Calibration of cine MRI segmentation probability for uncertainty estimation using a multi-task cross-task learning architecture

S. M. Kamrul Hasan[a, b] (✉) and Cristian A. Linte[a,b,c]

[a]Biomedical Modeling, Visualization and Image-guided Navigation (BiMVisIGN) Lab, RIT
[b]Center for Imaging Science, Rochester Institute of Technology, NY, USA
[c]Biomedical Engineering, Rochester Institute of Technology, NY, USA

ABSTRACT

While deep learning has shown potential in solving a variety of medical image analysis problems including segmentation, registration, motion estimation, etc., their applications in the real-world clinical setting are still not affluent due to the lack of reliability caused by the failures of deep learning models in prediction. Furthermore, deep learning models need a large number of labeled datasets. In this work, we propose a novel method that incorporates uncertainty estimation to detect failures in the segmentation masks generated by CNNs. Our study further showcases the potential of our model to evaluate the correlation between the uncertainty and the segmentation errors for a given model. Furthermore, we introduce a multi-task cross-task learning consistency approach to enforce the correlation between the pixel-level (segmentation) and the geometric-level (distance map) tasks. Our extensive experimentation with varied quantities of labeled data in the training sets justifies the effectiveness of our model 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 from cine MRI images available through the MICCAI 2017 ACDC Challenge Dataset. Our study serves as a proof-of-concept of how uncertainty measure correlates with the erroneous segmentation generated by different deep learning models, further showcasing the potential of our model to flag low-quality segmentation from a given model in our future study.

Keywords: Bayesian multi-task cross-task learning, Monte-Carlo sampling, uncertainty, error estimation, cine MR image, cardiac imaging, deep learning, image segmentation, ventricle blood-pool, myocardium

1. INTRODUCTION

Cardiac Magnetic Resonance Imaging (CMRI) has made a significant paradigm shift in medical imaging through the quantification of volumetric changes in the heart during the cardiac cycle, thanks to its capability of imaging different structures within the heart without ionizing radiation. Cine MRI can capture the full cardiac dynamics via multiple short-axis acquisitions. In today's clinical routine, though the manual delineation is the standard image segmentation approach, the huge benefits of these comprehensive measurements are still not exploited due to both the inter-and intra-user segmentation biases, as well as time inefficiency, suggesting the need for desirable automatic approaches for simultaneous multi-structure segmentation (LV, RV, Myo).

Recently, convolutional neural networks (CNNs) have shown emerging success in solving high-level computer vision tasks to develop machine learning tools that are capable of learning hierarchical features in an end-to-end manner.[1,2] Motivated by the superior performance of deep learning, the medical imaging community has also embraced the implementation of deep learning-based approaches for medical image segmentation.[3,4] However, a major challenge in adopting automated medical image segmentation in a clinical workflow is the lack of reliability and trustworthiness.

(✉) Code, pretrained models, and additional details are available at https://github.com/smkamrulhasan/BMTCTL.

Further author information:
S. M. Kamrul Hasan (E-mail: sh3190@rit.edu)
Cristian A. Linte (E-mail: calbme@rit.edu)

Medical Imaging 2022: Image-Guided Procedures, Robotic Interventions, and Modeling,
edited by Cristian A. Linte, Jeffrey H. Siewerdsen, Proc. of SPIE Vol. 12034, 120340T
 · 1605-7422 · doi: 10.1117/12.2612269

To date, most of these studies have been centered solely on automatic segmentation and there have only been very few research endeavors exploring the ambiguous predictions in some challenging regions generated by the deep learning models, increasing the model's uncertainty. An efficient method that can accurately identify the problematic segmentation generated by the models with the overall goal to avoid the review of all images and reducing errors in downstream analysis would be a great asset.

To date, a number of approaches have attempted to estimate uncertainty in CNNs for medical image segmentation including Monte Carlo (MC) Dropout,[5,6] Deep Ensembles[7] and techniques based on Learned Confidence.[8] Recent work by Wang et al.[9] observed positive correlations between segmentation accuracy and uncertainty measures. Heo *et al.*[10] proposed a method that allows the attention model to leverage uncertainty for the improvement of both model calibration as well as performance. However, many of these successes are achieved at the cost of a large pool of labeled datasets. Obtaining labeled images however is laborious as well as costly, impeding the adoption of large-scale deep learning models in clinical settings. To address the problem of limited access to labeled data, semi-supervised learning (SSL)[11] has been a growing trend for improving the deep learning model performance by utilizing unlabeled data. Furthermore, multi-task learning (MTL)[12] techniques have shown promising results for improving the generalizability of any models by jointly tackling multiple tasks through shared representation learning.[13] Although these methods were successful for cardiac segmentation and uncertainty estimation, the estimation of uncertainty calibration in a semi-supervised setting for medical image segmentation is still rarely reported.

We propose a novel semi-supervised module exploiting adversarial learning and task-based consistency regularization for jointly learning multiple tasks in a single backbone module – uncertainty estimation, geometric shape generation, and cardiac anatomical structure segmentation, illustrated in Figure 1. The network takes a 2D image as input and outputs an uncertainty map, a 2D distance map, and a segmentation map. The distance map is fed to a transformer to produce a segmentation map which is then used to share the supervisory signal from the predicted segmentation map. To leverage the unlabeled data, the distance map is fed to an adversarial discriminator network to distinguish the predicted distance map from the labeled data. The same encoder backbone is used to estimate the Bayesian uncertainty map by Bayesian Neural network with Monte Carlo (MC) sampling. As a departure from the recent work by Sander *et al.*,[6] 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 adjusted.

2. METHOD

We define the learning task as follows: given an (unknown) data distribution $p(x, y)$ over images and segmentation masks, we define a source domain having a training set, $\mathcal{D}_{\mathcal{L}} = \{(x_1^l, y_1), ..., (x_n^l, y_n\}$ with n labeled data and another domain having a training set, $\mathcal{D}_{\mathcal{UL}} = \{x_1^{ul}, ..., x_m^{ul}\}$ with m unlabeled data which are sampled i.i.d. from $p(x, y)$ and $p(x)$ distribution. Empirically, we want to minimize the target risk $\in_t (\phi, \theta) = min_{\phi,\theta} \mathcal{L}_{\mathcal{L}}(\mathcal{D}_{\mathcal{L}}, (\phi, \theta)) + \gamma \mathcal{L}_{\mathcal{UL}}(\mathcal{D}_{\mathcal{UL}}, (\phi, \theta))$, where $\mathcal{L}_{\mathcal{L}}$ is the supervised loss for segmentation, $\mathcal{L}_{\mathcal{UL}}$ is unsupervised loss defined on unlabeled images and ϕ and θ denote the learnable parameters of the overall network.

In this work, our architecture is composed of a shared encoder e and a main decoder d, which constitute the segmentation network $f = d \circ e$. We introduce a set of J auxiliary decoders d_a^j, with $j \in [1, J]$.

The overall objective function consists of different loss functions including distance loss, cross-task loss, adversarial loss, dice loss, and guidance loss. Our goal is to infer the *posterior* distribution $p(w|\mathcal{D})$ over the weights, instead of optimizing maximum likelihood using a Bayesian neural network (BNN). This posterior distribution represents uncertainty in the weights, which could be propagated to calculate uncertainty in the predictions. Unfortunately, the posterior probability distribution cannot be evaluated in closed form for neural networks, so one must resort to approximate inference based on variational inference[14] methods and stochastic regularization techniques using dropouts with an aim to find a surrogate distribution $q(w)$ by minimizing the *Kullback-Leibler (KL)* divergence between the approximate and the posterior probability distribution which is equivalent to maximizing the evidence lower bound (ELBO) as follows:

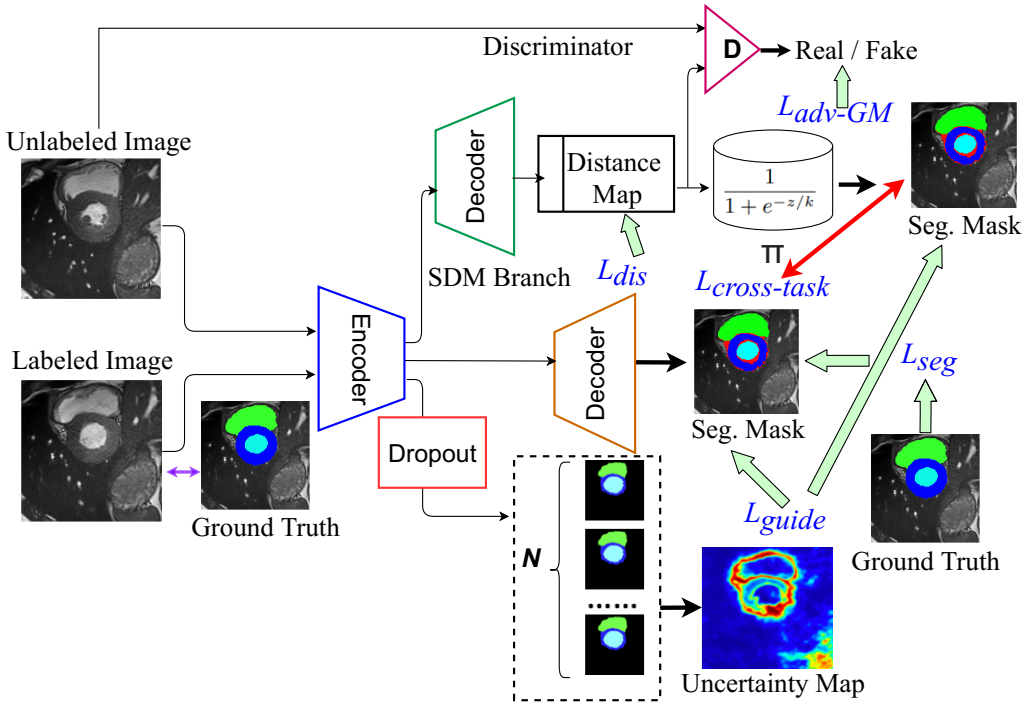

Figure 1: Schematic of the *BMT-CTL* model: we combine segmentation and uncertainty decoder who share the same backbone encoder – Deep Bayesian Neural Network.

$$\mathbb{E}_{q(w)}[\log p(Y|X,w)] - KL[q(w)||p(w)], \tag{1}$$

where $\mathbb{E}_{q(w)}[\cdot]$ denotes expectation over the approximate posterior $q(w)$, $\log p(Y|X,w)$ is the log-likelihood of the training data with given weights w, $p(w)$ represents the prior distribution of w, and $KL[\cdot]$ is the Kulback-Leibler divergence between two probability distributions.

2.1 UNCERTAINTY QUANTIFICATION

The uncertainty map is obtained by computing the maximum softmax probabilities with a number of samples N per voxel over all classes over the MC probability maps. The mean standard deviation of softmax probabilities are computed as follows:

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

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 sample. 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 $q(w)$. To increase the reliability of the segmentation, we have calibrated the probability based on *Brier score (BS)* 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:

$$BS = \sum_{i}\sum_{c=0}^{c-1}\Big[p(\hat{y}_i = c) - p(y = c)\Big]^2, \tag{3}$$

where p denotes the probability for a specific voxel with corresponding reference label y_i for class c.

2.2 Cardiac MRI Data

For this study, we used the Automated Cardiac Diagnosis Challenge (ACDC) dataset*, consisting of short-axis cardiac cine-MR images acquired for 100 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.[15] 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 a 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). Since the slice thickness was large and 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 such that the pixel values lie in between 0 and 1 according to the 5th and 95th percentile.

2.3 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 were performed. The training dataset was divided into 70% training data, 15% validation data, and 15% testing data

*https://www.creatis.insa-lyon.fr/Challenge/acdc/databases.h

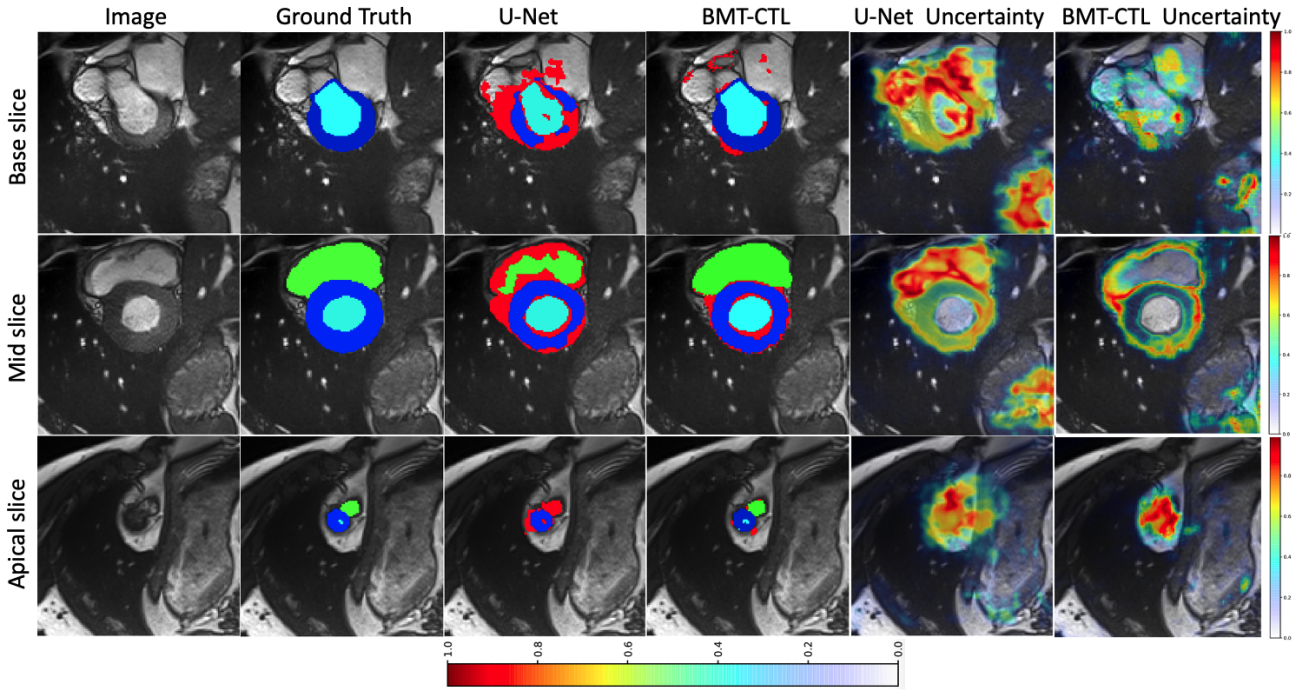

Figure 2: Representative segmentation results and uncertainty maps (red areas correspond to higher uncertainty as shown in the color bar) of a cardiac cycle from the base (top row) to apex (bottom row) showing RV blood-pool (green), LV blood-pool (cyan), LV-Myocardium (blue), and segmentation errors (red). The first column shows SSFP cine cardiac MR images. The second column shows ventricular structures of heart annotated by experts. The third column shows the MRI overlaid with segmentation predictions and errors (red) of U-Net architecture. The fourth column shows the segmentation predictions of our Bayesian BMT-CTL network trained with our custom loss. The fifth and sixth column show the Bayesian uncertainty maps for the Brier score.

with five non-overlapping folds for cross-validation. Networks implemented in PyTorch were initialized with He normal initializer[16] and trained for 100k 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 25k step. All experiments were run on a workstation equipped with two NVIDIA GTX 1080 Ti GPU (11GBs of memory).

3. RESULTS

Figure 2 shows a qualitative comparison of the segmentation, generated segmentation errors, and uncertainty maps, illustrating that our proposed model significantly improved the segmentation as well as the uncertainty estimation against the classical U-Net model. Upon visual assessment, the uncertainty maps of the U-Net model show high uncertainty in the periphery of the LV and LV-Myocardium and a larger area of high uncertainty in the RV blood pool region, whereas the uncertainty maps derived from our model have a low uncertainty gradient at the margins. Images in the third and fourth column visualize the segmentation errors (red) for the U-Net and BMT-CTL models respectively. We can observe from the error map (fourth column) as well as the uncertainty map (sixth column) that the estimated errors are accurately captured by the Bayesian uncertainty maps i.e. the errors are prominent on base and apical slices especially in the RV regions. For instance, U-Net has prominent red pixels in the regions where there are no actual RV regions segmented in the ground truth and this trend is also consistent with the information portrayed in the uncertainty maps. The redish color in the uncertainty map of U-Net model denotes higher uncertainty which is also visible in the U-Net segmentation errors regions. On the other hand, our proposed BMT-CTL model shows significantly less segmentation error around the LV boundary. Both the mid and apical slices exhibit similar effects.

4. CONCLUSION

In this paper, we propose a new paradigm for accurate LV, RV blood-pool, and LV-myocardium segmentation associated with uncertainty estimation from cine cardiac MR images by introducing a multi-task cross-task learning consistency approach to enforce the correlation between the pixel-level (segmentation) and the geometric-level (distance map) tasks. We have assessed the relationship between the uncertainty distribution and the size of the erroneous region by computing the correlation. We present model uncertainty estimation derived from a novel Bayesian multi-task cross-task learning model for the task of cardiac ventricle segmentation. Our focus is not to achieve state-of-the-art results on the segmentation tasks, but to exploit uncertainty measures to flag regions exhibiting sub-optimal segmentation. 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.

ACKNOWLEDGMENTS

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

REFERENCES

[1] Kirillov, A., Girshick, R., He, K., and Dollár, P., "Panoptic feature pyramid networks," in [*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*], 6399–6408 (2019).

[2] Hasan, S. M. K. and Linte, C. A., "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)*], 1217–1220, IEEE (2020).

[3] Ronneberger, O., Fischer, P., and Brox, T., "U-Net: Convolutional networks for biomedical image segmentation," in [*International Conference on Medical Image Computing and Computer-Assisted Intervention*], 234–241, Springer (2015).

[4] Hasan, S. K. and Linte, C. A., "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*], **11315**, 113151J, International Society for Optics and Photonics (2020).

[5] Roy, A. G., Conjeti, S., Navab, N., and Wachinger, C., "Inherent brain segmentation quality control from fully convnet Monte Carlo sampling," in [*International Conference on Medical Image Computing and Computer-Assisted Intervention*], 664–672, Springer (2018).

[6] Sander, J., de Vos, B. D., Wolterink, J. M., and Išgum, I., "Towards increased trustworthiness of deep learning segmentation methods on cardiac MRI," in [*Medical Imaging 2019: Image Processing*], **10949**, 1094919, International Society for Optics and Photonics (2019).

[7] Lakshminarayanan, B., Pritzel, A., and Blundell, C., "Simple and scalable predictive uncertainty estimation using deep ensembles," *Advances in neural information processing systems* **30** (2017).

[8] DeVries, T. and Taylor, G. W., "Learning confidence for out-of-distribution detection in neural networks," *arXiv preprint arXiv:1802.04865* (2018).

[9] Wang, G., Li, W., Ourselin, S., and Vercauteren, T., "Automatic brain tumor segmentation based on cascaded convolutional neural networks with uncertainty estimation," *Frontiers in Computational Neuroscience* **13**, 56 (2019).

[10] Heo, J., Lee, H. B., Kim, S., Lee, J., Kim, K. J., Yang, E., and Hwang, S. J., "Uncertainty-aware attention for reliable interpretation and prediction," *arXiv preprint arXiv:1805.09653* (2018).

[11] Ouali, Y., Hudelot, C., and Tami, M., "Semi-supervised semantic segmentation with cross-consistency training," in [*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*], 12674–12684 (2020).

[12] Caruana, R., "Multitask learning," *Machine Learning* **28**(1), 41–75 (1997).

[13] Zhang, Y., Wei, Y., and Yang, Q., "Learning to multitask," *arXiv preprint arXiv:1805.07541* (2018).

[14] Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J., "Stochastic variational inference.," *Journal of Machine Learning Research* **14**(5) (2013).

[15] 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 Transactions on Medical Imaging* **37**(11), 2514–2525 (2018).

[16] He, K., Zhang, X., Ren, S., and Sun, J., "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in [*Proceedings of the IEEE International Conference on Computer Vision*], 1026–1034 (2015).