Using Out-of-Distribution Detection for Model Refinement in Cardiac Image Segmentation*

Francesco Galati[1] and Maria A. Zuluaga[1][0000−0002−1147−766X]

Data Science Department
EURECOM, Sophia Antipolis, France
{galati,zuluaga}@eurecom.fr

Abstract. We introduce a new learning framework that builds upon the recent progress achieved by methods for quality control (QC) of image segmentation to address the poor generalisation of deep learning models in Out-of-Distribution (OoD) data. Under the assumption that the label space is consistent across data coming from different distributions, we use the information provided by a QC module as a proxy of the segmentation model's performance in unseen data. If the model's performance is poor, the QC information is used as feedback to refine the training of the segmentation model, thus adapting to the OoD data. Our method was evaluated in the context of the Multi-Disease, Multi-View & Multi-Center Right Ventricular Segmentation in Cardiac MRI Challenge reporting average Dice Score and Hausdorff distance of 0.905 and 10.472, respectively.

Keywords: Out-of-Distribution Detection · Quality Control · Semi-supervised Model Refinement · Cardiac Segmentation

1 Introduction

Deep learning (DL) techniques have demonstrated the capability to reproduce the analysis of an expert in cardiac image segmentation from cardiac magnetic resonance (CMR) imaging [3]. Despite their success, they still suffer from two major drawbacks that hinder their translation into clinical practice.

First, unlike experts, DL methods may generate anatomically impossible segmentation results [3], which are an important risk in clinical use. Automated quality control (QC) tools have been proposed [14, 15] to assess the quality of a segmentation in the absence of ground truth and flag erroneous segmentations, so they can be discarded from further clinical analysis. The information about the detected erroneous segmentations, which is a surrogate measure of the segmentation model's performance, is generally not incorporated as feedback to the model. Second, DL methods still fail to generalise to out-of-distribution (OoD)

* Supported by the French government, through the 3IA Côte d'Azur Investments in the Future project managed by the National Research Agency (ANR) (ANR-19-P3IA-0002).

samples, i.e. data from other domains than the one of the training set [2]. Unfortunately, this may often be the case at inference time, where new samples may come from a different scanner, acquisition protocol, or population demographics. Labeling data from the unseen OoD domain to re-train the original model is straightforward and yet expensive, labour-intensive, and not scalable to clinical scenarios. In fact, it is hardly viable to obtain an annotated training set that can faithfully represent anatomical variability, different demographics, pathologies, protocols, and scanners. Recently, multiple works have explored alternatives [5, 7, 10, 12, 13] to improve the generalisation capability of CMR segmentation models, typically, under the assumption that it is possible to know if the new data is OoD. In practice, this is not necessarily the case.

We hypothesize that segmentation quality measurements are a proxy of a model's generalisation capabilities. Therefore, this information can be used to improve a model's performance in OoD data. In this work, we propose to use quality measurements to refine a segmentation model's generalisation capabilities, when there is no knowledge about the distribution of the testing data. Under the assumption that the label space is consistent across different distributions, we train a QC assessment module to learn the variability of the ground truth training data. At inference, the QC module provides estimates of a segmentation model's performance on unseen data. The OoD detection information obtained from the QC module is used to refine the segmentation model, allowing it to improve its generalisation capabilities. The proposed method is evaluated in the context of the MICCAI Multi-Disease, Multi-View & Multi-Center Right Ventricular Segmentation in Cardiac MRI (M&Ms-2) Challenge.

Related Work. Previous works on CMR segmentation have addressed poor generalisation by reducing the model's complexity through regularisation [10] or by reducing the number of network parameters [12]. Although these techniques are very effective to tackle overfitting when the training sets are small, there is no guarantee they can mitigate poor generalisation to OoD data.

Data augmentation has been explored to enlarge the training set by simulating various possible data distributions across different domains, applying geometrical operations to the source training data. This technique, however, has shown variable performance across different scenarios [5]. Domain adaptation techniques propose enlarging the training set by combining labeled in-distribution data with OoD data [13, 7] using adversarial training, which is prone to instabilities due to problems, such as mode collapse and non-convergence [1]. Moreover, these methods assume that it is possible to discriminate between in- and out-of-distribution data, which in practice is not necessarily possible.

Our framework differs from previous works in the fact that, first, it is model agnostic. Any given network can be used for segmentation. Second, being formulated as a semi-supervised problem, our QC module avoids the adversarial setup of domain adaptation techniques, thus leading to improved robustness and stability. Finally, differently from all previous approaches, it makes no assumptions about the nature of the distribution of the unseen data.

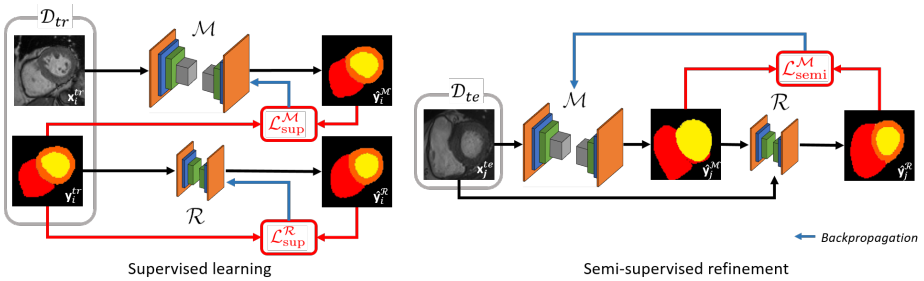

Fig. 1. During the supervised phase, the segmenter $\mathcal{M}$ and the QC module $\mathcal{R}$ are trained independently using $\mathcal{D}_{\text{tr}}$. At inference, if the QC module detects OoD data, a semi-supervised refinement step takes place. The segmenter $\mathcal{M}$ is used to segment unlabelled images from $\mathcal{D}_{\text{te}}$ that are then fed to $\mathcal{R}$. The difference between the reconstruction $\widehat{\mathbf{y}}^R$ and the model's segmentation $\widehat{\mathbf{y}}^M$ is backpropagated to update $\mathcal{M}$.

2 Method

Figure 1 presents an overview of the proposed framework. It consists of a segmentation network ($\mathcal{M}$) and a QC module ($\mathcal{R}$). The information provided by the QC module plays two roles. First, it detects erroneous segmentation masks. Second, the QC information from erroneous segmentation masks is used to refine the segmentation network in a semi-supervised setup to achieve increased performance on OoD data. In the following, we provide a detailed description of these two components (Sec. 2.1), the learning phases (Sec. 2.2), and the mechanism used for OoD data detection and quality control (Sec. 2.3).

2.1 Framework Components

Our framework consists of two elements: a segmentation network and a QC module. The segmentation network or segmenter $\mathcal{M}$ predicts segmentation masks using CMR images as input. The QC module measures the quality of the predicted segmentation masks in the absence of ground truth.

The segmenter $\mathcal{M}$. The segmenter $\mathcal{M}$ learns a function $f_M : X \to Y$, which is used to predict a segmentation mask $\widehat{\mathbf{y}}^M = f_M(\mathbf{x})$. In this sense, it is a standard segmentation network trained in a supervised setting with a training set $\mathcal{D}_{\text{tr}}$. It should be noted that the functioning of the framework is not conceived to depend on a specific segmenter architecture, for which several options in the literature are tailored to cardiac image segmentation [6].

The QC module $\mathcal{R}$. In our framework, we use the QC for image segmentations proposed in [9]. It consists in a convolutional autoencoder, which is trained using in-distribution samples to reconstruct the input segmentation masks through a function $f_R : Y \to Y$, $\widehat{\mathbf{y}}^R = f_R(\mathbf{y}) \approx \mathbf{y}$. In our framework, the in-distribution samples are the ground truth masks from $\mathcal{D}_{\text{tr}}$, i.e. samples without segmentation errors.

Under the assumption that the space Y is consistent across domains, $\mathcal{R}$ is used to obtain $\widehat{\mathbf{y}}^R = f_R(\widehat{\mathbf{y}}^M)$, where $\widehat{\mathbf{y}}^M$ is a predicted segmentation mask and $\widehat{\mathbf{y}}^R$ its reconstruction. Following [9], we use the degree of similarity between $\widehat{\mathbf{y}}^M$ and $\widehat{\mathbf{y}}^R$ as a surrogate measure for QC of the segmentation, which is also a measure of the performance of $\mathcal{M}$ on unseen data. Where the degree of similarity is low (i.e. $\widehat{\mathbf{y}}^R \not\approx \widehat{\mathbf{y}}^M$), the segmentation is considered poor and it is flagged as a potential OoD sample. On such samples, the QC information is backpropagated to refine $\mathcal{M}$. The procedure is detailed in the following section.

2.2 Learning Phases

Supervised Learning. During the supervised phase, $\mathcal{M}$ and $\mathcal{R}$ are trained individually on the available training data $\mathcal{D}_{\text{tr}}$. $\mathcal{M}$ is trained to minimise a loss function measuring the dissimilarity between the ground truth masks $\{\mathbf{y^{tr}}\}$ and the model's prediction $\widehat{\mathbf{y}}^M$, i.e.

$$\mathcal{L}^M_{\text{SUP}} = \mathcal{L}_{\text{GD}}(\widehat{\mathbf{y}}^M, \mathbf{y^{tr}}) + \mathcal{L}_{\text{CE}}(\widehat{\mathbf{y}}^M, \mathbf{y^{tr}}), \tag{1}$$

with $\mathcal{L}_{\text{GD}}(\cdot)$ the generalised Dice loss [16] and $\mathcal{L}_{\text{CE}}(\cdot)$ the cross-entropy loss. Differently from standard supervised training, we keep the best m_{best} segmenters, not the single best, according to their performance in a validation set $\mathcal{D}_{\text{val}}$.

The QC module uses the loss $\mathcal{L}^R = \mathcal{L}_{\text{MSE}}(\widehat{\mathbf{y}}^R, \mathbf{y^{tr}}) + \mathcal{L}_{\text{GD}}(\widehat{\mathbf{y}}^R, \mathbf{y^{tr}})$, where $\mathcal{L}_{\text{MSE}(\cdot)}$ is the mean squared error loss.

Semi-Supervised Refinement of the segmenter $\mathcal{M}$ takes place whenever $\mathcal{R}$ detects poor segmentation quality, i.e. when an unseen segmented sample $\widehat{\mathbf{y}}^M$ is flagged as OoD. The refinement phase (Figure 1) estimates and backpropagates a semi-supervised loss $\mathcal{L}^M_{\text{SEMI}}$, measuring the similarity between the predicted and the reconstructed masks, i.e.

$$\mathcal{L}^M_{\text{SEMI}} = \alpha\, \mathcal{L}_{\text{WGD}}(\widehat{\mathbf{y}}^M, \widehat{\mathbf{y}}^R) + \mathcal{L}^M_{\text{SUP}}, \tag{2}$$

where $\mathcal{L}_{\text{WGD}}$ is the weighted generalised Dice loss, giving more importance to the RV, and $\mathcal{L}^M_{\text{SUP}}$ the supervised loss from Eq. 1, which differently from $\mathcal{L}_{\text{WGD}}$ relies on annotated training data.

Finally, the scaling factor α controls the reliability of the pseudo ground truth $\widehat{\mathbf{y}}^R$. When the semi-supervised refinement process starts, $\widehat{\mathbf{y}}^R$ is highly reliable. As it advances, $\mathcal{R}$ becomes a less reliable source for QC and $\widehat{\mathbf{y}}^R$ should be less trusted. To account for this, we set $\alpha = 1/k$, where k is a learning epoch.

2.3 OoD Detection and QC-based Candidate Selection

At inference time, the selected m_{best} segmenters predict candidate segmentation masks $\widehat{\mathbf{y}}^M$ from unseen data $\mathcal{D}_{\text{te}}$, whereas $\mathcal{R}$ reconstructs $\widehat{\mathbf{y}}^R$ to provide a quality measure for each $\widehat{\mathbf{y}}^M$. Following [9], we use the Dice Coefficient (high is best) and the Hausdorff Distance (low is best) as segmentation quality pseudo-measures. We denote them pDC and pHD, respectively.

For increased robustness, we determine the OoD detection thresholds based on the distributions of pDC and pHD. For each set of estimated pseudo-measures, we obtain the first and third quartile, Q_1 and Q_3 and define the following lower and upper thresholds:

$$th_{\text{low}} = Q_1^{\text{pDC}} - 1.5(Q_3^{\text{pDC}} - Q_1^{\text{pDC}}), \tag{3}$$

$$th_{\text{up}} = Q_3^{\text{pHD}} + 1.5(Q_3^{\text{pHD}} - Q_1^{\text{pHD}}). \tag{4}$$

For a test image $\mathbf{x^{te}} \in \mathcal{D}_{\text{te}}$ with m_{best} predicted segmentations, if $\text{pDC} < th_{\text{low}}$ or $\text{pHD} > th_{\text{up}}$ $\forall$ $\{\widehat{\mathbf{y}}_i^M\}_{i=1}^{m_{\text{best}}}$, the sample is considered OoD and model refinement takes place. The procedure is performed until no more OoD samples are detected in $\mathcal{D}_{\text{te}}$ or until a number of semi-supervised refinement epochs K_{semi} is reached.

After semi-supervised refinement, the best $\widehat{\mathbf{y}}^M$ is chosen through a QC-based selection procedure balancing pDC and pHD. The procedure is as follows:

1. The m_{best} candidate models are ranked according to the pDC of their prediction $\widehat{\mathbf{y}}_i^M$.
2. The top ranked candidate model is assessed. If $\text{pHD} < th_{\text{up}}$, its prediction is selected.
3. Otherwise, we discard the model and consider the next best ranked model.
4. Repeat points 2-3 until a segmentation is chosen.
5. If none of the m_{best} models meets the requirements, $\widehat{\mathbf{y}}^M$ is set to the average between the best prediction according to the pDC and to the pHD.

We highlight that QC-based candidate selection is completely independent from semi-supervised refinement. While we use the thresholds th_{low} and th_{up} to both determine when $\mathcal{M}$ should be refined and select a model during inference, the latter procedure could be directly applied to already static trained models where no refinement is desired (equivalent to setting $K_{\text{semi}} = 0$).

3 Experiments and Results

3.1 Experimental Setup

Data & Setup. The proposed method was evaluated in the context of the M&Ms-2 challenge. The goal of the challenge was to segment the right ventricle (RV) from CMR images in two different views: short axis (SA) and long axis (LA). The challenge cohort was composed of 360 patients with different RV and left ventricle (LV) pathologies as well as healthy subjects, who were scanned in four clinical centres in two different countries using four different magnetic resonance scanner vendors. The training set contained 200 annotated images, of which 160 were used for training and 40 for validation on the challenge website. The remaining data were used for testing.

The accuracy of the segmentation masks was measured using the Dice Similarity Coefficient (DC) and the Hausdorff Distance (HD). Further details about the data and the challenge can be found in the challenge website[1].

[1] https://www.ub.edu/mnms-2/

!t

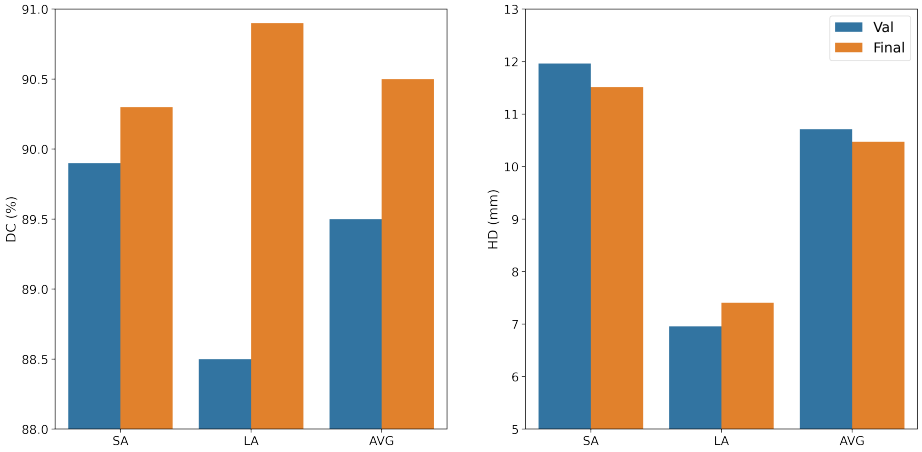

Fig. 2. Short axis (SA), long axis (LA) and average (AVG) dice score (DC) and Hausdorff Distance (HD) obtained on the validation (Val) and final submissions.

Implementation. We chose to use rather simple segmenter models $\mathcal{M}$. In particular, we tested our framework plugging 2 different 2D U-Net architectures, i.e. the 2D models from the 2D-3D U-Net ensembles winning respectively the M&Ms Challenge (BN1) [8] and the ACDC Challenge (BN2) [11]. For the QC module, we used the implementation from [9]. We obtained separate segmenter/QC module paired sets for each view. We set $m_{\text{best}} = 3$. The networks were implemented in PyTorch and trained on Google Colab Pro, alternating Tesla P100 and Tesla V100 GPUs. All our code is publicly available[2].

3.2 Results

Figure 2 presents the performance achieved by the model of each view and the overall average performance (AVG) in the validation and final submissions of the challenge. Despite using a simple segmenter model $\mathcal{M}$ (BN1), our results show competitive performances. In particular, performance in the SA is comparable to that reported for the RV in the ACDC Challenge [3] and M&Ms Challenge [4] using more complex 3D segmenter models.

Table 1 displays the performance of our final model stratified according to the conditions present in both the training and test sets. As expected, the performances look quite homogeneous, without any evident drop on the two conditions, tricuspidal regurgitation and dilated right ventricle, that were not present in the training set. Only the LA-view for dilated right ventricle shows a drop in performance indicating poor generalisation.

We performed an ablation study to gain an understanding of the properties of our framework. We analysed performance in three scenarios: 1) using the best

[2] https://github.com/robustml-eurecom/MnMs2

Table 1. Short axis (SA) and long axis (LA) dice score (DC) and Hausdorff Distance (HD) per pathology. Tricuspidal Regurgitation and Dilated Right Ventricle were excluded from the labelled training set by the challenge organisers.

	SA		LA	
	DC	HD	DC	HD
Normal	0.902 ± 0.064	10.89 ± 5.16	0.918 ± 0.062	7.05 ± 6.34
Tetralogy of Fallot	0.899 ± 0.065	12.80 ± 6.22	0.916 ± 0.038	8.33 ± 4.67
Interatrial Communication	0.869 ± 0.137	12.09 ± 4.98	0.912 ± 0.083	8.03 ± 7.48
Congenital Arrhytmogenesis	0.911 ± 0.069	12.70 ± 10.1	0.929 ± 0.029	5.60 ± 1.72
Hypertrophic Cardiomyopathy	0.908 ± 0.083	10.67 ± 5.91	0.916 ± 0.058	6.30 ± 3.30
Dilated Left Ventricle	0.897 ± 0.077	11.99 ± 6.53	0.906 ± 0.053	8.13 ± 9.30
Tricuspidal Regurgitation	0.909 ± 0.051	12.10 ± 8.08	0.910 ± 0.041	6.06 ± 2.98
Dilated Right Ventricle	0.913 ± 0.045	10.83 ± 4.82	0.881 ± 0.062	9.68 ± 6.58

segmenter, according to the validation set, after supervised training ($\mathcal{M}$); 2) using QC-based candidate selection to choose the best model from the m_{best} candidates at inference time without performing refinement ($\{\mathcal{M}\}$+QC); and 3) using the complete framework (full). Figure 3 summarizes the results using the two different base networks (BN1, BN2).

The results reveal the effects of QC on the segmentation performance. Overall, the use of QC information leads to improvements in the segmentation accuracy across different segmenter models, especially for the HD, which by definition is a measure more prone to anomalies. Figure 4 illustrates this through an example of failure in the segmentation process, solved after the proposed semi-supervised refinement. However, in one case (LA segmentation with segmenter model BN1), the use of the full framework reports a drop in performance. Since

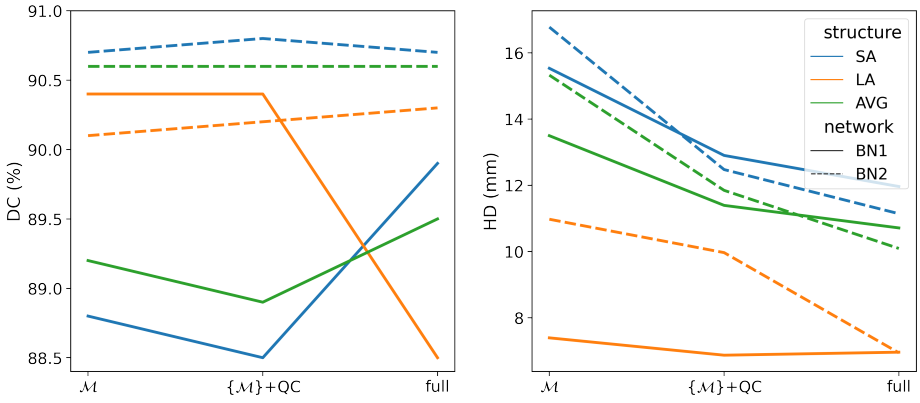

Fig. 3. Ablation study. DC (↑) and HD (↓) results for two segmenters (BN1, BN2) under three configurations: standard supervised learning ($\mathcal{M}$), QC-based model selection ($\{\mathcal{M}\}$+QC) and the full framework (full).

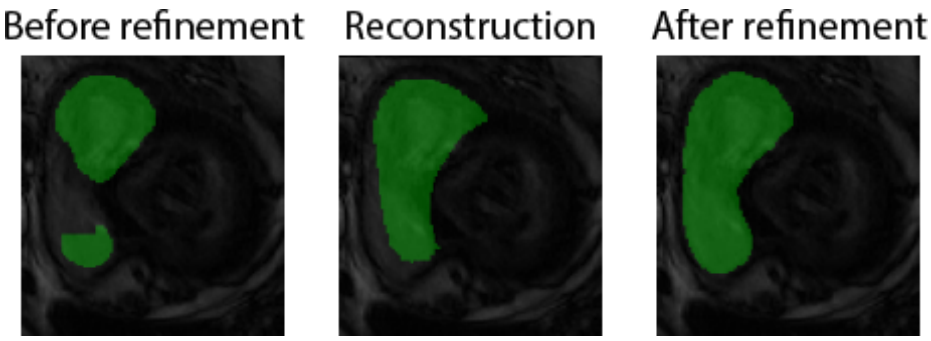

Fig. 4. A visual example of right ventricle segmentation before (left) and after semi-supervised refinement (right). In the middle, the output of the QC module when plugged with the initial segmentation on the left.

using QC to select a model does not degrade the performance, we explain this behavior as a failure of the semi-supervised training stage.

4 Conclusions

In this work, we investigate how to couple QC information to refine a segmentation model to increase its performance on unseen OoD data. Differently from previous works, our framework assumes no previous knowledge about the distribution of the unseen data, yet it is able to determine when the model should be refined (OoD images) or not (in-distribution images). We evaluated our framework within the M&Ms-2 Challenge reporting performances which are comparable to those of similar previous challenges [3, 4], while using a simple 2D segmenter model. This encourages us to pursue this line of work by investigating more sophisticated mechanisms to establish the OoD detection threshold. This, in fact, remains an open problem in the anomaly detection literature.

References

1. Arjovsky, M., Bottou, L.: Towards principled methods for training generative adversarial networks. In: 5th International Conference on Learning Representations, ICLR 2017. OpenReview.net, Toulon, France (2017)
2. Bai, W., Sinclair, M., Tarroni, G., Oktay, O., Rajchl, M., Vaillant, G., Lee, A.M., Aung, N., Lukaschuk, E., Sanghvi, M.M., Zemrak, F., Fung, K., Paiva, J.M., Carapella, V., Kim, Y.J., Suzuki, H., Kainz, B., Matthews, P.M., Petersen, S.E., Piechnik, S.K., Neubauer, S., Glocker, B., Rueckert, D.: Automated cardiovascular magnetic resonance image analysis with fully convolutional networks. Journal of Cardiovascular Magnetic Resonance **20**(1), 65 (2018)
3. Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.A., Cetin, I., Lekadir, K., Camara, O., Ballester, M.A.G., Sanroma, G., Napel, S., Petersen, S., Tziritas, G., Grinias, E., Khened, M., Kollerathu, V.A., Krishnamurthi, G.,

Rohe, M.M., Pennec, X., Sermesant, M., Isensee, F., Jager, P., Maier-Hein, K.H., Full, P.M., Wolf, I., Engelhardt, S., Baumgartner, C.F., Koch, L.M., Wolterink, J.M., Isgum, I., Jang, Y., Hong, Y., Patravali, J., Jain, S., Humbert, O., Jodoin, P.M.: Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved? IEEE Transactions on Medical Imaging **37**(11), 2514–2525 (2018)

4. Campello, V.M., Gkontra, P., Izquierdo, C., Martin-Isla, C., Sojoudi, A., Full, P.M., Maier-Hein, K., Zhang, Y., He, Z., Ma, J., Parreno, M., Albiol, A., Kong, F., Shadden, S.C., Acero, J.C., Sundaresan, V., Saber, M., Elattar, M., Li, H., Menze, B., Khader, F., Haarburger, C., Scannell, C.M., Veta, M., Carscadden, A., Punithakumar, K., Liu, X., Tsaftaris, S.A., Huang, X., Yang, X., Li, L., Zhuang, X., Vilades, D., Descalzo, M.L., Guala, A., Mura, L.L., Friedrich, M.G., Garg, R., Lebel, J., Henriques, F., Karakas, M., Cavus, E., Petersen, S.E., Escalera, S., Segui, S., Rodriguez-Palomares, J.F., Lekadir, K.: Multi-centre, multi-vendor and multi-disease cardiac segmentation: The M&Ms challenge. IEEE Transactions on Medical Imaging p. in press (2021)
5. Chen, C., Bai, W., Davies, R.H., Bhuva, A.N., Manisty, C.H., Augusto, J.B., Moon, J.C., Aung, N., Lee, A.M., Sanghvi, M.M., et al.: Improving the generalizability of convolutional neural network-based segmentation on CMR images. Frontiers in cardiovascular medicine **7**, 105 (2020)
6. Chen, C., Qin, C., Qiu, H., Tarroni, G., Duan, J., Bai, W., Rueckert, D.: Deep learning for cardiac image segmentation: A review. Frontiers in Cardiovascular Medicine **7**, 25 (2020)
7. Chen, J., Zhang, H., Zhang, Y., Zhao, S., Mohiaddin, R., Wong, T., Firmin, D., Yang, G., Keegan, J.: Discriminative consistent domain generation for semi-supervised learning. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 595–604 (2019)
8. Full, P.M., Isensee, F., Jäger, P.F., Maier-Hein, K.: Studying robustness of semantic segmentation under domain shift in cardiac mri. In: Puyol Anton, E., Pop, M., Sermesant, M., Campello, V., Lalande, A., Lekadir, K., Suinesiaputra, A., Camara, O., Young, A. (eds.) Statistical Atlases and Computational Models of the Heart. M&Ms and EMIDEC Challenges. pp. 238–249. Springer International Publishing, Cham (2021)
9. Galati, F., Zuluaga, M.A.: Efficient model monitoring for quality control in cardiac image segmentation. In: International Conference on Functional Imaging and Modeling of the Heart. pp. 101–111. Springer (2021)
10. Guo, F., Ng, M., Goubran, M., Petersen, S.E., Piechnik, S.K., Neubauer, S., Wright, G.: Improving cardiac mri convolutional neural network segmentation on small training datasets and dataset shift: A continuous kernel cut approach. Medical image analysis **61**, 101636 (2020)
11. Isensee, F., Jaeger, P.F., Full, P.M., Wolf, I., Engelhardt, S., Maier-Hein, K.H.: Automatic cardiac disease assessment on cine-MRI via time-series segmentation and domain specific features. In: Statistical Atlases and Computational Models of the Heart. pp. 120–129 (2017)
12. Khened, M., Kollerathu, V.A., Krishnamurthi, G.: Fully convolutional multi-scale residual densenets for cardiac segmentation and automated cardiac diagnosis using ensemble of classifiers. Medical image analysis **51**, 21–45 (2019)
13. Ouyang, C., Kamnitsas, K., Biffi, C., Duan, J., Rueckert, D.: Data efficient unsupervised domain adaptation for cross-modality image segmentation. In: Interna-

tional Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 669–677 (2019)
14. Puyol-Antón, E., Ruijsink, B., Baumgartner, C.F., Masci, P.G., Sinclair, M., Konukoglu, E., Razavi, R., King, A.P.: Automated quantification of myocardial tissue characteristics from native t1 mapping using neural networks with uncertainty-based quality-control. Journal of Cardiovascular Magnetic Resonance **22**(1) (2020)
15. Robinson, R., Valindria, V.V., Bai, W., Oktay, O., Kainz, B., Suzuki, H., Sanghvi, M.M., Aung, N., Paiva, J.M., Zemrak, F., Fung, K., Lukaschuk, E., Lee, A.M., Carapella, V., Kim, Y.J., Piechnik, S.K., Neubauer, S., Petersen, S.E., Page, C., Matthews, P.M., Rueckert, D., Glocker, B.: Automated quality control in image segmentation: application to the UK biobank cardiovascular magnetic resonance imaging study. Journal of Cardiovascular Magnetic Resonance **21**(1) (2019)
16. Sudre, C.H., Li, W., Vercauteren, T., Ourselin, S., Cardoso, M.J.: Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In: Deep learning in medical image analysis and multimodal learning for clinical decision support. pp. 240–248 (2017)