2022 44th Annual International Conference of
the IEEE Engineering in Medicine & Biology Society (EMBC)
Scottish Event Campus, Glasgow, UK, July 11-15, 2022

Joint Segmentation and Uncertainty Estimation of Ventricular Structures from Cardiac MRI using a Bayesian CondenseUNet

S. M. Kamrul Hasan[1], Cristian A. Linte[1,2]
[1]Center for Imaging Science, [2]Biomedical Engineering
Rochester Institute of Technology, Rochester, NY
{sh3190, calbme}@rit.edu

***Abstract*— While convolutional neural networks (CNNs) have shown potential in segmenting cardiac structures from magnetic resonance (MR) images, their clinical applications still fall short of providing reliable cardiac segmentation. As a result, it is critical to quantify segmentation uncertainty in order to identify which segmentations might be troublesome. Moreover, quantifying uncertainty is critical in real-world scenarios, where input distributions are frequently moved from the training distribution due to sample bias and non-stationarity. Therefore, well-calibrated uncertainty estimates provide information on whether a model's output should (or should not) be trusted in such situations. In this work, we used a Bayesian version of our previously proposed *CondenseUNet* [1] framework featuring both a learned group structure and a regularized weight-pruner to reduce the computational cost in volumetric image segmentation and help quantify predictive uncertainty. Our study further showcases the potential of our deep-learning framework to evaluate the correlation between the uncertainty and the segmentation errors for a given model. The proposed model was trained and tested on the Automated Cardiac Diagnosis Challenge (ACDC) dataset featuring 150 cine cardiac MRI patient dataset for the segmentation and uncertainty estimation of the left ventricle (LV), right ventricle (RV), and myocardium (Myo) at end-diastole (ED) and end-systole (ES) phases.**

***Index Terms*— Cine MRI, learned group-convolution, ventricle segmentation, uncertainty, segmentation errors**

I. INTRODUCTION

Deep neural networks (DNNs) have been widely used in practice in light of their recent accomplishments in a variety of domains. As a result, the predictive capabilities of these DNN models have been increasingly used to make decisions in crucial applications ranging from machine-learning-assisted medical diagnostics [2] to self-driving cars. In addition to class predictions, such high-stake applications necessitate reliable quantification of predictive uncertainty, i.e. meaningful confidence levels.

For assessing predicted uncertainty in DNNs, a number of approaches, including Bayesian and non-Bayesian, have been proposed. Despite the fact that Bayesian neural networks (BNNs) provide a theoretical foundation for generating well-calibrated uncertainty estimates, learning BNNs is difficult due to the intractable nature of integrating over the posterior in high-dimensional space. As such, approximate inference approaches such as Monte Carlo (MC) Dropout [3], [4], Deep Ensembles [5] and techniques based on Learned Confidence [6] are becoming increasingly prominent.

Recent work by Sander et al. [4] used MC Dropout on a CNN for cardiac MRI segmentation, demonstrating that training with a Brier loss or cross-entropy loss yielded well-calibrated pixel-wise uncertainty, and that correcting uncertain pixels may consistently enhance segmentation outcomes.

In this work, we study predictive uncertainty estimation for semantic segmentation with fully convolution network (FCN) and propose a Bayesian dropout for reliable predictive uncertainty estimation of segmented cardiac structures. The network takes a 2D image as input and outputs an uncertainty map, and a segmentation map, as illustrated in Fig. 1.

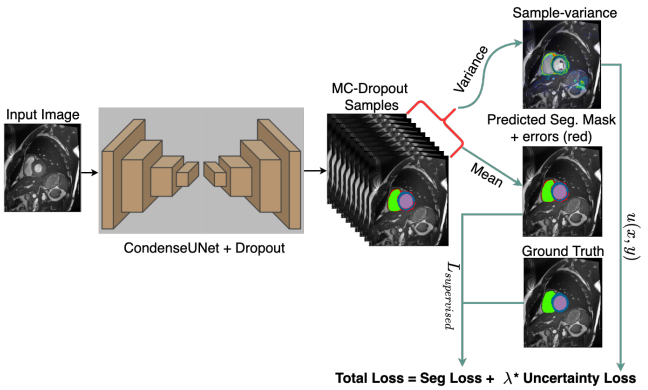

Fig. 1. System diagram for our proposed pipeline. A semantic segmentation network takes an input image and produces a segmentation prediction along with errors and an uncertainty map.

We computed a probability calibration to prove the concept that the generated probabilities represent the empirical probability of being correct due to the unavailability of human intervention in a timely manner. The overall goal of this work is to demonstrate how this method can be employed to evaluate uncertainty in cardiac MRI segmentation, to inform an expert whether and where the generated segmentation should be corrected and the extent to which it can be trusted.

II. METHOD

A. Imaging Data

For this study, we used the Automated Cardiac Diagnosis Challenge (ACDC) dataset, consisting of short-axis cardiac

Research reported in this publication was supported by the National Institute of General Medical Sciences Award No. R35GM128877 of the National Institutes of Health, and the Office of Advanced Cyber infrastructure Award No. 1808530 of the National Science Foundation.

cine-MR images acquired for 150 different patients divided into 5 evenly distributed subgroups according to their cardiac condition: normal- NOR; myocardial infarction- MINF; dilated cardiomyopathy- DCM; hypertrophic cardiomyopathy- HCM; and abnormal right ventricle- ARV; available as a part of the STACOM 2017 ACDC challenge [7]. The acquisitions were obtained over a 6 year period using two MRI scanners of different magnetic strengths (1.5T and 3.0T). The images were acquired using retrospective or prospective gating and the SSFP sequence with the following settings: thickness 5-8mm, inter-slice gap of 5 or 10mm, spatial resolution 1.37 to 1.68 mm2/pixel, 28 to 40 frames per cardiac cycle. The manual segmentation for RV blood-pool, LV myocardium, and LV blood-pool were performed by a clinical expert for the end-systole (ES) and end-diastole (ED) image frames. Since the slice thickness was large, ranged from 5 mm to 10 mm, we re-sampled the dataset to $1.4 \times 1.4\ mm^2$. The image intensity values are normalized pixel intensity values lie between 0 and 1 according to the 5th and 95th percentile.

B. Segmentation

The segmentation of the MR images is the first step towards extracting further information. In this study, we used our previously proposed *CondenseUNet* [1] segmentation method, which is both a modification of DenseNet, as well as a combination of CondenseNet [8] and U-Net [9]. The *CondenseUNet* framework substitutes the concept of both standard convolution and group convolution (G-Conv) with learned group convolution (LG-Conv). Our network learns the group convolution automatically during training through a multi-stage scheme. The capability of our network to learn the group structure allows multiple groups to re-use the same features via condensed connectivity. Moreover, the efficient weight-pruning methods we implemented lead to high computational savings without compromising segmentation accuracy [10].

As *CondenseUNet* is based on both U-Net and DenseNet, it comprises both a down-sampling and up-sampling path. The down-sampling path is similar to CondenseNet and the up-sampling path is comprised of transposed convolutions, condense blocks and skip-connections with a soft-max layer to generate the image mask. Concatenation in the skip-layer has been replaced by an element-wise addition operation to mitigate the problem of the feature-map explosion. We employ a number of layers per block as 2,3,4,5,4,3,2 with 32 initial feature maps, 3 max-pooling layers, a growth rate of k = 16, and condensation factor, C = 4. e. Softmax probabilities are calculated over the four tissue classes (LV, RV, MYO, background). By applying dropout after each convolutional layer during training and test time, the Monte Carlo dropout CondenseUNet approximates the probabilistic uncertainty similar to a Bayesian neural network from segmentation models. We construct 10 slightly different samples for each input, average the voxelwise probability over these samples to generate a final segmentation probability map, and then binarize this map to generate a final segmentation result for MC dropout CondenseUNet (MCOUNET) models.

The weights are updated during the back-propagation operation by minimizing the dual loss function, $\mathcal{L}_{Total}$ as mentioned in [10].

C. Segmentation Accuracy

We evaluated the performance of the segmentation using Dice similarity coefficient, and Hausdorff distance (HD). Given the set of all pixels in the image, set of foreground pixels by automated segmentation S_1^a, and the set of pixels for ground truth S_1^g, DICE score can be compared with [S_1^a, S_1^g]$\subseteq \Omega$, when a vector of ground truth labels T_1 and a vector of predicted labels P_1,

$$Dice(T_1, P_1) = \frac{2|T_1 \cap P_1|}{|T_1| + |P_1|} \tag{1}$$

DICE score will measure the similarity between two sets, T_1 and P_1 and $|T_1|$ denotes the cardinality of the set T_1 with the range of D(T_1,P_1) $\in$ [0,1].

Let, S_T and S_P, be surfaces (with N_T and N_P points, respectively) corresponding to two binary segmentation masks, T and P, respectively.

Hausdorff Distance (HD) is the maximum distance between two contours and is calculated as:

$$\text{HD} = \max\left(\max_{p \in S_T} d(p, S_P), \max_{q \in S_P} d(q, S_T)\right) \tag{2}$$

where,

$$d(p, S) = \min_{q \in S} d(p, q)$$

D. Uncertainty Quantification

In this work, we used the sample variance as the voxel-wise uncertainty measure, computed on a voxel-by-voxel basis. The metric is calculated as the variance of N Monte-Carlo prediction samples of a voxel (i.e. each voxel (x, y) has N softmax predictions $(p_1^{(x,y,c)} ... p_N^{(x,y,c)})$) over all classes of the MC probability maps. In Eq. 3, $u(x, y)$ is the sample variance of each voxel (x, y) of the image. The mean variance of softmax probabilities is computed as follows:

$$u(x,y) = \frac{1}{C}\sum_{c=1}^{C}\Big[\frac{1}{N-1}\sum_{n=1}^{N}\Big(p_n^{(x,y,c)} - \frac{1}{N}\sum_{n=1}^{N} p_n^{(x,y,c)}\Big)^2\Big], \tag{3}$$

where $p_n^{(x,y,c)}$ represents the softmax probability of the $c-$th class in the $n-$th time, C is the number of classes and N is the number of samples. We set the dropout rate to q = 0.1 and produce 10 MC samples. We employ dropout layers after every encoder and decoder block with a dropout rate to create a probabilistic encoder decoder network. By also using dropouts during testing, we obtain per voxel samples from the posterior distribution. The segmentation loss is the *Brier (B)* loss, which measures how closely the neural network segmentation probabilities represent the likelihood of being correct on a per-pixel basis by computing the mean squared error between the predicted and ground truth probabilities:

$$B_{seg} = \sum_{i}\sum_{c=0}^{c-1}\Big[p(\hat{y}_i = c) - p(y = c)\Big]^2, \tag{4}$$

where p denotes the probability for a specific voxel with corresponding reference label y_i for class c. Hence, the total loss is computed as a sum of the segmentation loss and uncertainty loss, $L_{Total} = B_{seg} + \lambda\ u(x, y)$ (**Fig. 1**).

E. Network Training and Testing

To solve the class-imbalance problem in multi-slice cardiac MR images, a patch of size 128×128 was extracted around the LV center from a full-sized cardiac MR and slice-wise normalization of voxel intensities was performed. The training dataset was divided into 70% training data, 15% validation data, and 15% testing data with five non-overlapping folds for cross-validation. Networks implemented in PyTorch were initialized with He normal initializer [11] and trained for 100,000 epochs with a batch size of 16. We used the Adam optimizer with a learning rate of 0.001 and decay rate of 0.1 after every 25,000 step. All experiments were performed on a workstation equipped with two NVIDIA GTX 1080 Ti GPU (11GBs of memory).

III. Results

A. Segmentation Evaluation

The proposed model was trained for the joint segmentation and uncertainty estimation of RV blood-pool, LV-Myocardium, and LV blood-pool from the ACDC challenge dataset. The provided reference segmentation and the corresponding automatic segmentation obtained from the MCOUNET model for a test patient is shown in **Fig. 2**.

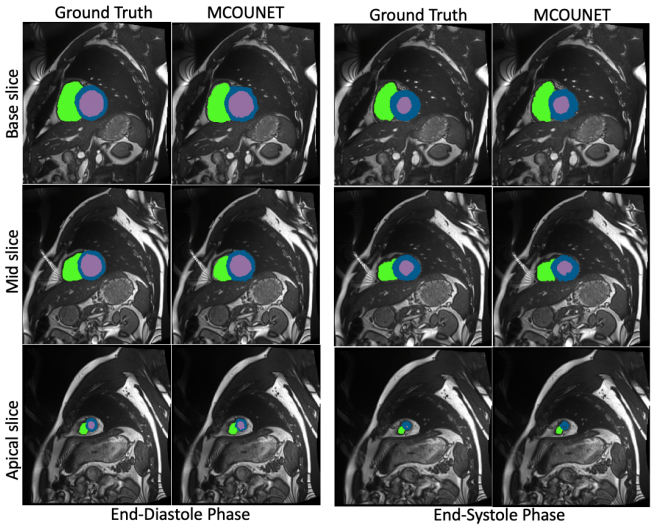

Fig. 2. Representative ED and ES frame segmentation results of a complete cardiac cycle from the base to apex showing RV blood-pool, LV blood-pool, and LV-Myocardium in green, purple, and blue respectively.

Automatic segmentation obtained from the model, for ED and ES phases, are evaluated against the reference segmentation and summarized in **Table I**; also shown as the graphical representation in **Fig. 3**. Our proposed model achieved Dice score (Std. Dev.) of 96.8%(0.01) and 95.1%(0.07) for the LV bloodpool, 89.5%(0.03) and 90.3%(0.03) for the LVMyocardium and 93.5%(0.02) and 88.3%(0.09) for the RV blood-pool in end-diastole and end-systole, respectively. Accordingly, the Hausdorff distance (Std. Dev.) for the LV bloodpool, 7.9mm (10.40) and 6.4mm (6.10) for the LV bloodpool, 8.9mm (8.92) and 9.1mm (10.17) for the LVMyocardium and 11.2mm (8.10) and 11.9mm (9.12) for the RV blood-pool in end-diastole and end-systole, respectively.

TABLE I

QUANTITATIVE EVALUATION OF THE SEGMENTATION RESULTS IN TERMS OF MEAN DICE SCORE (STD. DEV.) (%) AND HD - HAUSDORFF DISTANCE (STD. DEV.) (MM), ON THE ACDC DATASET FOR LV, RV BLOOD-POOL AND LV-MYOCARDIUM.

	End-Diastole (ED)			End-Systole (ES)		
	LV	MYO	RV	LV	MYO	RV
Dice	96.8 (0.01)	89.5 (0.03)	93.5 (0.02)	95.1 (0.07)	90.3 (0.03)	88.3 (0.09)
HD	7.9 (10.40)	8.9 (8.92)	11.2 (8.10)	6.4 (6.10)	9.1 (10.17)	11.9 (9.12)

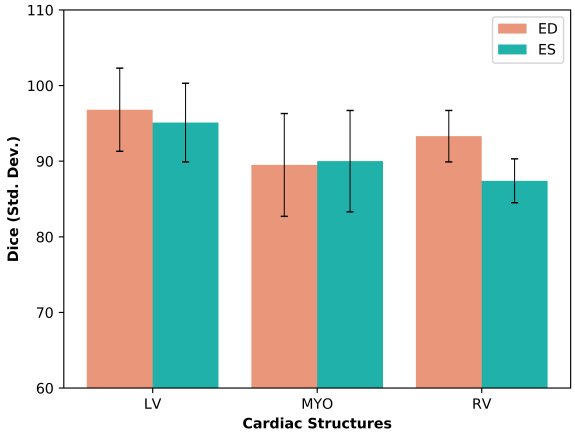

Fig. 3. Mean Dice score (%) and Hausdorff distance (mm) for LV blood-pool, LV myocardium, and RV blood-pool segmentation achieved on images from the ACDC dataset.

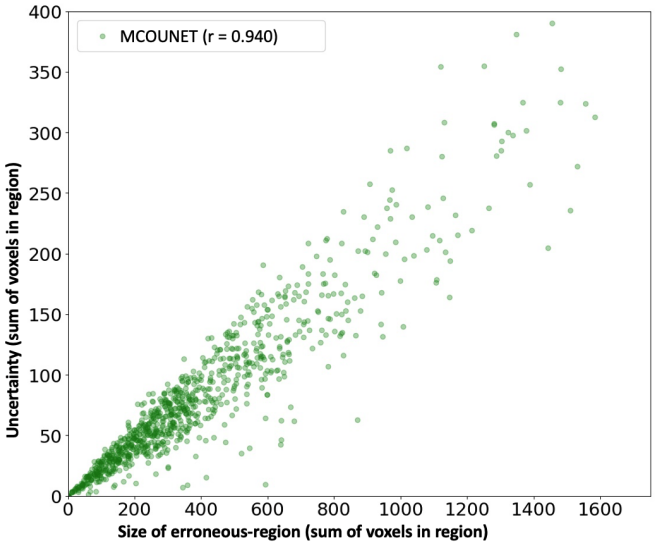

Fig. 4. Correlation between the segmentation error and model-predicted uncertainty.

B. Segmentation Uncertainty

Theoretically, incorrectly segmented voxels should be covered by higher uncertainty than correctly segmented voxels. The spatial uncertainty maps are perfectly calibrated in this scenario. **Fig. 4** illustrates the correlation between the erroneous pixels and the uncertainty. The error is calculated by finding the difference between the reference mask and the predicted mask. The computed correlation value of r=0.94 indicates that there is a very good correlation between the error and the uncertainty in terms of pixels. The qualitative uncertainty maps from our proposed model for both the ED and ES phases are visualized in **Fig. 5**.

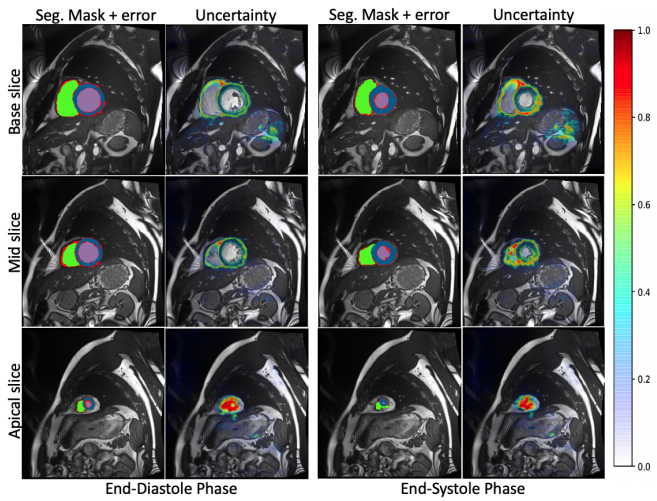

Fig. 5. Representative uncertainty maps (red areas correspond to higher uncertainty as shown in the color bar) of a cardiac cycle in ED and ES phase from the base to apex showing RV blood-pool (green), LV blood-pool (cyan), LV-Myocardium (blue), and segmentation errors (red). The first column shows SSFP cine cardiac MR images. The second column shows the MRI overlaid with segmentation predictions and errors (red) of U-Net architecture. The third column shows the errors in predictions of our model trained with our custom loss. The last column shows the Bayesian uncertainty maps for the Brier score.

As seen from Fig. 5, our model-predicted uncertainty maps closely match the regions where the segmentation algorithm under-performs compared to the ground truth. As such, these predictive maps show lower uncertainty in the periphery of the LV blood pool an LV myocardium, and higher uncertainty (on the order of 80%) close to the periphery of the RV blood pool. Similarly, these regions also show the greatest discrepancies between the proposed and ground truth segmentation masks.

One benefit of the uncertainty maps is their behavior in the regions featuring poor segmentation. The panels in columns 1 and 3 of Fig. 5 show the proposed and ground truth segmentation masks overlaid onto the ED and ES images slice, while columns 2 and 4 illustrate the segmentation uncertainties. These panels, when visualized side-by-side clearly show how that Bayesian uncertainty maps are highly indicative of the poorly segmented regions, confirming the 94% correlation between the erroneously segmented regions and the cumulative segmentation uncertainty regions shown in Fig. 4. Hence, these uncertainty maps are key to raising awareness and caution about the reliability of the segmentation at various locations.

IV. Conclusion

In this paper, we propose a segmentation pipeline that integrates a Monte Carlo dropout CondenseUNet model with inherent uncertainty estimation, with the overall goal to study the uncertainty associated with the obtained segmentations and errors, as a means to flag regions that feature less than optimal segmentation results. This overall pipeline will increase the reliability of automatic segmentation for both research and clinical use.

Our future research will explore the use of uncertainty measures to flag low-quality segmentation for automatic detection using a deep neural network in place of human review to detect and correct the low-quality segmentation maps.

References

[1] SM Kamrul Hasan and Cristian A Linte. CondenseUNet: A memory-efficient condensely-connected architecture for bi-ventricular blood pool and myocardium segmentation. In *Medical Imaging 2020: Image-Guided Procedures, Robotic Interventions, and Modeling*, volume 11315, page 113151J. International Society for Optics and Photonics, 2020.

[2] S M Kamrul Hasan and Cristian A Linte. A multi-task cross-task learning architecture for ad hoc uncertainty estimation in 3D cardiac MRI image segmentation. In *2021 Computing in Cardiology (CinC)*, volume 48, pages 1–4, 2021.

[3] Abhijit Guha Roy, Sailesh Conjeti, Nassir Navab, and Christian Wachinger. Inherent brain segmentation quality control from fully convnet Monte Carlo sampling. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 664–672. Springer, 2018.

[4] Jörg Sander, Bob D de Vos, Jelmer M Wolterink, and Ivana Išgum. Towards increased trustworthiness of deep learning segmentation methods on cardiac MRI. In *Medical Imaging 2019: Image Processing*, volume 10949, page 1094919. International Society for Optics and Photonics, 2019.

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

[6] Terrance DeVries and Graham W Taylor. Learning confidence for out-of-distribution detection in neural networks. *arXiv preprint arXiv:1802.04865*, 2018.

[7] Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al. Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved? *IEEE Transactions on Medical Imaging*, 37(11):2514–2525, 2018.

[8] Gao Huang, Shichen Liu, Laurens Van der Maaten, and Kilian Q Weinberger. CondenseNet: An efficient densenet using learned group convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2752–2761, 2018.

[9] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015.

[10] S M Kamrul Hasan and Cristian A Linte. L-CO-Net: Learned condensation-optimization network for segmentation and clinical parameter estimation from cardiac cine MRI. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 1217–1220. IEEE, 2020.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1026–1034, 2015.