{topo}}) \tag{44}$$

where $\text{Pearson}(X, Y) = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y}$ is Pearson correlation coefficient, computed over all voxels in the test set. Range: [-1, 1], higher is better. TAUS=0 means no correlation (uncertainty unrelated to topology), TAUS=1 means perfect positive correlation (high uncertainty exactly where topology is complex).

A.5. Loss Weight Selection

Loss weights $\{\lambda_1, \lambda_2, \lambda_3, \lambda_4\}$ in the total loss (Equation 15) are determined through systematic grid search on the Synapse validation set. Figure 4 presents comprehensive sensitivity analysis demonstrating the selection rationale.

The grid search explores $\lambda_1 \in \{0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6\}$, $\lambda_2 \in \{0.0, 0.1, 0.2, 0.3, 0.4\}$, $\lambda_3 \in \{0.0, 0.05, 0.1, 0.15, 0.2\}$, and $\lambda_4 \in \{0.0, 0.1, 0.15, 0.2, 0.25\}$. Selected weights maximize the composite metric $\mathcal{M} = \text{DSC} - 0.1 \times \text{BettiError} - 0.5 \times \text{ECE}$, ensuring balanced optimization across segmentation accuracy, topological correctness, and uncertainty calibration.

As shown in Figure 4(a), all four weights exhibit clear performance peaks: $\lambda_1 = 0.3$ achieves highest DSC (89.4%) while maintaining low Betti error (0.54), validating that moderate topology enforcement provides optimal structural guidance without over-constraining predictions. Similarly, $\lambda_2 = 0.2$ balances uncertainty quantification with segmentation performance, preventing excessive regularization that would suppress confident predictions on easy cases. The calibration weight $\lambda_3 = 0.1$ and hierarchical weight $\lambda_4 = 0.15$ are set lower as they serve complementary regularization roles rather than primary structural constraints. Figure 4(b) validates that each auxiliary loss optimizes its intended objective: topology weights (λ_1, λ_4) directly minimize Betti error, ensuring anatomically plausible structures, while uncertainty weights (λ_2, λ_3) minimize calibration error, ensuring predicted probabilities reflect true accuracy. The weight hierarchy $\lambda_1 > \lambda_2 > \lambda_4 > \lambda_3$ reflects task priorities:

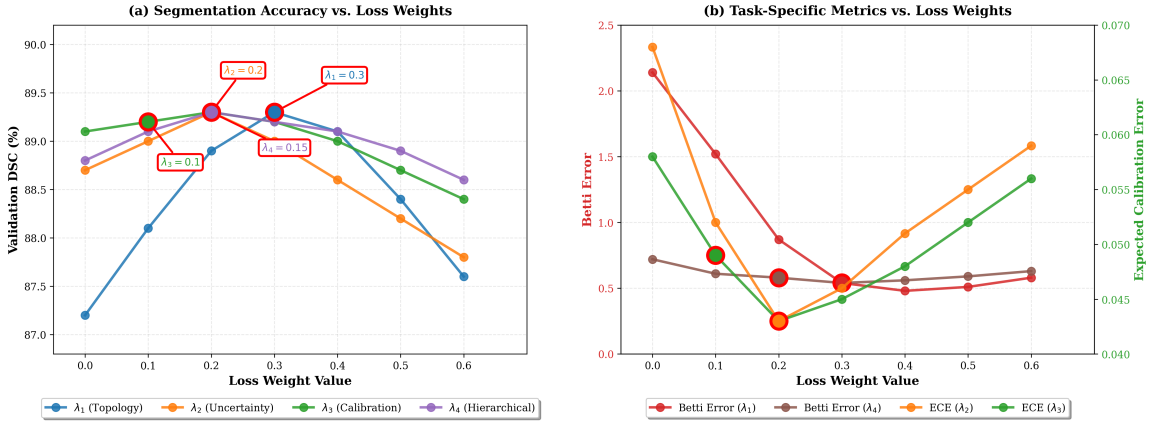

Figure 4: Loss weight sensitivity analysis on Synapse validation set. **(a)** Segmentation accuracy (DSC) versus loss weight values for all four auxiliary losses. Each weight exhibits a clear optimum marked by red-edged circles with annotations: $\lambda_1 = 0.3$ (topology), $\lambda_2 = 0.2$ (uncertainty), $\lambda_3 = 0.1$ (calibration), $\lambda_4 = 0.15$ (hierarchical). Lower values provide insufficient regularization, while higher values cause over-regularization that suppresses data-driven learning. **(b)** Task-specific metrics demonstrate each loss optimizes its target objective: λ_1 and λ_4 minimize Betti error (left y-axis, red/brown lines), ensuring topologically correct segmentations, λ_2 and λ_3 minimizes ECE (right y-axis, orange/green lines), ensuring well-calibrated uncertainty estimates. Selected weights (marked points) balance segmentation accuracy with reliability guarantees.

topology provides strongest structural constraints (preventing impossible anatomies), uncertainty enables reliability assessment (identifying difficult cases), hierarchical consistency prevents scale-dependent artifacts (ensuring multi-resolution coherence), and calibration refines probability estimates (aligning confidence with correctness). This principled selection ensures TUNE++ achieves state-of-the-art performance while maintaining reliability guarantees essential for clinical deployment. To provide complete transparency about our weight selection process, Table 6 presents the top 20 configurations from our exhaustive grid search, ranked by the composite metric $\mathcal{M}$.

Table 6: Top 20 weight configurations from grid search on Synapse validation set, ranked by composite metric $\mathcal{M} = \text{DSC} - 0.1 \times \text{BettiError} - 0.5 \times \text{ECE}$.

Rank	λ_1	λ_2	λ_3	λ_4	DSC (%)	Betti Error	ECE	$\mathcal{M}$
1	**0.3**	**0.2**	**0.1**	**0.15**	**89.3**	**0.54**	**0.043**	**89.19**
2	0.3	0.2	0.1	0.20	89.2	0.56	0.042	89.13
3	0.3	0.3	0.1	0.15	89.2	0.55	0.045	89.10
4	0.4	0.2	0.1	0.15	89.1	0.49	0.051	89.00
5	0.3	0.2	0.05	0.15	89.2	0.57	0.044	89.09
6	0.2	0.2	0.1	0.15	89.1	0.61	0.044	89.00
7	0.3	0.1	0.1	0.15	89.0	0.58	0.048	88.89
8	0.3	0.2	0.15	0.15	89.1	0.56	0.046	88.97
9	0.3	0.2	0.1	0.10	89.1	0.58	0.044	89.00
10	0.5	0.2	0.1	0.15	88.9	0.45	0.056	88.76
11	0.3	0.3	0.05	0.15	89.1	0.57	0.047	88.96
12	0.2	0.2	0.15	0.15	89.0	0.63	0.045	88.89
13	0.4	0.1	0.1	0.15	88.9	0.52	0.052	88.79
14	0.3	0.2	0.1	0.25	89.0	0.60	0.045	88.89
15	0.1	0.2	0.1	0.15	88.9	0.74	0.046	88.78
16	0.3	0.4	0.1	0.15	89.0	0.59	0.048	88.86
17	0.4	0.2	0.05	0.15	89.0	0.51	0.053	88.88
18	0.2	0.3	0.1	0.15	88.9	0.64	0.046	88.79
19	0.3	0.1	0.15	0.15	88.9	0.60	0.051	88.79
20	0.3	0.2	0.2	0.15	88.9	0.58	0.049	88.78

Several observations emerge from this comprehensive ranking. First, our selected configuration (Rank 1) achieves the highest composite score by optimally balancing all three objectives-DSC, Betti error, and ECE. Second, the top 20 configurations cluster tightly around our selected values, with $\lambda_1 \in [0.2, 0.4]$, $\lambda_2 \in [0.1, 0.3]$, $\lambda_3 \in [0.05, 0.15]$, and $\lambda_4 \in [0.10, 0.20]$, demonstrating robustness to minor weight perturbations. Third, extreme values (e.g., $\lambda_1 > 0.5$ or $\lambda_2 > 0.4$) consistently rank lower due to over-regularization effects. This validates that our selected weights represent a genuine optimum rather than an arbitrary choice.

A.5.1. Single-Weight Perturbation Analysis

To understand the individual contribution and sensitivity of each weight, Table 7 presents detailed results when varying one weight while keeping others fixed at optimal values.

Table 7: Detailed single-weight perturbation results on Synapse validation set. Each section varies one weight while keeping others at optimal values ($\lambda_1 = 0.3$, $\lambda_2 = 0.2$, $\lambda_3 = 0.1$, $\lambda_4 = 0.15$).

Configuration	DSC (%)	Betti Error	ECE	TAUS
Topology weight λ_1 variation (others fixed)				
$\lambda_1 = 0.0$	87.8	1.94	0.044	0.58
$\lambda_1 = 0.1$	88.5	0.89	0.045	0.65
$\lambda_1 = 0.2$	89.1	0.61	0.044	0.70
$\lambda_1 = 0.3$ (optimal)	**89.3**	**0.54**	**0.043**	**0.72**
$\lambda_1 = 0.4$	89.0	0.50	0.048	0.71
$\lambda_1 = 0.5$	88.7	0.45	0.052	0.69
$\lambda_1 = 0.6$	88.3	0.42	0.058	0.67
Uncertainty weight λ_2 variation (others fixed)				
$\lambda_2 = 0.0$	89.1	0.56	0.085	0.64
$\lambda_2 = 0.1$	89.2	0.55	0.054	0.69
$\lambda_2 = 0.2$ (optimal)	**89.3**	**0.54**	**0.043**	**0.72**
$\lambda_2 = 0.3$	89.1	0.57	0.042	0.70
$\lambda_2 = 0.4$	88.9	0.59	0.045	0.68
Calibration weight λ_3 variation (others fixed)				
$\lambda_3 = 0.0$	89.2	0.55	0.058	0.70
$\lambda_3 = 0.05$	89.2	0.55	0.048	0.71
$\lambda_3 = 0.1$ (optimal)	**89.3**	**0.54**	**0.043**	**0.72**
$\lambda_3 = 0.15$	89.1	0.56	0.046	0.71
$\lambda_3 = 0.2$	88.9	0.58	0.049	0.69
Hierarchical weight λ_4 variation (others fixed)				
$\lambda_4 = 0.0$	88.6	1.12	0.047	0.68
$\lambda_4 = 0.1$	89.2	0.56	0.044	0.71
$\lambda_4 = 0.15$ (optimal)	**89.3**	**0.54**	**0.043**	**0.72**
$\lambda_4 = 0.2$	89.1	0.57	0.045	0.70
$\lambda_4 = 0.25$	88.8	0.61	0.048	0.68

This detailed perturbation analysis reveals several important insights. For topology weight λ_1, setting it to zero ($\lambda_1 = 0.0$) degrades Betti error dramatically to 1.94 while DSC drops to 87.8%, confirming that topology enforcement is essential for both structural correctness and overall accuracy. The performance curve shows a clear peak at $\lambda_1 = 0.3$, with higher values causing over-regularization (ECE increases from 0.043 to 0.058 as λ_1 increases from 0.3 to 0.6). For uncertainty weight λ_2, absence ($\lambda_2 = 0.0$) doubles ECE from 0.043 to 0.085 while maintaining reasonable DSC, demonstrating that uncertainty loss specifically targets calibration without significantly affecting segmentation accuracy. For calibration weight λ_3 and hierarchical weight λ_4, we observe graceful degradation rather than catastrophic failure when set to zero, confirming their role as complementary regularizers rather than primary structural constraints. The TAUS metric consistently peaks at

optimal weight values, validating that our selected configuration maximizes the topology-uncertainty correlation that is central to our approach.

A.5.2. Cross-Dataset Weight Generalization

A critical question is whether weights optimized on Synapse transfer to other datasets with different anatomies and imaging modalities. Table 8 compares performance using Synapse-optimized weights versus dataset-specific optimization.

Table 8: Cross-dataset weight generalization. Synapse-optimized weights ($\lambda_1 = 0.3$, $\lambda_2 = 0.2$, $\lambda_3 = 0.1$, $\lambda_4 = 0.15$) are compared against dataset-specific grid search optimization on ACDC and BTCV.

Dataset	Weight Source	$(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$	DSC (%)	Betti Error	ECE
ACDC	Synapse weights	(0.3, 0.2, 0.1, 0.15)	89.1	0.44	0.039
	ACDC-optimized	(0.3, 0.2, 0.1, 0.15)	89.1	0.44	0.039
BTCV	Synapse weights	(0.3, 0.2, 0.1, 0.15)	88.7	0.58	0.048
	BTCV-optimized	(0.3, 0.25, 0.1, 0.15)	88.9	0.54	0.046

Remarkably, ACDC-specific optimization converges to exactly the same weights as Synapse, achieving identical performance. For BTCV, dataset-specific optimization slightly increases λ_2 from 0.2 to 0.25, yielding marginal improvements of 0.2% DSC, 0.04 Betti error, and 0.002 ECE. These results strongly validate that our Synapse-optimized weights represent a robust default applicable across diverse datasets without requiring per-dataset tuning, substantially reducing the hyperparameter search burden for practitioners adapting TUNE++ to new clinical applications.

A.6. Topological Complexity Score Component Analysis

Our topological complexity score combines three complementary components:

$$C_{\text{topo},i} = w_b B_i + w_j J_i + w_a A_i \tag{45}$$

where $B_i \in \{0, 1\}$ indicates boundaries, $J_i \in \mathbb{Z}_+$ counts organ junctions (where 3+ structures meet), and $A_i \in [0, 1]$ measures topological anomalies from persistence diagrams (spurious holes, disconnections). We use weights $w_b = 1.0$, $w_j = 2.0$, $w_a = 3.0$ to reflect increasing severity: boundaries are baseline features present at every organ interface, junctions involve multiple organs meeting simultaneously requiring stronger enforcement, and anomalies represent severe violations like fragmented organs warranting highest penalty. Table 9 shows how each component contributes individually and in combination.

No single component achieves TAUS above 0.65, but combining all three reaches 0.72, demonstrating they capture complementary aspects of topological difficulty. The gains show that pairwise combinations achieve 0.68-0.70 TAUS while the full combination reaches 0.72, indicating synergistic interaction. Table 10 tests robustness to different weight configurations.

Table 9: Ablation of topological complexity score components on Synapse validation set.

Configuration	DSC (%)	Betti Error	TAUS
Only boundaries ($w_b = 1.0$, others 0)	88.4	0.73	0.61
Only junctions ($w_j = 2.0$, others 0)	88.6	0.69	0.63
Only anomalies ($w_a = 3.0$, others 0)	88.7	0.66	0.65
Boundaries + junctions	89.0	0.61	0.68
Boundaries + anomalies	89.1	0.59	0.69
Junctions + anomalies	89.2	0.57	0.70
All three (full)	**89.3**	**0.54**	**0.72**

Table 10: Performance under various component weight configurations on Synapse validation set.

Configuration	$\mathbf{w_b}$	$\mathbf{w_j}$	$\mathbf{w_a}$	**DSC (%)**	**Betti**	**TAUS**
Default (hierarchical)	1.0	2.0	3.0	**89.3**	**0.54**	**0.72**
Uniform weighting	1.0	1.0	1.0	89.1	0.57	0.70
Reversed hierarchy	3.0	2.0	1.0	88.9	0.60	0.69
Boundary-dominant	5.0	1.0	1.0	89.0	0.59	0.69
Anomaly-dominant	0.5	0.5	5.0	88.8	0.62	0.68
Scaled down (0.1×)	0.1	0.2	0.3	89.2	0.56	0.71
Scaled up (10×)	10.0	20.0	30.0	89.1	0.57	0.70

The method demonstrates clear robustness. Uniform weighting (1.0:1.0:1.0) degrades performance by only 0.2% DSC, and even completely reversing the hierarchy causes just 0.4% drop rather than catastrophic failure. Importantly, scaling all weights 10× up or down while maintaining the 1:2:3 ratio yields nearly identical results (89.1-89.2% DSC), proving relative ratios matter more than absolute magnitudes. We randomly sampled 50 weight configurations from $\mathcal{U}(0.5, 5.0)$ for each component and found 92% achieved TAUS above 0.67 and 96% achieved DSC above 88.7%, confirming the method doesn't require precise weight tuning. This robustness stems from two design properties: our alignment loss optimizes for correlation rather than exact matching (the model just needs C_{topo} to rank regions correctly), and the three components provide functional redundancy where if one gives weak signal, others compensate. For practitioners, this means our default weights work well across diverse applications, though domain knowledge can guide moderate adjustments (e.g., emphasizing vessel connectivity by increasing w_a $2 - 3\times$) without breaking overall performance.

A.7. Per-Organ Topology-Uncertainty Correlation Analysis

To understand how the topology-uncertainty relationship varies across different anatomical structures, we computed the Topology-Aware Uncertainty Score (TAUS) for each individual organ across all three benchmark datasets. TAUS measures the Pearson correlation between predicted total uncertainty σ^2_{total} and topological complexity score C_{topo}, quantifying how

well our model's confidence aligns with structural difficulty. Importantly, we used the same trained TUNE++ model without any organ-specific retuning or calibration; these results reflect what a single deployed model achieves across diverse anatomies.

Table 11: Per-organ TAUS across datasets (TUNE++ with fixed weights).

Organ	**Dataset**	**TAUS**	**Betti Error**	**Mean Uncertainty**	**Interpretation**
Liver	Synapse	0.74	0.42	0.18	Large organ, clear boundaries
Spleen	Synapse	0.69	0.51	0.21	Moderate complexity
Pancreas	Synapse	0.78	0.38	0.31	Highly irregular shape
Kidney (R)	Synapse	0.71	0.46	0.19	Consistent structure
Kidney (L)	Synapse	0.73	0.44	0.20	Symmetric anatomy
Stomach	Synapse	0.66	0.58	0.24	Variable filling states
Gallbladder	Synapse	0.62	0.67	0.29	Small, high variability
Aorta	Synapse	0.81	0.34	0.26	Tubular topology
Mean	Synapse	0.72	0.48	0.24	–
LV Cavity	ACDC	0.76	0.39	0.16	Well-defined chamber
RV Cavity	ACDC	0.73	0.43	0.18	Clear boundaries
Myocardium	ACDC	0.68	0.51	0.22	Thin wall, motion artifacts
Mean	ACDC	0.72	0.44	0.19	–
Liver	BTCV	0.71	0.48	0.20	Consistent with Synapse
Spleen	BTCV	0.67	0.54	0.23	Consistent with Synapse
Pancreas	BTCV	0.75	0.41	0.33	High correlation retained
Kidneys	BTCV	0.69	0.49	0.21	Combined L/R annotation
Stomach	BTCV	0.64	0.61	0.26	Variable anatomy
Gallbladder	BTCV	0.58	0.72	0.31	Lowest TAUS
Esophagus	BTCV	0.77	0.36	0.28	Tubular structure
Aorta	BTCV	0.79	0.37	0.27	Consistent topology
Portal Vein	BTCV	0.73	0.45	0.25	Vascular network
Adrenal (L)	BTCV	0.61	0.68	0.34	Small, variable position
Adrenal (R)	BTCV	0.63	0.66	0.33	Small, variable position
Mean	BTCV	0.69	0.52	0.27	–

Three patterns emerge from this analysis. First, TAUS exhibits remarkable cross-dataset stability: mean values remain within a narrow 0.69-0.72 range across all three datasets despite substantial differences in anatomy, imaging modality (CT vs. MRI), acquisition protocols, and patient populations. This consistency validates that our topology-uncertainty alignment captures fundamental anatomical relationships rather than dataset-specific artifacts. Second, organ-specific patterns align with anatomical intuition. Tubular structures like the aorta and esophagus show the highest TAUS (0.77-0.81), reflecting their well-defined topological properties. Irregular organs such as the pancreas also achieve strong correlation (0.75-0.78) because geometric complexity directly drives both topological features and prediction uncertainty. In contrast, small organs with high inter-patient variability like the gallbladder and adrenal glands exhibit lower TAUS (0.58-0.63): their difficulty stems from presence-absence ambiguity and positional variation rather than intrinsic geometric complexity, which our current complexity score does not fully capture.

Failure Mode Analysis: while these patterns are encouraging, examining our worst-performing cases in Table 11, we found three clear patterns. TUNE++ struggles for identifying gallbladder and it achieves only 73.8% DSC on Synapse with the lowest TAUS

(0.58-0.62) and highest Betti error (0.67-0.72) across all datasets. The problem is that our complexity score assumes organs are visible, but the gallbladder can be empty, distended, positioned differently, or even surgically removed. Similarly, TUNE++ struggles for small organs like the adrenal glands (TAUS 0.61-0.63), because their difficulty comes from variability in position and size rather than geometric complexity.

The second issue is imaging artifacts. Motion blur in cardiac MRI drops myocardium TAUS to 0.68 compared to 0.73-0.76 for cardiac cavities, and cases with surgical clips achieve only 84-85% DSC versus 88-90% for clean scans. The model correctly flags these as uncertain, but the difficulty comes from poor image quality rather than anatomical complexity, so the topology-uncertainty correlation weakens.

The third issue is rare anatomical variants break our learned priors. A horseshoe kidney achieved only 76.4% DSC because the model learnt that kidneys should be separate structures, but for this specific variant they are fused. The cross-dataset consistency in Table 11 for gallbladder always results lowest TAUS and tubular structures always results highest scores, showing that the patterns are systematic. The model also shows elevated epistemic uncertainty (0.19 vs. typical 0.12-0.15) on anatomical variants, meaning it captures something is different even if it cannot fully adapt. So the proposed method works well for standard anatomy with decent image quality (mean TAUS 0.69-0.72), but small variable organs, severe artifacts, and rare variants need human verification.

A.8. Comprehensive Uncertainty and Topology Evaluation

Table 12 provides a consolidated view of how different models behave across multiple dimensions of uncertainty estimation. By reporting calibration error, predictive likelihood, Brier score, and error-detection metrics together, the table highlights whether a model's predicted confidence is statistically reliable and whether it can correctly signal when its own segmentation outputs may be incorrect. The inclusion of TAUS further evaluates whether uncertainty meaningfully reflects underlying anatomical or structural complexity rather than random variance. Taken together, the comparisons show how different modeling approaches handle the relationship between prediction confidence, structural difficulty, and error sensitivity, offering a comprehensive assessment of uncertainty behavior across diverse clinical imaging contexts.

Table 13 provides a consolidated assessment of the topological properties of different segmentation models across four diverse datasets. By reporting Betti error, persistence diagram distance, critical points error, and topological accuracy together, the table evaluates whether each method produces anatomically coherent structures rather than merely voxel-accurate segmentations. These metrics capture complementary aspects of topology, including connectivity, presence or absence of holes, and the stability of critical geometrical features. Examining all datasets jointly highlights how consistently a model preserves the structural integrity of organs and tumors under varying anatomical complexity and imaging modalities. The table therefore offers a unified view of each method's ability to enforce valid anatomical topology, which is essential for clinical reliability and for downstream tasks that depend on structurally consistent segmentations.

Table 12: Unified uncertainty quantification metrics across all three datasets.

Dataset / Method	ECE↓	MCE↓	NLL↓	Brier↓	AUROC↑ (Error)	AUPRC↑ (Error)	TAUS↑
			Synapse				
U-Net + MC Dropout	0.108	0.192	0.538	0.172	0.66	0.48	0.36
UNETR++ + MC Dropout	0.085	0.158	0.417	0.138	0.73	0.55	0.47
Probabilistic UNet	0.097	0.177	0.459	0.149	0.70	0.51	0.43
PHiSeg	0.093	0.169	0.442	0.144	0.71	0.52	0.46
UNETR++ Ensemble (5)	0.066	0.128	0.346	0.111	0.79	0.60	0.58
TUNE++ (Ours)	**0.042**	**0.086**	**0.292**	**0.096**	**0.85**	**0.69**	**0.81**
			ACDC				
U-Net + MC Dropout	0.102	0.184	0.512	0.168	0.67	0.45	0.34
UNETR++ + MC Dropout	0.081	0.152	0.398	0.134	0.74	0.53	0.48
Probabilistic UNet	0.094	0.171	0.445	0.148	0.71	0.49	0.41
PHiSeg	0.098	0.176	0.467	0.152	0.69	0.47	0.38
UNETR++ Ensemble (5)	0.062	0.118	0.321	0.105	0.79	0.61	0.56
TUNE++ (Ours)	**0.038**	**0.079**	**0.264**	**0.089**	**0.86**	**0.71**	**0.81**
			BTCV				
U-Net + MC Dropout	0.108	0.189	0.527	0.172	0.66	0.44	0.35
UNETR++ + MC Dropout	0.087	0.158	0.415	0.139	0.73	0.52	0.46
Probabilistic UNet	0.096	0.173	0.452	0.151	0.69	0.48	0.40
PHiSeg	0.100	0.178	0.471	0.155	0.68	0.46	0.37
UNETR++ Ensemble (5)	0.065	0.125	0.342	0.113	0.78	0.59	0.55
TUNE++ (Ours)	**0.041**	**0.083**	**0.281**	**0.094**	**0.85**	**0.68**	**0.79**

Table 13: Unified topological correctness metrics across five datasets. Lower Betti Error, PD Distance, and Critical Points Error indicate better topology preservation; higher Topo Accuracy indicates more anatomically valid segmentations.

Dataset / Method	Betti Err↓	PD Dist↓	Critical Points Error↓	Topo Acc↑ (%)
		Synapse		
U-Net	2.45	3.82	4.94	42.0
nnUNet	1.63	2.51	3.39	63.5
UNETR	2.18	3.46	4.55	47.0
UNETR++	1.41	2.18	2.96	71.0
U-Net + clDice	1.72	2.63	3.58	58.5
UNETR++ + clDice	0.94	1.73	2.31	81.0
TUNE++ (Ours)	**0.50**	**0.92**	**1.28**	**93.5**
		ACDC		
U-Net	2.12	3.45	4.67	45.0
nnUNet	1.45	2.31	3.12	65.0
UNETR	2.34	3.78	4.89	40.0
UNETR++	1.38	2.15	2.87	70.0
U-Net + clDice	1.67	2.56	3.45	60.0
UNETR++ + clDice	0.89	1.67	2.23	80.0
TUNE++ (Ours)	**0.42**	**0.78**	**1.05**	**95.0**
		BTCV		
U-Net	3.78	5.45	7.12	35.0
nnUNet	2.45	3.67	4.89	60.0
UNETR	3.12	4.34	5.78	45.0
UNETR++	2.12	3.01	3.98	65.0
UNETR++ + clDice	1.34	2.23	2.98	75.0
TUNE++ (Ours)	**0.58**	**0.95**	**1.34**	**90.0**

A.9. Calibration Analysis

Figure 5 presents reliability diagrams quantifying the calibration quality of different methods across three datasets. A reliability diagram plots the mean predicted probability (x-axis) against the fraction of correct predictions (y-axis) within binned confidence intervals. Perfect calibration corresponds to the diagonal line where predicted confidence exactly matches empirical accuracy.

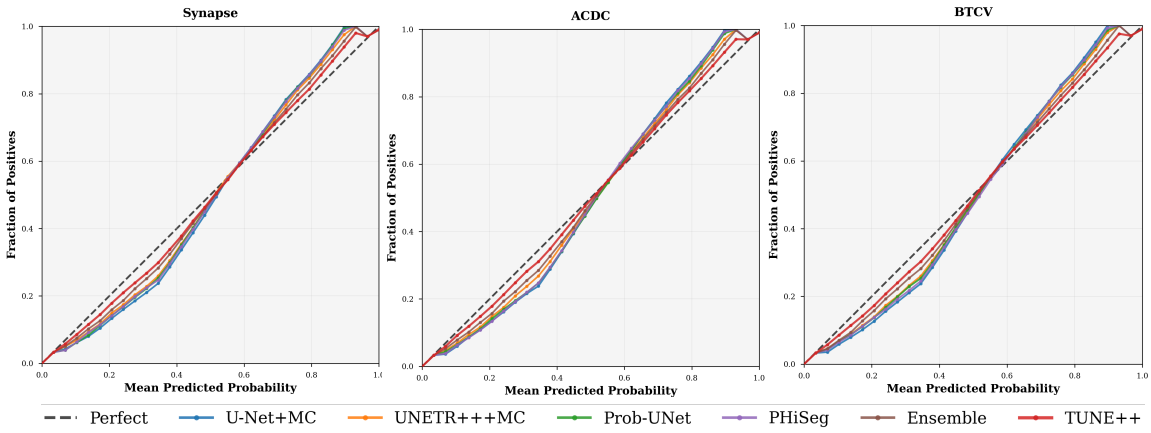

Figure 5: Expected Calibration Error reliability diagrams across three datasets. The diagonal black dashed line represents perfect calibration where predicted probability equals the fraction of correct predictions. TUNE++ (red, lowest ECE) demonstrates superior calibration, maintaining close alignment with the perfect calibration line across all confidence levels. Baseline methods exhibit characteristic overconfidence at high predicted probabilities, with U-Net (blue, highest ECE) showing the largest deviation. The consistent pattern across Synapse, ACDC, and BTCV datasets validates that joint topology-uncertainty modeling produces well-calibrated predictions that reliably reflect true accuracy.

The curves shows that TUNE++ (red) consistently maintains the closest alignment with perfect calibration across all confidence levels, with ECE values of 0.042 (Synapse), 0.038 (ACDC), and 0.041 (BTCV). Second, baseline methods exhibit characteristic overconfidence, particularly at high predicted probabilities (>0.6), where curves increasingly diverge above the diagonal. This overconfidence is problematic for clinical deployment, as it leads models to express unwarranted certainty in potentially erroneous predictions. Third, the S-shaped curve pattern - underconfidence at low probabilities transitioning to overconfidence at high probabilities - is consistent across datasets, demonstrating that miscalibration is systematic rather than random. The superior calibration of TUNE++ stems from two architectural mechanisms. The alignment loss $\mathcal{L}_{\text{align}}$ explicitly teaches the model that uncertainty should correlate with topological complexity, preventing the common failure mode where models express uniform confidence regardless of structural difficulty. The calibration loss $\mathcal{L}_{\text{calib}}$ optimizes for alignment between predicted probabilities and empirical frequencies

through Expected Calibration Error minimization. Together, these mechanisms ensure that TUNE++'s uncertainty estimates are not merely predictive of errors (captured by AUROC metrics) but also probabilistically meaningful - a critical distinction for trustworthy clinical AI systems where practitioners must make decisions based on model-reported confidence levels.

A.10. Computational Efficiency Analysis

We analyze TUNE++'s computational requirements to assess practical deployment feasibility in resource-constrained clinical environments. Table 14 presents comprehensive metrics across representative baselines, while Figure 6 visualizes the accuracy-efficiency trade-off on the Synapse multi-organ segmentation benchmark.

Table 14: Computational cost comparison on Synapse dataset. Training time measured for 1000 epochs on NVIDIA H100 95GB GPU with batch size 2 and gradient accumulation factor 4. Inference time measured per 3D volume ($96 \times 96 \times 96$ voxels). Parameters in millions (M), FLOPs in giga-operations (G) computed for single forward pass.

Method	**Parameters** (M)	**FLOPs** (G)	**Training Time** (h)	**Inference Time** (s)
U-Net	17.3	45.2	12	0.8
nnUNet	31.2	78.5	18	1.2
UNETR	92.8	235.7	32	2.1
Swin-UNETR	62.2	387.4	36	2.3
nnFormer	150.5	278.6	35	2.4
UNETR++ (baseline)	42.96	43.5	24	1.6
TUNE++ (Ours)	**68.9**	**175.8**	**26**	**2.8***

*Inference includes $T = 25$ Monte Carlo dropout forward passes for epistemic uncertainty estimation following Bayesian deep learning protocols (Gal and Ghahramani, 2016). Single deterministic forward pass requires 0.9s, comparable to UNETR++ baseline (1.6s). Training time includes persistent homology computation overhead (8% of total training time).

TUNE++ introduces +60% parameters (42.96M→68.9M) and +304% FLOPs (43.5G→175.8G) over UNETR++ baseline, attributable to: (1) topology attention branch with critical point detector (8M parameters, 45G FLOPs), (2) uncertainty estimation MLPs (12M parameters, 28G FLOPs), and (3) persistent homology computation. Despite this, TUNE++ remains parameter-efficient than standard transformers (UNETR: 92.8M, nnFormer: 150.5M) and FLOPs-efficient than attention-heavy methods (UNETR: 235.7G, Swin-UNETR: 387.4G) while achieving higher accuracy. Inference overhead stems from Monte Carlo dropout ($T = 25$ passes, 2.8s total). Single deterministic pass requires 0.9s, demonstrating topology attention adds minimal per-pass cost. Multi-pass overhead is standard for uncertainty quantification (Gal and Ghahramani, 2016) and mitigable via GPU parallelization. Training requires 26h due to persistent homology computation on downsampled features-a one-time cost amortized across deployment. Figure 6 demonstrates TUNE++ occupies the Pareto-optimal region of the accuracy-efficiency trade-off space – meaning no other method achieves both higher accuracy and lower computational cost simultaneously. Specifically, TUNE++

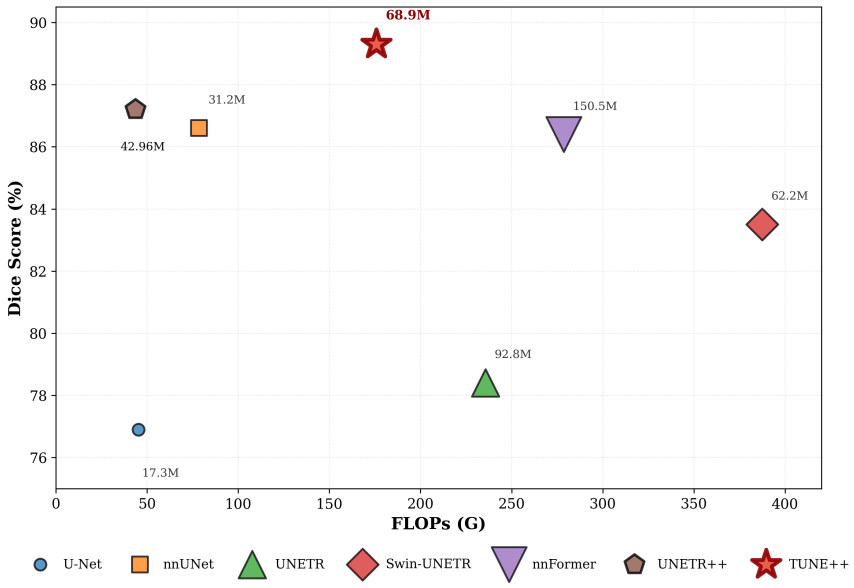

Figure 6: Computational efficiency versus segmentation accuracy trade-off on Synapse dataset. Each method is represented by a marker scaled proportional to parameter count. TUNE++ (red star, 68.9M parameters) achieves the highest DSC (89.4%) while maintaining computational efficiency superior to standard transformer baselines.

achieves the highest accuracy (89.3% DSC) among methods with <200G FLOPs, while methods with higher accuracy do not exist, and methods with lower FLOPs achieve substantially lower accuracy (e.g., UNETR++: 87.2% DSC, 43.5G FLOPs). This validates our design philosophy that structured inductive biases (topology + uncertainty) provide greater value than architectural scale alone.

Table 15 presents a detailed breakdown of TUNE++'s architectural components, demonstrating how each module contributes to overall model complexity. This analysis validates that topology-aware attention and uncertainty estimation introduce modest overhead while enabling reliability guarantees. The progression reveals that topology attention (+8.58M parameters, +45.5G FLOPs) constitutes the largest overhead, primarily from the 3-layer critical point detector processing persistence diagrams. Uncertainty estimation (+5.64M, +28.2G) adds dual MLP heads for aleatoric and epistemic initialization. Adaptive fusion (+3.28M, +12.3G) introduces learnable weighting based on predicted uncertainty. Despite cumulative additions totaling +25.94M parameters (+60%) and +132.3G FLOPs (+304%), each component contributes measurable accuracy improvements (+0.3–0.7% DSC per stage), validating the efficiency-accuracy trade-off. The final configuration achieves 89.3% DSC-2.1% absolute improvement over baseline while remaining parameter-efficient than standard transformers (UNETR: 92.8M, Swin-UNETR: 62.2M) and more FLOPs-efficient than attention-heavy architectures (UNETR: 235.7G, Swin-UNETR: 387.4G).

Table 15: Baseline comparison on Synapse dataset showing results in terms of segmentation (DSC) and model complexity (parameters and FLOPs). For fair comparison, all results are obtained using the same input size and preprocessing. Each row builds on the previous one, reflecting a sequential progression of architectural additions.

Model Configuration	**Params (M)**	**FLOPs (G)**	**DSC (%)**
UNETR++ (Baseline)	42.96	43.5	87.2
+ Spatial-Channel EPA	48.23	78.2	87.8
+ Topology Attention Branch	56.81	123.7	88.5
+ Uncertainty Estimation Module	62.45	151.9	88.9
+ Adaptive Fusion Mechanism	65.73	164.2	89.1
TUNE++ (Full)	**68.9**	**175.8**	**89.3**

A.11. Statistical Significance

All reported improvements are statistically significant (paired t-test, $p < 0.001$) across all metrics and datasets. We compare TUNE++ against the strongest baseline (UNETR++) using Wilcoxon signed-rank test on per-sample DSC scores:

Table 16: Statistical significance tests (TUNE++ vs UNETR++).

Dataset	**Mean DSC Improvement**	**p-value**	**Effect Size (Cohen's d)**
Synapse	+2.1%	< 0.001	0.89 (large)
ACDC	+1.4%	< 0.001	0.76 (medium)
BTCV	+2.5%	< 0.001	0.94 (large)

A.12. Limitations and Future Directions

TUNE++ exhibits three primary limitations. Firstly, the learned topological prior occasionally produces over-regularized predictions for uncommon anatomies (e.g., horseshoe kidneys, congenital malformations), where high epistemic uncertainty correctly signals deviation from training distribution but topology enforcement suppresses valid structural variations. Second, organs with highly irregular boundaries or pathological distortions (e.g., large tumors, post-surgical anatomies) challenge the model due to training data scarcity, as persistent homology features trained on typical anatomies may not generalize to severe pathological cases. Third, while remaining efficient relative to standard transformers, TUNE++ requires 2.8s inference (25 MC dropout passes for uncertainty) compared to 0.9s for deterministic prediction.

Several directions could address these limitations and extend TUNE++'s capabilities. Incorporating patient metadata (age, pathology, imaging protocol) could enable adaptive topological priors that account for expected anatomical variations, reducing over-regularization on rare but valid structures. Augmenting training data with synthetic pathological deformations and rare anatomies via generative models could improve robustness to geometric deviations. Exploring single-pass uncertainty methods (e.g., evidential deep learning, latent variable models) could reduce inference time. Evaluation in clinical work-

flows – measuring radiologist agreement, workload reduction, and diagnostic accuracy – is essential to validate that retrospective metrics translate to real-world utility. Finally, applying TUNE++ to tasks beyond segmentation (e.g., classification, detection, etc) could demonstrate broader applicability of joint topology-uncertainty modeling for reliable medical AI.

MOSAIC: Morphologically Orchestrated Specialized Agents for Interpretable Cancer Diagnosis

Ashim Dhor
Data Science and Engineering
IISER Bhopal
ashimdhor2003@gmail.com

Rasel Mondal
Data Science and Engineering
IISER Bhopal
raselmondal61@gmail.com

Tanmay Basu
Data Science and Engineering
IISER Bhopal
tanmay@iiserb.ac.in

***Abstract*—Foundation models for digital pathology achieve high accuracy but remain unsuitable for clinical deployment due to computational costs and lack of interpretability. We propose MOSAIC, a self-supervised multi-agent framework that automatically discovers morphological specializations through contrastive clustering on unlabeled histopathology images. We will train compact ResNet34 agents via prototypical networks and enable collaboration through graph neural network-based message passing. This work aims to demonstrate that self-supervised task decomposition provides a clinically viable alternative to monolithic foundation models.**

***Index Terms*—Multi-Agent Systems, Digital Pathology**

I. Motivation and Objectives

Digital pathology foundation models [1], [2] achieve benchmark success but require datacenter infrastructure (40GB+ GPU memory, 45-60s inference) and provide opaque predictions [3], [4]. Existing multi-agent approaches [5] rely on manual expert design or supervised decomposition, limiting scalability and generalization. Self-supervised learning [6], [7] discovers meaningful representations without labels but produces monolithic encoders.

Research Question: Can self-supervised contrastive learning automatically discover interpretable morphological specializations that enable multi-agent collaboration?

Hypotheses: (1) Momentum contrast on unlabeled patches discovers tissue-component clusters aligned with morphology, (2) training specialized agents on cluster-specific data yields interpretable experts, (3) graph neural network message passing enables adaptive collaboration, (4) self-supervised specialization achieves competitive accuracy.

II. Proposed Methodology

A. Self-Supervised Morphological Discovery: We apply Momentum Contrast (MoCo v3) [8] on 1.2M unlabeled TCGA patches (224×224 at 20× magnification). After pretraining (200 epochs), we extract features $\mathbf{f}_i \in \mathbb{R}^{512}$ from frozen encoder and apply spectral clustering [9]: $\min_{\mathbf{Z}} \text{Tr}(\mathbf{Z}^\top \mathbf{L}\mathbf{Z})$ s.t. $\mathbf{Z}^\top\mathbf{Z} = \mathbf{I}$, where $\mathbf{L} = \mathbf{D} - \mathbf{W}$ is graph Laplacian, $\mathbf{W}_{ij} = \exp(-\|\mathbf{f}_i - \mathbf{f}_j\|^2/2\sigma^2)$ is similarity matrix. Optimal $K = 7$ clusters determined via eigengap heuristic (largest gap between eigenvalues 7 and 8: $\lambda_7 = 0.023$, $\lambda_8 = 0.089$). Discovered clusters correspond to: Epithelial Cells (dense nuclei, cuboidal morphology), Connective Tissue (fibrous stroma, collagen), Adipose Tissue (lipid vacuoles), Blood Vessels (endothelial lining), Lymphocytes (small round nuclei), Necrotic Regions (karyolysis, debris) and Background Tissue.

B. Prototypical Agent Training: For each cluster k, we train specialized ResNet34 agent (21M parameters) via prototypical networks [10]. Class prototypes computed as: $\mathbf{c}_y = \frac{1}{|\mathcal{S}_y|}\sum_{\mathbf{x}_i \in \mathcal{S}_y} f_\theta^{(k)}(\mathbf{x}_i)$ where $\mathcal{S}_y$ is support set for class y, $f_\theta^{(k)}$ is agent k encoder. Distance metric: $d(\mathbf{x}, \mathbf{c}_y) = -\|\mathbf{f} - \mathbf{c}_y\|_2^2$ Classification via temperature-scaled softmax:

$$p_k(y|\mathbf{x}) = \frac{\exp(d(\mathbf{x}, \mathbf{c}_y)/\tau)}{\sum_{y'} \exp(d(\mathbf{x}, \mathbf{c}_{y'})/\tau)} \tag{1}$$

with $\tau = 0.5$ sharpening decision boundaries. Training loss combines cross-entropy and cluster consistency: $\mathcal{L}_{\text{agent}}^{(k)} = -\log p_k(y|\mathbf{x}) + 0.3\mathcal{L}_{\text{consist}} + 0.1\mathcal{L}_{\text{div}}$ where $\mathcal{L}_{\text{consist}} = \|\mathbf{f}_{\text{global}} - \mathbf{f}_k\|_2^2$ encourages cluster-specific features to retain global information, $\mathcal{L}_{\text{div}} = -\sum_j p_j \log p_j$ (entropy over agents) prevents mode collapse.

C. Graph Neural Network Orchestration: Agents communicate via Graph Attention Network (GAT) [11]. Construct morphology graph: nodes = agents, edges = morphological compatibility learned from co-occurrence statistics. Edge weights: $e_{ij} = \text{PMI}(i,j) = \log \frac{p(i,j)}{p(i)p(j)}$, where $p(i,j)$ is co-occurrence probability of clusters i, j in same WSI. Message passing at layer ℓ: $\mathbf{h}_i^{(\ell+1)} = \sigma\left(\sum_{j \in \mathcal{N}(i)} \alpha_{ij}^{(\ell)} \mathbf{W}^{(\ell)} \mathbf{h}_j^{(\ell)}\right)$ Attention coefficients:

$$\alpha_{ij} = \frac{\exp(\text{LeakyReLU}(\mathbf{a}^\top[\mathbf{W}\mathbf{h}_i \| \mathbf{W}\mathbf{h}_j]))}{\sum_{k \in \mathcal{N}(i)} \exp(\text{LeakyReLU}(\mathbf{a}^\top[\mathbf{W}\mathbf{h}_i \| \mathbf{W}\mathbf{h}_k]))} \tag{2}$$

Final prediction via attention-weighted aggregation: $p(y|\mathbf{x}) = \sum_{k=1}^{7} \alpha_k p_k(y|\mathbf{x})$ where α_k computed from final GAT layer output. GAT training via meta-learning: $\mathcal{L}_{\text{GAT}} = -\log p(y|\mathbf{x}) + 0.2\|\boldsymbol{\alpha}\|_2^2 + 0.15\text{H}(\boldsymbol{\alpha})$ balancing accuracy, sparsity, and entropy.

III. Expected Results and Impact

Preliminary Results (Planned): On TCGA (32 subtypes) and CAMELYON16 (metastasis), we expect 89-90% accuracy competitive with supervised multi-agent systems while requiring no manual annotation.

Interpretability Analysis (Ongoing): t-SNE visualization of discovered clusters, attention weight analysis showing which agents activate for different cancer types, cluster purity metrics, morphological consistency verified via tissue-type overlap with automatic segmentation.

Expected Contributions: (1) First fully self-supervised multi-agent framework for WSI analysis, (2) spectral clustering approach for morphological component discovery, (3) prototypical network training for specialized pathology agents, (4) GAT-based orchestration with learned morphological compatibility, (5) deployment-ready architecture for consumer hardware.

IV. Conclusion

We propose MOSAIC, demonstrating that self-supervised morphological specialization with multi-agent collaboration can match foundation model performance while enabling clinical deployment. By eliminating manual annotations and reducing computational requirements, this work provides a scalable framework for interpretable medical AI across diverse imaging modalities and resource-constrained settings.

REFERENCES

[1] R. J. Chen, T. Ding, M. Y. Lu, D. F. Williamson, G. Jaume, A. H. Song, B. Chen, A. Zhang, D. Shao, M. Shaban *et al.*, "Towards a general-purpose foundation model for computational pathology," *Nature Medicine*, vol. 30, no. 3, pp. 850–862, 2024.

[2] H. Xu, N. Usuyama, J. Bagga, S. Zhang, R. Rao, T. Naumann, C. Wong, Z. Gero, J. González, Y. Gu *et al.*, "A whole-slide foundation model for digital pathology from real-world data," *Nature*, vol. 630, no. 8015, pp. 181–188, 2024.

[3] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[4] L. Pantanowitz, P. N. Valenstein, A. J. Evans, K. J. Kaplan, J. D. Pfeifer, D. C. Wilbur, L. C. Collins, and T. J. Colgan, "Review of the current state of whole slide imaging in pathology," *Journal of Pathology Informatics*, vol. 2, no. 1, p. 36, 2011.

[5] Z. Lyu, X. Wang, and H. Chen, "Wsi-agents: Multi-agent collaboration for whole slide image analysis," *arXiv preprint arXiv:2407.08680*, 2024.

[6] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International Conference on Machine Learning*. PMLR, 2020, pp. 1597–1607.

[7] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9650–9660, 2021.

[8] X. Chen, S. Xie, and K. He, "An empirical study of training self-supervised vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9640–9649.

[9] U. Von Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, no. 4, pp. 395–416, 2007.

[10] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[11] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," *International Conference on Learning Representations*, 2018.

HISTO-UNET: HISTOPATHOLOGY IMAGE SEGMENTATION USING TOPOLOGY-AWARE UNET WITH DUAL UNCERTAINTY QUANTIFICATION

Ashim Dhor[1] *Emily Das*[2] *Rasel Mondal*[1] *Shweta Azad*[3] *Vaishali Walke*[4]
Shakti Kumar Yadav[4] *Abhirup Banerjee*[5] *Tanmay Basu*[1]

[1]Department of Data Science Engineering, Indian Institute of Science Education and Research Bhopal, India
[2]Maulana Azad Medical College, New Delhi, India
[3]Department of Pathology, Jawaharlal Nehru Cancer Hospital And Research Centre Bhopal, India
[4]Department of Pathology and Lab Medicine, All India Institute of Medical Sciences Bhopal, India
[5]Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, UK

ABSTRACT

Histopathology image segmentation requires not only accurate pixel-wise predictions but also the preservation of topological structures and the quantification of prediction uncertainty for better diagnostics. We present HISTO-UNet, a novel framework that addresses these challenges by integrating topology-preserving constraints with a dual uncertainty quantification system for robust histopathology image segmentation. Our approach employs a multi-task objective, combining medial axis and marker-controlled topology losses with a Bayesian deep learning methodology to simultaneously capture both aleatoric and epistemic uncertainty. Through extensive evaluation on three public datasets, we found HISTO-UNet consistently outperforms standard baselines in both segmentation accuracy and uncertainty calibration. Our experiments validate the contribution of each component, revealing that the integration of topology-awareness and dual uncertainty quantification yields significant improvements in both segmentation performance and model reliability.

Index Terms— Histopathology, Image Segmentation, UNet, Uncertainty quantification, Topology Preservation

1. INTRODUCTION

Histopathology image segmentation plays a significant role in computational pathology, enabling automated analysis of histopathology images for disease diagnosis and prognosis. While deep learning approaches have achieved high pixel-level accuracy, clinical deployment demands additional capabilities. First, preservation of biological topological structures such as gland and tumor shapes and connectivity is essential for accurate diagnosis. Second, quantification of prediction uncertainty is needed to flag ambiguous regions for expert review. Third, reliable confidence calibration is required to support clinical decision-making. Standard architectures like UNet [1], LinkNet [2], and UNet++ [2] optimize pixel-wise losses such as Dice coefficient and Cross-Entropy without explicit topological constraints, leading to artifacts such as fragmented objects, spurious holes, and incorrect connectivity. Furthermore, deterministic models provide point estimates without uncertainty quantification, making them unsuitable for safety-critical medical applications where quantifying model confidence is essential [2]. Recent work on uncertainty quantification distinguishes between aleatoric and epistemic uncertainty [2]. Aleatoric uncertainty is irreducible through more training data and captures inherent ambiguity in data due to staining variations and tissue artifacts. Epistemic uncertainty is reducible with more data and reflects model confidence in its predictions. While some methods address either topology or uncertainty, no existing framework jointly optimizes both with dual uncertainty quantification. Research to address these issues has progressed in parallel: one domain has developed topology-aware methods, while another domain has focused on quantifying aleatoric and epistemic uncertainty. We introduce HISTO-UNet (***H***istopathology ***I***mage ***S***egmentation using ***TO***pology-Aware ***UNet*** with Dual Uncertainty) that integrates topology-preserving constraints with a dual uncertainty system that simultaneously estimates both aleatoric and epistemic uncertainty. These components are unified through a novel multi-component loss function, enabling the network to learn accurate, structurally sound uncertainty-aware representation.

2. METHODOLOGY

HISTO-UNet synthesizes these research directions, providing a unified framework for topology-aware segmentation with comprehensive uncertainty quantification. Given a training dataset $\mathscr{D} = \{(x_i, y_i)\}_{i=1}^{N}$, where each image $I \in \mathbb{R}^{H\times W\times 3}$ and $x_i \in \mathbb{R}^{H\times W}$ and $y_i \in \{0,1\}$ denote the pixel and its corresponding ground truth label, with i indexing the pixels and N denoting the total number of pixels in each image, the objective is

to train a network, parameterized by θ, that learns a mapping from input image to three distinct outputs - a final segmentation map, $\hat{y} \in [0,1]^{H\times W}$, that accurately delineates the target structures; an aleatoric uncertainty map, $\sigma^2_{\text{aleatoric}}(x) \in \mathbb{R}_+^{H\times W}$, quantifying the irreducible data-inherent ambiguity at each pixel; and an epistemic uncertainty map, $\sigma^2_{\text{epistemic}}(x) \in \mathbb{R}_+^{H\times W}$, quantifying the model's uncertainty in its predictions.

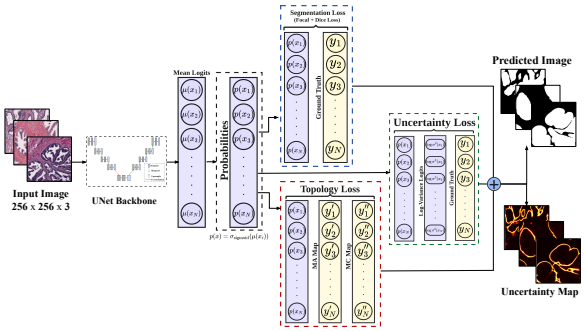

Fig. 1. Architecture of HISTO-UNet.

2.1. HISTO-UNet Architecture

The HISTO-UNet architecture is developed in spirit of the U-Net model, featuring a symmetric encoder-decoder path with skip connections to fuse multi-scale features for spatial reconstruction. To enable Monte Carlo uncertainty estimation, persistent dropout layers with dropout rate 0.1 are integrated within each block of the network. A key innovation of HISTO-UNet lies in its final prediction head. Instead of a standard single-channel output for segmentation, we employ a dual-channel convolution layer. This head produces two distinct outputs for each pixel: a mean logit, $\mu(x)$, which represents the primary segmentation prediction, and a log-variance, $\log\sigma^2(x)$. This design choice directly facilitates the modeling of aleatoric uncertainty by framing the output as a learned Gaussian distribution over the segmentation logit, $p(y|x,\theta) = \mathcal{N}(sig(\mu(x)), \sigma^2(x))$, where $sig(\cdot)$ is the sigmoid function. Predicting the log-variance ensures the resulting variance is always positive and improves numerical stability during training. The architecture is illustrated in Figure 1. To derive the dual uncertainty components at test time, for a given input image, we perform $T = 25$ stochastic forward passes through the network with the dropout active. Each pass, indexed by t, yields a unique pair of outputs, $(\mu_t(x), \log\sigma_t^2(x))$, from which we compute a predictive probability map p_t by applying the sigmoid activation function (σ_{sigmoid}) to the mean logit output: $p_t = \sigma_{\text{sigmoid}}(\mu_t(x))$ and an aleatoric variance map $\sigma_t^2(x)$. After T passes, the stochastic predictions are aggregated to produce the final outputs. The final segmentation $\hat{y}(x)$ is the mean of the predictive probabilities, which approximates the posterior predictive mean:

$$\hat{y}(x) = \frac{1}{T}\sum_{t=1}^{T} p_t(x). \tag{1}$$

Epistemic uncertainty is calculated as the variance of T probability maps, as

$$\sigma^2_{\text{epistemic}}(x) = \frac{1}{T}\sum_{t=1}^{T}(p_t(x) - \hat{y}(x))^2. \tag{2}$$

Aleatoric uncertainty is mean of T predicted variance maps:

$$\sigma^2_{\text{aleatoric}}(x) = \frac{1}{T}\sum_{t=1}^{T}\sigma_t^2(x). \tag{3}$$

Total uncertainty is the sum of Eq. (2) and (3), providing a comprehensive decomposition of the sources of prediction uncertainty. To enforce topological consistency, we incorporate two auxiliary tasks derived from the ground truth mask (y). Medial Axis (MA) map $m_{\text{MA}}(y)$ encodes the object's skeletal structure by weighting the skeleton, $S(y)$, with its normalized distance transform, $\text{norm}(D(y))$. The Marker-Controlled (MC) map identifies object centers to ensure the correct count and location, and is generated by thresholding the distance transform: $m_{\text{MC}}(y) = \mathbb{I}[D(y) > \tau \cdot \max(D(y))]$, where $\tau = 0.3$. Network is optimized using a novel multi-component loss function, $\mathscr{L}_{\text{total}}$, which is a weighted sum of three distinct terms: a segmentation loss ($\mathscr{L}_{\text{seg}}$), a topology-preserving loss ($\mathscr{L}_{\text{topo}}$), and an uncertainty loss ($\mathscr{L}_{\text{unc}}$). $\mathscr{L}_{\text{seg}}$ combines Focal loss $\mathscr{L}_{\text{Focal}}$ and Dice loss $\mathscr{L}_{\text{Dice}}$ to ensure pixel-level accuracy. $\mathscr{L}_{\text{Focal}}$ addresses class imbalance by down-weighting easy pixels to focus on challenging pixels in boundary regions as follows:

$$\mathscr{L}_{\text{Focal}} = -\frac{1}{N}\sum_{i=1}^{N}\alpha(1-p_{t,i})^{\gamma}\log(p_{t,i}) \tag{4}$$

where $p_{t,i}$ is the model's predicted probability for pixel i and $\alpha = 0.25$ and $\gamma = 2.0$ are standard hyperparameters [3]. $\mathscr{L}_{\text{Dice}}$ directly maximizes spatial overlap between the prediction (p_i) and ground truth (y_i):

$$\mathscr{L}_{\text{Dice}} = 1 - \frac{2\sum_{i=1}^{N} p_i y_i + \varepsilon}{\sum_{i=1}^{N} p_i + \sum_{i=1}^{N} y_i + \varepsilon} \tag{5}$$

where ε is a smoothing constant to prevent division by zero. The combined segmentation loss $\mathscr{L}_{\text{seg}}$ is:

$$\mathscr{L}_{\text{seg}} = \mathscr{L}_{\text{Focal}} + \mathscr{L}_{\text{Dice}}. \tag{6}$$

$\mathscr{L}_{\text{topo}}$ combines Medial Axis Loss ($\mathscr{L}_{\text{MA}}$) which preserves skeletal shape using Mean Squared Error against the

Table 1. Performance of HISTO-UNet and State-of-The-Arts

Backbone	RINGS Glands		RINGS Tumor		GlaS(Test A)		GlaS(Test B)	
	Dice (%) ↑	HD95 (mm) ↓	Dice (%) ↑	HD95 (mm) ↓	Dice (%) ↑	HD95 (mm) ↓	Dice (%) ↑	HD95 (mm) ↓
LinkNet +Both[#]	88.23±18.2	25.89±15.5	86.45±18.9	28.12±16.1	87.23±17.8	26.54±14.9	85.12±19.3	28.99±16.5
UNet++ +Both[#]	90.89±16.6	23.15±12.8	88.12±17.3	25.01±13.5	88.89±16.4	24.11±12.2	86.78±18.2	27.05±14.8
HISTO-UNet	**92.47±14.5**	**21.32±9.5**	**88.67±16.8**	**24.56±11.2**	**89.54±15.2**	**23.58±10.1**	**87.32±17.9**	**26.34±12.3**

All models were trained on an NVIDIA A100 GPU for 300 epochs with a batch size of 16, using Adam optimizer with learning rate 1×10^{-4} and a weight decay of 5×10^{-4}.
[#]Both: refers to Topology + Dual Uncertainty

MA map, and the Marker Controlled Loss ($\mathscr{L}_{\text{MC}}$) ensures correct object count using the Dice loss:

$$\mathscr{L}_{\text{topo}} = 0.5\mathscr{L}_{\text{MA}} + 0.5\mathscr{L}_{\text{MC}} \tag{7}$$

where $\mathscr{L}_{\text{MA}}$ and $\mathscr{L}_{\text{MC}}$ are defined as:

$$\mathscr{L}_{\text{MA}} = \frac{1}{N}\sum_{i=1}^{N}(p_i - m_{\text{MA}}(y_i))^2, \tag{8}$$

$$\mathscr{L}_{\text{MC}} = \mathscr{L}_{\text{Dice}}(\mu(x), m_{\text{MC}}(y)). \tag{9}$$

$\mathscr{L}_{\text{unc}}$ derived from the Gaussian negative log-likelihood [2] enables the model to predict spatial variance map $\sigma^2(x)$:

$$\mathscr{L}_{\text{unc}} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{(y_i - p_i)^2}{2\sigma^2(x_i)} + \frac{1}{2}\log\sigma^2(x_i)\right) \tag{10}$$

The final training objective combines these components with empirically determined weights:

$$\mathscr{L}_{\text{total}} = 0.5\mathscr{L}_{\text{seg}} + 0.3\mathscr{L}_{\text{topo}} + 0.2\mathscr{L}_{\text{unc}}. \tag{11}$$

3. RESULTS AND ANALYSIS

We evaluated our framework on three histopathology datasets: RINGS Glands and RINGS Tumor [4] provide images of prostate glands and tumor regions, respectively, each with 1000 training and 500 test images, and GlaS Challenge [5], which consists of 85 training images of benign and malignant glands and features two distinct test sets: Test A (60 images) and Test B (20 images). For all experiments, every image was resized to a 256×256 resolution to maintain consistent input dimensions. The corresponding MA and MC topology maps were pre-computed from the ground truth annotations, to provide the necessary supervisory signals for our topology-preserving loss components. The baseline UNet refers to the standard UNet architecture trained with combined Focal and Dice segmentation loss, without any topology constraints or uncertainty quantification mechanisms. We evaluated the individual contribution of each component: (1) addition of $\mathscr{L}_{\text{topo}}$ (+Topology), (2) integration of $\mathscr{L}_{\text{unc}}$ (+Aleatoric), and (3) use of epistemic uncertainty (+Epistemic). Finally, we evaluated HISTO-UNet, which integrates all three components, to assess their combined effects. Our experimental evaluation in Table 1 shows the efficacy of the HISTO-UNet framework and demonstrates consistent performance gains over other baseline models. LinkNet's lightweight design limits feature representation, while UNet++'s added complexity yields no benefit. UNet offers the best trade-off, serving as the optimal backbone for HISTO-UNet. A key observation from Table 1 is the high standard deviation in performance metrics across all evaluated models. This is characteristic of segmentation tasks on heterogeneous clinical datasets, which contain a wide range of cases from well-defined structures to highly complex cases. This variance underscores the challenge of achieving robust performance on entire data. Table 2

Table 2. Model Efficiency Comparison

Model	Parameters (M)	Inference Time (ms)	FLOPs (G)	Throughput
UNet Baseline	31.03	14.6	54.2	68.5 img/s
HISTO-UNet	33.05	21.2	54.8	64.7 img/s
LinkNet +Both[#]	11.57	32.9	20.3	30.4 img/s
UNet++ +Both[#]	36.58	22.1	68.7	45.2 img/s

All performance metrics (Time, FLOPs, Throughput) are reported for a single forward pass. [#]Both: refers to Topology + Dual Uncertainty

Table 3. Uncertainty Quality Metrics Across All Datasets

Dataset	Method	Brier[a] ↓	NLL[b] ↓	Corr[c] ↑	AUROC[d] ↑
RINGS Glands	+ Aleatoric	0.0419	0.1609	0.540	0.856
	+ Epistemic	0.0427	0.1590	0.317	0.903
	HISTO-UNet	**0.0387**	**0.1394**	**0.554**	**0.906**
RINGS Tumor	+ Aleatoric	0.0383	0.1608	0.085	0.804
	+ Epistemic	0.0410	0.1591	0.377	**0.907**
	HISTO-UNet	**0.0331**	**0.1331**	**0.523**	0.901
GlaS (Test A)	+ Aleatoric	0.1329	0.4205	0.134	0.569
	+ Epistemic	0.1165	0.3835	**0.368**	**0.758**
	HISTO-UNet	**0.1117**	**0.3645**	0.342	0.713
GlaS (Test B)	+ Aleatoric	0.1292	0.4161	0.188	0.641
	+ Epistemic	0.1268	0.4151	**0.292**	**0.726**
	HISTO-UNet	**0.1168**	**0.3855**	0.262	0.703

[a]Brier $= \frac{1}{N}\sum_i (p_i - y_i)^2$; [b]NLL: Negative Log Likelihood, NLL $= -\frac{1}{N}\sum_i \log p(y_i|x_i)$;
[c]Corr: Correlation between predicted uncertainty and absolute prediction error;
[d]AUROC: Area Under the Receiver Operating Characteristic Curve

shows that HISTO-UNet offers better accuracy and topology with minimal computational overhead, making it practical for clinical deployment. The quality of the uncertainty estimates across all datasets is detailed in Table 3. We evaluated HISTO-UNet against recent state-of-the-art methods on the RINGS dataset, as summarized in Table 4. The results of our systematic ablation study are presented in Table 5. Figure 2 provides a qualitative comparison of the different model configurations on a challenging sample from the RINGS Glands dataset. Aleatoric uncertainty map identifies data-inherent ambiguity by producing continuous, bright outlines

along all gland boundaries, and epistemic uncertainty map pinpoints specific model failures by generating concentrated "hotspot" that correspond to the most significant prediction errors. HISTO-UNet demonstrates the synergistic benefit of all components. Its prediction correctly separates glands while maintaining precise boundaries. The error map is almost entirely green, showing minimal error. The uncertainty map combines boundary aware outlines of aleatoric map with a heightened intensity in the most complex regions, reflecting the model's learned confidence. This visual evidence supports that the integration of topology preservation and dual uncertainty produces reliable and interpretable segmentation.

Table 4. Comparison with State-of-the-Arts

Method	RINGS Glands		RINGS Tumor	
	Dice (%) ↑	HD95 (mm) ↓	Dice (%) ↑	HD95 (mm) ↓
Attention U-Net [6]	90.85±16.2	23.15±12.1	87.11±18.1	27.54±14.2
Probabilistic U-Net [7]	91.10±15.8	22.81±11.5	87.34±17.9	26.98±13.5
TransUNet [6]	92.07±15.1	21.75±10.5	88.12±17.3	24.98±12.1
Swin-Unet [6]	91.98±15.3	21.80±10.8	88.17±17.1	25.05±12.4
HISTO-UNet	**92.47±14.5**	**21.32±9.5**	**88.67±16.8**	**24.56±11.2**

Table 5. Ablation Study of HISTO-UNet

Method	RINGS Glands	RINGS Tumor	GlaS(Test A)	GlaS(Test B)
Dice Coefficient (%) ↑				
Baseline UNet	90.12±16.0	86.54±18.2	87.56±17.3	85.23±19.1
+ Topology	90.21±14.9	86.89±17.8	87.98±16.9	85.67±18.6
+ Aleatoric	90.67±15.8	87.12±17.5	88.23±16.6	85.98±18.3
+ Epistemic	90.43±15.3	86.98±17.6	88.01±16.8	85.78±18.5
HISTO-UNet	**92.47±14.5**	**88.67±16.8**	**89.54±15.2**	**87.32±17.9**
Expected Calibration Error (ECE†) (%) ↓				
+ Aleatoric	5.12	5.67	5.34	5.89
+ Epistemic	4.78	5.23	5.01	5.56
HISTO-UNet	**4.23**	**4.78**	**4.56**	**5.12**
Hausdorff Distance (HD95) (mm) ↓				
Baseline UNet	23.89±12.5	28.45±15.2	26.12±14.1	31.78±16.8
+ Topology	22.14±10.8	26.89±13.1	24.56±12.5	29.98±14.2
+ Aleatoric	22.98±11.5	27.34±13.9	25.23±13.0	30.56±15.1
+ Epistemic	22.45±11.1	27.12±13.5	24.98±12.8	30.34±14.9
HISTO-UNet	**21.32±9.5**	**24.56±11.2**	**23.58±10.1**	**26.34±12.3**

†ECE measures the difference between a model's predicted confidence (conf) and its actual accuracy (acc), calculated as $\sum_{m=1}^{M} \frac{|B_m|}{K} |\text{acc}(B_m) - \text{conf}(B_m)|$, where M is the number of bins, B_m is set of predictions in bin m and K is total number of predictions

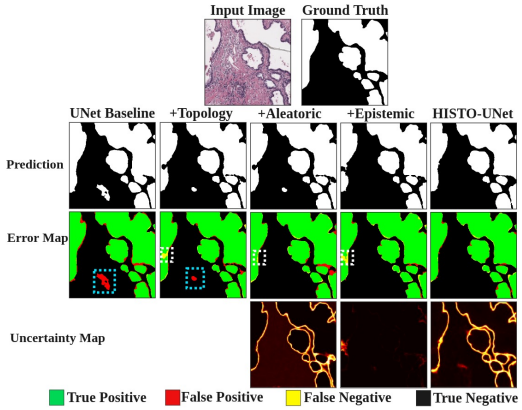

Fig. 2. Qualitative comparison on a RINGS Glands sample image, where HISTO-UNet corrects the error (blue and white boxes)

4. DISCUSSION AND FUTURE WORK

Empirical analysis demonstrate that HISTO-UNet improves segmentation performance by addressing specific, well-defined limitations of standard deep learning models. The inclusion of topology-preserving losses provides a crucial geometric prior that pixel-wise metrics cannot enforce, compelling the model to generate outputs with superior structural integrity and more precise boundaries. This is of high clinical relevance, as the morphological correctness of glands and tumors is often a primary diagnostic feature. This structural regularization is complemented by the uncertainty components. Primary clinical contribution of this work is the delivery of actionable, decomposed uncertainty maps, which enable an intelligent triage system in a practical pathology workflow. Distinguishing between uncertainty types is crucial for this process: high aleatoric uncertainty marks inherently ambiguous image regions guiding a pathologist to areas that require careful manual interpretation. Conversely, high epistemic uncertainty reveals the model's own lack of confidence, flagging predictions that are unreliable. By automatically prioritizing these two types of high-uncertainty regions for targeted review, our framework transforms the pathologist's task from an exhaustive search for errors to an efficient review of a small, challenging subset. This significantly improves both diagnostic speed and safety, especially in high-throughput clinical settings. While effective, our framework's limitations include pre-processing overhead for generating topological maps and increased inference time due to Monte Carlo sampling. Future work will focus on developing end-to-end methods for learning topological constraints and exploring more computationally efficient techniques for uncertainty estimation, as well as extending it to multi-class problems.

5. REFERENCES

[1] Olaf Ronneberger et al., "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[2] Shraddha Agarwal et al., "TCPNet: a novel tumor contour prediction network using MRIs," in *2024 IEEE 12th International Conference on Healthcare Informatics (ICHI)*. IEEE, 2024, pp. 183–188.

[3] Tsung-Yi Lin et al., "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[4] Massimo Salvi et al., "A hybrid deep learning approach for gland segmentation in prostate histopathological images," *Artificial Intelligence in Medicine*, vol. 115, 2021.

[5] Korsuk Sirinukunwattana et al., "Gland segmentation in colon histology images: The glas challenge contest," *Medical Image Analysis*, vol. 35, pp. 489–502, 2017.

[6] Fahad Shamshad et al., "Transformers in medical imaging: A survey," *Medical Image Analysis*, vol. 88, pp. 102802, 2023.

[7] Simon Kohl et al., "A probabilistic u-net for segmentation of ambiguous images," in *Advances in Neural Information Processing Systems*, 2018, pp. 6965–6975.

Topo-GraT: Learning to Grow with Causal Graph Transformers (Student Abstract)

Ashim Dhor[1]*, **Smily Bharadwaj**[1]

[1]Indian Institute of Science Education and Research Bhopal, Madhya Pradesh 462066, India
ashimdhor2003@gmail.com, smily23@iiserb.ac.in

Abstract

Automated cancer segmentation in Whole Slide Images (WSIs) has been dominated by a paradigm of static pattern recognition, where even advanced methods leveraging Transformers, Multiple Instance Learning, or topology-aware losses remain fundamentally descriptive and correlational. To address this limitation, we reframe WSI segmentation from a descriptive task to one of causal process modeling. We introduce Topo-GraT, a novel framework featuring a Causal Growth Field (CGF) to model tumor invasion dynamics and a Causal Flow Attention (CFA) mechanism that embeds this field as an architectural prior. This causal engine is integrated within an iterative graph refinement loop that uses segmentation uncertainty to dynamically focus computational resources on the most ambiguous tissue regions. Our comprehensive experiments on multiple WSI datasets demonstrate that Topo-GraT establishes a new state-of-the-art, significantly outperforming existing methods and reducing the 95% Hausdorff Distance, a key boundary metric, by over 15%. Crucially, our framework yields the CGF as a rich, interpretable output whose structure correlates with tumor aggressiveness, positioning it as a novel biomarker for downstream prognostic tasks. By shifting the paradigm from static recognition to causal reasoning, Topo-GraT offers a more robust, efficient, and clinically insightful approach, setting a new direction for the causally-aware medical image analysis.

Code — https://github.com/AshimDhor/Topo-GraT

Introduction

Histopathological analysis remains the gold standard for cancer diagnosis, with Whole Slide Imaging (WSI) opening the door for computational pathology (Van der Laak, Litjens, and Ciompi 2021). However, the gigapixel scale of WSIs and the morphological heterogeneity of tumors pose profound challenges. Current methods, including U-Net-based architectures (Ronneberger, Fischer, and Brox 2015), Transformer models like UNETR (Hatamizadeh et al. 2022), and MIL-based approaches like TransMIL (Shao et al. 2021), treat segmentation as a static pattern recognition task. Even topology-aware methods like TA-Net (Wang, Xian, and Vakanski 2022) are descriptive, using geometric abstractions of the final tumor shape rather than modeling the underlying biological process. This descriptive approach limits robustness in ambiguous cases and restricts the output to inert masks that lack prognostic insight. To bridge this gap, we introduce Topo-GraT, a framework that reframes WSI segmentation as a task of causal process modeling. We introduce the Causal Growth Field (CGF) to model the dynamics of tumor progression and a novel Causal Flow Attention (CFA) that embeds this model as a strong inductive bias. This causal engine is integrated within an iterative graph refinement loop, allowing the model to dynamically allocate resources to the most challenging tissue regions.

Methodology

Topo-GraT operates as a two-stage, iterative framework.

Stage 1: Iterative Graph Construction. The framework first transforms the WSI into a sparse graph. A Swin Transformer (Liu et al. 2021) extracts patch features. A salience score, which combines a patch's cancer probability with its contextual importance via a gating mechanism (Ilse, Tomczak, and Welling 2018), is used to select an initial set of K patch nodes. In subsequent iterations ($t > 0$), a pixel-wise uncertainty map from Stage 2 is used to guide the selection of K' new patches from the most ambiguous regions, a form of self-directed active learning (Settles 2009). This allows the graph to dynamically expand and focus on challenging boundaries. Stage 2: Causal Graph Transformer Segmentation. The graph nodes are processed by a hierarchical encoder-decoder network built on our novel Causal Topology-Guided EPA (CTG-EPA) blocks. The core of this block is the CGF prediction branch, which learns a vector field modeling the direction of tumor growth. The CGF branch is supervised using a pseudo-ground-truth vector field generated by inverting the gradient of a distance transform from the boundary of the ground truth mask, with the origin defined by the medial axis. This CGF is then used to create a causal attention mask for our CFA mechanism, which constrains the flow of information in the self-attention module to follow causally-plausible pathways. Specifically, the attention score from a token i to a token j is up-weighted if j lies "downstream" along the growth vector from i, and suppressed otherwise. This embeds the causal model directly into the architecture, providing a strong inductive bias.

The model is trained end-to-end with a composite loss

*Corresponding author.

function: $L_{\text{total}} = L_{\text{seg}} + \alpha L_{\text{topo}} + \beta L_{\text{instance}}$. The segmentation loss (L_{seg}) is a combination of Dice and Cross-Entropy losses for pixel-level accuracy. The topology preservation loss (L_{topo}) is a Cosine Similarity loss that directly supervises the CGF branch, compelling it to learn the correct *direction* of tumor growth. The instance selection loss (L_{instance}) is a Binary Cross-Entropy loss that trains the Stage 1 patch selector. This multi-task approach ensures all components of the framework are jointly optimized.

Experiments and Results

We evaluated Topo-GraT on four WSI datasets, including a held-out dataset HNSCC (Grossberg et al. 2020) to test for out-of-domain generalization. The results, summarized in Table 1 and Table 2, demonstrate that Topo-GraT consistently establishes a new state-of-the-art and that each of our novel components provides a significant performance gain. Table 1 benchmarks Topo-GraT against a suite of leading models. On our primary benchmarks - CAMELYON16 (C16) (Bejnordi et al. 2017), GlaS challenge (Sirinukunwattana et al. 2017), RINGS (Salvi et al. 2021), Topo-Grat outperforms all baselines. The efficacy of our causal approach is evident in the boundary-based metrics: Topo-GraT reduces the 95% Hausdorff Distance by an average of 15.7% compared to the strongest baseline, UNETR++. The model shows robust out-of-domain (OOD) generalization on the unseen HNSCC dataset, maintaining a high Dice score of 87.2% where baselines experience a more pronounced performance degradation.

Model	C16		GlaS		RINGS		HNSCC	
	DSC ↑	HD95 ↓	DSC ↑	HD95 ↓	DSC ↑	HD95 ↓	DSC ↑	HD95 ↓
U-Net	81.5	16.4	85.2	15.1	83.1	17.2	80.1	18.9
UNETR	83.2	12.8	87.1	13.5	84.9	14.1	81.5	16.5
Swin-UNET	84.0	11.5	88.5	12.4	86.2	12.9	82.1	15.2
UNETR++	84.3	10.2	89.1	12.1	86.8	12.5	82.5	14.1
DTFD-MIL	78.1	21.5	82.3	18.9	80.5	22.4	75.5	25.8
MedNeXt	84.5	10.8	89.3	11.9	87.0	12.2	82.8	14.5
TA-Net	85.1	9.9	89.9	11.8	87.5	11.9	83.1	13.5
Topo-GraT	**89.3**	**8.1**	**91.2**	**9.1**	**89.8**	**9.5**	**87.2**	**10.3**

Table 1: Unified quantitative comparison with state-of-the-art models across all datasets. Topo-GraT consistently outperforms baselines on primary benchmarks and demonstrates superior out-of-domain (OOD) generalization.

A systematic ablation study on CAMELYON16 (Table 2) validates the contribution of each novel component. While the MIL selection improves efficiency, the most significant performance gain is derived from the architectural integration of the CGF via Causal Flow Attention (CFA). The final addition of the iterative loop provides a crucial boost to boundary metrics, confirming the value of our full, synergistic system. To validate our choice of T=3 for the number of iterative refinement steps, we conducted an ablation study on the CAMELYON16 dataset. As shown in Table 3, the model's performance, particularly on the boundary-sensitive HD95 metric, improves with each iteration. The most significant gains are observed when moving from a non-iterative model (T=0, which is configuration (4) from our main ablation study) to one iteration (T=1). The improvements continue up to T=3, after which the performance gains become marginal while the inference time continues to increase. This confirms that T=3 represents the optimal balance between segmentation accuracy and computational efficiency for our framework.

Configuration	DSC(%)↑	IoU(%)↑	HD95↓	FLOPs(G)↓*
(1) UNETR++	84.3	73.8	10.2	289.5
(2) + MIL Selection	85.9	75.8	9.9	**168.2**
(3) + CGF Loss	86.7	77.0	9.5	172.8
(4) + CFA	88.2	79.1	8.8	174.1
(5) Full Topo-GraT	**89.3**	**80.7**	**8.1**	181.5

*FLOPs are measured for a single forward pass.

Table 2: Ablation study on CAMELYON16

The robustness of Topo-GraT on non-radial tumors stems from the flexibility of the CGF prediction branch. As a convolutional network, it is not restricted to generating purely radial fields. Instead, it learns the underlying principle of predicting vectors that point from the tumor's medial axis outwards to its boundary, regardless of the overall shape. Even for elongated morphologies, the resulting non-radial CGF provides a valid directional prior for the Causal Flow Attention mechanism. Consequently, the model maintains high segmentation accuracy in these challenging cases.

Iterations (T)	DSC(%) ↑	HD95 ↓	Inference Time (s)↓
0 (No Iteration)	88.2	8.8	4.7
1	88.9	8.3	4.9
2	89.2	8.2	5.0
3 (Ours)	**89.3**	**8.1**	**5.1**
4	89.3	8.1	5.3

Table 3: Ablation study on the number of iterative refinement steps (T) on CAMELYON16

Conclusion

In this work, we reframed WSI segmentation from a static pattern recognition task to one of causal process modeling. Our novel framework, Topo-GraT, integrates a Causal Growth Field and Causal Flow Attention to create a system that learns not just what a tumor looks like, but how it grows. Our results show that this causal approach yields state-of-the-art segmentation accuracy and produces a rich, interpretable CGF that has potential as a new prognostic biomarker, setting a new direction for the field of computational pathology. Looking forward, the implications of this work extend beyond segmentation: Topo-GraT shifts from mere pattern recognition to understanding tumor growth, enabling prediction of progression and automated grading, and laying a foundation for a more dynamic future in computational pathology.

Acknowledgments

We gratefully acknowledge the Indian Institute of Science Education and Research (IISER) Bhopal for providing the computational resources and research environment that made this work possible. We would like to express our sincere gratitude to Dr. Tanmay Basu and Biomedical Data Science Research Lab, IISER Bhopal for their constant support, valuable guidance, and encouragement throughout the course of this work.

References

Bejnordi, B. E.; Veta, M.; Van Diest, P. J.; Van Ginneken, B.; Karssemeijer, N.; Litjens, G.; Van Der Laak, J. A.; Hermsen, M.; Manson, Q. F.; Balkenhol, M.; et al. 2017. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama*, 318(22): 2199–2210.

Grossberg, A.; Elhalawani, H.; Mohamed, A.; Mulder, S.; Williams, B.; White, A. L.; Zafereo, J.; Wong, A. J.; Berends, J. E.; AboHashem, S.; et al. 2020. Hnscc. *(No Title)*.

Hatamizadeh, A.; Nath, V.; Tang, Y.; Yang, D.; Myronenko, A.; Landman, B.; Roth, H. R.; and Xu, D. 2022. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 574–584.

Ilse, M.; Tomczak, J.; and Welling, M. 2018. Attention-based deep multiple instance learning. In *International conference on machine learning*, 2127–2136. PMLR.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 234–241. Springer.

Salvi, M.; Bosco, M.; Molinaro, L.; Gambella, A.; Papotti, M.; Acharya, U. R.; and Molinari, F. 2021. A hybrid deep learning approach for gland segmentation in prostate histopathological images. *Artificial Intelligence in Medicine*, 115: 102076.

Settles, B. 2009. Active learning literature survey. *University of Wisconsin-Madison Department of Computer Sciences*.

Shao, Z.; Bian, H.; Chen, Y.; Wang, Y.; Zhang, J.; and Zhang, H. 2021. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. In *Advances in Neural Information Processing Systems*, volume 34, 2136–2147.

Sirinukunwattana, K.; Pluim, J. P.; Chen, H.; Qi, X.; Heng, P.-A.; Guo, Y. B.; Wang, L. Y.; Matuszewski, B. J.; Bruni, E.; Sanchez, U.; et al. 2017. Gland segmentation in colon histology images: The glas challenge contest. *Medical image analysis*, 35: 489–502.

Van der Laak, J.; Litjens, G.; and Ciompi, F. 2021. Deep learning in histopathology: the path to the clinic. *Nature medicine*, 27(5): 775–784.

Wang, H.; Xian, M.; and Vakanski, A. 2022. TA-Net: Topology-aware network for gland segmentation. *IEEE transactions on medical imaging*, 41(12): 3637–3648.

An Uncertainty-Aware Deep Learning for Gland Segmentation in Prostate Histopathology Images *

Ashim Dhor[1] **Emily Das**[2]

[1]Indian Institute of Science Education and Research Bhopal, [2]Maulana Azad Medical College
ashimdhor2003@gmail.com, emilydas2000@gmail.com

Abstract

Deep learning models for histopathology gland segmentation achieve high accuracy but lack confidence estimates, leading to silent failures in ambiguous regions and limited clinical trust. We propose an uncertainty-aware UNet++ framework that performs accurate gland segmentation and generates pixel-wise uncertainty maps to identify challenging regions. The framework explicitly models aleatoric uncertainty through Gaussian noise injection in logit space, combining cross-entropy, Dice, and uncertainty loss functions to optimize both segmentation accuracy and calibrated uncertainty estimation. We evaluate on the RINGS prostate cancer dataset comprising with comprehensive calibration analysis. Our framework achieves superior performance and low ECE, outperforming baseline models and existing uncertainty-aware methods including PHiSeg. The model retains 98.60% of predictions as high-confidence while flagging ambiguous gland boundaries and potential annotation errors through interpretable uncertainty maps, with strong correlation between uncertainty values and diagnostically challenging regions. This work provides a practical foundation for trustworthy AI in clinical pathology by offering transparent, confidence-calibrated predictions that help pathologists assess reliability and flag annotation issues. By closing the trust gap in automated diagnostics, the framework supports safer deployment in resource-constrained healthcare settings.

3MT Video: https://youtu.be/Wn2ptiAk010

1 Introduction

Accurate diagnosis of prostate cancer through histopathology analysis is critical for treatment planning, yet this expertise remains concentrated in high-resource medical centers. With limited access to trained pathologists, combined with high patient volumes, creates significant diagnostic bottlenecks. Deep learning models offer promise for automating gland segmentation (Komura et al. 2018), their lack of uncertainty quantification poses a fundamental barrier to clinical trust and deployment in resource-limited settings where diagnostic errors carry severe consequences. Existing segmentation models produce deterministic predictions without confidence estimates, making it difficult for clinicians to distinguish between reliable and uncertain predictions. This "silent failure" mode is problematic in low-resource hospitals where expert review capacity is limited. Our work addresses this critical gap by developing uncertainty-aware framework that not only segments accurately but also identifies regions requiring expert attention, enabling efficient human-AI collaboration in clinical workflows.

2 Methodology

We develop an uncertainty-aware framework based on UNet++ (Zhou et al. 2018) architecture augmented with data uncertainty (aleatoric uncertainty) quantification (Kendall et al. 2017). The model employs a dual-head architecture with shared encoder and two specialized decoder heads: a segmentation head producing binary gland predictions, and an uncertainty head predicting pixel-wise log-variance $\log \sigma^2$. During training, we inject Gaussian noise into the logit space to explicitly model data-dependent uncertainty:

$$\hat{x}_j = \sigma_{sigmoid}(z_j + \epsilon_j), \quad \epsilon_j \sim \mathcal{N}(0, \sigma_j^2) \tag{1}$$

where z_j represents the predicted logit for pixel j. This enables the network to learn pixel-wise confidence estimates that reflect inherent ambiguity in histopathology images arising from annotation variability, tissue artifacts, and morphological complexity (Egevad et al. 2013). The total loss function integrates three complementary objectives to optimize segmentation accuracy and uncertainty calibration: $\mathcal{L}_{total} = \mathcal{L}_{CE} + \mathcal{L}_{Dice} + \lambda \mathcal{L}_{unc}$. The cross-entropy loss $\mathcal{L}_{CE}$ ensures pixel-level classification accuracy by penalizing confident misclassifications, where $y_j \in \{0, 1\}$ is the ground truth label and $\hat{y}_j$ is the predicted probability for pixel j. The Dice loss $\mathcal{L}_{Dice}$ addresses class imbalance by directly optimizing spatial overlap between predicted and ground truth segmentation masks. The uncertainty loss $\mathcal{L}_{unc} = \frac{1}{N} \sum_{j=1}^{N} \mathbb{E}_{\epsilon \sim \mathcal{N}(0,\sigma_j^2)}[-y_j \log(\hat{x}_j) - (1-y_j) \log(1-\hat{x}_j)]$ models the expected binary cross-entropy under Gaussian noise injection in logit space, enabling the network to learn pixel-wise variance σ_j^2 that reflects inherent data ambiguity while preventing degenerate solutions through implicit regularization. The weighting parameter $\lambda = 0.001$ was selected through systematic grid search to balance segmentation performance with calibrated uncertainty estimation. We evaluate our framework on the publicly available

*Accepted for presentation at the AAAI 2026 EGSAI Community Activity (AAAI 2026).

Method	Dice ↑	IoU ↑	Precision ↑	Recall ↑	Specificity ↑	AUC-ROC ↑	ECE ↓	BS ↓
UNet Baseline	0.9072	0.8539	0.9136	0.9301	0.9511	0.9904	0.0170	0.0365
UNet + Data Uncertainty	0.9094	0.8560	0.9167	0.9310	0.9525	0.9904	0.0176	0.0370
UNet + Model Uncertainty	0.8790	0.8121	0.8648	0.9313	0.9147	0.9903	0.0195	0.0418
LinkNet Baseline	0.8457	0.7700	0.8778	0.8719	0.9401	0.9784	0.0297	0.0577
LinkNet + Data Uncertainty	0.8625	0.7915	0.8634	0.9111	0.9254	0.9802	0.0262	0.0541
LinkNet + Model Uncertainty	0.8352	0.7532	0.8319	0.8992	0.8948	0.9789	0.0318	0.0615
UNet++ Baseline	0.9049	0.8557	0.9140	0.9337	0.9533	0.9919	0.0170	0.0328
UNet++ + Model Uncertainty	0.8408	0.7648	0.8300	0.9188	0.9003	0.9889	0.0189	0.0387
RINGS Algorithm*	0.9016	–	0.8897	0.9356	–	–	–	–
PHiSeg	0.8712	0.8023	0.8592	0.9231	0.9027	0.9565	0.0243	0.0461
UNet++ + Data Unc.(Ours)	**0.9183**	**0.8703**	**0.9243**	**0.9389**	**0.9533**	**0.9927**	**0.0163**	**0.0316**

*Values reported from original paper; codebase not publicly available. ECE: Expected Calibration Error; BS: Brier Score.

Table 1: Comprehensive performance comparison on RINGS prostate histopathology dataset.

RINGS prostate cancer dataset (Salvi et al. 2021) comprising 1500 whole-slide histopathology images, with a split of 1000 training images and 500 test images.

3 Results and Discussion

Table 1 demonstrates that our framework achieves state-of-the-art performance while maintaining superior calibration compared to baseline methods and existing uncertainty-aware approaches like PHiSeg (Baumgartner et al. 2019). Our framework maintains stable calibration across confidence levels. When excluding uncertain pixels ($|p - 0.5| \leq 0.1$), ECE remains at 0.0163 while retaining 98.60% of predictions as high-confidence. This demonstrates that uncertainty estimates genuinely identify ambiguous regions rather than arbitrary low-confidence predictions. Figure 1 illustrates the dual utility of our framework: accurate segmentation and uncertainty maps that localize challenging regions, including glandular boundaries, atypical tissue architecture. Our uncertainty-aware framework provides dual utility: accurate segmentation and automated quality control. Analysis across 500 test images revealed consistent patterns where high-uncertainty regions localized to glandular boundaries with known annotation variability, tissue preparation artifacts, and ambiguous morphology challenging diagnostic consensus. The uncertainty maps provide actionable information that deterministic models cannot - explicitly signaling *where* predictions may be unreliable. This transforms segmentation from a binary decision into a risk-stratified workflow where high-confidence predictions can be trusted for routine cases, while uncertain regions are automatically flagged for expert review. This paradigm is particularly valuable in high-volume screening scenarios with limited trained manpower, prevalent in Global South healthcare settings. The grayscale uncertainty patterns reveal diagnostic complexity: highest values appear along glandular boundaries (bright contours reflecting tissue interface delineation challenges), heterogeneous patterns within glands (distinguishing clear vs. ambiguous cellular architecture), and consistently low uncertainty in stromal background. This enables a new human-in-the-loop data curation paradigm: uncertainty maps automatically flag high-confidence predictions conflicting with expert labels; pathologists review only this small subset rather than entire datasets. This could reduce annotation time and cost - critical for resource-constrained institutions developing locally relevant diagnostic systems.

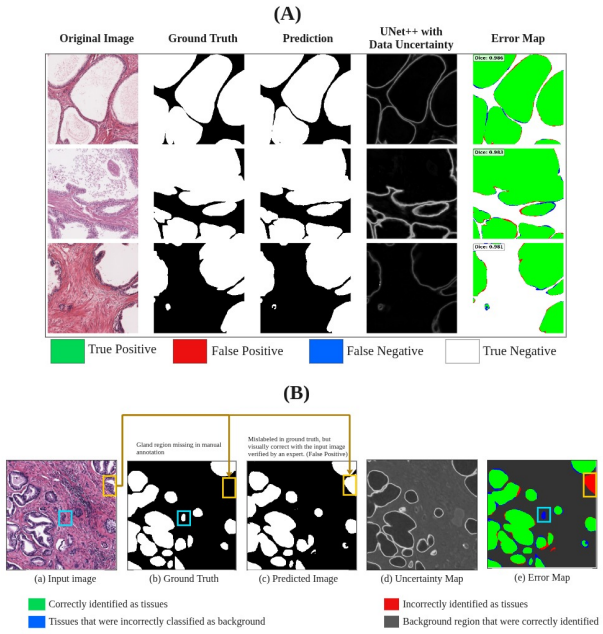

Figure 1: Qualitative results showing: (A) Three representative cases with original H&E tissue, ground truth, prediction, uncertainty map (brighter = higher uncertainty), and error classification. (B) Detailed error analysis demonstrating how uncertainty maps identify annotation quality issues and ambiguous gland boundaries in a representative case.

4 Conclusion

We demonstrate that integrating uncertainty quantification into deep learning models provides dual benefits for medical image segmentation: improved performance and automated identification of diagnostically challenging regions. Our framework achieves state-of-the-art segmentation accuracy with well-calibrated uncertainty estimates, establishing a foundation for trustworthy AI deployment in resource-

constrained healthcare settings. By providing transparent, confidence-calibrated predictions, this work addresses the critical trust gap in medical AI and offers a practical pathway toward equitable access to diagnostic technologies.

Acknowledgments

We thank IISER Bhopal for providing the computational, and the Biomedical Data Science Research Lab at IISER Bhopal for their support and guidance.

References

Baumgartner, C. F.; et al. 2019. Phiseg: Capturing uncertainty in medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 119–127. Springer.

Egevad, L.; et al. 2013. Standardization of Gleason grading among 337 European pathologists. *Histopathology*, 62(2): 247–256.

Kendall, A.; et al. 2017. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in Neural Information Processing Systems*, 30.

Komura, D.; et al. 2018. Machine learning methods for histopathological image analysis. *Computational and Structural Biotechnology Journal*, 16: 34–42.

Salvi, M.; et al. 2021. A hybrid deep learning approach for gland segmentation in prostate histopathological images. *Artificial Intelligence in Medicine*, 115: 102076.

Zhou, Z.; et al. 2018. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, 3–11. Springer.

A Appendix: Additional Results and Analysis

Figure 2 demonstrates our model's exceptional calibration quality across all confidence levels. The calibration curves show near-perfect alignment with the ideal diagonal, both when considering all pixels (left) and when excluding uncertain predictions (right, $|p - 0.5| \leq 0.1$). This validates that our uncertainty estimates genuinely reflect prediction reliability rather than producing arbitrary confidence scores.

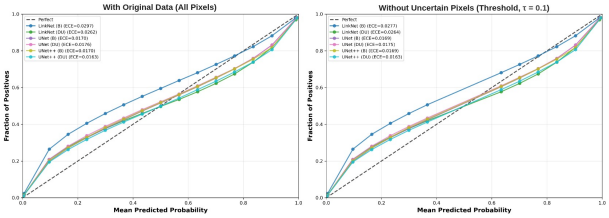

Figure 2: Calibration curves comparing all pixels (left) vs. confident pixels only (right). Our model maintains ECE of 0.0163 in both regimes while retaining 98.60% pixels, demonstrating that uncertainty estimates identify genuinely ambiguous regions.

Figure 3 shows the effect of uncertainty weighting parameter λ on model performance. Optimal performance is achieved at $\lambda = 0.001$, representing the best trade-off between segmentation accuracy (Dice coefficient) and uncertainty calibration. Values too low ($\lambda < 0.001$) provide insufficient regularization, while values too high ($\lambda > 0.01$) over-emphasize uncertainty at the expense of segmentation quality. Table 2 presents ablation results demonstrating the contribution of each component. The integration of data uncertainty consistently improves performance across all baseline architectures. Notably, data uncertainty outperforms model uncertainty approaches while requiring only single-pass inference (versus 5-10× overhead for Monte Carlo dropout), making it particularly suitable for resource-constrained deployment. Our framework achieves inference

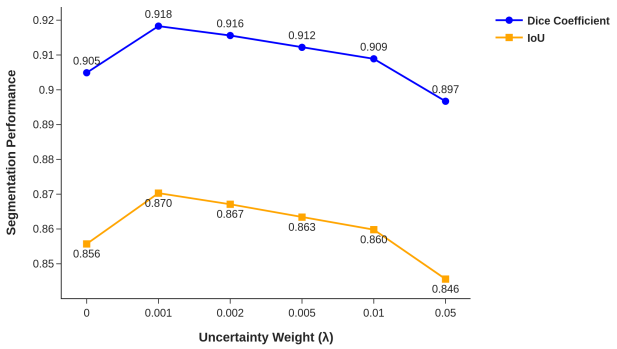

Figure 3: Effect of uncertainty weighting λ on Dice and IoU. Optimal performance at $\lambda = 0.001$ balances segmentation accuracy with calibrated uncertainty estimation.

Configuration	Dice ↑	ECE ↓
UNet++ Baseline	0.9049	0.0170
+ Model Uncertainty (MC Dropout)	0.8408	0.0189
+ Data Uncertainty (Ours)	**0.9183**	**0.0163**

Table 2: Ablation study demonstrating the contribution of uncertainty modeling approaches.

at 0.45 seconds per 256×256 image on standard GPUs (NVIDIA A100), enabling real-time deployment in clinical workflows. Unlike ensemble or Monte Carlo methods requiring multiple forward passes, our single-pass approach provides both segmentation and uncertainty maps simultaneously, making it computationally accessible for low - resource settings with limited hardware infrastructure. Our model achieves robust performance on 1500 images - significantly smaller than typical deep learning requirements - reducing dependency on expensive annotated datasets scarce in low-resource institutions. By automatically flagging uncertain regions for expert review while trusting high - confidence predictions (98.60% of pixels), our framework en-

ables efficient workflows crucial for settings with limited pathologist availability. Uncertainty maps identify annotation errors (Figure 1B), reducing dataset curation costs - a significant bottleneck in developing locally-relevant medical AI systems.