Improved post-hoc probability calibration for out-of-domain MRI segmentation

Cheng Ouyang[1](✉), Shuo Wang[2], Chen Chen[1], Zeju Li[1], Wenjia Bai[1,3,4], Bernhard Kainz[1,5] Daniel Rueckert[1,6]

[1] BioMedIA Group, Department of Computing, Imperial College London, UK
[2] School of Basic Medical Sciences, Fudan University, China
[3] Department of Brain Sciences, Imperial College London, UK
[4] Data Science Institute, Imperial College London, UK
[5] Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany
[6] Klinikum rechts der Isar, Technical University of Munich, Germany
c.ouyang@imperial.ac.uk

Abstract. Probability calibration for deep models is highly desirable in safety-critical applications such as medical imaging. It makes output probabilities of deep networks interpretable, by aligning prediction probability with the actual accuracy in test data. In image segmentation, well-calibrated probabilities allow radiologists to identify regions where model-predicted segmentations are unreliable. These unreliable predictions often occur to out-of-domain (OOD) images that are caused by imaging artifacts or unseen imaging protocols. Unfortunately, most previous calibration methods for image segmentation perform sub-optimally on OOD images. To reduce the calibration error when confronted with OOD images, we propose a novel post-hoc calibration model. Our model leverages the pixel susceptibility against perturbations at the local level, and the shape prior information at the global level. The model is tested on cardiac MRI segmentation datasets that contain unseen imaging artifacts and images from an unseen imaging protocol. We demonstrate reduced calibration errors compared with the state-of-the-art calibration algorithm.

1 Introduction

In safety-critical applications like medical imaging, segmentation models are required to produce accurate predictions on clean input data and are also expected to be *aware* of predictions for which the model has *low confidence*, when confronted with out-of-domain (OOD) data. In medical imaging, OOD data is often caused by imaging artifacts or changes in imaging protocols. The awareness of

Link to code: https://github.com/cheng-01037/Probability_Calibration_for_OOD_MRI_Segmentation

uncertainty allows to alert radiologists about potentially unreliable predictions. Unfortunately, deep models are found to be generally over-confident about predicted probabilities [1,2].

Probability calibration corrects over- or under-confident predictions, and makes prediction probability *interpretable*, by aligning it with the accuracy on the test dataset. For example, if a segmentation model yields a *confidence* (the probability of the highest-scored class) of 70% for each pixel in a test image, we say the model to be well-calibrated if 70% of the pixels are correctly predicted [3].

Unfortunately, most existing probability calibration methods cannot be directly applied to medical image segmentation due to the following reasons: First, the majority of existing methods are designed for image classification, which yield a single class probability per image [4,5,6,7,8]. Secondly, most previous methods assume training and testing images are from a same domain. However, we argue that it is the OOD image for which probability calibration is most desirable, while most calibration methods are shown to perform sub-optimally on OOD images [9]. Therefore, in this study we particularly focus on improving calibration for corrupted medical images.

In this work, we propose a new learning-based probability calibration model for medical image segmentation on out-of-domain (OOD) data. Particularly, we focus on the most flexible calibration setting: *post-hoc* calibration that can be applied to various frozen feed-forward networks. Specifically, our calibration model outputs a *tempeature map* that re-adjusts the prediction probability of the segmentation network [6,3], correcting over- or under-confident probabilities. Unlike the state-of-the-art method [3] that only considers the pixel values of input images and their logits, our model finds unreliable predictions by considering how susceptible the prediction of each pixel is, against small perturbations. Such susceptibility helps to reveal the uncertainty caused by the real-world perturbations that originate from imperfect acquisition process (device noise, patient movement *etc.*) or changes in imaging condition (machine vendors, imaging protocols *etc.*). The proposed model further takes advantage of global prior information about the shapes of segmentation targets. These local-level and global-level sources of information strengthens the calibration performance for OOD images. Our contributions can be summarized as follows:

- We systematically investigate post-hoc probability calibration for the safety-critical medical image segmentation on out-of-domain (OOD) images.
- We propose a new learning-based probability calibration model that incorporates the susceptibility information of pixel-level predictions against perturbations at the local level, and the shape prior information at the global level. The proposed method demonstrates improved performance on OOD testing images compared to the state-of-the-art method.
- We build a comprehensive testing environment for post-hoc calibration, on segmentation for out-of-domain MRI. It incorporates common imaging artifacts: motion artifacts, bias fields, ghosting artifacts, spikes in k-space, and an unseen imaging protocol: late gadolinium enhancement (LGE) sequence for MRI.

2 Related Work

Probability calibration for image segmentation: Most probability calibration methods can be categorized into three types: 1) training strategies that intrinsically improve calibration for the task network (classification, regression, *etc.*). These techniques include focal loss [10], multi-task learning [11], adversarial training [12]; 2) Bayesian methods that carefully model the uncertainties of model parameters, input data and/or labeling process [13,14,15,16,17]; 3) post-hoc methods that post-process the softmax output (probability) of an already-trained task network [4,5,6,3]. Our work follows the post-hoc framework due to its superior flexibility: being applicable to most of already-trained task networks.

More recently, several papers have discussed calibration for image segmentation: [16] evaluates the effects of segmentation losses, model ensembling and MC-dropout on calibration. [11] demonstrates that multi-task learning improves calibration. However, neither works contribute further to post-hoc calibration. Our idea of using data augmentation to estimate susceptibility of pixel-level predictions, which can be interpreted as aleatoric uncertainty estimation, is inspired by [15]. However, [15] does not investigate post-hoc calibration itself. Our method is built on the state-of-the-art local temperature scaling (LTS) [3]. To reduce the calibration error on OOD images, we extend LTS by incorporating pixel-level susceptibility and global-level shape prior information.

Segmenting out-of-domain medical images: A robust image segmentation model can usually be obtained by applying input-level or feature-level data augmentations [18,19,20], or by enforcing shape priors [21,22,23]. Unlike these works, our method focuses on the under-explored problem of promoting interpretability of prediction probabilities, especially for those on out-of-domain images.

Segmentation quality assessment: Segmentation quality assessment [24,25,26] predicts a global model performance score, and/or makes corrections to the predicted segmentation labels. Probability calibration is more challenging, as it is required to make continuous, pixel-wise adjustments to prediction probabilities.

3 Method

Model-based post-hoc calibration: We aim to align the prediction probability with the accuracy on the test dataset. To this end, in model-based post-hoc calibration, we build a separate calibration model $g_\phi(\cdot)$ for a pre-trained task model (in our case segmentation) $f_\theta(\cdot)$. To train the calibration model, the validation dataset for the task model is re-used for building $g_\phi(\cdot)$. We let $\mathbf{x}_i \in \mathbb{R}^{1\times M\times N}$ be the image, $\mathbf{y}_i \in \mathbb{R}^{C\times M\times N}$ the ground truth segmentation in the form of one-hot encoding, where (M, N) is the spatial size and C the number of classes. Note, it is usually desirable that the calibration process does not affect the categorical prediction $\hat{\mathbf{y}}_i$ for segmentation (therefore does not change the accuracy of $f_\theta(\cdot)$).

Temperature scaling: Temperature scaling [27,3] is one of the most simple and effective frameworks for probability calibration. It produces a temperature factor

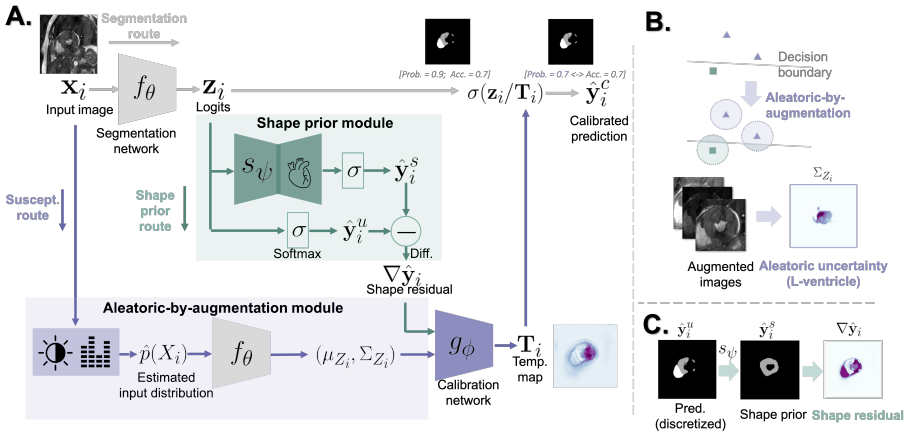

Fig. 1. A. Workflow of the proposed calibration technique: A temperate map $\mathbf{T}_i$ is used to adjust probabilities of a segmentation network. To do this, the image $\mathbf{x}_i$ is sent through a segmentation network $f_\theta(\cdot)$ to obtain the logits $\mathbf{z}_i$. Meanwhile, to obtain $\mathbf{T}_i$, $\mathbf{x}_i$ is sent through two calibration routes: In the susceptibility route, the estimated distribution $\hat{p}(X_i)$ of $\mathbf{x}_i$ is obtained by repeated data augmentations. The uncertainty $(\mu_{Z_i}, \Sigma_{Z_i})$ is computed by sending samples of $\hat{p}(X_i)$ to $f_\theta(\cdot)$. In the shape prior route, $\mathbf{z}_i$ is sent to the shape prior network $s_\psi(\cdot)$ to obtain a shape residual $\nabla\hat{\mathbf{y}}_i$ which highlights the regions where the prediction differs from the prior knowledge about plausible shapes of segmentation targets. The calibration network $g_\phi(\cdot)$ takes $(\mu_{Z_i}, \Sigma_{Z_i})$ and $\nabla\hat{\mathbf{y}}_i$ as inputs and estimates $\mathbf{T}_i$ for rescaling logits $\mathbf{z}_i$ of the segmentation. **B. Aleatoric uncertainty** reflects the susceptibility (shaded regions trespassing the decision boundary) of a prediction under small perturbations. **C. Shape prior and shape residual**, highlighting potentially unreliable predictions.

(or map) $\mathbf{T}_i > 0$ to weigh over-confident predictions down while boost under-confident ones. Formally, let $\mathbf{z}_i = f_\theta(\mathbf{x}_i)$, $\mathbf{z}_i \in \mathbb{R}^{C\times M\times N}$ be the output logits, let $\sigma(\cdot)$ denote the softmax function along the channel dimension, we naturally have the uncalibrated probability $\hat{\mathbf{y}}_i^u = \sigma(\mathbf{z}_i)$. While with the temperature map $\mathbf{T}_i \in \mathbb{R}^{C\times M\times N}$, the calibrated probability $\hat{\mathbf{y}}_i^c$ can be obtained by re-scaling the logits using $\mathbf{T}_i$, *i.e.* $\hat{\mathbf{y}}_i^c = \sigma(\mathbf{z}_i/\mathbf{T}_i)$[7].

Method overview: We aim to obtain a temperature-scaling-based calibration network $g_\phi(\cdot)$ that is suitable for out-of-domain (OOD) testing images. Examples of these OOD images are assumed to be *unseen* by the segmentation network $f_\theta(\cdot)$ and the calibration network $g_\phi(\cdot)$ during their training processes. To this end we propose to 1) provide the susceptibility of the prediction of each pixel against small perturbations caused by potential image corruption/artifact or a

[7] To ensure that the calibration does not affect the accuracy of the task network, for each spatial location (m, n) in $\mathbf{T}_i$, it is usually assumed that $\mathbf{T}_i(c_j, m, n) = \mathbf{T}_i(c_k, m, n)$, $\forall(c_j, c_k) \in \{1, 2, 3, ..., C\}$, *i.e.*, temperature values remain the same for different channels/classes) [6,3].

change in imaging protocol. This susceptibility reflects how likely the prediction of a pixel might be altered when real image artifacts or changes in imaging protocols are present. This is also known as *aleatoric uncertainty* [8] [14,15]. 2) We also provide the calibration network with prior information about the shape of the target segmentation. This shape prior is encoded by a denoising autoencoder $s_\psi(\cdot)$ and it provides a second opinion about the correctness of the prediction.

As shown in Fig. 1-A, to obtain the temperature map $\mathbf{T}_i$, the input $\mathbf{x}_i$ is fed to two modules: The *Aleatoric-by-augmentation* module (colored in purple) estimates the pixel-level susceptibility (aleatoric uncertainty) by repeated data augmentations. The *shape prior module* (colored in green) compares the uncalibrated prediction with the shape prior encoded in the denoising autoencoder $s_\psi(\cdot)$, and provides the calibration network $g_\phi(\cdot)$ with the residual between the uncalibrated prediction and the prior. The calibration network $g_\phi(\cdot)$ takes the outputs of the two modules, and estimates a temperature map for adjusting $\mathbf{z}_i$. Finally, the calibrated prediction is made by passing $\mathbf{z}_i/\mathbf{T}_i$ to a softmax layer.

Aleatoric uncertainty by augmentation: The aleatoric-by-augmentation module provides the calibration network $g_\phi(\cdot)$ with information about susceptibility of predictions for each pixel against small perturbations. Intuitively, if the prediction can be easily flipped by a small perturbation, the prediction of that pixel could be unreliable. In medical images, OOD images can also be viewed as being generated by perturbing intra-domain images [23].

To formally model this susceptibility, we resort to the concept of *aleatoric uncertainty* [13,14,15]. As shown in Fig. 1-B, it models images to have a distribution $p(X_i)$ arising from the acquisition process, rather than treating each image as a single data point (which is instead assumed by the state-of-the-art LTS [3]). This modeled distribution can be written as $p(X_i) = \int p(X_i|a)p(a)da$, where $p(X_i|a)$ represents the image acquisition process and $a \sim p(A)$ denotes the "randomness" within different possible acquisition processes [15]. Then, the susceptibility (uncertainty) can be estimated by propagating $p(X_i)$ through the segmentation model $f_\theta(\cdot)$.

In practice, inspired by [15,20], we employ data augmentation to obtain the estimation $\hat{p}(X_i)$ of the real $p(X_i)$. Specifically, for each image $\mathbf{x}_i$, we perform repeated augmentations to obtain $\{\mathbf{x}'_{i,l}|\mathbf{x}'_{i,l} = \mathcal{T}_{a'_l}(\mathbf{x}_i),\ a'_l \sim p(A')\}$, where $l = 1, 2, 3, ..., N_A$ is the index of augmented samples and $\mathcal{T}_{a'_l}(\cdot)$'s are photometric augmentations parameterized by a'_l's. To ensure fairness, $\mathcal{T}_{a'_l}(\cdot)$'s are configured to be the *same* types of photometric augmentations used for training $f_\theta(\cdot)$ and they *do not* incorporate the corruptions (artifacts) in the testing data. Then, the propagated uncertainty in the logits, in the form of mean μ_{Z_i} and variance Σ_{Z_i}, can be computed by sending $\{\mathbf{x}'_{i,l}\}$ to the segmentation network $f_\theta(\cdot)$. For simplicity, when computing Σ_{Z_i}, each pixel is assumed to be independent.

Shape prior: To provide a second opinion about the correctness of the segmentation, shape priors [24,28,21] are used. For probability calibration, if the pre-

[8] We do not explicitly highlight it as aleatoric uncertainty, since we do not have the ground truth to evaluate the accuracy of this estimation of aleatoric uncertainty.

dicted shape deviates largely from the prior information about plausible shapes, the prediction is likely to be unreliable.

Here, we employ a denoising autoencoder as the shape prior model. It memorizes correct shapes of segmentation targets in the validation dataset. As shown in the green block in Fig. 1-A, the autoencoder $s_\psi(\cdot)$ takes the uncalibrated logits $\mathbf{z}_i$ as the input and produces a denoised plausible shape $\hat{\mathbf{y}}_i^s$ of the segmentation target, in the form of probabilities. To highlight regions where the uncalibrated prediction $\hat{\mathbf{y}}_i^u = \sigma(\mathbf{z}_i)$ deviates from the plausible shape $\hat{\mathbf{y}}_i^s$, we send the shape residual $\nabla\hat{\mathbf{y}}_i = \hat{\mathbf{y}}_i^s - \hat{\mathbf{y}}_i^u$ to the calibration network. An example of a shape residual is shown in Fig. 1-C. In practice, to avoid learning an identity mapping, we apply heavy dropout to the encoder part of $s_\psi(\cdot)$ during training.

Unlike the shape priors in [21,23], which directly correct the prediction, we do not expect $s_\psi(\cdot)$ to provide highly accurate segmentations: As shown in Fig. 1-C., the right ventricle has been correctly predicted by $f_\theta(\cdot)$ while $s_\psi(\cdot)$ (erroneously) disagrees. Instead, we only expect the shape prior module to highlight potentially implausible regions. We leave the calibration network to make the final decision.

Calibration network: The calibration network $g_\phi(\cdot)$ produces a temperature map $\mathbf{T}_i$ that is specific to $\mathbf{x}_i$, by considering the pixel-level susceptibility (uncertainty) $(\mu_{Z_i}, \Sigma_{Z_i})$ and the shape residual $\nabla\hat{\mathbf{y}}_i$. Following the baseline LTS [3], we also send the image $\mathbf{x}_i$ and the uncalibrated logits $\mathbf{z}_i$ to $g_\phi(\cdot)$. After $\mathbf{T}_i$ is computed, the calibrated prediction $\hat{\mathbf{y}}_i^c$ is given by $\hat{\mathbf{y}}_i^c = \sigma(\mathbf{z}_i/\mathbf{T}_i)$, where $\mathbf{T}_i = g_\phi(\mu_{Z_i}, \Sigma_{Z_i}, \nabla\hat{\mathbf{y}}_i, \mathbf{z}_i, \mathbf{x}_i)$, and $\sigma(\cdot)$ denotes the softmax layer.

In practice, we configure $g_\phi(\cdot)$ as a shallow residual network, which we empirically found to yield comparable results to the decision-tree-inspired network in the vanilla LTS [3], but to be more flexible in terms of model architecture. A channel attention layer is used in $g_\phi(\cdot)$ to allow the network to adaptively weigh information from different sources.

Training objectives: Following the standard setting of post-hoc calibration, both the calibration network $g_\phi(\cdot)$ and the shape prior module $s_\psi(\cdot)$ are trained on the validation dataset used in building the segmentation network $f_\theta(\cdot)$. To avoid shortcut learning from $s_\psi(\cdot)$ to $g_\phi(\cdot)$, two networks are trained one by one. We first train the shape prior module using the cross entropy loss:

$$\mathcal{L}_s(\psi) = -\frac{1}{MN}\sum_{m,n}\sum_{c}\mathbf{y}_i(c,m,n)\log\left(\sigma(s_\psi(\mathbf{z}_i(c,m,n)))\right), \tag{1}$$

where $\mathbf{z}_i = f_\theta(\mathbf{x}_i)$, $(\mathbf{x}_i, \mathbf{y}_i) \in \mathcal{D}_{val}$ the validation set, the subscript c denotes the class index. After $s_\psi(\cdot)$ is trained, we close the gradient computation for $s_\psi(\cdot)$. We then train the calibration network $g_\phi(\cdot)$ using the negative log likelihood loss that is commonly used for training post-hoc calibration networks [6,3,27]:

$$\mathcal{L}_g(\phi) = -\frac{1}{MN}\sum_{m,n}\sum_{c}\mathbf{y}_i(c,m,n)\log\left(\sigma(\mathbf{z}_i(c,m,n)/\mathbf{T}_i(c,m,n))\right), \tag{2}$$

where $\mathbf{T}_i = g_\phi(\mu_{Z_i}, \Sigma_{Z_i}, \nabla\hat{\mathbf{y}}_i, \mathbf{z}_i, \mathbf{x}_i)$, $(\mu_{Z_i}, \Sigma_{Z_i})$'s are obtained by sending multiple augmented versions of $\mathbf{x}_i$ to $f_\theta(\cdot)$. This loss penalizes over-confident erroneous

Table 1. Quantitative results on expected calibration error (ECE) and static calibration error (SCE). Lower the better. Average Dice scores of the segmentation networks are appended for reference.

Method	ECE [%] ↓							SCE [%] ↓						
	Intra-dom.	Bias field	Motion	Ghosting	Spike	Artifact Avg.	Cross Seq.	Intra-dom.	Bias field	Motion	Ghosting	Spike	Artifact Avg.	Cross Seq.
UC	10.29	13.60	22.38	19.68	39.05	23.67	30.29	5.27	6.96	11.45	10.09	19.82	12.08	15.58
Alea. [15,14]	7.74	9.06	16.47	16.78	37.31	19.90	28.50	5.08	8.39	10.11	10.56	21.15	12.55	16.89
TS [6]	10.06	13.31	22.04	19.42	38.87	23.41	29.96	5.17	6.84	11.31	9.99	19.76	11.98	15.47
LTS [3]	3.22	5.46	10.21	10.61	31.60	14.48	16.78	3.63	4.91	7.93	7.80	17.80	9.61	11.52
Proposed	**3.12** (-0.10)	**4.65*** (-0.82)	**8.88**‡ (-1.33)	**9.23*** (-1.38)	**28.35**† (-3.26)	**12.78** (-1.70)	**15.37**† (-1.41)	**3.38** (-0.25)	**4.75*** (-0.16)	**7.23**† (-0.70)	**7.14**‡ (-0.67)	**16.45**‡ (-1.35)	**8.89** (-0.72)	**10.77**† (-0.75)

†: p-value < 0.01; ‡: p-value < 0.05; *: p-value > 0.05, compared with the results of LTS [3].

	Dice score [%] ↑						
	Intra-dom.	Bias field	Motion	Ghosting	Spike	Artifact Avg.	Cross Seq.
Seg. Net.	85.14	80.29	69.02	79.73	39.02	67.02	62.74

predictions while it encourages high confidence for correct predictions. Although Eq. 2 has similar form as cross-entropy, it essentially optimizes over ϕ via $\mathbf{T}_i$. Since at each location (m, n), $\mathbf{T}_i(c, m, n)$'s remain constant for all the classes c's, this loss does not affect the categorical segmentation result [6,3].

4 Evaluation and Results

Table 2. Ablating key components and the number of test-time augmentations, evaluated on artifact-corrupted images.

Alea.	Shape	ECE [%] ↓	SCE [%] ↓	No. of Aug.	ECE [%] ↓	SCE [%] ↓
×	×	14.48	9.61	15	13.22	9.00
✓	×	13.03	9.20	45	12.94	8.94
×	✓	13.50	9.18	90	12.85	8.92
✓	✓	12.78	8.89	180	12.78	8.89

Dataset: *Training and validation dataset*: We employ the ACDC cardiac MRI segmentation dataset (bSSFP sequence) [29] for building the segmentation model and the proposed calibration model. Specifically, we take the ES fold of ACDC and split it into training, validation and (intra-domain) testing sets of 60/20/20 cases. To simulate data-hungry medical image segmentation [23], each time we take 20 cases out of the training data for building the segmentation network, and 5 out of validation data for validating the segmentation network and for training the calibration model. We repeat this process for 3 times to cover all the training samples, and obtain 3 segmentation models. For each segmentation model, we repeat training the calibration model for 3 times.
Artifact-corrupted testing dataset: Inspired by [23], we simulate common MRI artifacts: bias field, motion artifact, ghosting artifact and k-space spikes, separately, to the 20 intra-domain testing cases mentioned above, using TorchIO [30]. Using this controlled environment allows us to observe the model behaviors under each type of artifacts.
Cross-sequence testing dataset: We further test the above ACDC-based models on the 40 LGE MRI of the testing fold of the MS-CMRSeg challenge [31]. As

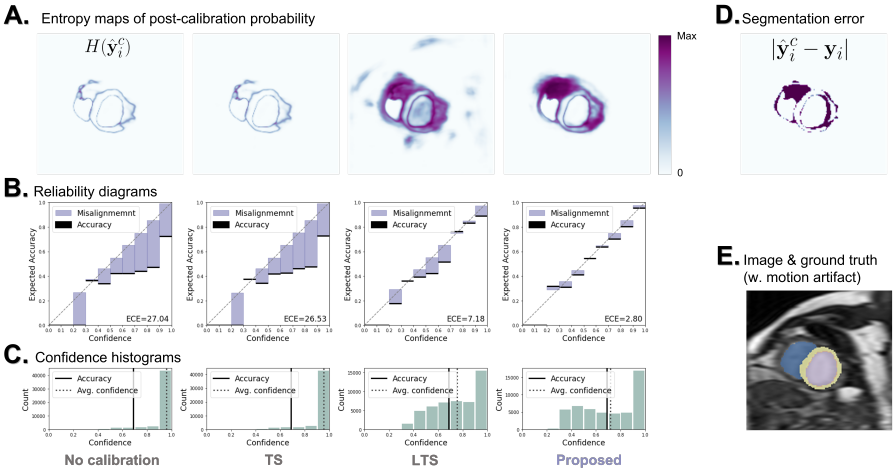

Fig. 2. A. For the proposed method, the entropy map which shows the doubt of the calibration model, agrees well with the actual segmentation error (shown in **D.**). **B.** Reliability map of the proposed method demostrates the least misalignment (purple bars) between confidence and accuracy. **C.** The confidence histogram shows that the proposed method has corrected over-confident predictions, compared with uncalibrated results. **E.** The motion-corrupted input image and its ground truth segmentation.

ACDC is based on bSSFP sequence, the segmentation and calibration models have never seen images from LGE sequence before testing.

Network architecture and training configurations: We employ a U-Net [32] as the segmentation network. For the calibration network $g_\phi(\cdot)$, we employ a shallow ResNet with 5 input branches for processing $\nabla\hat{\mathbf{y}}_i$, μ_{Z_i}, Σ_{Z_i}, $\mathbf{z}_i$, and $\mathbf{x}_i$ separately. These branches are merged by a channel attention block, followed by two ResNet blocks. The shape prior model $s_\psi(\cdot)$ is configured as a small U-Net with dropout (p=0.5) in its encoder. The Adam optimizer is used, with an initial learning rate of 1×10^{-3}, 800 epoches separately for $s_\psi(\cdot)$ and $g_\phi(\cdot)$. In each iteration, $(\mu_{Z_i}, \Sigma_{Z_i})$ are computed by repeating augmentations for 6 times.

Photometric transforms: brightness, contrast, gamma transform, random additive noises [23], and geometric transformations: affine transformation and elastic transformation are used as data augmentation for training the segmentation model and the calibration model (also for the LTS [3]). Importantly, these data augmentations *do not* include the corruptions in the testing data.

Quantitative and qualitative results: We employ commonly-used expected calibration error (ECE) [33] and static calibration error (SCE) [27] for evaluation (lower the better). Both of them measure the gap between prediction probability and the accuracy in test time, and the latter is a class-conditional version of the former. To account for the foreground-background class imbalance in ACDC, inspired by [16], these two metrics are computed over the region-of-interests

obtained by dilating (expanding) the ground truth segmentations with a kernel size of 10 pxiels.

As shown in Table 1, we compare the proposed method with the uncalibrated model (UC) and the state-of-the-art local temperature scaling (LTS) [3]. The proposed method demonstrates overall smaller calibration errors compared with LTS. Calibration errors of the estimated aleatoric uncertainty (Alea.) [14,15] and temperature scaling (TS) [6] are also presented. The segmentation performances measured in Dice scores of the segmentation networks are also attached.

We show in the first row of Fig. 2 the entropy maps $H(\hat{\mathbf{y}}_i^c)$'s of the calibrated probabilities, where higher values suggest stronger doubts by the calibration network. The entropy map produced by the proposed method has the best agreement with the actual segmentation error. We further show the reliability map in the second row, where the purple bars represent the gaps between confidence (x-axis) and accuracy (y-axis) at each confidence level. The proposed method also yields the smallest gaps. Confidence histograms of post-calibration probabilities are shown in the third row.

Ablation studies: We ablate the two key components of the proposed method: the susceptibility (aleatoric uncertainty) estimation and the shape prior. The results in Table 2 left show that the best performances are obtained when two components work together. We also ablate the number of repeated augmentations N_A used for estimating susceptibility during *test time*. As shown in Table 2 right, a larger N_A leads to more precise estimations, yielding less errors.

Conclusion: In this work we propose a new calibration method for out-of-domain MRI segmentation. Future works can be done by designing better shape prior models that can account for segmentation targets with more irregular shapes, like blood vessels and tumors.

Acknowledgments: This work was in part supported by EPSRC Programme Grants (EP/P001009/1, EP/W01842X/1) and in part by the UKRI London Medical Imaging and Artificial Intelligence Centre for Value Based Healthcare (No.104691). S.W. was also supported by the Shanghai Sailing Programs of Shanghai Municipal Science and Technology Committee (22YF1409300).

References

1. Nguyen, A., Yosinski, J., Clune, J.: Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In: Proceedings of the IEEE CVPR. (2015) 427–436
2. Gonzalez, C., Gotkowski, K., Bucher, A., Fischbach, R., Kaltenborn, I., Mukhopadhyay, A.: Detecting when pre-trained nnu-net models fail silently for covid-19 lung lesion segmentation. In: International Conference on MICCAI, Springer (2021) 304–314
3. Ding, Z., Han, X., Liu, P., Niethammer, M.: Local temperature scaling for probability calibration. In: Proceedings of the IEEE/CVF ICCV. (2021) 6889–6899
4. Platt, J., et al.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in large margin classifiers **10**(3) (1999) 61–74

5. Zadrozny, B., Elkan, C.: Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In: Icml. Volume 1., Citeseer (2001) 609–616
6. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: ICML, PMLR (2017) 1321–1330
7. Tomani, C., Buettner, F.: Towards trustworthy predictions from deep neural networks with fast adversarial calibration. In: Proceedings of the AAAI Conference. Volume 35. (2021) 9886–9896
8. Ji, B., Jung, H., Yoon, J., Kim, K., et al.: Bin-wise temperature scaling (bts): Improvement in confidence calibration performance through simple scaling techniques. In: 2019 IEEE/CVF ICCV Workshop, IEEE (2019) 4190–4196
9. Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., Snoek, J.: Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. Advances in NeurIPS **32** (2019)
10. Mukhoti, J., Kulharia, V., Sanyal, A., Golodetz, S., Torr, P., Dokania, P.: Calibrating deep neural networks using focal loss. Advances in NeurIPS **33** (2020) 15288–15299
11. Karimi, D., Gholipour, A.: Improving calibration and out-of-distribution detection in deep models for medical image segmentation. IEEE Trans. on Artificial Intelligence (2022)
12. Kireev, K., Andriushchenko, M., Flammarion, N.: On the effectiveness of adversarial training against common corruptions. arXiv preprint arXiv:2103.02325 (2021)
13. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: ICML, PMLR (2016) 1050–1059
14. Kendall, A., Gal, Y.: What uncertainties do we need in bayesian deep learning for computer vision? Advances in NIPS **30** (2017)
15. Wang, G., Li, W., Aertsen, M., Deprest, J., Ourselin, S., Vercauteren, T.: Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. Neurocomputing **338** (2019) 34–45
16. Mehrtash, A., Wells, W.M., Tempany, C.M., Abolmaesumi, P., Kapur, T.: Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. IEEE trans. on medical imaging **39**(12) (2020) 3868–3878
17. Baumgartner, C.F., Tezcan, K.C., Chaitanya, K., Hötker, A.M., Muehlematter, U.J., Schawkat, K., Becker, A.S., Donati, O., Konukoglu, E.: Phiseg: Capturing uncertainty in medical image segmentation. In: International Conference on MICCAI, Springer (2019) 119–127
18. Zhang, L., Wang, X., Yang, D., Sanford, T., Harmon, S., Turkbey, B., Wood, B.J., Roth, H., Myronenko, A., Xu, D., et al.: Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation. IEEE trans. on medical imaging **39**(7) (2020) 2531–2540
19. Chen, C., Qin, C., Qiu, H., Ouyang, C., Wang, S., Chen, L., Tarroni, G., Bai, W., Rueckert, D.: Realistic adversarial data augmentation for mr image segmentation. In: International Conference on MICCAI, Springer (2020) 667–677
20. Ouyang, C., Chen, C., Li, S., Li, Z., Qin, C., Bai, W., Rueckert, D.: Causality-inspired single-source domain generalization for medical image segmentation. arXiv preprint arXiv:2111.12525 (2021)
21. Larrazabal, A.J., Martínez, C., Glocker, B., Ferrante, E.: Post-dae: anatomically plausible segmentation via post-processing with denoising autoencoders. IEEE trans. on medical imaging **39**(12) (2020) 3813–3820
22. Liu, Q., Chen, C., Dou, Q., Heng, P.A.: Single-domain generalization in medical image segmentation via test-time adaptation from shape dictionary. (2022)

23. Chen, C., Hammernik, K., Ouyang, C., Qin, C., Bai, W., Rueckert, D.: Cooperative training and latent space data augmentation for robust medical image segmentation. In: International Conference on MICCAI, Springer (2021) 149–159
24. Robinson, R., Valindria, V.V., Bai, W., Suzuki, H., Matthews, P.M., Page, C., Rueckert, D., Glocker, B.: Automatic quality control of cardiac mri segmentation in large-scale population imaging. In: International Conference on MICCAI, Springer (2017) 720–727
25. Li, K., Yu, L., Heng, P.A.: Towards reliable cardiac image segmentation: Assessing image-level and pixel-level segmentation quality via self-reflective references. Medical Image Analysis **78** (2022) 102426
26. Wang, S., Tarroni, G., Qin, C., Mo, Y., Dai, C., Chen, C., Glocker, B., Guo, Y., Rueckert, D., Bai, W.: Deep generative model-based quality control for cardiac mri segmentation. In: International Conference on MICCAI, Springer (2020) 88–97
27. Nixon, J., Dusenberry, M.W., Zhang, L., Jerfel, G., Tran, D.: Measuring calibration in deep learning. In: CVPR Workshops. Volume 2. (2019)
28. Raju, A., Miao, S., Cheng, C., Lu, L., Han, M., Xiao, J., Liao, C., Huang, J., Harrison, A.P.: Deep implicit statistical shape models for 3d medical image delineation. arXiv (2021)
29. Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.A., Cetin, I., Lekadir, K., Camara, O., Ballester, M.A.G., et al.: Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? IEEE trans. on medical imaging **37**(11) (2018) 2514–2525
30. Pérez-García, F., Sparks, R., Ourselin, S.: Torchio: a python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. Computer Methods and Programs in Biomedicine **208** (2021)
31. Zhuang, X., Xu, J., Luo, X., Chen, C., Ouyang, C., Rueckert, D., Campello, V.M., Lekadir, K., Vesal, S., RaviKumar, N., et al.: Cardiac segmentation on late gadolinium enhancement mri: a benchmark study from multi-sequence cardiac mr segmentation challenge. Medical Image Analysis (2022) 102528
32. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on MICCAI, Springer (2015) 234–241
33. Naeini, M.P., Cooper, G., Hauskrecht, M.: Obtaining well calibrated probabilities using bayesian binning. In: Twenty-Ninth AAAI Conference. (2015)