Multi-Centre, Multi-Vendor and Multi-Disease Cardiac Segmentation: The M&Ms Challenge

Víctor M. Campello, Polyxeni Gkontra, Cristian Izquierdo, Carlos Martín-Isla, Alireza Sojoudi, Peter M. Full, Klaus Maier-Hein, Yao Zhang, Zhiqiang He, Jun Ma, Mario Parreño, Alberto Albiol, Fanwei Kong, Shawn C. Shadden, Jorge Corral Acero, Vaanathi Sundaresan, Mina Saber, Mustafa Elattar, Hongwei Li, Bjoern Menze, Firas Khader, Christoph Haarburger, Cian M. Scannell, Mitko Veta, Adam Carscadden, Kumaradevan Punithakumar, Xiao Liu, Sotirios A. Tsaftaris, Xiaoqiong Huang, Xin Yang, Lei Li, Xiahai Zhuang, David Viladés, Martín L. Descalzo, Andrea Guala, Lucia La Mura, Matthias G. Friedrich, Ria Garg, Julie Lebel, Filipe Henriques, Mahir Karakas, Ersin Çavuş, Steffen E. Petersen, Sergio Escalera, Santi Seguí, José F. Rodríguez-Palomares, and Karim Lekadir

***Abstract*— The emergence of deep learning has considerably advanced the state-of-the-art in cardiac magnetic resonance (CMR) segmentation. Many techniques have been proposed over the last few years, bringing the accuracy of automated segmentation close to human performance. However, these models have been all too often trained and validated using cardiac imaging samples from single clinical centres or homogeneous imaging protocols. This has prevented the development and validation of models that are generalizable across different clinical centres, imaging conditions or scanner vendors. To promote further research and scientific benchmarking in the field of generalizable deep learning for cardiac segmentation, this paper presents the results of the Multi-Centre, Multi-Vendor and Multi-Disease Cardiac Segmentation (M&Ms) Challenge, which was recently organized as part of the MICCAI 2020 Conference. A total of 14 teams submitted different solutions to the problem, combining various baseline models, data augmentation strategies, and domain adaptation techniques. The obtained results indicate the importance of intensity-driven data augmentation, as well as the need for further research to improve generalizability towards unseen scanner vendors or new imaging protocols. Furthermore, we present a new resource of 375 heterogeneous CMR datasets acquired by using four different scanner vendors in six hospitals and three different countries (Spain, Canada and Germany), which we provide as open-access for the community to enable future research in the field.**

***Index Terms*— Cardiovascular magnetic resonance, image segmentation, deep learning, generalizability, data augmentation, domain adaption, public dataset.**

This work was partly funded by the European Union's Horizon 2020 research and innovation program under grant agreement number 825903 (euCanSHare project).

V. M. Campello, P. Gkontra, C. Izquierdo, C. Martín-Isla, and K. Lekadir are with the Artificial Intelligence in Medicine Lab (BCN-AIM), Dept. de Matemàtiques i Informàtica, Universitat de Barcelona, Spain (e-mail: victor.campello@ub.edu).

A. Sojoudi is with Circle Cardiovascular Imaging, Canada.

P. M. Full and K. Maier-Hein are with the Division of Medical Image Computing, German Cancer Research Center, Germany.

Y. Zhang is with the Institute of Computing Technology, Chinese Academy of Sciences, China.

Z. He is with Lenovo Ltd., China.

J. Ma is with the Dept. of Mathematics, Nanjing University of Science and Technology, China.

M. Parreño is with the PRHLT Research Center, Universitat Politècnica de València, Spain.

A. Albiol is with the iTeam Research Institute, Universitat Politècnica de València, Spain.

F. Kong and S. C. Shadden are with the Dept. of Mechanical Engineering, University of California Berkeley, USA.

J. Corral Acero is with the Institute of Biomedical Engineering, Dept. of Engineering Science, University of Oxford, UK.

V. Sundaresan is with the Centre for the Functional MRI of the Brain, Nuffield Dept. of Clinical Neurosciences, University of Oxford, UK.

M. Saber and M. Elattar are with the Research and Development Division, Intixel Co. S.A.E., Egypt and M. E. is also with the Medical Imaging and Image Processing Group, Nile University, Egypt.

H. Li and B. Menze are with the Dept. of Computer Science, Technische Universität München, and H. L. is also with Orbem GmbH, Germany.

F. Khader and C. Haarburger are with ARISTRA GmbH, Germany.

C. M. Scannell is with the School of Biomedical Engineering and Imaging Sciences, King's College London, UK.

M. Veta is with the Department of Biomedical Engineering, Eindhoven University of Technology, the Netherlands.

A. Carscadden and K. Punithakumar are with the Dept. of Radiology & Diagnostic Imaging, University of Alberta, Canada and with the Servier Virtual Cardiac Centre, Mazankowski Alberta Heart Institute, Canada.

X. Liu and S. A. Tsaftaris are with the School of Engineering, University of Edinburgh, UK and S. A. T. is also with the Alan Turing Institute, Turing Fellow London, UK.

X. Huang and X. Yang are with the School of Biomedical Engineering and the Medical UltraSound Image Computing (MUSIC) Lab, Shenzhen University, China.

L. Li and is with School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, China.

X. Zhuang is with the School of Data Science, Fudan University, China.

D. Viladés and M. L. Descalzo are with the Cardiac Imaging Unit, Cardiology Service, Hospital de la Santa Creu i Sant Pau, Universitat Autònoma de Barcelona, Spain.

A. Guala and J. F. Rodríguez Palomares are with the Dept. of Cardiology, CIBERCV, Universitat Autònoma de Barcelona, Vall d'Hebron Institut de Recerca, H. Universitari Vall d'Hebron, Barcelona, Spain

L. La Mura is with the Dept. of Advanced Biomedical Sciences, University of Naples Federico II, Italy.

M. G. Friedrich, R. Garg, J. Lebel and F. Henriques are with the Dept. of Medicine and Diagnostic Radiology, McGill University, Canada.

M. Karakas, E. Çavuş are with the Dept. of Cardiology, University Heart & Vascular Center Hamburg, Hamburg, Germany and DZHK (German Center for Cardiovascular Research).

S. E. Petersen is with the Barts Heart Centre, Barts Health NHS Trust, UK and also with the William Harvey Research Institute, NIHR Barts Biomedical Research Centre, Queen Mary University of London, Charterhouse Square, UK.

S. Escalera and S. Seguí are with the Dept. de Matemàtiques i Informàtica, Universitat de Barcelona, Spain and S. E. is also with the Computer Vision Center, Universitat Autònoma de Barcelona, Spain.

I. Introduction

ACCURATE segmentation of cardiovascular magnetic resonance (CMR) images is an important pre-requisite in clinical practice to reliably diagnose and assess a number of major cardiovascular diseases [1], [2]. Currently, the process typically requires the clinician to provide a significant amount of manual input and correction to accurately and consistently annotate the cardiac boundaries across all image slices and cardiac phases. The automation of such a tedious and time-consuming task has been pursued for a long time by using multiple approaches, such as statistical shape models [3] or cardiac atlases [4]. In the last few years, the advent of the deep learning paradigm has motivated the development of many neural network based techniques for improved CMR segmentation, as listed in a recent review [5]. However, most of these techniques have been all too often trained and evaluated using cardiac imaging samples collected from single clinical centres using similar imaging protocols. While these works have advanced the state-of-the-art in deep learning based cardiac image segmentation, their high performances were reported on samples with relatively homogeneous imaging characteristics.

As an example, the CMR datasets from the Automated Cardiac Diagnosis Challenge (ACDC) dataset [6] have been extensively used to build and test new implementations of deep neural networks for cardiac image segmentation. The top performing technique in the ACDC challenge, proposed by Isensee et al. [7], obtained a very high segmentation accuracy for both the left and right ventricles. However, the ACDC datasets were compiled from 150 subjects scanned at a single clinical centre using the same imaging protocol, which limits the ability of the researchers to develop and test models that can generalize suitably across multiple centres and scanner vendors. Other researchers attempted to encode higher variability by building and testing their models based on much larger datasets obtained from the UK Biobank [8]. For instance, Bai et al. [9] implemented a fully convolutional network that achieved highly accurate results on this large dataset (over 4,875 cases), but the authors concluded that their model might not generalize well to other vendor or sequence datasets.

Some researchers proposed to improve CMR segmentation by training neural networks with images from multiple cohorts [10], [11], but these works do not include methods for addressing domain shifts between training and new unseen cohorts. Other works used data augmentation on models built from single cohorts such as the ACDC [12] or the UK Biobank [13], then tested their techniques on other existing public cohorts, including the Sunnybrook Cardiac Data [14], LV Segmentation Challenge Dataset (LVSC) [15] or RV Segmentation Challenge Dataset (RVSC) [16]. However, these studies are limited by the fact that these different CMR cohorts have been annotated with distinct standard operating procedures (SOPs), which makes it difficult to draw conclusions from the multi-cohort comparative results. Furthermore, such an approach requires a large training dataset from the single centre to model high variability across subjects. Another multi-centre and multi-vendor study conducted by Tao et al. [11] relied solely on private data, which makes it difficult to replicate the results and perform community-driven benchmarking. While these recent works confirmed the difficulties encountered by deep learning models to generalize beyond the training samples, they also support the need for well-defined heterogeneous public datasets that can be used by the community to improve model generalizability through scientific benchmarking.

In this context, the Multi-Centre, Multi-Vendor and Multi-Disease Cardiac Segmentation (M&Ms) Challenge was proposed and organized as part of the Statistical Atlases and Computational Modelling of the Heart (STACOM) Workshop, held in conjunction with the MICCAI 2020 Conference. The M&Ms challenge was set up as part of the euCanSHare international project[1], which is aimed at developing interoperable data sharing and analytics solutions for multi-centre cardiovascular research data. Together with clinical collaborators from six different hospitals in Spain, Canada and Germany, a public CMR dataset was established from 375 participants, scanned with four different scanners (Siemens, Philips, General Electric (GE) and Canon) and annotated using a consistent contouring SOP across centres.

To our knowledge, this dataset is the most diverse resource of CMR studies, which is provided as open-access[2] to promote further research and scientific benchmarking in the development and evaluation of future generalizable deep learning models in cardiac image segmentation. In this paper, we also present and discuss the results of the M&Ms challenge in detail, to which a total of 14 international teams submitted a range of solutions, including different strategies of transfer learning, domain adaptation and data augmentation, to accommodate for the differences in scanner vendors and imaging protocols. The obtained results show the extent of the problem, the promise of the proposed solutions, as well as the need for further research to build fully generalizable tools that can be translated reliably and deployed in routine clinical practice across the globe.

II. Challenge framework

A. Data preparation

TABLE I
INFORMATION FROM CENTRES INCLUDED IN THIS WORK.

	Name	City	Country
1	Hospital Vall d'Hebron	Barcelona	Spain
2	Clínica Sagrada Familia	Barcelona	Spain
3	Universitätsklinikum Hamburg-Eppendorf	Hamburg	Germany
4	Hospital Universitari Dexeus	Barcelona	Spain
5	Clínica Creu Blanca	Barcelona	Spain
6	McGill University Health Centre	Montreal	Canada

A total of six clinical centres from Spain, Canada and Germany (numbered 1 to 6 in this work) contributed to this challenge by providing a different number of CMR studies from different scanner vendors, as detailed in Table I. In total, 375 studies were included in this challenge. The subjects considered for this multi-disease study were selected among

[1] euCanSHare project website: www.eucanshare.eu

[2] The dataset is publicly available at www.ub.edu/mnms

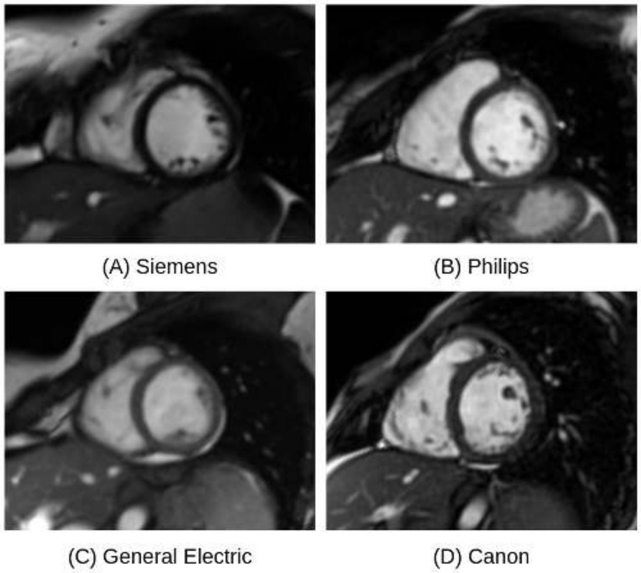
(A) Siemens (B) Philips (C) General Electric (D) Canon

Fig. 1. Visual appearance of a CMR short axis middle slice for anatomically similar subjects in the four different vendors considered.

TABLE II

DISTRIBUTION OF THE MOST FREQUENT PATHOLOGIES AND HEALTHY VOLUNTEERS BETWEEN CENTRES. THE ABBREVIATIONS CORRESPOND TO HYPERTROPHIC CARDIOMYOPATHY (HCM), DILATED CARDIOMYOPATHY (DCM), HYPERTENSIVE HEART DISEASE (HHD), ABNORMAL RIGHT VENTRICLE (ARV), ATHLETE HEART SYNDROME (AHS), ISCHEMIC HEART DISEASE (IHD) AND LEFT VENTRICLE NON-COMPACTION (LVNC).

Centre	1	2	3	4	5	6
Pathology						
Healthy vol.	22	33	32	21	14	3
HCM	25	37	14	8	15	4
DCM	37	-	5	-	9	-
HHD	-	4	-	19	1	1
ARV	12	-	-	2	1	1
AHS	-	-	-	3	-	-
IHD	-	-	-	4	1	3
LVNC	-	-	-	-	2	2
Other	-	-	-	18	7	15

groups of various cardiovascular diseases, such as hypertrophic cardiomyopathy, dilated cardiomyopathy, coronary heart disease, abnormal right ventricle, myocarditis and ischemic cardiomyopathy as well as healthy volunteers (see Table II for more details on the distribution of these cases). The specific scanner manufacturers are: 1) Siemens (Siemens Healthineers, Germany), 2) Philips (Philips Healthcare, Netherlands), 3) General Electric (GE, GE Healthcare, USA) and 4) Canon (Canon Inc., Japan). These four manufacturers were coded as A, B, C and D during the challenge, respectively. The CMR images derived from these four vendors are illustrated in Fig. 1. More specific details on the studies are given in Table III.

Every CMR study was annotated manually by an expert clinician from the centre of origin, with experiences ranging from 3 to more than 10 years. Following the clinical protocol, short-axis views were annotated at the end-diastolic (ED) and end-systolic (ES) phases, as they correspond to the phases used to compute the relevant clinical biomarkers for cardiac diagnosis and follow-up. Three main regions were considered: the left and right ventricle (LV and RV, respectively) cavities and the left ventricle myocardium (MYO). In order to reduce the inter-observer and inter-centre variability in the contours, in particular at the apical and basal regions, a detailed revision of the provided segmentations was performed by four researchers in pairs. They applied the same SOP across all CMR datasets to obtain the final ground truth. To generate consistent annotations for the research community, we chose to apply the SOP that was already used by the ACDC challenge, as follows:

a) The LV and RV cavities must be completely covered, including the papillary muscles.
b) No interpolation of the MYO boundaries must be performed at the basal region.
c) The RV must have a larger surface at the ED time-frame compared to ES.
d) The RV does not include the pulmonary artery.

Clinical delineations as well as later corrections were performed using CVI42 software (Circle Cardiovascular Imaging Inc., Calgary, Alberta, Canada). All studies were provided in DICOM format and contours were extracted in cvi42 workspace format (.cvi42ws). An in-house software was then used to extract the contours and transform the images into the NIFTI format, representing the final files delivered to the challenge participants.

B. *Model training*

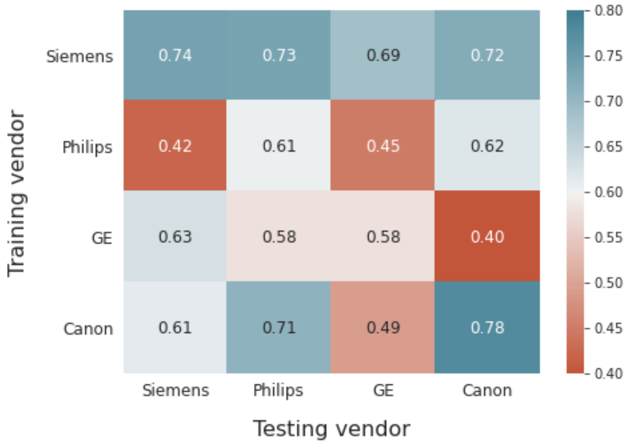

Fig. 2. Degree of generalizability of models trained from the four vendors. Four 2D UNet models [17] were trained with datasets from the four vendors separately (rows) and subsequently tested their segmentation performance on datasets from all vendors (columns). The heatmap shows the Dice similarity coefficient, with a color scale that goes from blue (good generalizability) to red (poor generalizability). The results are the average of 5 models cross-validated on subsets of 30 training subjects.

The 375 CMR studies were divided into three sets, namely training, validation and testing, as detailed in Table IV. To decide on a particular subdivision, we first estimated the degree of generalizability of models trained from the four vendors, as shown in Figure 2. We have thus decided to combine the datasets from vendors A, which generalize relatively well, with

TABLE III
AVERAGE SPECIFICATIONS FOR THE IMAGES ACQUIRED IN THE DIFFERENT CENTRES.

Centre	Vendor	Model	Field strength (T)	In-plane resolution (mm)	Slice thickness (mm)	Number of slices	Number of time frames
1	Siemens	MAGNETOM Avanto	1.5	1.32	9.2	12	25
2	Philips	Achieva	1.5	1.20	9.9	10	30
3	Philips	Achieva	1.5	1.45	9.9	11	26
4	GE	Signa Excite	1.5	1.36	10	12	25
5	Canon	Vantage Orian	1.5	0.85	10	13	29
6	Siemens	MAGNETOM Skyra	3.0	0.98	9.7	12	29

TABLE IV
NUMBER OF STUDIES FOR EACH STEP OF THE CHALLENGE PRESENTED BY CENTRE AND SCANNER VENDOR.

	Siemens		Philips		GE	Canon	Total
Label	A		B		C	D	
Centres	1	6	2	3	4	5	
Training	75	0	50	25	25	0	175
Validation	5	5	5	5	10	10	40
Testing	16	24	19	21	40	40	160
Overall	96	29	74	51	75	50	375

datasets from B, which generalize poorly to new vendors, as training datasets. The participants received the 175 training cases on 1st May 2020, including 75 annotated CMRs from vendor A, 75 annotated CMRs from vendor B, 25 CMRs from vendor C but without any annotations (only the raw images) and no datasets from vendor D, in order to test generalizability to different situations (e.g. image protocol included or not included in the training). Note that in the case of vendor A, the 75 CMRs were included from centre 1 but none from centre 6, to test generalizability across vendors but also across centres for the same vendors. Regarding vendor B, we included more training datasets from centre 2 (50 cases) than from centre 3 (25 cases) to assess the impact of imbalanced training data and fairness in multi-centre cardiac image segmentation. For optimizing the models, the participants were allowed to remotely validate against 40 additional CMRs, i.e. 10 from each of the four vendors. A maximum of 7 submissions were allowed per team during the validation process. Note that during training, it was not allowed to use any external datasets or pre-trained models, to enable a fair comparison between the proposed solutions.

C. Model evaluation

The testing period for the challenge started on 8th June 2020 and concluded on 15th July 2020. The participants had to evaluate their models remotely to ensure the unseen datasets were totally hidden from the segmentation methods. As such, for example, the participants had no prior information on the images provided by vendor D. In order to evaluate the models, the participants were asked to build a Singularity image[3] and share it with the organizers via a MEGA[4] folder shared by the organizers or by any other secure cloud storage service. This Singularity image allows its execution on a similar architecture machine without the need to install all the diversity of used libraries. The necessary computing power was sponsored by NVIDIA, who provided the organizers with access to an NVIDIA V100 GPU card with 16GB of memory, as well as the Barcelona Supercomputing Center (BSC) who provided access to two K80 NVIDIA GPU cards.

In order to assess the quality of the automatically segmented masks P with respect to the ground truth G, four measures were proposed, namely:

(i) Dice similarity coefficient (DSC):

$$DSC(P,G) = \frac{2|P \cap G|}{|P| + |G|} \tag{1}$$

that measures the degree of overlapping of two volumes.

(ii) Jaccard index (JI):

$$JI(P,G) = \frac{|P \cap G|}{|P \cup G|} = \frac{|P \cap G|}{|P| + |G| - |P \cap G|} \tag{2}$$

that measures overlapping as well but is more sensitive to results with average performance.

(iii) Average symmetric surface distance (ASSD):

$$ASSD(P,G) = \frac{1}{|P| + |G|} \left(\sum_{p \in P} d(p,G) + \sum_{g \in G} d(g,P) \right)$$
$$d(p,G) := \inf_{g \in G} d(p,g) \tag{3}$$

that measures the average distance between the two volumes.

(iv) Hausdorff distance (HD):

$$HD(P,G) = \max \left\{ \sup_{p \in P} d(p,G), \sup_{g \in G} d(g,P) \right\} \tag{4}$$

that measures the largest disagreement between the volumes and it is useful for identifying small outliers. All these metrics were computed using the public library medpy[5].

These metrics were computed for the three target labels: LV, RV, and MYO, resulting in a total of 12 measures. In case one participant had a prediction missing for a specific subject, a value of zero was assumed for DSC and JI and maximum values of 150 and 50 milimetres were assumed for HD and ASSD, respectively, based on the worst results obtained by the participating methods. Any value above the thresholds on surface distances was set to the maximum value.

To obtain the final ranking for each team, a weighted average was computed giving a greater importance to the unlabelled and unseen scanner vendors. Therefore, if v_A and v_B are defined as the labelled vendors, v_C, the unlabelled one

[3] https://sylabs.io
[4] https://mega.nz
[5] https://github.com/loli/medpy

and v_D, the unseen one, the weighted sum for a metric M is obtained as follows:

$$M = \frac{1}{6}M_{v_A} + \frac{1}{6}M_{v_B} + \frac{1}{3}M_{v_C} + \frac{1}{3}M_{v_D} \tag{5}$$

Then, a min-max normalization was applied across participants for each measure and a final average over the normalized metrics yielded the performance (P) ranging from 0 to 1, being 1 the value that a team would obtain if it had the best results for every metric.

III. Participating methods

In total, 80 teams registered to download the M&Ms training dataset, 16 submitted a solution for the final testing phase and 14 teams submitted their methodology as a paper to the STACOM Workshop (see Table V for details on these teams). All participants used deep learning as their segmentation approach. Table VI summarizes the main characteristics of the submitted techniques, including the backbone architectures and domain adaptation strategies, which are described in more detail in the following subsections. Furthermore, details on the hardware used during training and the times that each method took for training and inference as well as the number of parameters for each model are presented in Table VII.

A. Backbone architectures

There is a degree of variability in the backbone architectures used between the different participants, as shown in Table VI. Four teams used the nnUNet [33] (which includes UNet architectures in 2D and 3D as well as a cascaded UNet) as their baseline segmentation model (P1-P3 & P9). Four participants used a traditional UNet [17] (P6, P10, P13, P14), while other variants of UNets were adopted by the rest of the teams. In particular, UNets combined with residual connections were applied by three teams (P4, P8, P11), with P8 preferring a residual UNet with dilated convolutions (DRUNet) [34]. P5 proposed the use of an attention UNet [35], while P7 developed a modified UNet based on multi-gate and dilated inception blocks to extract multi-scale features. Lastly, one team (P12) proposed a modified Spatial Decomposition Network (SDN) [36] with an AdaIN [37] decoder.

As pre-processing techniques, all models that provided detailed information about this step performed either image normalization to a unit Gaussian distribution or pixel value rescaling to the range [0,1] (only P6 chose the range [0,255] instead). With regards to image resolution, images were resized based on target size or pixel resolution values in 10 out of 14 methods, while the other methods preferred to keep the original image resolution (P4, P7, P8, P11). In order to obtain squared images, cropping and zero padding were used depending on the desired image size for each case. Additionally, some methods applied intensity clipping between varying ranges to get rid of bright artifacts (P5, P6, P11). Finally, P8 was the only method to apply also a non-local means denoising filter prior to the training process.

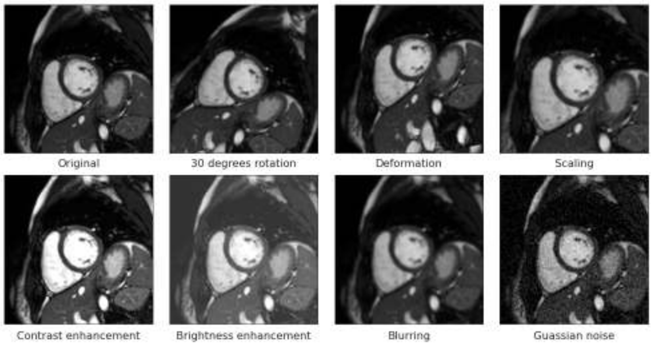

Fig. 3. The effect of data augmentation on a single CMR slice. In the top row, the original image and spatial augmentations are shown. In the bottom row, intensity-based augmentations.

B. Data augmentation

All participants in the challenge (except P11) used some form of data augmentation to enhance their models. Specifically, two families of data augmentations were considered: (1) spatial transformations to increase sample size through rotation, flipping, scaling or deformation of the original images; (2) intensity-driven techniques, which maintain the spatial configuration of the anatomical structures but modify their image appearance. The second type of augmentation seems particularly relevant for the M&Ms as it may increase the variability in image appearance, with the hypothesis that this may lead to improved adaptation to varying imaging protocols and scanner vendors. Two teams performed data augmentation using only spatial transformations (P4, P6). Eleven teams additionally implemented intensity-based transformations using one of two main approaches: (i) standard image transformations such as histogram matching, blurring, change in brightness, gamma and contrast, or addition of Gaussian noise (P1-P3, P7-P8, P10, P13) (see 3 for a visualization of a subset of these transformations on a training slice); (ii) advanced image synthesis by using generative adversarial networks (GANs) (P5, P8, P14) or variational auto-encoders (VAE) (P12). For the latter one, the generation of synthetic images for the unseen vendor D is not feasible since it was not included in the training. Note that the majority of the teams participating in the challenge (10 out of 14) relied solely on data augmentation of the training sample to address the domain-shift problem posed by the M&Ms challenge.

Additionally, some teams (P1-P3, P9, P13) applied test-time augmentation techniques, which consist of passing to the model two or more transformed versions of the same inference image to obtain several predictions. These predictions are then combined to obtain one final outcome, usually by averaging them. This method has been shown to improve the final performance in small data size scenarios and a net improvement with a scale effect that depends on the model architecture [38].

C. Domain adaptation

Of all participants, only three teams (P4, P6, P10) implemented a method to explicitly address the differences in the image distributions between the unseen and trained vendors. At training, P4 constructed a classifier to distinguish between

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2021.3090082, IEEE Transactions on Medical Imaging

TABLE V

LIST AND DETAILS OF THE PARTICIPATING TEAMS IN THE CHALLENGE.

Team	Institution	Location	Name during challenge	Reference
P1	German Cancer Research Center (DKFZ)	Heidelberg, Germany	Mountain goat	[18]
P2	Chinese Academy of Sciences	Beijing, China	Dugong	[19]
P3	Nanjing University of Science and Technology	Nanjing, China	Opossum	[20]
P4	Universitat Politècnica de València	València, Spain	Ox	[21]
P5	University of California	Berkeley, USA	Monkey	[22]
P6	University of Oxford	Oxford, UK	Donkey	[23]
P7	Nile University	Cairo, Egypt	Porpoise	[24]
P8	Technical University of Munich	Munich, Germany	Owl	[25]
P9	Aristra GmbH	Berlin, Germany	Lovebird	[26]
P10	King's College London	London, UK	Mandrill	[27]
P11	University of Alberta	Edmonton, Canada	Muskox	[28]
P12	University of Edinburgh	Edinburgh, UK	Springbok	[29]
P13	Shenzhen University	Shenzhen, China	Seagull	[30]
P14	Fudan University	Shanghai, China	Steer	[31]

TABLE VI

CHARACTERISTICS OF PARTICIPATING MODELS. ABBR: ROTATIONS (R), FLIPPING (F), SCALING (S), DEFORMATIONS (D), HISTOGRAM MATCHING (HM), GAUSSIAN NOISE (GN), BRIGHTNESS (B), GAMMA (G), TEST TIME AUGMENTATION (TTA).

Method	Backbone architecture	Data augmentation										TTA	Domain adaptation
		Spatial augmentations				Intensity-based augmentations							
		R (°)	F	S	D	HM	GN	B	G	Synthesis	Others		
P1	nnUNet	±180	✓	✓	✓		✓	✓	✓		contrast	✓	No
P2	nnUNet	±180	✓	✓	✓	✓			✓		label propagation	✓	No
P3	nnUNet	±180	✓	✓	✓	✓			✓			✓	No
P4	UNet (ResNet-34)	±45	✓		✓						translations		Yes
P5	Attention UNet	±10		✓						CycleGAN	low-level frequency		No
P6	UNet+DA+DUNN	±180	✓								translations		Yes
P7	UNet	±15	✓	✓				✓					No
P8	DRUNet	±15		✓	✓			✓	✓	CycleGAN	blurring		No
P9	nnUNet	±180	✓		✓							✓	No
P10	UNet	±22.5		✓	✓		✓	✓			translations		Yes
P11	UNet++ (ResNet101)												No
P12	SDNet			✓						VAE			No
P13	UNet	±90	✓	✓				✓		WaveCT-AIN [32]	contrast	✓	No
P14	UNet									CycleGAN			No

TABLE VII

TRAINING AND INFERENCE TIME, AND HARDWARE USED, FOR ALL PARTICIPATING METHODS. H, M, S AND MIL. STAND FOR HOURS, MINUTES, SECONDS AND MILLIONS, RESPECTIVELY.

Team	Training time	Inference time (s)	Model parameters (Mil.)	GPU (NVIDIA)
P1	60 h	26	30	Titan XP
P2	48 h	4.8	30	Tesla V100
P3	96-120 h	n/a	30	Tesla V100
P4	6 h	0.35	36	RTX 2080
P5	11 h	10.4	33	GTX 1080 Ti
P6	15 m/epoch	10	28	Tesla V100
P7	8 h	0.0022	6	GTX 1080 Ti
P8	8 h	10	9	Titan V 12GB
P9	96 h	1.2	30	GTX 1080 Ti
P10	10 h	1	4	Tesla K20
P11	11 h	4.48	38	Tesla P100 12GB
P12	3.4 h	0.014	18	GTX 1080 Ti
P13	3 h	0.087	20	GTX 2080 Ti
P14	n/a	15	24	Titan X GPU

scanner vendors and used it to modify the training images (through error propagation) until the classifier could not distinguish between the domain. In other words, this method resulted in training images and a trained model that are less dependent on the specific vendors. P6 and P10 proposed to train two models simultaneously with shared features, one for segmentation and one for classification, such that the classification loss is high while the segmentation loss is low, generating features that are robust to vendor-specific variations as well as optimal for segmentation.

IV. RESULTS

As shown in Table IV, a balanced dataset across the four vendors was prepared for evaluating the final submissions (40 CMRs per vendor, total 160 datasets). In this section, we analyze the obtained results per (1) team, (2) vendor, (3) clinical center, and (4) show some qualitative results. For analysing the obtained results, we also implemented two baseline models to better appreciate the added value of the data augmentation and domain adaptation techniques used in this challenge:

B1: A 2D UNet without any data augmentation as described in the original reference [17], trained with weighted cross entropy loss.

B2: The nnUNet pipeline, with a 2D UNet module and default parameters as given in [33] (the best fold according to the validation set was selected).

In particular, B2 differed from those in P1-P3 in that it only included one architecture type [2D UNet] and ±180

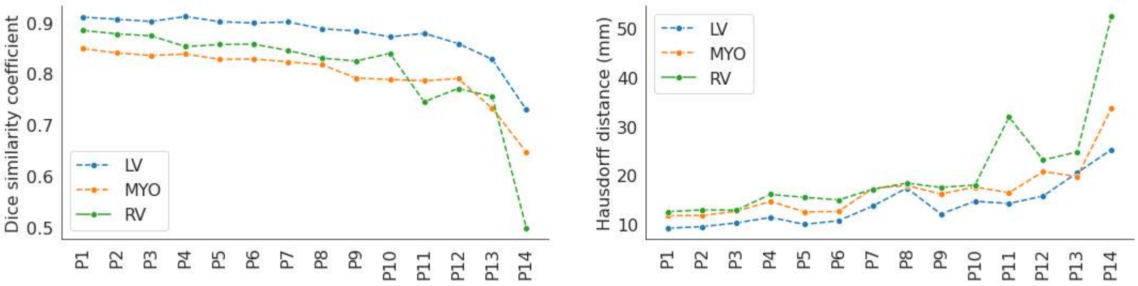

Fig. 4. Weighted average DSC and HD for all participating methods, according to equation (5).

degrees rotations, flippings, scalings, deformations, gamma transformations and test-time augmentation as data augmentation. In contrast, P1, P2 and P3 methods included further augmentation techniques such as histogram matching, noise addition, brightness modification, contrast modification and pseudo-label generation by label propagation in time space.

A. Analysis per team

Fig. 4 displays the results of the challenge for all participants and according to two evaluation metrics (DSC and HD). It can be seen that the curves are flat for about half of the participating teams, which indicates comparable performances overall. Note that these methods (P1 to P7) are also the ones that performed better than the baseline methods and we hypothesize that the other models (P8 to P14) suffered from some form of over-fitting (see also the shapes of the curves in Fig. 4). Team P1 provided the most consistent results across all metrics. However, the difference with respect to other teams was relatively small and in many cases not statistically significant, as presented in Table VIII. The three best performing teams, P1 to P3, used nnUNet as the baseline pipeline, as well as standard intensity-based data augmentation (e.g. blurring, noise addition, histogram matching), but no domain adaptation, showing a significative improvement with respect to the standard nnUNet implementation B2. For a similar performance, P5 used an Attention UNet as the backbone architecture and CycleGANs for data augmentation through image synthesis. P4 and P6 also obtained similar performances overall, but implemented instead domain adaptation methods and no image-driven data augmentation.

Fig. 5 displays the average DSC for all participating teams organised this time per pathology, showing better segmentation performance for healthy cases and dilated cardiomyopathy (DCM), followed by hypertrophic cardiomyopathy (HCM) and other pathologies. It can be seen that the performances of the 14 techniques relative to each other do not change when analysed per pathology.

B. Analysis per vendor

Fig. 6 summarizes the segmentation results for all teams for each vendor separately (A, B, C & D). It can be seen that overall, the differences in the segmentation errors between the vendors are reduced with respect to the results obtained by the two baseline methods as detailed in Table IX. Specifically, it can be seen that for the baseline methods there is a loss of accuracy of up to -6% in the segmentation of images from vendors C and D compared to A and B. However, this loss is reduced, for example, to -1.5% for P1 (e.g. from DSC = 0.92 for vendor A to 0.90 in vendor C and D, for the LV), -2.1% for P2 (e.g. from DSC = 0.87 in vendor B to 0.82 in vendor D, for the RV), and almost to 0% for P7. This indicates that while there is a need for further research to bring segmentation accuracy in unseen and unlabelled vendors at the same level of the one obtained in trained vendors, data augmentation and data adaptation enable to close the gap and improve the generalizability of deep learning models.

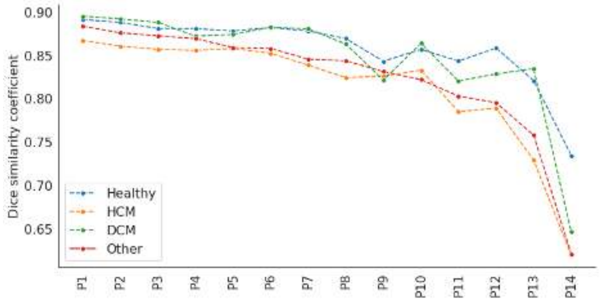

Fig. 5. Average DSC for all participants for the most common pathologies in the dataset. HCM and DCM stand for hypertrophic and dilated cardiomyopathy, respectively.

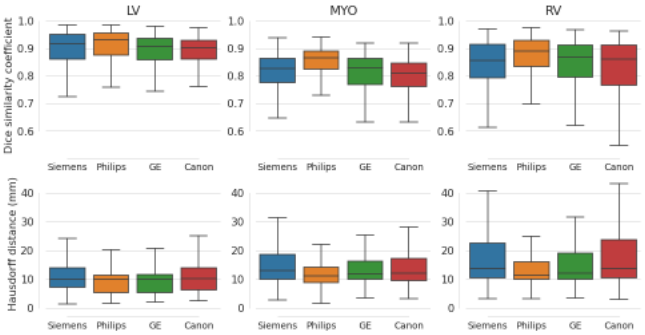

Fig. 6. Boxplots with vendor-wise results for DSC and HD when all participants predictions are considered. Vendors are presented in order: Siemens (A), Philips (B), GE (C) and Canon (D).

C. Analysis per centre

In the previous subsection, centres were combined in the analysis despite having different machines or scanning protocols. In doing so, possible variabilities between centres using the same scanner may be overstated, making it necessary

TABLE VIII

DSC AND HD FOR THE FINAL SUBMISSIONS OF ALL PARTICIPANTS AND THE TWO BASELINE MODELS. BOLD FACE NUMBERS ARE THE BEST RESULTS FOR EACH COLUMN AND BLUE NUMBERS ARE NON-SIGNIFICANTLY LOWER RESULTS WHEN COMPARED TO THE P1 RESULTS (P-VALUE > 0.01 FOR THE WELCH'S T-TEST). HD IS MEASURED IN MILIMETERS.

	ED						ES					
Method	LV		MYO		RV		LV		MYO		RV	
	DSC	HD	DSC	HD	DSC	HD	DSC	HD	DSC	HD	DSC	HD
P1	**0.939**	**9.1**	**0.839**	**12.8**	**0.910**	**11.8**	**0.886**	**9.1**	**0.867**	**10.6**	**0.860**	**12.7**
P2	0.938	9.3	0.830	12.9	0.909	12.3	0.880	9.5	0.861	10.8	0.850	13.0
P3	0.935	9.5	0.825	13.3	0.906	12.3	0.875	10.5	0.856	11.6	0.844	13.0
P4	**0.939**	11.3	0.826	15.2	0.886	15.4	0.884	11.4	0.856	14.0	0.829	16.7
P5	0.931	10.0	0.816	13.7	0.893	14.3	0.877	9.8	0.850	11.3	0.827	15.2
P6	0.927	11.2	0.815	14.0	0.892	13.6	0.877	9.7	0.852	11.1	0.834	15.0
P7	0.933	13.4	0.812	17.1	0.876	15.7	0.867	14.0	0.839	18.2	0.815	18.1
P8	0.922	15.5	0.809	18.0	0.867	16.6	0.857	17.5	0.836	17.2	0.802	19.1
P9	0.914	12.1	0.768	17.2	0.850	17.5	0.853	12.0	0.814	15.2	0.794	17.0
P10	0.905	13.6	0.772	17.2	0.876	16.2	0.848	15.5	0.820	17.5	0.809	19.6
P11	0.913	14.5	0.776	17.8	0.791	30.7	0.851	13.0	0.809	14.5	0.732	32.9
P12	0.889	16.0	0.785	22.1	0.814	22.1	0.835	14.2	0.808	18.9	0.758	22.0
P13	0.896	15.7	0.761	17.9	0.820	21.0	0.772	23.0	0.721	20.2	0.698	29.5
P14	0.797	21.9	0.668	31.6	0.552	49.1	0.716	25.8	0.673	33.0	0.517	52.0
B1	0.918	12.9	0.801	15.5	0.881	15.7	0.866	11.5	0.842	12.6	0.817	16.3
B2	0.930	10.8	0.817	15.7	0.889	14.8	0.863	13.2	0.835	14.8	0.818	16.8

TABLE IX

DSC RESULTS STRATIFIED BY VENDOR AND HEART SUBSTRUCTURE. THE LAST TWO COLUMNS ARE THE AVERAGE DSC LOSS FOR VENDORS C AND D WITH RESPECT TO THE COMBINED AVERAGE DSC RESULTS FROM VENDORS A AND B.

Method	Vendor A			Vendor B			Vendor C			Vendor D			DSC % loss	DSC % loss
	LV	MYO	RV	LV	MYO	RV	LV	MYO	RV	LV	MYO	RV	for vendor C	for vendor D
P1	**0.923**	**0.857**	**0.887**	0.915	**0.876**	**0.888**	0.903	**0.842**	**0.884**	0.909	**0.838**	**0.882**	-1.7	-1.6
P2	0.919	0.848	0.885	**0.916**	0.872	0.887	0.899	0.834	0.876	0.903	0.827	0.871	-2.0	-2.4
P3	0.915	0.843	0.877	0.914	0.868	0.879	0.894	0.827	0.873	0.898	0.824	0.870	-2.0	-2.1
P4	0.908	0.831	0.864	0.913	0.867	0.879	**0.906**	0.833	0.870	**0.918**	0.833	0.816	-0.9	-2.4
P5	0.912	0.834	0.869	0.910	0.859	0.870	0.891	0.817	0.819	0.903	0.820	**0.882**	-3.8	-0.8
P6	0.912	0.837	0.880	0.912	0.858	0.877	0.893	0.816	0.861	0.892	0.823	0.833	-2.6	-3.4
P7	0.891	0.804	0.820	0.904	0.859	0.870	0.898	0.821	0.838	0.908	0.817	0.853	-0.7	+0.1
P8	0.889	0.821	0.817	0.900	0.854	0.877	0.880	0.799	0.842	0.889	0.815	0.802	-2.3	-2.9
P9	0.879	0.765	0.800	0.889	0.816	0.827	0.881	0.787	0.831	0.885	0.797	0.829	+0.5	+1.0
P10	0.894	0.812	0.860	0.887	0.822	0.841	0.849	0.753	0.803	0.877	0.796	0.865	-6.1	-0.8
P11	0.885	0.781	0.778	0.899	0.846	0.846	0.875	0.787	0.773	0.869	0.758	0.650	-3.3	-9.8
P12	0.831	0.769	0.795	0.909	0.860	0.867	0.859	0.786	0.792	0.847	0.771	0.690	-3.1	-8.3
P13	0.820	0.712	0.684	0.885	0.823	0.858	0.868	0.779	0.803	0.762	0.650	0.691	+2.5	-12.1
P14	0.805	0.668	0.492	0.872	0.818	0.794	0.822	0.740	0.703	0.528	0.456	0.147	+2.3	-50.9
B1	0.908	0.834	0.861	0.901	0.850	0.865	0.863	0.790	0.800	0.894	0.813	0.870	-6.0	-1.3
B2	0.905	0.832	0.860	0.902	0.846	0.857	0.890	0.806	0.836	0.886	0.821	0.861	-2.7	-1.3

to consider also Fig. 7, where the segmentation results are summarized according to the six clinical centres. Here too, it can be seen that there remains some degree of variation in the segmentation of the CMR images from the different centres. In more detail, there is a decrease in segmentation accuracy between centres 1 and 6 even though their images are from the same scanner vendor A. However, this difference can be explained by two facts: 1) the scanners in these two centres are different models and have different field strengths, as shown in Table III, and 2) all the 75 datasets included during training for vendor A were from centre 1 (Spain) and none from centre 6 (Canada). In this case, even though the images are from the same vendor, differences in scanner specifications resulted in the lack of generalizability. In contrast, images from both centres 2 and 3 were included in the training of vendor B, which resulted in segmentation accuracies for these two centres that are comparable. Finally, the datasets from centres 4 and 5 correspond to vendors C and D, respectively, which were not included in the training, which explain the loss of accuracy compared to centres 1, 2 and 3. In Fig. 8, the results are grouped for all centres according to their inclusion (or not) in the training. Clearly, it can be seen that the segmentation accuracy is the highest for centres that are part of the training together with their labels, followed by those with images but no labels, and finally the performance is the lowest and most variable for images from fully unseen centres. This result confirms the need for further developments to optimize the generalizability of deep learning solutions in future tools for cardiac image segmentation.

D. Qualitative results

Fig. 9 presents the effect of the slice position in the final segmentation DSC for the top three performing teams, quantifying the loss of accuracy, especially prominent in the

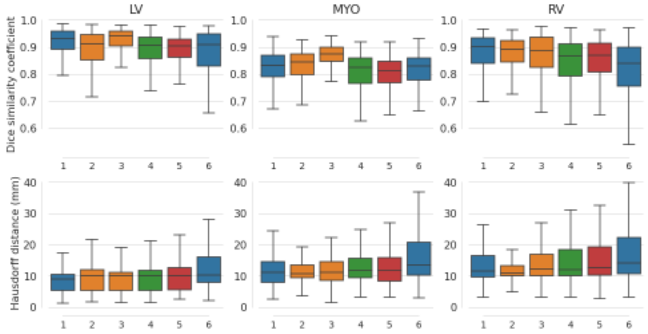

Fig. 7. Boxplots with centre-wise results for DSC and HD when all participants predictions are considered. Same color-coding as in Fig. 6 is used for scanner vendors.

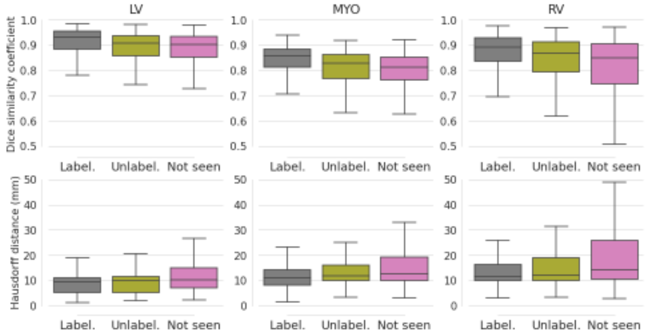

Fig. 8. Boxplots for DSC and HD results for centres that had labelled samples in the training set, unlabelled samples in the training set and no samples at all.

apical and basal slices. To illustrate this, Fig. 10 provides some visual examples from team P1 to further show the added value of the implemented techniques, as well as their limitations when applied to unseen vendors. In the two examples above, the segmentation techniques enabled to accurately identify the cardiac boundaries even though these imaging protocols were not included in the training set. However, in the two examples below, despite the use of data augmentation and domain adaptation, the models were unsuccessful in the segmentation of these unseen cases and diverged more notably from the ground truth in basal slices. These examples illustrate the need for future work to further improve the generalizability of deep learning models in cardiac image segmentation.

V. Discussion

In this paper, we presented a comprehensive analysis of a range of deep learning solutions for the automated segmentation of multi-centre, multi-vendor and multi-disease CMR datasets. Roughly speaking, the 14 participants in the challenge developed varying workflows combining a baseline neural network, intensity-based and/or spatial data augmentation, and in some cases a data adaptation strategy. In addition to a relatively large sample of 175 cases for training, the authors were given a total of seven attempts for optimising the parameters and characteristics of their models during the validation process, to ensure an optimal design of the solutions.

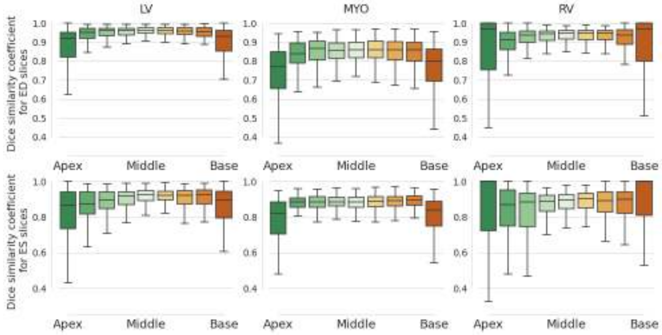

Fig. 9. Boxplots for DSC results for the top 3 performing methods depending on different cardiac structures (LV, MYO and RV) and different slice position for both ED and ES. The apex and the base are defined as the last and first annotated slices, respectively. The middle slice is the slice located in between the apex and base slices. The remaining slices are defined based on their relative position with respect to the middle slice.

A. Analysis of the methods

The obtained results, first of all, indicate that data augmentation, though its primary purpose is to increase training size and reduce over-fitting, can perform well in addressing some of the differences in image appearance between vendors. In particular, by varying the parameters and types of intensity transformations (e.g. histogram matching, contrast modification, noise addition, image synthesis), one can generate new training images that enhance the generalizability of the models. As an example, one can look at the performance of the baselines models B1 and B2 and augmented models, such as P1, P2 and P3. While for the baseline models, the results do not differ significantly for specific cases, such as at ES, P1-P3 used many more data augmentation types, such as histogram matching, noise addition, brightness modification and contrast modification, and obtained a more marked improvement (*e.g.* the DSC for the myocardium at ES increased from 0.84 for B1 to 0.86 for P1, the DSC for the RV at ES increased from 0.81 for B1 to 0.84 for P3). This indicates the added value of more advanced image-driven data augmentation for multi-vendor image segmentation as well as that the domain shift between different scanners or protocols can be potentially solved by using an exhaustive set of image transformations during training. However, the results also clearly show that the obtained segmentations remain generally more stable in trained vendors compared to unseen vendors, as intensity-driven data augmentation alone cannot enable a full coverage of the variety of imaging protocols that can exist across clinical centres.

As for domain adaptation, while it is theoretically suitable for multi-vendor image segmentation, as it can adapt on the spot to the imaging distribution of the unseen images, it did not result in better segmentations than when using exhaustive data augmentation alone. In fact, the three first techniques in the ranking did not use any domain adaptation, though it is important to reiterate that the first seven solutions obtained relatively similar results overall. It is worth noting that the choice of the baseline model may play a role, as again the first three techniques used the same model, namely the

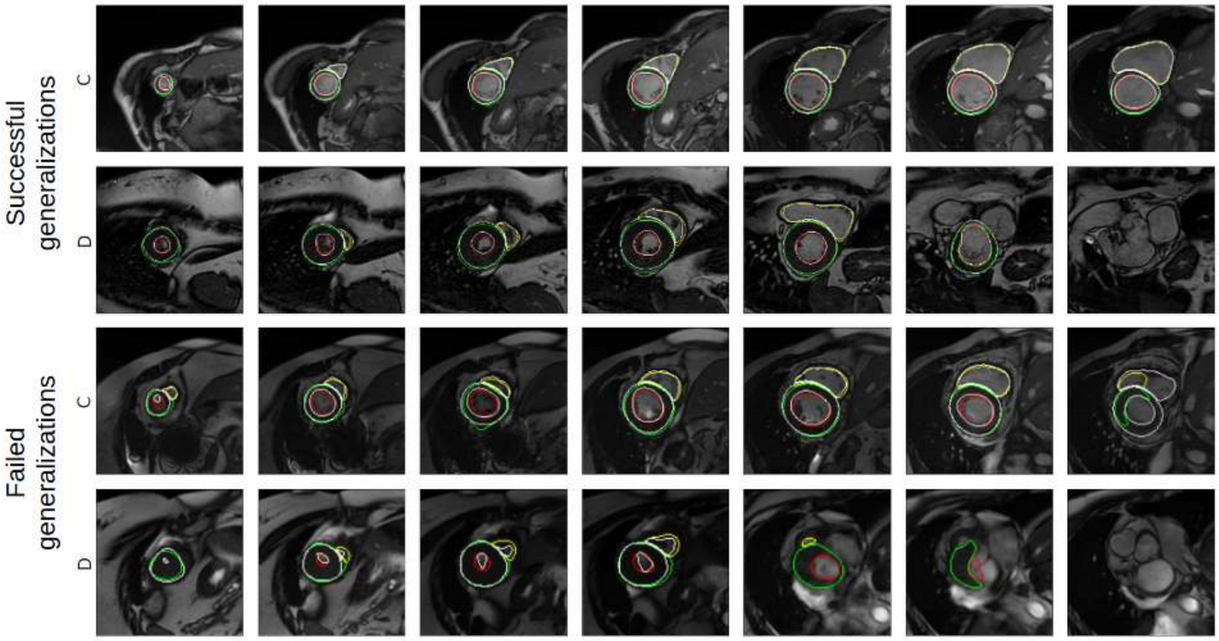

Fig. 10. Prediction examples for method P1 for vendors C (GE) and D (Canon). Top two rows show satisfactory results, while the two bottom rows present some error in the final contours. Color correspondence: left ventricle endocardium (red), left ventricle epicardium (green) and right ventricle endocardium (yellow). Ground truth is drawn in white color.

nnUNet. Finally, while the results indicate the potential of data augmentation and domain adaption, they also show that there is still a loss in segmentation accuracy when segmenting labelled versus unlabelled or unseen image samples. Note also that training and testing a model on two datasets from the same vendor does not guarantee a good generalizability. This is particularly true if the two sets of images are from two different centres and scanner types, such as 1.5T (e.g. centre 1) and 3T (e.g. centre 6) as shown in Figure 7.

The results also show that advanced workflows integrating, for instance, data augmentation or generative adversarial networks, are not guaranteed to lead to robust segmentations. In fact, half of the submitted techniques had a lower performance than the two baselines implemented for comparison. This shows that over-fitting remains a challenge that requires special attention during the calibration and validation of complex deep learning solutions for cardiac image segmentation, in particular in the presence of highly heterogeneous data.

Lastly, the presented methods show a vast diversity in hardware performance, with training times ranging from 6 to 100 hours and inference times from tenths of seconds to almost half a minute. However, the amount of training and inference time do not correlate well with the final accuracy, indicating an excessive use of computational power for some techniques. For example, the methods implemented by P1 and P2, despite using the same baseline model than P3, needed around half the time for training and obtained slightly better results (1.2% average improvement in DSC), while P4 used around one tenth of computing time for similar loss of accuracy with respect to P1 (1.6% average loss in DSC). Furthermore, clinical centres usually lack dedicated hardware for deep learning models thus increasing even more the segmentation time. In this sense, a good equilibrium between accuracy and processing time needs to be attained, with methods such as P4 serving as a good example with a competitive performance and a prediction rate of around 3 images per second.

In summary, the main findings are:

a) Exhaustive data augmentation reduced considerably the domain gap, although the results were still more stable within the domains used during training.
b) Domain adaptation did not result in better performance when compared to nnUNet models trained with spatial and intensity-driven data augmentation.
c) Complex workflows did not always lead to better results, resulting sometimes in an excessive use of computing resources.

B. Analysis of the segmentation results

Compared to other publicly available and annotated multi-structure (LV, MYO, RV) datasets in the field of CMR segmentation, M&Ms is the largest as well as the most diverse (375 cases from four vendors, six centres and three countries, vs. 150 cases for ACDC from one centre). However, given that ACDC is an established database, we selected to use its contouring SOP in this challenge to derive standardized annotations for the community, as well as to enable the combination of these datasets in future studies.

Note that our study, while it focuses on multi-scanner generalizable segmentation, confirms several of the results already obtained by the ACDC challenge and other previous works. Specifically:

a) The segmentations at ED were more accurate than at ES for LV and RV cavities, but not for the myocardium, which becomes thicker and therefore easier to segment when the heart contracts.
b) The segmentation accuracy according to the DSC was the highest for the LV blood pool, followed by the RV and MYO, in this order, but it was the lowest for the RV for the distance-based measures, given its shape complexity.
c) The segmentation accuracy was at its maximum at the mid-ventricular slices, while the performance decreased for the apical and basal slices, where there is higher variability and complexity.

On average, the best performing method in this challenge obtained 0.88 as DSC and 11 mm as HD versus the values 0.93 and 9 mm obtained in the ACDC challenge, respectively, with the greatest difference shown at ES. This gap can be easily explained by the single-centre nature of the ACDC studies in comparison to a multi-centre scenario in this work, although other effects such as the training size may play a role and should be assessed (150 vs. 100 studies, respectively).

C. Future work

In addition to the results and analyses presented in this paper on multi-scanner cardiac image segmentation, we also provide the M&Ms dataset open-access for the community, which can be downloaded from the M&Ms website[6]. It represents one of the most heterogeneous datasets ever compiled in cardiac image analysis, comprising CMRs from a variety of imaging protocols and cardiology units, and including a range of cardiovascular diseases as distinct as coronary heart disease, cardiomyopathies, abnormal right ventricle or myocarditis. We thus hope the dataset will be of high value for the community to address a number of research topics in the field, such as multi-scanner image registration, multi-structure segmentation, cardiac quantification, motion analysis and image synthesis.

It is important to note that a follow-up challenge is being organised on multi-centre, multi-vendor and multi-disease cardiac diagnosis. The diagnoses for the 375 cases are being gathered from the different hospitals in a legally compliant manner and the clinical information will be made available after the end of the next challenge, thus allowing the community to work on cardiac image analysis as well as on computer-aided diagnosis in a multi-centre setting. Note that the participants had less than three months to implement, optimize and test their techniques, which did not allow to go beyond the existing state-of-the-art techniques in data augmentation and domain adaptation. With more time at their disposal beyond the constraints of the challenge, we expect that researchers will have a valuable resource with the M&Ms dataset to investigate, develop and test new theories and frameworks for addressing the difficulties posed by domain-shift in cardiac image analysis.

D. Conclusions

The M&Ms challenge is the first study to evaluate a range of deep learning solutions for the automated segmentation of multi-centre, multi-vendor and multi-disease cardiac images. The results show the promise of existing data augmentation and domain adaptation methods, but also calls for further research to develop highly generalizable solutions given the inherent heterogeneity in cardiac imaging between centres, vendors and protocols. More generally, there is a need for more research and development to realise the much-needed shift from single-centre image analysis towards multi-domain approaches that will enable wider translation and usability of future artificial intelligence tools in cardiac imaging and clinical cardiology.

Acknowledgment

The authors would like to thank NVIDIA and the Barcelona Super Computing (BSC) centre for providing the necessary computational resources for this work.

References

[1] I. Cetin, G. Sanroma, S. E. Petersen, S. Napel, O. Camara, M.-A. G. Ballester, and K. Lekadir. A radiomics approach to computer-aided diagnosis with cardiac cine-MRI. In *Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges*, pages 82–90, Cham, 2018. Springer International Publishing.

[2] C. Martín-Isla, V. M. Campello, C. Izquierdo, Z. Raisi-Estabragh, B. Baessler, S. Petersen, and K. Lekadir. Image-based cardiac diagnosis with machine learning: A review. *Frontiers in Cardiovascular Medicine*, 7, 2020.

[3] X. Albà, K. Lekadir, M. Pereañez, P. Medrano-Gracia, A. Young, and A. Frangi. Automatic initialization and quality control of large-scale cardiac MRI segmentations. *Medical Image Analysis*, 43:129–141, 2018.

[4] W. Bai, W. Shi, C. Ledig, and D. Rueckert. Multi-atlas segmentation with augmented features for cardiac MR images. *Medical image analysis*, 19 1:98–109, 2015.

[5] C. Chen, C. Qin, H. Qiu, G. Tarroni, J. Duan, W. Bai, and D. Rueckert. Deep learning for cardiac image segmentation: A review. *Frontiers in Cardiovascular Medicine*, 7, 2020.

[6] O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, X. Yang, P. Heng, I. Cetin, K. Lekadir, O. Camara, M. A. G. Ballester, G. Sanroma, S. Napel, S. Petersen, G. Tziritas, E. Grinias, M. Khened, V. Kollerathu, G. Krishnamurthi, M. Rohé, X. Pennec, M. Sermesant, F. Isensee, P. Jäger, K. Maier-Hein, P. M. Full, I. Wolf, S. Engelhardt, C. F. Baumgartner, L. Koch, J. Wolterink, I. Išgum, Y. Jang, Y. Hong, J. Patravali, S. Jain, O. Humbert, and P.-M. Jodoin. Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved? *IEEE Transactions on Medical Imaging*, 37:2514–2525, 2018.

[7] F. Isensee, P. F. Jaeger, P. M. Full, I. Wolf, S. Engelhardt, and K. H. Maier-Hein. Automatic cardiac disease assessment on cine-mri via time-series segmentation and domain specific features. In *Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges*, pages 120–129, Cham, 2018. Springer International Publishing.

[8] I. Cetin, Z. Raisi-Estabragh, S. E. Petersen, S. Napel, S. K. Piechnik, S. Neubauer, M. A. Gonzalez Ballester, O. Camara, and K. Lekadir. Radiomics signatures of cardiovascular risk factors in cardiac MRI: Results from the UK Biobank. *Frontiers in Cardiovascular Medicine*, 7:232, 2020.

[9] W. Bai, M. Sinclair, G. Tarroni, O. Oktay, M. Rajchl, G. Vaillant, A. Lee, N. Aung, E. Lukaschuk, M. M. Sanghvi, F. Zemrak, K. Fung, J. Paiva, V. Carapella, Y. Kim, H. Suzuki, B. Kainz, P. Matthews, S. Petersen, S. Piechnik, S. Neubauer, B. Glocker, and D. Rueckert. Automated cardiovascular magnetic resonance image analysis with fully convolutional networks. *Journal of Cardiovascular Magnetic Resonance*, 20, 2018.

[10] P. Tran. A fully convolutional neural network for cardiac segmentation in short-axis MRI. *ArXiv*, abs/1604.00494, 2016.

[6] www.ub.edu/mnms

[11] Q. Tao, W. Yan, Y. Wang, E. H. M. Paiman, D. P. Shamonin, P. Garg, S. Plein, L. Huang, L. Xia, M. Sramko, J. Tintera, A. de Roos, H. J. Lamb, and R. J. van der Geest. Deep learning–based method for fully automatic quantification of left ventricle function from cine MR images: A multivendor, multicenter study. *Radiology*, 290(1):81–88, 2019. PMID: 30299231.

[12] M. Khened, A. Varghese, and G. Krishnamurthi. 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] C. Chen, W. Bai, R. Davies, A. Bhuva, C. Manisty, J. Moon, N. Aung, A. Lee, M. M. Sanghvi, K. Fung, J. Paiva, S. Petersen, E. Lukaschuk, S. Piechnik, S. Neubauer, and D. Rueckert. Improving the generalizability of convolutional neural network-based segmentation on CMR images. *Frontiers in Cardiovascular Medicine*, 7, 2020.

[14] P. Radau, Y. Lu, K. Connelly, G. Paul, A. Dick, and G. Wright. Evaluation framework for algorithms segmenting short axis cardiac MRI. *The MIDAS Journal-Cardiac MR Left Ventricle Segmentation Challenge*, 49, 2009.

[15] A. Suinesiaputra, B. Cowan, A. O. Al-Agamy, M. A. Alattar, N. Ayache, A. Fahmy, A. Khalifa, P. Medrano-Gracia, M. Jolly, A. H. Kadish, D. Lee, J. Margeta, S. Warfield, and A. Young. A collaborative resource to build consensus for automated left ventricular segmentation of cardiac MR images. *Medical image analysis*, 18 1:50–62, 2014.

[16] C. Petitjean, M. A. Zuluaga, W. Bai, J.-N. Dacher, D. Grosgeorge, J. Caudron, S. Ruan, I. B. Ayed, M. J. Cardoso, H.-C. Chen, D. Jimenez-Carretero, M. J. Ledesma-Carbayo, C. Davatzikos, J. Doshi, G. Erus, O. M. Maier, C. M. Nambakhsh, Y. Ou, S. Ourselin, C.-W. Peng, N. S. Peters, T. M. Peters, M. Rajchl, D. Rueckert, A. Santos, W. Shi, C.-W. Wang, H. Wang, and J. Yuan. Right ventricle segmentation from cardiac MRI: A collation study. *Medical Image Analysis*, 19(1):187 – 202, 2015.

[17] O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Lecture Notes in Computer Science*, pages 234–241. Springer International Publishing, 2015.

[18] P. M. Full, F. Isensee, P. F. Jäger, and K. Maier-Hein. Studying robustness of semantic segmentation under domain shift in cardiac MRI. In *Statistical Atlases and Computational Models of the Heart. M&Ms and EMIDEC Challenges*, pages 238–249, Cham, 2021. Springer International Publishing.

[19] Y. Zhang, J. Yang, F. Hou, Y. Liu, Y. Wang, J. Tian, C. Zhong, Y. Zhang, and Z. He. Semi-supervised cardiac image segmentation via label propagation and style transfer. In *Statistical Atlases and Computational Models of the Heart. M&Ms and EMIDEC Challenges*, pages 219–227, Cham, 2021. Springer International Publishing.

[20] J. Ma. Histogram matching augmentation for domain adaptation with application to multi-centre, multi-vendor and multi-disease cardiac image segmentation. In *Statistical Atlases and Computational Models of the Heart. M&Ms and EMIDEC Challenges*, pages 177–186, Cham, 2021. Springer International Publishing.

[21] M. Parreño, R. Paredes, and A. Albiol. Deidentifying MRI data domain by iterative backpropagation. In *Statistical Atlases and Computational Models of the Heart. M&Ms and EMIDEC Challenges*, pages 277–286, Cham, 2021. Springer International Publishing.

[22] F. Kong and S. C. Shadden. A generalizable deep-learning approach for cardiac magnetic resonance image segmentation using image augmentation and attention U-Net. In *Statistical Atlases and Computational Models of the Heart. M&Ms and EMIDEC Challenges*, pages 287–296, Cham, 2021. Springer International Publishing.

[23] J. Corral Acero, V. Sundaresan, N. Dinsdale, V. Grau, and M. Jenkinson. A 2-step deep learning method with domain adaptation for multi-centre, multi-vendor and multi-disease cardiac magnetic resonance segmentation. In *Statistical Atlases and Computational Models of the Heart. M&Ms and EMIDEC Challenges*, pages 196–207, Cham, 2021. Springer International Publishing.

[24] M. Saber, D. Abdelrauof, and M. Elattar. Multi-center, multi-vendor, and multi-disease cardiac image segmentation using scale-independent multi-gate UNET. In *Statistical Atlases and Computational Models of the Heart. M&Ms and EMIDEC Challenges*, pages 259–268, Cham, 2021. Springer International Publishing.

[25] H. Li, J. Zhang, and B. Menze. Generalisable cardiac structure segmentation via attentional and stacked image adaptation. In *Statistical Atlases and Computational Models of the Heart. M&Ms and EMIDEC Challenges*, pages 297–304, Cham, 2021. Springer International Publishing.

[26] F. Khader, J. Schock, D. Truhn, F. Morsbach, and C. Haarburger. Adaptive preprocessing for generalization in cardiac MR image segmentation. In *Statistical Atlases and Computational Models of the Heart. M&Ms and EMIDEC Challenges*, pages 269–276, Cham, 2021. Springer International Publishing.

[27] C. M. Scannell, A. Chiribiri, and M. Veta. Domain-adversarial learning for multi-centre, multi-vendor, and multi-disease cardiac MR image segmentation. In *Statistical Atlases and Computational Models of the Heart. M&Ms and EMIDEC Challenges*, pages 228–237, Cham, 2021. Springer International Publishing.

[28] A. Carscadden, M. Noga, and K. Punithakumar. A deep convolutional neural network approach for the segmentation of cardiac structures from MRI sequences. In *Statistical Atlases and Computational Models of the Heart. M&Ms and EMIDEC Challenges*, pages 250–258, Cham, 2021. Springer International Publishing.

[29] X. Liu, S. Thermos, A. Chartsias, A. O'Neil, and S. A. Tsaftaris. Disentangled representations for domain-generalized cardiac segmentation. In *Statistical Atlases and Computational Models of the Heart. M&Ms and EMIDEC Challenges*, pages 187–195, Cham, 2021. Springer International Publishing.

[30] X. Huang, Z. Chen, X. Yang, Z. Liu, Y. Zou, M. Luo, W. Xue, and D. Ni. Style-invariant cardiac image segmentation with test-time augmentation. In *Statistical Atlases and Computational Models of the Heart. M&Ms and EMIDEC Challenges*, pages 305–315, Cham, 2021. Springer International Publishing.

[31] L. Li, V. A. Zimmer, W. Ding, F. Wu, L. Huang, J. A. Schnabel, and X. Zhuang. Random style transfer based domain generalization networks integrating shape and spatial information. In *Statistical Atlases and Computational Models of the Heart. M&Ms and EMIDEC Challenges*, pages 208–218, Cham, 2021. Springer International Publishing.

[32] Z. Liu, X. Yang, R. Gao, S. Liu, H. Dou, S. He, Y.-H. Huang, Y. Huang, H. Luo, Y. Zhang, Y. Xiong, and D. Ni. Remove appearance shift for ultrasound image segmentation via fast and universal style transfer. *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 1824–1828, 2020.

[33] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. Maier-Hein. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 2020.

[34] H. Li, A. Zhygallo, and B. Menze. Automatic brain structures segmentation using deep residual dilated U-Net. *ArXiv*, abs/1811.04312, 2018.

[35] O. Oktay, J. Schlemper, L. L. Folgoc, M. J. Lee, M. Heinrich, K. Misawa, K. Mori, S. G. McDonagh, N. Hammerla, B. Kainz, B. Glocker, and D. Rueckert. Attention U-Net: Learning where to look for the pancreas. *ArXiv*, abs/1804.03999, 2018.

[36] A. Chartsias, T. Joyce, G. Papanastasiou, S. Semple, M. Williams, D. E. Newby, R. Dharmakumar, and S. A. Tsaftaris. Disentangled representation learning in cardiac image analysis. *Medical image analysis*, 58:101535, 2019.

[37] X. Huang and S. Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1510–1519, 2017.

[38] D. Shanmugam, D. W. Blalock, G. Balakrishnan, and J. Guttag. When and why test-time augmentation works. *ArXiv*, abs/2011.11156, 2020.