WaCT-US: Cyclic-Alignment Scaled Contour Tubes for Efficient, Calibrated Uncertainty in Echocardiography

Mohammad Mahdi Kazemi Esfeh*[1,2], Zahra Gholami[1], Renjie Liao[1,2,4], Christina Luong[1,3], Teresa Tsang[1,3], and Purang Abolmaesumi[1]

[1] The University of British Columbia, Vancouver, BC, Canada
mahdik@ece.ubc.ca
[2] Vector Institute, Toronto, ON, Canada
[3] Vancouver General Hospital, Vancouver, BC, Canada
[4] Canada CIFAR AI Chair.

Abstract. Uncertainty in echocardiography segmentation is predominantly geometric: speckle, shadowing, and inter-scanner variation create boundary ambiguity that propagates to clinical measurements. We propose WaCT-US, a contour-native uncertainty representation that outputs a calibrated prediction set ("contour tube") around a single predicted boundary. Cyclic landmark alignment resolves closed-curve indexing, and a learned heteroscedastic scale profile specifies where the boundary is intrinsically ambiguous. Split conformal calibration on held-out data yields one global scalar that meets a user-chosen coverage level with finite-sample, distribution-free marginal validity. Inference requires one forward pass with light post-processing. We evaluate on CAMUS and EchoNet-Dynamic across three seeds at nominal 0.9 conformal coverage. While matching Dice, WaCT-US achieves the best mean calibrated uncertainty scale (representing conformal efficiency and tube thickness) while holding comparable coverage and improving uncertainty-guided error localization. This information can estimate how wrong the underlying segmentation model is on each input image. The code will be publicly available at https://github.com/MohammadMahdiKazemi/WaCT-US.

Keywords: Ultrasound · Echocardiography · Uncertainty Quantification · Conformal Calibration · Contours

1 Introduction

Echocardiography segmentation underpins routine cardiac quantification, yet ultrasound artifacts (speckle, shadowing, foreshortening) and scanner/vendor variation can yield plausible-looking contours with clinically meaningful errors. Because downstream measurements are geometric and highly sensitive to boundary perturbations, reliable uncertainty estimates are needed for quality control, deferral, and safe large-scale deployment. Most segmentation uncertainty

* Corresponding author.

quantification (UQ) methods are either *pixel-centric* (e.g., entropy- or evidence-based maps, calibration layers) or *sampling-heavy* (e.g., MC dropout, Bayesian deep learning, deep ensembles, Laplace/SWAG) [3, 7, 8, 2, 10, 16, 4]. While useful, these approaches do not directly match echocardiography practice: clinicians and downstream pipelines typically consume contours, landmarks, and derived metrics rather than per-pixel probabilities, and repeated test-time sampling is costly for real-time and video-scale settings. Echo-specific contour UQ has therefore been explored in multi-stage pipelines and metric-propagation studies [5, 6], but they generally do not provide an explicit, *calibrated set* of plausible contours with user-controlled coverage. Conformal prediction offers a complementary interface: it constructs prediction sets with finite-sample marginal validity under exchangeability [14, 17]. Recent work has adapted conformal ideas to segmentation and structured outputs [11, 12, 18], yet two challenges remain for contour-based echocardiography: (i) mask-oriented constructions do not naturally respect closed-curve geometry and indexing ambiguity, and (ii) a single global conformal inflation can be overly conservative without modeling *local* (heteroscedastic) boundary ambiguity. We address these gaps with **WaCT-US** (Wrap-aligned Contour Tubes for ultrasound), introducing *cyclic-alignment scaled contour tubes*: contour-space prediction sets centered on a single predicted boundary with a learned, spatially varying thickness profile and one globally conformalized scale. Cyclic landmark alignment resolves the start-index ambiguity of closed contours, enabling stable training, calibration, and evaluation in landmark space. The resulting uncertainty object is contour-native, compute-aware (single forward pass with lightweight post-processing), and provides an explicit coverage–efficiency trade-off. We evaluate on CAMUS [9] and EchoNet-Dynamic [13], two publically available echo datasets, showing improved coverage–efficiency and stronger uncertainty utility for error localization relative to post-hoc baselines.

Contributions. Our main contributions are:

- **Contour-native prediction sets:** we formulate *contour tubes* as set-valued outputs around predicted landmarks, with a learned heteroscedastic scale profile.
- **Cyclic alignment for closed curves:** we introduce cyclic landmark alignment to resolve contour indexing ambiguity and define stable contour-space discrepancies.
- **Single-scalar conformal calibration:** we calibrate one global scale via split conformal prediction, yielding finite-sample marginal coverage for the chosen tube family with single-pass inference.

Section 2 formulates WaCT-US and the contour-tube construction, followed by theoretical guarantees and complexity analysis, and Sections 3 and 4 report experiments on CAMUS and EchoNet-Dynamic.

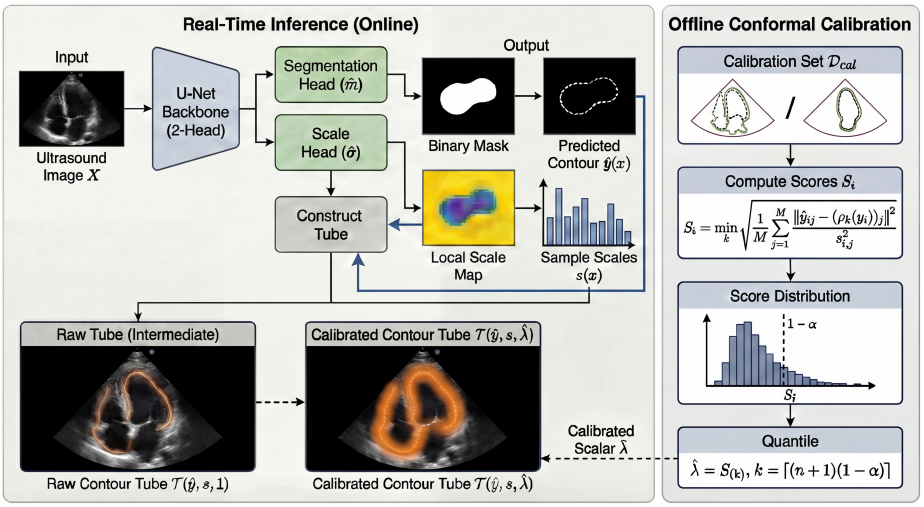


Fig. 1. WaCT-US overview. A single forward pass predicts a segmentation and local contour-scale profile. Split-conformal calibration on a held-out split yields one global scalar $\hat{\lambda}$, and test-time output is the calibrated contour tube $\mathcal{T}(\hat{y}, s, \hat{\lambda})$.

2 Method: Cyclic-Alignment Scaled Contour Tubes

2.1 Setup and Goal

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $X : \Omega \to \mathcal{X}$ an ultrasound frame (including echo), and $Y : \Omega \to \mathcal{Y}$ its ground-truth contour. We represent a contour as M ordered landmarks $y = (y_1, \ldots, y_M) \in \mathcal{Y} := (\mathbb{R}^2)^M$. We seek a set-valued predictor $C : \mathcal{X} \to 2^{\mathcal{Y}}$ such that, for a chosen $\alpha \in (0, 1)$,

$$\mathbb{P}\big(Y \in C(X)\big) \geq 1 - \alpha. \tag{1}$$

2.2 Cyclic-Alignment Discrepancy and Tube

To handle the unknown starting index of landmarks, let ρ_k denote a cyclic shift: $(\rho_k(y))_j := y_{((j+k-1) \bmod M)+1}$ for $k \in \{0, \ldots, M-1\}$. For $k \in \{0, \ldots, M-1\}$ define the shift-aligned discrepancy

$$D_k(y, \tilde{y}) := \left(\frac{1}{M} \sum_{j=1}^{M} \|y_j - (\rho_k(\tilde{y}))_j\|_2^2 \right)^{1/2}. \tag{2}$$

We take the best cyclic alignment, $W(y, \tilde{y}) := \min_k D_k(y, \tilde{y})$. To model heteroscedastic contour ambiguity, we use a positive scale profile $s \in (0, \infty)^M$ and define the scaled discrepancy

$$d_s(y, \tilde{y}) := \min_k \left(\frac{1}{M} \sum_{j=1}^{M} \frac{\|y_j - (\rho_k(\tilde{y}))_j\|_2^2}{s_j^2} \right)^{1/2}. \tag{3}$$

Given a predicted contour $\hat{y}$, scale profile s, and scalar $\lambda \geq 0$, the scaled contour tube is $\mathcal{T}(\hat{y}, s, \lambda) := \{y \in \mathcal{Y} : d_s(\hat{y}, y) \leq \lambda\}$. We interpret s as a local heteroscedastic ambiguity profile and λ as a global calibration factor. Because (3) is an RMS constraint, λs_j is a *local scale proxy* (used for visualization/relative ranking), not an independent per-point feasible radius.

Hence, WaCT-US decomposes uncertainty into ***where*** the contour is ambiguous (shape of s) and ***how much*** global inflation is needed for calibration (λ).

2.3 Architecture, Training, Inference

We use a single-pass backbone (e.g., U-Net [15]) with two heads: (i) segmentation probability map $\hat{m}(x)$ and (ii) positive boundary scale map $\hat{\sigma}(x)$ (Note that WaCT-US is architecture agnostic and can be applied to any segmentation backbone). We extract M landmarks $\hat{y}(x)$ from the 0.5-level set of $\hat{m}(x)$ as follows: threshold/level-set extraction, keep the anatomically central largest component, resample by arc length to M points, and enforce a canonical contour orientation for deterministic preprocessing. This orientation step fixes clockwise/counterclockwise direction and a stable traversal convention, while cyclic alignment in Eqs. (2)–(3) still absorbs residual start-index ambiguity during training, calibration, and evaluation. Scales $s_j(x)$ are sampled from $\hat{\sigma}(x)$ at landmark locations.

For training, we use Dice + binary cross-entropy (BCE) for segmentation and an alignment-aware pinball loss for scales. Let $k^\star \in \arg\min_k D_k(\hat{y}, y)$ and residuals $r_j := \|\hat{y}_j - (\rho_{k^\star}(y))_j\|_2$. With pinball parameter $\tau \in (0,1)$, letting $(.)+ \coloneqq \max(., 0)$ denote the positive-part operator, the pinball loss is $\ell_\tau(u) = \tau u_+ + (1-\tau)(-u)_+$,

$$\mathcal{L}_{\text{scale}} := \frac{1}{M} \sum_{j=1}^{M} \ell_\tau(r_j - s_j). \tag{4}$$

Calibration and guarantees. Given a calibration set $\{(x_i, y_i)\}_{i=1}^n$, define scores $S_i := d_{s(x_i)}(\hat{y}(x_i), y_i)$ and let $k := \lceil (n+1)(1-\alpha) \rceil$. We set $\hat{\lambda}$ to the k-th smallest of $(S_1, \ldots, S_n)$ and output $\mathcal{T}(\hat{y}(x), s(x), \hat{\lambda})$. Let P_{tr} and P_{te} denote the train and test joint distributions on (X, Y); for any event B, $P_{\text{tr}}(B)$ and $P_{\text{te}}(B)$ denote event probabilities under these distributions. Let $(X_{n+1}, Y_{n+1}) \sim P_{\text{tr}}$ be a fresh in-distribution pair and define its score $S_{n+1} := d_{s(X_{n+1})}(\hat{y}(X_{n+1}), Y_{n+1})$.

Theorem 1 (Finite-sample coverage and TV-shift diagnostic bound). *Assume $\alpha \in (0,1)$ with $\alpha \geq 1/(n+1)$ (equivalently, $k \leq n$), assume exchangeability of $(S_1, \ldots, S_n, S_{n+1})$, and assume no ties almost surely. Then split conformal with $k = \lceil (n+1)(1-\alpha) \rceil$ satisfies*

$$P_{\text{tr}}\big(d_{s(X)}(\hat{y}(X), Y) \leq \hat{\lambda}\big) \geq 1 - \alpha.$$

If additionally the total variation distance $\text{TV}(P_{\text{tr}}, P_{\text{te}}) \leq \delta$, then

$$P_{\text{te}}\big(d_{s(X)}(\hat{y}(X), Y) \leq \hat{\lambda}\big) \geq 1 - \alpha - \delta.$$

Proof sketch. The first statement is the standard split-conformal order-statistic argument under exchangeability. The second follows by applying $|P_{\mathrm{te}}(B) - P_{\mathrm{tr}}(B)| \leq \mathrm{TV}(P_{\mathrm{tr}}, P_{\mathrm{te}})$ to $B = \{d_{s(X)}(\hat{y}(X), Y) \leq \hat{\lambda}\}$. Since $\mathbb{P}(d_{s(X)}(\hat{y}(X), Y) \leq \lambda)$ is monotone in λ, the tightest population tube corresponds to the $(1-\alpha)$-quantile $\lambda^\star$ of $S := d_{s(X)}(\hat{y}(X), Y)$; split conformal estimates it via the order statistic $\hat{\lambda}$. Operationally, this means the method is calibrated for the in-distribution population represented by the calibration split; under shift, calibration may degrade according to the TV term in Theorem 1. Robust weighted conformal alternatives under distribution drift are discussed in [1]. Conformal guarantees are marginal/population statements over calibration-split randomness; finite test sets can therefore over- or under-shoot the nominal level. In implementation, we use the standard conservative event check $S \leq \hat{\lambda}$; this tie-safe rule preserves finite-sample validity under discrete/tied scores. Accordingly, we interpret empirical coverage together with exact binomial confidence intervals in the experiments section. Brute-force calibration costs $O(nM^2 + n \log n)$ (min over M shifts per score + sorting) and inference adds $O(M)$ arithmetic beyond a single forward pass; with $M = 64$, this post-processing is negligible relative to CNN inference in practice.

3 Experiments

3.1 Datasets and Splits

We evaluate WaCT-US on two public echocardiography *segmentation* datasets: CAMUS [9] (2D echocardiography; 2-chamber and 4-chamber (2CH/4CH) views), and EchoNet-Dynamic [13] (echo Left Ventricle (LV) tracing subset). Unless stated otherwise, we resize each image to 256×256 and normalize intensities per image. We use train/val/test splits. Under a strict full-validation policy, for every reported method we use the whole validation split for checkpoint selection and conformal calibration of $\hat{\lambda}$, and we use the whole test split for final evaluation and reporting the results. For CAMUS, splits are patient-disjoint to avoid leakage across views and phases. CAMUS uses a 60:20:20 patient-disjoint split. For EchoNet, we use the official video-level benchmark split (train/val/test). For segmentation and contour-tube experiments, we report mean±std over three random seeds ($\{0, 1, 2\}$). Our primary evaluation unit is frame-level coverage (Cov(frame)). Both datasets are publicly released for research use under their original data-use terms; we use only de-identified public data and introduce no new patient recruitment or intervention.

3.2 Model, Training, and Calibration

We use a two-head U-Net backbone (segmentation logits + positive scale map) with a single forward pass at test time. Training minimizes Dice + binary cross-entropy (BCE) for segmentation and an alignment-aware pinball loss (Eq. 4) for the scale head. Unless otherwise specified, we train for 100 epochs with Adam

(10^{-3} learning rate, batch size 8), $M = 64$ landmarks, pinball parameter $\tau = 0.9$, and scale-loss weight 0.01. We select the checkpoint with the best validation Dice on the full validation split. For calibration, we compute nonconformity scores $S_i = d_{s(x_i)}(\hat{y}(x_i), y_i)$ on that same full validation split and set $\hat{\lambda}$ to the conformal quantile. We report results at nominal $1 - \alpha = 0.9$ coverage (i.e., $\alpha = 0.1$). All reported coverages in the main tables are frame/image level (Cov(frame)). In Table 1, we define

$$\text{Scale} := \frac{1}{n} \sum_{i \in \mathcal{V}} \hat{\lambda} \left(\frac{1}{M} \sum_{j=1}^{M} s_{ij} \right),$$

where $\mathcal{V} \subseteq \{1, \ldots, n_{\text{test}}\}$ indexes test frames and $n = |\mathcal{V}|$. We compute Cov(frame) conservatively over all n_{test} test frames. Note that, Scale is reported in mm-equivalent units for CAMUS (from spacing metadata) and in resized-grid pixel units for EchoNet (absolute spacing metadata unavailable). For the main CAMUS/EchoNet comparisons, this full-validation checkpoint-selection + calibration policy and whole-test evaluation are applied identically to WaCT-US and all reported baselines.

3.3 Baselines and Shift Tests

For segmentation, we compare post-hoc uncertainty baselines: deterministic entropy of p, evidential deep learning, temperature scaling [4], MC dropout [3] (T=10), last-layer Laplace [2], and SWAG [10]. For each baseline, we build a calibrated *contour tube* with the same full-validation policy (checkpoint selection and calibration on the whole validation split; evaluation on the whole test split). Entropy and temperature scaling use contour-sampled pixel entropy as the local scale map. Evidential deep learning uses contour-sampled Dirichlet vacuity $2/(\alpha_+ + \alpha_-)$. MC dropout, Laplace, and SWAG use contour-sampled predictive standard deviation. Our Laplace/SWAG implementations use the publically available implementations. Main-text comparisons report both single-pass (WaCT-US, entropy, evidential, temperature scaling) and multi-pass (MC dropout, Laplace, SWAG) baselines.

4 Results

4.1 Calibrated Tubes and Metrics

Table 1 reports Dice/Hausdorff and contour-tube metrics (coverage, mean scale proxy) on CAMUS and EchoNet. We separate interpretation into segmentation quality (Dice/Hausdorff), set validity (Cov(frame)), and set efficiency (Scale). At matched achieved coverage, smaller scale indicates sharper uncertainty sets.

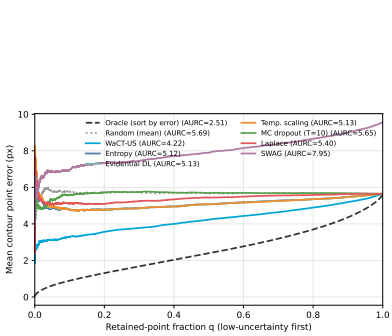
Segmentation quality. Table 1 shows that WaCT-US remains competitive on Dice/Hausdorff (since it doesn't alter the main network architecture) against the included post-hoc alternatives. In this table, Haus refers to *max* Hausdorff distance; we also log HD95 in our experiment logs.

Table 1. Compute-aware segmentation and calibrated contour-set metrics ($\alpha = 0.1$, $M = 64$; test metrics at best validation Dice checkpoint; mean±std across seeds). Method compute is shown as ($\times n$), where n is the number of stochastic samples/forward passes per test frame (WaCT-US: $\times 1$). (CAMUS: mm unit from metadata-aware spacing; EchoNet: resized-grid pixel unit).

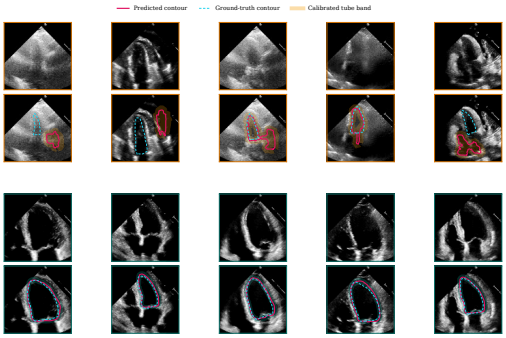
Dataset	Method	Dice↑	Haus↓	Cov(frame)↑	Cov 95% CI	Scale↓
CAMUS	**WaCT-US** (×1)	0.910±0.003	12.20±1.52	0.903±0.015	[0.881, 0.917]	8.09±0.41
	Entropy (×1)	0.910±0.003	12.20±1.52	0.901±0.016	[0.880, 0.915]	9.30±0.44
	Evidential DL (×1)	0.910±0.003	12.20±1.52	0.912±0.007	[0.886, 0.919]	9.00±0.10
	MC dropout (×10)	0.910±0.003	12.20±1.52	0.905±0.009	[0.882, 0.918]	8.98±0.11
	Temp. scaling (×1)	0.910±0.003	12.20±1.52	0.904±0.008	[0.884, 0.920]	9.02±0.10
	Laplace (×20)	0.910±0.003	12.20±1.51	0.903±0.008	[0.880, 0.917]	9.95±0.95
	SWAG (×10)	0.906±0.005	13.05±0.74	0.900±0.026	[0.888, 0.923]	10.48±0.73
EchoNet	**WaCT-US** (×1)	0.901±0.004	8.53±0.67	0.914±0.010	[0.897, 0.929]	5.10±0.36
	Entropy (×1)	0.901±0.004	8.53±0.67	0.916±0.006	[0.899, 0.931]	5.55±0.33
	Evidential DL (×1)	0.901±0.004	8.53±0.67	0.917±0.019	[0.900, 0.932]	5.27±0.32
	MC dropout (×10)	0.901±0.004	8.53±0.67	0.916±0.013	[0.899, 0.931]	5.47±0.26
	Temp. scaling (×1)	0.901±0.004	8.53±0.67	0.917±0.020	[0.900, 0.932]	5.28±0.34
	Laplace (×20)	0.901±0.004	8.51±0.65	0.915±0.005	[0.898, 0.930]	5.98±0.37
	SWAG (×10)	0.903±0.003	8.65±0.72	0.913±0.010	[0.896, 0.929]	8.07±0.76

Set validity and efficiency. Coverage is reported with exact Clopper–Pearson 95% confidence intervals (pooled across seeds) to reflect finite-sample variability around the nominal conformal target. On EchoNet, all included methods remain close to the nominal 0.9 target (Cov(frame) range, with no marked undercoverage under the strict full-validation policy. Scale is reported in resized-grid pixel units on EchoNet and mm units on CAMUS and should be interpreted as a calibrated scale proxy (not an independent per-point radius under Eq. 3). On both datasets, WaCT-US reduces scale (uncertainty sharpens) relative to all the baselines, including the non-deterministc uncertainty baselines that need orders of magnitude more compute at inference time while holding better or comparable empirical conformal coverage.

Decision utility. Finally, Fig. 2 (left) quantifies usefulness for error triage: WaCT-US achieves the lowest risk–coverage area among the included baselines, meaning its uncertainty ranking best localizes larger contour errors. At sample level, the per-sample decile analysis shows consistent monotone alignment between predicted uncertainty and contour MAE, while the right subfigure provides qualitative EchoNet examples where higher uncertainty aligns with visibly poorer frame quality and larger contour disagreement. For an actionable quality-control policy, uncertainty-ranked deferral is effective: deferring the top 20% most uncertain CAMUS samples reduces retained-set contour MAE by 15.7% for WaCT-US in our single-seed triage analysis.



(a) CAMUS risk–coverage localization.



(b) EchoNet uncertainty extremes (top: most uncertain; bottom: least uncertain), with calibrated tube and contour overlays.

Fig. 2. Uncertainty utility across quantitative and qualitative views. (a) CAMUS usefulness of uncertainty for error localization: contour points are sorted by predicted uncertainty (low to high), and mean contour-point error is plotted versus retained-point fraction q (risk–coverage curve). Lower area indicates better concentration of high-error points into high-uncertainty regions; oracle and random orderings are shown for reference. (b) EchoNet-Dynamic qualitative ranking for WaCT-US using calibrated scale proxy $\hat{\lambda}\,\overline{s}$: the most uncertain frames (top row pair) exhibit lower visual quality and larger contour mismatch, while the least uncertain frames (bottom row pair) show cleaner boundaries and tighter agreement.

5 Conclusion

WaCT-US provides contour-native uncertainty via cyclic-alignment scaled contour tubes with explicit local ambiguity profiles and global conformal calibration. It provides finite-sample calibration via split conformal prediction with real-time complexity. Practically, this gives clinicians and developers a calibrated, contour-level reliability object: maintain segmentation quality while obtaining sharper uncertainty sets and better localization of high-error boundary regions. Our current discrepancy uses Euclidean landmark residuals; normal-direction discrepancies are a relevant extension for future clinically weighted contour errors. While we focus on segmentation, the same contour-tube formalism applies to other structured geometric outputs with observer variability (e.g., landmark/diameter annotations), where local scale can be calibrated against repeated labels and validated across internal/external cohorts (e.g., PLAX LVID).

Acknowledgments. This work was funded, in part, by the Canadian Institutes of Health Research (CIHR) and the Natural Sciences and Engineering Research Council of Canada (NSERC), the Vector Institute for AI, Canada CIFAR AI Chairs, and a Google Gift Fund. Resources used in preparing this research were provided, in part, by the Province of Ontario, the Government of Canada through the Digital Research Alliance of Canada www.alliancecan.ca, and companies sponsoring the Vector Institute www.vectorinstitute.ai/#partners, and Advanced Research Computing at the University of

British Columbia. Additional hardware support was provided by John R. Evans Leaders Fund CFI grant. Mohammad Mahdi Kazemi Esfeh is also supported by UBC Four Year Doctoral Fellowships and Canada Graduate Scholarship-Doctoral (CGS-D).

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

References

1. Barber, R.F., Candès, E.J., Ramdas, A., Tibshirani, R.J.: Conformal prediction beyond exchangeability. The Annals of Statistics **51**(2), 816–845 (2023), https://doi.org/10.1214/23-AOS2276
2. Daxberger, E., Kristiadi, A., Immer, A., Eschenhagen, R., Bauer, M., Hennig, P.: Laplace redux - effortless bayesian deep learning. In: Advances in Neural Information Processing Systems. vol. 34, pp. 20089–20103 (2021)
3. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: Balcan, M.F., Weinberger, K.Q. (eds.) Proceedings of The 33rd International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 48, pp. 1050–1059. PMLR, New York, New York, USA (20–22 Jun 2016), https://proceedings.mlr.press/v48/gal16.html
4. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: Precup, D., Teh, Y.W. (eds.) Proceedings of the 34th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 70, pp. 1321–1330. PMLR (2017), https://proceedings.mlr.press/v70/guo17a.html
5. Judge, T., Bernard, O., Cho Kim, W.J., Gomez, A., Beqiri, A., Chartsias, A., Jodoin, P.M.: Uncertainty propagation for echocardiography clinical metric estimation via contour sampling (2025), https://arxiv.org/abs/2502.12713
6. Judge, T., Bernard, O., Cho Kim, W.J., Gomez, A., Chartsias, A., Jodoin, P.M.: Asymmetric contour uncertainty estimation for medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2023. Lecture Notes in Computer Science, vol. 14222, pp. 210–220. Springer, Cham (2023), https://doi.org/10.1007/978-3-031-43898-1_21
7. Kazemi Esfeh, M.M., Luong, C., Behnami, D., Tsang, T., Abolmaesumi, P.: A deep bayesian video analysis framework: Towards a more robust estimation of ejection fraction. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. Lecture Notes in Computer Science, vol. 12262, pp. 582–590. Springer, Cham (2020), https://doi.org/10.1007/978-3-030-59713-9_56
8. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. In: Advances in Neural Information Processing Systems. vol. 30, pp. 6402–6413 (2017)
9. Leclerc, S., Smistad, E., Pedrosa, J., Ostvik, A., Cervenansky, F., Espinosa, F., Espeland, T., Berg, E.A.R., Jodoin, P.M., Grenier, T., Lartizien, C., Dhooge, J., Lovstakken, L., Bernard, O.: Deep learning for segmentation using an open large-scale dataset in 2d echocardiography. IEEE Transactions on Medical Imaging **38**(9), 2198–2210 (2019), https://doi.org/10.1109/TMI.2019.2900516
10. Maddox, W.J., Izmailov, P., Garipov, T., Vetrov, D.P., Wilson, A.G.: A simple baseline for bayesian uncertainty in deep learning. In: Advances in Neural Information Processing Systems. vol. 32, pp. 13153–13164 (2019)

11. Mossina, L., Dalmau, J., Andéol, L.: Conformal semantic image segmentation: Post-hoc quantification of predictive uncertainty. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 3574–3584. IEEE (Jun 2024), https://doi.org/10.1109/CVPRW63382.2024.00361
12. Mossina, L., Friedrich, C.: Conformal prediction for image segmentation using morphological prediction sets. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2025. Lecture Notes in Computer Science, vol. 15963, pp. 78–88. Springer, Cham (2026), https://doi.org/10.1007/978-3-032-04965-0_8
13. Ouyang, D., He, B., Ghorbani, A., Yuan, N., Ebinger, J., Langlotz, C.P., Heidenreich, P.A., Harrington, R.A., Liang, D.H., Ashley, E.A., Zou, J.Y.: Video-based ai for beat-to-beat assessment of cardiac function. Nature **580**(7802), 252–256 (Mar 2020), https://doi.org/10.1038/s41586-020-2145-8
14. Romano, Y., Patterson, E., Candes, E.: Conformalized quantile regression. In: Advances in Neural Information Processing Systems. vol. 32, pp. 3543–3553. Curran Associates, Inc. (2019)
15. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Lecture Notes in Computer Science, vol. 9351, pp. 234–241. Springer (2015), https://doi.org/10.1007/978-3-319-24574-4_28
16. Sensoy, M., Kaplan, L., Kandemir, M.: Evidential deep learning to quantify classification uncertainty. In: Advances in Neural Information Processing Systems. vol. 31, pp. 3179–3189 (2018)
17. Vovk, V., Gammerman, A., Shafer, G.: Algorithmic Learning in a Random World. Springer, New York, NY, USA (2005), https://doi.org/10.1007/b106715
18. Zhang, W.: Distribution-free uncertainty quantification for contour objects with application to tumour segmentation in pet imaging. In: Proceedings of the Thirteenth Symposium on Conformal and Probabilistic Prediction with Applications. Proceedings of Machine Learning Research, vol. 230, pp. 536–537. PMLR (2024), https://proceedings.mlr.press/v230/zhang24a.html