Received: 14 August 2019 | Revised: 23 September 2019 | Accepted: 24 September 2019

DOI: 10.1002/mrm.28043

FULL PAPER

Magnetic Resonance in Medicine

Accuracy, uncertainty, and adaptability of automatic myocardial ASL segmentation using deep CNN*

Hung P. Do[1] | **Yi Guo**[1] | **Andrew J. Yoon**[2] | **Krishna S. Nayak**[1]

[1]Ming Hsieh Department of Electrical and Computer Engineering, Viterbi School of Engineering, University of Southern California, Los Angeles, California

[2]Long Beach Memorial Medical Center, University of California Irvine, Irvine, California

Correspondence
Hung P. Do, Canon Medical Systems USA, 2441 Michelle Drive, Tustin, CA 92780.
Email: hdo@us.medical.canon

Funding information
National Institutes of Health (R01HL130494-01A1) and the Whittier Foundation (0003457-00001).

Purpose: To apply deep convolution neural network to the segmentation task in myocardial arterial spin labeled perfusion imaging and to develop methods that measure uncertainty and that adapt the convolution neural network model to a specific false-positive versus false-negative tradeoff.

Methods: The Monte Carlo dropout U-Net was trained on data from 22 subjects and tested on data from 6 heart transplant recipients. Manual segmentation and regional myocardial blood flow were available for comparison. We consider 2 global uncertainty measures, named "Dice uncertainty" and "Monte Carlo dropout uncertainty," which were calculated with and without the use of manual segmentation, respectively. Tversky loss function with a hyperparameter β was used to adapt the model to a specific false-positive versus false-negative tradeoff.

Results: The Monte Carlo dropout U-Net achieved a Dice coefficient of 0.91 ± 0.04 on the test set. Myocardial blood flow measured using automatic segmentations was highly correlated to that measured using the manual segmentation ($R^2 = 0.96$). Dice uncertainty and Monte Carlo dropout uncertainty were in good agreement ($R^2 = 0.64$). As β increased, the false-positive rate systematically decreased and false-negative rate systematically increased.

Conclusion: We demonstrate the feasibility of deep convolution neural network for automatic segmentation of myocardial arterial spin labeling, with good accuracy. We also introduce 2 simple methods for assessing model uncertainty. Finally, we demonstrate the ability to adapt the convolution neural network model to a specific false-positive versus false-negative tradeoff. These findings are directly relevant to automatic segmentation in quantitative cardiac MRI and are broadly applicable to automatic segmentation problems in diagnostic imaging.

KEYWORDS

automatic segmentation, Bayesian, deep convolutional neural network, false-positive and false-negative tradeoff, Monte Carlo dropout, uncertainty measure

*A preliminary version of this manuscript was presented as an oral presentation at the ISMRM Workshop on Machine Learning Part II in Washington DC, October 25th–28th, 2018.

1 | INTRODUCTION

Myocardial arterial spin labeling (ASL) is a noncontrast quantitative perfusion technique that can assess coronary artery disease.[1] Manual segmentation of left ventricular myocardium is a required step in the postprocessing pipeline and is a major bottleneck due to the low and inconsistent SNR and blood-myocardium contrast-to-noise ratio (CNR) in the source images. More generally, segmentation of left ventricular myocardium is a key step in the postprocessing pipeline of all quantitative myocardial imaging. Segmentation masks are needed to make volumetric measurements, to provide semantic delineation of different tissues (e.g., myocardium vs. blood vs. epicardial fat), and in many cases to map measurements to a bullseye plot for convenient visualization.[2]

Convolutional neural networks (CNNs) have been applied successfully to automatic segmentation in several MRI applications.[3-6] For example, Bai et al recently demonstrated that CNN can provide a performance on par with human experts in analyzing cine cardiovascular magnetic resonance (CMR) data.[5] Cine CMR data typically have high spatial temporal resolution, excellent SNR, and consistent blood–myocardium CNR throughout the cardiac cycle, which is why cine CMR is the gold standard for assessment of ventricular function, volumes, mass, and ejection fraction.[7] In contrast to cine CMR, quantitative myocardial CMR techniques (e.g., myocardial ASL, myocardial BOLD, myocardial first-pass perfusion, multiparametric myocardial relaxometry, myocardial DTI)[8-10] often consist of images with substantially lower spatial resolution, SNR, and CNR, which can vary between images due to factors such as variability in contrast preparation or heart rate. These are all reasons why automatic segmentation in quantitative CMR remains a significant challenge.

Automatic segmentation has been developed and used for many years,[11,12] but the postprocessing pipeline remains semi-automatic, as a human operator is required to verify segmentation mask quality before commencing to the next step in the pipeline. A global score of model uncertainty is therefore desired for automatic quality control at production, for model improvement through active learning[13] and for out-of-distribution detection. Model uncertainty can be estimated using a Bayesian approach, in which not only parameters are estimated but also their posterior distributions. Dropout has been demonstrated as a Bayesian approximation, which provides model uncertainty through Monte Carlo (MC) dropout at test time.[14-16] Monte Carlo dropout has been used to measure model uncertainty in many segmentation problems.[17-20] These studies demonstrate that pixel-wise uncertainty maps can be achieved using MC dropout (MCD) at test time, which allows qualitative assessment of predicted segmentations. A global quantitative score for model uncertainty, however, is desirable for automating the quality assessment, which may enable the automatic postprocessing pipeline.

For quantitative CMR, the America Heart Association 17-segment model[2] is often used in the form of a bullseye plot for visualization and diagnosis. To generate the bullseye plot, segmentation of the left ventricular myocardium is required. Left ventricular myocardium is surrounded by ventricular blood pools and epicardial fat, which have distinct physical properties as well as spin (magnetization) history.[21] Careful and conservative manual segmentation of myocardium is often required to minimize partial volume effects.[22-25] Therefore, a model with a lower false-positive rate may be preferred over that with a higher false-positive rate. Note that false positive means pixels predicted by an algorithm that are not present in the reference segmentation (i.e., excessive segmentation).

This study aimed to (1) apply deep CNN for automatic segmentation of myocardial ASL data, which have low and inconsistent SNR and CNR, (2) measure a global score of model uncertainty without the use of the reference segmentation using MCD, and (3) adapt the network to the specific false-positive and false-negative needs of the application using Tversky loss function. Additionally, model uncertainty was calculated using the probabilistic U-Net[26] and then compared with that calculated using MCD U-Net.

2 | METHODS

2.1 | Network architecture

We implemented a CNN model based on the U-Net architecture[27] called the MCD U-Net with the following modifications: (1) increased filter size from 3×3 to 5×5, (2) added batch normalization[28] after every convolutional layer, and (3) added dropout[15] with dropout rate of 50% at the end of every resolution scale (Figure 1). Similar to the original U-Net architecture, the number of base feature maps per convolutional layer in the first resolution scale was 64, which was doubled and halved in the next resolution scale in the encoding path and the decoding path, respectively. The MCD U-Net was implemented in Keras[29] with TensorFlow[30] backend.

2.2 | Data set

Our data set included 478 ASL images (control and labeled images) from 22 subjects. These were randomly divided into training and validation sets of 438 and 40 images, respectively. Trained networks were tested on 144 unseen ASL images acquired at rest and during adenosine stress from 6 heart transplant recipients. Training, validation, and testing data were drawn from previously published studies,[31-33] in which manual segmentation and quantitative reginal myocardial blood flow (MBF) from all subjects were readily available.

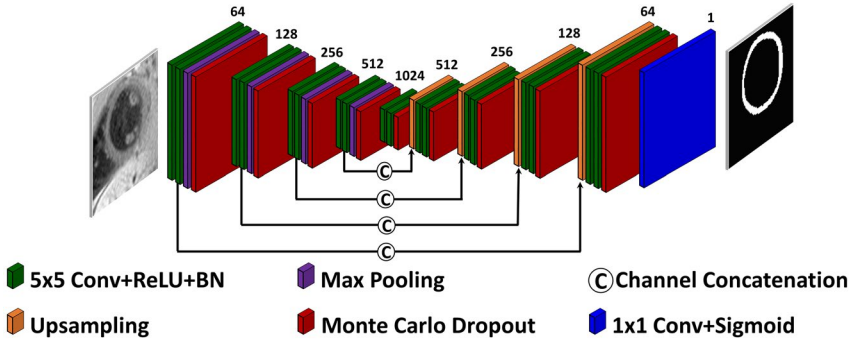

FIGURE 1 The network architecture of Monte Carlo dropout (MCD) U-Net. The modifications are that (1) 5 × 5 filter was used for all convolutional layers; (2) batch normalization was added after each convolutional layer; and (3) dropout rate of 50% was added after each convolutional scale of the down-sampled and up-sampled paths. The number of convolutional kernels is noted on top of the convolutional layers. Abbreviations: BN, batch normalization; Conv, convolutional layer; ReLU, rectified linear unit

The ASL acquisition details are summarized here, and a complete description can be found in Do et al.[31] Each ASL data set consists of 6 pairs of control and labeled images. Each pair of control and labeled images was acquired in a 12-second breath-hold using balanced SSFP (bSSFP) image acquisition preceded by a flow-sensitive alternating inversion recovery[34,35] control (slab-selective inversion) and labeled (nonselective inversion) pulse, respectively. The control/labeled pulse and the balanced SSFP image acquisition were cardiac-triggered to occur at middiastole in consecutive heartbeats. Balanced SSFP parameters were TR/TE = 3.2/1.5 ms, flip angle = 50º, slice thickness = 10 mm, matrix size = 96 × 96, and parallel acceleration factor of 2 for SENSE[36] or 1.6 for GRAPPA.[37] Buxton's general kinetic model[38] was used for MBF quantification. Physiological noise was defined as a short-term variability of MBF measurement and calculated as a SD of 6 MBF values measured from 6 pairs of control and labeled images.[39] The data set was normalized to have zero mean and unit variance. No data augmentation was used in the training of the MCD U-Net.

2.3 | Monte Carlo dropout for uncertainty measure

The concept of using MCD to evaluate model uncertainty was first introduced by Hinton el al in an online lecture[14] and manuscript,[15] respectively. Model uncertainty can be calculated from a network trained with a dropout rate of 50% in every (hidden) layer. At test time, multiple predictions of the same input can be obtained by running the stochastic model several times. Final prediction and model uncertainty is simply the mean and SD of the stochastic predictions, respectively. Subsequently, Gal et al[16] demonstrated that any neural network with dropout added in every weight layer is mathematically equivalent to an approximation of the Bayesian model. Hence, model uncertainty can be estimated given the posterior distribution of the trained weights.

Monte Carlo dropout has been applied to evaluate model uncertainty in semantic segmentation tasks in both computer vison and medical imaging applications.[20,40] In these studies, the typical output of the model uncertainty is a SD pixel-by-pixel map that provides spatial information detailing where, within the image, the model is uncertain. However, a global quantitative score of model uncertainty on a specific input is desired for automatic quality assessment, for triaging images for active learning, and for out-of-distribution detection. In this study, we introduce and evaluate 2 global quantitative scores of model uncertainty, which are named "Dice uncertainty" and "MCD uncertainty." These are the measures of model uncertainty estimated with and without the use of manual segmentation, respectively.

2.4 | Tversky loss function for model adaptability

Binary cross-entropy (BCE) and Dice loss functions are often used to train a CNN model for automatic segmentation.[5,41-43] The definition of the BCE is as follows:

$$L_{BCE} = -\frac{1}{K}\sum_{i=1}^{K}[y_i \cdot \log(\hat{y_1}) + (1-y_1)\cdot \log(1-\hat{y_1})],$$

where K is the total number of pixels in the image, y_i and $\hat{y_1}$ are values of the reference and predicted mask at the ith pixel. Below is the definition of Dice loss function:

$$L_{Dice} = 1 - \frac{2\cdot|A\cap B|}{|A|+|B|},$$

where A is the predicted segmentation and B is the reference segmentation. Alternatively, Dice loss function can also be defined in terms of false positive and false negative, as follows:

$$L_{Dice} = 1 - \frac{TP}{TP+0.5\cdot FP+0.5\cdot FN},$$

where TP is true positive, and FP and FN are false positive and false negative, respectively. As we can see from this equation, Dice coefficient weights false positive and false negative equally, which may not be optimal for myocardial segmentation because of partial volume effects. Myocardium is surrounded by ventricular blood pools and epicardial fat, which have very different physical properties and spin (magnetization) history compared with myocardium. Therefore, false negative may be preferred over false positive. To adapt the network to the desired false-positive versus false-negative tradeoff, Tversky[44] loss function could be used and is defined as

$$L_{\text{Tversky}} = 1 - \frac{\text{TP}}{\text{TP} + (1-\beta)\cdot\text{FP} + \beta\cdot\text{FN}},$$

where β is a hyperparameter that could be set during training. By adjusting β during training, one could adapt the network to output the specific false-positive versus false-negative tradeoff.

2.5 | Experiments

All experiments were performed on a NVIDIA K80 GPU with 12 GB of RAM. Network architectures were implemented using Keras[29] with TensorFlow backend.[30] Common training parameters are number of epochs = 150, batch size = 12, learning rate = 1^{-4}, dropout rate = 50%, and adaptive moment estimation (Adam) optimizer.[45]

2.5.1 | Accuracy

The MCD U-Net architecture was trained with Dice loss function. Training and validation loss were recorded. Dice coefficients of the test set were calculated to evaluate model accuracy. Quantitative MBF measured using automatic segmentation was compared against that measured using the reference manual segmentation using linear regression analysis.

To investigate the efficacy different training procedures, MCD U-Net was also trained on the control images only and then fine-tuned to labeled images. The Dice accuracy was then calculated and compared to that from the model trained using control and labeled images simultaneously.

2.5.2 | Uncertainty

The MCD U-Net architecture was trained with Dice loss function, and MCD was turned on during inference. At test time, MCD inference was applied N = 1115 times (N is the number of MC trials) on each and every test image (total of 144 test images). The mean and SD of 1115 predicted masks yielded the final predicted segmentation and pixel-by-pixel uncertainty map for each test image, respectively.

In this study, we propose 2 global scores of model uncertainty: Dice uncertainty and MCD uncertainty, which are calculated with and without the use of reference segmentation, respectively. Dice uncertainty is defined as the SD of Dice coefficients calculated from N stochastic predicted segmentations, given the ground-truth reference segmentation. Higher Dice uncertainty means the model is less certain of its predictions and experiences higher variability.

Monte Carlo dropout uncertainty provides a global score of how certain or uncertain the model is, given an input image, and is defined as the sum of values of all pixels in the uncertainty map normalized by the volume of the predicted mask. Similar to the mean squared error or structural similarity index, MCD uncertainty is expected to provide a global view on how certain or uncertain the model is for a specific test image. This potentially allows automatic quality assessment of automatic segmentation without the need of the reference ground-truth segmentation, which is typically not available at production.

Time penalty is a major consideration when using MCD for model uncertainty measure.[18] We carried out 2 experiments that studied the effects of the number of MC trials and batch size on uncertainty measure and inference time, respectively. In the first experiment, we performed MCD on 1 test case with 16 384 MC trials. Smaller MC trials were retrospectively bootstrapped from a distribution of 16 384 samples 1024 times. This allows calculation of a confidence interval (i.e., SD) of uncertainty measure as a function of the number of MC trials. In the second experiment, we performed N = 1024 MC trials on a single test case to measure inference time, mean prediction, and Dice uncertainty as a function of batch size.

2.5.3 | Adaptability

The MCD U-Net architecture was trained with Tversky loss function, which has a hyperparameter β. The MCD U-Net model was trained with 9 different values of β ranging from 0.1 to 0.9 with step size of 0.1. False-positive and false-negative rates were defined as an average number of false-positive and false-negative pixels per image, respectively. False-positive and false-negative rates were calculated and compared to that from the model trained with BCE loss and Dice loss.

To demonstrate the consequence of partial volume effects, "thin mask" and "thick mask" were generated using a "*bwmorph*" function in MATLAB, which removes and adds 1 pixel from both sides of the reference masks in the test set, respectively. The average Dice coefficient, false-positive rate, and false-negative rate were calculated and compared with that calculated from CNN models. Quantitative MBF measured using "thin mask" and "thick mask" were compared against that from the reference masks to demonstrate that false positive is more detrimental than false negative.

2.6 | Probabilistic U-Net

Probabilistic U-Net (prob U-Net)[26] has recently been proposed for segmentation of ambiguous images including medical imaging. The primary goal of the prob U-Net is to efficiently generate many (even infinite) plausible segmentation hypotheses for a given input image. It combines a U-Net architecture with a conditional variational autoencoder. Monte Carlo Dropout U-Net and prob U-Net are both probabilistic models that could generate unlimited number of plausible predictions by sampling the posterior distributions of trained weights and variational latent representation (lower-dimensional space), respectively.

The prob U-Net[46] model was adapted and trained on the ASL images with the same data training, validation, and testing splits. Training details were similar to the medical imaging example in the original work.[26] The training was performed with randomly initialized weights for over 5400 stochastic gradient decent iterations; the initial learning rate was 1^{-4}, which was lowered to 5^{-5} after one-third and to 1^{-5} after two-thirds of the iterations. Batch size was 12. Adam optimizer with default parameters[45] was used in combination with 1^{-5} multiplier of weight decay. In addition to BCE loss, Kullback-Leibler divergence with a multiplier λ was added to compose the total loss. The Kullback-Leibler divergence penalizes the differences between the posterior and the prior distributions. Several values of λ (0.1, 0.5, 0.75, 1.0, 2.0, 5.0, and 10.0) were tested and $\lambda = 10$ was chosen. given the minimum validation loss and Kullback-Leibler divergence. In this experiment, 6-dimensional latent space was used. Random elastic deformation and random rotation were used for data augmentation.

2.7 | Data analysis

2.7.1 | Accuracy

To access accuracy of the automatic segmentation method, Dice coefficients of a test set were calculated. Furthermore, MBF calculated using the automatic segmentations was compared against that calculated using the reference manual segmentations using linear regression and concordance correlation analyses.

To investigate whether there is a correlation between CNR and Dice accuracy, CNR was calculated for all images in the test set. Linear regression analysis between CNR and Dice accuracy was performed.

2.7.2 | Uncertainty measure

For each test case, MCD inference was performed N = 1115 times resulted in 1115 stochastic predictions. Mean predicted segmentation, pixel-by-pixel uncertainty map, Dice uncertainty, and MCD uncertainty were calculated as described previously. Linear regression analysis was carried out to study the relationship between Dice uncertainty and MCD uncertainty. Linear regression analysis was performed to determine whether there is a relationship between MCD uncertainty and physiological noise of the MBF measurement.

2.7.3 | Adaptability

False-positive and false-negative rate, defined as an average number of false-positive and false-negative pixels per image, were calculated given the reference masks and predicted masks. False-positive and false-negative rate from networks trained with BCE, Dice, and Tversky losses and that from the "thick mask" and "thin mask" were compared.

2.7.4 | Probabilistic U-Net

For each test case, prob U-Net was inferenced N = 128 times, resulting in 128 stochastic predictions. Mean predicted segmentation, pixel-wise uncertainty map, Dice uncertainty, and MCD uncertainty associated with prob U-Net were calculated in the same manner as those for MCD U-Net. Dice accuracy associated with prob U-Net was calculated and compared with that of MCD U-Net. Linear regression analysis was carried out to study the relationship between Dice uncertainty and MCD uncertainty calculated from prob U-Net. Furthermore, MCD uncertainty calculated using MCD U-Net and prob U-Net were compared using linear regression analysis.

3 | RESULTS

3.1 | Accuracy

The model was trained for 150 epochs. The training and validation loss are shown in Supporting Information Figure S1. Representative segmentation masks and MBF maps generated using the CNN model in comparison with that using manual segmentation are shown in Figure 2. The average Dice coefficient for the test set was 0.91 ± 0.04. For quantitative imaging, accuracy assessment using clinically relevant quantity is desired. Quantitative regional MBF measured using automatic segmentation is highly correlated with that calculated using manual segmentation ($R^2 = 0.96$; Figure 3). The concordance correlation coefficient was 0.98.

Examples of control and labeled images are shown in Supporting Information Figure S2. The average CNR of control and labeled images were 21.44 ± 3.81 and 6.46 ± 1.03, respectively. The average Dice accuracy from the control

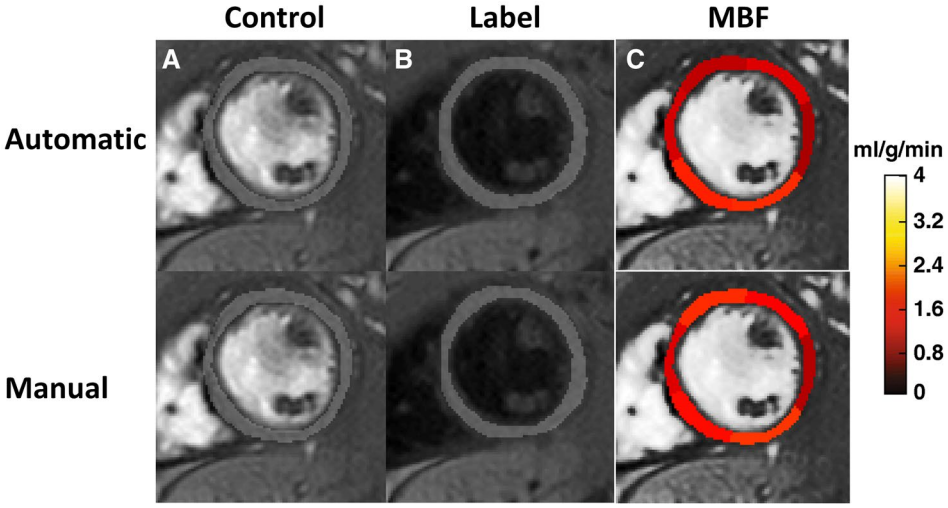

FIGURE 2 Examples of image quality, segmentation, and myocardial blood flow (MBF) maps. Shown are representative control images (A), labeled images (B), and regional MBF (C) maps. Segmentation masks and MBF maps generated by the convolution neural network (CNN; top row) are comparable to those from manual segmentation (bottom row)

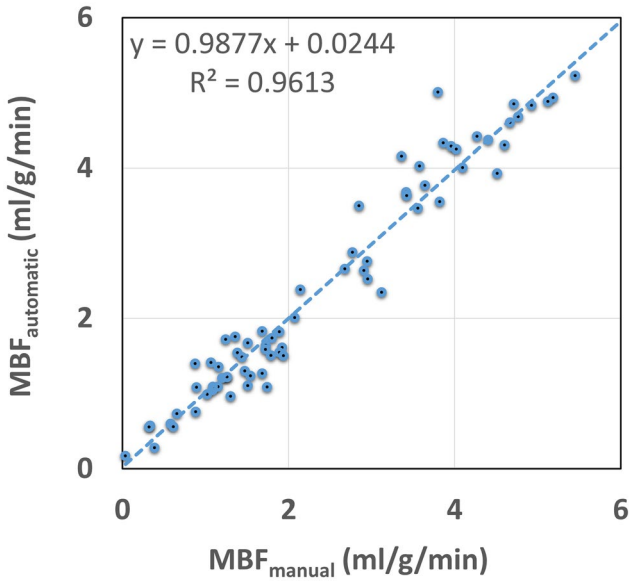

FIGURE 3 Evaluation of automatic segmentation. Regional MBF measured using automatic segmentation (y-axis) is highly correlated with that measured using manual segmentation (x-axis) ($R^2 = 0.96$). The concordance correlation coefficient was 0.98

images (0.92 ± 0.02) was significantly higher than that (0.90 ± 0.04) from the labeled images ($P < .001$); however, the difference was small. There was no significant correlation between CNR and Dice accuracy ($R^2 = 0.09$).

As indicated in Table 1, the model trained using control and labeled images simultaneously demonstrated higher Dice accuracy compared with that obtained from the model trained on the control images and then fine-tuned on the labeled images.

3.2 | Uncertainty

Given the manual segmentation, the Dice coefficient distributions calculated from 2 test cases are shown in Figure 4 and in Supporting Information Figure S3. Dice coefficient distributions of 4 other test cases are shown in Supporting Information Figure S4. The variance of the Dice coefficient distribution represents how the stochastic predictions fluctuate. Therefore, we proposed to use the SD (named Dice uncertainty) of the Dice coefficient distribution as a measure of model uncertainty, given the manual segmentation.

TABLE 1 Comparison of Dice accuracy from the 2 training procedures

	Pretrain with control images and fine-tune with the labeled images			Train with both control and labeled images	
	Without fine-tuning		Fine-tuning		
	Control	Labeled	Labeled	Control	Labeled
Dice accuracy	0.87 ± 0.04	0.14 ± 0.08	0.86 ± 0.06	0.92 ± 0.02	0.90 ± 0.04

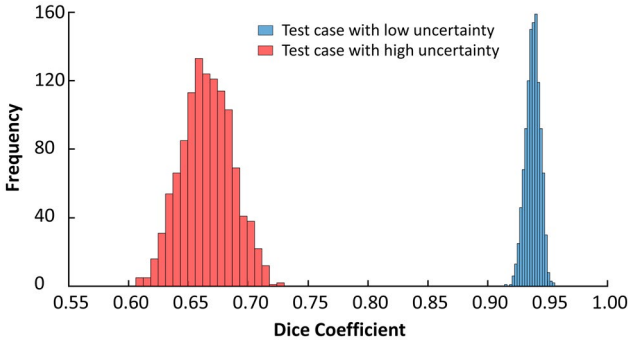

FIGURE 4 Illustration of the proposed Dice uncertainty metric. Shown are the Dice coefficient distributions of 2 test cases after MCD inference. The distribution is broader in the test case with high uncertainty (red) compared to that with low uncertainty (blue)

Dice uncertainty and MCD uncertainty are in good agreement ($R^2 = 0.64$; Figure 5A). Dice uncertainty is defined as the SD of Dice coefficients calculated from N stochastic predicted segmentations, given the ground-truth reference segmentation. Higher Dice uncertainty means the model is less certain of its predictions and experiences higher variability. Calculation of Dice uncertainty requires the use of the ground-truth segmentation, which is often not available at production. Therefore, we proposed MCD uncertainty as an alternative that is calculated without the use of the ground-truth segmentation. The goal of Figure 5A is to demonstrate that MCD uncertainty also represents model uncertainty.

Figure 5B,C shows the predicted segmentation and uncertainty map of a test case with low global uncertainty score (blue circle in Figure 5A). In the test case with low uncertainty, the automatic segmentation (yellow lines) is in good agreement with the manual segmentation (green lines). Figure 5D,E shows the predicted segmentation and uncertainty map of a test case with high global uncertainty score (red circle in Figure 5A). In this case, discrepancies (yellow arrows) can be seen between the automatic segmentation and the manual segmentation. Additionally, the model uncertainty map provides spatial information when the model is most uncertain, as seen in the area indicated by the yellow arrows. Monte Carlo dropout uncertainty is weakly correlated ($R^2 = 0.13$) to physiological noise, as seen in Supporting Information Figure S5.

As batch size increases, inference time is significantly decreased (Supporting Information Figure S6A) without altering the mean prediction and uncertainty measure, as seen in Supporting Information Figure S6B. The time reduction experienced diminishing return around a batch size of 256 with the 12-GB-memory NVIDIA K80 GPU used in this study. With a more powerful GPU, the inference time is expected to be further decreased by using a larger batch size.

As the number of MC trials decreases, the confidence in uncertainty estimation is decreased, as seen in Figure 6. In this study, we simply chose N = 1115 MC trials for our uncertainty analysis; however, it is worth noting that there is a trade-off between inference time and confidence interval of uncertainty measure, and N larger than 100 can provide accurate uncertainty measure.

3.3 | Adaptability

Supporting Information Figure S7 shows an example of a "thin mask," a "thick mask," and a reference manual mask. Thin mask and thick mask data have very similar Dice coefficients, which are 0.80 ± 0.04 and 0.81 ± 0.02, respectively. However, the false-positive and false-negative rates are completely opposite (Figure 7). Thin mask had a negligible effect on the endpoint clinically relevant quantitative MBF, whereas thick mask introduced a significant overestimation,

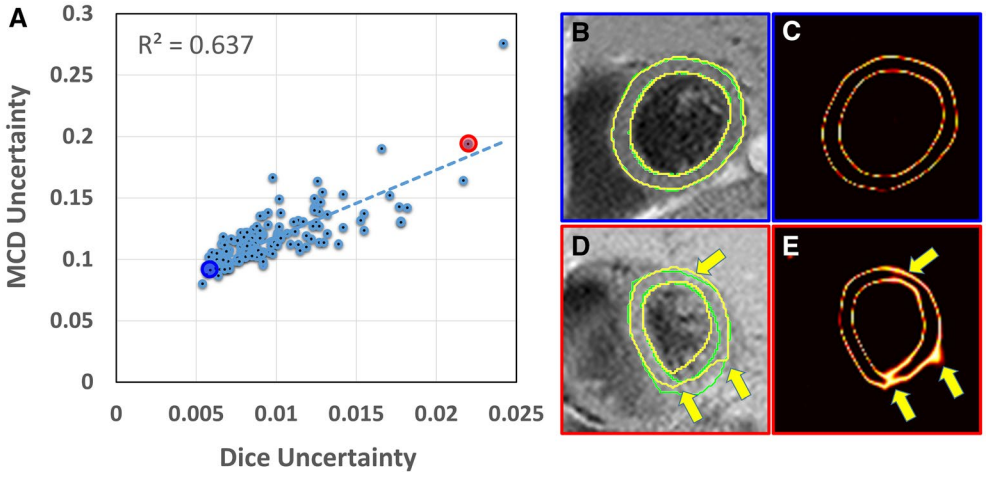

FIGURE 5 Uncertainty metrics. A, Dice uncertainty (x-axis) and MCD uncertainty (y-axis) are correlated ($R^2 = 0.637$). Examples of predicted segmentations and uncertainty maps of test cases with low (B,C; blue circle in [A]) and high (D,E; red circle in [A]) uncertainty are shown. Manual segmentations are in green and automatic segmentations are in yellow. In the high uncertainty case, the uncertainty map (E) provides the specific spatial locations where the model is most uncertain, which aligns with segmentation errors (D), identified by yellow arrows

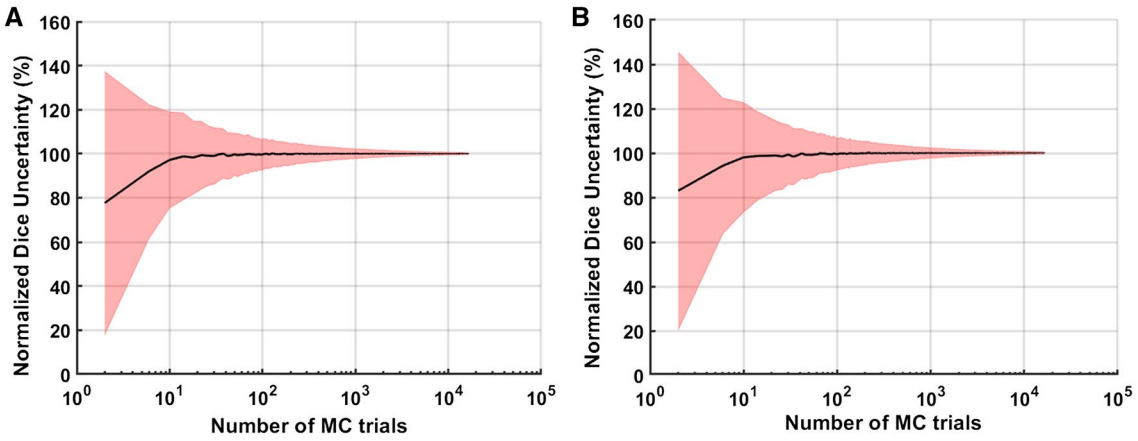

FIGURE 6 Impact of the number of Monte Carlo (MC) trials (x-axis) on normalized uncertainty measure (y-axis) from a test case with high uncertainty (A) and another case with low uncertainty (B). Mean (black line) ± 1 SD (shaded pink) is plotted. Note that more than 100 Monte Carlo trials are needed to achieve accurate uncertainty measures

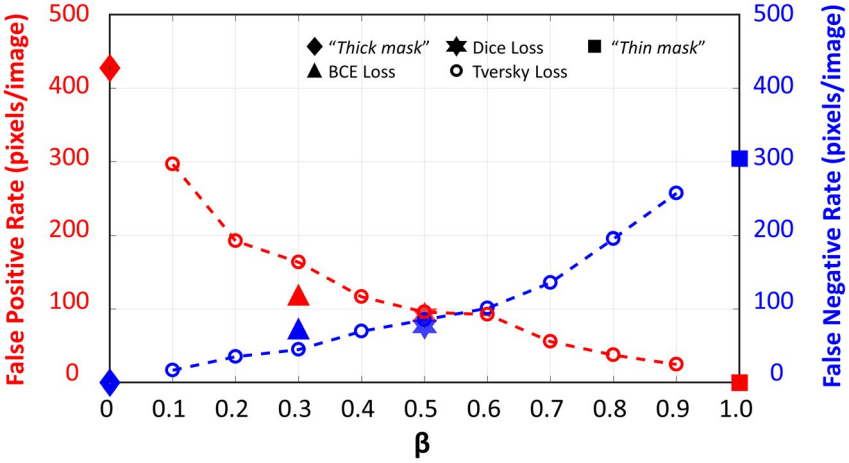

FIGURE 7 Impact of β on the false-positive (left y-axis) and false-negative (right y-axis) rate. Tversky loss function with different β shows a range of false-positive versus false-negative tradeoff. For quantitative cardiovascular magnetic resonance, β larger than 0.5 may be preferred. Abbreviation: BCE, binary cross-entropy

as shown in Supporting Information Figure S7D,E, respectively. The overestimation is a consequence of partial volume effects (i.e., contamination signal from ventricular blood pools and/or epicardial fat).

Supporting Information Figure S8A,B shows the number of false-positive pixels subtracted by number of false-negative pixels and average Dice coefficient for the entire test set as a function of β. Figure 7 demonstrates the adaptability of the CNN trained with Tversky loss function. As β increases, false-positive and false-negative rates are monotonically decreasing and increasing, respectively. By varying β during training, specific false-positive versus false-negative tradeoffs can be made. For comparison purposes, the false-positive and false-negative rate from the thin mask, thick mask, and CNN models trained with BCE and Dice loss are shown in Figure 7.

3.4 | Probability U-Net

The averaged Dice coefficient from the test set was 0.86 ± 0.04, which is lower than that obtained from MCD U-Net (0.91 ± 0.04). Analysis similar to that in Figure 5A was performed for prob U-Net. The resulting linear regression analysis shows a weaker correlation ($R^2 = 0.18$) between Dice uncertainty and MCD uncertainty using prob U-Net. Furthermore, there was no correlation between MCD uncertainty calculated using MCD U-Net and prob U-Net ($R^2 = 0.0021$).

4 | DISCUSSION

The contributions of this manuscript are 3-fold. First, we demonstrated that it is possible to train a single CNN model to segment control and labeled ASL images, which have substantially different SNR, CNR, and contrast, and whose contrast may vary substantially among images due to heart-rate variation. This challenges a common view that a CNN model must be tailored to a specific contrast (with the help of transfer learning). Second, we introduce and evaluate 2 novel approaches to measure model uncertainty. We denote these as Dice uncertainty and MCD uncertainty, and calculate them

with and without the need for reference manual segmentation, respectively. Monte Carlo dropout uncertainty may be valuable for automatic quality control at production, model improvement through active learning, and/or out-of-distribution detection. Third, we introduce the use of Tversky loss function to adapt the CNN to specific false-positive versus false-negative needs of an application. This is useful for quantitative cardiac MRI, as left ventricular myocardium is surrounded by blood pool and epicardial fat, which have distinct physical properties and spin history compared with that of myocardium.

4.1 | Accuracy

The proposed model achieved good Dice accuracy of 0.91 ± 0.04, similar to that reported in the literature.[5,47-49] This was in spite of facing additional challenges compared with cardiac cine imaging, namely (1) lower spatial resolution, (2) lower and inconsistent SNR and blood–myocardium CNR, and (3) SNR and CNR differences between the control and labeled series. Our study also found the quantitative MBF measured using automatic segmentation to be highly correlated with MBF measured using manual segmentation ($R^2 = 0.96$).

4.2 | Uncertainty

Spatial uncertainty maps were calculated as the SD of all stochastic predictions. A global uncertainty score is needed to perform automatic quality assessment without a human observer. In this study, we introduced and evaluated 2 simple yet intuitive approaches, denoted as Dice uncertainty and MCD uncertainty. These were calculated with and without the need for manual segmentation, respectively. We demonstrated that Dice uncertainty and MCD uncertainty were in good agreement ($R^2 = 0.64$). Dice uncertainty intuitively represents model uncertainty, suggesting that MCD uncertainty also represents model uncertainty on an input image. Monte Carlo dropout uncertainty does not require the use of reference segmentation; therefore, it could be used for automatic quality control, for automatic triage of images for active learning, and for out-of-distribution detection.

Uncertainty measures fit several clinical scenarios. In the context of image segmentation, many so-called "automatic" segmentation methods are not "fully automatic," because there lacks a quality metric that could inform the user on how confident the prediction is. Consequently, the human operator often has to review every segmentation generated by an algorithm before beginning the next step in the postprocessing pipeline (i.e., quantification step). This is still an open challenge for the medical imaging community. This work attempted to address the challenge by introducing the MCD uncertainty measure that may enable a "fully automatic" postprocessing pipeline, as the human operator may only need to double-check cases with high uncertainty. Human intervention in the highly uncertain cases helps to prevent propagations of segmentation error/uncertainty to the perfusion quantification step in the first place, avoiding the possibilities that uncertainty/errors from the segmentation step confound the integrity of the myocardial perfusion quantification.

There is a time penalty for using MCD to compute uncertainty. This study demonstrated that increased batch size significantly decreased inference time without altering the mean prediction and uncertainty measure, as shown in Supporting Information Figure S6. A batch size of 64 or 128 yielded the minimum inference time. given the GPU used in this study (NVIDIA K80 GPU). With more powerful GPUs, it is likely that a larger batch size will be feasible and provide shorter inference time. A complimentary approach to reduce inference time is to use a smaller number of MC trials. Figure 6 demonstrates the confidence of the uncertainty measure versus the number of MC trials (i.e., inference time). One SD of less than 2% from the expected uncertainty measure was observed with the number of MC trials larger than 1000. In this work, MC trials of 1115 were used to calculate uncertainty metrics; however, accurate uncertainty calculation is achieved with MC trials around 100.

There are additional measures of model uncertainty that remain to be explored. Teye et al demonstrated that Bayesian uncertainty could be estimated from batch-normalized deep networks.[50] Ayhan et al demonstrated that model uncertainty could be measured using test-time augmentation.[51] Like MCD, these methods are efficient and practical, requiring no modification to the existing network architecture. Monte Carlo dropout may be the most efficient among the 3, as it does not require any postprocessing (i.e., data augmentation or mini-batch preparation). Furthermore, a single input can be replicated and bundled in a mini batch, which is very graphics processing unit (GPU)-efficient. Higher computational capacity allows for larger possible batch sizes, leading to significantly reduced MC inference times.

This work considered only 2 uncertainty metrics, Dice uncertainty and MCD uncertainty, which are both calculated using the variance of the outputs. Neither captures pixel correlations or non-Gaussianity of the output distribution. Further research is needed to investigate metrics that capture pixel correlations and/or consider higher-order statistics of the output distribution, such as skewness and kurtosis, which would account for asymmetry and tail extremity, respectively.

4.3 | Adaptability

Quantitative myocardial measurements can be easily biased if there is even a small contamination with signal from the

ventricular blood pool and/or epicardial fat. To improve diagnostic efficacy, false positive may be penalized heavier than false negative during training, leading to a model favorably producing output predictions with a lower false-positive rate. In this study, we demonstrated that model adaptability could be achieved using Tversky loss function.

The false-positive and false-negative rate produced by a network depends on the false-positive and false-negative weighting in the loss function. Dice loss function weights false positive and false negative equally, which is why we observed that false-positive and false-negative rate from a network trained with Dice loss are very similar, as shown in Figure 7. On the other hand, false negative is weighted heavier in BCE loss function, leading to a false-positive rate that is higher than the false-negative rate. Tversky loss function is a generalized form of Dice loss function with a hyperparameter β that controls the false-positive versus false-negative tradeoff. As shown in Figure 7, the network trained with Tversky loss function was able to produce ranges of false-positive versus false-negative tradeoff. For quantitative CMR, including myocardial ASL, β larger than 0.5 may be favorable. The false-positive versus false-negative tradeoff of a network could be adapted more broadly in 2D space with 2 positive hyperparameters, which independently weight false-positive and false-negative terms in the Tversky loss function.

4.4 | Probability U-Net

The prob U-Net is an innovative and attractive concept, as it can accurately produce many (even infinitely) plausible hypotheses, given an input image. That is especially useful for ambiguous images including medical imaging. One open question for prob U-Net is how to evaluate the quality of each and every plausible hypothesis, to identify the best prediction given an input image. For analyzing results obtained from prob U-Net, we simply used the similar methods as in MCD U-Net to calculate the mean prediction, uncertainty map, Dice uncertainty and MCD uncertainty, and compared them to that calculated from MCD U-Net. We hypothesize that the reason for lower Dice accuracy of prob U-Net (compared with MCD U-Net) is that it is not designed to predict the best (i.e., prediction with least BCE or Dice loss) but rather to generate many plausible predictions.

4.5 | Limitations

The primary limitation of this study is that it was performed on a relatively small sample size collected from a single MRI vendor, a single institution, and a single graphical prescription protocol (mid short-axis). This is primarily because ASL-based human myocardial perfusion imaging is an emerging and still experimental technique. Although the sample size is small in absolute terms, it is among the largest myocardial ASL data sets from human subjects to date. This study is also based on the U-Net model, which has been validated in many medical applications. Dice accuracy on the unseen test set was consistent with those reported in the literature.[5,47-49] Model retraining is often required when applied to different data sets, tasks, or applications, even with large training data.[5] Therefore, we expect that the results from this study can be translated to more variable data sets or other CMR applications through retraining.

5 | CONCLUSIONS

We demonstrate the feasibility of deep CNN fully automatic segmentation of the left ventricular myocardium in myocardial ASL perfusion imaging, with good accuracy in terms of Dice coefficients and regional MBF quantification. We introduce 2 simple yet powerful methods for measuring a global uncertainty score, both with and without the use of manual segmentation, named Dice uncertainty and MCD uncertainty, respectively. We also demonstrate the ability to adapt the CNN model to a specific false-positive versus false-negative tradeoff using the Tversky loss function. These findings are directly relevant to automatic segmentation in quantitative cardiac MRI and are broadly applicable to automatic segmentation problems in diagnostic imaging.

ACKNOWLEDGMENTS

This study was supported by the National Institutes of Health (R01HL130494-01A1) and the Whittier Foundation (0003457-00001).

ORCID

Hung P. Do https://orcid.org/0000-0002-8141-8630
Krishna S. Nayak https://orcid.org/0000-0001-5735-3550

REFERENCES

1. Kober F, Jao T, Troalen T, Nayak KS. Myocardial arterial spin labeling. *J Cardiovasc Magn Reson*. 2016;18:22.
2. Cerqueira MD, Weissman NJ, Dilsizian V, et al. Standardized myocardial segmentation and nomenclature for tomographic imaging of the heart. *Circulation*. 2002;105:539–542.
3. Kamnitsas K, Ledig C, Newcombe VFJ, et al. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Med Image Anal*. 2017;36:61–78.
4. Liu F, Zhou Z, Jang H, Samsonov A, Zhao G, Kijowski R. Deep convolutional neural network and 3D deformable approach for tissue segmentation in musculoskeletal magnetic resonance imaging. *Magn Reson Med*. 2018;79:2379–2391.

5. Bai W, Sinclair M, Tarroni G, et al. Automated cardiovascular magnetic resonance image analysis with fully convolutional networks. *J Cardiovasc Magn Reson*. 2018;20:65.
6. Milletari F, Navab N, Ahmadi S-A. V-Net: fully convolutional neural networks for volumetric medical image segmentation. In: Proceedings of the Fourth International Conference on 3D Vision (3DV), Stanford, California, 2016. pp 565–571.
7. Puntmann VO, Valbuena S, Hinojar R, et al. Society for Cardiovascular Magnetic Resonance (SCMR) expert consensus for CMR imaging endpoints in clinical research. Part I: Analytical validation and clinical qualification. *J Cardiovasc Magn Reson*. 2018;20:67.
8. Salerno M, Kramer CM. Advances in parametric mapping with CMR imaging. *JACC Cardiovasc Imaging*. 2013;6:806–822.
9. Messroghli DR, Moon JC, Ferreira VM, et al. Clinical recommendations for cardiovascular magnetic resonance mapping of T1, T2, T2* and extracellular volume: a consensus statement by the Society for Cardiovascular Magnetic Resonance (SCMR) endorsed by the European Association for Cardiovascular Imaging. *J Cardiovasc Magn Reson*. 2017;19:75.
10. Nguyen C, Fan Z, Sharif B, et al. In vivo three-dimensional high resolution cardiac diffusion-weighted MRI: a motion compensated diffusion-prepared balanced steady-state free precession approach. *Magn Reson Med*. 2014;72:1257–1267.
11. Atkins MS, Mackiewich BT. Fully automatic segmentation of the brain in MRI. *IEEE Trans Med Imaging*. 1998;17:98–107.
12. Petitjean C, Dacher J-N, Petitjean C, Dacher J-N. A review of segmentation methods in short axis cardiac MR images. *Med Image Anal*. 2011;15:169–184.
13. Gal Y, Islam R, Ghahramani Z. Deep Bayesian active learning with image data. In: Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, 2017. pp 1183–1192.
14. Hinton G. Lecture 10.5 – Dropout: an efficient way to combine neural nets. *COURSERA Neural Networks Mach Learning*. 2012;34–41.
15. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. *J Mach Learn Res*. 2014;15:1929–1958.
16. Gal Y, Ghahramani Z, Uk ZA, Ghahramani Z, Uk ZA. Dropout as a Bayesian approximation: representing model uncertainty in deep learning. In: Proceedings of the International Conference on Machine Learning, New York, New York, 2016. pp 1050–1059.
17. Kendall A, Badrinarayanan V, Cipolla R.Bayesian SegNet: model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. arXiv:1511.02680.
18. Kendall A, Gal Y. What uncertainties do we need in Bayesian deep learning for computer vision? In: *Advances in Neural Information Processing Systems*. Cambridge, Massachusetts: MIT Press; 2017:5574–5584.
19. Kampffmeyer M, Salberg A-B, Jenssen R. Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks. In: Proceedings of the 29th IEEE Conference on Computer Vision and Pattern Recognition Workshops, Las Vegas, Nevada, 2016. pp 1–9.
20. Zhao G, Liu F, Oler JA, Meyerand ME, Kalin NH, Birn RM. Bayesian convolutional neural network based MRI brain extraction on nonhuman primates. *NeuroImage*. 2018;175:32–44.
21. Kellman P, Hansen MS. T1-mapping in the heart: accuracy and precision. *J Cardiovasc Magn Reson*. 2014;16:2.
22. Kellman P, Wilson JR, Xue H, et al. Extracellular volume fraction mapping in the myocardium. Part 2: Initial clinical experience. *J Cardiovasc Magn Reson*. 2012;14.
23. Kellman P, Wilson JR, Xue H, Ugander M, Arai AE. Extracellular volume fraction mapping in the myocardium. Part 1: Evaluation of an automated method. *J Cardiovasc Magn Reson*. 2012;14:63.
24. Bulluck H, Rosmini S, Abdel-Gadir A, et al. Automated extracellular volume fraction mapping provides insights into the pathophysiology of left ventricular remodeling post-reperfused ST-elevation myocardial infarction. *J Am Heart Assoc*. 2016;5:1–12.
25. Ferreira VM, Wijesurendra RS, Liu A, et al. Systolic ShMOLLI myocardial T1-mapping for improved robustness to partial-volume effects and applications in tachyarrhythmias. *J Cardiovasc Magn Reson*. 2015;17:1–11.
26. Kohl S, Romera-Paredes B, Meyer C, et al. A probabilistic U-Net for segmentation of ambiguous images. In: *Advances in Neural Information Processing Systems*. Cambridge, Massachusetts: MIT Press; 2018:6965–6975.
27. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In: Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 2015. pp 234–241.
28. Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. arXiv:1502.03167.
29. Chollet F. Keras. github.com. 2015. https://github.com/keras-team/keras. Accessed January 05, 2018.
30. Abadi M, Barham P, Chen J, et al. TensorFlow: a system for large-scale machine learning. In: Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation, Savannah, Georgia, 2016. pp 265–283.
31. Do HP, Jao TR, Nayak KS. Myocardial arterial spin labeling perfusion imaging with improved sensitivity. *J Cardiovasc Magn Reson*. 2014;16:15.
32. Do HP, Yoon AJ, Fong MW, Saremi F, Barr ML, Nayak KS. Double-gated myocardial ASL perfusion imaging is robust to heart rate variation. *Magn Reson Med*. 2017;77:1975–1980.
33. Yoon AJ, Do HP, Cen S, et al. Assessment of segmental myocardial blood flow and myocardial perfusion reserve by adenosine-stress myocardial arterial spin labeling perfusion imaging. *J Magn Reson Imaging*. 2017;46:413–420.
34. Kim S-G. Quantification of relative cerebral blood flow change by flow-sensitive alternating inversion recovery (FAIR) technique: application to functional mapping. *Magn Reson Med*. 1995;34:293–301.
35. Kwong KK, Chesler DA, Weisskoff RM, et al. MR perfusion studies with T1-weighted echo planar imaging. *Magn Reson Med*. 1995;34:878–887.
36. Pruessmann KP, Weiger M, Scheidegger MB, Boesiger P. SENSE: sensitivity encoding for fast MRI. *Magn Reson Med*. 1999;42:952–962.
37. Griswold MA, Jakob PM, Heidemann RM, et al. Generalized autocalibrating partially parallel acquisitions (GRAPPA). *Magn Reson Med*. 2002;47:1202–1210.
38. Buxton RB, Frank LR, Wong EC, Siewert B, Warach S, Edelman RR. A general kinetic model for quantitative perfusion imaging with arterial spin labeling. *Magn Reson Med*. 1998;40:383–396.

39. Zun Z, Wong EC, Nayak KS. Assessment of myocardial blood flow (MBF) in humans using arterial spin labeling (ASL): feasibility and noise analysis. *Magn Reson Med*. 2009;62:975–983.
40. 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*. 2019;338:34–45.
41. Milletari F, Navab N, Ahmadi S-A. V-Net: fully convolutional neural networks for volumetric medical image segmentation. In: Proceedings of the Fourth International Conference on 3D Vision (3DV), Stanford, California, 2016. pp 565–571.
42. Badrinarayanan V, Kendall A, Cipolla R. SegNet: a deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans Pattern Anal Mach Intell*. 2017;39:2481–2495.
43. Long J, Shelhamer E, Darrell T, Long J, Darrell T. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, Massachusetts, 2015. pp 3431–3440.
44. Tversky A. Features of similarity. *Psychol Rev*. 1977;84:327–352.
45. Kingma DP, Ba J, Lei Ba J.Adam: a method for stochastic optimization. arXiv:1412.6980.
46. Kohl S. Probabilistic U-Net. github.com. 2018. https://github.com/SimonKohl/probabilistic_unet. Accessed January 10, 2019.
47. Zotti C, Luo Z, Humbert O, Lalande A, Jodoin P-M. GridNet with automatic shape prior registration for automatic MRI cardiac segmentation. In: Proceedings of the 8th International Workshop on Statistical Atlases and Computational Models of the Heart, Quebec City, Canada, 2017. pp 73–81.
48. Patravali J, Jain S, Chilamkurthy S. 2D–3D fully convolutional neural networks for cardiac MR segmentation. In: Proceedings of the 8th International Workshop on Statistical Atlases and Computational Models of the Heart, Quebec City, Canada, 2017. pp 130–139.
49. Baumgartner CF, Koch LM, Pollefeys M, Konukoglu E, Given NA. An exploration of 2D and 3D deep learning techniques for cardiac MR image segmentation. In: Proceedings of the 8th International Workshop on Statistical Atlases and Computational Models of the Heart, Quebec City, Canada, 2017. pp 111–119.
50. Teye M, Azizpour H, Smith K.Bayesian uncertainty estimation for batch normalized deep networks. arXiv:1802.06455.
51. Ayhan MS, Berens P. Test-time data augmentation for estimation of heteroscedastic aleatoric uncertainty in deep neural networks. In: International Conference on Medical Imaging with Deep Learning. 2018.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

FIGURE S1 Training and validation loss

FIGURE S2 Data characteristics of myocardial arterial spin labeling (ASL). Representative examples of labeled (A) and control (B) images from the test set. Labeled and control images are distinctly different in SNR and contrast-to-noise ratio (CNR). The SNR and CNR within each image type are variable due to changes in heart rate during experiments and from patient to patient. Abbreviations: RV, right ventricle; LV, left ventricle

FIGURE S3 Dice coefficients calculated from N = 1115 stochastic predictions associated with 2 representative test cases in Figure 4, which have low uncertainty (blue circles) and high uncertainty (red triangles)

FIGURE S4 Dice coefficient distributions of 4 other test cases

FIGURE S5 Monte Carlo dropout uncertainty versus physiological noise (PN). The MCD uncertainty is weakly correlated ($R^2 = 0.13$) to PN

FIGURE S6 Monte Carlo dropout inference time with different batch size (N = 1024 MC trials). Inference time is significantly reduced with increase in batch size (A) without alteration to uncertainty measure (B)

FIGURE S7 Demonstration of partial-volume effects. "Thin mask" (A) and "thick mask" (C) were generated using the *bwmorph* MATLAB function, which removes and adds 1 pixel from both sides of the manual mask (B), respectively. Despite having distinctly different false-positive and false-negative rates, the thin mask and thick mask data have very similar mean Dice coefficients, which is approximately 0.8. D, Thin mask does not introduce any bias into the endpoint quantitative MBF measure. E, However, significant overestimation was observed with thick mask due to the partial-volume effects

FIGURE S8 Number of false-positive (FP) pixels subtracted by the number of false-negative (FN) pixels (A) and average Dice coefficient (B) on a test set as a function of β, which is the hyperparameter in the Tversky loss function

How to cite this article: Do HP, Guo Y, Yoon AJ, Nayak KS. Accuracy, uncertainty, and adaptability of automatic myocardial ASL segmentation using deep CNN. *Magn Reson Med*. 2019;00:1–12. https://doi.org/10.1002/mrm.28043