RETHINKING MODEL CALIBRATION THROUGH SPECTRAL ENTROPY REGULARIZATION IN MEDICAL IMAGE SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep neural networks for medical image segmentation often produce overconfident predictions, posing clinical risks due to miscalibrated uncertainty estimates. In this work, we rethink model calibration from a frequency-domain perspective and identify two critical factors causing miscalibration: spectral bias, where models overemphasize low-frequency components, and confidence saturation, which suppresses overall power spectral density in confidence maps. To address these challenges, we propose a novel frequency-aware calibration framework integrating spectral entropy regularization and power spectral smoothing. The spectral entropy term promotes a balanced frequency spectrum and enhances overall spectral power, enabling better modeling of high-frequency boundary and low-frequency structural uncertainty. The smoothing module stabilizes frequency-wise statistics across training batches, reducing sample-specific fluctuations. Extensive experiments on six public medical imaging datasets and multiple segmentation architectures demonstrate that our approach consistently improves calibration metrics without sacrificing segmentation accuracy.

1 INTRODUCTION

Deep neural networks (DNNs) have become essential in medical image segmentation, enabling tasks such as tumor delineation and organ identification Hatamizadeh et al. (2021); Chen et al. (2021); Isensee et al. (2021). Despite their high accuracy, DNNs often exhibit overconfident predictions—particularly near lesion boundaries, where predicted confidence substantially exceeds actual correctness Yeung et al. (2023); Larrazabal et al. (2023); Murugesan et al. (2025). This overconfidence undermines the reliability of automated systems in clinical settings and leads to misinformed diagnostic decisions Begoli et al. (2019).

To mitigate such risks, confidence calibration aims to align the predicted confidence with the true likelihood of correctness. Existing calibration methods in medical segmentation primarily operate in the spatial domain. Post-hoc techniques such as temperature scaling and Platt scaling apply dataset-level adjustments to logits Guo et al. (2017); Platt et al. (1999), but they often fail to adapt to anatomical and modality variability Tomani et al. (2021). Training-time methods, including label smoothing Müller et al. (2019), Focal Loss Mukhoti et al. (2020), and regional uncertainty regularization Murugesan et al. (2023a; 2024), add explicit or implicit constraints on output confidence Pereyra et al. (2017). While these approaches partially alleviate overconfidence, they predominantly focus on spatial domain characteristics and neglect potential frequency domain information, which can lead to suboptimal calibration performance.

Recent efforts suggest that frequency-domain information can enhance calibration and robustness. For instance, style-invariant Fourier domain adaptation has been used to stabilize model behavior across domains, and hybrid spatial-spectral representations have shown promise in improving generalization and uncertainty estimation Chen et al. (2023b); Li et al. (2024); Tang et al. (2021). However, such works primarily target domain shift or style variation, rather than addressing overconfidence in confidence maps for medical segmentation.

From a spectral perspective, neural networks exhibit an inherent **Spectral Bias** during optimization, preferentially learning low-frequency patterns while inadequately capturing high-frequency details

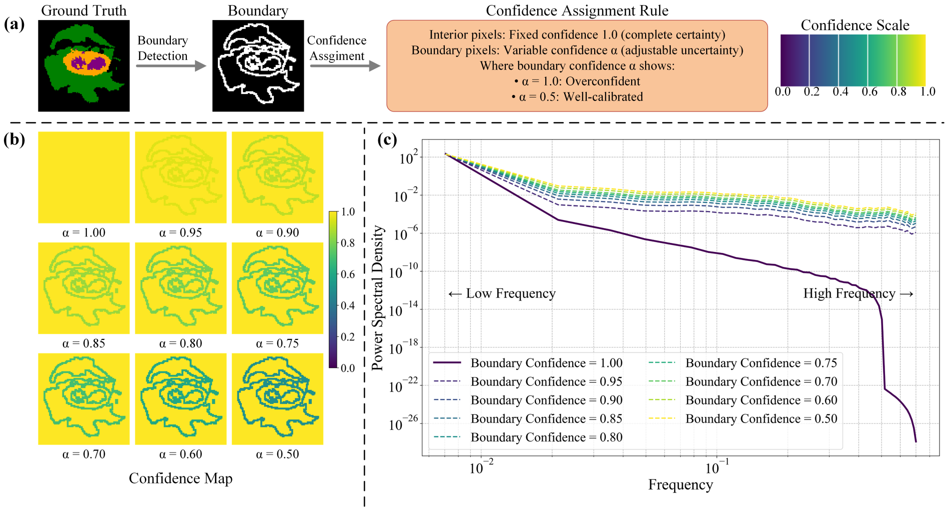


Figure 1: Frequency-domain analysis of overconfidence in medical image segmentation. **(a)** Synthesized confidence maps for a binary (foreground/background) segmentation example. Boundary pixels are assigned variable foreground confidence levels α, while interior pixels have full confidence (confidence $= 1.0$). **(b)** Resulting confidence maps as boundary confidence α varies from overconfident (1.0) to well-calibrated (0.5). In a binary context, $\alpha = 0.5$ represents maximum uncertainty, the ideal state for a boundary pixel, as the model is equally confident in the foreground and background classes. Confidence values $\alpha < 0.5$ are not shown, as they would correspond to an incorrect background prediction, which is outside the scope of calibrating a correct foreground prediction. **(c)** Power spectral density (PSD) analysis. A Hamming window Hamming & Stearns (1979) is applied before the FFT to reduce spectral leakage caused by edge artifacts, ensuring a more accurate spectrum. As boundary confidence decreases toward the well-calibrated state, the PSD becomes richer across all frequency bands.

Rahaman et al. (2019); Xu et al. (2019), which further exacerbates overconfidence in boundary regions where high-frequency information is crucial for uncertainty estimation. Besides, as illustrated in Fig. 1, we create synthetic confidence maps by assigning different confidence levels to boundary pixels, while keeping interior pixels fixed. The power spectral density (PSD) analysis shows that overconfident maps (with boundary confidence $\alpha = 1.0$) have sparse spectral energy. In contrast, well-calibrated maps (e.g., $\alpha = 0.5$) display richer power across frequency bands. This suggests that overconfidence is linked to both spectral bias and low spectral complexity. Better-calibrated predictions distribute spectral power more evenly, capturing both high-frequency boundary uncertainty and low-frequency structural variations. We term this effect of spectral suppression in confidence maps as **Confidence Saturation**.

To address these two issues toward trustworthy medical image segmentation models, this paper introduces a novel adaptive calibration method from a spectral perspective. Our main contributions are as follows:

1. We conduct the first comprehensive spectral analysis of overconfidence in segmentation, revealing a strong link between prediction uncertainty and spectral power distributions.

2. We introduce a **power spectral smoothing** mechanism that stabilizes frequency-wise power statistics by aggregating spectral information across historical training batches, helping maintain segmentation performance while promoting consistent spectral patterns.

3. We design a **spectral entropy regularization** mechanism that enhances overall spectral power and improves frequency balance in confidence maps, leading to significantly improved calibration quality.

2 RELATED WORK

This section reviews research areas closely related to our approach, including confidence calibration methods, uncertainty estimation approaches, and frequency domain approaches in medical image analysis.

2.1 CONFIDENCE CALIBRATION METHODS

Post-hoc Calibration Methods. These methods improve calibration by adjusting output distributions after model training without retraining. Temperature Scaling Guo et al. (2017) adjusts softmax outputs by optimizing a global temperature parameter. Platt Scaling Platt et al. (1999) applies logistic regression to transform logits, while non-parametric methods such as Histogram Binning Zadrozny & Elkan (2001) and Isotonic Regression Fawcett & Niculescu-Mizil (2007) map predicted probabilities to empirical accuracies. However, these approaches lack region-specific calibration capabilities for different anatomical structures and show limited generalization under domain shifts across modalities Tomani et al. (2021). **In-training Calibration Methods.** These methods modify training objectives with regularization strategies to optimize calibration during training. Label Smoothing Müller et al. (2019) softens target labels to reduce overconfidence, while Focal Loss Mukhoti et al. (2020) reduces the influence of easily segmented samples. Entropy Regularization Pereyra et al. (2017) encourages higher predictive entropy to suppress overconfident predictions. Recent methods such as MarginLoss Murugesan et al. (2023a) and CRaC Murugesan et al. (2024) incorporate regional information and adaptive constraints for spatial-aware calibration. While these approaches demonstrate advantages in medical image tasks, they primarily rely on spatial domain features and lack systematic frequency domain analysis for calibration.

2.2 PROBABILISTIC UNCERTAINTY ESTIMATION METHODS

Uncertainty estimation approaches in deep learning are primarily based on probabilistic modeling. Bayesian neural networks (BNNs) Vadera et al. (2021), exemplified by MC-Dropout Gal & Ghahramani (2016), approximate Bayesian posteriors by retaining dropout during inference. Deep ensembles Lakshminarayanan et al. (2017) train multiple models with different initializations and aggregate predictions, while evidential deep learning Sensoy et al. (2018) directly parameterizes probability distributions rather than point estimates. These methods have been applied in medical image analysis, with Nair et al. Nair et al. (2020) evaluating approaches in brain tumor segmentation and Jungo et al. Jungo et al. (2018) examining the impact of uncertainty estimation on segmentation accuracy. However, computational complexity and multiple inference requirements limit their real-time clinical applicability Murugesan et al. (2023b).

2.3 FREQUENCY DOMAIN APPROACHES IN MEDICAL IMAGE ANALYSIS

Frequency analysis has been widely applied in medical image analysis, from traditional Fourier-based denoising and enhancement to recent deep learning integration. Recent works have explored diverse applications including data augmentation impact analysis Yin et al. (2019), enhanced translation invariance Zhu et al. (2021), skin lesion segmentation optimization Li et al. (2021), MRI organ segmentation with spectral constraints Chen et al. (2023a), and cross-domain few-shot segmentation through frequency-aware matching Bo et al. (2025). Notably, Guang et al. Gorade et al. (2024) used a spectral loss that matches the prediction's spectrum to the ground truth's to improve segmentation accuracy. In contrast, we focus on confidence calibration and directly regularize the intrinsic spectral properties of the model's output itself, without relying on a ground truth spectrum. To our knowledge, this use of frequency analysis for confidence calibration in medical image segmentation remains underexplored.

3 METHOD

3.1 SPECTRAL ANALYSIS OF MODEL UNCERTAINTY

To motivate our frequency-aware calibration framework, we first conduct a comprehensive spectral analysis to investigate how model confidence relates to uncertainty estimation in medical image segmentation. Specifically, we analyze the power spectral density (PSD) of predicted confidence maps and observe that miscalibrated models often exhibit unbalanced spectral power distributions, characterized by the dominance of low-frequency components and suppressed high-frequency responses, as shown in Fig. 1. This **spectral bias** limits the model's ability to capture detailed boundary uncertainty and subtle anatomical structures. We observe that uncalibrated models often produce confidence maps that saturate toward uniformly high values, a behavior we define as **confidence saturation**. This effect results in reduced overall power spectral density (PSD), diminishing the frequency richness required to represent model uncertainty, especially in regions with boundary ambiguity. In contrast, well-calibrated models yield confidence maps with higher overall PSD, preserving richer spectral characteristics that better reflect structural variability and predictive uncertainty.

These findings reveal a strong connection between model calibration and the spectral characteristics of prediction confidence. The presence of spectral bias and confidence saturation suggests that current models may fail to adequately represent uncertainty across spatial scales, and particularly at high-frequency boundaries. This motivates the design of a calibration strategy that explicitly incorporates spectral structure, aiming to rebalance frequency distributions and increase overall spectral power, thereby enhancing model calibration and improving segmentation reliability.

3.2 SPECTRAL DECOMPOSITION AND WINDOW SMOOTHING

Building on our prior spectral uncertainty analysis, we now introduce a spectral decomposition and temporal smoothing scheme that extracts reliable frequency-domain confidence profiles, mitigating per-sample variability and enabling stable, frequency-aware calibration. Given input images $x \in \mathbb{R}^{B\times C_{\text{in}}\times D\times H\times W}$, where B is the batch size, C_{in} denotes the number of input channels, and D, H, W denote depth, height, and width respectively, the neural network $f_\theta(\cdot)$ outputs prediction features (logits before softmax) $z = f_\theta(x) \in \mathbb{R}^{B\times C\times D\times H\times W}$, where C is the number of classes. A scalar confidence map is constructed by taking the voxel-wise maximum logit value across the class dimension. While calibration is ultimately assessed on post-softmax probabilities, we operate on logits because they represent the network's raw, unnormalized evidence and offer a more sensitive signal for spectral analysis. Softmax outputs can saturate near 0 or 1, which can mask underlying spectral characteristics, whereas logits retain a wider dynamic range. This logit-based map therefore directly captures the precursors to overconfidence, where excessively high values are a primary cause of miscalibration.

$$z_b(d,h,w) = \max_{c=1,\ldots,C} z_{b,c}(d,h,w), \tag{1}$$

For each sample b, we transform $z_b(d,h,w)$ into the frequency domain via a 3D Fast Fourier Transform (FFT) to analyze its spectral properties:

$$F_b(u,v,w) = \text{SC}\left(\mathcal{F}_{3D}\left[z_b(d,h,w)\right]\right) \tag{2}$$

$$E_b(u,v,w) = \left|F_b(u,v,w)\right|^2, \tag{3}$$

where (u,v,w) are frequency domain coordinates, and $E_b(u,v,w)$ represents the power spectral density at each frequency for sample b. Here, SC denotes Spectrum Centering, i.e., zero-frequency component shift to center the spectrum.

To characterize spectral behavior at different scales, we partition the frequency domain into K concentric spherical shells $\mathcal{I}_k$ based on frequency magnitude, following the band decomposition strategy in Bo et al. (2025):

$$\mathcal{I}_k = \left\{(u,v,w) \;\middle|\; r_{k-1} < \sqrt{u^2+v^2+w^2} \le r_k\right\}, \quad k = 1,\ldots,K, \tag{4}$$

with $r_k = \frac{k}{K}\frac{\min(D,H,W)}{2}$, which normalizes the frequency bands relative to the spatial dimensions, allowing scale-invariant spectral partitioning. This decomposition enables frequency-aware characterization of low-frequency structural content and high-frequency boundary information. For each sample b and frequency band k, we aggregate the spectral power within the corresponding shell:

$$S_b^{(k)} = \sum_{(u,v,w)\in\mathcal{I}_k} E_b(u,v,w). \tag{5}$$

To reduce statistical fluctuations and enhance training stability, we apply temporal smoothing over a sliding window of recent batches:

$$\widetilde{S}^{(k)} = \frac{1}{W} \sum_{w=1}^{W} \frac{1}{|\mathcal{B}_w|} \sum_{b \in \mathcal{B}_w} S_{b,w}^{(k)}, \tag{6}$$

where W is the window size, $w = W$ corresponds to the current batch, and $w = 1$ corresponds to the oldest batch within the window. This smoothing aggregates spectral power statistics over multiple samples and batches to reduce noise and sample-specific variance. Let $\widetilde{S} = (\widetilde{S}^{(1)}, \dots, \widetilde{S}^{(K)})$ denote the vector of smoothed spectral power across all K frequency bands. During early training, when historical batches are unavailable, we replicate the current batch statistics to initialize the window and ensure numerical stability. This temporally smoothed spectral profile provides a consistent and robust reference for regulating spectral power distributions in frequency bands throughout training, thereby supporting frequency-aware calibration.

3.3 POWER SPECTRAL ENTROPY REGULARIZATION

Leveraging the temporally smoothed spectral profile vector $\widetilde{S}$, we introduce a power spectral entropy regularization term, which encourages the model to learn balanced frequency representations that alleviate spectral bias. It also promotes higher, more informative spectral power to counteract confidence saturation, thereby facilitating calibrated segmentation.

To quantify spectral diversity, we first convert the raw spectral power vectors into probability distributions. For the smoothed spectral profile $\widetilde{S}$ and each sample's profile S_b, we define their normalized counterparts, $\widetilde{P}$ and P_b, as:

$$P^{(k)} = \frac{S^{(k)}}{\sum_{j=1}^{K} S^{(j)} + \varepsilon}, \tag{7}$$

where S is either $\widetilde{S}$ or S_b, and ε is a small constant (e.g., 10^{-6}) to ensure numerical stability in both the normalization and entropy calculation steps. The spectral entropy is then computed using the standard Shannon entropy formula:

$$H_{\text{spec}}(P) = -\sum_{k=1}^{K} P^{(k)} \log(P^{(k)} + \varepsilon). \tag{8}$$

Here, P can be $\widetilde{P}$ or P_b. We select Shannon entropy as it is a principled measure from information theory for quantifying the uniformity of a probability distribution. A higher spectral entropy indicates that power is more evenly distributed across all frequency bands rather than being concentrated in a few. This directly aligns with our goal of encouraging richer, more balanced frequency representations for better uncertainty estimation. The entropy $H_{\text{spec}}(P_b)$ is computed independently for each sample b in a batch. While this compresses the spectrum into a summary statistic, it intentionally quantifies the spectral uniformity of each sample to address spectral bias. By penalizing this per-sample value, our method enhances the frequency representation of each specific prediction, thereby preserving its unique characteristics rather than obscuring them.

To address spectral bias and the resulting confidence saturation, we introduce a hinge-like spectral regularization loss. It encourages each sample's spectral entropy $H_{\text{spec}}(P_b)$ to approach or exceed the dynamically updated target entropy $H_{\text{spec}}(\widetilde{P})$:

$$\mathcal{L}_{\text{Spectral}} = \frac{1}{B} \sum_{b \in B} \left[\max\Big(0, H_{\text{spec}}(\widetilde{P}) - H_{\text{spec}}(P_b)\Big) \right]^2. \tag{9}$$

The mechanism of this loss directly counteracts spectral bias. The hinge formulation creates a targeted learning signal: a penalty is incurred only when a sample's spectral entropy is deficient compared to the dynamic target. Minimizing this penalty formally encourages a more uniform spectral distribution by pushing power from dominant low-frequency bands into underrepresented high-frequency bands. While entropy itself is scale-invariant, this re-balancing forces the model to increase the total spectral energy of the logit map to avoid compromising the task loss. This prevents the generation of overly smooth, low-energy confidence maps, directly mitigating the effects of confidence saturation.

The $\max(\cdot, 0)$ operation ensures that only samples with spectral entropy below the adaptive target incur penalties. This selective penalization encourages the model to enhance frequency-domain diversity while maintaining calibration performance for already well-calibrated samples.

3.4 OVERALL TRAINING OBJECTIVE

We integrate our spectral regularization with the conventional segmentation objective through a weighted combination:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \lambda \mathcal{L}_{\text{Spectral}}, \tag{10}$$

where $\mathcal{L}_{\text{CE}}$ is the cross-entropy loss for pixel-wise accuracy, and $\lambda > 0$ balances the trade-off between segmentation performance and uncertainty calibration. The hyperparameter λ allows users to balance segmentation accuracy and calibration strength. This flexibility is important because different organs or lesions exhibit varying levels of uncertainty, requiring tailored calibration strategies to reflect the underlying uncertainty characteristics in each case optimally.

4 EXPERIMENTS

4.1 DATASETS AND IMPLEMENTATION DETAILS

We conducted experiments on six public medical imaging datasets, covering diverse anatomical structures and imaging modalities, to comprehensively evaluate our method. These datasets are Brain Tumor Segmentation (BraTS) 2020 Menze et al. (2014), Automated Cardiac Diagnosis Challenge (ACDC) Bernard et al. (2018), Fast and Low GPU memory Abdominal oRgan sEgmentation (FLARE) 2021 Ma et al. (2022), MICCAI Grand Challenge on 6-month Infant Brain MRI Segmentation (iSeg) 2017 Sun et al. (2021), A Tumour and Liver Automatic Segmentation (ATLAS) 2023 Quinton et al. (2023), and Prostate MR Image Segmentation (PROMISE) 2012 Litjens et al. (2014). All datasets were split into training, validation, and test sets with a ratio of 4:1:1.

Experiments were conducted on Ubuntu 22.04 with an NVIDIA RTX 4090 GPU, utilizing the U-Net architecture Ronneberger et al. (2015). We adopted consistent training settings across all experiments, including a batch size of 2, input patch size of $96 \times 96 \times 96$, and the SGD optimizer (initial learning rate 0.01). Input data were normalized to $[0, 1]$.

4.2 EVALUATION METRICS AND COMPARISON METHODS

Segmentation performance was evaluated using Dice Similarity Coefficient (DSC), 95% Hausdorff Distance (HD95), and Average Surface Distance (ASD), which collectively assess overlap accuracy, boundary error, and surface deviation, respectively. To evaluate calibration, we used Expected Calibration Error (ECE) as the primary metric, following the protocol in Nixon et al. (2019), and included Static Calibration Error (SCE) and Thresholded Adaptive Calibration Error (TACE). SCE generalizes ECE to multi-class settings via class-wise binning, while TACE improves robustness through equal-sized binning and probability thresholding. Both ECE and SCE were computed using $B = 15$ bins, and TACE employed a threshold $\epsilon = 0.001$ as in Murugesan et al. (2024).

All comparison methods are in-training calibration strategies built upon the cross-entropy loss, consistent with our approach. We implement these baselines using the optimal hyperparameter configurations reported in their respective original publications, a standard practice for fair benchmarking. These include Focal Loss (FL, $\gamma = 3$) Lin et al. (2017), Label Smoothing (LS, $\alpha = 0.1$) Szegedy et al. (2016), Margin-based Label Smoothing (MbLS, $\lambda = 0.1$, $m = 5$) Murugesan et al. (2023b), Spatially Variant Label Smoothing (SVLS, $\sigma = 2$) Islam & Glocker (2021), and Class and Region-Adaptive Constraints (CRaC, $\gamma = 1.2$, $\mu = 0.9$, $R = 2$, $margin = 5$) Murugesan et al. (2024).

5 RESULTS AND DISCUSSION

Quantitative Evaluation in Calibration and Segmentation. Table 1 shows that our method achieves the best and second-best segmentation performance across all metrics, outperforming the

Table 1: Quantitative comparison of segmentation performance across multiple datasets (DSC↑, HD95↓, ASD↓). ↑ indicates higher is better, ↓ indicates lower is better. **Bold** denotes the best result, and *italics* indicate the second-best.

Dataset	Method & Publication																				
	DSC↑ (%)							HD95↓							ASD↓						
	CE	FL	LS	MbLS	SVLS	CRaC	Ours	CE	FL	LS	MbLS	SVLS	CRaC	Ours	CE	FL	LS	MbLS	SVLS	CRaC	Ours
	-	ICCV	CVPR	MedIA	IPMI	MICCAI	-	-	ICCV	CVPR	MedIA	IPMI	MICCAI	-	-	ICCV	CVPR	MedIA	IPMI	MICCAI	-
BraTS2020	*86.9*	84.6	86.7	85.9	85.9	85.9	**87.2**	*2.27*	2.80	2.52	2.91	2.46	2.58	**2.23**	0.40	0.44	*0.37*	0.41	0.47	0.39	**0.37**
iSeg2017	*94.2*	94.0	94.1	94.1	*94.2*	*94.2*	**94.4**	*1.01*	1.46	1.13	1.15	1.15	1.17	**0.85**	0.06	0.06	0.06	*0.06*	0.06	0.06	**0.05**
FLARE2021	91.5	87.7	91.3	*91.5*	91.2	91.4	**92.5**	8.29	*6.17*	8.37	7.38	8.40	7.48	**4.55**	2.64	2.22	2.49	*2.09*	2.22	2.10	**1.34**
ACDC	*91.1*	89.4	91.0	90.7	*91.1*	91.0	**91.3**	*0.75*	0.90	0.88	0.83	0.80	0.81	**0.74**	0.10	0.12	0.11	*0.10*	0.11	0.11	**0.10**
ATLAS2023	68.7	65.9	69.1	*70.5*	68.3	69.8	**71.8**	22.20	20.19	19.76	*18.84*	20.18	20.31	**17.82**	4.99	4.53	4.46	**3.54**	5.21	3.87	*3.57*
PROMISE2012	*80.2*	78.9	79.6	79.1	79.5	80.1	**81.2**	19.80	*18.79*	19.93	22.05	20.33	18.95	**18.24**	7.36	8.51	5.22	8.99	8.86	*4.32*	**4.18**

Table 2: Quantitative comparison of calibration performance across multiple datasets (ECE↓, SCE↓, TACE↓).

Dataset	Method & Publication																				
	ECE↓ ($\times 10^{-3}$)							SCE↓ ($\times 10^{-3}$)							TACE↓ ($\times 10^{-3}$)						
	CE	FL	LS	MbLS	SVLS	CRaC	Ours	CE	FL	LS	MbLS	SVLS	CRaC	Ours	CE	FL	LS	MbLS	SVLS	CRaC	Ours
	-	ICCV	CVPR	MedIA	IPMI	MICCAI	-	-	ICCV	CVPR	MedIA	IPMI	MICCAI	-	-	ICCV	CVPR	MedIA	IPMI	MICCAI	-
BraTS2020	9.1	5.5	25.2	*1.9*	2.1	2.2	**1.5**	10.7	6.4	13.8	*4.0*	5.0	5.1	**0.8**	42.7	37.7	61.4	*31.5*	32.8	39.4	**20.2**
iSeg2017	4.5	6.5	17.7	*2.1*	2.6	2.6	**2.0**	11.9	4.0	8.5	*2.3*	2.9	2.9	**2.0**	15.6	17.6	33.8	*11.8*	13.5	13.4	**6.4**
FLARE2021	25.5	19.4	6.5	*2.2*	2.5	2.4	**0.8**	18.1	17.3	13.0	11.5	9.1	9.4	**0.6**	37.8	52.7	80.0	30.4	*24.2*	28.5	**3.1**
ACDC	32.5	32.3	71.9	*23.2*	27.9	28.1	**2.1**	16.6	16.7	38.1	*10.6*	13.3	**1.1**	**1.1**	32.0	38.0	48.1	26.7	33.9	**5.3**	*18.5*
ATLAS2023	24.9	9.4	50.4	7.2	*6.8*	7.0	**5.5**	15.2	22.9	44.3	9.6	22.7	*5.5*	**4.2**	*18.1*	40.7	89.9	20.8	37.8	37.0	**15.9**
PROMISE2012	*11.7*	12.6	20.2	12.8	20.7	20.8	**10.8**	25.7	34.2	52.8	*18.0*	54.7	52.6	**14.8**	48.1	45.6	55.0	*41.7*	64.1	**41.2**	55.7

CE baseline and leading segmentation methods. As shown in Table 1, our method demonstrates consistently strong and robust performance across Dice, HD95, and ASD metrics, validating its effectiveness across diverse segmentation tasks. As shown in Table 2, our method demonstrates state-of-the-art calibration, achieving the lowest ECE and SCE scores across all datasets. These results highlight the reliability and robustness of our approach in producing well-calibrated confidence estimates across the full probability spectrum. CRaC's superior TACE performance on ACDC and PROMISE2012 stems from its task-specific calibration tailored for fine-grained confidence adjustments in small, well-defined anatomical structures such as heart chambers and the prostate. By leveraging spatial consistency and localized uncertainty modeling, CRaC effectively reduces subtle miscalibrations in high-certainty regions emphasized by TACE. However, this focused calibration strategy can sometimes compromise segmentation accuracy. In contrast, our method achieves a more balanced calibration reflected in ECE and SCE metrics, alongside consistently superior segmentation performance across diverse datasets.

Qualitative Analysis in Calibration and Spectral Analysis. Fig. 2 illustrates how our method enhances uncertainty calibration and segmentation performance on BraTS2020. The CE baseline yields disconnected boundaries in the confidence map (Fig. 2(a)), resulting in overconfident behavior in the high-confidence region of the reliability diagram (Fig. 2(b)). Other methods, particularly FL and LS, generate low-confidence boundaries but apply overly aggressive confidence suppression, resulting in diffuse and spatially imprecise uncertainty regions on the confidence map, which manifest as systematically underconfident predictions in the lower-confidence region of the reliability diagram. Our spectral regularization and smoothing promote a balanced frequency representation, enabling sharper and more precise boundary localization in the confidence maps. Fig. 2 illustrates how our method enhances uncertainty calibration and segmentation performance on BraTS2020. The CE baseline yields disconnected boundaries in the confidence map (Fig. 2(a)), resulting in systematic overconfidence (under the ideal curve) in the high-confidence region of the reliability diagram (Fig. 2(b)). On the other hand, FL applies overly aggressive suppression to generate low-confidence boundaries, leading to diffuse and spatially imprecise uncertainty regions in the confidence map, which manifest as consistent underconfidence (above the ideal curve) in the high-confidence region of the reliability diagram. The rest of the methods are experiencing underconfidence in their lower-confident predictions to some extent, resulting in significantly higher ECE than our method.

Fig. 3 provides a spectral perspective on calibration by linking the power spectral density (PSD) distribution to miscalibration patterns observed in the reliability diagram (Fig. 2(b)). The CE baseline exhibits the lowest overall spectral power, reflecting strong confidence saturation and a steep PSD slope that indicates significant spectral imbalance. This spectral bias corresponds to its overcon-

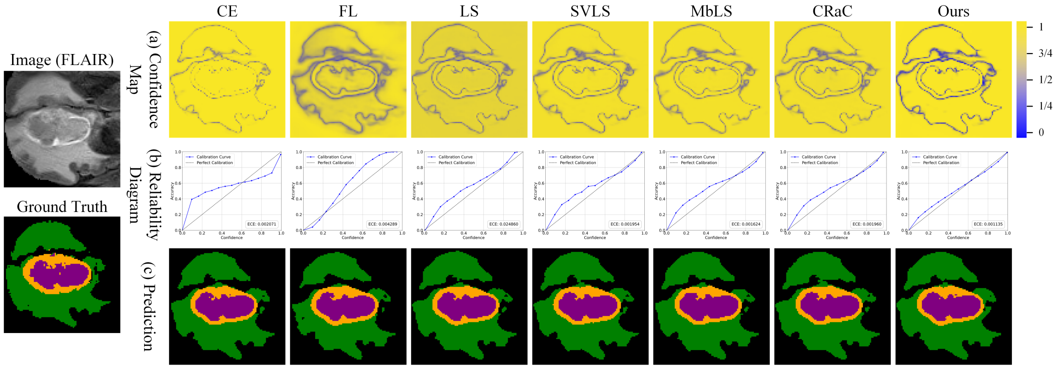


Figure 2: BraTS2020 brain tumor segmentation: (a) confidence maps with blue (low confidence) and orange (high confidence), (b) reliability diagrams showing calibration quality against the ideal diagonal baseline, and (c) segmentation results comparison with black (background), green (edema), purple (non-enhancing necrosis), and orange (enhancing tumor).

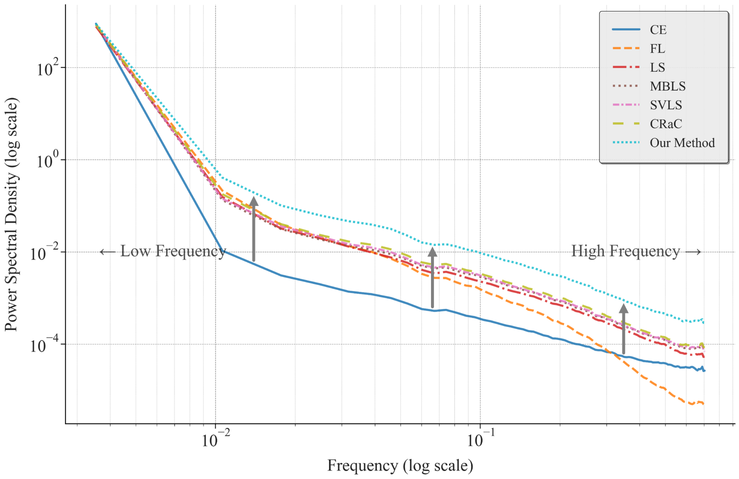


Figure 3: Comparison of power spectral density of confidence maps on the BraTS2020 dataset across different approaches.

fident predictions in high-confidence regions and underconfident predictions in low-confidence regions in the reliability diagram. Focal Loss reinforces low-frequency dominance, which moderately enhances the PSD in low-frequency regions but sharply suppresses high-frequency components—its high-frequency PSD is even lower than CE. This imbalance hinders its ability to represent boundary-level uncertainty, resulting in underconfident predictions in high-confidence regions. Other methods show comparable overall spectral power and insufficient high-frequency representation, which aligns with their mild underconfidence in the reliability diagram. In contrast, our method achieves the highest overall spectral power and the most balanced frequency distribution, particularly in the high-frequency regions critical for boundary uncertainty. This enables accurate modeling of detailed confidence variations and contributes to the best overall calibration performance on multiple datasets.

To validate the generality of our approach, we applied it across a diverse set of widely adopted segmentation architectures (Fig. 4), including nnUNet Isensee et al. (2021), SwinUNETR Hatamizadeh et al. (2021), UNet++ Zhou et al. (2019), AttentionUNet Oktay et al. (2018), and TransUNet Chen et al. (2021). Using a fixed set of hyperparameters for all experiments, our method consistently improved both segmentation accuracy and calibration. This result demonstrates the robustness and plug-and-play applicability of our approach.

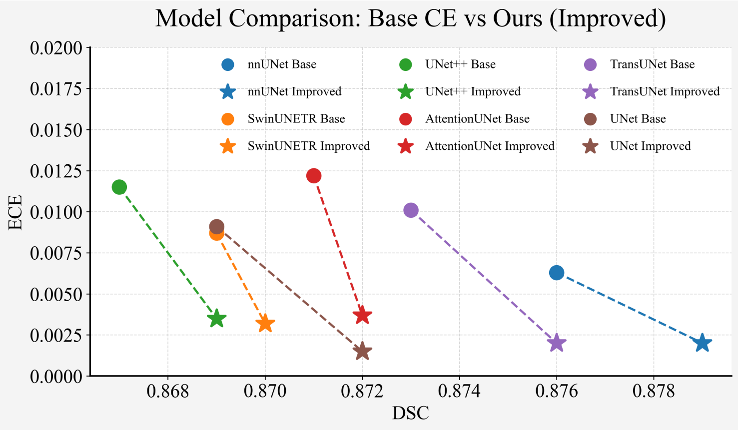


Figure 4: Generalizability of the method across segmentation architectures: improvements in segmentation (DSC) and calibration (ECE) performance compared to the CE baseline.

Table 3: Ablation study results of our method on the BraTS2020 and FLARE2021 datasets. We evaluate the contribution of each key component across segmentation metrics (DSC, HD95, ASD) and calibration metrics (ECE, SCE, TACE). The configurations are: Baseline ($\mathcal{L}_{CE}$), adding spectral entropy regularization without temporal smoothing ($\mathcal{L}_{CE}$ w/ $\mathcal{L}_{\text{Spectral}}$ w/o W), and our full method with smoothing ($\mathcal{L}_{CE}$ w/ $\mathcal{L}_{\text{Spectral}}$).

Configuration	BraTS2020						FLARE2021					
	DSC↑	HD95↓	ASD↓	ECE↓	SCE↓	TACE↓	DSC↑	HD95↓	ASD↓	ECE↓	SCE↓	TACE↓
Baseline ($\mathcal{L}_{CE}$)	0.869	2.27	0.40	0.0091	0.0107	0.0427	0.915	8.29	2.64	0.0255	0.0181	0.0378
$\mathcal{L}_{CE}$ w/ $\mathcal{L}_{\text{Spectral}}$ w/o W	0.870	2.25	0.39	0.0065	0.0078	0.0315	0.921	6.15	1.98	0.0170	0.0123	0.0210
$\mathcal{L}_{CE}$ w/ $\mathcal{L}_{\text{Spectral}}$	0.872	2.23	0.37	0.0015	0.0008	0.0202	0.925	4.55	1.34	0.0008	0.0006	0.0031

Ablation Studies. Our ablation studies (Table 3) evaluate the contribution of each component. Introducing unsmoothed spectral regularization ($\mathcal{L}_{\text{CE}}$ w/ $\mathcal{L}_{\text{Spectral}}$ w/o W) improves both segmentation and calibration metrics over the baseline, confirming the benefit of frequency/domain analysis. However, the gains in boundary/sensitive metrics (HD95, ASD) are modest, suggesting that instantaneous spectral feedback is noisy. Our full method, which adds temporal smoothing of the power spectrum ($\mathcal{L}_{\text{CE}}$ w/ $\mathcal{L}_{\text{Spectral}}$), yields substantial improvements across all metrics. This demonstrates that stabilizing the spectral target is crucial for achieving superior segmentation accuracy and robust calibration.

6 CONCLUSION

In this work, we introduced a frequency-aware calibration framework to address the persistent challenge of confidence miscalibration in medical image segmentation. By analyzing model uncertainty through the lens of frequency domain behavior, we identified spectral bias and confidence saturation as two fundamental causes of miscalibration. Our method combines spectral entropy regularization, which promotes a balanced frequency spectrum and increases overall spectral power, with a temporal spectral smoothing module that stabilizes frequency-wise statistics across training batches. This joint design enables the model to more accurately calibrate the predictions of both high-frequency boundary and low-frequency structure. Experiments across six public datasets and multiple segmentation architectures confirm that our approach consistently improves calibration quality without degrading segmentation performance. These results demonstrate the potential of frequency domain calibration as a robust foundation for developing more reliable and trustworthy medical image segmentation systems.

REFERENCES

Edmon Begoli, Tanmoy Bhattacharya, and Dimitri Kusnezov. The need for uncertainty quantification in machine-assisted medical decision making. *Nature Machine Intelligence*, 1(1):20–23, 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.

Yuntian Bo, Yazhou Zhu, Lunbo Li, and Haofeng Zhang. Famnet: Frequency-aware matching network for cross-domain few-shot medical image segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 1889–1897, 2025.

Chaoyu Chen, Xin Yang, Rusi Chen, Junxuan Yu, Liwei Du, Jian Wang, Xindi Hu, Yan Cao, Yingying Liu, and Dong Ni. Ffpn: Fourier feature pyramid network for ultrasound image segmentation. In *International Workshop on Machine Learning in Medical Imaging*, pp. 166–175. Springer, 2023a.

Fang Chen, Jian Wang, Han Liu, Wentao Kong, Zhe Zhao, Longfei Ma, Hongen Liao, and Daoqiang Zhang. Frequency constraint-based adversarial attack on deep neural networks for medical image classification. *Computers in biology and medicine*, 164:107248, 2023b.

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.

Tom Fawcett and Alexandru Niculescu-Mizil. Pav and the roc convex hull. *Machine Learning*, 68: 97–106, 2007.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059. PMLR, 2016.

Vandan Gorade, Sparsh Mittal, Debesh Jha, Rekha Singhal, and Ulas Bagci. Harmonized spatial and spectral learning for generalized medical image segmentation. In *International Conference on Pattern Recognition*, pp. 178–193. Springer, 2024.

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

Richard W Hamming and Samuel D Stearns. Digital filters. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):67–67, 1979.

Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI brainlesion workshop*, pp. 272–284. Springer, 2021.

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

Mobarakol Islam and Ben Glocker. Spatially varying label smoothing: Capturing uncertainty from expert annotations. In *Information Processing in Medical Imaging: 27th International Conference, IPMI 2021, Virtual Event, June 28–June 30, 2021, Proceedings 27*, pp. 677–688. Springer, 2021.

Alain Jungo, Raphael Meier, Ekin Ermis, Marcela Blatti-Moreno, Evelyn Herrmann, Roland Wiest, and Mauricio Reyes. On the effect of inter-observer variability for a reliable estimation of uncertainty of medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I*, pp. 682–690. Springer, 2018.

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

Agostina J Larrazabal, César Martínez, Jose Dolz, and Enzo Ferrante. Maximum entropy on erroneous predictions: Improving model calibration for medical image segmentation. In *International conference on medical image computing and computer-assisted intervention*, pp. 273–283. Springer, 2023.

Jiaming Li, Hongtao Xie, Jiahong Li, Zhongyuan Wang, and Yongdong Zhang. Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6458–6467, 2021.

Li Li, Qiyuan Liu, Xinyi Shi, Yujia Wei, Huanqi Li, and Hanguang Xiao. Ucfiltransnet: Cross-filtering transformer-based network for ct image segmentation. *Expert Systems with Applications*, 238:121717, 2024.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.

Geert Litjens, Robert Toth, Wendy Van De Ven, Caroline Hoeks, Sjoerd Kerkstra, Bram Van Ginneken, Graham Vincent, Gwenael Guillard, Neil Birbeck, Jindang Zhang, et al. Evaluation of prostate segmentation algorithms for mri: the promise12 challenge. *Medical image analysis*, 18 (2):359–373, 2014.

Jun Ma, Yao Zhang, Song Gu, Xingle An, Zhihe Wang, Cheng Ge, Congcong Wang, Fan Zhang, Yu Wang, Yinan Xu, et al. Fast and low-gpu-memory abdomen ct organ segmentation: the flare challenge. *Medical Image Analysis*, 82:102616, 2022.

Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34 (10):1993–2024, 2014.

Jishnu Mukhoti, Viveka Kulharia, Amartya Sanyal, Stuart Golodetz, Philip Torr, and Puneet Dokania. Calibrating deep neural networks using focal loss. *Advances in neural information processing systems*, 33:15288–15299, 2020.

Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? *Advances in neural information processing systems*, 32, 2019.

Balamurali Murugesan, Sukesh Adiga Vasudeva, Bingyuan Liu, Herve Lombaert, Ismail Ben Ayed, and Jose Dolz. Trust your neighbours: Penalty-based constraints for model calibration. In *International conference on medical image computing and computer-assisted intervention*, pp. 572–581. Springer, 2023a.

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

Balamurali Murugesan, Julio Silva-Rodriguez, Ismail Ben Ayed, and Jose Dolz. Class and region-adaptive constraints for network calibration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 57–67. Springer, 2024.

Balamurali Murugesan, Sukesh Adiga Vasudeva, Bingyuan Liu, Hervé Lombaert, Ismail Ben Ayed, and Jose Dolz. Neighbor-aware calibration of segmentation networks with penalty-based constraints. *Medical Image Analysis*, pp. 103501, 2025.

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

Jeremy Nixon, Michael W Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. Measuring calibration in deep learning. In *CVPR workshops*, 2019.

Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.

Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. Regularizing neural networks by penalizing confident output distributions. *arXiv preprint arXiv:1701.06548*, 2017.

John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

Félix Quinton, Romain Popoff, Benoît Presles, Sarah Leclerc, Fabrice Meriaudeau, Guillaume Nodari, Olivier Lopez, Julie Pellegrinelli, Olivier Chevallier, Dominique Ginhac, et al. A tumour and liver automatic segmentation (atlas) dataset on contrast-enhanced magnetic resonance imaging for hepatocellular carcinoma. *Data*, 8(5):79, 2023.

Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *International conference on machine learning*, pp. 5301–5310. PMLR, 2019.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pp. 234–241. Springer, 2015.

Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31, 2018.

Yue Sun, Kun Gao, Zhengwang Wu, Guannan Li, Xiaopeng Zong, Zhihao Lei, Ying Wei, Jun Ma, Xiaoping Yang, Xue Feng, et al. Multi-site infant brain segmentation algorithms: the iseg-2019 challenge. *IEEE Transactions on Medical Imaging*, 40(5):1363–1376, 2021.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

Xianlun Tang, Jiangping Peng, Bing Zhong, Jie Li, and Zhenfu Yan. Introducing frequency representation into convolution neural networks for medical image segmentation via twin-kernel fourier convolution. *Computer Methods and Programs in Biomedicine*, 205:106110, 2021.

Christian Tomani, Sebastian Gruber, Muhammed Ebrar Erdem, Daniel Cremers, and Florian Buettner. Post-hoc uncertainty calibration for domain drift scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10124–10132, 2021.

Meet P Vadera, Soumya Ghosh, Kenney Ng, and Benjamin M Marlin. Post-hoc loss-calibration for bayesian neural networks. In *Uncertainty in Artificial Intelligence*, pp. 1403–1412. PMLR, 2021.

Zhi-Qin John Xu, Yaoyu Zhang, Tao Luo, Yanyang Xiao, and Zheng Ma. Frequency principle: Fourier analysis sheds light on deep neural networks. *arXiv preprint arXiv:1901.06523*, 2019.

Michael Yeung, Leonardo Rundo, Yang Nan, Evis Sala, Carola-Bibiane Schönlieb, and Guang Yang. Calibrating the dice loss to handle neural network overconfidence for biomedical image segmentation. *Journal of Digital Imaging*, 36(2):739–752, 2023.

Dong Yin, Raphael Gontijo Lopes, Jon Shlens, Ekin Dogus Cubuk, and Justin Gilmer. A fourier perspective on model robustness in computer vision. *Advances in Neural Information Processing Systems*, 32, 2019.

Bianca Zadrozny and Charles Elkan. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *Icml*, 2001.

Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE transactions on medical imaging*, 39(6):1856–1867, 2019.

Yiqin Zhu, Jianyong Chen, Lingyu Liang, Zhanghui Kuang, Lianwen Jin, and Wayne Zhang. Fourier contour embedding for arbitrary-shaped text detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3123–3131, 2021.

A APPENDIX

This appendix provides supplementary materials to support the main manuscript. We offer a more detailed interpretation of the multi-dimensional performance comparison and a comprehensive sensitivity analysis of our method's key hyperparameters. A statement regarding the use of Large Language Models (LLMs) in preparing this manuscript is provided in Section A.1.

A.1 STATEMENT ON LLM USAGE

In accordance with conference guidelines, we disclose the use of a Large Language Model (LLM) during the preparation of this manuscript. The LLM was employed exclusively for copyediting to improve the clarity, conciseness, and grammatical correctness of the text. All scientific contributions—including research ideation, methodological design, experimental setup, data analysis, and interpretation of results—were conducted entirely by the authors. The LLM did not generate scientific content or insights. The authors have reviewed, edited, and assume full responsibility for the final content of this paper.

A.2 DATA PREPROCESSING

A standard preprocessing pipeline was applied to all medical image data to ensure consistency across datasets and stable model training. This pipeline was uniformly applied to all methods under comparison to ensure a fair evaluation.

Image Resampling and Spacing Normalization: All images were resampled to a unified isotropic voxel spacing, defined as the median voxel spacing of the dataset. Trilinear interpolation was used for intensity images, whereas nearest-neighbor interpolation was applied to segmentation masks to preserve discrete label integrity.

Intensity Normalization: For CT images, Hounsfield Unit (HU) values were clipped to a clinically relevant range of [-1000, 400] HU, followed by Z-score normalization (i.e., subtracting the mean and dividing by the standard deviation). For MRI data, Z-score normalization was applied exclusively to foreground voxels (i.e., non-zero values) to prevent background regions from skewing the statistics.

Data Augmentation: During training, a suite of data augmentation techniques was employed to enhance model robustness and generalization, including gamma transformation, additive Gaussian noise, Gaussian blurring, and brightness adjustment.

A.3 THEORETICAL MOTIVATION

From Confidence Maps to 3D Signals. Our framework formulates the model's confidence map as a 3D signal. Given a 3D medical image, a segmentation model f_θ produces logits $z(\mathbf{x})$ at each spatial location $\mathbf{x} \in \mathbb{R}^3$. A scalar confidence map, $c(\mathbf{x})$, is derived from these logits (e.g., via the maximum logit value) to represent the model's spatial confidence distribution. The 3D Fourier Transform decomposes this signal into its constituent frequency components:

$$\hat{c}(\mathbf{k}) = \mathcal{F}\{c(\mathbf{x})\} = \iiint_{\mathbb{R}^3} c(\mathbf{x}) e^{-i2\pi\mathbf{k}\cdot\mathbf{x}} d\mathbf{x} \tag{11}$$

where $\mathbf{k}$ is the 3D frequency vector.

Spectral Signatures of Segmentation Quality. The Power Spectral Density (PSD), $E(\mathbf{k}) = |\hat{c}(\mathbf{k})|^2$, quantifies the distribution of the signal's energy across frequencies. By Parseval's theorem, the total energy is conserved between the spatial and frequency domains. Low-frequency components correspond to large, smooth structures in $c(\mathbf{x})$, whereas high-frequency components capture sharp transitions and fine details. An effective model should produce confidence maps that are smooth within homogeneous regions (low frequencies) but sharp at anatomical boundaries (high frequencies). A common failure mode, particularly for over-confident models, is the suppression of high-frequency detail, yielding overly smooth or "blob-like" predictions. This manifests as a PSD pathologically concentrated in the low-frequency bands.

Spectral Entropy as an Information-Theoretic Regularizer. We employ spectral entropy to quantify the structural complexity of the confidence map. The frequency space is partitioned into K disjoint bands $\{\mathcal{I}_k\}_{k=1}^{K}$, and the power in each band is defined as $S^{(k)} = \int_{\mathbf{k}\in\mathcal{I}_k} E(\mathbf{k})d\mathbf{k}$. Normalizing these values yields a probability distribution $P = \{P^{(k)}\}_{k=1}^{K}$, where $P^{(k)} = S^{(k)}/\sum_j S^{(j)}$. The spectral entropy is the Shannon entropy of this distribution:

$$H(P) = -\sum_{k=1}^{K} P^{(k)} \log(P^{(k)} + \varepsilon) \tag{12}$$

This metric aligns with the **Principle of Maximum Entropy**, where a higher-entropy distribution is more uniform and embodies fewer structural assumptions. Low spectral entropy indicates a degenerate, low-information output (e.g., over-smoothed) characterized by an overly simplistic confidence structure.

Objective as a Corrective Regularizer. Our objective does not maximize entropy directly but instead acts as a corrective measure against spectral collapse. The loss, $\mathcal{L}_{\text{Spectral}} \propto \max(0, H_{\text{target}} - H_{\text{sample}})^2$, penalizes a sample only if its spectral entropy H_{sample} falls below a dynamic target H_{target}. This design is theoretically motivated to enforce a minimum level of structural complexity, thereby preventing the aforementioned failure mode without discouraging valid, spectrally rich predictions. It steers the model away from degenerate solutions characterized by overconfidence and excessive structural simplicity.

A.4 ALGORITHM PSEUDOCODE

Algorithm 1 details the proposed training procedure. Each training batch undergoes three main steps:

1. **Per-Sample Spectral Analysis:** Following a standard forward pass and cross-entropy loss computation, a scalar confidence map is generated from the maximum logits for each sample. This map is transformed into the frequency domain, and its PSD is aggregated into K frequency bands to produce a spectral power vector.
2. **Dynamic Target Generation:** Spectral power vectors from the current batch are averaged and appended to a historical buffer of size W. Averaging this buffer yields a temporally smoothed spectral profile, providing a stable estimate of the desired spectral characteristics. The Shannon entropy of this profile defines the dynamic target entropy for the current iteration.
3. **Regularization and Optimization:** The spectral entropy of each sample is compared against the dynamic target. A hinge-like loss penalizes samples with entropy below this target, thereby promoting a more informative spectral distribution. The final objective is a weighted sum of the cross-entropy loss and this spectral regularization loss.

A.5 COMPREHENSIVE PERFORMANCE COMPARISON

Fig. 5 presents a holistic performance evaluation, visualizing the trade-off between segmentation accuracy and uncertainty calibration. In this visualization, superior methods enclose a larger, more regular polygon. The ideal method would maximize this area, indicating superior performance across all metrics. Our proposed method exhibits a notably balanced and expansive shape, demonstrating consistently high segmentation scores (high DSC, low HD95/ASD) and strong calibration performance (low ECE/SCE/TACE). In contrast, competing methods often display "spiky" profiles; for instance, some achieve a high DSC at the cost of significantly worse calibration metrics (e.g., ECE, TACE), indicating poor uncertainty estimation. This visual analysis underscores the primary

Algorithm 1 Frequency-Aware Calibration Training Procedure

```
Require: Model $f_\theta$, training data loader $\mathcal{D}$, loss weight $\lambda$, window size $W$, number of frequency bands $K$.
Ensure: Trained model parameters $\theta$.
Initialize model parameters $\theta$.
Initialize historical spectral power buffer $\mathcal{S}_{\text{hist}} \leftarrow []$.
for each batch $\{x, y\}$ in $\mathcal{D}$ do
    // Forward Pass and Standard Segmentation Loss
    $z \leftarrow f_\theta(x)$    ▷ Get logits, shape: $(B, C, D, H, W)$
    $\mathcal{L}_{\text{CE}} \leftarrow$ CrossEntropyLoss$(z, y)$
    // Per-Sample Spectral Analysis
    Initialize batch spectral loss $\mathcal{L}_{\text{Spectral}} \leftarrow 0$.
    Let $S_{\text{batch}}$ be an empty list to store spectral vectors for the current batch.
    for $b = 1, \ldots, B$ do    ▷ Iterate over each sample in the batch
        $z_b \leftarrow \max_c z_{b,c}$    ▷ Generate scalar confidence map from logits
        $F_b \leftarrow$ Shift(FFT3D$(z_b)$)    ▷ Transform to frequency domain
        $E_b \leftarrow |F_b|^2$    ▷ Compute Power Spectral Density (PSD)
        Let $S_b = (S_b^{(1)}, \ldots, S_b^{(K)})$ be the spectral power vector for sample $b$.
        for $k = 1, \ldots, K$ do
            $S_b^{(k)} \leftarrow \sum_{(u,v,w) \in \mathcal{I}_k} E_b(u, v, w)$    ▷ Aggregate power in frequency band $k$
        end for
        Append $S_b$ to $S_{\text{batch}}$.
    end for
    // Update Smoothed Spectral Profile and Target Entropy
    $\bar{S}_{\text{batch}} \leftarrow \frac{1}{B} \sum_{b=1}^{B} S_{\text{batch}}[b]$    ▷ Average spectral power for the current batch
    Append $\bar{S}_{\text{batch}}$ to $\mathcal{S}_{\text{hist}}$.
    if $|\mathcal{S}_{\text{hist}}| > W$ then
        Remove oldest entry from $\mathcal{S}_{\text{hist}}$.
    end if
    $\widetilde{S} \leftarrow \frac{1}{|\mathcal{S}_{\text{hist}}|} \sum_{S' \in \mathcal{S}_{\text{hist}}} S'$    ▷ Compute smoothed spectral profile
    $\widetilde{P} \leftarrow$ Normalize$(\widetilde{S})$    ▷ Convert to probability distribution
    $H_{\text{target}} \leftarrow$ ShannonEntropy$(\widetilde{P})$    ▷ Compute target entropy
    // Compute Spectral Regularization Loss for the Batch
    for $b = 1, \ldots, B$ do
        $S_b \leftarrow S_{\text{batch}}[b]$
        $P_b \leftarrow$ Normalize$(S_b)$
        $H_{\text{sample}} \leftarrow$ ShannonEntropy$(P_b)$
        $\text{loss}_b \leftarrow \max(0, H_{\text{target}} - H_{\text{sample}})^2$
        $\mathcal{L}_{\text{Spectral}} \leftarrow \mathcal{L}_{\text{Spectral}} + \text{loss}_b$
    end for
    $\mathcal{L}_{\text{Spectral}} \leftarrow \frac{1}{B} \mathcal{L}_{\text{Spectral}}$
    // Update Model Parameters
    $\mathcal{L}_{\text{total}} \leftarrow \mathcal{L}_{\text{CE}} + \lambda \mathcal{L}_{\text{Spectral}}$
    Update $\theta$ using gradients from $\mathcal{L}_{\text{total}}$.
end for
```

strength of our approach: achieving a robust balance between segmentation accuracy and uncertainty calibration, a critical requirement for clinical applications where both prediction correctness and confidence reliability are paramount.

A.6 HYPERPARAMETER SENSITIVITY ANALYSIS

Table 4 presents a sensitivity analysis of the key hyperparameters, revealing consistent trends across the BraTS2020 and FLARE2021 datasets. The regularization weight λ exhibits a clear trade-off: small values provide insufficient regularization (high ECE), whereas excessively large values degrade segmentation performance (low DSC). Optimal values (e.g., $\lambda = 0.01$ for BraTS2020,

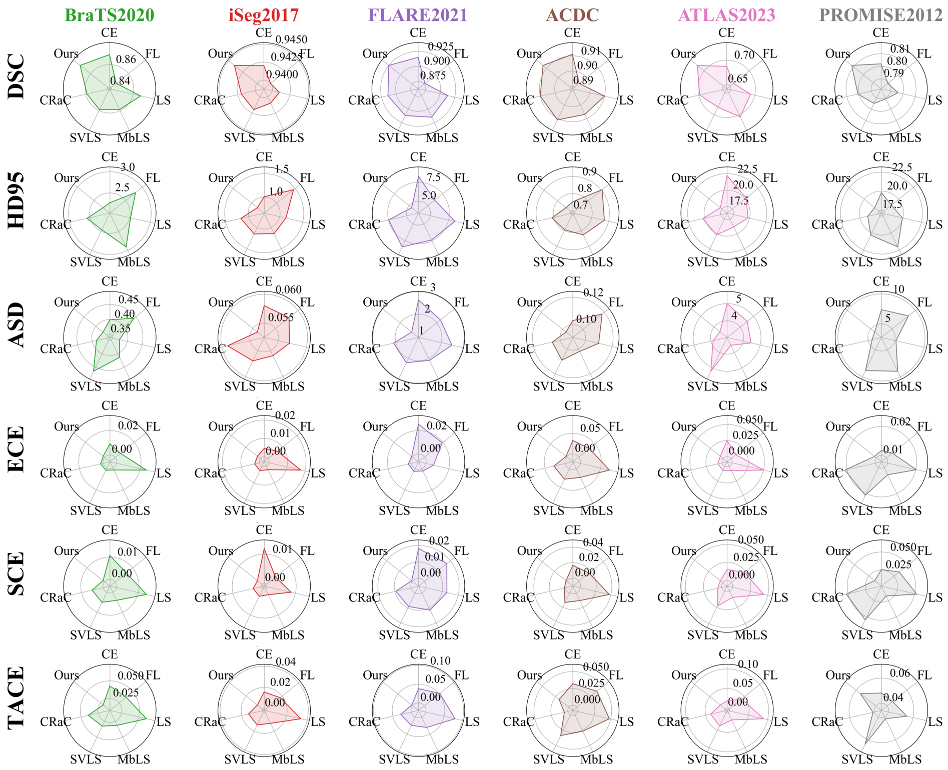


Figure 5: Multi-dimensional radar chart comparing eight methods across segmentation (DSC, HD95, ASD) and calibration (ECE, SCE, TACE) metrics. A larger, more regular area indicates better overall performance.

Table 4: Sensitivity analysis of the key hyperparameters (λ, W, K) on the BraTS2020 and FLARE2021 datasets.

Varying Hyperparameter	Value	BraTS2020		FLARE2021	
		DSC↑	ECE↓	DSC↑	ECE↓
λ (Fixed $W = 50, K = 5$)	0.001	0.870	0.0045	0.922	0.0028
	0.005	0.871	0.0020	0.924	0.0015
	0.010	0.872	0.0015	0.925	0.0008
	0.050	0.871	0.0013	0.925	0.0007
	0.100	0.865	0.0012	0.920	0.0008
W (Fixed $\lambda = 0.01, K = 5$)	10	0.870	0.0025	0.922	0.0014
	25	0.871	0.0018	0.924	0.0010
	50	0.872	0.0015	0.925	0.0008
	75	0.872	0.0016	0.925	0.0009
	100	0.871	0.0022	0.923	0.0013
K (Fixed $\lambda = 0.01, W = 50$)	2	0.870	0.0028	0.923	0.0016
	3	0.871	0.0019	0.925	0.0008
	5	0.872	0.0015	0.925	0.0009
	7	0.871	0.0017	0.924	0.0011
	10	0.870	0.0023	0.922	0.0015

$\lambda = 0.05$ for FLARE2021) balance both objectives effectively. Similarly, the temporal smoothing window W is crucial for a stable entropy target. A small window ($W = 10$) creates a noisy target, whereas a large one ($W = 100$) adapts slowly to distributional shifts; a moderate size ($W \in$) achieves the best results. Finally, the number of frequency bands K determines the granularity of spectral analysis. Too few bands ($K = 2$) oversimplify the spectrum, while too many ($K = 10$) are sensitive to noise. A value between 3 and 7 provides a robust balance, consistent with related work Bo et al. (2025). These results confirm that the parameter response is stable and predictable, providing clear guidance for application to new datasets.

A.7 Distribution Analysis

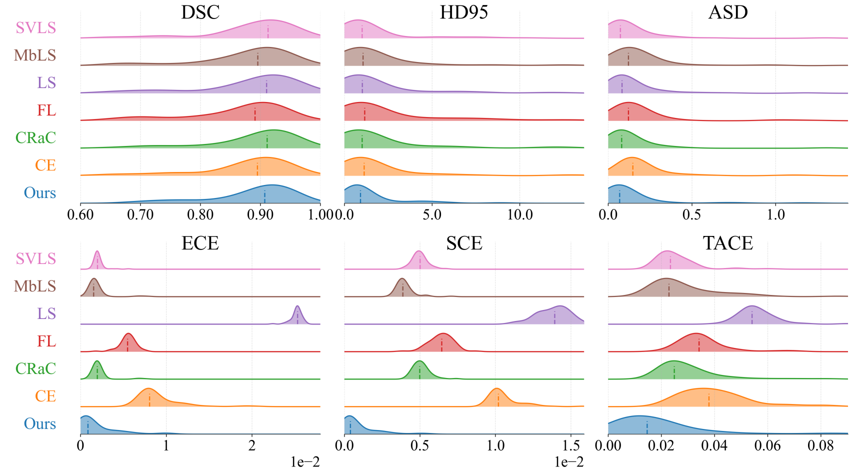


Figure 6: Distribution analysis of segmentation (DSC) and calibration (ECE) metrics on the BraTS2020 dataset. Our method demonstrates improved mean performance and reduced variance compared to the baseline.

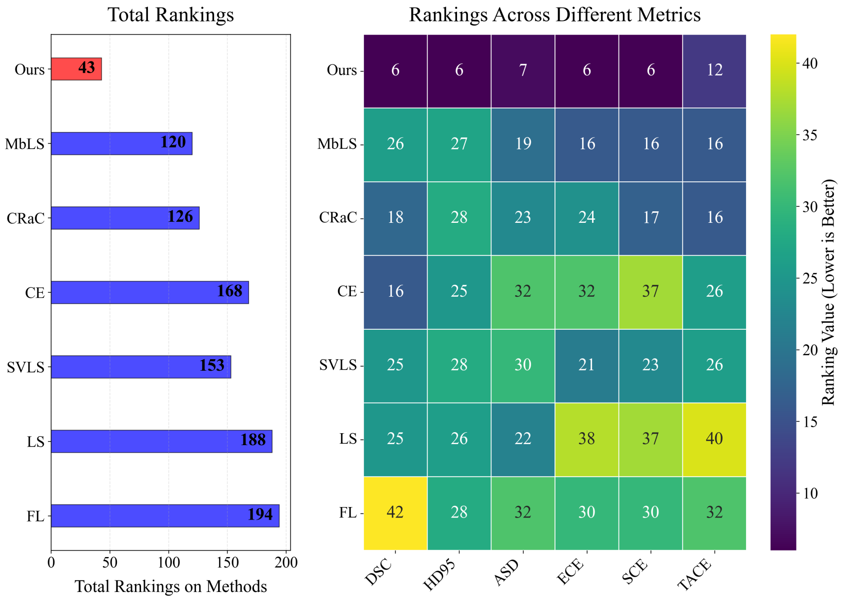
Fig. 6 illustrates the performance distributions for the Dice Score (DSC) and Expected Calibration Error (ECE). Our method's distributions exhibit two key characteristics: (1) a rightward shift in the mean for DSC and a leftward shift for ECE, indicating superior average performance, and (2) reduced variance (a narrower, taller peak). The latter is particularly important, as it demonstrates improved model robustness and reliability. The smaller spread signifies a reduction in outlier predictions and catastrophic failures, where a baseline model might perform poorly on challenging cases. This enhanced consistency is a direct result of our spectral regularization, which prevents the model from adopting overly simplistic, spectrally biased solutions and ensures more reliable performance across the test population.

A.8 Comprehensive Performance Ranking

This analysis evaluates the generalization capabilities of our method across multiple datasets and metrics. As shown in Fig. 7, we aggregate performance ranks to provide a holistic comparison. For aggregation, ranks are assigned to each method on every dataset-metric pair (rank 1 is best) and then summed to yield an overall score (lower is better). The left panel confirms our method's superior aggregate rank. The heatmap on the right is more revealing: competing methods often exhibit specialized performance, excelling on certain metrics or datasets while faltering on others, suggesting potential overfitting to specific data characteristics. In contrast, our method achieves consistently strong performance across the board. This demonstrates that by operating in the frequency domain, our regularizer targets a fundamental aspect of the model's confidence representation rather than dataset-specific spatial statistics, leading to a more robust and holistically effective model.

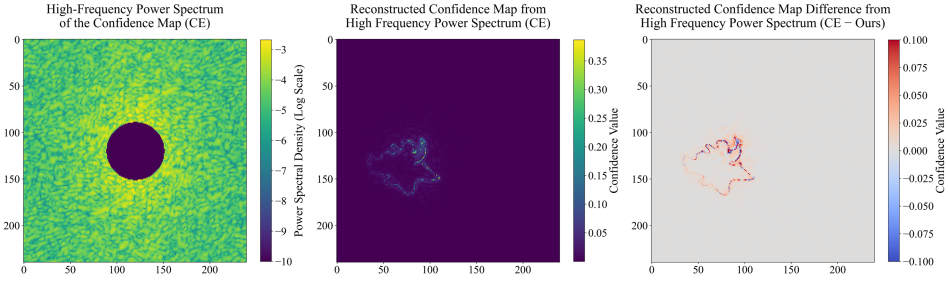
A.9 Visualizing the Mitigation of Spectral Bias

This analysis provides a direct visual link between the theoretical motivation of spectral bias and our method's corrective effect. We isolate the high-frequency components from a baseline model's logit output for a challenging sample. Applying an inverse Fourier transform to these components reveals their spatial manifestation: misplaced, high-confidence hotspots, demonstrating how unchecked high-frequency power can lead to erroneous predictions at complex boundaries (Fig. 8a-b). The confidence difference map (Baseline − Ours) in Fig. 8c highlights regions where our



	DSC	HD95	ASD	ECE	SCE	TACE
Ours	6	6	7	6	6	12
MbLS	26	27	19	16	16	16
CRaC	18	28	23	24	17	16
CE	16	25	32	32	37	26
SVLS	25	28	30	21	23	26
LS	25	26	22	38	37	40
FL	42	28	32	30	30	32

Figure 7: Cross-dataset performance ranking. Left: Aggregate ranking scores (lower is better). Right: Heatmap of individual metric ranks (darker is better). Our method achieves a consistently high rank across diverse metrics and datasets.



(a) (b) (c)

Figure 8: Analysis of high-frequency components on a BraTS2020 sample. (a) High-frequency power from the baseline model's logits. (b) Confidence map reconstructed from (a), showing misplaced high-confidence "hotspots." (c) Confidence difference map (Baseline − Ours), where red indicates overconfidence corrected by our approach, correlating strongly with the hotspots in (b).

method reduces overconfidence. The strong spatial correlation between these regions and the baseline's hotspots serves as direct evidence that our spectral regularization is working as intended. It successfully suppresses problematic high-frequency power, leading to better-calibrated and more structurally coherent confidence maps, particularly in anatomically complex regions.