Deep Learning Segmentation of the Right Ventricle in Cardiac MRI: The M&Ms Challenge

Carlos Martín-Isla, Víctor M. Campello, Cristian Izquierdo, Kaisar Kushibar, Carla Sendra-Balcells, Polyxeni Gkontra, Alireza Sojoudi, Mitchell J Fulton, Tewodros Weldebirhan Arega, Kumaradevan Punithakumar, Lei Li, Xiaowu Sun, Yasmina Al Khalil, Di Liu, Sana Jabbar, Sandro Queirós, Francesco Galati, Moona Mazher, Zheyao Gao, Marcel Beetz, Lennart Tautz, Christoforos Galazis, Marta Varela, Markus H "ullebrand, Vicente Grau, Xiahai Zhuang, Domenec Puig, Maria A. Zuluaga, Hassan Mohy-ud-Din, Dimitris Metaxas, Marcel Breeuwer, Rob J. van der Geest, Michelle Noga, Stephanie Bricq, Mark E. Rentschler, Andrea Guala, Steffen E. Petersen, Sergio Escalera, José F. Rodríguez Palomares, and Karim Lekadir

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

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

A. Sojoudi is with Circle Cardiovascular Imaging, Canada.

M. J Fulton and Mark E. Rentschler are with the University of Colorado Boulder, USA.

T. Arega and S. Bricq are with the ImViA Laboratory, Université Bourgogne Franche-Comté, France.

L. Li is with the School of Data Science, Fudan University, China.

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

Y. Al Khalil and M. Breeuwer are with the Eindhoven University of Technology, the Netherlands.

X. Sun and R. J. van der Geest are with the Division of Image Processing, Department of Radiology, Leiden University Medical Center, the Netherlands.

D. Liu and D. Metaxas are with the Department of Computer Science, Rutgers University, Piscataway, USA.

S. Jabbar and H. Mohy-ud-Din are with the Department of Electrical Engineering, Syed Babar Ali School of Science and Engineering, Lahore, Pakistan.

S. Queirós is with the Life and Health Sciences Research Institute (ICVS), School of Medicine, University of Minho, Braga, Portu-gal and with the ICVS/3 B's - PT Government Associate Laboratory, Braga/Guimarães, Portugal.

F. Galati and M. A. Zuluaga are with the Data Science Department EURECOM, Sophia Antipolis, France.

M. Mazher and D. Puig are with the Department of Computer Engineering and Mathematics, University Rovira i Virgili, Spain.

Z. Gao and X. Zhuang are with the School of Data Science, Fudan University, Shanghai, China.

M. Beetz and V. Grau are with the Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, UK.

L. Tautz and M. Hüllebrand are with Charité - Universitätsmedizin Berlin, Berlin, Germany, and with Fraunhofer MEVIS, Bremen, Germany

C. Galazis and M. Varela are with the National Heart & Lung Institute, Imperial College London, UK, and C. Galazis is also with the Department of Computing, Imperial College London, UK

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

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 is 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.

Abstract — In recent years, several deep learning models have been proposed to accurately quantify and diagnose cardiac pathologies. These automated tools heavily rely on the accurate segmentation of cardiac structures in MRI images. However, segmentation of the right ventricle is challenging due to its highly complex shape and ill-defined borders. Hence, there is a need for new methods to handle such structure's geometrical and textural complexities, notably in the presence of pathologies such as Dilated Right Ventricle, Tricuspid Regurgitation, Arrhythmogenesis, Tetralogy of Fallot, and Inter-atrial Communication. The last MICCAI challenge on right ventricle segmentation was held in 2012 and included only 48 cases from a single clinical center. As part of the 12th Workshop on Statistical Atlases and Computational Models of the Heart (STACOM 2021), the M&Ms-2 challenge was organized to promote the interest of the research community around right ventricle segmentation in multi-disease, multi-view, and multi-center cardiac MRI. Three hundred sixty CMR cases, including short-axis and long-axis 4-chamber views, were collected from three Spanish hospitals using nine different scanners from three different vendors, and included a diverse set of right and left ventricle pathologies. The solutions provided by the participants show that nnU-Net achieved the best results overall. However, multi-view approaches were able to capture additional information, highlighting the need to integrate multiple cardiac diseases, views, scanners, and acquisition protocols to produce reliable automatic cardiac segmentation algorithms.

Index Terms — Cardiovascular magnetic resonance, image segmentation, data augmentation, multi-view segmentation, public dataset.

I. Introduction

The role of the right ventricle (RV) in circulation has historically been overshadowed by that of the left ventricle (LV). For years, RV dysfunction was thought to not contribute significantly to cardiac output and pressures, while LV was considered the key player in cardiac hemodynamics [3]. This led to RV receiving limited attention, and often being described as the "forgotten ventricle" [4]. However, in the past few decades, the misconception regarding the lack of impact of the RV dysfunction in cardiac function has changed

TABLE I
AUTOMATIC CMR SEGMENTATION CHALLENGES IN FIGURES

Challenge	Year	Cases	Number of scanners	Target Regions	Multiview	Techniques used	Number of pathologies	Stratified by pathology
M&Ms-2	2021	360	9	RV	✓	Deep Learning	8	✓
RVSC	2012	48	1	RV	X	Atlas-based	6	X
M&Ms	2020	375	5	LV/RV/MYO	X	Deep Learning	6	X
ACDC	2017	150	1	LV/RV/MYO	X	Deep Learning	5	✓
LVSC [1]	2011	200	-	MYO	X	Atlas-based	1	X
Sunnybrook [2]	2009	45	1	LV/MYO	X	Atlas-based	4	✓

[5]–[10]. A significant amount of research has progressively demonstrated the pivotal role of RV in cardiac function, and its implication and prognostic value in high-burden diseases, such as heart failure and/or pulmonary hypertension [11]–[13], dilated cardiomyopathy [14], tricuspid regurgitation [15], tetralogy of fallot [16], to name a few.

Given the prognostic significance of RV, the clinical interest has shifted in recent years from a simple visual inspection of the RV from cardiac magnetic resonance imaging (CMR), the reference modality for RV assessment [17], to extracting quantitative RV parameters by first segmenting the structure. Despite this renewed interest of the medical community to quantitatively assess the RV [18], the artificial intelligence community has lagged in providing fully automated solutions for RV segmentation from CMR, that are as accurate as for LV [19], and in benchmarking deep learning (DL) algorithms, the current state-of-the-art in medical imaging.

More precisely, the last challenge focused on RV segmentation using CMR data was the Right Ventricle Segmentation Challenge Dataset (RVSC) [20]. Prior to the RVSC, challenges solely focused on the myocardium and LV (Table I). Despite its significance, the RVSC challenge was organized back in 2012 when DL was still in its early development and not yet adopted for CMR segmentation [21]. Therefore, none of the seven participants in the challenge used DL. Three approaches were atlas-based, two prior-based, and the other two based on cardiac motion without needing prior information. The best semi-automated methods achieved a dice accuracy of 80% and a Hausdorff distance of 1 cm. At the same time, automated approaches demonstrated a similar performance at the expense of higher computational costs. At those times, this performance level was competitive, but it is now considered far from what the current state-of-the-art DL-based models could achieve.

The early application of DL in CMR segmentation using a Fully Convolutional Network (FCN) by Tran [22] showed improved results compared to prior CMR segmentation methods in RVSC, LVSC, and Sunnybrook datasets. At the same time, the U-Net [23] architecture, which added a symmetric decoding path to the FCN architecture, started gaining inertia along the biomedical imaging segmentation community. However, Lieman et al. [24] shown that there was no statistical difference in CMR segmentation performance between the two architectures, with FCN outperforming U-Net in LV volume prediction using a large sample size of 1,143 subjects. This was further validated by Bai et al. [25] in a large-scale study using 4,875 cases for the bi-ventricular segmentation task.

In recent CMR segmentation challenges such as the Automated Cardiac Diagnosis Challenge (ACDC) [21] and the MMs challenge [19], the U-Net architecture has emerged as the dominant choice. In the ACDC, research by Baumgartner et al. [26] showed that U-Net outperformed the Fully Convolutional Network (FCN) in all proposed segmentation tasks, except for the RV end diastolic average symmetric surface distance. The early stages of the nnU-Net framework [27], which is capable of optimizing preprocessing, network architecture, training, inference, and post-processing automatically without manual intervention, were also demonstrated by Isensee et al. [28] in such challenge. The Top-3 participants in the MMs challenge [19] used nnU-Net.

Nonetheless, both aforementioned challenges were focused on cardiac multi-structure segmentation, and the best performance was achieved for the LV and the myocardium. The reduced accuracy in the RV segmentation can be explained by the additional challenges posed by the complex geometry and appearance of the RV. These include its irregular shape, the heterogeneity in the appearance and thickness of its free wall, and its complex trabeculations [20]. As a result, several works have been recently proposed to improve RV segmentation. [25], [29]–[37]. Nonetheless, the scarcity of relevant public CMR data has resulted in the vast majority of current state-of-the-art methods using the data provided by the RVSC challenge which comprises solely 48 cases from a single clinical center. Moreover, while the cohort includes diverse pathologies, the considered diseases are not directly related to the RV. Lastly, the complementary long-axis 4-chamber views, particularly helpful for improving RV apical and basal slices segmentation, were not released. Other relevant works using larger datasets, such as that of Chen et al. [38] based on 145 cases, although important, rely on private cohorts and, therefore, do not allow for benchmarking.

In response to the gap in public datasets and evalua-tion frameworks for computational approaches focused on automated RV segmentation from CMR, the Multi-Disease, Multi-View & Multi-Center Right Ventricular Segmentation Challenge (M&Ms-2) was organized as part of the Statistical Atlases and Computational Models of the Heart (STACOM) Workshop held in conjunction with the MICCAI 2021 Conference. This is the first work to provide a public multi-center, multi-disease, multi-view CMR dataset, associated contours, and an evaluation framework to benchmark DL algorithms for RV segmentation. Moreover, the dataset complements the

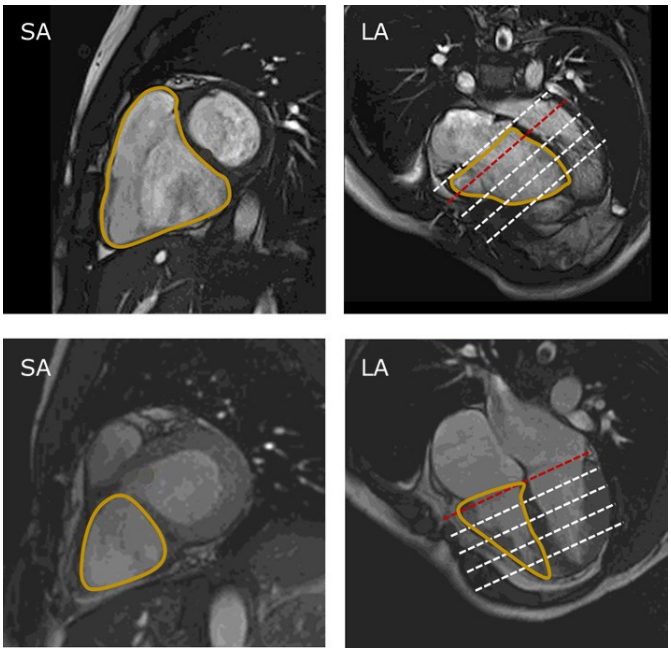

Fig. 1. Visual appearance of short-axis (SA) and long-axis (LA) views of pathological (upper row) and healthy (lower row) subjects. Dashed lines (white) correspond to the projection of SA slices into the LA view. The red dashed line corresponds to the projection of the SA slice shown in the first column. The yellow line corresponds to ground truth delineations.

dataset of the challenge's first edition [19], a reference dataset for multi-structure segmentation, by providing multi-view information and other diseases relevant to RV dysfunction. In total, the M&Ms-2 challenge dataset comprises CMR data from 360 participants originating from three Spanish hospitals. The data were acquired by nine different scanners from three different vendors (Siemens, Philips, and General Electric). The dataset was built in close collaboration with clinicians and accounts for seven different pathologies, while it also includes a control group of 75 healthy participants. It should be noted that the short-axis studies were annotated using the same Standard Operation Procedure (SOP) as previous reference challenges, while the complementary long-axis 4-chamber acquisitions for precise basal and apical delineation were also made publicly available.

In this paper, we present and discuss the results of the M&Ms-2 challenge in detail. The obtained results show the challenging nature of the task of automatically segmenting the RV from CMR images and the promise of the proposed solutions. Moreover, the findings of the challenge highlight the need for further research to build tools that can integrate multi-view cardiac information for the RV segmentation task in the presence of a diverse set of pathologies.

II. Challenge framework

A. Data preparation

A total of three clinical centers from Spain contributed to this challenge by providing several CMR studies with different left and right ventricular pathologies, namely:

Dilated Left Ventricle (DLV): LV is considered dilated when the LV end-diastolic volume measured in CMR is >214mL (>105mL/m2) in men or 179mL (>96mL/m2) in women.

Dilated Right Ventricle (DRV): RV is considered dilated when RV end-diastolic volume measured in CMR is >250mL (>121mL/m2) in men or 201mL (>112mL/m2) in women.

Hypertrophic cardiomyopathy (HCM) is an inherited heart disease defined by increased LV wall thickness (>15mm in one or more LV myocardial segments) that cannot be explained by abnormal loading conditions. In CMR, left ventricular mass typical values are 62-176g in men and 56-140g in women, and right ventricular mass typical values are 25-57g in men and 50-56g in women.

Arrhythmogenic cardiomyopathy (ARR) , inherited heart disease with a loss of myocytes and fibrofatty replacement of right ventricular myocardium; biventricular involvement is often observed. Diagnosis includes global RV dilatation and regional wall motion abnormalities with or without a decreased ejection fraction.

Tetrology of Fallot (FALL) is characterized by the following four features: a nonrestrictive ventricular septal defect, overriding aorta; right ventricle outflow tract obstruction and/or branch pulmonary artery stenosis; and consequent RV hypertrophy.

Inter-atrial communication (CIA) , a defect in the septum that separates the two atria. CMR is rarely required but may be useful for assessment of RV volume overload, identification of inferior sinus venous defect in the long-axis 4-chamber view, quantification of pulmonary to systemic flow ratio, and evaluation of pulmonary venous connection.

Tricuspid regurgitation (TRI) consists of the insufficiency of the tricuspid valve, causing blood flow from the RV to the right atrium during systole. In CMR, TRI appears as one or more flow jets emanating from the tricuspid valve and projecting into the RV. Jets are often holosystolic and readily apparent on the long-axis 4-chamber view.

In total, 360 studies were included. Images were acquired with different scanners, field strengths, and resolutions for both short-axis (SA) and long-axis 4-chamber (LA) views. Most images were acquired from scanners with magnetic strength of 1.5T and a small fraction of 3.0T. The specific vendors are 1) Siemens (Siemens Healthineers, Germany) – including Avanto (AVA), Avanto Fit (AVF), Symphony (SYM), SymphonyTim (SYT), and TrioTim (TRT) scanners; 2) Philips (Philips Healthcare, Netherlands) – including Achieva (ACH) scanners; and 3) General Electric (GE, GE Healthcare, USA) – including Signa Excite (EXC), Signa Explorer (EXP), and Signa HDxt (HDXT) scanners. More specific details on the collected studies are given in Table III.

The subjects included in this multi-disease study were selected among groups of the aforementioned cardiovascular diseases and healthy volunteers (NOR). The distribution of pathologies within the dataset partitions and scanners are specified in Table II and Figure 2.

Each CMR imaging study was annotated manually by an expert clinician from the corresponding center, with clinical

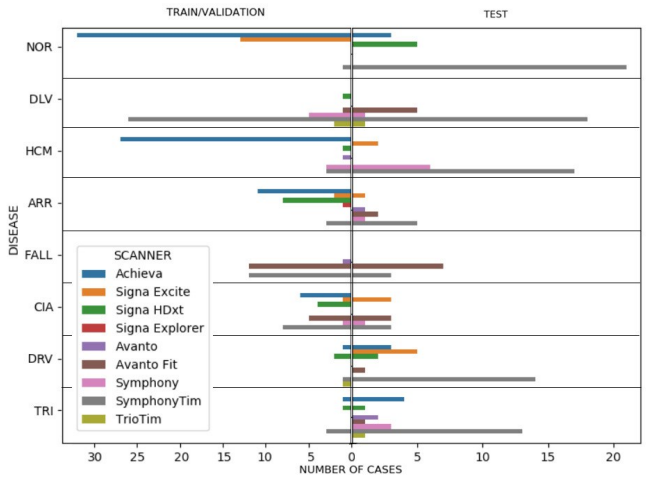

Fig. 2. Distribution per pathology and scanner along train, validation, and test sets.

TABLE II
NUMBER OF STUDIES PER PATHOLOGY IN EACH DATASET PARTITION

	Number of studies		
Pathology	Training	Validation	Test
Normal subjects	40	5	30
Dilated Left Ventricle	30	5	25
Hypertrophic Cardiomyopathy	30	5	25
Congenital Arrhythmogenesis	20	5	10
Tetralogy of Fallot	20	5	10
Interatrial Communication	20	5	10
Dilated Right Ventricle	0	5	25
Tricuspidal Regurgitation	0	5	25
Total	160	40	160

experience ranging from 3 to over 10 years. The annotation process involved marking the short-axis and long-axis 4-chamber views at both end-diastolic (ED) and end-systolic (ES) phases, which correspond to the phases used to calculate clinically relevant biomarkers such as ejection fraction and myocardial mass, for cardiac diagnosis and monitoring. Furthermore, the basal slice of the RV at ED/ES was inferred from the position of the tricuspid annulus as defined on the long-axis 4-chamber view at ED/ES. The apical slice was defined as the last slice with a detectable ventricular cavity. Three main regions were provided: the left and right ventricular cavities and the left ventricle myocardium (MYO). However, the evaluation was performed exclusively on the RV. Two additional researchers performed a detailed revision of the provided segmentation to reduce inter-observer and inter-center variability in the contours, particularly in the apical and basal regions. Discrepancies were resolved by consensus between the observers. Such observers applied the same SOP across all CMR datasets to obtain the final ground truth. To generate consistent annotations, we chose to apply the SOP that was already used by the ACDC and M&Ms challenges with an additional rule (d) as follows:

a) The LV and RV cavities, including the papillary muscles, must be completely covered.
b) No interpolation of the myocardial boundaries must be performed at the basal region.
c) The RV must have a larger volume at the ED time frame compared to ES.
d) Additionally, long-axis view is used as a reference to delimit the basal and apical regions, as stated above.

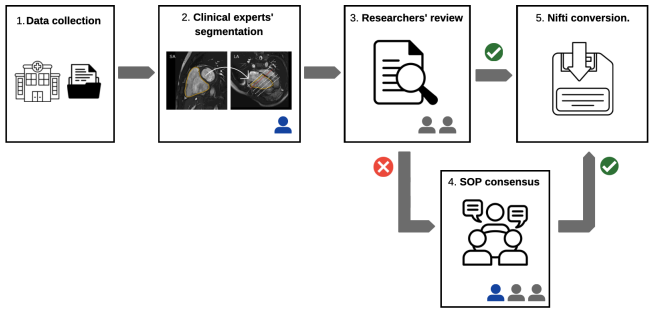

Fig. 3. Data collection and pre-processing pipeline.

Clinical delineations and subsequent corrections were performed using the 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). In-house software was then used to create the contours and transform the images into NIFTI format, and this final file format was delivered to the challenge participants. The inter-view correspondence was preserved during pre-processing. Figure 3 presents the data collection and pre-processing pipeline.

B. Model training and validation

The 360 CMR studies were divided into training, validation, and testing, as detailed in Table II. The participants received the 160 training cases with annotations for short and long-axis views and 40 validation cases without annotation on May 10th, 2021. Two pathologies, DRV and TRI, were excluded from the training dataset to test the generalization capability of the models to unseen pathologies. In order to optimize the models, the participants were allowed to automatically inspect their models' performance against 40 validation CMR cases, i.e. 5 from each of the pathologies, and publish their validation scores using the Codalab platform [39]. A maximum of 20 submissions per team were allowed during the validation process. Note that it was not permitted to use any external datasets or pre-trained models during training.

C. Model evaluation

The testing phase started on July 1st, 2021, and concluded on July 20th, 2021. The participants were forced to evaluate their models remotely to ensure the unseen test set was hidden from the segmentation methods. The organizers' GPU server infrastructure with five NVIDIA 3090 RTX GPUs was provided to evaluate the submissions. The participants were asked to assess their models by submitting their trained models to the Codalab platform and executing them using a Docker[1] image.

[1]https://www.docker.com/

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

Center*	Vendor	Model	In-plane res. (mm) (SA/LA)	In-plane dim. (pixels)	Slice thickness (mm)	Number of slices	Field Strength (T)	Number of studies
A	Philips	Achieva	1.18/1.19	332±32/288 ± 38	10	10	1.5	88
B	GE	Signa Excite	1.40/1.58	270±28/258 ± 6	9.8	12	1.5	27
B	GE	Signa HDxt	0.98/1.08	420±124/420 ± 124	10	12	1.5/3.0	25
B	GE	Signa Explorer	0.78/0.78	512±0/512 ± 0	10	13	1.5	1
C	Siemens	Avanto	1.21/1.15	232±24/240 ± 20	14	9	1.5	5
C	Siemens	Avanto Fit	1.13/1.24	234±24/234 ± 24	9.9	11	1.5	37
C	Siemens	Symphony	1.27/1.27	232±24/240 ± 18	9.7	10	1.5	21
C	Siemens	SymphonyTim	1.34/1.24	230±36/238 ± 26	9.7	12	1.5	151
C	Siemens	TrioTim	1.15/1.20	234±24/238 ± 18	8.6	13	3.0	5

* A: Clínica Sagrada Familia, B: Hospital Universitari Dexeus, C: Hospital Vall d'Hebron.

TABLE IV
LIST AND DETAILS OF THE PARTICIPATING TEAMS IN THE CHALLENGE.

Team	Institution	Location
P1	University of Colorado Boulder	Boulder, USA
P2	ImViA Laboratory, Université Bourgogne Franche-Comté	Dijon, France
P3	Dept. of Radiology and Diagnostic Imaging, University of Alberta	Edmonton, Canada
P4	School of Data Science, Fudan University	Shanghai, China
P5	Department of Radiology, Leiden University Medical Center	Leiden, Netherlands
P6	Eindhoven University of Technology	Eindhoven, Netherlands
P7	Department of Computer Science, Rutgers University	Piscataway, USA
P8	Department of Electrical Engineering, Syed Babar Ali School of Science and Engineering	Lahore, Pakistan
P9	Life and Health Sciences Research Institute (ICVS), School of Medicine, University of Minho	Braga, Portugal
P10	Data Science Department, EURECOM	Sophia Antipolis, France
P11	Department of Computer Engineering and Mathematics, University Rovira i Virgili	Tarragona, Spain
P12	School of Data Science, Fudan University	Shanghai, China
P13	Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford	Oxford, UK
P14	Charité - Universitatsmedizin Berlin,	Berlin, Germany
P15	Department of Computing, Imperial College London	London, UK

To assess the quality of the automatic segmentations (P) against the ground truth (G), two measures were used:

(i) Dice similarity coefficient (DSC) – degree of overlapping of two volumes:

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

(ii) Hausdorff distance (HD) – largest disagreement between the contours, useful for identifying small outliers:

$$HD(P, G) = \max \left(\sup_{p \in P} d(p, G), \sup_{g \in G} d(g, P) \right) \quad (2)$$

where sup represents the supremum, inf the infimum, and

$$d(a, B) = \inf_{b \in B} d(a, b) \quad (3)$$

quantifies the distance from a point $a \in X$ to the subset $B \subseteq X$.

These metrics were computed for the RV segmentation from both SA and LA views, resulting in 4 measures for each cardiac phase. If one participant had a prediction missing for a specific subject, a zero value was assumed for DSC. A distance of 50mm was considered for HD, 10mm above the maximum HD distance computed across all participants and cases.

To obtain the final ranking, HD was min-max normalized across all subjects ($\hat{HD}$) to get a number between 0 and 1 for ED and ES phases in both SA and LA views independently. Due to the difference in dimensionality between SA and LA views, a weighted average was performed. The weighted metric, M, was obtained as follows:

$$M = \frac{0.75(DSC_{SA} + \hat{HD}_{SA}) + 0.25(DSC_{LA} + \hat{HD}_{LA})}{2} \quad (4)$$

where DSC and HD are the average of the corresponding metrics in ED and ES:

$$DSC = \frac{DSC_{ED} + DSC_{ES}}{2} \text{ and } HD = \frac{HD_{ED} + HD_{ES}}{2} \quad (5)$$

The normalized metrics returned a performance between 0 and 1, being 1 the value that a team would obtain if it had perfect results for every metric.

III. PARTICIPATING METHODS

More than 120 teams registered to download the M&Ms-2 training dataset, 17 submitted a solution for the final testing phase, and 15 teams presented their methodology as a paper to the STACOM Workshop (see Table IV for the participant details). Table V summarizes the main features of the submitted techniques, which are described in more detail in the following subsections.

A. Backbone architectures

There is a degree of diversity in the backbone architectures employed by the various participants (as depicted in Table V). This subsection will provide a comprehensive overview of the various architectures implemented by the participants.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2023.3267857

TABLE V

CHARACTERISTICS OF PARTICIPATING MODELS. SPATIAL AUGMENTATION INCLUDES ROTATIONS, FLIPPING, SCALING, AND DEFORMATIONS. INTENSITY AUGMENTATION INCLUDES GAUSSIAN NOISE, BRIGHTNESS, GAMMA, AND CONTRAST.

	Architecture			Data Augmentation		
Method	Backbone	Additional Features	Multiview	Spatial	Intensity	Other
P1 [40]	nnU-Net	Deformable Bayesian Convolutions	X	✓	✓	
P2 [41]	nnU-Net	Dropout + Batch Normalization	X	✓	✓	MRI-Specific
P3 [42]	nnU-Net	Default configuration	X	✓	✓	
P4 [43]	nnU-Net	Cross-view ROI detection LA → SA	✓	✓	✓	
P5 [44]	nnU-Net	Spatial and temporal Multi-channel input	X	✓	✓	Label propagation
P6 [45]	nnU-Net	ROI detection, Intensity-based Multi-channel input	X	✓	✓	SPADE Synthesis
P7 [46]	DLA	Cross-view refinement network	✓	✓	✓	Histogram Matching
P8 [47]	U-Net	Shared Bottleneck between views	✓	✓	X	
P9 [48]	xU-Net	3D Unit + 2D Unit with cross-view mid-fusion	✓	✓	✓	Test Time Augmentation
P10 [49]	U-Net	OoD detection and refinement (Convolutional Autoencoder)	X	X	X	
P11 [50]	U-Net	Single 2D network. Expansion, depth-wise, projection block	X	✓	X	Test Time Augmentation
P12 [51]	U-Net	Tranformer encoder in the bottleneck, cross-view consistency loss	✓	✓	✓	
P13 [52]	AttU-Net	ROI detection, cross-over Attention	✓	✓	✓	Histogram Matching, Fourier
P14 [53]	U-Net	Multi-view 3D mesh reconstruction	✓	X	X	
P15 [54]	MPFP+ViT	Multi-scale Feature Pyramid, Geometric Spatial Transformer	✓	✓	✓	In-painting

1) nnU-Net architectures: Six teams used the nnU-Net [55] framework as their baseline segmentation models (P1–P6). The nnU-Net framework includes 2D, 3D and cascaded 2D/3D U-Net [56] architectures. The choice of base architecture for a specific segmentation problem is left to the user. In the case of 3D short-axis (SA) volumes, the variations among P1-P6 models were primarily in terms of the input dimensionality, with some additional minor modifications to the base architecture. All of these methods produced separate models for each view.

P1 adopted a 2D nnU-Net for both SA and LA views and replaced its bottleneck convolutions with deformable Bayesian convolutions. Deformable convolutions enable an increased and adaptable receptive field without requiring additional convolution layers, while Bayesian convolutions improve generalisability and training speed.

P2 used a 3D nnU-Net for the SA view and a 2D nnU-Net for LA views and added batch normalisation instead of the default nnU-Net instance normalisation. P2 also added dropout of 0.2 to the intermediate layers of the network.

P3 used an ensemble of 2D and 3D nnU-Nets for the SA segmentation task and a regular 2D nnU-Net for the LA view. The default nnU-Net architectures were used.

P4 trained a default 2D nnU-Net for LA views and used its output to delimit SA views along the z axis and trained a default 3D nnU-Net with the extracted region of interest.

P5 and P6 used multi-channel late fusion approaches in their independent default 2D nnU-Nets for both SA and LA views.

P5 used stacks of three registered CMR consecutive images to train a three input channel 2D nnU-Net. While spatial and temporal information were porposed to generate the SA stacks, the LA stacks only incorporated temporal information.

P6 used six filtered versions of each 2D image as input for 2D nnU-Net with six input channel. The images feeded to this network were pre-processed extracting the region of interest by means of a regression CNN that delimited them to their bounding box.

2) U-Net architectures: Seven participants (P8–P14) constructed their architectures on top of a traditional U-Net.

P8 generated a multi-view SA-LA model consisting of two 2D U-Net structures with a shared bottleneck. Each of the out-of-plane 2D SA slices belonging to the same subject received the same complementary LA view and their features were concatenated in the bottleneck, training simultaneously both SA and LA views in a single end-to-end model.

P9 combined a 2D U-Net with a 3D U-Net in a unified model. In order to achieve this goal, both views are centered around the mean position of their original centroids. Moreover, both images are rotated to align their axes, where the LAx image is rotated to make its Y-axis match the Z-axis of the SAx stack. To take advantage of the complementary spatial context offered by both aligned views, a set of 3 cross-view modules were placed at the end of the three lowest levels in the compression path. Each cross-view module concatenated SA and LA information and retrieved a new set of spatially significant features using a 1x1 convolution layer. At inference time, SA and LA views were reoriented to their original pose.

P10 implemented for each view two 2D U-Nets and a 2D autoencoder. The architecture used in the implementation of the U-Net networks corresponds to the best methods presented in [19] and [21], while the architecture used in the implementation of the autoencoder network can be found in [57]. While the segmentation network used pairs of input images and their manual delineations, the autoencoder was trained to reconstruct delineations of the training set. The autoencoder loss was used as a quality control measure, being backpropagated to the U-Net when a poor quality was detected. At inference time, the best segmentation network was selected for each subject, taking in consideration the quality assessment of the autoencoder.

P11 used a single 2D U-Net for both views and replaced the standard convolutional blocks of its decoder with expansion,

depth-wise, and projection blocks. These blocks extract helpful information with less computational complexity and thus allowed to increase the number of channels in the decoding stage. Channels are then combined via depthwise convolutions and finally collapsed to the original depth in the projection stage. Additionally, P11 added residual blocks to the standard U-Net skip connections.

P12 proposed co-training a pair of 2D U-Nets end-to-end. The main modification of the backbone U-Net architecture used in each branch consist of the addition of an transformer module to the bottleneck that established self-attention mechanisms on high-level convolutional features. At training time, a SA slice and its complementary LA view were simultaneously fed to the paired U-Net. The segmentations obtained were then mapped between views using their complementary affine transformations. The final loss consisted of a combination of per-view standard DSC score and the co-segmentation SA to LA and LA to SA inter-view DSC scores.

P13 used Attention U-Net [58] as backbone. Initially, two 2D Attention U-Nets were utilized to extract the heart's location in both LA and SA views. The information from both views was then combined into one volume. For LA segmentation, the cropped LA slice and three mid-cavity SA slices were joined together. For SA segmentation, the cropped SA slice is combined with the cropped LA slice, allowing access to additional anatomy information in the basal and apical heart regions. Finally, each volume is processed as a multi-channel input through a separate Attention U-Net to produce the final segmentation masks for each view.

P14 used independent 2D U-Nets for LA and SA views and combined them into a 3D deformable model to improve quantification and volumetry. An initial 3D deformable model was triangulated directly from the SA segmentation contour points obtained from the network. SA apical and basal planes were estimated from the obtained LA segmentation and used to reconstruct the final SA volume.

3) Other architectures: P7 used 2D Deep Layer Aggregation (DLA) networks as a backbone for both SA and LA views. Being the backbone the same presented in [59]. The implementation consists of two stages: initially, two individual networks were employed to segment the SA and LA images independently. In the following stage, the results are then jointly refined using two additional networks. Four networks were trained independently in total, all having a similar structure except for the refinement networks, whose input comprised the original image, the respective 2D segmentation, and the aligned segmentation obtained from the complementary SA/LA view. Both stages were trained independently.

P15 propose a new hybrid 2D/3D geometric spatial Transformer Multi-Pass feature pyramid to simultanenously segment SA and LA views. The architecture consists of 2D SA/LA feature pyramid [60], independent 3D (SA) and 2D (LA) branches and finally a geometric spatial transformer (GST). The feature pyramid receives individual 2D in-plane complementary slices for both the SA and LA as inputs and extracts features at different downsampling levels. Then, the SA features are regrouped in a 3D SA stack, and a segmentation is obtained by means of a simple 3D convolutional residual block. LA features pyramids follow the same procedure on its 2D counterpart.

The GST takes as input the pre-computed affine matrix and the complementary LA and SA views. After projecting SA volume to its complementary LA view, both are concatenated and merged via a 2D convolutional block to obtain a refined LA prediction.

B. Data augmentation

Data augmentation (DA) is a widely utilized technique that helps to enhance the performance of deep learning algorithms through improved generalization and regularization. Its utilization in the medical imaging field has been well documented [61], and it has been consistently shown that incorporating DA can greatly benefit segmentation tasks in cardiovascular magnetic resonance imaging [19], [62].

All participants in the challenge (except P10 and P14) used some form of data augmentation to enhance their models. Specifically, two kinds of data augmentations were considered: (1) spatial transformations to increase sample size through flipping, rotation, 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. Both augmentation families seem particularly relevant for the M&Ms-2: while spatial transformations can reduce the gap between seen and unseen anatomies and pathologies, intensity-driven techniques are useful in the presence of heterogeneous imaging protocols and scanner vendors. Two teams performed data augmentation using only spatial transformations (P8, P11). Nine teams utilized intensity-based augmentations using standard image transformations such as blurring, change in brightness and contrast, or addition of Gaussian noise (P1–P9, P12–P13, P15). P3, P6 and P7 added histogram matching to their pool of intensity transformations. Additionally, P2 used MRI-specific augmentations such as random bias fields, random ghosting, and random motion artifacts to increase the textural variability of the images.

P5 and P6 added more sophisticated augmentations to their pipeline, and both methods used multi-channel inputs.

P5 registered temporal (SA and LA views) and spatial (z-axis SA view) and propagated the label information to unlabeled temporal phases to increase the training set. As described in the previous subsection, triplets of consecutive unlabeled images were effectively used to pretrain each SA and LA multi-channel net, taking as ground truth a registered label from an annotated cardiac phase. Since the propagated masks are not as accurate as the manual segmentations, the network was fine-tuned using the real labeled images and the adjacent registered cardiac phases.

P6 applied advanced image synthesis by using Generative Adversarial Networks (GANs). In particular, P6 used the method proposed in SPADE [63] to increase the number of samples per vendor and per cardiac region in an anatomically consistent way. The augmentation consisted of morphological manipulations of the segmentation masks to obtain synthetic images with the desired RV cavity shape. Multi-channel augmentations were then applied on top of synthesis, as a stack of intensity transformed channels and the the original (real or

synthetic) image. The transformed images were obtained using Laplacian, posterization, and edge-preserving filters.

On the other hand, P6 also proposed two data balancing strategies: (1) For SA stacks, the mid-ventricular slices cover most of the 3D volume, generating unbalance between basal, mid-ventricular and apical regions when using a 2D segmentation model. Approaches such as [64] alleviated this effect using balanced batches of the different short axial regions i.e. apical, basal and mid-ventricular regions. Following the same principle, P6 generated synthetic basal samples from randomly deformed segmentations.

(2) Since the provided dataset is acquired using 9 different scanners with a different number of samples per scanner and vendor, it is appropriate to consider some degree of unbalance related to domain shifts. Approaches such as [65], [66] tried to minimize the domain shift negative effects using domain adversarial training. P6 instead identified a set of outliers for each vendor based on the computed RV cardiac indices. Then, each vendor was synthetically augmented up to 1000 times, incorporating a 50% of outliers and a 50% of regular cases.

Finally, P13 added Fourier Domain Adaptation [67] to alleviate vendor differences.

IV. Results

As shown in Table II, a diverse testing set integrating nine scanners and eight cardiac pathologies was prepared for evaluating the final submissions with a total of 160 subjects. We show the obtained results per team, per cardiac region, per pathology and per clinical indices. Additionally, we show some qualitative results.

To understand and analyse the participating methods in this challenge, we have performed the following experimental comparisons. Firstly, we rank the participants exactly as it was presented during the challenge workshop. Secondly, we further dissect the results to emphasize different aspects and qualities of cardiac segmentation, such as pathological groups, cardiac regions or clinical indices. Thirdly, we perform a qualitative comparison of the approaches of the participants.

A. Team Ranking

The results of the challenge, as displayed in Table VI, present the evaluation of all participants using two relevant segmentation metrics (DSC and HD) for both SA and LS acqusitions. Additionally, the average inference time is included in terms of volumes per second for SA acquisitions and images per second for LA acquisitions . The inference time for methods using an unified model whose inference time could not be computed independently (P8, P9 and P15) for each view present a single inference time . Lastly, a Welch's t-test was conducted to determine statistical significance between participants' evaluation performance.

B. Results per Pathology

Figure 4 summarizes the average DSC per pathology according to equations (4) and (5). This dissection is particular relevant since accurate segmentation of different pathologies is critical for several clinical applications, including diagnosis, treatment planning, and monitoring disease progression.

In order to evaluate the ability of the proposed methods to generalize to new, unseen pathological groups, subjects with Dilated Right Ventricle and Tricuspidal Regurgitation were omitted from the training phase. A Mann-Whitney U rank test, with a significance level of 0.05, was conducted for each participant to compare their segmentation DSC scores for known and unknown pathologies. The results of this analysis are presented in Figure 5 in an organized manner, separated by imaging view and cardiac phase.

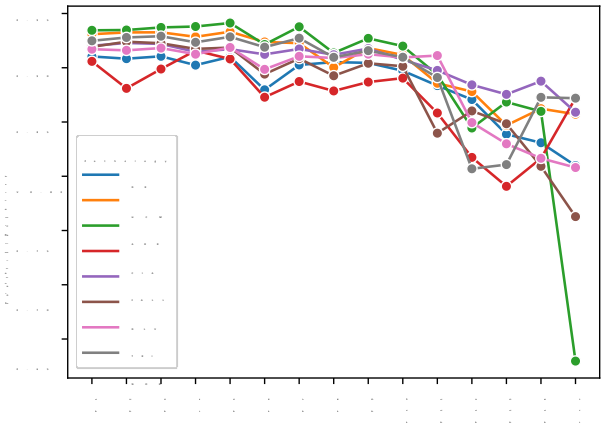

Fig. 4. Weighted average DSC per pathology according to equations (4) and (5).

C. Results per Cardiac Region

The examination of various segments of the heart, including the apical, basal, and mid-ventricular regions, is crucial for determining the individual impact each region may have on the segmentation error. To illustrate such impact, Figure 6 shows the average performance of P1–P5 in SA volumes from basal to apical planes. Further analysis is presented in relation to the detection of the basal plane, whose contribution to segmentation accuracy is greater than another regions: detection rate of the basal plane, as shown in Figure 7, presents the number of subjects per participant where there was a disagreement regarding the manual delineation in the detected first basal slice.

D. Clinical Measurements

In the assessment of cardiac function, clinical metrics such as End-Diastolic and End-Systolic Volumes, and Right Ventricle Ejection Fraction (RVEF) are commonly utilized indices. However, geometrical metrics, such as DSC and HD, may not always correlate with these indices. This lack of correlation is attributed to the scalar, rather than spatial, nature of the clinical indices, which can result in good estimations of volumes and RVEF even when the contour is not accurately defining the cardiac structures. For such reason, the beforementioned clinical measurements are presented in Table VII, in term of

TABLE VI

DSC AND HD, MEAN AVERAGE VOLUME ERROR AND INFERENCE TIME FOR THE FINAL SUBMISSIONS OF ALL PARTICIPANTS. HD IS MEASURED IN MILLIMETERS. VOLUME ERROR IS MEASURED IN MILILITRES. INFERENCE TIME IS MEASURED IN SECONDS PER VOLUME.

	SA					LA				
	ED		ES			ED		ES		
Method	DSC	HD	DSC	HD	Inference (s)	DSC	HD	DSC	HD	Inference (s)
P1	0.934	9.610	0.910	10.032	1.72	0.935	6.227	0.904	5.935	0.34
P2	0.932	10.078	0.910	9.782	0.86	0.935	6.028	0.905	6.188	0.11
P3	0.940	10.122	0.914	9.987	1.8	0.931	6.337	0.904	5.976	0.42
P4	0.933	10.563	0.907	10.050	2.22	0.930	6.246	0.902	6.097	0.54
P5	0.937	10.879	0.913	10.300	2.43	0.935	6.056	0.903	6.031	0.17
P6	0.927	9.941	0.897	10.307	2.74	0.907	8.444	0.883	7.265	0.56
P7	0.932	10.517	0.903	10.880	4.11	0.923	7.371	0.902	6.019	1.23
P8	0.923	11.258	0.897	11.062	2.23	0.910	7.757	0.882	6.933	--
P9	0.924	11.327	0.898	11.447	2.89	0.922	7.173	0.900	6.391	0.34
P10	0.916	11.681	0.890	11.347	2.12	0.923	7.846	0.894	6.970	0.18
P11	0.909	15.275	0.880	14.606	0.67	0.888	9.323	0.854	8.347	0.42
P12	0.844	15.495	0.821	16.750	2.34	0.887	9.733	0.851	9.659	0.30
P13	0.873	16.682	0.791	18.499	3.12	0.852	11.325	0.829	9.591	0.67
P14	0.883	17.024	0.838	17.803	4.27	0.839	13.303	0.809	13.716	-
P15	0.852	19.430	0.821	19.117	1.54	0.814	18.629	0.781	17.198	

Boldface numbers are the best results for each column. Blue numbers represent results are not significantly different compared to the top-performing method for each column (p-value > 0.01 for Welch's t-test)

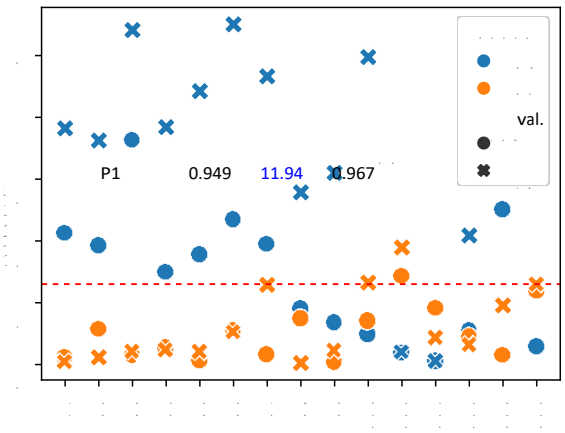

Fig. 5. Statistical difference according to the Mann-Whitney U rank test for DSC scores between seen and unseen pathologies. The red dashed line stands for the 0.05 significance threshold.

TABLE VII

CLINICAL METRICS FOR THE 15 PARTICIPATING METHODS.

	Volume ED		Volume ES		RVEF		
Method	corr. mL	mae val.	corr.	mae mL	corr. val.	mae %	bias %±σ
				7.63	0.878	4.81	-0.31±6.9
				7.63	0.873	4.54	0.89±6.7
P2	0.952	11.14	0.967	7.63	0.891	4.4	0.65±6.2
P3	0.963	10.16	0.967	8.3	0.87	4.67	-0.02±6.8
P4	0.958	11.07	0.965	7.22	0.892	4.36	0.18 ±6.5
P5	0.955	10.83	0.97	9.16	0.864	4.77	-0.14±7.5
P6	0.915	13.49	0.936	8.52	0.892	4.64	-0.63±6.4
P7	0.951	11.61	0.964	9.04	0.855	5.26	0.36±8.0
P8	0.95	11.94	0.954	8.93	0.871	4.74	0.41 ±6.6
P9	0.954	12.12	0.959	10.98	0.764	6.65	0.6±9.7
P10	0.944	14.55	0.93	11.01	0.772	4.96	1.38 ±9.3
P11	.913	15.79	0.917	15.24	0.491	11.45	-7.88±15.0
P12	0.744	32.18	0.823	17.06	0.674	9.06	1.42±13.2
P13	0.883	21.37	0.865	13.00	0.671	7.74	-1.1 ±11.7
P14	0.898	16.99	0.886	14.87	0.55	11.7	4.19±18.1
P15	0.732	23.10	0.825				

Boldface numbers are the best results for each column. Blue numbers represent results that are not significantly different compared to the top-performing method for each column (p-value > 0.01 for Welch's t-test)

i) correlation (corr), ii) mean average error (mae), and iii) bias. Note that, it may be the case where ED volume is not accurately predicted. In such case the RVEF, defined as $(Vol_{ED} - Vol_{ES})/Vol_{ED}$ can increase considerably or be infinite. In such cases a RVEF of 100% was considered.

Outliers play a crucial role on integrating automatic segmentation methods in clinical practice, as a single missed case or a significant discrepancy in a few instances can have a greater impact than a small average improvement that may not make a noticeable difference in diagnostic tasks. In Table VIII, the number of cases exhibiting an RV ejection fraction above various thresholds is presented, alongside the number of cases in which computation was not feasible due to a missing segmentation in some of the cardiac phases.

TABLE VIII

NUMBER OF PATIENTS ABOVE DIFFERENT RVEF ERROR THRESHOLDS.

	RV Ejection Fraction Mean Average Error				
Method	≥ 5%	≥ 10%	≥ 15%	≥ 20%	Missing
P1	60	19	6	3	0
P2	59	16	4	2	1
P3	55	16	5	2	0
P4	57	14	4	3	0
P5	47	17	4	1	0
P6	49	15	7	3	0
P7	52	19	5	2	0
P8	58	22	9	6	0
P9	67	12	1	1	0
P10	59	34	13	8	1
P11	59	20	6	3	0
P12	116	70	36	14	1
P13	87	49	29	18	1
P14	79	37	15	11	2
P15	67	34	23	23	16

E. Qualitative results

Figure 8 provides some visual examples from different teams to discuss the possible limitations and strengths of the implemented methods. In the first row, complex basal regions for short-axis views are correctly captured by various multi-view approaches. All of these examples were not segmented by the top 5 non-multi-view strategies. In the second row, a pathological subject with a high degree of remodeling in the RV is not correctly segmented by the best-performing methods, capturing the surrounding tissue instead of the cardiac structure. P10 captured the cardiac structure as well as the surrounding tissue. P13 and P8 delineated only the cardiac structure with different degrees of accuracy. These methods merged SA and LA views in their networks without additional cross-view affine projections. Finally, the last two rows show highly remodeled right ventricular cavities correctly segmented by top-performing methods.

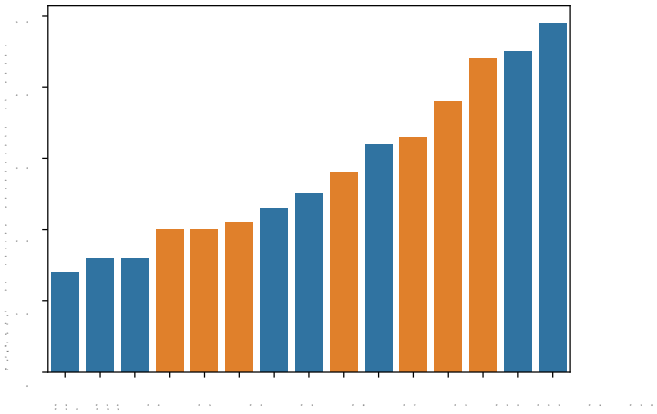

Fig. 7. Number of not segmented slices at basal region. In blue, multi-view approaches. In orange, non-multi-view approaches.

V. Discussion

This study presents a comprehensive evaluation of various automatic deep learning-based methods for multi-disease, multi-view, and multi-center right ventricular segmentation in cardiac magnetic resonance imaging (CMR). The 15 participants employed a diverse range of methodologies, including the choice of backbone architecture, number of stages, multi-view fusion, and data augmentation strategies. In addition to a large training sample of 160 cases, the authors were given 20 opportunities to optimize the parameters and characteristics of their models during the validation process using a well-stratified validation set of 40 cases. A Codalab-based automatic submission system was provided to allow for public comparison of performance and promote fair and dynamic competition between participants.

A. Summary of the challenge results

It can be concluded that the performance of the different proposals, and in particular for P1-P5, is relatively comparable. Statistical analysis has shown limited significant differences between the methods, with no clear advantage for any of the participants.

From a general point of view, our study supports several observations found in the previous edition of the challenge and other studies based on different CMR datasets. Specifically, the results confirm that end-diastolic segmentations are more accurate than end-systolic segmentations for the right ventricle. Additionally, the accuracy of segmentation decreases in the basal regions that are susceptible to under-segmentation and also is impacted in the apical regions due to their smaller size relative to the rest of the ventricular cavity.

The accuracy of segmentation is more stable across cardiac phases in comparison to previous challenges such as MMs-1 or ACDC, with an improvement of 0.042 in average DSC over MMs-1 and a comparable performance with ACDC (+0.004 average DSC), despite being MMs-2 a heterogeneous cohort.

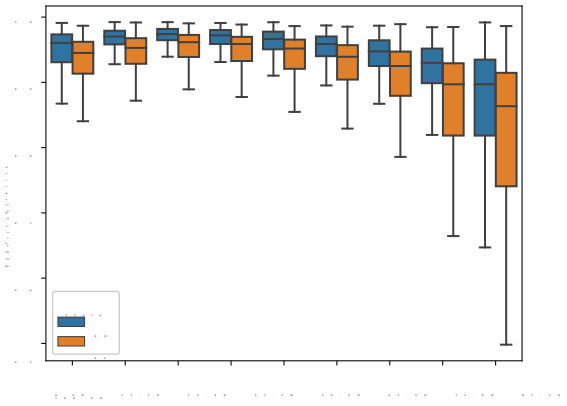

Fig. 6. Average performance of the top 5 ranked methods in SA from basal (0%) to the apical (100%) regions.

B. Analysis of Pathologies

One of the relevant aspects of the challenge consists on evaluating the generalization capacity of the proposed methods to new, unseen pathologies. For this reason, the participants trained their models without access to subjects belonging to the Dilated Right Ventricle and Tricuspidal Regurgitation groups. Figure 4 shows that unseen pathologies perform consistently worse with exception of Inter-atrial communication. It is remarkable that three out of the ten subjects belonging to this group had a closure device visible in the basal region of the image.

We investigated in more detail the statistical differences between both, seen and unseen groups by analyzing independently the two annotated cardiac phases and the two views available in each study. The results in Figure 5 present some degree of statistical significance between seen and unseen pathologies for both SA and LA end-diastolic phases. This finding reveals the need for including diverse cardiac morphologies to improve model generalisability.

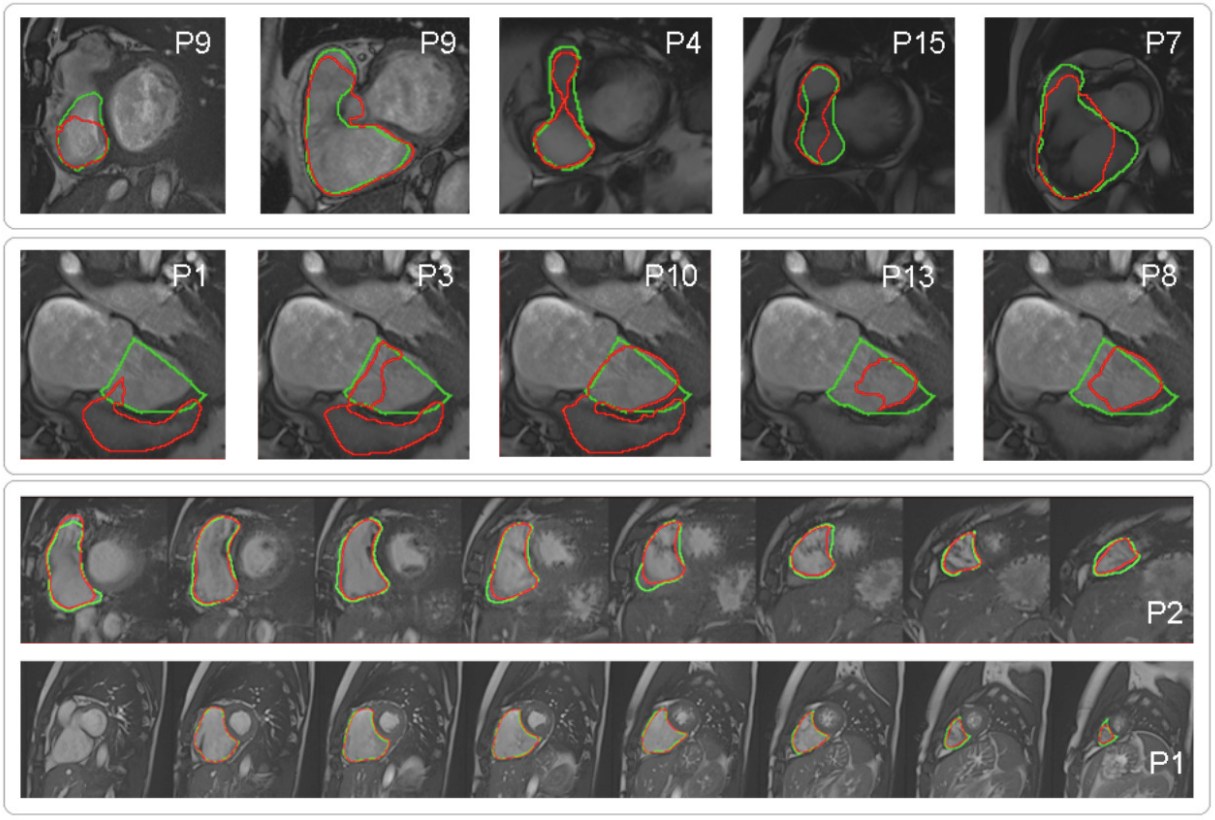

Fig. 8. Prediction examples for some of the presented methods. The first row shows satisfactory segmentations at conflictive basal regions for SA images that were missed by non-multi-view approaches but correctly captured by multi-view methods. The second row shows a pathological subject with severe right ventricular dilation that was only correctly captured by multi-view methods. The last two rows show pathological subjects from unseen pathologies correctly segmented by top-ranked methods. Color correspondence: ground truth (green), prediction (red).

C. Single- versus Multi-view Models

Regarding multi-view approaches, eight out of fifteen participants utilized the complementary information between views. Although a definitive conclusion cannot be drawn on the general benefits, the evaluation suggests that multi-view methods have the potential to improve basal plane detection in certain circumstances. Specifically, participants P9, P14, and P4 achieved a lower number of not-segmented basal slices. Additionally, some of the multi-view approaches presented a better RVEF stability. The solution proposed by P9 obtained the lowest number of cases with a RVEF error ≥ 10%, improving significantly the results obtained by P1-P5, as expressed in Table VIII. Further research is required to incorporate multi-view techniques into thoroughly optimized frameworks such as nnU-Net.

D. Impact on clinical indices

We also assesed the participating methods by computing the clincal indices derived from the generated segmentations. The results were consistent with the ranking presented in Table VI, with almost any statistical difference between the Top-10 ranked methods. Interestingly, the multi-view approach P9 presented a more consistent Ejection Fraction across patients, with fewer cases with RVEF error greater than 10%. This point is specially relevant for diagnostic tasks.

E. Further considerations

Due to the high heterogeneity of the presented dataset, one could argue that there are many sensible parameters affecting the segmentation performance. Different image dimensions, in-plane resolutions or field strengths may be critical parameters for a DL segmentation algorithms.

Field strenght: only five out of twelve samples obtained using 3T scanners were included in the test set. Despite the small sample size, there were no substantial differences in the segmentation performance between 1.5T and 3T acquisitions.

In-plane resolution: In Table III, we presented a wide range of in-plane resolutions and volume dimensions directly related to the acquisition scanners. When comparing the average performance of P1-P5 across different scanners, we obtain a stable DSC of 0.912±0.016 for long-axis 4-chamber images and 0.922±0.011 for short-axis volumes. Interestingly, the learning methods were able to generalise correctly to the heterogeneous set of scanners, resolutions and protocols present during the training stage.

F. Future work

In addition to the analyses and results presented in this paper, we also provide the M&Ms-2 dataset open-access for the community, which can be downloaded from the M&Ms-2

website[2]. In conjunction with M&Ms-1, it represents the most heterogeneous dataset ever compiled in CMRs image analysis, comprising CMRs from various imaging protocols and cardiology units. It also includes a wide range of cardiovascular diseases and multi-view information. It is anticipated that the scientific community will embrace the dataset as a comprehensive resource to support a wide range of automated cardiac imaging research initiatives, including automatic pathology assessment, multi-scanner and multi-view image registration, multi-structure segmentation, cardiac imaging quantification, strain and motion analysis, and image synthesis. Further efforts will focus on incorporating 2-chamber and 3-chamber long-axis views to fully leverage the multi-view aspect of cardiac magnetic resonance studies. The integration of diverse disease characterization with these various views will also be pivotal in facilitating automatic evaluation and diagnosis.

VI. Conclusions

To summarize, the key conclusions are:

1) The main findings correlate with the obtained results in previous CMR segmentation challenges: end-systolic phase and basal and apical cardiac regions are more conflictive than their counterparts.
2) nnU-Net based approaches proved to be more effective overall. Additional effort is required to incorporate complex models into optimized frameworks such as nnU-Net for a fair evaluation of different architectural proposals.
3) Further research is needed regarding generalisation: it is essential to develop methods that can generalize well across a wide range of pathologies and patient populations.The results highlight the need to integrate a variety of cardiac diseases, centers, scanners, and acquisition protocols to generate robust DL approaches in the biomedical imaging analysis domain.
4) Regarding multi-view methods, it cannot be definitively concluded that they bring a significant improvement to the CMR RV segmentation problem. However, further study is necessary in order to perform a conclusive assessment of their impact and potential.

Ethical Approval

The study was approved by the ethics committee of the three centers involved: Vall d'Hebron Hospital, Sagrada Familia Hospital, and Dexeus Hospital. Written informed consent was obtained from all participants.

Acknowledgment

This work was partly funded from the European H2020 programme under grant agreement no. 825903 (euCanSHare project) and grant agreement no. 965345 (HealthyCloud project). This work has been partially supported by the Spanish project PID2019-105093GB-I00 and by ICREA under the ICREA Academia programme. K. Lekadir is supported by the Ramon y Cajal Program of the Spanish Ministry of Economy and Competitiveness under grant no. RYC-2015- 17183. A. Guala has received funding from the Spanish Ministry of Science, Innovation and Universities (IJC2018-037349-I) and from "la Caixa" Foundation (LCF/BQ/PR22/11920008). S. Queirós is supported by National funds, through the Foundation for Science and Technology (FCT, Portugal; project PTDC/EMD-EMD/1140/2020 and grant CEECIND/03064/2018).

[2]www.ub.edu/mnms-2

References

[1] 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.

[2] 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.

[3] F. Haddad, R. Doyle, D. J. Murphy, and S. A. Hunt. Right ventricular function in cardiovascular disease, part II. Circulation, 117(13):1717–1731, apr 2008.

[4] V. H. Rigolin, P. A. Robiolio, J. S. Wilson, J. K. Harrison, and T. M. Bashore. Impact of right ventricular involvement on mortality and morbidity in patients with inferior myocardial infarction. Catheterization and Cardiovascular Diagnosis, 35:18–28, 1995.

[5] M. Amsallem, O. Mercier, Y. Kobayashi, K. Moneghetti, and F. Haddad. Forgotten no more: a focused update on the right ventricle in cardiovascular disease. JACC: Heart Failure, 6(11):891–903, 2018.

[6] M. I. Burgess, N. Mogulkoc, R. J. Bright-Thomas, P. Bishop, J. J. Egan, and S. G. Ray. Comparison of echocardiographic markers of right ventricular function in determining prognosis in chronic pulmonary disease. Journal of the American Society of Echocardiography, 15(6):633–639, 2002.

[7] S. R. Mehta, J. W. Eikelboom, M. K. Natarajan, R. Diaz, C. Yi, R. J. Gibbons, and S. Yusuf. Impact of right ventricular involvement on mortality and morbidity in patients with inferior myocardial infarction. Journal of the American College of Cardiology, 37(1):37–43, 2001.

[8] P. de Groote, A. Millaire, C. Foucher-Hossein, O. Nugue, X. Marchandise, G. Ducloux, and J.-M. Lablanche. Right ventricular ejection fraction is an independent predictor of survival in patients with moderate heart failure. Journal of the American College of Cardiology, 32(4):948–954, 1998.

[9] J. M. Del Rio, L. Grecu, and A. Nicoara. Right ventricular function in left heart disease. In Seminars in Cardiothoracic and Vascular Anesthesia, volume 23, pages 88–107. SAGE Publications Sage CA: Los Angeles, CA, 2019.

[10] K. Keramida, G. Lazaros, and P. Nihoyannopoulos. Right ventricular involvement in hypertrophic cardiomyopathy: Patterns and implications. Hellenic Journal of Cardiology, 61(1):3–8, 2020.

[11] L. Bosch, C. S. Lam, L. Gong, S. P. Chan, D. Sim, D. Yeo, F. Jaufeerally, K. T. G. Leong, H. Y. Ong, T. P. Ng, A. M. Richards, F. Arslan, and L. H. Ling. Right ventricular dysfunction in left-sided heart failure with preserved versus reduced ejection fraction. European Journal of Heart Failure, 19(12):1664–1671, jun 2017.

[12] N. M. Fine, L. Chen, P. M. Bastiansen, R. P. Frantz, P. A. Pellikka, J. K. Oh, and G. C. Kane. Outcome prediction by quantitative right ventricular function assessment in 575 subjects evaluated for pulmonary hypertension. Circulation: Cardiovascular Imaging, 6(5):711–721, sep 2013.

[13] S. Ghio, P. L. Temporelli, C. Klersy, A. Simioniuc, B. Girardi, L. Scelsi, A. Rossi, M. Cicoira, F. T. Genta, and F. L. Dini. Prognostic relevance of a non-invasive evaluation of right ventricular function and pulmonary artery pressure in patients with chronic heart failure. European Journal of Heart Failure, 15(4):408–414, apr 2013.

[14] A. Gulati, T. F. Ismail, A. Jabbour, F. Alpendurada, K. Guha, N. A. Ismail, S. Raza, J. Khwaja, T. D. Brown, K. Morarji, E. Liodakis, M. Roughton, R. Wage, T. C. Pakrashi, R. Sharma, J.-P. Carpenter, S. A. Cook, M. R. Cowie, R. G. Assomull, D. J. Pennell, and S. K. Prasad. The prevalence and prognostic significance of right ventricular systolic dysfunction in nonischemic dilated cardiomyopathy. Circulation, 128(15):1623–1633, oct 2013.

[15] B. S. Rana, S. Robinson, R. Francis, M. Toshner, M. J. Swaans, S. Agarwal, R. de Silva, A. A. Rana, and P. Nihoyannopoulos. Tricuspid regurgitation and the right ventricle in risk stratification and timing of intervention. Echo Research &amp Practice, 6(1):R26–R40, mar 2019.

[16] L. Lopez, M. S. Cohen, R. H. Anderson, A. N. Redington, D. G. Nykanen, D. J. Penny, J. E. Deanfield, and B. W. Eidem. Unnatural history of the right ventricle in patients with congenitally malformed hearts. Cardiology in the Young, 20(S3):107–112, dec 2010.

[17] L. L. Mertens and M. K. Friedberg. Imaging the right ventricle—current state of the art. Nature Reviews Cardiology, 7(10):551–563, aug 2010.

[18] N. Jones, A. T. Burns, and D. L. Prior. Echocardiographic assessment of the right ventricle–state of the art. Heart, Lung and Circulation, 28(9):1339–1350, sep 2019.

[19] V. M. Campello, P. Gkontra, C. Izquierdo, C. Martın-Isla, A. Sojoudi, P. M. Full, K. Maier-Hein, Y. Zhang, Z. He, J. Ma, et al. Multi-centre, multi-vendor and multi-disease cardiac segmentation: the M&Ms challenge. IEEE Transactions on Medical Imaging, 40(12):3543–3554, 2021.

[20] 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.

[21] 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.

[22] P. V. Tran. A fully convolutional neural network for cardiac segmentation in short-axis mri.

[23] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, pages 234–241. Springer, 2015.

[24] J. Lieman-Sifry, M. Le, F. Lau, S. Sall, and D. Golden. Fastventricle: cardiac segmentation with enet. In Functional Imaging and Modelling of the Heart: 9th International Conference, FIMH 2017, Toronto, ON, Canada, June 11-13, 2017, Proceedings, pages 127–138. Springer, 2017.

[25] 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.

[26] C. F. Baumgartner, L. M. Koch, M. Pollefeys, and E. Konukoglu. An exploration of 2d and 3d deep learning techniques for cardiac mr image segmentation. In Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges: 8th International Workshop, STACOM 2017, Held in Conjunction with MICCAI 2017, Quebec City, Canada, September 10-14, 2017, Revised Selected Papers 8, pages 111–119. Springer, 2018.

[27] F. Isensee, J. Petersen, S. A. A. Kohl, P. Jäger, and K. Maier-Hein. nnu-net: Breaking the spell on successful medical image segmentation. ArXiv, abs/1904.08128, 2019.

[28] F. Isensee, P. Jaeger, P. M. Full, I. Wolf, S. Engelhardt, and K. Maier-Hein. Automatic cardiac disease assessment on cine-mri via time-series segmentation and domain specific features. ArXiv, abs/1707.00587, 2017.

[29] M. A. Zuluaga, M. J. Cardoso, and S. Ourselin. Automatic right ventricle segmentation using multi-label fusion in cardiac mri. arXiv preprint arXiv:2004.02317, 2020.

[30] M. G. Oghli, A. Mohammadzadeh, R. Kafieh, and S. Kermani. A hybrid graph-based approach for right ventricle segmentation in cardiac mri by long axis information transition. Physica Medica, 54:103–116, 2018.

[31] G. Luo, R. An, K. Wang, S. Dong, and H. Zhang. A deep learning network for right ventricle segmentation in short-axis mri. In 2016 Computing in Cardiology Conference (CinC), pages 485–488. IEEE, 2016.

[32] H. Yang, Z. Liu, and X. Yang. Right ventricle segmentation in short-axis mri using a shape constrained dense connected u-net. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 532–540. Springer, 2019.

[33] J. Li, Z. L. Yu, Z. Gu, H. Liu, and Y. Li. Dilated-inception net: multi-scale feature aggregation for cardiac right ventricle segmentation. IEEE Transactions on Biomedical Engineering, 66(12):3499–3508, 2019.

[34] G. Borodin and O. Senyukova. Right ventricle segmentation in cardiac mr images using u-net with partly dilated convolution. In International Conference on Artificial Neural Networks, pages 179–185. Springer, 2018.

[35] K. Huang, L. Xu, Y. Zhu, and P. Meng. Au-snake based deep learning network for right ventricle segmentation. Medical Physics, 2022.

[36] M. R. Avendi, A. Kheradvar, and H. Jafarkhani. Automatic segmentation of the right ventricle from cardiac mri using a learning-based approach. Magnetic resonance in medicine, 78(6):2439–2448, 2017.

[37] S. Karimi-Bidhendi, A. Arafati, A. L. Cheng, Y. Wu, A. Kheradvar, and H. Jafarkhani. Fully-automated deep-learning segmentation of pediatric cardiovascular magnetic resonance of patients with complex congenital heart diseases. Journal of cardiovascular magnetic resonance, 22(1):80, 2020.

[38] J. Chen, H. Zhang, W. Zhang, X. Du, Y. Zhang, and S. Li. Correlated regression feature learning for automated right ventricle segmentation. IEEE journal of translational engineering in health and medicine, 6:1–10, 2018.

[39] A. Pavao, I. Guyon, A.-C. Letournel, X. Baró, H. Escalante, S. Escalera, T. Thomas, and Z. Xu. CodaLab Competitions: An open source platform to organize scientific challenges. PhD thesis, Université Paris-Saclay, FRA., 2022.

[40] M. J. Fulton, C. R. Heckman, and M. E. Rentschler. Deformable bayesian convolutional networks for disease-robust cardiac mri segmentation. In International Workshop on Statistical Atlases and Computational Models of the Heart, pages 296–305. Springer, 2021.

[41] T. W. Arega, F. Legrand, S. Bricq, and F. Meriaudeau. Using mri-specific data augmentation to enhance the segmentation of right ventricle in multi-disease, multi-center and multi-view cardiac mri. In International Workshop on Statistical Atlases and Computational Models of the Heart, pages 250–258. Springer, 2021.

[42] K. Punithakumar, A. Carscadden, and M. Noga. Automated segmentation of the right ventricle from magnetic resonance imaging using deep convolutional neural networks. In International Workshop on Statistical Atlases and Computational Models of the Heart, pages 344–351. Springer, 2021.

[43] L. Li, W. Ding, L. Huang, and X. Zhuang. Right ventricular segmentation from short-and long-axis mris via information transition. In International Workshop on Statistical Atlases and Computational Models of the Heart, pages 259–267. Springer, 2021.

[44] X. Sun, L.-H. Cheng, and R. J. Geest. Right ventricle segmentation via registration and multi-input modalities in cardiac magnetic resonance imaging from multi-disease, multi-view and multi-center. In International Workshop on Statistical Atlases and Computational Models of the Heart, pages 241–249. Springer, 2021.

[45] Y. Al Khalil, S. Amirrajab, J. Pluim, and M. Breeuwer. Late fusion u-net with gan-based augmentation for generalizable cardiac mri segmentation. In International Workshop on Statistical Atlases and Computational Models of the Heart, pages 360–373. Springer, 2021.

[46] D. Liu, Z. Yan, Q. Chang, L. Axel, and D. N. Metaxas. Refined deep layer aggregation for multi-disease, multi-view & multi-center cardiac mr segmentation. In International Workshop on Statistical Atlases and Computational Models of the Heart, pages 315–322. Springer, 2021.

[47] S. Jabbar, S. T. Bukhari, and H. Mohy-ud Din. Multi-view sa-la net: A framework for simultaneous segmentation of rv on multi-view cardiac mr images. In International Workshop on Statistical Atlases and Computational Models of the Heart, pages 277–286. Springer, 2021.

[48] S. Queirós. Right ventricular segmentation in multi-view cardiac mri using a unified u-net model. In International Workshop on Statistical Atlases and Computational Models of the Heart, pages 287–295. Springer, 2021.

[49] F. Galati and M. A. Zuluaga. Using out-of-distribution detection for model refinement in cardiac image segmentation. In International Workshop on Statistical Atlases and Computational Models of the Heart, pages 374–382. Springer, 2021.

[50] M. Mazher, A. Qayyum, A. Benzinou, M. Abdel-Nasser, and D. Puig. Multi-disease, multi-view and multi-center right ventricular segmentation in cardiac mri using efficient late-ensemble deep learning approach. In International Workshop on Statistical Atlases and Computational Models of the Heart, pages 335–343. Springer, 2021.

[51] Z. Gao and X. Zhuang. Consistency based co-segmentation for multi-view cardiac mri using vision transformer. In International Workshop

on Statistical Atlases and Computational Models of the Heart, pages 306–314. Springer, 2021.

[52] M. Beetz, J. Corral Acero, and V. Grau. A multi-view crossover attention u-net cascade with fourier domain adaptation for multi-domain cardiac mri segmentation. In International Workshop on Statistical Atlases and Computational Models of the Heart, pages 323–334. Springer, 2021.

[53] L. Tautz, L. Walczak, C. Manini, A. Hennemuth, and M. Hullebrand. 3d right ventricle reconstruction from 2d u-net segmentation of sparse short-axis and 4-chamber cardiac cine mri views. In International Workshop on Statistical Atlases and Computational Models of the Heart, pages 352–359. Springer, 2021.

[54] C. Galazis, H. Wu, Z. Li, C. Petri, A. A. Bharath, and M. Varela. Tempera: Spatial transformer feature pyramid network for cardiac mri segmentation. In International Workshop on Statistical Atlases and Computational Models of the Heart, pages 268–276. Springer, 2021.

[55] F. Isensee, P. F. Jager, S. A. A. Kohl, J. Petersen, and K. Maier-Hein. Automated design of deep learning methods for biomedical image segmentation. arXiv: Computer Vision and Pattern Recognition, 2019.

[56] 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.

[57] P. Bergmann, S. Lowe, M. Fauser, D. Sattlegger, and C. Steger. Improving unsupervised defect segmentation by applying structural similarity to autoencoders. arXiv preprint arXiv:1807.02011, 2018.

[58] 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.

[59] F. Yu, D. Wang, E. Shelhamer, and T. Darrell. Deep layer aggregation. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, June 2018.

[60] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2117–2125, 2017.

[61] P. Chlap, H. Min, N. Vandenberg, J. Dowling, L. Holloway, and A. Haworth. A review of medical image data augmentation techniques for deep learning applications. Journal of Medical Imaging and Radiation Oncology, 65(5):545–563, 2021.

[62] C. Chen, W. Bai, R. H. Davies, A. N. Bhuva, C. H. Manisty, J. B. Augusto, J. C. Moon, N. Aung, A. M. Lee, M. M. Sanghvi, et al. Improving the generalizability of convolutional neural network-based segmentation on cmr images. Frontiers in cardiovascular medicine, 7:105, 2020.

[63] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu. Semantic image synthesis with spatially-adaptive normalization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019.

[64] J. Mariscal-Harana, N. Kifle, R. Razavi, A. P. King, B. Ruijsink, and E. Puyol-Anton.´ Improved ai-based segmentation of apical and basal slices from clinical cine cmr. In Statistical Atlases and Computational Models of the Heart. Multi-Disease, Multi-View, and Multi-Center Right Ventricular Segmentation in Cardiac MRI Challenge: 12th International Workshop, STACOM 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Revised Selected Papers, pages 84–92. Springer, 2022.

[65] 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: 11th International Workshop, STACOM 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4, 2020, Revised Selected Papers 11, pages 196–207. Springer, 2021.

[66] 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: 11th International Workshop, STACOM 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4, 2020, Revised Selected Papers 11, pages 228–237. Springer, 2021.

[67] Y. Yang and S. Soatto. Fda: Fourier domain adaptation for semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 4085–4095, 2020.