B-spline Activations as Intrinsic Uncertainty Estimators in Medical Image Segmentation

Wenju Du[1], Zhiwei Wu[1], Bing Zeng[2], Nini Rao[3], and Wei Liu[1⋆]

[1] College of Information and Artificial Intelligence (College of Industrial Software), Yangzhou University, Yangzhou 225127, China
wenjudu@yzu.edu.cn, wzw2502350654@gmail.com, weiliu@yzu.edu.cn

[2] School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China
eezeng@uestc.edu.cn

[3] School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu 610054, China
raonn@uestc.edu.cn

Abstract. A fundamental tension underlies uncertainty estimation for clinical segmentation: Bayesian methods—MC Dropout, Deep Ensembles, TTA—demand repeated inference incompatible with real-time workflows, while single-pass alternatives either alter the training objective (EDL) or impose global architectural constraints that risk distorting learned representations (DUQ, SNGP, DDU). The core scientific question is whether prediction uncertainty can be read directly from the *activation geometry* of a learned function in a single forward pass, without auxiliary losses, repeated sampling, or global architectural change. We identify a principled answer in the *local-support* property of B-spline basis functions within Kolmogorov–Arnold Network (KAN) layers: because each basis function has compact support, activation *concentration* directly proxies training-data density—well-covered inputs produce winner-take-all activation while distribution boundaries and misclassified sites produce dispersed patterns. We verify this statistically (Mann–Whitney $p<0.001$, Cohen's d up to 1.61 across colonoscopy, ultrasound, and dermoscopy) and operationalize it via the Residual KAN Head, a plug-and-play classifier adding <0.01M parameters that preserves segmentation accuracy (Dice change <0.3%). On four medical imaging benchmarks, our single-pass method achieves the best selective prediction performance in 4 of 12 backbone×dataset settings at 6.9× faster inference than multi-pass alternatives.

Keywords: Uncertainty estimation · Medical image segmentation · Kolmogorov–Arnold Networks · B-spline basis functions

⋆ Corresponding author.

1 Introduction

Clinical deployment of deep segmentation models demands reliable uncertainty estimates: models that flag uncertain predictions enable clinicians to prioritize review of suspicious cases in time-critical workflows such as intraoperative navigation and endoscopic analysis [12–14]. Although uncertainty correlates with segmentation error across architectures [17], turning it into a practical real-time signal remains a key bottleneck [18].

Multi-pass methods dominate: MC Dropout [1] runs T=10–20 stochastic passes, Deep Ensembles [2] trains N models, and TTA [3] augments inputs at test time—all incurring a $T/N/K\times$ overhead impractical for latency-sensitive settings. Single-pass alternatives carry trade-offs: softmax entropy [4] captures only output-level confidence, insensitive to distributional coverage; EDL [5] places Dirichlet priors but requires a modified loss and can be poorly calibrated [10, 11]; DUM methods (DUQ [15], SNGP [26], DDU [16]) leverage feature-space geometry but impose global architectural constraints that may interfere with backbone representations. This raises a natural question: can a *local*, lightweight module yield a deterministic uncertainty signal without touching the backbone or the training objective?

The recent adoption of KAN layers in medical image segmentation [7] points to an answer: KAN [6] replaces fixed activations with learnable B-spline functions whose *local-support* property means well-covered inputs concentrate activation on few basis functions while sparse or out-of-distribution inputs spread it across many—a *deterministic distributional-coverage* signal in a single forward pass requiring no changes to training or backbone.

We make three contributions: a statistical characterization of B-spline activation dispersion as a pixel-level uncertainty signal (Cohen's d up to 1.61, p<0.001 across all modalities and seeds); the Residual KAN Head, a plug-and-play module that preserves segmentation accuracy (Dice change <0.3%) while exposing this signal, adding <0.01M parameters; and an empirical demonstration that the resulting single-pass method achieves the best E-AURC (defined in Sec. 2.3) in 4 of 12 backbone×dataset settings—including the overall best result (18.7 on ISIC/UNet)—at 6.9× faster inference than multi-pass alternatives.

2 Method

Fig. 1 illustrates the proposed approach: the Residual KAN Head fuses a Conv branch for accurate segmentation with an α-KAN branch whose B-spline basis activations serve as an intrinsic uncertainty sensor; per-pixel activation spread is aggregated to an image-level score via mean pooling for use in selective prediction.

2.1 B-spline Basis Functions and Uncertainty

Mechanism. Each input dimension in a KAN layer is transformed through B-spline basis functions $\{B_k(x)\}_{k=1}^{K}$ where K=G+S (G: grid size, S: spline or-

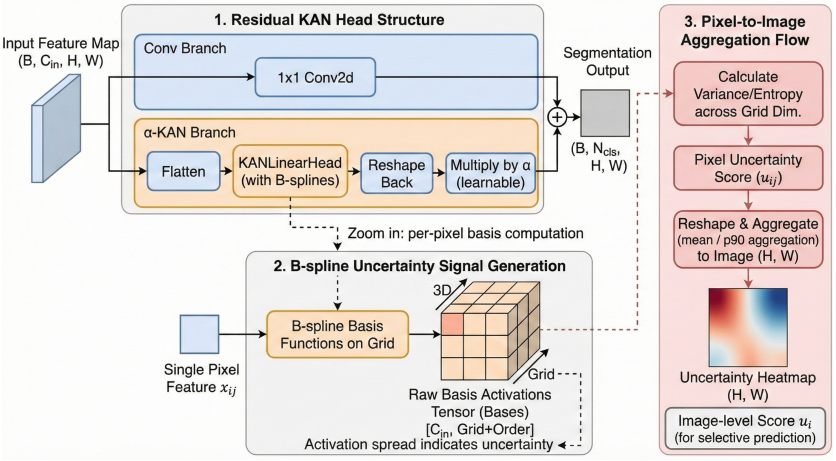


Fig. 1. Method overview. The Residual KAN Head replaces the final 1×1 convolution. The Conv branch provides the main prediction while the $\alpha \cdot$KAN branch learns residual increments via B-spline basis functions. Uncertainty is extracted from activation patterns and aggregated to image-level scores during a single forward pass.

der). Each B_k is a degree-S piecewise polynomial with *compact (local) support* (Cox–de Boor recursion): nonzero only over $S{+}1$ adjacent grid intervals, so any input activates at most $S{+}1$ neighbouring bases, localizing it within the grid (plotted in Fig. 4a). Features from well-covered training regions concentrate activation on a few basis functions (*winner-take-all*), whereas sparse or unseen inputs spread activation across many (*dispersed pattern*). We emphasize that this signal captures an *epistemic* component of predictive reliability arising from limited training-data coverage, rather than a full Bayesian uncertainty decomposition.

Fig. 2 illustrates this on a challenging case (Dice = 0.712): a correct pixel ❶ shows concentrated activation, a misclassified pixel ❷ a dispersed pattern, and a boundary pixel ❸ an intermediate one.

As shown in Fig. 3, across colonoscopy, ultrasound, and dermoscopy, misclassified pixels show higher activation entropy than correct ones, and boundary pixels higher than interior ones (Cohen's $d = 0.07$–1.61 and 0.71–1.43, $p<0.001$ throughout).

Uncertainty metrics. *Activation entropy* measures normalized activation uniformity:

$$U_{\text{ent}}(x) = -\sum_{k=1}^{K} p_k \log p_k, \qquad p_k = \frac{b_k}{\sum_j b_j + \epsilon} \tag{1}$$

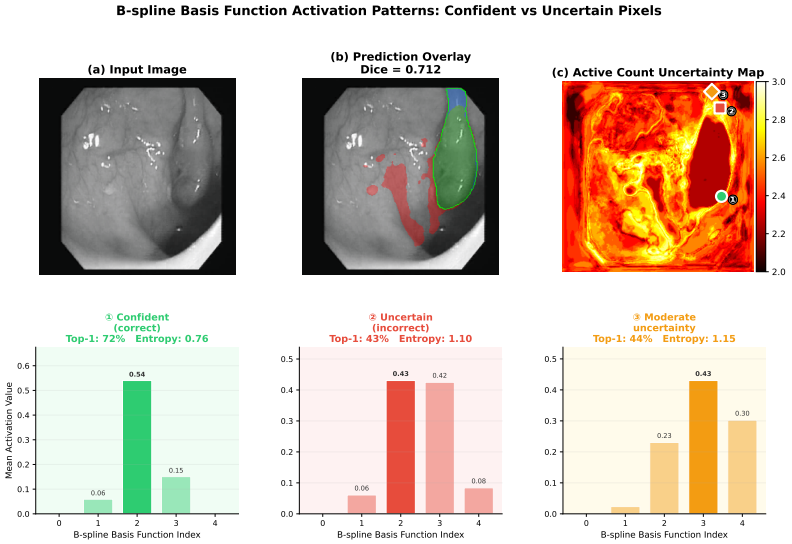


Fig. 2. B-spline activation patterns for three representative pixels (CVC-ClinicDB/U-KAN, seed 42, Dice = 0.712). *Top*: input, prediction overlay, and Active Count uncertainty map. *Bottom*: mean activation vectors for a confident pixel ❶ (concentrated), a misclassified pixel ❷ (dispersed), and a boundary pixel ❸ (intermediate).

Active count counts basis functions exceeding threshold τ:

$$U_{\text{ac}}(x) = \sum_{k=1}^{K} \mathbf{1}[b_k(x) > \tau], \qquad \tau = 0 \tag{2}$$

Entropy provides the most consistent pixel-level separation; active count with mean aggregation yields stronger image-level ranking.

2.2 Residual KAN Head

Extracting this signal requires a module that exposes the activations without sacrificing segmentation accuracy. Directly replacing the final convolution with a KAN layer degrades accuracy (Dice drops by ∼0.9% on BUSI). We propose the Residual KAN Head:

$$\hat{y} = f_{\text{conv}}(x) + \alpha \cdot f_{\text{kan}}(x) \tag{3}$$

where f_{conv} is a standard 1×1 convolution, f_{kan} is a KAN linear layer, and α is a learnable scalar initialized to 0, allowing the head to reduce to a pure convolution at training onset. We use grid_size=3 and spline_order=2, adding <0.01M parameters.

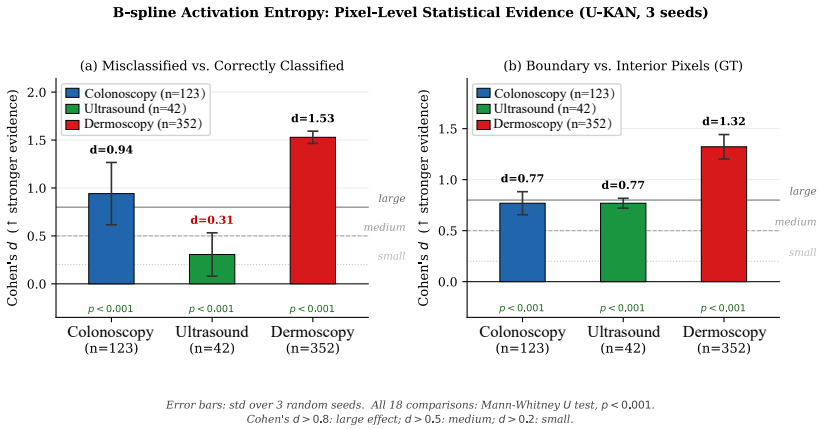


Fig. 3. Pixel-level statistical evidence across three medical imaging modalities (U-KAN backbone; error bars: seed range). *(a)* Misclassified vs. correct pixels: Cohen's d = 0.07–1.61, p<0.001 in all comparisons. *(b)* Boundary vs. interior pixels: d = 0.71–1.43, p<0.001 in all comparisons. Dashed lines mark small/medium/large effect size thresholds.

2.3 Evaluation Framework

Per-pixel maps are mean-pooled to image-level scores, and we frame uncertainty evaluation as selective prediction [8, 9]: given n validation images indexed from most certain to least certain, the Area Under the Risk-Coverage curve (AURC) integrates selective risk over all coverage levels:

$$\text{AURC} = \frac{1}{n}\sum_{i=1}^{n}\bar{r}_i, \qquad \bar{r}_i = \frac{1}{i}\sum_{j=1}^{i}\left(1 - \text{Dice}_j\right) \tag{4}$$

where $\bar{r}_i$ is the mean risk when the i most confident images are retained. To isolate ranking quality from baseline accuracy, we report the Excess AURC (E-AURC):

$$\text{E-AURC} = \text{AURC}_{\text{method}} - \text{AURC}_{\text{oracle}} \quad (\downarrow) \tag{5}$$

where $\text{AURC}_{\text{oracle}}$ is obtained by sorting images by their true Dice loss.

3 Experiments

3.1 Setup

Datasets. (1) CVC-ClinicDB [19]: 612 colonoscopy polyp images, 80/20 split (123 val). (2) BUSI [20]: breast ultrasound, 80/20 split (42 val). (3) ISIC 2018 [21]: 1762 dermoscopic images, 80/20 split (352 val). (4) REFUGE [22]: 1200 retinal fundus images, standard split (240 val).

Table 1. Segmentation Dice (3-seed mean±std). Res-KAN preserves accuracy (change <0.3%). EDL Dice drops are concentrated on REFUGE (up to −2.5%); changes on CVC/BUSI/ISIC are within 1%.

Head	Backbone	CVC	BUSI	ISIC	REFUGE
Conv	U-KAN	0.869±0.011	0.741±0.032	0.895±0.005	0.879±0.002
	UNet	0.879±0.010	0.730±0.005	0.894±0.002	0.881±0.003
	TransUNet	0.877±0.009	0.744±0.025	0.900±0.001	0.879±0.006
Res-KAN (Ours)	U-KAN	0.873±0.008	0.751±0.027	0.895±0.005	0.881±0.002
	UNet	0.877±0.011	0.732±0.011	0.896±0.003	0.883±0.002
	TransUNet	0.870±0.014	0.746±0.011	0.901±0.002	0.878±0.002
EDL	U-KAN	0.859±0.004	0.743±0.018	0.894±0.006	0.854±0.003
	UNet	0.876±0.005	0.724±0.008	0.897±0.003	0.865±0.006
	TransUNet	0.879±0.008	0.739±0.023	0.900±0.003	0.870±0.001

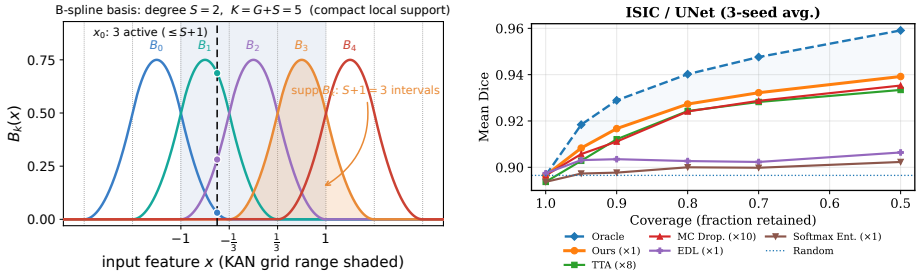
Backbones. U-KAN [7] (embed_dims=[128,160,256]), UNet [23] (base_c=32), and TransUNet [24], each paired with Conv head and Residual KAN Head.
Training. All images resized to 256×256. Adam [25] (lr=10^{-4}), cosine annealing, BCE + Dice loss, 200 epochs, 3 seeds (42, 1187, 2981). Training is fully end-to-end with no frozen components.
Baselines. MC Dropout [1] (T=10), TTA [3] (×8), EDL [5], Softmax Entropy [4]. Our method uses active count (τ=0) with mean aggregation.
Reproducibility. Code is available at: `https://github.com/Wenju-Du/B-spline-uncertainty-segmentation`.

3.2 Segmentation Performance

Table 1 confirms the Residual KAN Head preserves accuracy across all four datasets and three backbones (Dice change <0.3%).

3.3 Uncertainty Quality: Selective Prediction

We further investigate whether the B-spline activation signal ranks images by prediction reliability. Table 2 summarizes E-AURC across all settings. Fig. 4(b) shows representative selective prediction curves on ISIC 2018/UNet.
Overall picture. Across 12 backbone×dataset settings, our single-pass method wins 4 (CVC/UNet, BUSI/U-KAN, ISIC/UNet, REFUGE/TransUNet), including the lowest E-AURC in the table (18.7). TTA (×8) wins 5 and MC Dropout (×10) wins 2, but at 7–10× more computation. Among single-pass methods, ours and EDL each win 6 of 12; however, EDL requires a modified loss and incurs Dice degradation of up to 2.5% on REFUGE, whereas our module preserves accuracy across all settings, with greater cross-seed stability than TTA on CVC/UNet (±5.5 vs. ±58.3).
Statistical significance. A Wilcoxon signed-rank test over the 12 backbone×dataset settings shows our single-pass score significantly outperforms Softmax entropy



(a) B-spline basis $\{B_k(x)\}_{k=1}^{K}$ (b) Selective prediction (ISIC 2018/UNet)

Fig. 4. (a) The B-spline basis $\{B_k(x)\}_{k=1}^{K}$: each has compact support, nonzero over only $S+1$ adjacent grid intervals, so any input activates only a few neighbouring bases. (b) Selective prediction curves (3-seed average); at lower coverage our single-pass score yields higher selective Dice than TTA (×8) and MC Dropout (×10), approaching oracle ranking.

(better in 11/12 settings, p=0.0024) and reaches statistical parity with the strongest multi-pass baselines—TTA×8 (p=0.52) and MC Dropout×10 (p=0.79)—as well as with EDL (p=0.79). Cross-seed variation on the strongest settings is small (e.g., ISIC/UNet: 18.7±1.2).

Data-sufficiency threshold scales with model complexity. Fixing the dataset at CVC (n=123), UNet and U-KAN achieve low E-AURC (29.3 and 29.1) while TransUNet degrades to 74.7. Fixing TransUNet and increasing data reveals clear recovery: 81.4 (BUSI, n=42) → 74.7 (CVC) → 29.6 (ISIC, n=352) → 37.9 (REFUGE, n=240), whereas UNet recovers already at n=123. TransUNet's larger parameter count requires broader coverage before B-spline activations develop discriminative concentration gradients; with sufficient data it yields the best single-setting result (TransUNet/REFUGE: 37.9).

3.4 Inference Efficiency

Our efficiency claim concerns *sequentially repeated* inference (Table 3): TTA (×8) and MC Dropout (T=10) incur ∼8–10× overhead, whereas our single forward pass adds only 1.1× over a plain Conv head (68.5 FPS), just above Softmax and EDL. Yet Softmax ranks far worse (Wilcoxon p=0.0024) and EDL degrades Dice by up to 2.5%, while ours preserves accuracy (<0.3%) without altering training.

3.5 Ablation Studies

Two design choices most affect performance: grid size and the residual structure. Table 4 isolates their effects on BUSI/U-KAN. Reducing grid size from 5 to 3 improves both E-AURC (82.1→50.8) and Dice (0.729→0.742): wider basis functions span a larger input range, so inputs between training clusters activate multiple functions at once, sharpening the concentrated-vs-dispersed contrast. The

Table 2. Uncertainty quality (E-AURC $\times 10^{-3}$, ↓, 3-seed mean±std). Best per column **bolded**. †Single-pass methods.

		CVC-ClinicDB			BUSI			ISIC			REFUGE		
Method	Pass	*U-KAN*	*UNet*	*Trans*	*U-KAN*	*UNet*	*Trans*	*U-KAN*	*UNet*	*Trans*	*U-KAN*	*UNet*	*Trans*
Ours†	×1	29.1 ±9.0	**29.3** ±5.5	74.7 ±23.7	**45.7** ±23.2	80.7 ±6.3	81.4 ±12.8	35.1 ±9.4	**18.7** ±1.2	29.6 ±8.1	46.1 ±4.1	45.6 ±3.3	**37.9** ±4.0
TTA	×8	45.3 ±18.7	57.1 ±58.3	**26.1** ±3.5	46.8 ±21.3	**45.9** ±12.9	64.5 ±14.0	**23.3** ±1.8	23.9 ±2.8	**28.2** ±2.5	**37.9** ±3.8	41.3 ±0.7	50.3 ±6.0
MC Drop.	×10	46.0 ±20.5	53.0 ±55.1	32.8 ±21.8	69.0 ±14.0	75.3 ±17.8	**56.5** ±18.7	27.3 ±4.3	28.3 ±4.8	31.8 ±5.0	46.4 ±4.6	**39.7** ±2.8	56.5 ±3.3
EDL†	×1	**22.5** ±3.0	29.9 ±11.5	63.3 ±10.3	70.6 ±26.4	53.1 ±20.0	66.2 ±2.3	48.9 ±1.7	50.9 ±2.8	52.2 ±7.1	42.6 ±6.2	42.0 ±3.6	42.5 ±3.7
Softmax†	×1	123.3 ±42.9	79.8 ±40.9	64.0 ±20.0	72.4 ±13.7	81.5 ±28.7	93.2 ±26.0	49.3 ±3.8	53.3 ±12.2	49.3 ±4.5	56.7 ±11.3	65.0 ±6.0	55.1 ±2.0

Table 3. Inference time on CVC/U-KAN (RTX 3080 Ti, 256×256). Ours runs at 68.5 FPS, a 6.9× advantage over TTA.

Method	Time (ms)	Overhead	FPS
Conv Forward	12.79	1.0×	78.2
Softmax Entropy	13.12	1.0×	76.2
EDL	13.5	1.1×	74.1
Ours	**14.60**	**1.1×**	**68.5**
TTA ×8	100.69	7.9×	9.9
MC Dropout T=10	129.48	10.1×	7.7

Residual KAN Head improves further (E-AURC 50.8→45.7, Dice 0.742→0.751). The α=0 initialization is key: the KAN branch, bypassed at onset, never competes with the Conv branch for the primary signal and instead specializes on its residual—a lower-amplitude, locally structured signal where B-spline dispersion best discriminates coverage. Accuracy and uncertainty thus improve jointly, not trading off. We fix the spline order S=2; since K=G+S, the grid size G sets the basis resolution and is the primary lever, as the G=5→3 effect (E-AURC 82.1→50.8) confirms. Disentangling spline-order sensitivity is left to future work.

4 Conclusion

B-spline activation dispersion provides a deterministic, single-pass uncertainty signal requiring no backbone modifications or altered training loss. Across four datasets and three backbones, the Residual KAN Head achieves the best E-AURC in 4 of 12 settings, including the overall lowest value (18.7 on ISIC/UNet), at 6.9× faster inference than multi-pass alternatives. Our framework targets

Table 4. Ablation on BUSI/U-KAN (3-seed mean). E-AURC$\times 10^{-3}$ (↓), Dice (↑).

Factor	Variant	E-AURC	Dice
Head design	Pure KAN, Grid=5	82.1	0.729
	Pure KAN, Grid=3	50.8	0.742
	Res-KAN (Ours)	**45.7**	**0.751**

ranking-based selective prediction—deferring uncertain cases for review; as the signal is a continuous scalar, it stays compatible with post-hoc calibration (e.g., temperature scaling) when threshold-based automation is required. The signal strength scales with training-data coverage relative to backbone capacity, and extension to multi-class and volumetric settings remains future work.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (Grant Nos. 62271127, 61872405, and 61720106004), the China Postdoctoral Science Foundation (Grant No. 2025M772885), and the Yangzhou Municipal Science and Technology Program (Grant Nos. SZR2025000010 and YZ2025208).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Gal, Y., Ghahramani, Z.: Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In: ICML. pp. 1050–1059 (2016)
2. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. In: NIPS. pp. 6402–6413 (2017)
3. Wang, G., Li, W., Aertsen, M., et al.: Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. Neurocomputing **338**, 34–45 (2019)
4. Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. In: ICLR (2017)
5. Sensoy, M., Kaplan, L., Kandemir, M.: Evidential deep learning to quantify classification uncertainty. In: NeurIPS. pp. 3179–3189 (2018)
6. Liu, Z., Wang, Y., Vaidya, S., et al.: KAN: Kolmogorov–Arnold Networks. arXiv:2404.19756 (2024)
7. Li, C., Liu, X., Li, W., et al.: U-KAN makes strong backbone for medical image segmentation and generation. In: AAAI. vol. 39, pp. 4652–4660 (2025)
8. Geifman, Y., Uziel, G., El-Yaniv, R.: Bias-reduced uncertainty estimation for deep neural classifiers. In: ICLR (2019)
9. Galil, I., Dabbah, M., El-Yaniv, R.: What can we learn from the selective prediction and uncertainty estimation performance of 523 ImageNet classifiers? In: ICLR (2023)
10. Bengs, V., Hüllermeier, E., Waegeman, W.: Pitfalls of epistemic uncertainty quantification through loss minimisation. In: NeurIPS (2022)
11. Ulmer, D., Hardmeier, C., Frellsen, J.: Prior and posterior networks: A survey on evidential deep learning methods for epistemic uncertainty estimation. Trans. Mach. Learn. Res. (2023)

12. Jungo, A., Reyes, M.: Assessing reliability and challenges of uncertainty estimations for medical image segmentation. In: MICCAI. LNCS, vol. 11765, pp. 48–56. Springer (2019)
13. Nair, T., Precup, D., Arnold, D.L., Arbel, T.: Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation. Med. Image Anal. **59**, 101557 (2020)
14. Judge, T., Bernard, O., Porumb, M., Chartsias, A., Beqiri, A., Jodoin, P.M.: CRISP – Reliable uncertainty estimation for medical image segmentation. In: MICCAI. LNCS, vol. 13438, pp. 492–502. Springer (2022)
15. Van Amersfoort, J., Smith, L., Teh, Y.W., Gal, Y.: Uncertainty estimation using a single deep deterministic neural network. In: ICML. pp. 9690–9700 (2020)
16. Mukhoti, J., Kirsch, A., van Amersfoort, J., Torr, P.H.S., Gal, Y.: Deep deterministic uncertainty: A new simple baseline. In: CVPR. pp. 24384–24394 (2023)
17. Czolbe, S., Arnavaz, K., Krause, O., Feragen, A.: Is segmentation uncertainty useful? In: IPMI. LNCS, vol. 12729, pp. 715–726. Springer (2021)
18. Lambert, B., Forbes, F., Doyle, S., Dehaene, H., Dojat, M.: Trustworthy clinical AI solutions: A unified review of uncertainty quantification in deep learning models for medical image analysis. Artif. Intell. Med. **150**, 102830 (2024)
19. Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., et al.: WM-DOVA maps for accurate polyp highlighting in colonoscopy. IEEE Trans. Med. Imaging **34**(7), 1488–1500 (2015)
20. Al-Dhabyani, W., Gomaa, M., Khaled, H., Fahmy, A.: Dataset of breast ultrasound images. Data Brief **28**, 104863 (2020)
21. Codella, N., Rotemberg, V., Tschandl, P., et al.: Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the ISIC. arXiv:1902.03368 (2019)
22. Orlando, J.I., Fu, H., Breda, J.B., et al.: REFUGE challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. Med. Image Anal. **59**, 101570 (2020)
23. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: MICCAI. LNCS, vol. 9351, pp. 234–241. Springer (2015)
24. Chen, J., Mei, J., Li, X., et al.: TransUNet: Rethinking the U-Net architecture design for medical image segmentation through the lens of transformers. Med. Image Anal. **97**, 103280 (2024)
25. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (2015)
26. Liu, J.Z., Lin, Z., Padhy, S., Tran, D., Bedrax-Weiss, T., Lakshminarayanan, B.: Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. In: NeurIPS. pp. 7498–7512 (2020)