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: Under Review for 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, 2020) estimate prediction confidence but ignore anatomical structures. Topology-preserving approaches (Hu et al., 2019; Shit et al., 2021; Clough et al., 2022) 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. Method

Given a 3D medical image $\mathbf{x} \in \mathbb{R}^{H\times W\times D}$ where H, W, D are height, width, and depth, respectively, we learn a mapping $f_\theta : \mathbb{R}^{H\times W\times D} \to \mathcal{Y} \times \mathcal{U} \times \mathcal{T}$ that simultaneously produces segmentation $\mathbf{y} \in \{0,1\}^{H\times W\times D\times C}$ for C anatomical classes, aleatoric uncertainty $\boldsymbol{\sigma}_a^2 \in \mathbb{R}^{+H\times W\times D\times C}$, epistemic uncertainty $\boldsymbol{\sigma}_e^2 \in \mathbb{R}^{+H\times W\times D\times C}$, and a topological descriptor $\mathbf{t} \in \mathcal{T}$ encoding structural properties. We formalize the learning objective as:

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

where $\mathcal{L}_{\text{align}}$ enforces correlation between uncertainty $\boldsymbol{\sigma}$ and topological complexity $\mathbf{t}$ – our key innovation creating synergy between uncertainty and topology.

2.1. Architecture Overview:

TUNE++ adopts a hierarchical encoder-decoder architecture with four spatial resolutions $\{H/4, H/8, H/16, H/32\}$ and corresponding channel dimensions. This multi-scale design captures both global organ-level context at coarse resolutions and fine boundary details at higher resolutions. An overview of the model architecture is illustrated in Figure 1. Following ViT (Dosovitskiy et al., 2021), we divide the input volume into non-overlapping 3D patches of size $(P_h, P_w, P_d) = (4, 4, 2)$ where P_h, P_w, P_d denote patch height, width, and depth respectively, creating $N = \frac{H\cdot W\cdot D}{P_h\cdot P_w\cdot P_d} = \frac{HWD}{32}$ tokens. The asymmetric patch size accommodates anisotropic medical image resolution where axial slices are typically thicker than in-plane pixels. Each patch is linearly projected to $C_1 = 32$ dimensions with learnable positional embeddings encoding spatial structure. Each encoder stage contains – Downsampling layer: stride-2 with $3 \times 3 \times 3$ convolution reducing spatial dimensions while doubling

channels, TUPA block: topology-uncertainty aware paired attention, and Topology extraction: persistent homology computation on feature maps yielding persistence diagram PD_s at scale s. The topology extraction module computes Euclidean Distance Transform (EDT) of intermediate features, treating them as height functions for sublevel set filtration. This yields persistence diagrams $\text{PD}_s = \{(b_i, d_i)\}$ where birth-death pairs (b_i, d_i) represent topological features (connected components, holes, voids) with persistence $|d_i - b_i|$ indicating significance. The decoder mirrors the encoder with four stages – upsampling uses $2 \times 2 \times 2$ transposed convolutions with stride 2, doubling spatial dimensions while halving channel count. Skip connections from corresponding encoder stages are concatenated before TUPA blocks, following UNet's design (Ronneberger et al., 2015). The final decoder output (resolution $H/2 \times W/2 \times D/2$, channels $C_1 = 32$) is processed by four parallel heads: a segmentation head applying $\text{Conv}3 \times 3 \times 3(\text{Conv}1 \times 1 \times 1(\cdot))$, followed by bilinear upsampling to the original resolution, softmax, and producing $\mathbf{y} \in \mathbb{R}^{H \times W \times D \times C}$; an aleatoric uncertainty Head using $\text{MLP}(256 \to 128 \to C)$ followed by softplus to produce $\boldsymbol{\sigma}_a^2 \in \mathbb{R}+^{H \times W \times D \times C}$, where softplus ensures positivity of variance; an epistemic uncertainty Initialization Head using $\text{MLP}(256 \to 128 \to C)$ followed by softplus to generate $\boldsymbol{\sigma}_e^2$, serving as a learned prior with final epistemic uncertainty computed via Monte Carlo (MC) dropout at inference; and a topology descriptor head that computes persistence diagram on final segmentation $\mathbf{y}$.

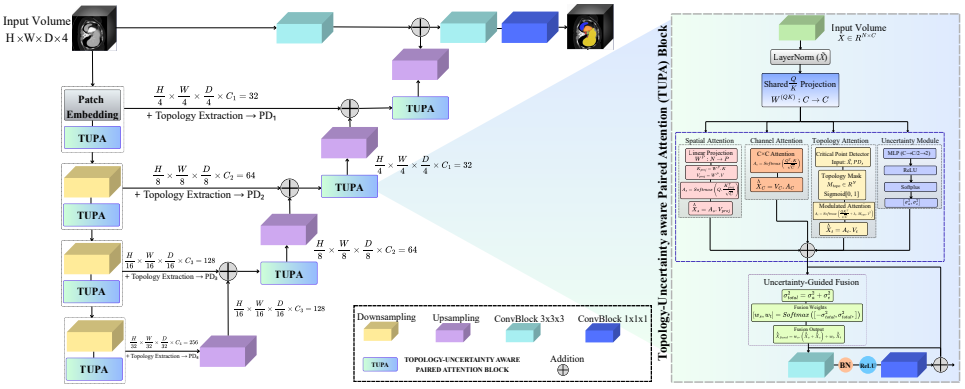

Figure 1: TUNE++ architecture overview.

2.2. Topology-Uncertainty Aware Paired Attention (TUPA)

TUPA block is the core of our framework. 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 and C is the channel dimension, we first apply layer normalization:

$$\tilde{\mathbf{X}} = \text{LayerNorm}(\mathbf{X}). \tag{2}$$

We then compute *shared* query and key projections as,

$$\mathbf{Q}_{\text{shared}} = \mathbf{K}_{\text{shared}} = \mathbf{W}^{QK}\tilde{\mathbf{X}} \in \mathbb{R}^{N \times C} \tag{3}$$

where $\mathbf{W}^{QK} \in \mathbb{R}^{C \times C}$ is a learnable projection matrix. The standard self-attention has $\mathcal{O}(N^2 C)$ complexity due to computing the $N \times N$ attention matrix. We reduce this to

$\mathcal{O}(NPC)$ where $P \ll N$ by projecting keys and values to a lower-dimensional space as,

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

$$\mathbf{V}_{\text{proj}}^{\text{spatial}} = (\mathbf{W}^P (\mathbf{V}_{\text{spatial}})^T)^T \in \mathbb{R}^{P \times C}, \tag{5}$$

$$\hat{\mathbf{X}}_{\text{spatial}} = \text{Softmax}\left(\frac{\mathbf{Q}_{\text{shared}} \mathbf{K}_{\text{proj}}^T}{\sqrt{C}}\right) \mathbf{V}_{\text{proj}}^{\text{spatial}}, \tag{6}$$

where $\mathbf{V}_{\text{spatial}} = \mathbf{W}^{V_{\text{spatial}}} \tilde{\mathbf{X}} \in \mathbb{R}^{N \times C}$ is spatial value projection, $\mathbf{W}^{V_{\text{spatial}}} \in \mathbb{R}^{C \times C}$ is value projection matrix, and $\mathbf{W}^P \in \mathbb{R}^{P \times N}$ is the learned dimensionality reduction. Further channel attention models feature interdependencies can be computed as,

$$\mathbf{A}_c = \text{Softmax}\left(\frac{\mathbf{Q}_{\text{shared}}^T \mathbf{K}_{\text{shared}}}{\sqrt{C}}\right) \in \mathbb{R}^{C \times C}, \tag{7}$$

$$\hat{\mathbf{X}}_c = \mathbf{V}_c \mathbf{A}_c, \tag{8}$$

where $\mathbf{V}_c = \mathbf{W}^{V_c} \tilde{\mathbf{X}}$ is channel value projection.

2.2.1. Topology Attention

We identify topologically critical regions using a Critical Point Detector:

$$\mathbf{M}_{\text{topo}} = \text{CriticalPointDetector}(\tilde{\mathbf{X}}, \mathbf{t}_s) \in \mathbb{R}^N \tag{9}$$

where $\mathbf{t}_s$ is the persistence diagram extracted at stage s. The detector (3-layer CNN) assigns high scores to organ boundaries, junctions, and critical points. Topology-modulated attention is computed as:

$$\mathbf{A}_t = \text{Softmax}\left(\frac{\mathbf{Q}_{\text{shared}} \mathbf{K}_{\text{shared}}^T}{\sqrt{C}} + \lambda_t \mathbf{M}_{\text{topo}} \mathbf{1}^T\right) \in \mathbb{R}^{N \times N}, \tag{10}$$

$$\hat{\mathbf{X}}_t = \mathbf{A}_t \mathbf{V}_t, \tag{11}$$

where $\mathbf{V}_t = \mathbf{W}^{V_t} \tilde{\mathbf{X}} \in \mathbb{R}^{N \times C}$ and $\lambda_t = 0.3$ controls topological bias strength, empirically determined via grid search over $\{0.1, 0.2, 0.3, 0.4, 0.5\}$ to balance data-driven attention with structural constraints – values below 0.2 provide insufficient topological enforcement (high Betti error), while values above 0.4 cause over-regularization that suppresses adaptive learning. The broadcast operation $\mathbf{M}_{\text{topo}} \mathbf{1}^T$ upweights attention to anatomically critical regions, with large persistence $|d_i - b_i|$ indicating significant structures.

2.2.2. Uncertainty Estimation

Parallel to attention branches, we estimate voxel-wise uncertainty:

$$[\boldsymbol{\sigma}_a^2, \boldsymbol{\sigma}_{e,\text{init}}^2] = \text{MLP}(\tilde{\mathbf{X}}) \in \mathbb{R}^{N \times 2} \tag{12}$$

where $\sigma_{e,\text{init}}^2$ initializes epistemic uncertainty, with final σ_e^2 computed via MC dropout at inference. Fusion weights are computed via:

$$[w_s, w_t] = \text{Softmax}([-\boldsymbol{\sigma}_{\text{total}}^2, \boldsymbol{\sigma}_{\text{total}}^2]) \in \mathbb{R}^{N \times 2} \tag{13}$$

where $\mathbf{w}_s, \mathbf{w}_t \in \mathbb{R}^{N\times 1}$ are per-voxel weights for spatial/channel and topology attention. Low uncertainty yields balanced fusion ($w_s \approx w_t \approx 0.5$); high uncertainty emphasizes topology ($w_t \to 1$, $w_s \to 0$). The three attention outputs are fused as:

$$\hat{\mathbf{X}}_{\text{fused}} = w_s \odot (\hat{\mathbf{X}}_s + \hat{\mathbf{X}}_c) + w_t \odot \hat{\mathbf{X}}_t \tag{14}$$

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

$$\mathbf{X}_{\text{out}} = \text{Conv}_{1\times1\times1}(\text{BN}(\text{ReLU}(\text{Conv}_{3\times3\times3}(\hat{\mathbf{X}}_{\text{fused}})))). \tag{15}$$

2.3. Loss Functions

Training TUNE++ balances five complementary objectives through a weighted total loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{seg}} + \lambda_1\mathcal{L}_{\text{topo}} + \lambda_2\mathcal{L}_{\text{unc}} + \lambda_3\mathcal{L}_{\text{calib}} + \lambda_4\mathcal{L}_{\text{hier}} \tag{16}$$

where $\lambda_1 = 0.3$, $\lambda_2 = 0.2$, $\lambda_3 = 0.1$, and $\lambda_4 = 0.15$ are determined through grid search over the validation set, prioritizing topology (λ_1) as it provides structural constraints, followed by uncertainty (λ_2) for reliability, with calibration (λ_3) and hierarchical consistency (λ_4) as regularizers. Ablation analysis (Table 4) demonstrates each component's contribution. Segmentation loss combines Dice and cross-entropy loss functions as,

$$\mathcal{L}_{\text{seg}} = \mathcal{L}_{\text{Dice}} + \mathcal{L}_{\text{CE}} = 1 - \frac{1}{C}\sum_{c=1}^{C}\frac{2\sum_{i=1}^{N_{\text{vox}}} p_{i,c}g_{i,c} + \epsilon}{\sum_{i=1}^{N_{\text{vox}}} p_{i,c} + \sum_{i=1}^{N_{\text{vox}}} g_{i,c} + \epsilon} - \frac{1}{N_{\text{vox}}}\sum_{i=1}^{N_{\text{vox}}}\sum_{c=1}^{C} g_{i,c}\log(p_{i,c}) \tag{17}$$

where N_{vox} is the number of voxels, C is the number of classes, $p_{i,c} \in [0,1]$ is the predicted probability, $g_{i,c} \in \{0,1\}$ is ground truth, and $\epsilon = 10^{-5}$ is a smoothing constant to prevent division by zero. The Topology Preservation Loss combines persistent homology, Betti numbers, and critical points as:

$$\mathcal{L}_{\text{topo}} = \mathcal{L}_{\text{PH}} + 0.5\mathcal{L}_{\text{Betti}} + 0.3\mathcal{L}_{\text{critical}}. \tag{18}$$

Persistent Homology Loss measures 2-Wasserstein distance between persistence diagrams PD_{pred} and PD_{gt} as follows,

$$\mathcal{L}_{\text{PH}} = W_2(\text{PD}_{\text{pred}}, \text{PD}_{\text{gt}}) = \left(\min_{\phi}\sum_{x\in\text{PD}_{\text{pred}}} \|x - \phi(x)\|_2^2\right)^{1/2} \tag{19}$$

where ϕ is an optimal bijection computed via Kuhn-Munkres algorithm (Munkres, 1957). Betti Number Loss penalizes incorrect topological invariants:

$$\mathcal{L}_{\text{Betti}} = \sum_{k=0}^{2} |\beta_k(Y_{\text{pred}}) - \beta_k(Y_{\text{gt}})| \tag{20}$$

where $Y_{\text{pred}}, Y_{\text{gt}} \in \{0,1\}^{H\times W\times D\times C}$ are binarized segmentation masks, and $\beta_0, \beta_1, \beta_2$ count connected components, loops/holes, and voids respectively. Critical Points Loss penalizes displacement of topologically critical points from EDT:

$$\mathcal{L}_{\text{critical}} = \frac{1}{|\mathcal{C}_{\text{gt}}|} \sum_{j\in\mathcal{C}_{\text{gt}}} \min_{j'\in\mathcal{C}_{\text{pred}}} \|\mathbf{c}_j^{\text{gt}} - \mathbf{c}_{j'}^{\text{pred}}\|_2 \tag{21}$$

where $\mathcal{C}_{\text{gt}}$ and $\mathcal{C}_{\text{pred}}$ are sets of critical points in ground truth and predicted segmentations respectively, and $\mathbf{c}_j$ denotes the 3D spatial coordinates of critical point j. Uncertainty Loss decomposes uncertainty – aleatoric, epistemic, and alignment components:

$$\mathcal{L}_{\text{unc}} = \mathcal{L}_{\text{aleatoric}} + \mathcal{L}_{\text{epistemic}} + 0.5\mathcal{L}_{\text{align}}. \tag{22}$$

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

$$\mathcal{L}_{\text{aleatoric}} = \frac{1}{N_{\text{vox}}C} \sum_{i=1}^{N_{\text{vox}}} \sum_{c=1}^{C} \left(\frac{\|p_{i,c} - g_{i,c}\|^2}{2\sigma_{a,i,c}^2} - \log(\sigma_{a,i,c}^2 + \epsilon) \right). \tag{23}$$

Epistemic Uncertainty enforces consistency across $K = 25$ MC dropout samples:

$$\mathcal{L}_{\text{epistemic}} = \text{KL}(p_{\text{single}} \| p_{\text{MC}}) = \frac{1}{N_{\text{vox}}C} \sum_{i,c} p_{i,c}^{\text{single}} \log \frac{p_{i,c}^{\text{single}}}{p_{i,c}^{\text{MC}}}. \tag{24}$$

Topology-Uncertainty Alignment creates synergy by correlating uncertainty with topological complexity:

$$\mathcal{L}_{\text{align}} = \frac{1}{N_{\text{vox}}} \sum_{i=1}^{N_{\text{vox}}} \|\sigma_{\text{total},i}^2 - C_{\text{topo},i}\|_2^2, \quad C_{\text{topo},i} = w_b B_i + w_j J_i + w_a A_i, \tag{25}$$

where $B_i \in \{0,1\}$ indicates boundaries, $J_i \in \mathbb{Z}_+$ counts organ junctions, $A_i \in [0,1]$ measures topological anomalies, with weights $w_b = 1.0$, $w_j = 2.0$, and $w_a = 3.0$ reflecting increasing topological complexity – boundaries are baseline structural features, junctions involve multiple organs requiring stronger enforcement, and topological anomalies (spurious holes, disconnections) represent the most severe violations warranting highest penalty. The Calibration Loss combines Expected Calibration Error (ECE) and Brier Score as,

$$\mathcal{L}_{\text{calib}} = \text{ECE} + 0.5 \cdot \text{Brier} = \sum_{m=1}^{M} \frac{|B_m|}{N_{\text{vox}}} |\text{acc}(B_m) - \text{conf}(B_m)| + \frac{0.5}{N_{\text{vox}}C} \sum_{i,c} (p_{i,c} - g_{i,c})^2 \tag{26}$$

where predictions are discretized into $M = 10$ confidence bins B_m. Hierarchical Topology Consistency ensures topology consistency across encoder stages:

$$\mathcal{L}_{\text{hier}} = \sum_{s=2}^{4} W_2(\text{PD}_s, \text{Downsample}(\text{PD}_{s-1})) \tag{27}$$

where PD_s is extracted at stage s and $\text{Downsample}(\cdot)$ recomputes the persistence diagram at the next scale. At inference, we perform MC dropout with $K = 25$ stochastic forward passes, each producing a prediction $\mathbf{y}_k$ and aleatoric estimate $\boldsymbol{\sigma}_{a,k}^2$. The final prediction is obtained as the mean output $\mathbf{y}_{\text{final}} = \frac{1}{K}\sum_{k=1}^{K} \mathbf{y}_k$, while aleatoric uncertainty is estimated as $\boldsymbol{\sigma}_a^2 = \frac{1}{K}\sum_{k=1}^{K} \boldsymbol{\sigma}_{a,k}^2$. Epistemic uncertainty is computed from the predictive variance $\boldsymbol{\sigma}_e^2 = \frac{1}{K}\sum_{k=1}^{K} (\mathbf{y}_k - \mathbf{y}_{\text{final}})^2$, and total uncertainty is finally obtained as $\boldsymbol{\sigma}_{\text{total}}^2 = \boldsymbol{\sigma}_a^2 + \boldsymbol{\sigma}_e^2$.

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:

$$\text{TAUS} = \text{Pearson}(\sigma^2_{\text{total}}, C_{\text{topo}}) \tag{28}$$

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.3.

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.

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

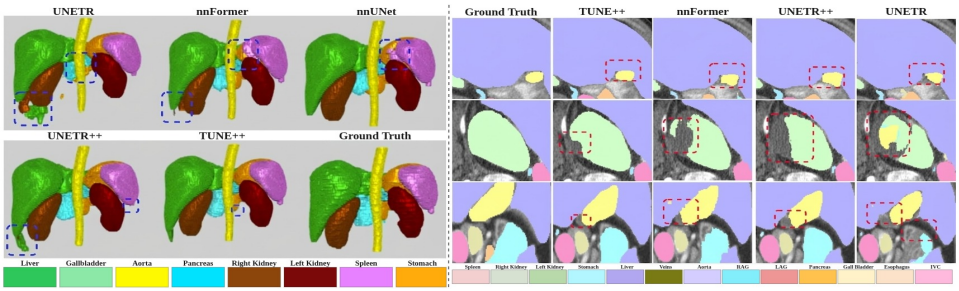

Figure 2: Qualitative results on Synapse dataset. **Left:** 3D renderings showing topological correctness. Blue dashed boxes highlight inaccurate segmentation. **Right:** 2D slices showing segmentation.

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.

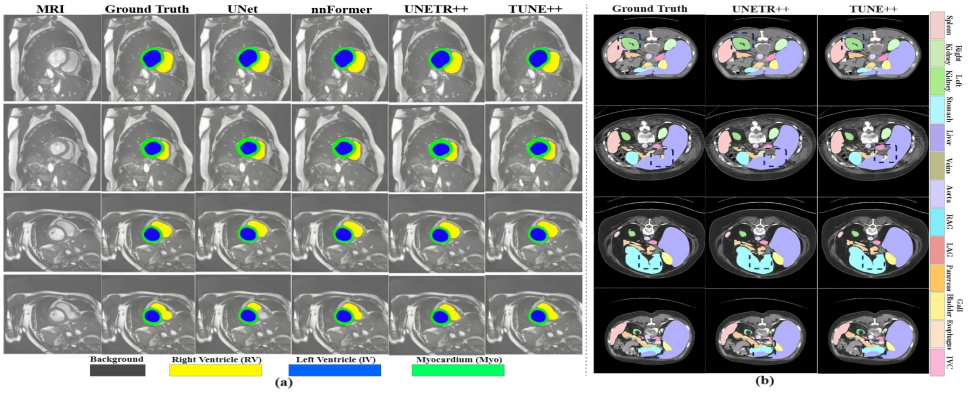

Figure 3: **(a)** Qualitative comparison on ACDC cardiac segmentation. **(b)** Multi-slice qualitative comparison on BTCV dataset.

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 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.

Detailed uncertainty (Table 6) and topology (Table 7) evaluations are provided in Appendix A.5, confirming TUNE++'s superior calibration and topological correctness across all datasets. Table 4 presents ablation across all datasets and reveals critical insights: adding

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.

Table 4: Comprehensive ablation study across all datasets.

Configuration	Synapse DSC	ACDC DSC	BTCV DSC	Mean DSC↑	Mean Betti↓	Mean ECE↓	Mean TAUS↑
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**

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 7, Appendix A.5) 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 10, Appendix A.8).

Loss weight selection (Appendix A.4, Figure 16) 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.6) 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 7) 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 8, Figure 6, Appendix A.7) 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.9.

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.

References

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. 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, 2022.

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. *Medical Image Computing and Computer Assisted Intervention*, pages 48–56, 2020.

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 et al. Metrics reloaded: Recommendations for image analysis validation. *arXiv preprint arXiv:2206.01653*, 2022.

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, Yakub 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*, 2021.

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, and Christian Wachinger. 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) 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., 2022) 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: $1.37\times1.37\times10$ 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 (± 15), scaling (0.9–$1.1\times$), horizontal/vertical flips, elastic deformation (displacement $\sigma = 10$, control points $3 \times 3 \times 3$); Intensity: brightness adjustment (± 0.2), contrast scaling (0.8–$1.2\times$), 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. Evaluation Metrics

A.3.1. Segmentation Accuracy Metrics

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

$$\text{DSC} = \frac{2|Y_{\text{pred}} \cap Y_{\text{gt}}|}{|Y_{\text{pred}}| + |Y_{\text{gt}}|} \tag{29}$$

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., 2021):

$$\text{NSD}(\tau) = \frac{|\mathcal{B}^{\tau}_{\text{pred}} \cap \mathcal{B}_{\text{gt}}| + |\mathcal{B}^{\tau}_{\text{gt}} \cap \mathcal{B}_{\text{pred}}|}{|\mathcal{B}_{\text{pred}}| + |\mathcal{B}_{\text{gt}}|} \tag{30}$$

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. (2022).

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

$$\text{HD95} = \max\{d_{95}(Y_{\text{pred}}, Y_{\text{gt}}), d_{95}(Y_{\text{gt}}, Y_{\text{pred}})\} \tag{31}$$

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:

$$\text{MASD} = \frac{1}{2}\left(\frac{1}{|\mathcal{B}_{\text{pred}}|}\sum_{b\in\mathcal{B}_{\text{pred}}} d(b, \mathcal{B}_{\text{gt}}) + \frac{1}{|\mathcal{B}_{\text{gt}}|}\sum_{b\in\mathcal{B}_{\text{gt}}} d(b, \mathcal{B}_{\text{pred}})\right) \tag{32}$$

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

A.3.2. Uncertainty Quantification Metrics

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

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N} |\text{acc}(B_m) - \text{conf}(B_m)| \tag{33}$$

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):

$$\text{MCE} = \max_{m=1,\ldots,M} |\text{acc}(B_m) - \text{conf}(B_m)| \tag{34}$$

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:

$$\text{NLL} = -\frac{1}{N}\sum_{i=1}^{N} \log p(y_i^{\text{gt}}|x_i) \tag{35}$$

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, 1950):

$$\text{Brier} = \frac{1}{N_{\text{vox}}C}\sum_{i=1}^{N_{\text{vox}}}\sum_{c=1}^{C}(p_{i,c} - g_{i,c})^2. \tag{36}$$

Brier score combines calibration with sharpness.

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

$$\text{AUROC} = P(\sigma_{\text{total}}(x_{\text{error}}) > \sigma_{\text{total}}(x_{\text{correct}})) \tag{37}$$

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):

$$\text{AUPRC} = \int_0^1 \text{Precision}(r)\, dr \tag{38}$$

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.3.3. Topological Correctness Metrics

Betti Number Error measures discrepancy in topological invariants:

$$\text{BettiError} = \sum_{k=0}^{2} |\beta_k(Y_{\text{pred}}) - \beta_k(Y_{\text{gt}})| \tag{39}$$

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:

$$\text{PD Dist} = W_2(\text{PD}_{\text{pred}}, \text{PD}_{\text{gt}}). \tag{40}$$

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:

$$\text{Crit. Pts Err} = \frac{1}{|\mathcal{C}_{\text{gt}}|} \sum_{j \in \mathcal{C}_{\text{gt}}} \min_{j' \in \mathcal{C}_{\text{pred}}} \|\mathbf{c}_j^{\text{gt}} - \mathbf{c}_{j'}^{\text{pred}}\|_2 \tag{41}$$

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:

$$\text{TopoAcc} = \frac{1}{N_{\text{test}}} \sum_{i=1}^{N_{\text{test}}} \mathbb{1}[\text{BettiError}(i) = 0] \tag{42}$$

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:

$$\text{TAUS} = \text{Pearson}(\sigma^2_{\text{total}}, C_{\text{topo}}) \tag{43}$$

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.4. Loss Weight Selection

Loss weights $\{\lambda_1, \lambda_2, \lambda_3, \lambda_4\}$ in the total loss (Equation 16) are determined through systematic grid search on the Synapse validation set. Figure 4 presents comprehensive sensitivity analysis demonstrating the selection rationale.

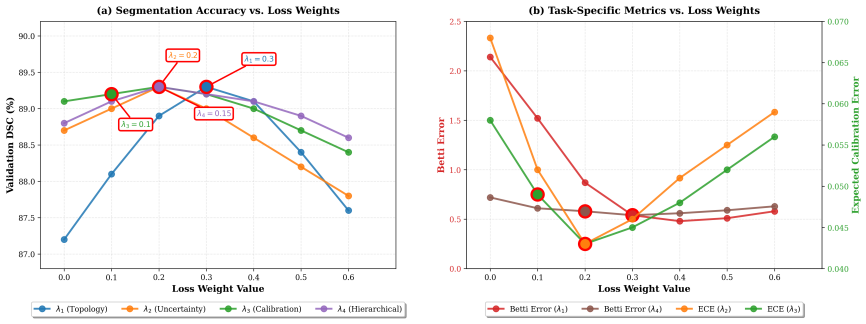

Figure 4: Loss weight sensitivity analysis on Synapse validation set. **(a)** Segmentation accuracy (DSC) versus loss weight values for all four auxiliary losses. Each weight exhibits a clear optimum marked by red-edged circles with annotations: $\lambda_1 = 0.3$ (topology), $\lambda_2 = 0.2$ (uncertainty), $\lambda_3 = 0.1$ (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.

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 mod-

erate 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: 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.

A.5. Comprehensive Uncertainty and Topology Evaluation

Table 6 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 7 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 6: 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 7: 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.6. 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.

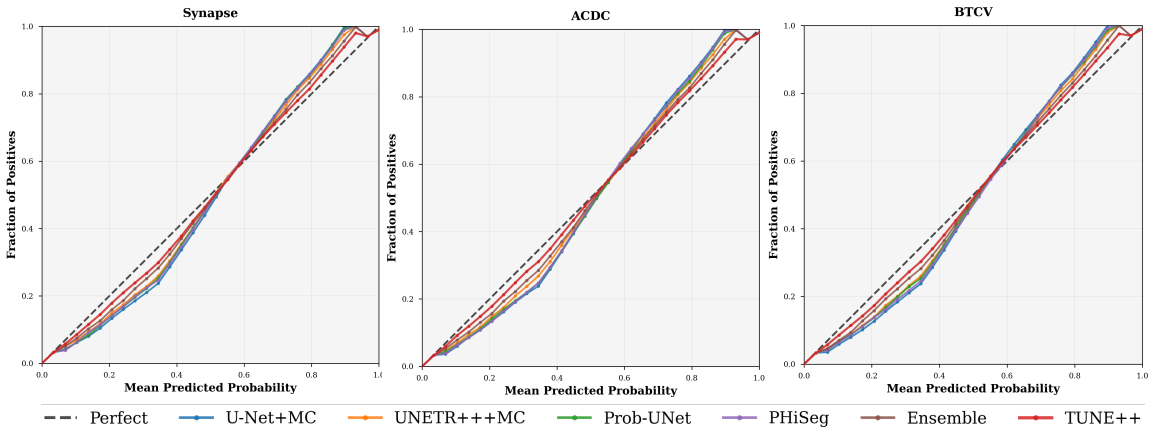

Figure 5: Expected Calibration Error reliability diagrams across three datasets. The diagonal black dashed line represents perfect calibration where predicted probability equals the fraction of correct predictions. TUNE++ (red, lowest ECE) demonstrates superior calibration, maintaining close alignment 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.7. Computational Efficiency Analysis

We analyze TUNE++'s computational requirements to assess practical deployment feasibility in resource-constrained clinical environments. Table 8 presents comprehensive metrics across representative baselines, while Figure 6 visualizes the accuracy-efficiency trade-off on the Synapse multi-organ segmentation benchmark.

Table 8: 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++

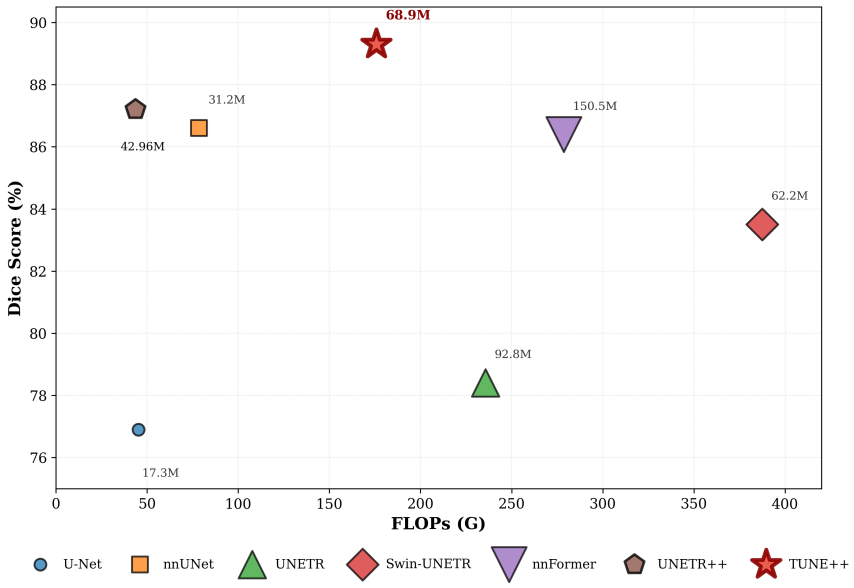

Figure 6: Computational efficiency versus segmentation accuracy trade-off on Synapse dataset. Each method is represented by a marker scaled proportional to parameter count. TUNE++ (red star, 68.9M parameters) achieves the highest DSC (89.4%) while maintaining computational efficiency superior to standard 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 9 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 9: 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.8. 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 10: 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.9. 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.