

A 2-step Deep Learning method with Domain Adaptation for Multi-Centre, Multi-Vendor and Multi-Disease Cardiac Magnetic Resonance Segmentation

Jorge Corral Acero^{1*}, Vaanathi Sundaresan^{2*}, Nicola Dinsdale², Vicente Grau^{1†}, Mark Jenkinson^{2,3,4†}

¹ Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, Oxford OX3 7DQ, UK

² Wellcome Centre for Integrative Neuroimaging, Oxford Centre for Functional MRI of the Brain, Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford OX3 9DU, UK

³ Australian Institute of Machine Learning, Department of Computer Science, University of Adelaide, SA 5005, Australia

⁴ South Australian Health and Medical Research Institute (SAHMRI), North Terrace, Adelaide, SA 5000, Australia

`jor.corral@eng.ox.ac.uk`

Abstract. Segmentation of anatomical structures from Cardiac Magnetic Resonance (CMR) is central to the non-invasive quantitative assessment of cardiac function and structure. Anatomical variability, imaging heterogeneity and cardiac dynamics challenge the automation of this task. Deep learning (DL) approaches have taken over the field of automatic segmentation in recent years, however they are limited by data availability and the additional variability introduced by differences in scanners and protocols. In this work, we propose a 2-step fully automated pipeline to segment CMR images, based on DL encoder-decoder frameworks, and we explore two domain adaptation techniques, domain adversarial training and iterative domain unlearning, to overcome the imaging heterogeneity limitations. We evaluate our methods on the MICCAI 2020 Multi-Centre, Multi-Vendor & Multi-Disease Cardiac Image Segmentation Challenge training and validation datasets. The results show the improvement in performance produced by domain adaptation models, especially among the seen vendors. Finally, we build an ensemble of baseline and domain adapted networks, that reported state-of-art mean Dice scores of 0.912, 0.857 and 0.861 for left ventricle (LV) cavity, LV myocardium and right ventricle cavity, respectively, on the externally validated Challenge dataset, including several unseen vendors, centers and cardiac pathologies.

Keywords: Segmentation, Cardiac Magnetic Resonance, Deep Learning, Domain Adaptation, Data harmonization

* Equally contributed to the work.

† Equally co-supervised the work.

1 Introduction

Cardiac magnetic resonance (CMR) imaging allows for an accurate non-invasive quantification of cardiac function and structure [1–3]. The valuable information that it provides in cardiovascular disease management has been repeatedly shown [4–9]. Nevertheless, the anatomical variability and the intrinsic complexity of cardiac dynamics and geometry represent a challenge, and CMR analysis remains manual in clinical practice [3, 10–12]. Deep learning (DL) has revolutionized medical image analysis in recent years, progressing towards automatic segmentation. However, these approaches are challenged by (1) the limited data availability due to technical, ethical and financial constraints along with confidentiality issues, especially for specific pathologies and (2) the imaging heterogeneity introduced by anatomical variability and the use of different scanners and protocols [13–16].

Various techniques have been proposed for improved and robust performance of DL models under limited training data. The most common technique is data augmentation, including affine and non-affine transformations to populate the space of shape variability [17–20]. Other techniques incorporate modifications in the network architecture such as reducing the number of network parameters to avoid overfitting or using a 2-step segmentation pipeline that improves class balance by zooming into the anatomical region of interest (ROI) [21, 22]. Scanner-induced variation in datasets echoes ‘domain shift’ which leads to the performance of models trained on data from one scanner (the source domain) degrading when applied to another (the target domain). Techniques such as transfer learning [23] have proven successful. However, fine-tuning for every unseen domain is still required. Solutions based on domain adaptation (DA) techniques, which aim to create a single feature representation for all domains which is invariant to domain but discriminative for the task of interest [24], have been proposed to overcome this limitation. One of the successful DA approaches is *domain adversarial training of neural networks* (DANN) [25]. DANN assumes that predictions must be based on domain-invariant features, and jointly optimizes the underlying features to simultaneously minimize the loss of the label predictor and maximize the loss of a domain predictor. An alternative approach was proposed using an iterative framework [26] of *domain unlearning* (DU) for adversarial adaptation, creating a classifier which is more uniformly uninformative across domains [27]. This framework has been successfully applied to harmonization of MRI for brain tissue segmentation [28].

The MICCAI 2020 Multi-Centre, Multi-Vendor & Multi-Disease Cardiac Image Segmentation Challenge (M&Ms 2020), including a dataset with a wide variety in centers and scanner vendors and granting external validation, provides a benchmark to assess the generalizability of the segmentation algorithms [29]. We propose a fully automated pipeline to segment CMR short-axis (SAX) stacks, based on the 2-step DL framework that was awarded 1st prize in the LVQuan19 Challenge [22, 30]. We evaluate the segmentation performance of the proposed method, and explore two DA techniques, DANN and DU, for multi-vendor and multi-center applications on the M&Ms 2020 Challenge. The results show that the DA techniques contribute to performance improvement and suggest more experimentation for improvement on unseen vendors

applications. A final model ensemble was built, achieving state-of-art performance for both seen and unseen vendors.

2 Materials and Methods

In essence, the proposed pipeline first locates where the heart is (1st neural network - NN) and then focuses on that ROI to produce a fine segmentation (2nd NN). Three alternative implementations are proposed for this latter step. The pipeline is coupled with pre-processing and post-processing stages, as illustrated on fig. 1.

2.1 Data

We deployed and evaluated our methods on the publicly available M&Ms 2020 Challenge dataset [29], which involves 4 acquisition centers and consists of 150 annotated SAx images from two different MRI vendors, A and B (75 each), and 25 unannotated images from a third vendor, C. The annotated images are segmented only at end-diastole (ED) and end-systole (ES), including left ventricular (LV) cavity, LV endocardium and right ventricle (RV) cavity masks. The image resolution varies from 192x240 to 384x384 pixels, and the pixel spacing, from 0.977 to 1.625 mm. An additional testing set of 200 cases from 6 acquisition centers, including vendors A, B, C and a fourth unseen vendor, D, in equal proportions, is held by the organizers for external validation (20% of the set – 40 cases) and final challenge results (80% of the set). Details on the acquisition and annotation protocols can be checked in [29].

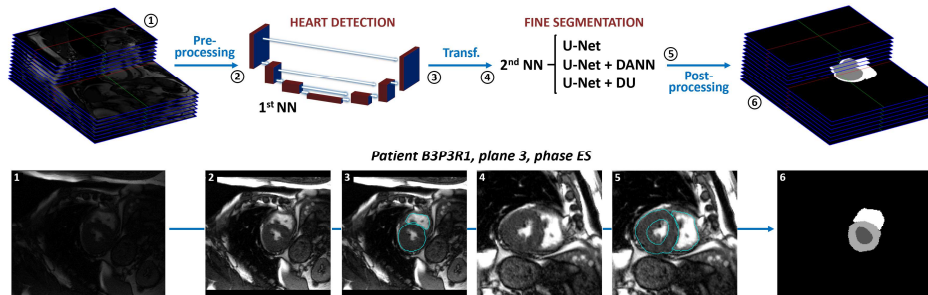


Fig. 1. Scheme of the proposed segmentation pipeline (top) along with a step-by step explanatory illustration of its application on patient B3P3R1 (bottom).

2.2 1st NN: Pre-processing, heart detection and transformation.

The 3 most central ES and ED slices of each patient are normalized in intensity and resolution (pre-processing – fig.1, step 1-2) and fed to the 1st NN that segments the LV epicardium and RV endocardium (heart detection – fig.1, step 2-3), so that a ROI can be defined (transformation – fig.1, step 3-4). The pre-processing consists of a linear interpolation to a 2D pre-defined template (256x256 pixels with symmetrical 1.12mm

pixel spacing) centered in the image, followed by an intensity clipping (10th and 96th percentiles, selected empirically to avoid clipping intensities in important structures) and a minmax normalization to 0-255 (uint8). The 2D mass center coordinates of the LV and RV are provided by the 6 prediction outputs of the 1stNN, and a transformation is calculated from these to align the images to a smaller pre-defined template (144x144 pixels, 1.32mm pixel spacing – balance between small size and high resolution while image large enough to cover the heart) centered in the heart. This same transformation is applied to each slice of the ES and ED SAx stacks. Details of the architecture and implementation of the 1stNN, based on a standard U-Net, can be checked in [22].

2.3 2ndNN: Fine segmentation and postprocessing

The transformed images, linear interpolated to the smaller template and normalized in intensities (5th and 93th percentiles clipping – cutoffs empirically adjusted to the intensities of the structures in the ROI space), are fed to the 2ndNN for a fine segmentation of the LV endo- and epicardium and the RV endocardium (fig.1, step 4-5). The LV myocardium is computed as the region between LV endo and epicardium, and the predictions are back interpolated to the original resolution and rearranged into the original 3D setting in a post-processing step (fig.1, step 5-6). This final step also accounts for segmentation quality enhancement (binarizing predictions, filling holes and removing stray clusters). Three architectures are proposed for the 2ndNN as described below:

U-Net. Baseline method.

A U-Net, fine-tuned for cardiac segmentations is applied for baseline comparisons. Details are available in [19].

Domain adversarial training of neural networks (DANN)

The DANN model, proposed by [25], consists of a feature extractor network with a domain predictor and label predictor. The gradient-reversal layer, placed between the feature extractor and the domain predictor, reverses the gradient direction during back-propagation and maximizes the domain prediction loss, thus minimizing the shift between the domains, while simultaneously making the model discriminative towards the main task of label prediction.

Domain unlearning (DU)

The DU model* [28] is based on the iterative unlearning framework [26] for adversarial adaptation, which rather than using a gradient reversal layer, optimizes two opposing loss functions in a sequential manner: one to maximize the performance of a label predictor given the fixed feature representation, and another to update the feature representation in order to maximally confuse the domain classifier.

* DU code available at: https://github.com/nkdinsdale/Unlearning_for_MRI_harmonisation

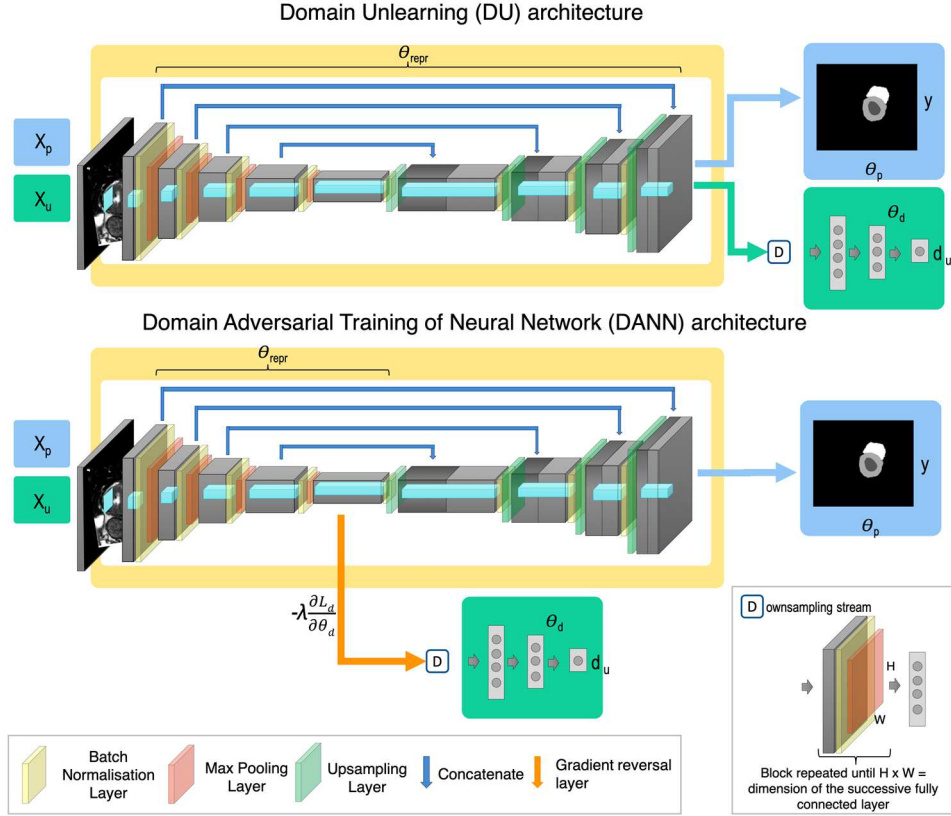


Fig. 2. Architectures of DU (top) and DANN (bottom) models. The models consist of the feature extractor, label predictor and domain predictor with corresponding training parameters θ_{repr} , θ_p and θ_d . The models take input features X_p and input domain information X_u and predicts output labels y , while unlearning output domains d_u . The DU model updates the label predictor, feature extractor and domain predictor in a sequential manner, while in DANN, label prediction and domain unlearning occur simultaneously.

2.4 Implementation details for DANN and DU

For the label predictor of the DANN and DU models, we used the Adam optimizer, with the same parameters as used in the baseline U-Net method. For the domain predictor in DANN and DU, we trained with the Momentum optimizer (with a momentum value of 0.9) and the Adam optimizer respectively. We used a batch size of 16, with a learning rate of 10^{-3} and 10^{-5} for DANN and DU respectively. In addition, in DU we used a beta value of 1000 (a factor used for weighting the domain confusion α). These training hyper parameters were chosen empirically, and a criterion based on a patience (number of epochs to wait for progress on validation set) value of 25 epochs was used to determine model convergence (early stopping). Both models were run on an NVIDIA

6

Tesla V100 GPU. The DANN and DU models took around 50 secs and 45 secs per epoch respectively for training.

2.5 Data augmentation

Standard data augmentation based on random rotations (0 to 360°), translation (1stNN: +/-20mm. 2ndNN: +/-8mm) and flipping was applied during the training of any of the architectures. For each epoch, all the original training images are randomly transformed as described prior to be fed to the NN. Thus, the originals are never used directly, and each instance fed to the NN is only seen once.

2.6 Performance evaluation

Volumetric Dice scores between predicted and original segmentations in the native image space are calculated to assess the fine segmentation performance of the three proposed approaches.

3 Experiments and Results

The 1stNN was trained on the 150 labeled images (A and B vendors) following a 5-fold cross-validation strategy with a training-validation-testing split ratio of 107-13-30 images. The 5 resulting models were used in combination (majority voting) for the prediction of unseen test subjects.

To assess the quantitative performance of the 3 proposed methods on the 2ndNN fine segmentation, for seen and unseen vendors, we carried out 2 experiments:

- Training on the 150 labeled images from vendors A and B (A+B) and cross-validation evaluation on the same set, A+B (seen vendors assessment).
- Training on vendors A and C (A+C), and evaluation on vendor B (unseen vendor assessment). The baseline method was only trained on A, since unlabeled data (vendor C) cannot be incorporated into training, as done in the other 2 methods.

Table 1. Volumetric Dice results from the two proposed experiments. Data presented as median (interquartile range). The best performing models are highlighted in bold. Significant differences ($p < 0.05$) of mean values with respect to the baseline are marked with an asterisk (*).

Train	TEST	METHOD	LV cavity	LV myocardium	RV cavity
A+B	5-fold validation A+B	BASELINE	0.927 (0.886-0.954)	0.845 (0.819-0.879)	0.889 (0.833-0.927)
		DANN	0.948 (0.918-0.964)*	0.873 (0.849-0.896)*	0.926 (0.897-0.943)*
		DU	0.953 (0.927-0.965)*	0.878 (0.855-0.897)*	0.928 (0.898-0.945)*
A+C	B	BASELINE	0.913 (0.869-0.956)	0.829 (0.800-0.853)	0.895 (0.841-0.930)
		DANN	0.927 (0.875-0.957)*	0.838 (0.813-0.866)*	0.891 (0.845-0.936)*
		DU	0.914 (0.864-0.956)	0.835 (0.811-0.862)*	0.892 (0.842-0.934)

In both cases, and for all the 3 methods, we followed a 5-fold cross-validation approach with the split ratio described above. The results are shown in table 1. While both DA techniques significantly improved the segmentation performance for seen vendors, the improvement in performance decreased for unseen vendors and only DANN reports a significantly better performance on average (t -test) than the baseline in segmenting any structure. Indeed, baseline is superior in RV median Dice under unseen vendors.

Finally, the 3 models were trained on A+B+C (only A+B for baseline), following the 5-fold training and split ratio described, and submitted to the M&Ms 2020 Challenge for external validation on unseen samples from vendors A, B, C and D. In addition, an ensemble with the combination in majority voting of the 15-resulting baseline, DANN and DU models (5folds x 3methods) was submitted. The external results, shown in table 2, are consistent with our experiments and confirm the improvement in performance created by DA, especially for unseen vendors. The differences were not significant due to the size of the sets (only 10 patients per vendor). The best performance on average was obtained with the ensemble model in any of the structures.

Table 2. Volumetric Dice results from the external validation set, stratified by vendor (median and interquartile range) and aggregated (average). The best performing models are highlighted in bold.

Method	Vendor	LV cavity	LV myocardium	RV cavity
Baseline	A	0.916 (0.857-0.957)	0.835 (0.808-0.879)	0.862 (0.713-0.896)
	B	0.951 (0.922-0.964)	0.864 (0.852-0.902)	0.942 (0.860-0.951)
	C	0.909 (0.828-0.952)	0.860 (0.827-0.888)	0.870 (0.837-0.921)
	D	0.911 (0.880-0.935)	0.828 (0.793-0.860)	0.871 (0.803-0.897)
	Average	0.904	0.844	0.852
DANN	A	0.936 (0.876-0.961)	0.863 (0.827-0.889)	0.878 (0.823-0.909)
	B	0.951 (0.923-0.968)	0.885 (0.856-0.906)	0.942 (0.895-0.957)
	C	0.903 (0.810-0.956)	0.860 (0.828-0.886)	0.844 (0.803-0.894)
	D	0.914 (0.874-0.944)	0.843 (0.798-0.869)	0.873 (0.816-0.908)
	Average	0.908	0.856	0.853
DU	A	0.918 (0.876-0.957)	0.857 (0.826-0.875)	0.878 (0.822-0.909)
	B	0.954 (0.922-0.970)	0.879 (0.853-0.902)	0.939 (0.890-0.951)
	C	0.898 (0.828-0.955)	0.862 (0.821-0.891)	0.856 (0.832-0.908)
	D	0.926 (0.863-0.947)	0.840 (0.807-0.870)	0.882 (0.811-0.909)
	Average	0.906	0.854	0.855
Ensemble	A	0.923 (0.866-0.961)	0.856 (0.825-0.886)	0.886 (0.807-0.920)
	B	0.954 (0.923-0.967)	0.882 (0.855-0.907)	0.946 (0.884-0.956)
	C	0.911 (0.834-0.956)	0.865 (0.839-0.896)	0.863 (0.833-0.920)
	D	0.922 (0.894-0.944)	0.844 (0.801-0.867)	0.878 (0.820-0.904)
	Average	0.912	0.857	0.861

Based on the previous results, the ensemble was submitted for final testing and M&Ms 2020 Challenge participation (remaining 160 unseen cases). Results along with a best and a worst case, provided by the Challenge organizers, in comparison with the rest of the participants are depicted below in table 3 and figs. 3 and 4. A drop in performance in comparison to the external validation set is observed, especially in vendor B.

8

Table 3. Volumetric Dice results from the final testing set, stratified by vendor (median and inter-quartile range) and aggregated (average). Significant differences ($p < 0.05$) of mean values with respect to the ensemble external validation results are marked with an asterisk (*).

Method	Vendor	LV cavity	LV myocardium	RV cavity
Ensemble	A	0.929 (0.887-0.960)	0.853 (0.810-0.875)	0.892 (0.832-0.925)
	B	0.933 (0.887-0.957)	0.866 (0.838-0.893)	0.903 (0.854-0.936)*
	C	0.924 (0.882-0.946)	0.845 (0.806-0.868)	0.887 (0.824-0.929)
	D	0.914 (0.875-0.929)	0.831 (0.784-0.867)	0.892 (0.847-0.922)
Average		0.902	0.833*	0.863

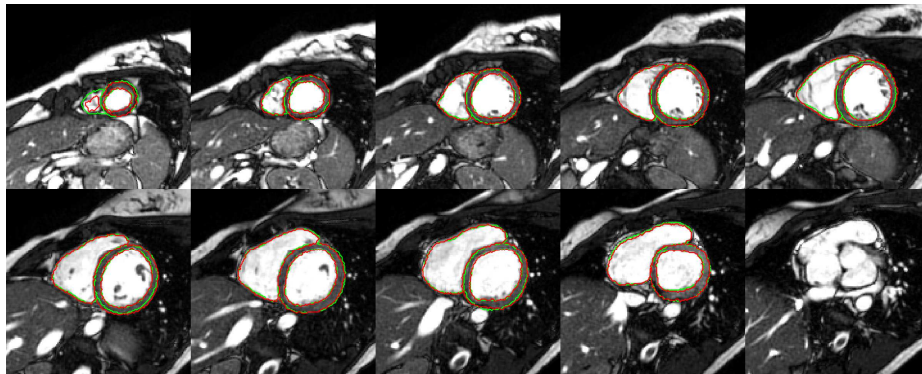


Fig. 3. Best case sample from the test set. Patient G1K1V3, vendor B, ED phase. Ground truth in green, ensemble prediction in red. Dice results of 0.979, 0.891 and 0.954 for LV cavity, LV myocardium and RV cavity, respectively.

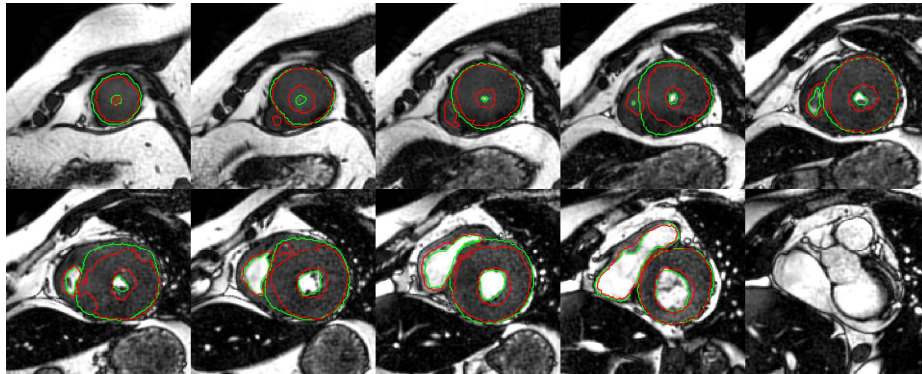


Fig. 4. Worst case sample from the test set. Patient L8N7Z0, vendor B, ES phase. Ground truth in green, ensemble prediction in red. Dice results of 0.653, 0.795 and 0.771 for LV cavity, LV myocardium and RV cavity, respectively.

4 Discussion and conclusions

Our contributions in this work are as follows: (1) we have developed a 2-step DL-based completely automatic pipeline for segmenting LV cavity, LV myocardium and RV cavity on CMR images; (2) we have implemented two DA techniques, domain adversarial training and domain unlearning, to overcome image heterogeneity and scanner-related variations; and (3) we have externally validated our methods on a challenging dataset, including unseen scanners, centers and pathologies, obtaining mean Dice scores of 0.902, 0.833 and 0.863 for LV cavity, LV myocardium and RV cavity, respectively.

The results show that the implemented DA methods significantly improve segmentation performance on seen labelled vendors. However, the 2 proposed DA methods only achieved comparable or slightly superior performance to baseline on the unseen vendors, and so the potential improvement that these 2 methods offer remains inconclusive for multi-vendor applications.

The DA model improvements for seen vendors, especially with DU, illustrated by the A+B experiment in Table 1 can be explained by the fact that both DA models learn a generic feature representation (by unlearning domain-specific information) enabling information from both domains to be incorporated. This essentially increases the size of the training dataset and its variability, improving the label prediction. We believe that the DU model provided better results compared to DANN since the minmax optimization (with gradient reversal) could occasionally become unstable and get trapped in local maxima. On the other hand, feature representations are updated to unlearn domain-specific information with each iteration in DU models. Thus, the label predictor is consistently improved using these feature representations, making the DU model comparatively more stable.

However, despite training DANN and DU models with data from vendor C, there was a lack of improvement on that vendor when compared to baseline results, according to the external validation experiments. A plausible explanation is the additional variability introduced by unseen pathological cases and centers within the same vendor domain. Further investigation is required, after the test data is released, to understand this behavior and to propose changes accordingly via domain shift analysis and hyperparameter fine-tuning.

Comparing the A+B and A+C experiment results, the relatively small drop in performance in the baseline models on the prediction of vendor B suggests the robustness of the baseline method for unseen domain prediction. This is further confirmed by the external validation experiments (see predictions on vendor C and D, table 2). Credit for this robustness should be given to the 2-step implementation, which normalizes for orientation, resolution and appearance and crops the original images to the ROI, leveraging the label imbalance (background vs structure) towards segmentation enhancement.

In an ensemble model, different models are trained with different initializations and each of them learn different aspects of the data. For instance, while the baseline model learns the salient features for label prediction, DANN and DU learn domain invariant features, still retaining task-specific features. Hence, we achieved better overall results in the external validation dataset by combining the predictions using an ensemble model compared to any of the individual models. The dice scores obtained by this model

ensemble in the validation dataset (see table 2) are comparable to those reported in [21], trained on 4875 theoretically healthy subjects, while we only used 150 labelled subjects for training and evaluated on data from multiple centers and vendors, including pathological cases.

While according to the M&Ms 2020 Challenge description the external validation (40 cases) and the final testing set (160 cases) come from the same test set, including samples from different vendors, centers and pathologies in the same proportion, significant differences in performance are obtained for the myocardium structure and for vendor B. A plausible explanation is that the size of the validation set might not be large enough to cover the variability of the test set. We believe that the discrepancies between the testing and validation set across participants is worth exploring to check this hypothesis and to improve the representativeness of the sets in future editions of the Challenge.

To sum up, we have evaluated our method and explored various DA techniques on heterogeneous data from multiple vendors and centers, achieving the best performance with the ensemble of DA and baseline models. We have made our methods publicly available as Singularity containers that may serve as an independent testing tool for the community*. Future directions include further exploration of DA models for improving model generalizability.

Acknowledgements

The authors of this paper declare that the segmentation method they implemented for participation in the M&Ms challenge has not used any pre-trained models nor additional MRI datasets other than those provided by the organizers. This work was supported by the EU's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie (g.a. 764738 to JCA). VS is supported by Wellcome Centre for Integrative Neuroimaging, which has core funding from the Wellcome Trust (203139/Z/16/Z). ND is supported by the Engineering and Physical Sciences Research Council and Medical Research Council (EP/L016052/1). MJ is supported by the National Institute for Health Research (NIHR), Oxford Biomedical Research Centre (BRC) and Wellcome Trust (215573/Z/19/Z). The computational aspects of this research were supported by the Wellcome Trust Core Award (203141/Z/16/Z) and the NIHR Oxford BRC. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health.

* The CMR SAx Segmentation Models presented in this paper are available at:
https://figshare.com/articles/software/A_2-step_Deep_Learning_method_with_Domain_Adaptation_for_Multi-Centre_Multi-Vendor_and_Multi-Disease_Cardiac_Magnetic_Resonance_Segmentation/13140554

References

1. M. B. Stokes and R. Roberts-Thomson, "The role of cardiac imaging in clinical practice.," *Aust. Prescr.*, vol. 40, no. 4, pp. 151–155, Aug. 2017.
2. M. A. Peterzan, O. J. Rider, and L. J. Anderson, "The Role of Cardiovascular Magnetic Resonance Imaging in Heart Failure.," *Card. Fail. Rev.*, vol. 2, no. 2, pp. 115–122, Nov. 2016.
3. P. Peng, K. Lekadir, A. Gooya, L. Shao, S. E. Petersen, and A. F. Frangi, "A review of heart chamber segmentation for structural and functional analysis using cardiac magnetic resonance imaging.," *MAGMA*, vol. 29, no. 2, pp. 155–95, Apr. 2016.
4. E. Dall'Armellina, "From Recognized to Novel Quantitative CMR Biomarkers of LV Recovery," *JACC Cardiovasc. Imaging*, vol. 10, no. 9, pp. 1000–1002, Sep. 2017.
5. S. Desch *et al.*, "Cardiac magnetic resonance imaging parameters as surrogate endpoints in clinical trials of acute myocardial infarction.," *Trials*, vol. 12, p. 204, Sep. 2011.
6. M. Reindl, I. Eitel, and S. J. Reinstadler, "Role of Cardiac Magnetic Resonance to Improve Risk Prediction following Acute ST-elevation Myocardial Infarction," *J. Clin. Med.*, vol. 9, no. 4, p. 1041, Apr. 2020.
7. H. Bulluck, R. Dharmakumar, A. E. Arai, C. Berry, and D. J. Hausenloy, "Cardiovascular Magnetic Resonance in Acute ST-Segment–Elevation Myocardial Infarction," *Circulation*, vol. 137, no. 18, pp. 1949–1964, May 2018.
8. J. L. Anderson and D. A. Morrow, "Acute Myocardial Infarction," *N. Engl. J. Med.*, vol. 376, no. 21, pp. 2053–2064, May 2017.
9. S. Zhang, L. Li, W. Guo, L. Wang, X. Yan, and C. Zhang, "Observation on left ventricular remodeling in acute myocardial infarction.," *Zhonghua nei ke za zhi*, vol. 38, no. 2, pp. 107–9, Feb. 1999.
10. S. E. Petersen *et al.*, "Reference ranges for cardiac structure and function using cardiovascular magnetic resonance (CMR) in Caucasians from the UK Biobank population cohort," *J. Cardiovasc. Magn. Reson.*, vol. 19, no. 1, p. 18, Dec. 2017.
11. A. Suinesiaputra *et al.*, "Quantification of LV function and mass by cardiovascular magnetic resonance: multi-center variability and consensus contours," *J. Cardiovasc. Magn. Reson.*, vol. 17, no. 1, p. 63, Dec. 2015.
12. B. Ibanez *et al.*, "2017 ESC Guidelines for the management of acute myocardial infarction in patients presenting with ST-segment elevation: The Task Force for the management of acute myocardial infarction in patients presenting with ST-segment elevation of the European Society," *Eur. Heart J.*, vol. 39, no. 2, pp. 119–177, Aug. 2017.
13. J. S. Rumsfeld, K. E. Joynt, and T. M. Maddox, "Big data analytics to improve cardiovascular care: promise and challenges," *Nat. Rev. Cardiol.*, vol. 13, no. 6, pp. 350–359, Jun. 2016.
14. K. Shameer, K. W. Johnson, B. S. Glicksberg, J. T. Dudley, and P. P. Sengupta, "Machine learning in cardiovascular medicine: are we there yet?," *Heart*, vol. 104, no. 14, pp. 1156–1164, Jul. 2018.
15. G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, Dec. 2017.
16. A. Hosny, C. Parmar, J. Quackenbush, L. H. Schwartz, and H. J. W. L. Aerts, "Artificial intelligence in radiology," *Nat. Rev. Cancer*, vol. 18, no. 8, pp. 500–510, Aug. 2018.
17. Z. Hussain, F. Gimenez, D. Yi, and D. Rubin, "Differential Data Augmentation Techniques for Medical Imaging Classification Tasks.," *AMIA Symp.*, vol. 2017, pp. 979–984, 2017, Accessed: Jan. 20, 2019.

12

18. L. Perez and J. Wang, "The Effectiveness of Data Augmentation in Image Classification using Deep Learning," *CoRR*, vol. abs/1712.0, 2017.
19. J. Corral Acero, E. Zacur, H. Xu, R. Ariga, A. Bueno-Orovio, P. Lamata and V. Grau, "SMOD - Data Augmentation Based on Statistical Models of Deformation to Enhance Segmentation in 2D Cine Cardiac MRI," in *FIMH2019, LNCS*, vol. 11504 LNCS, pp. 361–369, Jun. 2019
20. J. Lemley, S. Bazrafkan, and P. Corcoran, "Smart Augmentation Learning an Optimal Data Augmentation Strategy," *IEEE Access*, vol. 5, pp. 5858–5869, 2017
21. W. Bai *et al.*, "Automated cardiovascular magnetic resonance image analysis with fully convolutional networks," *J. Cardiovasc. Magn. Reson.*, vol. 20, no. 1, p. 65, Dec. 2018.
22. J. Corral Acero, H. Xu, E. Zacur, P. Lamata, A. Bueno-Orovio and V. Grau, "Left Ventricle Quantification with Cardiac MRI: Deep Learning Meets Statistical Models of Deformation," in *STACOM 2019, LNCS*, vol. 12009 LNCS, pp. 384–394, Oct. 2020.
23. S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.
24. S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," *Mach. Learn.*, vol. 79, no. 1–2, pp. 151–175, Oct. 2010.
25. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., *et al.*, "Domain-Adversarial Training of Neural Networks," *Journal of Machine Learning Research*, 2015.
26. E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko, "Simultaneous Deep Transfer Across Domains and Tasks," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 4068–4076.
27. M. Alvi, A. Zisserman, and C. Nellåker, "Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings," in *LNCS*, vol. 11129 LNCS, pp. 556–572, 2019
28. N. K. Dinsdale, M. Jenkinson, and A. I. L. Namburete, "Unlearning Scanner Bias for MRI Harmonisation in Medical Image Segmentation," Springer, Cham, 2020, pp. 15–25.
29. V. M. Campello, "Multi-Centre, Multi-Vendor & Multi-Disease Cardiac Image Segmentation Challenge (M&Ms) MICCAI 2020." <https://www.ub.edu/mnms/> (accessed Jul. 01, 2020).
30. G. Yang, T. Hua, W. Xue, and L. Shuo, "Left Ventricle Full Quantification Challenge MICCAI 2019," 2019. <https://lvquan19.github.io/> (accessed Jul. 01, 2020).