Supplementary Appendix: SegWithU

Anonymous Authors

A. Formula Reference

Backbone/probes. $z = f_\theta(M)$, $r_m = \mathrm{Resize}(W_m h^{(m)})$, $h = \phi_{\mathrm{fuse}}([r_1, \ldots, r_M])$, $v = \psi(h)$, $\sigma_r = \mathrm{softplus}(\alpha_r) + \epsilon$. For zero-indexed k, r, $q_{kr} = \cos\{\pi(k+1)(r+0.5)/R\}$, $u_r^{(k)} = +1$ if $q_{kr} \geq 0$ and -1 otherwise, $\Delta z(u^{(k)}) = A((\sigma \odot u^{(k)}) \odot v)$, $p^{(k)} = \mathrm{softmax}(z + \Delta z(u^{(k)}))$. $U_{\mathrm{epi}} = \sum_c \mathrm{Var}_k(p_c^{(k)})$, $U_{\mathrm{probe}} = R^{-1} \sum_r v_r^2$, $U_{\mathrm{res}} = C^{-1} \sum_c \Delta z_c^2$.

Calibration/ranking. $p = \mathrm{softmax}(z)$, $m = p_{(1)} - p_{(2)}$, $w = \exp(-\gamma m)$, $U_{\mathrm{ale}} = \mathrm{softplus}(\psi_{\mathrm{ale}}(h))$.

$$
\begin{aligned}
U_{\mathrm{cal}} &= \mathrm{softplus}\{\psi_{\mathrm{cal}}[\log(1 + U_{\mathrm{epi}} + U_{\mathrm{res}}),\\
&\qquad \log(1 + U_{\mathrm{ale}}), m]\}, \quad \tilde{z} = z/\sqrt{1 + U_{\mathrm{cal}}},\\
U_{\mathrm{anchor}} &= \log(1 + U_{\mathrm{epi}}) + \tfrac{1}{2}\log(1 + U_{\mathrm{res}}) + \tfrac{1}{4}\log(1 + U_{\mathrm{cal}})\\
&\qquad + \tfrac{1}{4} H(p)/\log C + w,\\
U_{\mathrm{rnk}} &= (1 + 0.1\tanh a) U_{\mathrm{anchor}} + b + \mathrm{softplus}(c) w.
\end{aligned}
$$

Objective. $\mathcal{L} = 0.5\mathcal{L}_{\mathrm{NLL}} + 0.25(\mathcal{L}_{\mathrm{EC}} + \mathcal{L}_{\mathrm{Pairwise}} + \mathcal{L}_{\mathrm{Tail}}) + 0.05(\mathcal{L}_{\mathrm{Trust}} + \mathcal{L}_{\mathrm{Anchor}} + \mathcal{L}_{\mathrm{Residual}})$, with $e_i = \mathbf{1}[\arg\max_c z_{i,c} \neq y_i]$ and $u_i = U_{\mathrm{rnk},i}$.

Loss definitions. Let $\hat{y}_i = \arg\max_c z_{i,c}$, $e_i = \mathbf{1}[\hat{y}_i \neq y_i]$, $u_i = U_{\mathrm{rnk},i}$, $\bar{u}_i = (u_i - \mu_u)/(\sigma_u + \epsilon)$, $\tilde{p}_i = \mathrm{softmax}(\tilde{z}_i)$, $p = \mathrm{softmax}(z)$, $p^{\Delta} = \mathrm{softmax}(z + \Delta z)$, $\bar{a}_i = (U_{\mathrm{anchor},i} - \mu_a)/(\sigma_a + \epsilon)$, voxels $\mathcal{V}$, and sampled error/correct pairs $\mathcal{P}$. Then $\mathcal{L}_{\mathrm{NLL}} = -|\mathcal{V}|^{-1} \sum_i \log \tilde{p}_{i,y_i}$; $\mathcal{L}_{\mathrm{EC}} = |\mathcal{V}|^{-1} \sum_i \mathrm{BCE}(\bar{u}_i/\tau, e_i)$; $\mathcal{L}_{\mathrm{Pairwise}} = |\mathcal{P}|^{-1} \sum_{(i,j)\in\mathcal{P}} \mathrm{softplus}((\bar{u}_j - \bar{u}_i + \delta)/\tau)$; $q_i = \exp(-u_i/T)/\sum_j \exp(-u_j/T)$, $\mathcal{L}_{\mathrm{Tail}} = \sum_i q_i e_i$; $\mathcal{L}_{\mathrm{Trust}} = (|\mathcal{V}|C)^{-1} \sum_{i,c} (\Delta z_{i,c})^2 + (4|\mathcal{V}|C)^{-1} \sum_{i,c} (p_{i,c}^{\Delta} - p_{i,c})^2$; $\mathcal{L}_{\mathrm{Anchor}} = |\mathcal{V}|^{-1} \sum_i \mathrm{SmoothL1}(\bar{u}_i, \bar{a}_i)$; and $\mathcal{L}_{\mathrm{Residual}} = |\mathcal{V}|^{-1} \sum_i (1 - w_i) U_{\mathrm{res},i}$. Here BCE denotes BCEWithLogits.

B. Experimental Protocol

Datasets and splits. ACDC: 3D cardiac MRI, 4 classes, official 200/100 train/test split, resampled, patch 64×128×128. BraTS2024: 4-channel brain MRI, 5 classes, 200 train and 100 held-out cases sampled from the public training set, patch 128^3. LiTS: abdominal CT, 3 classes, final 10 cases held out, patch 128^3. Foreground patches are oversampled at 33%. ACDC is resampled because through-plane spacing is coarse; BraTS and LiTS use no spacing alignment in the reported implementation. All datasets use dataset-specific intensity normalization.

Backbones and inference. All backbones are 3D DynUNet models trained on five non-overlapping folds plus an all-data model. Fold/all Dice: ACDC 88.21–90.51/90.35%; BraTS2024 55.87–60.07/62.75%; LiTS 71.77–76.75/78.21%. Backbone training uses Dice+CE, SGD momentum 0.9, LR 0.01, polynomial decay $p = 0.9$, batch sizes 16/8/8, and 200 epochs. All 3D inference uses sliding windows with Gaussian blending, overlap 0.5, window batch size 2, and the dataset ROI above.

Baselines. Deep Ensembles average the five fold backbones and score uncertainty from predictive-distribution mutual information. TTA uses the 8 axis-aligned 3D flip combinations, inverse-flips logits, averages probabilities, and scores entropy/variance. MC Dropout enables only dropout at inference on a dropout-enabled DynUNet and averages repeated stochastic sliding-window predictions. Temperature Scaling fits one scalar temperature on validation data by NLL; fitted temperatures are close to 1.5, and entropy of tempered probabilities is the ranking score. DUQ uses BCE plus gradient penalty $\lambda = 0.5$, SGD momentum 0.9, LR 0.05, MultiStep decay, centroid size 8, length scale 0.1, and EMA $\gamma = 0.999$. DDU-Seg trains Dice+CE, then fits class-conditional Gaussian densities on penultimate features using streaming Chan covariance estimates. DUE uses an SVGP head with RBF kernel, $M = 16$ inducing points for ACDC and $M = 64$ for BraTS/LiTS, K-means initialization, auxiliary CE weight 0.5, spectral normalization, 32 posterior samples per spatial location, and spatial subsampling to 32,768 locations per step. ConfidNet learns a confidence head that regresses the model's true-class probability (TCP) from intermediate features of the DynUNet.

SegWithU implementation. SegWithU trains only a frozen-backbone uncertainty head. Multi-scale taps are `upsamples.1.conv_block`, `upsamples.2.conv_block`, and `output_block`, with channels $(256, 128, 32)$ fused to 32 channels. Default probes use $R = K = 8$, $\sigma_{\mathrm{init}} = 0.1$, margin weight $\gamma = 4$, and the aleatoric branch enabled. Optimization uses AdamW, LR 10^{-3}, $\beta = (0.9, 0.999)$, $\epsilon = 10^{-8}$, cosine annealing to 3×10^{-4}, gradient clipping at 12, up to 200 epochs, early-stopping tolerance 10, and no deep supervision. Loss weights are $\lambda_{\mathrm{nll}} = 0.5$, $\lambda_{\mathrm{ec}} = \lambda_{\mathrm{pair}} = \lambda_{\mathrm{tail}} = 0.25$, $\lambda_{\mathrm{trust}} = \lambda_{\mathrm{anchor}} = \lambda_{\mathrm{res}} = 0.05$, and $\lambda_{\mathrm{seg}} = 0$.

Evaluation details. Ranking maps are: ensemble mutual information for Deep Ensembles; predictive entropy for MC Dropout, TTA, and TS; native aleatoric/deterministic scores for DUQ, DDU-Seg, and DUE; U_{rnk} for SegWithU. $1 - \widehat{\mathrm{TCP}}$ for ConfidNet, where its auxiliary confidence head approximates the voxel-wise true-class probability. SegWithU uses U_{cal} only for probability tempering in Brier/NLL-style scores. AURC sorts voxels by increasing uncertainty and numerically integrates risk–coverage; risk–coverage plots also use random-rejection and oracle references. Accuracy-threshold plots use 101 thresholds on $[0, 1]$.

Comparison caveats. DUQ, DDU, and DUE were originally classification methods; dense segmentation treats voxels as conditionally independent samples and changes computational scaling. ConfidNet has an optional stage that fine-tunes a copy of the segmentation model after training the confidence network. We omit this fine-tuning stage to maintain the same frozen-backbone constraint as SegWithU; additionally, in our experiments, the fine-tuned variant did not improve any metric on ACDC, BraTS, or LiTS compared to the original. Original baselines did not use MONAI DynUNet, so absolute values may differ from their papers. DUQ hyperparameters were inherited from low-resolution classification; Deep Ensembles omit adversarial training; MC Dropout depends on where dropout appears in DynUNet; DUE inference is stochastic unless the GP sampling seed is fixed.

C. ACDC Ablations

Group	Variant	Brier ↓	AUROC ↑	AURC ↓
Map	Cal. only	.0114±.0006	.7078±.0050	28.0538±1.4905
Map	Rank only	.0122±.0006	.9824±.0026	2.9853±.7933
Loss	No NLL	.0113±.0006	.9782±.0035	4.4299±1.3205
Loss	No EC	.0113±.0006	.9746±.0034	4.5017±1.1033
Loss	No pair	.0116±.0006	.9275±.0062	16.7523±2.4787
Loss	No tail	.0114±.0006	.9795±.0031	3.4856±.9728
Loss	No trust	.0113±.0006	.9823±.0024	2.7331±.6607
Probe	Direct	.0114±.0006	.9768±.0031	4.0557±.9928
Probe	Fixed σ	.0113±.0006	.9813±.0026	3.0440±.7501
Probe	No ale.	.0114±.0006	.9731±.0038	4.8666±1.2919
Full	SegWithU	**.0113±.0006**	**.9838±.0022**	**2.4885±.6077**

Dice is 0.9035±0.0044 for all rows because the backbone and hard segmentation are frozen. AURC is in 10^{-4} units. Calibration-only preserves Brier but fails ranking; ranking-only recovers ordering but worsens Brier; pairwise loss is the largest single loss contributor; direct/fixed/no-aleatoric variants show that the learned perturbation probe is not a generic auxiliary head.

D. Extra Sensitivity Ablations

R	Brier	AUROC	AURC	γ	Brier	AUROC	AURC
4	.0113	.9795	3.6617	1	.0113	.9783	3.7818
16	.0113	.9815	3.0500	2	.0113	.9797	3.8112
32	.0113	.9806	3.3402	8	.0114	.9784	3.9717
8	**.0113**	**.9838**	**2.4885**	**4**	**.0113**	**.9838**	**2.4885**

Default $R = 8$ and $\gamma = 4$ give the best ACDC ranking quality, suggesting a compact-probe bias–variance tradeoff and a moderate ambiguity-emphasis optimum. Single-tap vs multi-scale tapping: single late tap gives Brier/AUROC/AURC 0.0116/0.9743/3.9136; multi-scale tapping improves to 0.0113/0.9838/2.4885, indicating that uncertainty cues benefit from both coarse semantic and finer spatial features.

E. Statistical and Case-Level Details

Pairwise statistics use per-case Wilcoxon signed-rank tests with Holm correction within each dataset–metric family. The score vectors below follow (DE, TTA, MCDO, TS, DUQ, DDU-Seg, DUE, ConverSeg, SegWithU).

Dataset	DE/TTA/MCDO/TS	DUQ/DDU/DUE/ConfidNet/SegWithU
ACDC	+16, −20, −20, −1	−15, −5, +14, +6, +25
BraTS	−2, +3, −29, −5	+6, −6, +9, 0, +24
LiTS	0, 0, 0, 0	0, 0, 0, 0, 0
All	+14, −17, −49, −6	−9, −11, +23, +6, +49

LiTS is set to zero for corrected significance scoring because $n = 10$ is underpowered. SegWithU is the only method never significantly outperformed on ACDC/BraTS2024.

Qualitative case metrics.

Case	Dice	Brier	AUROC	AURC
ACDC 7	.8867	.0087	.9905	.7100
BraTS 44	.5554	.0017	.9974	.0347
LiTS 8	.9254	.0222	.9695	5.5866

Qualitatively, SegWithU usually keeps uncertainty attached to anatomy and produces fewer isolated high-uncertainty regions away from anatomy than Deep Ensembles or DUQ; DUE is often closest.

Risk and confidence curves. Per-case risk–coverage curves show SegWithU usually in the favorable low-risk group, though not always the single best curve at every coverage. It more reliably avoids early degradation than MC Dropout and often DDU/TTA. ConverSeg is competitive on ACDC 7 and BraTS2024 44, but SegWithU provides better error ranking across all three selected cases with the clearest separation on LiTS 8. In the accuracy-threshold curves, retained voxels generally become more accurate as confidence thresholds rise, supporting selective acceptance and targeted review rather than only scalar calibration.

LiTS per-case SegWithU summary.

Metric	Mean±SD	Range
Dice	.7821±.0343	–
Brier	.0067±.0020	–
AUROC	.9925±.0025	.9695–.9985
AURC	.8193±.5117	.0136–5.5866

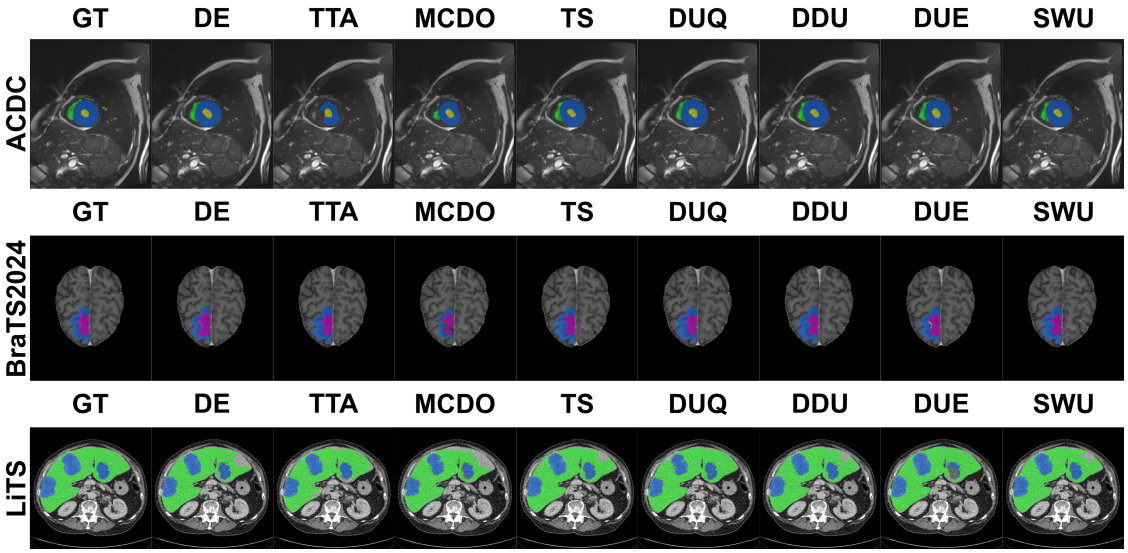
Volume 129 is hardest for uncertainty ranking despite high Dice, illustrating that segmentation overlap and uncertainty ordering are related but distinct targets.

Runtime and efficiency.

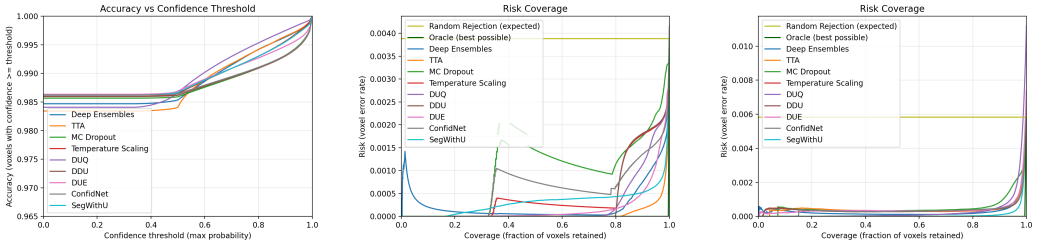
Method	Pass	Time A/B/L
Hardware: RTX Pro 6000 GPU		
DE	5	1.72/2.63/39.48
MCDO	20	1.25/2.96/90.38
TTA	8	0.50/1.18/31.31
TS	1	0.13/0.30/7.83
DUQ	1	0.13/0.31/8.12
DDU	1	0.13/0.30/7.82
DUE	1	0.92/3.24/45.52
ConfidNet	1	0.28/0.64/18.41
SegWithU	1	0.23/0.54/14.46

Time is measured in seconds. SegWithU adds 0.1M parameters.

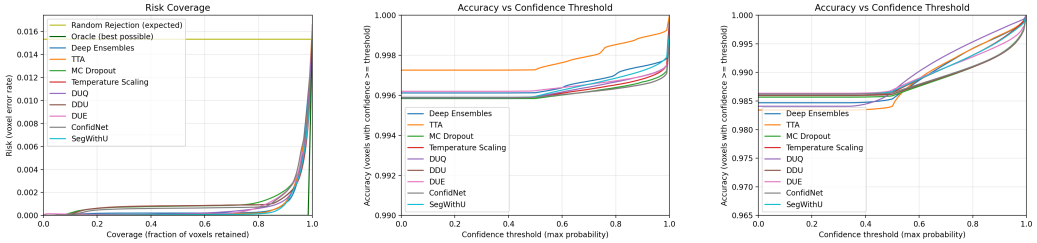
F. Additional Visual Plots



Hard-mask slices on ACDC, BraTS2024, and LiTS.



Risk–coverage examples from ACDC, BraTS2024, and LiTS.



Accuracy–threshold examples from the same selected cases.

The mask slices show that SegWithU preserves the frozen-backbone segmentation quality while adding reliability maps. Bar plots summarize aggregate metrics; risk–coverage and accuracy–threshold curves show selective-prediction behavior, with the leading method varying by case and coverage.