Disagreement-Driven Uncertainty Quantification in Late Gadolinium Enhancement Cardiac MRI

Matthias Schwab[1 [0009−0006−7976−6640]], Markus Haltmeier[2 [0000−0001−5715−0331]], and Agnes Mayr[1 [0000−0001−9363−873X]]

[1] Medical University of Innsbruck, Innsbruck, Tirol, Austria
[2] University of Innsbruck, Innsbruck, Tirol, Austria

Abstract. Accurate segmentation of infarcted myocardium and microvascular obstruction (MVO) in late gadolinium enhancement (LGE) cardiac magnetic resonance (CMR) imaging is critical for risk assessment in myocardial infarction patients. However, the task is challenging due to anisotropic CMR resolution, complex enhancement patterns, and severe class imbalance. In this work, we propose a cascaded deep learning framework that combines a 2D slice-wise segmentation network with a 3D correction network to provide enhanced voxel-wise uncertainty estimation. We introduce a novel uncertainty estimation approach that leverages disagreement between the 2D and 3D models as a proxy for segmentation uncertainty. We quantify this via a Soft Correction Score (SCS), based on probabilistic discrepancies, and a Discrete Correction Map (DCM), which encodes interpretable label corrections between networks. We evaluate our framework on the publicly available EMIDEC dataset and on a large in-house clinical dataset. Across both datasets, our framework achieves superior segmentation accuracy and provides uncertainty estimates comparable to established methods such as Monte Carlo Dropout, test-time augmentation, and deep ensembles. The proposed uncertainty measures correlate strongly with prediction errors and offer interpretable insights into ambiguous regions, enhancing both the reliability and clinical utility of automated LGE CMR analysis.

Keywords: Cardiac MRI · Late gadolinium enhancement · Myocardial infarction · Uncertainty estimation · Deep learning · Segmentation.

1 Introduction

Cardiovascular diseases remain the leading cause of mortality worldwide [12,17], with myocardial infarction being a primary contributor. Accurate characterization of infarcted myocardium and microvascular obstruction (MVO) using late gadolinium enhancement (LGE) cardiac magnetic resonance (CMR) imaging has shown to be critical for risk stratification and clinical decision-making [7,3,2]. However, manual annotation of LGE images is time-consuming and prone to inter-observer variability, which has driven the development of automated segmentation methods.

Recent advances in artificial intelligence (AI) have significantly improved automatic segmentation of LGE CMR images [10,6,8,15,9]. Nonetheless, the segmentation of infarcted tissue and MVO remains challenging due to several factors. First, LGE images typically exhibit high anisotropy, with slice thickness being roughly seven times greater than the in-plane resolution. Second, the spatial distribution of enhancement is often complex, lacking sharp boundaries between healthy and infarcted tissue, which introduces ambiguity. Third, MVO regions are extremely underrepresented, appearing in only approximately 50% to 60% of cases [11] and typically cover less than 10% of the myocardial volume, leading to severe class imbalance.

For clinical deployment, segmentation methods must address these LGE CMR specific challenges and additionally provide uncertainty estimates in regions prone to ambiguity. In prior work, we introduced a cascaded segmentation framework that combines a 2D convolutional neural network (CNN) for per-slice segmentation with a 3D CNN for volumetric refinement [16]. While this architecture effectively addresses key challenges in LGE CMR segmentation, such as anisotropic resolution and complex enhancement patterns, we build upon this established framework to introduce a novel method for uncertainty estimation. While several strategies have been proposed to quantify segmentation uncertainty, including approximate Bayesian inference via Monte Carlo (MC) Dropout [1], deep ensembles [4], or test-time augmentation [18], in this work we exploit the disagreement between the 2D and 3D networks as a proxy for uncertainty. We formalize this intuition and introduce a voxel-wise uncertainty measure which we call Soft Correction Score. Additionally, we introduce an interpretable Discrete Correction Map built to localize clinically relevant misclassifications. We validate our approach on two LGE CMR datasets and demonstrate robust segmentation performance and reliable uncertainty estimation compared to other state-of-the-art uncertainty quantification methods.

2 Method

Our framework comprises three stages: An initial 2D segmentation network, a 3D correction network and a hybrid uncertainty estimation strategy that combines model disagreement with test-time ensemble variance.

2.1 2D Slice-Wise Segmentation

The first stage of the cascade consists of a 2D U-Net [13] trained on individual CMR slices. This network produces initial predictions for all target classes (blood pool, healthy myocardium, infarcted tissue, and MVO) on a per-slice basis. To ensure robustness and generalization, we apply data augmentation including geometric and intensity-based transformations. Preprocessing steps are minimal, involving center cropping around the left ventricle and per-image normalization.

2.2 3D Contextual Correction Network

The second stage refines the 2D predictions using a 3D U-Net, which takes the original volumetric LGE data and the 2D-generated segmentation maps for infarction and MVO as input. This network leverages inter-slice context to correct typical 2D segmentation errors that arise from the absence of 3D spatial information. To train the 3D network effectively, we employ a perturbation module that injects synthetic errors into the 2D segmentation maps. These perturbations simulate realistic error patterns, such as missing or incomplete annotations in individual slices, and encourage the 3D network to learn consistent volumetric corrections. The design ensures that the 3D network does not simply copy 2D predictions, especially by excluding blood pool and healthy myocardium masks from the 2D inputs. For full architectural and training details, we refer the reader to [16].

2.3 Uncertainty Estimation

To enable voxel-wise uncertainty estimation utilizing the cascaded segmentation architecture, we introduce two complementary strategies that capture the disagreement between the 2D and 3D networks. These methods quantify how much the 3D contextual correction network alters the initial 2D predictions, under the assumption that strong corrections indicate low confidence.

Soft Correction Score

As a first uncertainty measure we introduce a *Soft Correction Score* (SCS), which captures the magnitude of change between the softmax probabilities predicted by the 2D and 3D models across multiple folds and stochastic passes via test-time augmentation. Let $P_{2D}, P_{3D} \in \mathbb{R}^{F\times C\times H\times W\times D}$ denote the probabilistic outputs of the 2D and 3D networks across F runs, for C classes and a volume of shape $H \times W \times D$. We define the voxel-wise Soft Correction Score $\mathrm{SCS} \in \mathbb{R}^{H\times W\times D}$ as

$$\mathrm{SCS} = \frac{1}{F}\sum_{f=1}^{F}\left(\frac{1}{C}\sum_{c=1}^{C}\left|P_{2D}^{(f,c)} - P_{3D}^{(f,c)}\right|\right). \tag{1}$$

This score quantifies how much the 3D model modifies the 2D probability distributions, averaged over all predictions and classes. Larger values indicate greater model disagreement and hence higher uncertainty.

Correction Map for Structured Uncertainty

In addition, we compute a *Discrete Correction Map* (DCM) that captures structured disagreements between the 2D and 3D networks using aggregated predictions across folds. This map identifies and encodes correction patterns made by

the 3D network relative to the 2D network, with the goal of providing interpretable insights into model uncertainty that could ultimately assist clinicians in daily practice. Given 5-fold predictions from both the 2D and 3D networks, we compute the mean of the softmax outputs at each voxel and then apply the arg max function to obtain segmentation volumes from both networks, denoted as S_{2D} and $S_{3D} \in \{0, \dots C\}^{H \times W \times D}$. The Discrete Correction Map (DCM) is then defined by comparing these volumes voxel-wise. Specifically, for each pixel x we set

$$\mathrm{DCM}(x) = \begin{cases} S_{2D}(x) \cdot C + S_{3D}(x), & \text{if } S_{2D}(x) \neq S_{3D}(x) \\ 0, & \text{if } S_{2D}(x) = S_{3D}(x) \end{cases},$$

where C is the total number of segmentation classes. In this way, each specific class transition corresponds to a unique correction label in the Discrete Correction Map and reflects the models' disagreement between two classes, such as uncertainty between myocardial scar and healthy myocardium. To improve interpretability and robustness, we apply connected component analysis [14] to each correction pattern and filter out small, isolated components by enforcing a minimum region size threshold of 10 pixels. This ensures that only meaningful, spatially coherent corrections are retained in the final correction map.

3 Experiments

3.1 Datasets

We evaluated our method on the publicly available EMIDEC dataset [5] and on an in-house dataset acquired at the University Clinic of Radiology, Innsbruck. The EMIDEC dataset consists of 100 LGE CMR scans, including both healthy subjects (1/3) and patients with myocardial infarction (2/3). The in-house dataset consists of 442 clinically acquired LGE CMR scans from patients diagnosed with ST-segment elevation myocardial infarction. For both datasets, manual ground truth annotations were provided, including segmentation masks for the left ventricular blood pool, healthy myocardium, myocardial scar, and MVO, following standardized clinical guidelines.

To evaluate segmentation performance and uncertainty quantification, we withheld a total of 60 scans (20 from EMIDEC and 40 from the in-house dataset) as test datasets. The remaining data were used for training.

3.2 Implementation Details

All models were implemented in PyTorch and trained on NVIDIA A40 GPUs using stochastic gradient descent with Nesterov momentum ($\mu = 0.99$). The 2D and 3D U-Nets were trained sequentially with batch sizes of 16 and 4, respectively, using Dice loss as the objective function. Model training followed a 5-fold cross-validation scheme and all models were trained for 750 epochs. Uncertainty estimation was performed using the ensemble of 5 models trained on

the different cross-validation folds. Additionally, test-time augmentation involving flipping along the horizontal and vertical axes was applied to each model in the ensemble.

3.3 Reference Methods for Uncertainty Quantification

We evaluated our disagreement-based uncertainty estimation against several uncertainty quantification strategies used in medical image segmentation:

- **MC Dropout** [1]: We applied dropout layers ($p = 0.1$) during inference with 10 stochastic forward passes and computed the voxel-wise variance from the resulting class probabilities to obtain an uncertainty map.
- **Test-Time Augmentation (TTA)** [18]: We generated predictions over 3 augmented variants of the input by flipping the volume along horizontal and vertical axes, and computed voxel-wise variance of the softmax outputs as the uncertainty score.
- **Deep Ensembles** [4]: We trained an ensemble of 5 independently initialized models, each on a different fold of the data. The variance across the ensemble outputs served as an uncertainty estimate.
- **Deep Ensembles + TTA**: We applied TTA to each model in the deep ensemble. We then computed the voxel-wise variance across all augmented ensemble outputs to produce a more robust and comprehensive uncertainty estimate.

3.4 Results

Uncertainty Evaluation

To evaluate the performance of our method in terms of both segmentation accuracy and uncertainty estimation, we used several standard metrics. Segmentation accuracy was quantified using the Dice similarity coefficient (DCS), reported as mean Dice coefficient across all classes and patients. For uncertainty estimation, we assess the quality of the predicted uncertainty maps by measuring how well they correlate with actual prediction errors. Specifically, we report the Area Under the Receiver Operating Characteristic Curve (AUC), which reflects how well uncertainty values can discriminate between correct and incorrect predictions. Additionally, we computed Pearson and Spearman correlation coefficients between voxel-wise uncertainty scores and absolute segmentation errors to measure the consistency and monotonicity of their relationship.

Visual examples of segmentation results alongside the corresponding uncertainty heatmaps generated by our Soft Correction Score are shown in Figure 1. All together, the cascaded segmentation pipeline in combination with the proposed Soft Correction Score demonstrated strong performance in both segmentation accuracy and uncertainty estimation. On the EMIDEC dataset (Table 1), the error correcting 3D network was able to improve the mean Dice score over all classes compared to the 2D baseline methods. Further, AUC (0.85) achieved by

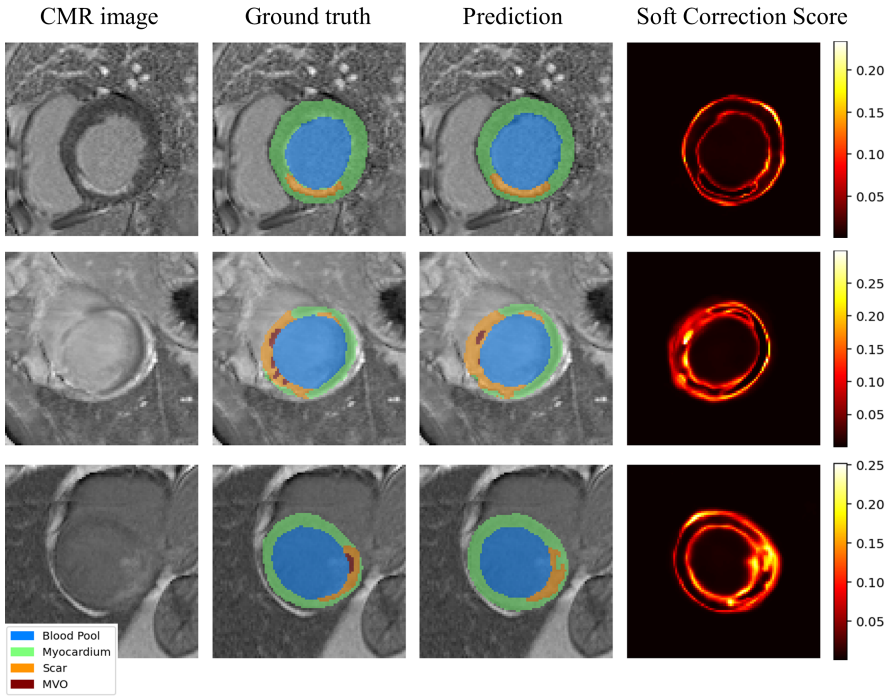

Fig. 1. Examples of segmentation of our framework compared to ground truth and corresponding heatmap using the proposed Soft Correction Score. High uncertainty is observed at class boundaries and in regions with challenging scar and MVO patterns.

the proposed method was comparable to ensemble-based and TTA approaches. Additionally, our method yielded slightly higher Pearson correlation (0.40), indicating a more consistent relationship between uncertainty estimates and voxel-wise prediction errors.

Table 1. Segmentation and uncertainty estimation results on the EMIDEC dataset for different metrics. All scores are reported as mean ± standard deviation across patients. Best values are marked in bold, second-best are underlined.

Method	**DCS (%)**	**AUC**	**Spearman**	**Pearson**
MC Dropout (2D)	71.2 ± 10.7	0.82 ± 0.07	0.39 ± 0.07	0.37 ± 0.06
TTA (2D)	70.9 ± 12.2	0.85 ± 0.03	0.42 ± 0.05	0.33 ± 0.09
Ensemble (2D)	71.8 ± 11.3	$\mathbf{0.86 \pm 0.03}$	$\underline{0.43 \pm 0.04}$	$\underline{0.39 \pm 0.06}$
TTA+Ensemble (2D)	$\underline{73.4 \pm 11.2}$	$\mathbf{0.86 \pm 0.03}$	$\mathbf{0.44 \pm 0.03}$	$\underline{0.39 \pm 0.06}$
Cascade+SCS (Ours)	$\mathbf{76.0 \pm 11.1}$	0.85 ± 0.03	0.39 ± 0.04	$\mathbf{0.40 \pm 0.07}$

Table 2. Segmentation and uncertainty estimation results on the in-house dataset. All scores are reported as mean ± standard deviation across patients. Best values are marked in bold, second-best are underlined.

Method	DCS (%)	AUC	Spearman	Pearson
MC Dropout (2D)	71.4 ± 10.3	0.82 ± 0.12	0.33 ± 0.12	$\underline{0.39 \pm 0.12}$
TTA (2D)	78.6 ± 9.6	0.83 ± 0.12	0.32 ± 0.13	0.31 ± 0.12
Ensemble (2D U-Net)	$\underline{80.2 \pm 8.8}$	0.86 ± 0.09	$\mathbf{0.36 \pm 0.11}$	0.36 ± 0.14
TTA+Ensemble (2D)	79.1 ± 9.2	$\underline{0.86 \pm 0.08}$	0.35 ± 0.10	0.36 ± 0.14
Cascade+SCS (Ours)	$\mathbf{82.4 \pm 8.4}$	$\mathbf{0.88 \pm 0.03}$	$\underline{0.35 \pm 0.06}$	$\mathbf{0.40 \pm 0.08}$

A similar trend was observed on the in-house dataset (Table 2), where our correction-based method yielded the best segmentation results (82.4%) and the highest AUC (0.88) among all evaluated methods. This indicates that incorporating 3D contextual corrections not only improves segmentation quality but also can serve as a very reliable signal for identifying uncertainty in error-prone regions.

Qualitative Analysis

To further explore model uncertainty and error modes, we visualize the Discrete Correction Maps introduced in Section 2.3. These maps highlight structured disagreements between the 2D and 3D networks, offering interpretable insights into where and how the 3D model systematically alters initial 2D predictions. Each non-zero entry represents a specific class transition, thereby distinguishing the two principal classes of uncertainty and connected component filtering ensures that the map highlights only spatially coherent and clinically relevant corrections. As illustrated in Figure 2 the structures displayed in the correction maps correspond with anatomically plausible regions of inherent ambiguity. Representative cases include boundary regions between myocardial scar tissue and adjacent structures such as the left ventricular blood pool or epicardial fat, where signal intensity contrast is inherently low (see Figure 2, first and second line). Another common source of ambiguity arises at the interface between viable myocardium and MVO, particularly when complex LGE patterns lead to overlapping intensity distributions between the two tissues (see Figure 2, third line).

Because the Discrete Correction Map localizes clinically relevant misclassifications, it can also guide targeted refinements of the pipeline. The prevalence of border corrections around scar tissue, for example, suggests augmenting the training data with additional boundary cases. Likewise, displaying maps during manual ground truth annotation could alert experts to regions that warrant particular care.

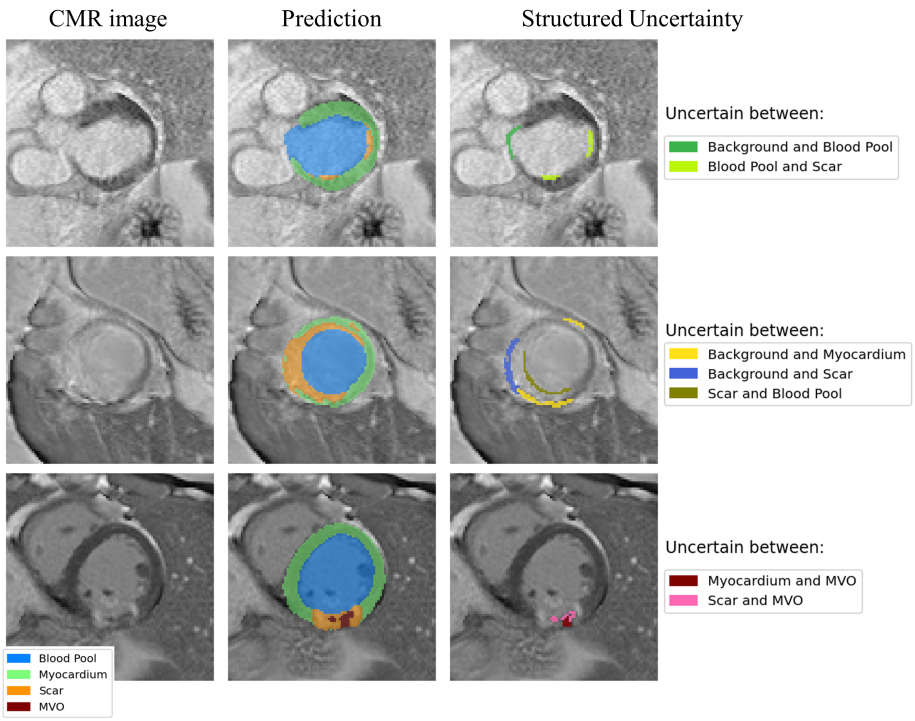

Fig. 2. Discrete Correction Map produced by the 3D contextual model. Voxels in color indicate connected components of class labels that were changed with respect to the 2D baseline.

3.5 Limitations and Future Work

Despite the strong performance of our method in uncertainty-aware segmentation, several limitations remain. First, our approach leverages 3D spatial context, whereas the baseline uncertainty quantification methods operate on 2D architectures. Consequently, some of the observed performance improvements may arise from differences in model capacity rather than the uncertainty estimation strategy itself. A fairer comparison using 3D U-Net backbones with state-of-the-art uncertainty quantification methods is planned for future work. Additionally, our current framework is designed as a two-stage pipeline involving an explicit correction model, which may constrain its applicability to end-to-end or single-model segmentation systems. Generalizing our approach to broader segmentation and classification architectures could be a promising direction. Finally, while we report high correlations between uncertainty scores and segmentation errors, comparing different methods remains challenging due to their distinct underlying assumptions. Further work is needed to examine the interpretability and calibration of the used uncertainty evaluation metrics.

4 Conclusion

We proposed a cascaded segmentation framework for LGE cardiac MR images that integrates voxel-wise uncertainty estimation through model disagreement. By combining a 2D slice-wise network with a 3D contextual correction network, our method addresses key challenges in LGE segmentation, including anisotropic resolution and ambiguous tissue boundaries. To quantify uncertainty, we introduced a Soft Correction Score and a Discrete Correction Map, which capture both probabilistic and structured disagreements between models. These measures correlate strongly with prediction errors and highlight clinically relevant regions of ambiguity. Our approach outperforms standard uncertainty methods across two datasets, achieving high segmentation accuracy and interpretable uncertainty estimates. Overall, our method enhances both the reliability and transparency of infarct and MVO segmentation, bringing it closer to practical clinical use.

Acknowledgments. This research was funded by the Austrian Science Fund (FWF) 10.55776/DOC110.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: Balcan, M.F., Weinberger, K.Q. (eds.) Proceedings of The 33rd International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 48, pp. 1050–1059. PMLR, New York, New York, USA (20–22 Jun 2016), https://proceedings.mlr.press/v48/gal16.html
2. Gerber, B.L., Rousseau, M.F., Ahn, S.A., Le Polain De Waroux, J.B., Pouleur, A.C., Phlips, T., Vancraeynest, D., Pasquet, A., Vanoverschelde, J.L.J.: Prognostic value of myocardial viability by delayed-enhanced magnetic resonance in patients with coronary artery disease and low ejection fraction: impact of revascularization therapy. Journal of the American College of Cardiology **59**(9), 825–835 (2012). https://doi.org/10.1016/j.jacc.2011.09.073
3. Klug, G., Mayr, A., Schenk, S., Esterhammer, R., Schocke, M., Nocker, M., Jaschke, W., Pachinger, O., Metzler, B.: Prognostic value at 5 years of microvascular obstruction after acute myocardial infarction assessed by cardiovascular magnetic resonance. Journal of Cardiovascular Magnetic Resonance **14**(1), 52 (2012). https://doi.org/10.1186/1532-429X-14-46
4. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017), https://proceedings.neurips.cc/paper_files/paper/2017/file/9ef2ed4b7fd2c810847ffa5fa85bce38-Paper.pdf

5. Lalande, A., Chen, Z., Decourselle, T., Qayyum, A., Pommier, T., Lorgis, L., de la Rosa, E., Cochet, A., Cottin, Y., Ginhac, D., Salomon, M., Couturier, R., Meriaudeau, F.: Emidec: A database usable for the automatic evaluation of myocardial infarction from delayed-enhancement cardiac MRI. Data **5**(4) (2020). https://doi.org/10.3390/data5040089
6. Lalande, A., Chen, Z., Pommier, T., Decourselle, T., Qayyum, A., Salomon, M., Ginhac, D., Skandarani, Y., Boucher, A., Brahim, K., de Bruijne, M., Camarasa, R., Correia, T.M., Feng, X., Girum, K.B., Hennemuth, A., Huellebrand, M., Hussain, R., Ivantsits, M., Ma, J., Meyer, C., Sharma, R., Shi, J., Tsekos, N.V., Varela, M., Wang, X., Yang, S., Zhang, H., Zhang, Y., Zhou, Y., Zhuang, X., Couturier, R., Meriaudeau, F.: Deep learning methods for automatic evaluation of delayed enhancement-MRI. The results of the EMIDEC challenge. Medical Image Analysis **79**, 102428 (2022). https://doi.org/10.1016/j.media.2022.102428
7. Larose, E., Rodés-Cabau, J., Pibarot, P., Rinfret, S., Proulx, G., Nguyen, C.M., Déry, J.P., Gleeton, O., Roy, L., Noël, B., Barbeau, G., Rouleau, J., Boudreault, J.R., Amyot, M., De Larochellière, R., Bertrand, O.F.: Predicting late myocardial recovery and outcomes in the early hours of ST-segment elevation myocardial infarction: traditional measures compared with microvascular obstruction, salvaged myocardium, and necrosis characteristics by cardiovascular magnetic resonance. Journal of the American College of Cardiology **55**(22), 2459–2469 (2010). https://doi.org/10.1016/j.jacc.2010.02.033
8. Li, L., Wu, F., Wang, S., Luo, X., Martín-Isla, C., Zhai, S., Zhang, J., Liu, Y., Zhang, Z., Ankenbrand, M.J., et al.: Myops: A benchmark of myocardial pathology segmentation combining three-sequence cardiac magnetic resonance images. Medical Image Analysis **87**, 102808 (2023). https://doi.org/10.1016/j.media.2023.102808
9. Li, S., Wu, C., Feng, C., Bian, Z., Dai, Y., Wu, L.M.: Segmentation of the left ventricle and its pathologies for acute myocardial infarction after reperfusion in LGE-CMR images. IEEE Transactions on Medical Imaging pp. 1–1 (2025). https://doi.org/10.1109/TMI.2025.3573706
10. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 fourth international conference on 3D vision (3DV). pp. 565–571. Ieee (2016). https://doi.org/10.1109/3DV.2016.79
11. Niccoli, G., Scalone, G., Lerman, A., Crea, F.: Coronary microvascular obstruction in acute myocardial infarction. European heart journal **37**(13), 1024–1033 (2016). https://doi.org/10.1093/eurheartj/ehv484
12. Nowbar, A.N., Gitto, M., Howard, J.P., Francis, D.P., Al-Lamee, R.: Mortality from ischemic heart disease: Analysis of data from the world health organization and coronary artery disease risk factors from NCD risk factor collaboration. Circulation: Cardiovascular Quality and Outcomes **12**(6), e005375 (2019). https://doi.org/10.1161/CIRCOUTCOMES.118.005375
13. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. pp. 234–241. Springer (2015). https://doi.org/10.1007/978-3-319-24574-4_28
14. Rosenfeld, A., Pfaltz, J.L.: Sequential operations in digital picture processing. Journal of the ACM (JACM) **13**(4), 471–494 (1966). https://doi.org/10.1145/321356.321357

15. Schwab, M., Pamminger, M., Kremser, C., Haltmeier, M., Mayr, A.: Deep learning pipeline for fully automated myocardial infarct segmentation from clinical cardiac mr scans. Radiology Advances **2**(4), umaf023 (07 2025). https://doi.org/10.1093/radadv/umaf023, https://doi.org/10.1093/radadv/umaf023
16. Schwab, M., Pamminger, M., Kremser, C., Obmann, D., Haltmeier, M., Mayr, A.: Error correcting 2d-3d cascaded network for myocardial infarct scar segmentation on late gadolinium enhancement cardiac magnetic resonance images. Medical Image Analysis **103**, 103594 (2025). https://doi.org/j.media.2025.103594
17. Vos, T., Lim, S.S., Abbafati, C., Abbas, K.M., Abbasi, M., Abbasifard, M., Abbasi-Kangevari, M., Abbastabar, H., Abd-Allah, F., Abdelalim, A., et al.: Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the global burden of disease study 2019. The Lancet **396**(10258), 1204–1222 (2020). https://doi.org/10.1016/S0140-6736(20)30925-9
18. 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). https://doi.org/10.1016/j.neucom.2019.01.103