Modeling Aleatoric Uncertainty in Cardiac MRI Segmentation: Probabilistic Detection and Contour Regression

Yidong Zhao, Yi Zhang, João Tourais, Sebastian Weingärtner, Avan Suinesiaputra, Alistair Young, Yuchi Han, Orlando Simonetti, and Qian Tao

***Abstract*— Accurate segmentation of cardiac MRI is essential for assessment of cardiac function through biomarkers such as the left and right ventricular ejection fraction (LVEF, RVEF). Although AI methods have achieved high average segmentation accuracy, the precision of biomarkers for individual patients–quantified by estimation variance, remains critical for reliable diagnosis. Calibrated biomarkers, whose uncertainty accurately reflects the true variability, are highly desirable. However, existing evaluations predominantly focus on population-level segmentation accuracy, leaving biomarker-level uncertainty and calibration largely underexplored. Intrinsic anatomical ambiguity and annotation variability are major sources of biomarker variability and cannot be fully eliminated, even when training on a single annotation set. To address this, we propose a probabilistic segmentation framework that explicitly models aleatoric uncertainty with the goal of improving calibration in the biomarker space. The framework disentangles two key sources of uncertainty: 1) *detection uncertainty*, arising from ambiguous inclusion of basal or apical slices in 2D cardiac MRI, modeled via objectness probabilities; and 2) *contour uncertainty*, reflecting variability in ventricular boundary delineation, modeled through mean–variance regression of elliptic Fourier descriptors, a compact representation of closed contours. By propagating these uncertainties to derived biomarkers, the proposed method produces more informative and better-calibrated confidence estimates for ejection fraction. Compared to conventional pixel-wise approaches, our framework improves biomarker reliability, particularly in realistic settings dominated by annotation ambiguity and limited domain shift.**

Received 2 February 2026; accepted 5 June 2026. Date of publication 12 June 2026; date of current version 4 August 2026. This work was supported in part by the Delft AI Initiative at Delft University of Technology, in part by the Dutch Research Council (NWO) under Grant NGF.1609.241.028, in part by the European Research Council under Grant 101078711, in part by the Dutch Heart Foundation under Dekker Beurs Grant 03-004-2022-0079, and in part by the NVIDIA Academic Grant Program. Recommended by Associate Editor X. Zhuang. *(Corresponding author: Qian Tao.)*

Yidong Zhao, Yi Zhang, João Tourais, Sebastian Weingärtner, and Qian Tao are with the Department of Imaging Physics, Delft University of Technology, 2628 CN Delft, The Netherlands (e-mail: y.zhao-8@tudelft.nl; y.zhang-43@tudelft.nl; j.l.silvacanaveiratourais@tudelft.nl; s.weingartner@tudelft.nl; q.tao@tudelft.nl).

Avan Suinesiaputra and Alistair Young are with the Department of Biomedical Computing, King's College London, SE1 7EU London, U.K. (e-mail: avan.suinesiaputra@kcl.ac.uk; alistair.young@kcl.ac.uk).

Yuchi Han and Orlando Simonetti are with the Cardiovascular Division, The Ohio State University Wexner Medical Center, Columbus, OH 43210 USA (e-mail: Yuchi.Han@osumc.edu; orlando.simonetti@osumc.edu).

Digital Object Identifier 10.1109/TMI.2026.3702822

***Index Terms*— Uncertainty, segmentation, cardiac MRI.**

I. Introduction

SEGMENTATION is an important task in medical image analysis and has made tremendous progress over the past decade [1], [2], [3]. The self-configuring nnU-Net [4], [5] represents the state of the art in medical image segmentation, achieving outstanding segmentation accuracy measured by geometric metrics such as the Dice score and Hausdorff distance [2], [3]. However, in practice, clinicians are often more concerned with quantitative biomarkers derived from segmentation masks. In the case of cardiac magnetic resonance imaging (MRI), left and right ventricular ejection fractions (LVEF, RVEF) of cine MRI are essential biomarkers for the diagnosis of cardiovascular disease [2], [3]. Other examples include cell counts on histological images [6], tumor size quantification on neuro MRI [7], tissue volumes on prostate MRI [8], and many more. In this regard, medical image segmentation is merely an intermediate step in assessing biomarkers that are ultimately clinically used.

For quantitative biomarkers, both accuracy and precision are clinically relevant [9], [10], [11]. Accuracy describes the deviation from the true value, while precision is the measure of estimation uncertainty, expressed by, e.g., variance [9]. Due to inconsistent manual annotations and structure ambiguities, quantitative biomarkers can oftentimes only be determined up to some uncertainty, thus a limited precision. Knowing the precision of individual predictions is important for clinical diagnosis. For example, LVEF requires a precision of at least 5% for meaningful diagnosis purposes [12], since a difference of 5% may lead to different clinical decisions [13]. Calibrating the precision of the estimated biomarkers, that is, estimating the distribution of possible biomarker quantification, would be important in such cases. Unfortunately, the population-level high accuracy of pre-trained models in testing datasets does not equal high precision for individual cases [14], which is crucial for personalized diagnosis and treatment.

The precision of segmentation or biomarkers derived from segmentation is related to two types of uncertainty [15]. The first is *epistemic uncertainty* [15], [16], which increases when the test data deviates from the training distribution. Epistemic uncertainty can be reduced by extending the coverage of the training data. The second is *aleatoric uncertainty*, which originates from the data-generating process [2], [17], [18],

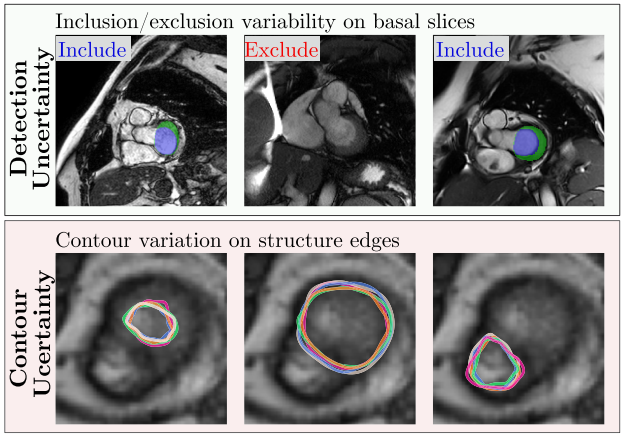

Fig. 1. Aleatoric uncertainty in cardiac MRI segmentation: (Upper–detection uncertainty) Expert annotations of three alike basal slices. The decision of inclusion/exclusion differs although the anatomical patterns are similar. The expert may include the slice as containing the left ventricle blood pool (purple) and myocardium (green), or exclude it (upper middle). (Below–contour uncertainty) Contours by different annotators, indicated by different colors. From left to right: LV endocardial contour, LV epicardial contour, RV endocardial contour.

[19], reflecting the intrinsic ambiguity associated with the data and the task. For example, the aleatoric uncertainty of segmentation is reflected by the inter- and intra-observer variability in manual annotations, even from experts. When image quality degrades or anatomical complexity increases, this uncertainty grows. Aleatoric uncertainty cannot be reduced by adding more training data [15], [20].

Even in deployment scenarios with limited domain shift where aleatoric uncertainty dominates, imperfect annotations lead to frequent discrepancies between network predictions and human annotations [2], [3]. Under such conditions, understanding the precision of cardiac biomarkers—namely, quantifying their plausible range of variation—becomes particularly important. In this work, we focus on calibrating the uncertainty of cardiac MRI biomarkers to enable reliable interpretation at the subject level. Although stochastic inference methods for segmentation [20], [21], [22], [23] have been widely studied, calibration of derived biomarkers remains relatively underexplored. Calibration in the segmentation space does not directly translate to calibration in the biomarker space, highlighting the non-trivial nature of uncertainty propagation.

In cardiac MRI, clinical biomarkers such as ventricular volume and ejection fraction are computed by aggregating segmentation areas across short-axis slices. In this process, two types of aleatoric uncertainty exist. The first is the *contour uncertainty*, which arises from the variable definition of the contours on the LV and RV borders, blurred by the partial volume effect (Figure 1). It is caused by the inter- or intra-observer variability of manually drawn contours [17], [18], [19]. The second is the *detection uncertainty*. This uncertainty arises when images contain complex 3D anatomical structures mapped onto the 2D image planes. For example, a short-axis cine at the base may or may not contain LV or RV depending on the cardiac phase, making it difficult to decide whether or not to include LV or RV in 2D analysis (Figure 1). Consequently, the detection of basal LV and RV exhibits a non-deterministic behavior, which is a widely recognized problem in cardiac MRI segmentation [2]. In practice, every center has its own convention for annotating the base part. Compared to the first type of contour uncertainty, detection uncertainty is more subtle and often overlooked but has a greater quantitative effect on the final biomarkers.

The aleatoric uncertainty in cardiac MRI segmentation can also be reflected in the manual annotation process: the observer first decides whether LV and RV exist in the given image and then delineates the contours around them. Variability in the first step represents the detection uncertainty, while variability in the second step represents contour uncertainty. Furthermore, annotation is done by following the contours around the LV and RV, instead of painting the pixels inside them (pixel classification) [24].

In this work, our primary goal is not to improve segmentation accuracy, but to enable reliable uncertainty quantification and calibration of segmentation-derived cardiac biomarkers, particularly ventricular volumes and ejection fraction. We treat segmentation as an intermediate representation and focus on modeling and propagating aleatoric uncertainty in a way that yields well-calibrated biomarker-level confidence estimates.

To this end, we propose a probabilistic segmentation framework that departs from conventional pixel-wise classification. Inspired by the manual annotation process, our method decomposes aleatoric uncertainty into (i) slice-level detection uncertainty, capturing ambiguous inclusion or exclusion of basal and apical slices, and (ii) contour uncertainty, modeling variability in ventricular boundary delineation. Detection uncertainty is modeled via probabilistic objectness prediction, while contour uncertainty is captured through joint mean–variance regression of elliptic Fourier descriptors, a compact representation of closed contours. These uncertainties are explicitly propagated to volume and ejection fraction estimates, enabling calibrated biomarker-level uncertainty quantification.

II. Related Work

Cardiac MRI segmentation has been extensively studied for automated assessment of cardiac function [2], [3], with modern deep learning frameworks such as nnU-Net [4], [5] and nnFormer [25] achieving strong performance across multiple benchmarks. On the ACDC dataset, state-of-the-art methods report mean absolute errors (MAE) of approximately 2.1% for LVEF and 5.4% for RVEF, indicating high accuracy at the population level. However, predictions frequently differ from human annotations on individual slices, particularly in anatomically ambiguous regions such as the basal and apical slices [2], [3]. As cardiac biomarkers are derived by aggregating segmentation results across short-axis slices, such subject-level discrepancies can propagate non-linearly and lead to variability in clinically relevant biomarkers like ejection fractions, despite excellent average performance.

Most existing segmentation uncertainty methods follow the pixel-space formulation by extending pixel-wise deterministic models to pixel-wise probabilistic models. A simple way is to temper the logits to prevent overconfident softmax outputs,

and the temperature can be either input-independent like temperature scaling [26] or input-dependent like LogitNorm [27]. Bayesian approaches are also widely used, essentially learning an ensemble of networks, either explicitly through Deep Ensembles [28], [29], Hamiltonian Monte Carlo [30], [31], or implicitly through MC-Dropout [15], [16]. These methods primarily capture epistemic uncertainty, which yields high uncertainty when the test data deviate from the training data [32]. For in-domain data, however, the predicted uncertainty tends to diminish [32].

To quantify aleatoric uncertainty, several methods [21], [22], [23], [33] have been proposed by introducing stochasticity into the network training and inference stage. The probabilistic U-Net [33] and Phi-Seg [22] methods utilize a variational latent space to generate the distribution of segmentation annotations. Stochastic Segmentation Networks (SSN) [23] predict logit map together with a low-rank approximation of the spatial correlation. Gao et al. propose a mixture of stochastic experts (MoSE) [21] to capture the aleatoric uncertainty from annotations. However, these methods rely on multiple interobserver annotations of the training data. When trained with annotations from a single observer, the uncertainty tends to degrade [22], [33]. Another method of quantifying aleatoric uncertainty is test time augmentation (TTA) [20], [34], treating the input as a sample of a generative distribution, which is heuristically simulated by random image augmentations. TTA can capture the uncertainty caused by border ambiguity [20], but cannot capture the uncertainty caused by anatomical ambiguity.

Even when only a single set of annotations is available—which represents the most common clinical setting—the intrinsic ambiguity of cardiac MRI segmentation implies that biomarker predictions can only be determined up to a finite precision, reflected by the frequent deviation from the human annotation [2], [3]. We argue that the current predominant pixel-space formulation is intrinsically insufficient for modeling aleatoric uncertainty: first, pixel-wise models are constrained by locality, collapsing uncertainty to thin boundaries and often generating anatomically implausible samples. Second, the detection uncertainty describes the binary behavior of a collection of pixels, while pixel-wise models have no straightforward mechanism to model such collective behavior.

Therefore, a new formulation of segmentation, which is amenable to aleatoric uncertainty modeling, is needed. In particular, there is a prominent line of work for object detection in computer vision, namely YOLO and its variants [35], [36]. YOLO first predicts an "objectness" score and subsequently predicts bounding boxes of the objects, specified by centers and rectangular dimensions. The global objectness transforms the pixel-wise segmentation to a slice-level classification problem. The ambiguous images such as basal slices with similar patterns but significantly different annotations of inclusion or exclusion are naturally pushed towards the global detection decision boundary [37]. As a result, the objectness prediction on these slices produce moderate probabilities, reflecting elevated uncertainty. For organ segmentation, rectangles are insufficient, but appropriate parameterization, such as elliptical Fourier descriptors (EFDs) [38], can accurately characterize arbitrary closed shapes [39], [40]. This inspires us to develop a novel detection-and-regression method for cardiac MRI segmentation, with built-in mechanisms for modeling aleatoric uncertainty.

TABLE I
SUMMARY OF MATHEMATICAL SYMBOLS

Symbol	Description
x, y	continuous spatial contour coordinates
t	contour parameter
h, w	feature map spatial indices
A	EFD cofficient matrix
k	EFD order
$\boldsymbol{a}$	flattened EFD cofficients
$\boldsymbol{p}, \boldsymbol{\mu}, \boldsymbol{\sigma}$	Dense objectness and EFD mean and variance map
$p, \mu_{\hat{\boldsymbol{a}}}, \sigma_{\hat{\boldsymbol{a}}}$	Aggregated objectness, EFD mean and EFD variance
v	2D slice volume
V	3D volume aggregated across slices
μ_c, σ_c	Contour induced 2D volume mean and standard deviation
EF	Ejection fraction
m	model index in ensemble
α	nominal confidence level
z_α	standard normal quantile

III. CONTRIBUTIONS

We make the following contributions:

- We identify and disentangle two distinct sources of aleatoric uncertainty: detection uncertainty and contour uncertainty. Moreover, we demonstrate that substantial biomarker variability persists even with single-annotation supervision, which pixel-wise uncertainty fails to reflect at the biomarker level.
- We propose a novel probabilistic detection-and-regression framework for cardiac MRI segmentation, which mimics the expert annotation process and accurately models aleatoric uncertainty. Our framework deviates from the conventional pixel-wise formulation and explicitly decouples detection uncertainty and contour uncertainty. It enables well-calibrated quantification of precision for segmentation-derived function biomarkers, including stroke volume, LVEF, and RVEF by providing confidence interval.
- We studied both aleatoric and epistemic uncertainty on datasets with different degrees of domain shifts. We show that while epistemic uncertainty grows with domain shift, aleatoric uncertainty dominates when there is little to no domain shift. In the latter cases, the precision of biomarkers is severely underestimated by current methods that focus on epistemic uncertainty. This underscores the significance of aleatoric uncertainty modeling, because limited domain shift represents the majority of deployment scenarios in clinical practice.

IV. METHOD

The proposed method models aleatoric uncertainty through slice-level detection and probabilistic contour regression. In particular, the network predicts compact EFD representations of contours (Sec. IV-A). Detection uncertainty captures ambiguous slice inclusion via objectness prediction, while contour uncertainty is modeled by joint mean–variance regression

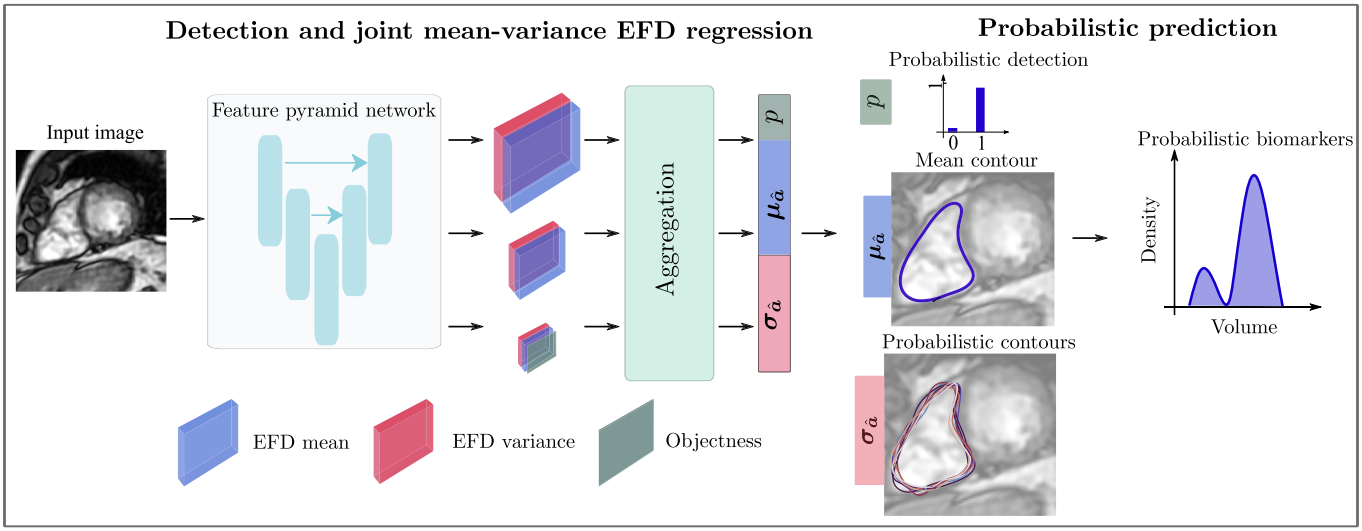

Fig. 2. Probabilistic contour regression network pipeline. The feature pyramid network extracts features at multiple resolution levels. The DC and AC components of low-, mid-, and high-frequency are predicated at different resolution levels, respectively. The network jointly estimates the mean and variance of elliptical Fourier descriptor (EFD) coefficients, which model the probabilistic contours.

of elliptic Fourier descriptors (Sec. IV-B). Both uncertainty sources are propagated to the biomarker space to obtain uncertainty estimates for volume and ejection fraction (Sec. IV-C). We provide a concise summary of notations in Table. I.

A. Contours as EFDs

1) EFD Representation: A 2D continuous contour $\mathcal{C}$ can be represented by a set of N_c points: $\mathcal{C} = \{(x_j, y_j) \mid j = 1, 2, \cdots, N_c\}$. Direct regression of these points is problematic due to variations in the number and order of the points N_c. A better way to represent closed contours with an arbitrary number of points is EFD [38], which decomposes contours into a superposition of ellipses rotating at various frequencies. The EFD of a contour $\mathcal{C} = \{(x(t), y(t)) \mid t \in [0, 1)\}$ can be formulated as:

$$\begin{bmatrix} x(t) \\ y(t) \end{bmatrix} = A_0 + \sum_{k=1}^{K} A_k \cdot \rho_k(t), \tag{1}$$

$$A_0 := \begin{bmatrix} a_0 \\ c_0 \end{bmatrix}, A_k := \begin{bmatrix} a_k & b_k \\ c_k & d_k \end{bmatrix}, \rho_k(t) := \begin{bmatrix} \cos(2k\pi t) \\ \sin(2k\pi t) \end{bmatrix} \tag{2}$$

where the DC component A_0 determines the centroid, the EFD coefficient matrix A_k characterizes the ellipse of k-th order, and $\rho_k(t)$ defines a rotating unit circle at an angular frequency of $2k\pi$. EFD translates segmentation from a pixel classification task to a regression task. The contours do not need further standardization or alignment as in other contour representation methods, such as active shape models (ASM) [41].

2) Distance in the EFD Space: We note that the EFD of a contour can only be determined up to an unknown initial phase ϕ. This will cause a metric problem: the same contours with different initial phases can have completely different EFDs and high Euclidean distances, which fail to reflect their true difference. Training the network for EFD regression requires the distance metric to reflect the true distance between contours, invariant to the initial phase. Therefore, given the EFD of the ground-truth contour $\boldsymbol{a} = [A_1, A_2, \cdots, A_K]$ and a predicted EFD $\hat{\boldsymbol{a}} = [\hat{A}_1, \hat{A}_2, \cdots, \hat{A}_K]$, we need to measure their distance with synchronised initial phase:

$$d(\boldsymbol{a}_k, \hat{\boldsymbol{a}}_k) = \left\| \hat{A}_k - A_k R_{k\phi^*} \right\|_F^2, \tag{3}$$

$$\phi^* = \arg\min_{\phi} \left\| \hat{A}_1 - A_1 R_\phi \right\|_F^2, \tag{4}$$

where $\boldsymbol{a}_k$ and $\hat{\boldsymbol{a}}_k$ denote the flattened EFD coefficients A_k and $\hat{A}_k$, R_ϕ is the rotation matrix of angle ϕ, $A_1 = \begin{bmatrix} a_1 & b_1 \\ c_1 & d_1 \end{bmatrix}$ and $\hat{A}_1 = \begin{bmatrix} \hat{a}_1 & \hat{b}_1 \\ \hat{c}_1 & \hat{d}_1 \end{bmatrix}$. By setting $\frac{\partial \|\hat{A}_1 - A_1 R_\phi\|_F^2}{\partial \phi} = 0$, we derive the closed-form solution to the optimization problem (4), with the optimal ϕ^*:

$$\phi^* = \tan^{-1}\left[\frac{\hat{a}_1 b_1 + \hat{c}_1 d_1 - \hat{b}_1 a_1 - \hat{d}_1 c_1}{\hat{a}_1 a_1 + \hat{c}_1 c_1 + \hat{b}_1 b_1 + \hat{d}_1 d_1} \right], \tag{5}$$

B. Detection and Probabilistic Contour Regression

We design the network to predict the EFD components at various resolution levels. The DC and low-frequency components in EFD are more related to the contour shapes and therefore depend on global image features. The high-frequency components reflect contour details and can be extracted from local image information. In total, we extract $K = 24$ harmonic components. The DC and low frequency (order $1 - 8$) are predicted from the $\frac{1}{4}$-resolution feature map, together with the objectness map. The midfrequency (order $9 - 16$) and highfrequency components (order $17 - 24$) are predicted with the features of half-resolution and full-resolution, respectively, as shown in Fig. 2.

The compact parametrization of contours by EFD enables the modeling of probabilistic predictions by inferring parameter distributions [39], [40]. We consider the predicted EFD coefficients $\hat{\boldsymbol{\mu}}$ as the mean values, shown as blue blocks in Figure 2. In addition to mean values, their predictive uncertainty is of interest. Therefore, we estimate their variances $\hat{\boldsymbol{\sigma}}^2$ besides the mean, depicted as the red blocks. Moreover, we

predict objectness $\boldsymbol{p}$ to model the detection probability, shown as green blocks in Figure 2.

To obtain a single contour prediction, we aggregated the predicted objectness map $\boldsymbol{p}$ and the mean variance predictions $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\sigma}}^2$. The following aggregation rule is designed as follows:

$$p = \max_{h,w} \boldsymbol{p}(h,w), \tag{6}$$

$$\mu_{\hat{\boldsymbol{a}}} = \frac{1}{\|\boldsymbol{p}\|_1} \sum_{h,w} \boldsymbol{p}(h,w) \cdot \hat{\boldsymbol{\mu}}(h,w), \tag{7}$$

$$\sigma_{\hat{\boldsymbol{a}}} = \frac{1}{\|\boldsymbol{p}\|_1} \sum_{h,w} \boldsymbol{p}(h,w) \left[\hat{\boldsymbol{\sigma}}(h,w)^2 + \hat{\boldsymbol{\mu}}(h,w)^2\right] - \mu_{\hat{\boldsymbol{a}}}^2. \tag{8}$$

where $(h,w) \in R^H \times R^W$ denotes the pixel location. Each pixel has a corresponding mean and variance prediction $\hat{\boldsymbol{\mu}}(h,w)$ and $\hat{\boldsymbol{\sigma}}(h,w)$. We treat the mean and variance predicted on the 2D grid as a mixture of $H \times W$ Gaussians weighted by objectness $\boldsymbol{p}(h,w)$. An individual detection head is used after feature extraction for each tissue type. Segmentation is generated by first thresholding the objectness scores p and suppressing the predictions with a lower objectness. Afterward, we transform the predicted EFD mean $\mu_{\hat{\boldsymbol{a}}}$ into the contour space for the predictions with a higher objectness and derive contour variations from the predicted variance $\sigma_{\hat{\boldsymbol{a}}}$. An overview of the network is shown in Figure 2.

1) Loss Function: We train the network with a loss function that consists of three parts. The first part $\mathcal{L}_{\text{obj}}$ guides the prediction of objectness with re-weighted binary-cross entropy (BCE) loss $\mathcal{L}_{\text{obj}}$, defined as:

$$\mathcal{L}_{\text{obj}} = \text{wBCE}\left(\boldsymbol{p}, \boldsymbol{y}\right), \tag{9}$$

where $\boldsymbol{y}$ is extracted from the ground truth segmentation map with 1 for the pixel that contains the centroid contour and 0 otherwise. The objectness map end up with only one bright pixel at the contour centroid and all rest pixels are 0. To counteract the class imbalancement, we reweight the BCE loss of class 1 by $\frac{511}{512}$ and 0 by $\frac{1}{512}$.

The second part $\mathcal{L}_{\text{reg}}$ guides the probabilistic regression of EFDs with negative logarithmic likelihood (NLL) loss [28]. The distance is used with the proposed synchronized distance metric d as presented in the previous section for likelihood modeling. The regression loss $\mathcal{L}_{\text{reg}}$ is given by:

$$\mathcal{L}_{\text{reg}} = \sum_{t=1}^{4K+2} \frac{1}{2} \log \hat{\sigma}_{\hat{\boldsymbol{a}}_k}^2 + \frac{d\left(\mu_{\hat{\boldsymbol{a}}_k}, \boldsymbol{a}_k^*\right)}{2\hat{\sigma}_{\hat{\boldsymbol{a}}_k}^2}, \tag{10}$$

where $d\left(\mu_{\hat{\boldsymbol{a}}_k}, \boldsymbol{a}_k^*\right)$ is the distance metric with phase correction defined in (3). The regression loss is minimized if the predicted variance $\hat{\sigma}_{\hat{\boldsymbol{a}}_t}^2$ matches the regression error $d\left(\mu_{\hat{\boldsymbol{a}}_k}, \boldsymbol{a}_k^*\right)$ and thus captures the contour uncertainty [42], which partly contributes to the aleatoric uncertainty.

In addition, the network is encouraged to reduce the L_2 distance between the prediction and the ground truth in the contours. The contour loss $\mathcal{L}_{\text{con}}$ minimizes the contour difference:

$$\mathcal{L}_{\text{con}} = \sum_{i}^{N_c} \inf_j \left\|\hat{\mathcal{C}}(i) - \mathcal{C}(j)\right\|_2^2, \tag{11}$$

where $\hat{\mathcal{C}}$ is derived from the predicted EFD mean $\mu_{\hat{\boldsymbol{a}}}$ and $\mathcal{C}$ from the ground truth EFD $\hat{\boldsymbol{a}}$. The total loss is the weighted combination of the three:

$$\mathcal{L} = \lambda_{\text{obj}}\mathcal{L}_{\text{obj}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}} + \lambda_{\text{con}}\mathcal{L}_{\text{con}}. \tag{12}$$

C. Uncertainty Propagation to the Volume Space

The objectness scores and the predicted probabilistic contours jointly capture the aleatoric uncertainty of segmentation. To simultaneously capture the epistemic uncertainty, we use the well-established Bayesian approach by training ensembles [29]. Because volumetric estimation of ventricles is more clinically relevant than segmentation maps, we propagate the uncertainty from segmentation to volume space and obtain the probabilistic distributions of stroke volume, LVEF, and RVEF.

1) Probabilistic Volume Prediction: The aleatoric uncertainty of the volume prediction also arises from two parts: detection uncertainty and contour uncertainty. We model the volume variation caused by the contour variability as a Gaussian $\mathcal{N}(\mu_c, \sigma_c^2)$. The mean volume μ is calculated from the mean EFD $\mu_{\hat{\boldsymbol{a}}}$. We use Monte-Carlo to sample possible EFDs $\hat{a}_n$:

$$\hat{\boldsymbol{a}}_n = \mu_{\hat{\boldsymbol{a}}} + \sigma_{\hat{\boldsymbol{a}}} \cdot \epsilon, \quad n \in \{1, 2, \cdots, N\}, \tag{13}$$

where ϵ is a Gaussian with zero mean and identity covariance. Then we generate contours based on the N EFD samples, from which the 2D volumes $\{\hat{v}^{(1)}, \hat{v}^{(2)}, \cdots \hat{v}^{(N)}\}$ are calculated over all the contour samples. The contour uncertainty in the volume space is then given by estimating the empirical variance in the area space:

$$\sigma^2 = \text{Var}\left(v^{(1)}, v^{(2)}, \cdots v^{(N)}\right). \tag{14}$$

We further model the volume variation caused by the detection uncertainty. The objectness score p determines whether the predicted slice volume will contribute to the total volume. Volume is computed by summing slice-wise contributions, where each slice first undergoes a stochastic inclusion decision and, if included, contributes a stochastic contour-derived area. We model the stochastic inclusion/exclusion process as a Bernoulli random variable that yield binary decision process, and the contour-induced volume variation as a Gaussian distribution. Combining the two stages, the ventricular volume per slice is then a product of a Bernoulli distribution that models the stochastic inclusion parameterized by p and a Gaussian distribution $\mathcal{N}(\mu_c, \sigma_c^2)$ that accounts for the contour-induced volume variability from (14). The variance of the resultant distribution is given by:

$$\sigma_v^2 = p(1-p)\mu_c^2 + p\sigma_c^2. \tag{15}$$

The final volume variance consists of two parts: $p(1-p)\mu_c^2$ that accounts for the detection uncertainty and the contour uncertainty $p\sigma_c^2$. The detection uncertainty approaches its maximum if p is close to 50%.

We follow the Bayesian approach to quantify the epistemic uncertainty [15], [28], [29], [30], [31], which can be efficiently estimated by evaluating the diversity in predictions made by ensembles of M detection-and-regression models. Each of the

M networks predicts an objectness p_m and contour uncertainty $\mathcal{N}(\mu_m, \sigma^2_{c,m})$. We estimate the aleatoric uncertainty by:

$$\sigma^2_{\text{alea}} = \mathbb{E}_m \left[p_m(1-p_m)\mu^2 + p_m \sigma^2_{c,m} \right]. \tag{16}$$

After ensembling, we get the mean statistics:

$$\overline{p} = \frac{1}{M}\sum_{m=1}^{M} p_m,\ \overline{\mu} = \frac{1}{M}\sum_{m=1}^{M} \mu_{c,m}, \overline{\sigma}^2_c = \frac{1}{M}\sum_{m=1}^{M} \sigma^2_{c,m}. \tag{17}$$

Ensembling introduces an additional contour uncertainty $\sigma^2_\mu = \text{Var}(\mu_1, \mu_2, \cdots, \mu_m)$, which characterizes the mean prediction variance across the M models. The total volume uncertainty can be estimated by:

$$\sigma^2_{\text{total}} = \overline{p}(1-\overline{p}) \cdot \overline{\mu}^2 + \overline{p} \cdot (\overline{\sigma}^2_c + \sigma^2_\mu). \tag{18}$$

So far we can decompose the two types of uncertainty explicitly: The aleatoric uncertainty as computed in (16), and the epistemic uncertainty σ^2_{epi} as the difference between the total and aleatoric uncertainty $\sigma^2_{\text{epi}} = \sigma^2_{\text{total}} - \sigma^2_{\text{alea}}$.

2) Uncertainty for EF Predictions: Based on the volumetric predictions from LV and RV segmentation, the EF of LV and RV can be calculated as:

$$\text{EF} = \frac{V_{\text{ED}} - V_{\text{ES}}}{V_{\text{ED}}} = \left(1 - \frac{V_{\text{ES}}}{V_{\text{ED}}}\right) \times 100\%, \tag{19}$$

where V_{ED} and V_{ES} denote the end-diastolic (ED) and end-systolic (ES) volumes of LV or RV, respectively. The uncertainty of volume can be further propagated to EF by deriving the variance of the ratio distribution $\frac{V_{\text{ES}}}{V_{\text{ED}}}$. Let $V_{\text{ES}} \sim \mathcal{N}(\mu_{\text{ES}}, \sigma^2_{\text{ES}})$ and $V_{\text{ED}} \sim \mathcal{N}(\mu_{\text{ED}}, \sigma^2_{\text{ED}})$, the EF variance can be approximated by

$$\sigma^2_{\text{EF}} \approx \frac{\mu^2_{\text{ES}}}{\mu^2_{\text{ED}}} \left(\frac{\sigma^2_{\text{ES}}}{\mu^2_{\text{ES}}} + \frac{\sigma^2_{\text{ED}}}{\mu^2_{\text{ED}}} \right), \tag{20}$$

under the assumption that the mean volume prediction is much larger than its uncertainty: $\mu_{\text{ES}} \gg \sigma_{\text{ES}}$ and $\mu_{\text{ED}} \gg \sigma_{\text{ED}}$ [43].

3) Modeling the Variability in Expert Annotation: In practice, human experts can make different annotations on the same image [18], [19]. Our proposed aleatoric uncertainty from probabilistic contours provides a straightforward way to model the level of contour variability in expert annotations. We can simulate the variation in human labels by perturbing the ground truth EFD $\boldsymbol{a}$ with additive noise:

$$\boldsymbol{a}' = \boldsymbol{a} + \gamma \sigma_{\boldsymbol{a}} \cdot \epsilon, \tag{21}$$

where ϵ is a sample from a normal distribution, $\sigma_{\boldsymbol{a}}$ is the average variance of each component in EFD estimated from the training dataset, and γ controls the level of variability.

4) Calibration of the Volume and EF Prediction: The predictive calibration reflects the quality of uncertainty estimation. To evaluate the calibration of the volume uncertainty obtained, we calculate the NLL in the test set with the predicted volume $\hat{V}$, predicted variance $\hat{\sigma}^2_V$ and the ground truth volume V. NLL in volume space is defined as:

$$\text{NLL}_V = \frac{1}{2}\log \hat{\sigma}^2_V + \frac{(\hat{V}-V)^2}{2\hat{\sigma}^2_V}. \tag{22}$$

Lower values of NLL indicate that the uncertainty $\hat{\sigma}_V$ covers the range of the possible error. A perfect calibration is achieved when $\hat{\sigma}^2_V = (\hat{V}-V)^2$. For EF predictions, we also compute NLL to measure the calibration performance:

$$\text{NLL}_{\text{EF}} = \frac{1}{2}\log \hat{\sigma}^2_{\text{EF}} + \frac{(\hat{\text{EF}} - \text{EF})^2}{2\hat{\sigma}^2_{\text{EF}}}, \tag{23}$$

where the EF uncertainty σ_{EF} is defined in (20).

The EF uncertainty σ_{EF} naturally provides a confidence interval (CI) for the EF prediction. a two-sided CI at nominal confidence level α is constructed as $\hat{\text{EF}} \pm z_\alpha \sigma_{\text{EF}}$, where z_α denotes the corresponding quantile of the standard normal distribution. To assess the reliability of the predicted uncertainty across different confidence levels, we evaluate calibration using multi-level prediction interval coverage probability (PICP) curves [44]. Specifically, for each $\alpha \in (0,1)$, the empirical coverage PICP_α measures the fraction of samples whose ground-truth EF falls within the predicted CI at level α, defined as:

$$\text{PICP}_\alpha = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left(\text{EF}_i \in \left[\hat{\text{EF}}_i - z_\alpha \sigma_{\text{EF}_i},\ \hat{\text{EF}}_i + z_\alpha \sigma_{\text{EF}_i}\right]\right) \tag{24}$$

where $\mathbf{1}(\cdot)$ is the indicator function, and z_α denotes the $(1+\alpha)/2$ quantile of the standard normal distribution. Plotting the empirical coverage PICP against nominal confidence yields the multi-level PICP curve, where perfect calibration corresponds to the identity line.

To summarize calibration performance with a single scalar metric, we further compute the average calibration error (ACE) [44], defined as the mean absolute deviation between empirical and nominal coverage over multiple confidence levels:

$$\text{ACE} = \int_0^1 |\text{PICP}_\alpha - \alpha| \, d\alpha. \tag{25}$$

V. Experiments

A. Datasets

We use the public ACDC dataset [2] for training, which consists of short-axis cardiac cine MRI images of 150 subjects with a training split of 100 subjects and a test split of 50 subjects for in-domain performance evaluation, where epistemic uncertainty is expected to be minimal and aleatoric effects dominate.

Additionally, we include cine datasets with varying degrees of domain shift to study the relative contributions of aleatoric and epistemic uncertainty to downstream biomarker uncertainty under realistic deployment conditions. To this end, we use the Multi-center Multi-vendor (M&M) cardiac MRI [3], which contains cine scans of 320 subjects from 6 medical centers, introducing moderate domain shift due to scanner and site variability, enabling evaluation under increased epistemic uncertainty. For even larger domain changes, we included an internal cine data set of eight patients with cardiac implantable electronic devices (CIED). Our CIED data set was acquired with the regular SSFP sequence for 1 patient and the alternative gradient echo (GRE) sequence for 7 patients, which has a markedly lower contrast and signal-to-noise ratio compared to SSFP.

TABLE II
GEOMETRIC SEGMENTATION ACCURACY ON CINE DATASETS MEASURED BY THE DICE SCORE, HAUSDORFF DISTANCE, AND THE VOLUME PREDICTION MEAN AVERAGE ERROR (MAE)

	Methods	ACDC			M&M			Device		
		LV	MYO	RV	LV	MYO	RV	LV	MYO	RV
Dice [%]	MoSE	94.0 ± 5.2	89.9 ± 3.2	91.7 ± 5.7	89.6 ± 7.8	82.8 ± 6.2	85.8±11.9	74.2±19.6	61.0±17.5	59.3±27.6
	nnUNet-Ens.	94.6 ± 4.5	90.8 ± 2.3	**92.0 ± 5.5**	90.1 ± 6.9	**83.7 ± 5.6**	**87.2 ± 9.7**	86.3 ± 5.6	**71.9 ± 7.6**	76.3±11.3
	Ens. + TTA	**95.3 ± 3.4**	**91.0 ± 2.1**	91.8 ± 5.4	**90.2 ± 8.0**	82.8 ± 9.0	85.8±13.6	86.9 ± 5.4	71.3 ± 8.1	75.9±14.7
	Proposed	93.2 ± 6.4	88.3 ± 2.9	91.0 ± 5.2	89.4 ± 7.4	82.3 ± 5.9	85.7±10.1	**87.3 ± 4.3**	70.5 ± 7.4	**78.1 ± 8.8**
HD [mm]	MoSE	8.7 ± 7.4	18.7±36.5	12.9±11.7	14.8±17.9	24.5±34.5	18.1±19.8	37.6±27.1	46.6±32.7	32.3±14.4
	nnUNet-Ens.	8.7 ± 10.1	9.8 ± 10.0	**10.5 ± 4.9**	13.7±16.6	17.5±18.7	13.5±10.8	19.6 ± 7.0	21.1 ± 8.4	32.1±24.3
	Ens. + TTA	**6.8 ± 5.2**	**8.2 ± 6.6**	10.7 ± 5.7	11.3±12.6	14.1±15.4	14.0±12.7	18.3 ± 7.4	16.8 ± 6.8	28.5±23.7
	Proposed	8.0 ± 5.8	9.0 ± 5.9	12.2 ± 6.9	**10.2 ± 7.1**	**12.0 ± 8.2**	**12.4 ± 7.8**	**14.1 ± 2.8**	**15.5 ± 4.8**	**22.6 ± 8.6**
MAE [mL]	MoSE	5.0 ± 4.6	8.3 ± 11.2	7.1 ± 9.2	9.2 ± 9.1	13.8±15.9	14.0±19.4	39.9±35.4	36.9±33.0	83.8±73.6
	nnUNet-Ens.	4.5 ± 4.1	**6.6 ± 5.7**	**7.1 ± 8.6**	9.6 ± 8.2	13.9±10.5	**12.7 ± 16.2**	21.1±14.6	23.8±20.1	46.3±44.5
	TTA	4.4 ± 4.2	6.9 ± 7.3	8.5 ± 10.6	9.6 ± 10.7	14.4±15.5	16.3±21.7	19.4±16.2	21.0±17.0	40.1±40.6
	Proposed	6.0 ± 5.6	8.6 ± 8.3	8.6 ± 10.0	10.4 ± 9.6	**12.8 ± 11.0**	14.4±17.8	**18.7 ± 15.5**	**15.6 ± 9.8**	**35.4 ± 34.6**

Finally, we include the SCMR Consensus dataset [19] to compare the generated contours against the inter-observer variability across 7 independent experts on cine scans of 15 subjects.

B. Methods in Comparison

1) **PhiSeg (aleatoric)** We implemented PhiSeg [22] which captures the aleatoric uncertainty in the nnU-Net framework with 6 resolution levels, 5 latent levels, and depth of latent features 4. In inference time, we draw $M = 20$ segmentations.
2) **LogitNorm (aleatoric)** We introduce LogitNorm with temperature $\tau = 0.1$ that regularizes the norm of the output logits such that the network can produce moderate probabilities.
3) **Mixture of Stochastic Experts (MoSE, aleatoric)** We use the MoSE network with 4 experts for the stochastic inference of segmentation masks and train it under the nnUNet framework.
4) **Stochastic Segmentation Network (SSN, aleatoric)**: We leverage the training strategy that models the correlation between pixels by a low-rank approximation with rank=4.
5) **Deep Ensemble (epistemic)** We follow the nnU-Net ensembling strategy [4] which trains a 5-network ensemble with different initialization on 5-fold cross-validation sets. The disagreement in the ensembles captures the epistemic uncertainty.
6) **TTA (aleatoric)** We use TTA as another baseline for aleatoric characterization. In practice, we perform inference 20 times per image with random flipping, rotation of random angle in $[0, 2\pi)$, and additive Gaussian noise with noise power in $[0.8, 1.2]$.
7) **TTA+Ensemble (aleatoric+epistemic)** The total uncertainty in U-Nets can be characterized by combining the aleatoric and epistemic uncertainty and applying TTA in the ensembles. We draw 4 samples per network in the 5-model ensemble, forming 20 predictive predictions.

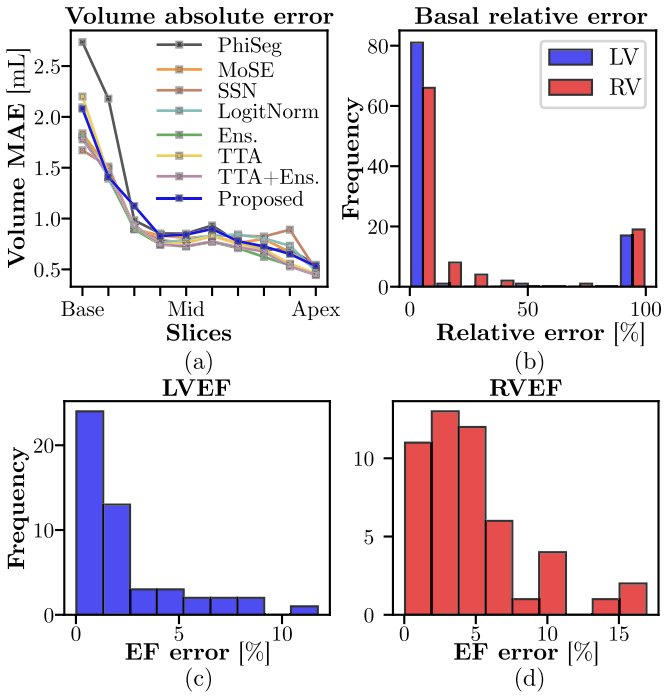

Fig. 3. (a) Per-slice volume error: basal slices contribute the largest errors across all methods. (b) Basal relative volume errors show a peak at 100%, reflecting frequent detection (inclusion/exclusion) errors. (c) LVEF error distribution: 8 subjects (16%) exhibit > 5% error. (d) RVEF errors: 16 subjects (32%) exhibit > 5% error. An EF error of > 5% may lead to misdiagnosis in clinical practice.

C. Implementation Details

The models were trained on the ACDC training split and tested on the ACDC testing split, M&M, and the device datset, respectively. We report the Dice score, Hausdorff distance (HD) and mean mean average error (MAE) of volume and EF. We trained our network with the same 5-fold cross-validation scheme as in nnU-Net. The loss items in (12) were weighted by setting $\lambda_0 = 1.0$, $\lambda_{\text{reg}} = 0.1$, and $\lambda_c = 0.1$, respectively. All implementation details are available from our public code repository.[1]

[1]https://github.com/Ido-zh/ProbContour

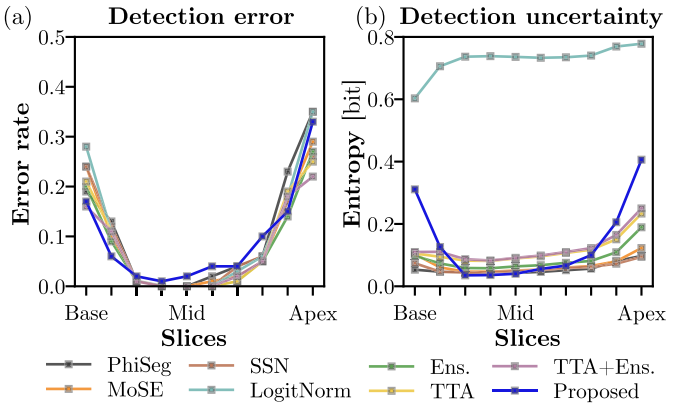

Fig. 4. (a) Detection errors show a consistent U-shaped distribution across slices for all methods. (b) Notably, only the proposed method (blue) yields the detection uncertainty aligned with (a), with elevated uncertainty at basal and apical slices.

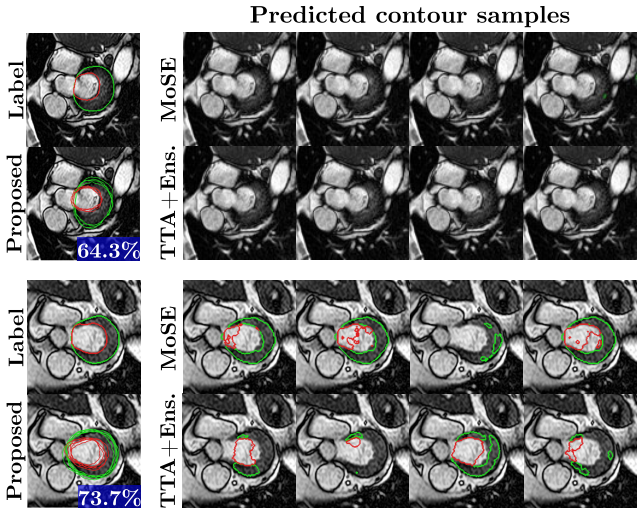

Fig. 5. Comparison of basal slice segmentation performance between MoSE, TTA+Ensemble (TTA+Ens.) and the proposed method. Pixel-wise segmentation networks (MoSE and TTA+Ens.) show similar failure patterns: (upper) confident missing of basal slices, which are included in manual annotation; (Lower) fragmented segmentations that violate anatomy. In contrast, the proposed method yields moderate objectness scores (64.3% and 73.7%) and produces anatomically feasible contours.

VI. Results

A. Segmentation Accuracy and Structured Error Patterns

In Table II, we report the segmentation accuracy of representative pixel-wise classification networks measured by both geometric scores including the Dice coefficient and Hausdorff distance together with the volumetric mean absolute error. All methods show similar geometric and volumetric accuracy on the ACDC and M&M datasets with a difference of <2% in mean Dice score and 1 – 1.5 mL MAE difference. This is also reflected in Fig. 3-(a), where all approaches achieve comparable per-slice volume errors on the in-domain ACDC dataset. Across all methods, errors are consistently largest at basal slices, while mid ventricular slices exhibit substantially lower errors.

The relative basal volume error distribution in Fig. 3-(b) reveals a pronounced peak at 100%, corresponding to the complete inclusion and exclusion of basal slices. This behavior reflects a dominant failure mode caused by anatomical ambiguity rather than pixel-wise misclassifications. Despite the high geometric accuracy: 94.6% for LV 92.0% for RV, a non-negligible portion of subjects still exhibit EF errors exceeding 5%, as is shown in Fig. 3, highlighting a disconnect between segmentation metrics and downstream biomarker reliability.

B. Detection Uncertainty: Objectness Modeling Versus Pixel-Wise Classification

We next examine the detection uncertainty that reflects the ambiguity of slice-level inclusion/exclusion. Fig. 4-(a) shows that the detection error rate follows a consistent U-shaped distribution across slices for all methods, with higher error rates at the basal and apical slices. This trend is largely method-agnostic, suggesting that the detection failures are driven by anatomical ambiguity rather than model design.

However, detection uncertainty exhibits significantly different behavior across methods, shown in Fig. 4-(b). For pixel-classification methods such as PhiSeg, MoSE and TTA+Ensemble, we report the average entropy of foreground pixels as the detection uncertainty. Most pixel-classification-based methods yield nearly uniform foreground uncertainty over slices with a slightly elevated entropy on apical slices. LogitNorm strongly penalizes high-norm logits by outputting logits close to 0 and thus yields moderate probabilities. However, it equally produce low-confident probabilities and a exceedingly high level of detection uncertainty on all slices, failing to align with the true detection error distribution. In comparison, the proposed method produces uncertainty patterns with elevated levels on basal and apical slices, which align with the true detection error distribution.

Qualitative examples in Fig. 5 further illustrate this phenomenon. Pixel-wise segmentation networks often produce confident consistent predictions on ambiguous basal slices, even when stochastic inference is applied. Another typical failure of pixel-wise segmentation networks is the implausible stochastic samples. In contrast, the proposed method expresses uncertainty in regions where detection decisions are inherently ambiguous, providing a structured representation of detection uncertainty.

C. Contour Variations: Regression-Based Probabilistic Contours Versus Pixel-Wise Classification

To isolate the effect of contour variability, we evaluated the EF uncertainty in cases free from detection errors by excluding the cases with any slice-level inclusion/exclusion error. Fig. 6 presents multi-level prediction interval coverage probability (PICP) curves on the ACDC and M&M datasets. The proposed probabilistic contour model yields coverage curves closely aligned with nominal confidence levels for both LVEF and RVEF, while representative pixel-wise uncertainty methods like MoSE and TTA+Ens. exhibit systematic under-coverage. These findings are quantitatively confirmed in Table III, where the proposed method achieves substantially improved 95%

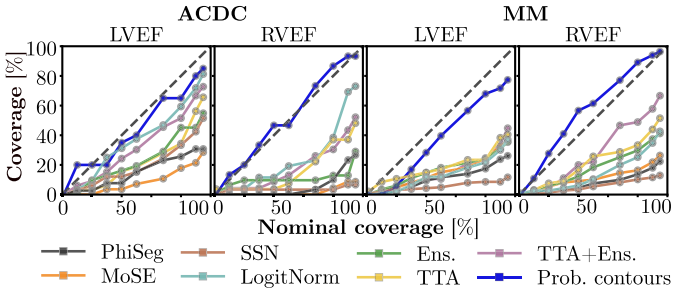

Fig. 6. The prediction interval coverage probability (PICP) curves, which show the nominal coverage of the predicted EF confidence interval and the actual empirical coverage, in the ACDC and MM datasets, respectively. To highlight the influence of contour uncertainty, we excluded cases with detection error. The dashed diagonal indicates perfect calibration. Pixel-classification networks produce systematic lower coverage while the probabilistic contours (prob. contours) produce the best calibration in the EF space.

TABLE III
95% CONFIDENCE INTERVAL (CI) COVERAGE AND AVERAGE CALIBRATION ERROR (ACE) ON CASES WITHOUT DETECTION ERRORS OF THE STOCHASTIC INFERENCE NETWORK REPRESENTATIVE MOSE, TTA+ENSEMBLE (TTA+ENS.), AND THE PROPOSED METHOD

	Methods	ACE [%]		95%−CI Coverage[%]	
		LV_{EF}	RV_{EF}	LV_{EF}	RV_{EF}
ACDC	MoSE	39.1	46.2	28.9	8.7
	TTA+Ens.	16.3	29.7	72.7	52.2
	Contour	**7.4**	**3.5**	**85.0**	**93.3**
M&M	MoSE	33.1	37.0	35.7	26.4
	TTA+Ens.	32.3	20.6	44.7	66.7
	Contour	**11.9**	**7.2**	**77.4**	**96.4**

confidence interval (CI) coverage and lower average calibration error (ACE) of both LVEF and RVEF compared to baselines.

Fig. 7 provides qualitative examples of probabilistic contour samples generated from different methods. The figure demonstrates that the stochastic inference network representative MoSE and TTA+Ensemble tend to produce contours that strongly overlap with each other and as a result, they underestimate the contour variation and fail to induce meaningful volume variability. However, the proposed regression-based method produces probabilistic contours with coherent volume distributions that correctly capture reference annotations.

D. Comparison With Inter-Observer Contour Variability

To validate whether the predicted contour uncertainty reflects clinically meaningful variability, we compare the probabilistic contours generated on the consensus dataset against the inter-observer variability derived from annotations by seven experts. Fig. 8 shows the volume variance induced by pixel-classification networks, such as the stochastic inference network representative MoSE and TTA+Ensemble, and by the proposed regression-based probabilistic contours. All methods were trained with simulated contour annotation noise by adding noise $\gamma = 0.08$ to the ground truth EFDs described in (21).

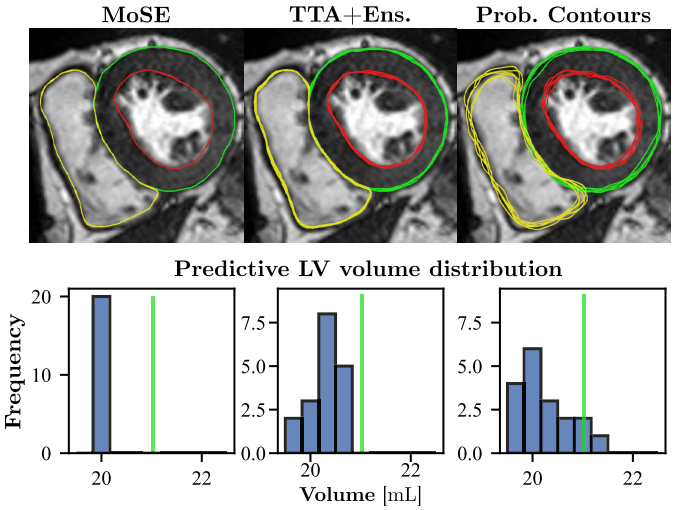

Fig. 7. Typical probabilistic contours on a middle slice generated by pixel-wise classification networks, including MoSE and TTA+Ens and the proposed regression-based method. The volume variations are shown in the lower panel. The green line shows the clinician's reference value. In the MoSE and TTA+Ens cases, the reference value was not covered due to the limited variation in probabilistic contours.

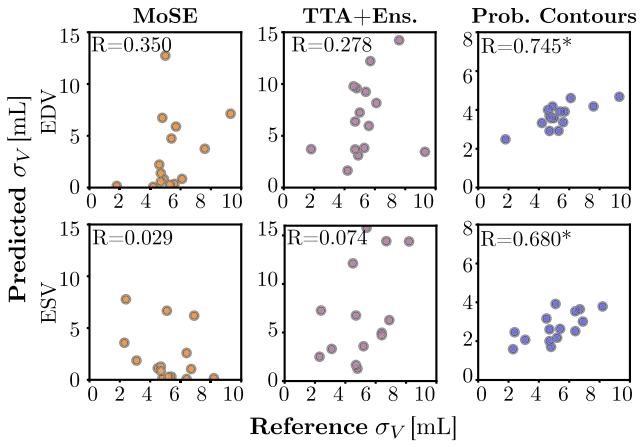

Fig. 8. Predicted LV volume variance from probabilistic contours versus reference volume variation across 7 experts on 15 subjects in the SCMR Consensus dataset [19]. Asterisks indicate the Pearson correlation factors (R) with statistical significance (p-value < 0.05). It can be observed that our method produces the highest correlation. This suggests that our method better captures expert annotation variation, which pixel-wise segmentation methods fail to model.

Fig. 8 demonstrated that the proposed contour regression yields a variation that correlates strongly with expert-derived variance for both EDV and ESV, with Pearson correlation factors of 0.745 and 0.680 with statistical significance (p-value <0.05), respectively. In contrast, the pixel-wise baselines (MoSE and TTA+Ensemble) exhibit weak or negligible correlation. Qualitative comparisons in Fig. 9 further demonstrate that pixel-wise uncertainty produces a thin LV contour variation band on the clean image in Fig. 9-(a) and incoherent contour variations on the low-contrast image in Fig. 9-(b), while the proposed method generates structured contour variability. And the probabilistic contour spread increases from the clean image to the image with low-contrast.

TABLE IV

CALIBRATION OF VOLUME AND EF PREDICTION MEASURED BY NLL ON THREE CINE DATASETS. THE BEST CALIBRATION IS BOLD AND LABELED WITH * IF THERE IS A STATISTICALLY SIGNIFICANT DIFFERENCE FROM THE SECOND BEST

Methods	ACDC		M&M		Device	
	$\mathrm{LV}_V \downarrow$	$\mathrm{RV}_V \downarrow$	$\mathrm{LV}_V \downarrow$	$\mathrm{RV}_V \downarrow$	$\mathrm{LV}_V \downarrow$	$\mathrm{RV}_V \downarrow$
PhiSeg	103.6 ± 214.1	108.3 ± 236.8	226.2 ± 542.7	202.2 ± 459.3	283.9 ± 427.2	88.6 ± 83.3
SSN	105.8 ± 164.4	323.6 ± 1156.5	428.8 ± 1057.7	1732.6 ± 14628.9	250.4 ± 464.1	2347.1 ± 5410.5
MoSE	99.8 ± 178.5	112.2 ± 255.0	160.6 ± 422.7	792.4 ± 7585.0	25.2 ± 33.3	234.4 ± 484.5
LogitNorm	5.5 ± 7.9	12.3 ± 46.4	25.6 ± 55.6	55.1 ± 210.3	156.4 ± 300.3	174.7 ± 523.6
Ens.	44.8 ± 170.4	28.0 ± 70.9	66.3 ± 196.0	88.6 ± 412.5	66.3 ± 148.3	21.0 ± 27.0
TTA	16.6 ± 34.7	14.6 ± 26.3	36.3 ± 90.1	55.6 ± 175.5	7.4 ± 7.2	7.5 ± 6.5
Ens. + TTA	11.5 ± 31.7	6.5 ± 12.4	16.8 ± 42.4	20.8 ± 59.2	$\mathbf{5.3 \pm 4.0}$	6.8 ± 4.9
Proposed	$\mathbf{2.7 \pm 1.0}^*$	$\mathbf{3.0 \pm 1.2}^*$	$\mathbf{3.3 \pm 2.2}^*$	$\mathbf{3.6 \pm 2.4}^*$	5.6 ± 6.0	$\mathbf{4.6 \pm 2.3}$
	$\mathrm{LV}_{\mathrm{EF}}$	$\mathrm{RV}_{\mathrm{EF}}$	$\mathrm{LV}_{\mathrm{EF}}$	$\mathrm{RV}_{\mathrm{EF}}$	$\mathrm{LV}_{\mathrm{EF}}$	$\mathrm{RV}_{\mathrm{EF}}$
PhiSeg	32.6 ± 42.3	83.4 ± 79.0	120.5 ± 300.4	79.6 ± 190.3	167.9 ± 166.5	16.5 ± 21.4
SSN	53.4 ± 91.3	177.2 ± 348.7	195.9 ± 357.4	290.8 ± 609.6	83.4 ± 84.4	238.0 ± 337.4
MoSE	21.0 ± 63.0	54.4 ± 97.0	59.6 ± 160.5	114.8 ± 434.1	4.8 ± 3.8	9.0 ± 6.3
LogitNorm	4.3 ± 6.3	15.5 ± 71.9	12.2 ± 22.7	15.0 ± 37.9	80.3 ± 124.0	58.1 ± 71.7
Ens.	7.5 ± 14.4	17.3 ± 22.7	29.0 ± 36.8	83.0 ± 194.8	12.7 ± 6.6	14.8 ± 5.1
TTA	5.5 ± 13.9	9.5 ± 14.9	21.5 ± 59.3	18.9 ± 53.8	5.6 ± 4.0	3.9 ± 1.5
Ens. + TTA	4.8 ± 7.3	5.0 ± 18.4	11.0 ± 32.9	8.9 ± 33.9	$\mathbf{4.9 \pm 3.7}$	3.8 ± 1.9
Proposed	$\mathbf{2.3 \pm 1.2}^*$	$\mathbf{2.3 \pm 0.8}^*$	$\mathbf{2.4 \pm 1.5}^*$	$\mathbf{2.7 \pm 1.1}^*$	5.8 ± 7.9	$\mathbf{3.0 \pm 5.2}$

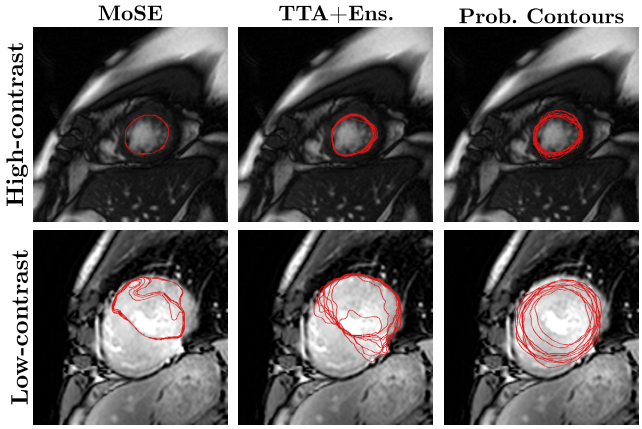

Fig. 9. Predicted probabilistic contours of MoSE, TTA+Ensemble (TTA+Ens.), and the proposed method on a high-contrast case (upper) and a low-contrast case (lower) in the SCMR Consensus dataset [19]. Pixel-wise segmentation methods like MoSE and TTA+Ensemble produce low contour variations on high-contrast images (upper) and implausible contours on low-contrast images (lower). In contrast, our proposed method produces anatomically feasible contours in both cases. Furthermore, as expected, the estimated variation grows as image contrast decreases.

E. Calibration of Probabilistic Volume and EF Predictions on Full Test Sets

Finally, we provide a comprehensive comparison across all competing approaches using negative log-likelihood (NLL) as a probabilistic calibration metric on the full test sets. Table IV shows the calibration performance of stroke volume prediction and EF prediction for both LV and RV, measured by NLL. It is observed that most stochastic inference frameworks including PhiSeg, SSN and MoSE have difficulty predicting well-calibrated volume and EF predictions, as reflected by their high NLL in both the volume and EF space. Deep ensembles produced better calibration than PhiSeg but worse than TTA. Among the pixel-wise classification networks, TTA+Ens. achieves the best calibration performance. In the two SSFP datasets (ACDC and M&M) with limited domain shift, our proposed network consistently yielded the best calibration over the pixel classification networks, achieving NLL's ranging from 2.7 to 3.6. In the device data set with a relatively large domain shift, we observed that both TTA and TTA + assemble achieved good calibration, with performance comparable to our proposed method.

We next evaluate the quality of the generated confidence intervals of EF estimations, where both detection failures and contour variability are present. Fig. 11 lists the multi-level PICP curves of LVEF and RVEF on the high-quality cine datasets ACDC and M&M. The calibration curves shows that representative pixel-wise stochastic segmentation methods like TTA+Ens. and MoSE exhibit systematic under-coverage of LVEF and RVEF variation on both datasets. In contrast, the proposed method consistently yields coverage curves closest to the ideal diagonal across both datasets and for both LVEF and RVEF. Table III quantitatively confirms these observations, with the proposed method achieving the lowest ACE and 95%-CI coverage closest to the nominal level. These results demonstrate that jointly modeling detection and contour variability is essential for reliable EF uncertainty estimation in realistic settings.

Beyond calibration, we assess the correlation between the predicted EF uncertainty and the EF error. Fig. 12 shows the relationship between predicted EF uncertainty and absolute EF error across ACDC and M&M datasets. We split the predicted σ_{EF} into 4 bins and present the distribution of EF error in each bin. The figure illustrates that higher predicted uncertainty consistently corresponds to larger EF errors, with statistically significant correlations for both ventricles. This monotonic relationship indicates that the proposed uncertainty not only achieves population-level calibration but also effectively stratifies prediction reliability at the individual-subject level.

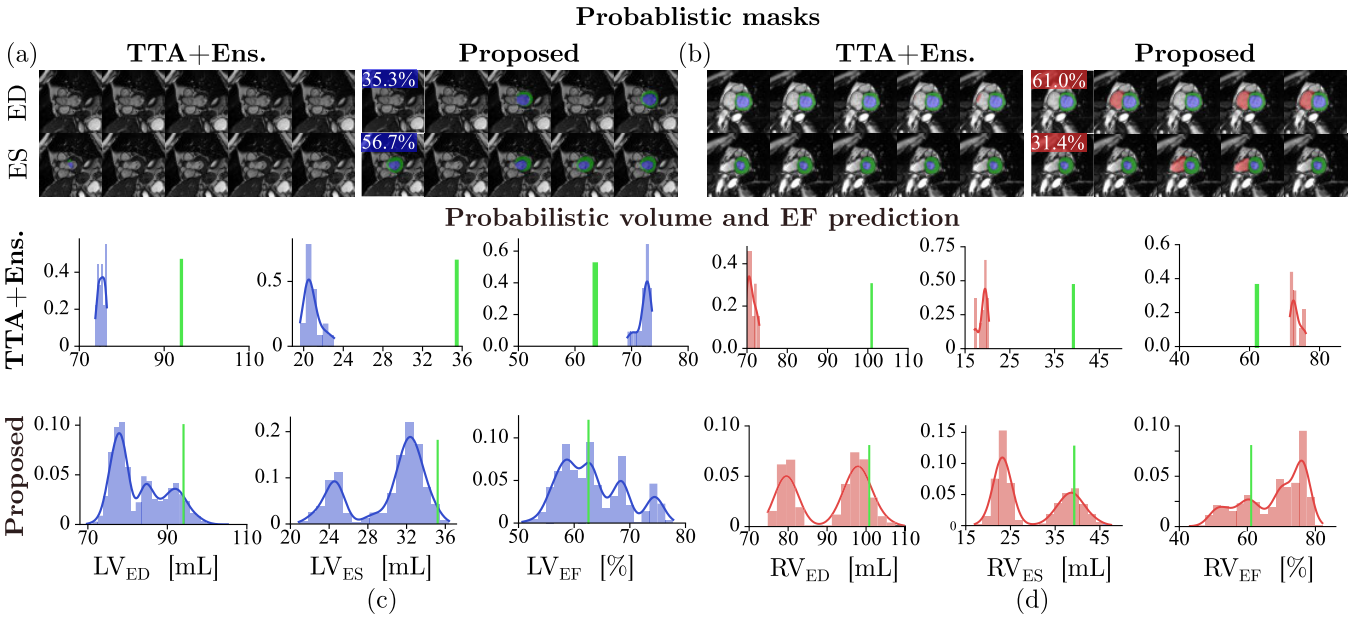

Fig. 10. Probabilistic masks for LV prediction (a) and RV prediction (b) on the ED and ES frames. We display the objectness score for the proposed method in the top-left corner. (c)-(d): Probabilistic volume and EF predictions by TTA+Ensemble and the proposed method. Green lines indicate the clinician's reference value. In both cases, the reference LV and RV volumes were correctly captured by our method but not covered by TTA+Ensemble.

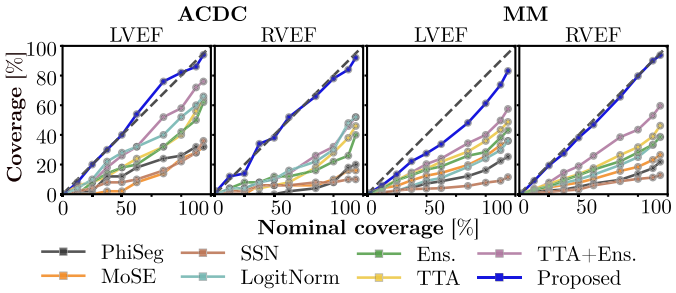

Fig. 11. The prediction interval coverage probability (PICP) curves, which show the nominal coverage of the predicted EF confidence interval and the actual empirical coverage, in the ACDC and MM datasets, respectively. All test cases were used. For both LVEF and RVEF, our proposed method yields empirical coverages closest to the ideal diagonal, whereas pixel-wise classification networks consistently under-estimate uncertainty, resulting in systematically low coverage.

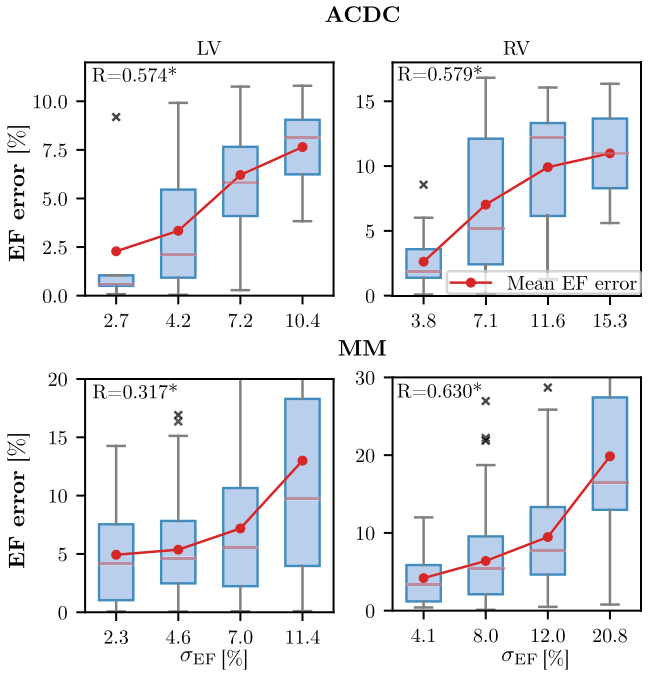

Fig. 12. Relationship between predicted EF uncertainty (σ_{EF}) and absolute EF error in the ACDC and MM datasets. Higher predicted σ_{EF} corresponds to larger EF errors, indicating meaningful uncertainty–error coupling. Pearson correlation coefficients (R) for ACDC are 0.574 (LVEF) and 0.579 (RVEF), and for MM are 0.317 (LVEF) and 0.630 (RVEF), all statistically significant ($p < 0.05$).

Figure 10 shows two representative examples of the predictive distribution by TTA+Ensemble and compares it with our proposed method. The difficulty that limits precision lies in defining the base LV in Figure 10-(a), and RV in Figure 10-(b). TTA+Ensemble did not capture this ambiguity and produced segmentation maps with limited variation (low uncertainty). As a result, the volume distribution has a single mode, which does not cover the true reference value (shown in green lines). In contrast, our proposed method has a detection score of 35.3% and 56.7% for the LV case in the ED and ES phases, respectively, reflecting the aleatoric "detection uncertainty". This leads to a two-modal distribution in EDV and ESV, correctly capturing the high uncertainty with the true EF value in its coverage.

F. Decomposition of Uncertainty Sources

Table VI shows the calibration performance of EF prediction in different configurations evaluated in ACDC. The contour uncertainty alone cannot yield calibrated confidence intervals, achieving 44% coverage with the 95% confidence interval for both LVEF and RVEF. This is in accordance with the observation in Fig. 3 that the error is more pronounced by the detection errors of basal slices. Detection uncertainty contributes significantly to the calibration of EF predictions

TABLE V
COVERAGE OF THE 95% CONFIDENCE INTERVAL (CI) AND AVERAGE CALIBRATION ERROR (ACE) IN ACDC AND M&M DATASETS

	Methods	ACE [%] ↓		95%−CI Coverage [%]	
		LV_{EF}	RV_{EF}	LV_{EF}	RV_{EF}
ACDC	PhiSeg	32.4	43.2	32.0	20.0
	SSN	35.0	42.6	36.0	10.0
	MoSE	37.4	41.8	32.0	16.0
	LogitNorm	17.0	29.0	66.0	52.0
	Ensemble	23.8	33.6	62.0	40.0
	TTA	22.4	33.0	64.0	46.0
	TTA+Ens.	12.6	29.0	76.0	52.0
	Proposed	**2.4**	**2.8**	**94.0**	**92.0**
M&M	PhiSeg	38.4	40.2	25.3	21.9
	SSN	43.4	42.7	11.6	12.8
	MoSE	32.4	36.9	35.9	26.6
	LogitNorm	34.7	32.8	34.7	35.9
	Ensemble	28.7	31.4	43.1	38.8
	TTA	26.0	28.8	48.8	46.2
	TTA+Ens.	22.9	21.9	57.5	59.6
	Proposed	**11.5**	**1.4**	**83.1**	**93.8**

TABLE VI
CONTRIBUTION OF THE DETECTION UNCERTAINTY (DET.), THE CONTOUR UNCERTAINTY (COUT.) TO THE EF CALIBRATION ON THE ACDC DATASET

	Variants	ACE [%]		95%−CI Coverage[%]	
		LV_{EF}	RV_{EF}	LV_{EF}	RV_{EF}
ACDC	Cont.	27.2	29.2	44.0	44.0
	Det.	2.8	4.4	90.0	88.0
	Total	**2.4**	**2.8**	**94.0**	**92.0**

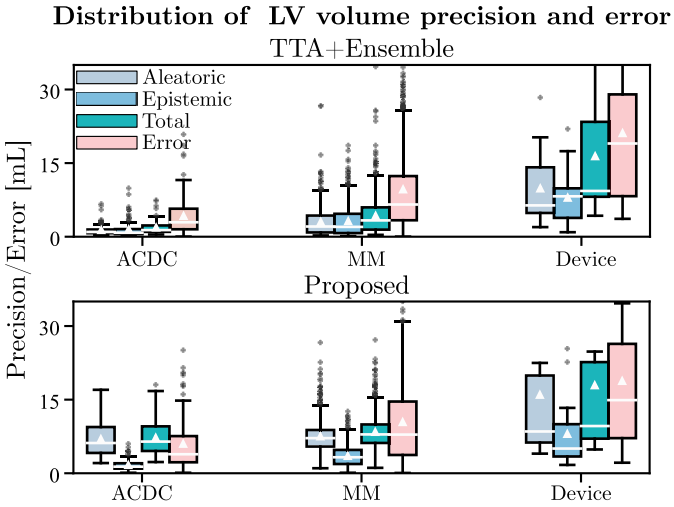

Fig. 13. Distribution of the uncertainty (aleatoric, epistemic, and total) and the error of LV volume. With a small domain shift (ACDC and MM), pixel-wise networks produce low aleatoric uncertainty, and the total uncertainty fails to match the magnitude of errors. In comparison, our proposed method outputs higher aleatoric uncertainty, which accounts for the major source of uncertainty in this case, thereby matching the magnitude of the true error. Across all datasets, including the Device dataset with a large domain shift, both methods yield similar epistemic uncertainty estimates.

with 90% and 88% coverage. The combination of detection and contour uncertainty further improves the calibration on LVEF and RVEF estimation reflected by the reduced ACE and increased 95%-CI coverage.

The low level of epistemic uncertainty is reflected in Figure 13, where we observe that the pixel-space models systematically underestimate volume uncertainty in both the ACDC (in-domain) and MM (small-domain shift) datasets. As a result, the estimated precision cannot match the true magnitude of error and thus often fails to cover the reference manual annotation. In comparison, our proposed method produces high aleatoric uncertainty that matches the actual error, which is crucial for calibrated predictions. The epistemic uncertainty gradually increase with domain-shift, yet the aleatoric uncertainty is the dominant source that contributes the uncertainty.

VII. DISCUSSION

This work investigates aleatoric uncertainty modeling for cardiac MRI segmentation with a focus on the reliability of derived clinical biomarkers, particularly ejection fraction (EF). While segmentation accuracy has been extensively studied, our results demonstrate that high geometric accuracy alone is insufficient to ensure reliable biomarker estimation. Instead, uncertainty must be modeled in a manner that reflects the underlying failure modes of segmentation and their propagation to downstream clinical metrics. Quantifying the uncertainty of segmentation and segmentation-derived biomarkers is highly relevant for the reliability of clinical diagnosis. However, in most deployment scenarios, the input data have little or limited domain shifts (e.g. similar MRI sequence and image contrast). In these cases, the difficulty does not arise from a low image quality, but rather from the intrinsic ambiguity of defining LV and RV structures, hence aleatoric uncertainty. The aleatoric uncertainty determines the precision of biomarkers, but principled methods to model aleatoric uncertainty in medical image segmentation are currently lacking.

A central finding is the distinction between detection uncertainty and contour uncertainty, which mimics the expert annotation process and models the aleatoric uncertainty therein. By separating detection uncertainty from contour uncertainty, we effectively transform pixel-wise classification into a global objectness classification problem, thereby alleviating the inherent locality limitation of pixel-wise predictions. This allows detection ambiguity to be expressed at the slice level, where inclusion or exclusion decisions are semantically meaningful, as is shown in Fig. 4. The detection uncertainty is the dominant contributor to calibration (Table. VI), as detection errors account for the largest portion of EF error. Contour uncertainty plays a complementary but essential role by refining uncertainty estimation once detection is correct.

Contour uncertainty becomes critical once detection is correct, as demonstrated in Fig 6. By isolating cases without detection errors, we show that contour-level uncertainty yields well-calibrated EF confidence intervals, whereas pixel-wise stochastic methods consistently underestimate uncertainty. This indicates that pixel-wise variability does not translate into meaningful variability of downstream clinical metrics.

Furthermore, contour regression introduces an implicit shape prior through a compact contour representation, which constrains the space of plausible segmentation samples and avoids anatomically implausible samples commonly produced by pixel-wise stochastic methods (Fig. 5, 9). Finally, joint mean–variance regression over contour parameters prevents stochastic contour realizations from collapsing to a one-pixel-wide boundary like in pixel-wise segmentation networks (Fig. 7), ensuring that uncertainty is expressed as coherent structural variability rather than local pixel noise.

Beyond calibration, the predicted EF uncertainty is informative at the subject level. Higher uncertainty consistently corresponds to larger EF errors, enabling effective risk stratification (Fig. 12). Furthermore, the experiments on the consensus dataset demonstrate that the contour-induced variance correlates strongly with inter-observer variability, providing external validation that the proposed uncertainty captures clinically meaningful ambiguity rather than arbitrary noise (Fig. 8).

Our empirical results confirm that, in cases of limited domain shifts, uncertainty stems mainly from the intrinsic ambiguity of data and difficulty of task (aleatoric), instead of domain shifts (epistemic). This is confirmed by the observation in Fig. 3 that different segmentation networks end up with similar volume error and slice-wise error distributions. Experiment results showed that current pixel-space uncertainty estimation methods fail to capture such aleatoric uncertainty and severely underestimate uncertainty, leading to "over-confident" biomarkers, whereas our method can produce well-calibrated uncertainty in such cases.

We note that aleatoric uncertainty arises from the intrinsic ambiguity in data, which cannot be reduced by adding more training data. In the case of cardiac MRI, the aleatoric uncertainty is not only related to the partial volume effect, but also associated with the speed limitation of MRI: the full heart is typically imaged by multiple 2D+t sequences, each in a separate breath hold. This causes shifts between slices, making it difficult to apply 3D models. Therefore, current cine MRI segmentation relies on 2D models, which encounter high uncertainty at the basal part where the anatomy appears complex on the 2D plane. Improved imaging techniques, such as 3D+t cine MRI, could effectively reduce imaging-induced aleatoric uncertainty by enabling 3D segmentation.

VIII. Conclusion

In conclusion, we proposed a probabilistic detection-and-regression framework for cardiac MRI segmentation that explicitly models uncertainty arising from slice inclusion ambiguity and contour variability. By disentangling these sources and propagating them to the biomarker space, the method enables more reliable estimation of the precision of clinically important biomarkers, including stroke volume, LVEF, and RVEF. Explicit modeling of detection and contour uncertainty leads to better-calibrated biomarker confidence intervals in common deployment scenarios with limited domain shift. While not designed for domain generalization, the framework exhibits robust behavior under domain shift due to its structured contour representation. More broadly, this work highlights the importance of uncertainty-aware biomarker assessment and provides a general paradigm applicable to other segmentation tasks requiring reliable biomarker quantification.

Acknowledgment

The authors also gratefully acknowledge the computational resources provided by the NVIDIA Academic Grant Program.

References

[1] R. Azad et al., "Medical image segmentation review: The success of U-Net," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 12, pp. 10076–10095, Dec. 2024.

[2] O. Bernard et al., "Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved?," *IEEE Trans. Med. Imag.*, vol. 37, no. 11, pp. 2514–2525, Nov. 2018.

[3] V. M. Campello et al., "Multi-centre, multi-vendor and multi-disease cardiac segmentation: The M&Ms challenge," *IEEE Trans. Med. Imag.*, vol. 40, no. 12, pp. 3543–3554, Dec. 2021.

[4] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation," *Nature methods*, vol. 18, no. 2, pp. 203–211, 2021.

[5] F. Isensee et al., "nnU-Net revisited: A call for rigorous validation in 3D medical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2024, pp. 488–498.

[6] R. M. Thomas and J. John, "A review on cell detection and segmentation in microscopic images," in *Proc. Int. Conf. Circuit, Power Comput. Technol. (ICCPCT)*, Apr. 2017, pp. 1–5.

[7] A. Wadhwa, A. Bhardwaj, and V. S. Verma, "A review on brain tumor segmentation of MRI images," *Magn. Reson. Imag.*, vol. 61, pp. 247–259, Sep. 2019.

[8] M.-K. Fassia et al., "Deep learning prostate MRI segmentation accuracy and robustness: A systematic review," *Radiol., Artif. Intell.*, vol. 6, no. 4, Jul. 2024, Art. no. 230138.

[9] W. M. Stallings and G. M. Gillmore, "A note on 'accuracy' and 'precision,'" *J. Educ. Meas.*, vol. 8, no. 2, pp. 127–129, 1971.

[10] P. Kellman and M. S. Hansen, "T1-mapping in the heart: Accuracy and precision," *J. Cardiovascular Magn. Reson.*, vol. 16, no. 1, p. 2, 2014.

[11] S. Weingärtner et al., "Development, validation, qualification, and dissemination of quantitative MR methods: Overview and recommendations by the ISMRM quantitative MR study group," *Magn. Reson. Med.*, vol. 87, no. 3, pp. 1184–1206, Mar. 2022.

[12] N. Kawel-Boehm et al., "Reference ranges ('normal values') for cardiovascular magnetic resonance (CMR) in adults and children: 2020 update," *J. Cardiovascular Magn. Reson.*, vol. 22, no. 1, p. 87, 2020.

[13] P. A. Heidenreich et al., "AHA/ACC/HFSA guideline for the management of heart failure: A report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines," *J. Amer. College Cardiol.*, vol. 79, no. 17, pp. e263–e421, 2022.

[14] Y. Zhao, Y. Zhang, O. Simonetti, Y. Han, and Q. Tao, "Lost in tracking: Uncertainty-guided cardiac cine MRI segmentation at right ventricle base," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2024, pp. 415–424.

[15] A. Kendall and Y. Gal, "What uncertainties do we need in Bayesian deep learning for computer vision?," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 30, 2017, pp. 5574–5584.

[16] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proc. 33rd Int. Conf. Mach. Learn.*, Jun. 2016, pp. 1050–1059.

[17] L. Joskowicz, D. Cohen, N. Caplan, and J. Sosna, "Inter-observer variability of manual contour delineation of structures in CT," *Eur. Radiol.*, vol. 29, no. 3, pp. 1391–1399, Mar. 2019.

[18] A. Suinesiaputra et al., "A collaborative resource to build consensus for automated left ventricular segmentation of cardiac MR images," *Med. Image Anal.*, vol. 18, no. 1, pp. 50–62, Jan. 2014.

[19] A. Suinesiaputra et al., "Quantification of LV function and mass by cardiovascular magnetic resonance: Multi-center variability and consensus contours," *J. Cardiovascular Magn. Reson.*, vol. 17, no. 1, p. 63, Jan. 2015.

[20] G. Wang, W. Li, M. Aertsen, J. Deprest, S. Ourselin, and T. Vercauteren, "Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks," *Neurocomputing*, vol. 338, pp. 34–45, Apr. 2019.

[21] Z. Gao, Y. Chen, C. Zhang, and X. He, "Modeling multimodal aleatoric uncertainty in segmentation with mixture of stochastic experts," 2022, *arXiv:2212.07328*.

[22] C. F. Baumgartner et al., "PHiSeg: Capturing uncertainty in medical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, Shenzhen, China. Cham, Switzerland: Springer, 2019, pp. 119–127.

[23] M. Monteiro et al., "Stochastic segmentation networks: Modelling spatially correlated aleatoric uncertainty," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, 2020, pp. 12756–12767.

[24] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, vol. 9351. Cham, Switzerland: Springer, 2015, pp. 234–241.

[25] H.-Y. Zhou, J. Guo, Y. Zhang, L. Yu, L. Wang, and Y. Yu, "nnFormer: Interleaved transformer for volumetric segmentation," 2021, *arXiv:2109.03201*.

[26] F. Guo et al., "Improving cardiac MRI convolutional neural network segmentation on small training datasets and dataset shift: A continuous kernel cut approach," *Med. Image Anal.*, vol. 61, May 2020, Art. no. 101636.

[27] H. Wei, R. Xie, H. Cheng, L. Feng, B. An, and Y. Li, "Mitigating neural network overconfidence with Logit normalization," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 23631–23644.

[28] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," 2016, *arXiv:1612.01474*.

[29] A. Mehrtash, W. M. Wells, C. M. Tempany, P. Abolmaesumi, and T. Kapur, "Confidence calibration and predictive uncertainty estimation for deep medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 39, no. 12, pp. 3868–3878, Dec. 2020.

[30] Y. Zhao, C. Yang, A. Schweidtmann, and Q. Tao, "Efficient Bayesian uncertainty estimation for nnU-Net," in *Proc. Int. Conf. Med. Image Comput. Assist. Intervent.* Cham, Switzerland: Springer, 2022, pp. 535–544.

[31] Y. Zhao et al., "Bayesian uncertainty estimation by Hamiltonian Monte Carlo: Applications to cardiac MRI segmentation," 2024, *arXiv:2403.02311*.

[32] M. Ng et al., "Estimating uncertainty in neural networks for cardiac MRI segmentation: A benchmark study," *IEEE Trans. Biomed. Eng.*, vol. 70, no. 6, pp. 1955–1966, Jun. 2022.

[33] S. A. A. Kohl et al., "A hierarchical probabilistic U-Net for modeling multi-scale ambiguities," 2019, *arXiv:1905.13077*.

[34] M. S. Ayhan and P. Berens, "Test-time data augmentation for estimation of heteroscedastic aleatoric uncertainty in deep neural networks," in *Proc. Med. Imag. Deep Learn.*, 2018, pp. 1–9.

[35] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[36] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.

[37] T. Hastie, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York, NY, USA: Springer, 2009.

[38] F. P. Kuhl and C. R. Giardina, "Elliptic Fourier features of a closed contour," *Comput. Graph. Image Process.*, vol. 18, no. 3, pp. 236–258, Mar. 1982.

[39] E. Upschulte, S. Harmeling, K. Amunts, and T. Dickscheid, "Contour proposal networks for biomedical instance segmentation," *Med. Image Anal.*, vol. 77, Jan. 2022, Art. no. 102371.

[40] C. Chen et al., "FFPN: Fourier feature pyramid network for ultrasound image segmentation," in *Proc. Int. Workshop Mach. Learn. Med. Imag.* Cham, Switzerland: Springer, 2023, pp. 166–175.

[41] S. Tilborghs, J. Bogaert, and F. Maes, "Shape constrained CNN for segmentation guided prediction of myocardial shape and pose parameters in cardiac MRI," *Med. Image Anal.*, vol. 81, Jun. 2022, Art. no. 102533.

[42] M.-H. Laves, S. Ihler, J. F. Fast, L. A. Kahrs, and T. Ortmaier, "Recalibration of aleatoric and epistemic regression uncertainty in medical imaging," 2021, *arXiv:2104.12376*.

[43] E. Díaz-Francés and F. J. Rubio, "On the existence of a normal approximation to the distribution of the ratio of two independent normal random variables," *Stat. Papers*, vol. 54, pp. 309–323, Apr. 2013.

[44] I. K. Bazionis and P. S. Georgilakis, "Review of deterministic and probabilistic wind power forecasting: Models, methods, and future research," *Electricity*, vol. 2, no. 1, pp. 13–47, Jan. 2021.