Temporal Uncertainty Localization to Enable Human-in-the-Loop Analysis of Dynamic Contrast-Enhanced Cardiac MRI Datasets

Dilek M. Yalcinkaya[1,2], Khalid Youssef[1,3], Bobak Heydari[4],
Orlando Simonetti[5], Rohan Dharmakumar[3,6], Subha Raman[3,6],
and Behzad Sharif[1,3,6](✉)

[1] Laboratory for Translational Imaging of Microcirculation, Indiana University School of Medicine (IUSM), Indianapolis, IN, USA

[2] Elmore Family School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN, USA

[3] Krannert Cardiovascular Research Center, IUSM/IU Health Cardiovascular Institute, Indianapolis, IN, USA

bsharif@iu.edu

[4] Stephenson Cardiac Imaging Centre, University of Calgary, Alberta, Canada

[5] Department of Internal Medicine, Division of Cardiovascular Medicine, Davis Heart and Lung Research Institute, The Ohio State University, Columbus, OH, USA

[6] Weldon School of Biomedical Engineering, Purdue University, West Lafayette, IN, USA

Abstract. Dynamic contrast-enhanced (DCE) cardiac magnetic resonance imaging (CMRI) is a widely used modality for diagnosing myocardial blood flow (perfusion) abnormalities. During a typical free-breathing DCE-CMRI scan, close to 300 time-resolved images of myocardial perfusion are acquired at various contrast "wash in/out" phases. Manual segmentation of myocardial contours in each time-frame of a DCE image series can be tedious and time-consuming, particularly when non-rigid motion correction has failed or is unavailable. While deep neural networks (DNNs) have shown promise for analyzing DCE-CMRI datasets, a "dynamic quality control" (dQC) technique for reliably detecting failed segmentations is lacking. Here we propose a new space-time uncertainty metric as a dQC tool for DNN-based segmentation of free-breathing DCE-CMRI datasets by validating the proposed metric on an external dataset and establishing a human-in-the-loop framework to improve the segmentation results. In the proposed approach, we referred the top 10% most uncertain segmentations as detected by our dQC tool to the human expert for refinement. This approach resulted in a significant increase in the Dice score ($p < 0.001$) and a notable decrease in the number of images with failed segmentation (16.2% to 11.3%) whereas the alternative approach of randomly selecting the same number of segmentations for human referral did not achieve any significant improvement. Our results suggest that the proposed dQC framework has the potential to accurately identify poor-quality segmentations and may enable efficient

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14222, pp. 453–462, 2023.
https://doi.org/10.1007/978-3-031-43898-1_44

DNN-based analysis of DCE-CMRI in a human-in-the-loop pipeline for clinical interpretation and reporting of dynamic CMRI datasets.

Keywords: Cardiovascular MRI · Dynamic MRI · Image Segmentation · Quality control · Uncertainty Quantification · Human-in-the-loop A.I.

1 Introduction

Dynamic contrast-enhanced (DCE) cardiac MRI (CMRI) is an established medical imaging modality for detecting coronary artery disease and stress-induced myocardial blood flow abnormalities. Free-breathing CMRI protocols are preferred over breath-hold exam protocols due to the greater patient comfort and applicability to a wider range of patient cohorts who may not be able to perform consecutive breath-holds during the exam. Once the CMRI data is acquired, a key initial step for accurate analysis of the DCE scan is contouring or segmentation of the left ventricular myocardium. In settings where non-rigid motion correction (MoCo) fails or is unavailable, this process can be a time-consuming and labor-intensive task since a typical DCE scan includes over 300 time frames.

Deep neural network (DNN) models have been proposed as a solution to this exhausting task [3,23,26,28]. However, to ensure trustworthy and reliable results in a clinical setting, it is necessary to identify potential failures of these models. Incorporating a quality control (QC) tool in the DCE image segmentation pipeline is one approach to address such concerns. Moreover, QC tools have the potential to enable a human-in-the-loop framework for DNN-based analysis [15], which is a topic of interest especially in medical imaging [2,18]. In a human-A.I collaboration framework, time/effort efficiency for the human expert should be a key concern. For free-breathing DCE-CMRI datasets, this time/effort involves QC of DNN-derived segmentations for each time frame. Recent work in the field of medical image analysis [5,9,14,20,24,25] and specifically in CMRI [7,16,17,22,27] incorporate QC and uncertainty assessment to assess/interpret DNN-derived segmentations. Still, a QC metric that can both temporally and spatially localize uncertain segmentation is lacking for dynamic CMRI.

Our contributions in this work are two-fold: (i) we propose an innovative spatiotemporal dynamic quality control (dQC) tool for model-agnostic test-time assessment of DNN-derived segmentation of free-breathing DCE CMRI; (ii) we show the utility of the proposed dQC tool for improving the performance of DNN-based analysis of external CMRI datasets in a human-in-the-loop framework. Specifically, in a scenario where only 10% of the dataset can be referred to the human expert for correction, although random selection of cases does not improve the performance ($p = \text{n.s.}$ for Dice), our dQC-guided selection yields a significant improvement ($p < 0.001$ for Dice). To the best of our knowledge, this work is the first to exploit the test-time agreement/disagreement between spatiotemporal patch-based segmentations to derive a dQC metric which, in turn, can be used for human-in-the-loop analysis of dynamic CMRI datasets.

2 Methods

2.1 Training/testing Dynamic CMRI Datasets

Our training/validation dataset (90%/10% split) consisted of DCE CMRI (stress first-pass perfusion) MoCo image-series from 120 subjects, which were acquired using 3T MRI scanners from two medical centers over 48–60 heartbeats in 3 short-axis myocardial slices [29]. The training set was extensively augmented by simulating breathing motion patterns and artifacts in the MoCo image-series, using random rotations (±50°), shear (±10°), translations (±2 pixels), scaling (range: [0.9, 1.1]), flat-field correction with 50% probability ($\sigma \in [0,\ 5]$), and gamma correction with 50% probability ($\gamma \in [0.5,\ 1.5]$). To assess the generalization of our approach, an external dataset of free-breathing DCE images from 20 subjects acquired at a third medical center was used. Local Institutional Review Board approval and written consent were obtained from all subjects.

2.2 Patch-Based Quality Control

Patch-based approaches have been widely used in computer vision applications for image segmentation [1,4] as well as in the training of deep learning models [6,10,12,13,21]. In this work, we train a spatiotemporal (2D+time) DNN to segment the myocardium in DCE-CMRI datasets. Given that each pixel is present in multiple patches, we propose to further utilize this patch-based approach at test-time by analyzing the discordance of DNN inference (segmentation output) of each pixel across multiple overlapping patches to obtain a dynamic quality control map.

Let $\Theta(w)$ be a patch extraction operator decomposing dynamic DCE-CMRI image $I \in \mathbb{R}^{M\times N\times T}$ into spatiotemporal patches $\theta \in \mathbb{R}^{K\times K\times T}$ by using a sliding window with a stride w in each spatial direction. Also, let $\Gamma_{m,n}$ be the set of overlapping spatiotemporal patches that include the spatial location (m,n) in them. Also, $p^i_{m,n}(t) \in \mathbb{R}^T$ denotes the segmentation DNN's output probability score for the i^{th} patch at time t and location (m,n). The binary segmentation result $\mathcal{S} \in \mathbb{R}^{M\times N\times T}$ is derived from the mean of the probability scores from the patches that are in $\Gamma_{m,n}$ followed by a binarization operation. Specifically, for a given spatial coordinate (m,n) and time t, the segmentation solution is:

$$\mathcal{S}_{m,n}(t) = \begin{cases} 1, & \text{if } \frac{1}{|\Gamma_{m,n}|}\sum_{i=1}^{|\Gamma_{m,n}|} p^i_{m,n}(t) \geq 0.5. \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

The patch-combination operator, whereby probability scores from multiple overlapping patches are averaged, is denoted by $\Theta^{-1}(w)$.

The dynamic quality control (dQC) map $\mathcal{M} \in \mathbb{R}^{M\times N\times T}$ is a space-time object and measures the discrepancy between different segmentation solutions obtained at space-time location (m,n,t) and is computed as:

$$\mathcal{M}_{m,n}(t) = \mathtt{std}(p^1_{m,n}(t), p^2_{m,n}(t), \ldots, p^{|\Gamma_{m,n}|}_{m,n}(t)) \tag{2}$$

where `std` is the standard deviation operator. Note that to obtain $\mathcal{S}$ and $\mathcal{M}$, the same patch combination operator $\Theta^{-1}(w)$ was used with $w_{\mathcal{M}} < w_{\mathcal{S}}$. Further, we define 3 quality-control metrics based on $\mathcal{M}$ that assess the segmentation quality at different spatial levels: pixel, frame, and slice (image series). First, $\mathcal{Q}^{\text{pixel}}_{m,n}(t) \in \mathbb{R}$ is the value of $\mathcal{M}$ at space-time location (m, n, t) normalized by the segmentation area at time t:

$$\mathcal{Q}^{\text{pixel}}_{m,n}(t) := \frac{\mathcal{M}_{m,n}(t)}{\sum_{m,n} \mathcal{S}_{m,n}(t)} \tag{3.1}$$

Next, $\mathcal{Q}^{\text{frame}}(t) \in \mathbb{R}^T$ quantifies the *per-frame segmentation uncertainty* as per-frame energy in $\mathcal{M}$ normalized by the corresponding per-frame segmentation area at time t:

$$\mathcal{Q}^{\text{frame}}(t) := \frac{\|\mathcal{M}(t)\|_F}{\sum_{m,n} \mathcal{S}_{m,n}(t)} \tag{3.2}$$

where $\|\cdot\|_F$ is the Frobenius norm and $\mathcal{M}(t) \in \mathbb{R}^{M\times N}$ denotes frame t of the dQC map $\mathcal{M}$. Lastly, $\mathcal{Q}^{\text{slice}}$ assesses the overall segmentation quality of the acquired myocardial slice (image series) as the average of the per-frame metric along time:

$$\mathcal{Q}^{\text{slice}} := \frac{1}{T}\sum_{t=1}^{T} \mathcal{Q}^{\text{frame}}(t) \tag{3.3}$$

2.3 DQC-Guided Human-in-the-Loop Segmentation Correction

As shown in Fig. 1, to demonstrate the utility of the proposed dQC metric, low confidence DNN segmentations in the test set, detected by the dQC metric $\mathcal{Q}^{\text{frame}}$, were referred to a human expert for refinement who was instructed to correct two types of error: (i) anatomical infeasibility in the segmentation (e.g., non-contiguity of myocardium); (ii) inclusion of the right-ventricle, left-ventricular blood pool, or regions outside of the heart in the segmented myocardium.

2.4 DNN Model Training

We used a vanilla U-Net [19] as the DNN time frames stacked in channels, and optimized cross-entropy loss using Adam. We used He initializer [8], batch size of 128, and linear learning rate drop every two epochs, with an initial learning rate of 5×10^{-4}. Training stopped after a maximum of 15 epochs or if the myocardial Dice score of the validation set did not improve for five consecutive epochs. MATLAB R2020b (MathWorks) was used for implementation on a NVIDIA Titan RTX. CMRI images were preprocessed to a size of $128 \times 128 \times 25$ after localization around the heart. Patch size of $64 \times 64 \times 25$ was used for testing and training, with a patch combination stride of $w_{\mathcal{S}} = 16$ and $w_{\mathcal{M}} = 2$ pixels.

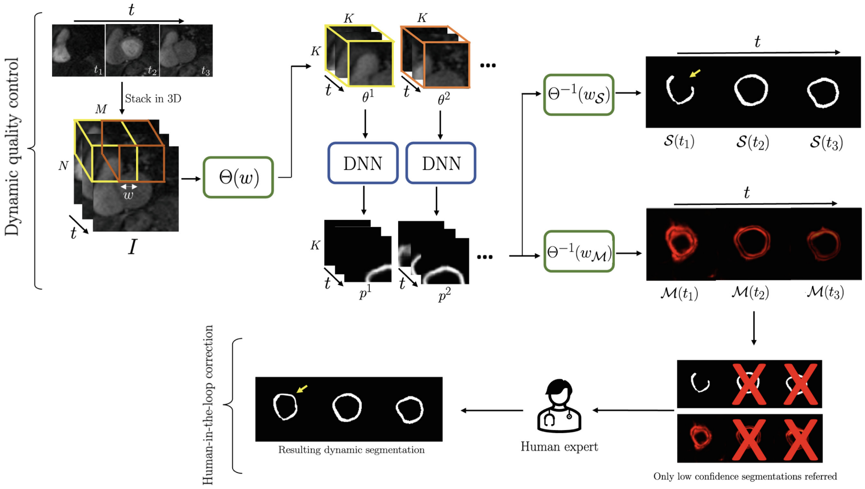

Fig. 1. Pipeline for the proposed dynamic quality control (dQC)-guided human-in-the-loop correction. With patch-based analysis, dQC map $\mathcal{M}$ is obtained and segmentation uncertainty is quantified as a normalized per-frame energy. Only low-confidence segmentations are referred to (and are corrected by) the human.

3 Results

3.1 Baseline Model Performance

The "baseline model" performance, i.e., the DNN output without the human-in-the-loop corrections, yielded an average spatiotemporal (2D+time) Dice score of 0.767 ± 0.042 for the test set, and 16.2% prevalence of non-contiguous segmentations, which is one of the criteria for failed segmentation (e.g., $\mathcal{S}(t_1)$ in Fig. 1) as described in Sect. 2.3. Inference times on a modern workstation for segmentation of one acquired slice in the test set and for generation of the dQC-map were 3 s and 3 min, respectively.

3.2 Human-in-the-Loop Segmentation Correction

Two approaches were compared for human-in-the-loop framework: (i) referring the top 10% most uncertain time frames detected by our proposed dQC tool (Fig. 1), and (ii) randomly selecting 10% of the time frames and referring them for human correction. The initial prevalence of non-contiguous (failed) segmentations among the dQC-selected vs. randomly-selected time frames was 46.8% and 17.5%, respectively. The mean 2D Dice score for dQC-selected frames was 0.607 ± 0.217 and, after human expert corrections, it increased to 0.768 ± 0.147 ($p < 0.001$). On the other hand, the mean 2D Dice for randomly selected frames was initially 0.765 ± 0.173 and, after expert corrections, there was only a small increase to 0.781 ± 0.134 (p = n.s.). Overall, the human expert corrected 87.1% of the dQC-selected and 40.3% of the randomly-selected frames.

Table 1. Spatiotemporal (2D+time) cumulative results comparing the two methods for human-in-the-loop image segmentation.

	Baseline	Random	dQC-guided
Dice score	0.767 ± 0.042	0.768 ± 0.042	0.781 ± 0.039
failure prevalence	16.2%	14.4%	11.3%

Table 1 shows spatiotemporal (2D+time) cumulative results which contain all time frames including not selected frames for correction demonstrating that dQC-guided correction resulted in a notable reduction of failed segmentation prevalence from 16.2% to 11.3%, and in a significant improvement of the mean 2D+time Dice score. In contrast, the random selection of time frames for human-expert correction yielded a nearly unchanged performance compared to baseline. To calculate the prevalence of failed segmentations with random frame selection, a total of 100 Monte Carlo runs were carried out.

3.3 Difficulty Grading of DCE-CMRI Time Frames vs. $\mathcal{Q}^{\text{frame}}$

To assess the ability of the proposed dQC tool in identifying the most challenging time frames in a DCE-CMRI test dataset, a human expert reader assigned "difficulty grades" to each time frame in our test set. The criterion for difficulty was inspired by clinicians' experience in delineating endo- and epicardial contours. Specifically, we assigned the following two difficulty grades: (i) Grade 1: both the endo- and epicardial contours are difficult to delineate from the surrounding tissue; (ii) Grade 0: at most one of the endo- or epicardial contours are challenging to delineate.

To better illustrate, a set of example time frames from the test set and the corresponding grades are shown in Fig. 2. The frequency of Grade 1 and Grade 0 time frames in the test set was 14.7% and 85.3%, respectively. Next, we compared the agreement of $\mathcal{Q}^{\text{frame}}$ values with difficulty grades through a binary classifier whose input is dynamic $\mathcal{Q}^{\text{frame}}$ values for each acquired slice. Note that each $\mathcal{Q}^{\text{frame}}$ yields a distinct classifier due to variation in heart size (hence in dQC maps $\mathcal{M}$) across the dataset. In other words, we obtained as many classifiers as the number of slices in the test set with a data-adaptive approach. The classifiers resulted in a mean area under the receiver-operating characteristics curve of 0.847 ± 0.109.

3.4 Representative Cases

Figure 3 shows two example test cases with segmentation result, dQC maps, and $\mathcal{Q}^{\text{frame}}$. In (a), the highest $\mathcal{Q}^{\text{frame}}$ was observed at $t = 22$, coinciding with the failed segmentation result indicated by the yellow arrow (also see the peak in the adjoining plot). In (b), the segmentation errors in the first 6 time frames (yellow arrows) are accurately reflected by the $\mathcal{Q}^{\text{frame}}$ metric (see adjoining plot) after

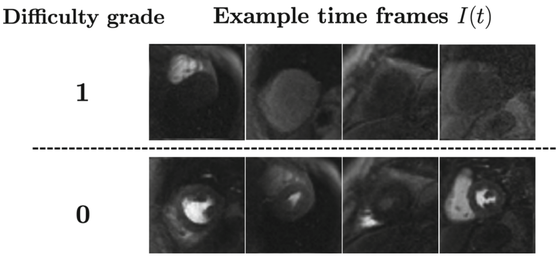

Fig. 2. Examples of DCE time frames corresponding to the two difficulty grades.

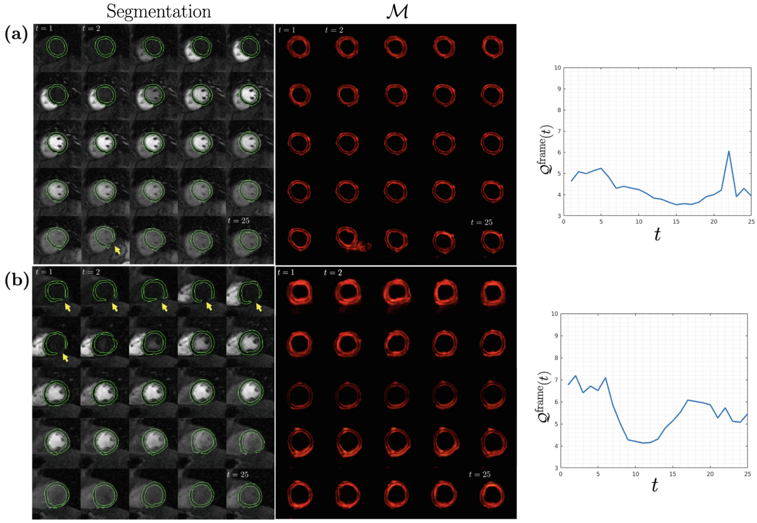

Fig. 3. Two representative DCE-CMRI test cases are shown in along with segmentation, dQC maps $\mathcal{M}$, and the change of dQC metric $\mathcal{Q}^{\text{frame}}(t)$ with time.

which the dQC metric starts to drop. Around $t = 15$ it increases again, which corresponds to the segmentation errors starting at $t = 16$.

4 Discussion and Conclusion

In this work, we proposed a dynamic quality control (dQC) method for DNN-based segmentation of dynamic (time resolved) contrast enhanced (DCE) cardiac MRI. Our dQC metric leverages patch-based analysis by analyzing the discrepancy in the DNN-derived segmentation of overlapping patches and enables automatic assessment of the segmentation quality for each DCE time frame.

To validate the proposed dQC tool and demonstrate its effectiveness in temporal localization of uncertain image segmentations in DCE datasets, we considered a human-A.I. collaboration framework with a limited time/effort budget

(10% of the total number of images), representing a practical clinical scenario for the eventual deployment of DNN-based methods in dynamic CMRI.

Our results showed that, in this setting, the human expert correction of the dQC-detected uncertain segmentations results in a significant performance (Dice score) improvement. In contrast, a control experiment using the same number of randomly selected time frames for referral showed no significant increase in the Dice score, showing the ability of our proposed dQC tool in improving the efficiency of human-in-the-loop analysis of dynamic CMRI by localization of the time frames at which the segmentation has high uncertainty. In the same experiment, dQC-guided corrections resulted in a superior performance in terms of reducing failed segmentations, with a notably lower prevalence vs. random selection (11.3% vs. 14.4%). This reduced prevalence is potentially impactful since quantitative analysis of DCE-CMRI data is sensitive to failed segmentations.

A limitation of our work is the subjective nature of the "difficulty grade" which was based on feedback from clinical experts. Since the data-analysis guidelines for DCE CMRI by the leading society [11] do not specify an objective grading system, we were limited in our approach to direct clinical input. Any such grading system may introduce some level of subjectivity.

Acknowledgement. This work was supported by the NIH awards R01-HL153430 & R01-HL148788, and the Lilly Endowment INCITE award (PI: B. Sharif).

References

1. Bai, W., et al.: A probabilistic patch-based label fusion model for multi-atlas segmentation with registration refinement: application to cardiac MR images. IEEE Trans. Med. Imaging **32**(7), 1302–1315 (2013)
2. Budd, S., Robinson, E.C., Kainz, B.: A survey on active learning and human-in-the-loop deep learning for medical image analysis. Med. Image Anal. **71**, 102062 (2021)
3. Chen, C., et al.: Improving the generalizability of convolutional neural network-based segmentation on CMR images. Front. Cardiovasc. Med. **7**, 105 (2020)
4. Coupé, P., Manjón, J.V., Fonov, V., Pruessner, J., Robles, M., Collins, D.L.: Patch-based segmentation using expert priors: application to hippocampus and ventricle segmentation. Neuroimage **54**(2), 940–954 (2011)
5. DeVries, T., Taylor, G.W.: Leveraging uncertainty estimates for predicting segmentation quality. arXiv preprint arXiv:1807.00502 (2018)
6. Fahmy, A.S., et al.: Three-dimensional deep convolutional neural networks for automated myocardial scar quantification in hypertrophic cardiomyopathy: a multicenter multivendor study. Radiology **294**(1), 52–60 (2020)
7. Hann, E., et al.: Deep neural network ensemble for on-the-fly quality control-driven segmentation of cardiac MRI T1 mapping. Med. Image Anal. **71**, 102029 (2021)
8. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1026–1034 (2015)
9. Hoebel, K., et al.: An exploration of uncertainty information for segmentation quality assessment. In: Medical Imaging 2020: Image Processing, vol. 11313, pp. 381–390. SPIE (2020)

10. Hou, L., Samaras, D., Kurc, T.M., Gao, Y., Davis, J.E., Saltz, J.H.: Patch-based convolutional neural network for whole slide tissue image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2424–2433 (2016)
11. Hundley, W.G., et al.: Society for cardiovascular magnetic resonance (SCMR) guidelines for reporting cardiovascular magnetic resonance examinations. J. Cardiovasc. Magn. Reson. **24**(1), 1–26 (2022)
12. Kuo, W., Häne, C., Mukherjee, P., Malik, J., Yuh, E.L.: Expert-level detection of acute intracranial hemorrhage on head computed tomography using deep learning. Proc. Natl. Acad. Sci. **116**(45), 22737–22745 (2019)
13. Lu, M.Y., Williamson, D.F., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nat. Biomed. Eng. **5**(6), 555–570 (2021)
14. Mehrtash, A., Wells, W.M., Tempany, C.M., Abolmaesumi, P., Kapur, T.: Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. IEEE Trans. Med. Imaging **39**(12), 3868–3878 (2020)
15. Mozannar, H., Sontag, D.: Consistent estimators for learning to defer to an expert. In: International Conference on Machine Learning, pp. 7076–7087. PMLR (2020)
16. Ng, M., et al.: Estimating uncertainty in neural networks for cardiac MRI segmentation: a benchmark study. IEEE Trans. Biomed. Eng. **70**(6), 1955–1966 (2019)
17. Puyol-Antón, E., et al.: Automated quantification of myocardial tissue characteristics from native T1 mapping using neural networks with uncertainty-based quality-control. J. Cardiovasc. Magn. Reson. **22**, 1–15 (2020)
18. Rajpurkar, P., Lungren, M.P.: The current and future state of AI interpretation of medical images. N. Engl. J. Med. **388**(21), 1981–1990 (2023)
19. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
20. Roy, A.G., Conjeti, S., Navab, N., Wachinger, C., Initiative, A.D.N., et al.: Bayesian QuickNAT: model uncertainty in deep whole-brain segmentation for structure-wise quality control. Neuroimage **195**, 11–22 (2019)
21. Rudie, J.D., et al.: Three-dimensional U-Net convolutional neural network for detection and segmentation of intracranial metastases. Radiol. Artif. Intell. **3**(3), e200204 (2021)
22. Sander, J., de Vos, B.D., Išgum, I.: Automatic segmentation with detection of local segmentation failures in cardiac MRI. Sci. Rep. **10**(1), 21769 (2020)
23. Scannell, C.M., et al.: Deep-learning-based preprocessing for quantitative myocardial perfusion MRI. J. Magn. Reson. Imaging **51**(6), 1689–1696 (2020)
24. Wang, G., Li, W., Aertsen, M., Deprest, J., Ourselin, S., Vercauteren, T.: Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. Neurocomputing **338**, 34–45 (2019)
25. Wickstrøm, K., Kampffmeyer, M., Jenssen, R.: Uncertainty and interpretability in convolutional neural networks for semantic segmentation of colorectal polyps. Med. Image Anal. **60**, 101619 (2020)
26. Xue, H., et al.: Automated inline analysis of myocardial perfusion MRI with deep learning. Radiol. Artif. Intell. **2**(6), e200009 (2020)
27. Yalcinkaya, D.M., Youssef, K., Heydari, B., Zamudio, L., Dharmakumar, R., Sharif, B.: Deep learning-based segmentation and uncertainty assessment for automated analysis of myocardial perfusion MRI datasets using patch-level training and

advanced data augmentation. In: 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pp. 4072–4078. IEEE (2021). https://doi.org/10.1109/EMBC46164.2021.9629581

28. Youssef, K., et al.: A patch-wise deep learning approach for myocardial blood flow quantification with robustness to noise and nonrigid motion. In: 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pp. 4045–4051. IEEE (2021). https://doi.org/10.1109/EMBC46164.2021.9629630
29. Zhou, Z., et al.: First-pass myocardial perfusion MRI with reduced subendocardial dark-rim artifact using optimized cartesian sampling. J. Magn. Reson. Imaging **45**(2), 542–555 (2017)

Image Segmentation I