



UNIVERSITY OF LEEDS

This is a repository copy of *Cardiac segmentation on late gadolinium enhancement MRI: A benchmark study from multi-sequence cardiac MR segmentation challenge*.

White Rose Research Online URL for this paper:

<https://eprints.whiterose.ac.uk/190049/>

Version: Accepted Version

Article:

Zhuang, X, Xu, J, Luo, X et al. (20 more authors) (2022) Cardiac segmentation on late gadolinium enhancement MRI: A benchmark study from multi-sequence cardiac MR segmentation challenge. *Medical Image Analysis*, 81. 102528. ISSN 1361-8415

<https://doi.org/10.1016/j.media.2022.102528>

© 2022 Elsevier B.V. All rights reserved. This manuscript version is made available under the CC-BY-NC-ND 4.0 license <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

Reuse

This article is distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs (CC BY-NC-ND) licence. This licence only allows you to download this work and share it with others as long as you credit the authors, but you can't change the article in any way or use it commercially. More information and the full terms of the licence here: <https://creativecommons.org/licenses/>

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.



eprints@whiterose.ac.uk
<https://eprints.whiterose.ac.uk/>

Cardiac Segmentation on Late Gadolinium Enhancement MRI: A Benchmark Study from Multi-Sequence Cardiac MR Segmentation Challenge

Xiahai Zhuang^{1*}, Jiahang Xu^{1*}, Xinzhe Luo¹, Chen Chen³, Cheng Ouyang³, Daniel Rueckert³, Victor M. Campello⁴, Karim Lekadir⁴, Sulaiman Vesal⁵, Nishant RaviKumar⁵, Yashu Liu⁶, Gongning Luo⁶, Jingkun Chen⁷, Hongwei Li⁸, Buntheng Ly⁹, Maxime Sermesant⁹, Holger Roth¹⁰, Wentao Zhu¹⁰, Jiexiang Wang¹¹, Xinghao Ding¹¹, Xinyue Wang¹², Sen Yang^{12,13}, Lei Li^{1,2*}

¹*School of Data Science, Fudan University, Shanghai, China*

²*School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, China*

³*Biomedical Image Analysis Group, Imperial College London, London, UK*

⁴*Department Mathematics & Computer Science, Universitat de Barcelona, Barcelona, Spain*

⁵*Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany*

⁶*School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China*

⁷*Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen, China*

⁸*Department of Informatics, Technical University of Munich, Germany*

⁹*INRIA, Université Côte d'Azur, Sophia Antipolis, France*

¹⁰*NVIDIA, Bethesda, USA*

¹¹*School of Informatics, Xiamen University, Xiamen, China*

¹²*College of Electrical Engineering, Sichuan University, Chengdu, China*

¹³*Tencent AI Lab, Shenzhen, China*

arXiv:2006.12434v2 [eess.IV] 17 Jul 2021

Abstract

Accurate computing, analysis and modeling of the ventricles and myocardium from medical images are important, especially in the diagnosis and treatment management for patients suffering from myocardial infarction (MI). Late gadolinium enhancement (LGE) cardiac magnetic resonance (CMR) provides an important protocol to visualize MI. However, automated segmentation of LGE CMR is still challenging, due to the indistinguishable boundaries, heterogeneous intensity distribution and complex enhancement patterns of pathological myocardium from LGE CMR. Furthermore, compared with the other sequences LGE CMR images with gold standard labels are particularly limited, which represents another obstacle for developing novel algorithms for automatic segmentation of LGE CMR. This paper presents the selective results from the Multi-Sequence Cardiac MR (MS-CMR) Segmentation challenge, in conjunction with MICCAI 2019. The challenge offered a data set of paired MS-CMR images, including auxiliary CMR sequences as well as LGE CMR, from 45 patients who underwent cardiomyopathy. It was aimed to develop new algorithms, as well as benchmark existing ones for LGE CMR segmentation and compare them objectively. In addition, the paired MS-CMR images could enable algorithms to combine the complementary information from the other sequences for the segmentation of LGE CMR. Nine representative works were selected for evaluation and comparisons, among which three methods are unsupervised methods and the other six are supervised. The results showed that the average performance of the nine methods was comparable to the inter-observer variations. Particularly, the top-ranking algorithms from both the supervised and unsupervised methods could generate reliable and robust segmentation results. The success of these methods was mainly attributed to the inclusion of the auxiliary sequences from the MS-CMR images, which provide important label information for the training of deep neural networks. The challenge continues as an ongoing resource, and the gold standard segmentation as well as the MS-CMR images of both the training and test data are available upon registration via its homepage (www.sdspeople.fudan.edu.cn/zhuangxiahai/0/mscmrseg/).

Keywords: Multi-Sequence, Cardiac MRI Segmentation, Benchmark, Challenge

1. Introduction

Myocardial infarction (MI) is a major cause of deaths and disability worldwide (Thygesen et al., 2008). Assess-

URL: zxh@fudan.edu.cn (Xiahai Zhuang^{1*}),
jhxu18@fudan.edu.cn (Jiahang Xu^{1*}), lilei.sky@sjtu.edu.cn
(Lei Li^{1,2*})

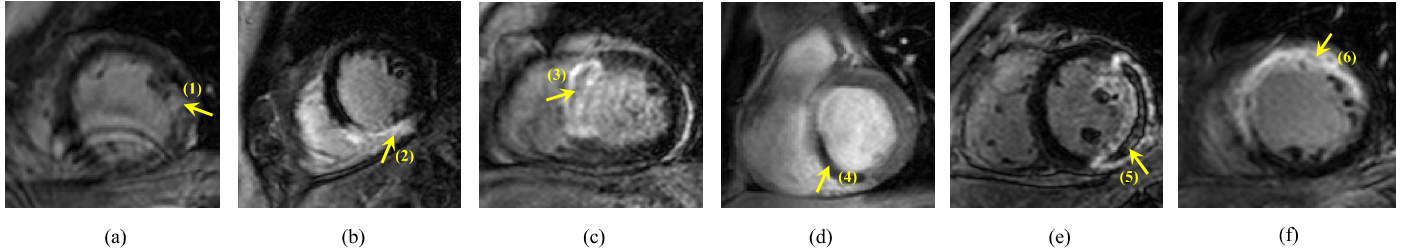


Figure 1: Examples of LGE CMR slices, where scars are pointed out by arrows. One can see the enhancement patterns can be complex, and the scars are diverse in location, shape and size. Note that scars in (3), (4) and (5) include MVO appearing dark inside, and the boundary between blood pools and infarcts can be generally indistinct, particularly in (1) and (3).

ment of myocardial viability is essential in the diagnosis and treatment management for patients suffering from MI. Late gadolinium enhancement (LGE) cardiac magnetic resonance (CMR) sequence can visualize MI, attributed to the slow washout kinetics of the gadolinium in the infarcted areas, which appear with distinctive brightness compared with the healthy tissues. The sensitivity and accuracy of LGE CMR for the assessment of the presence, location, and extent of MI have been well validated (Amado et al., 2004; Wagner et al., 2003).

Before the delineation of MI, accurate segmentation of myocardium is generally required (Flett et al., 2011). However, automating this segmentation remains challenging, due to the large shape variability of heart, indistinguishable boundaries, and the possible poor image quality, as Fig. 1 shows. Particularly, there are three challenges for the automatic analysis of LGE CMR. Firstly, the intensity distribution of the pathological myocardium is heterogeneous. Hence, it is difficult to model the intensity distribution using parametric models, which usually work well for other CMR sequences. Secondly, the enhancement patterns can be complex. The location, shape and size of infarcts vary greatly across different patients, as Fig. 1 shows. Particularly, microvascular obstruction (MVO), which could occur in a high proportion of MI patients, appears particularly dark inside the bright infarct region, due to the lack of uptake of contrast agent. Thirdly, the intensity distributions of the myocardium and its surroundings are conjoining, making it even more difficult to differentiate the boundaries, such as the boundaries between scars and blood pools and the boundaries between healthy myocardium and liver.

To the best of our knowledge, limited work has been done for the fully automatic segmentation from LGE CMR, in contrast to the rich literature of cardiac segmentation on the balanced-Steady State Free Precession (bSSFP) CMR.

1.1. State-of-the-art of LGE CMR segmentation

Fully automated segmentation of LGE CMR could either solely rely on the information of the LGE CMR themselves, or combine the complementary information from other sequences, such as bSSFP and T2 CMR. For the former scheme, shape models or constraints are generally required. Wei et al. (2013) proposed a 1D parametric

model to detect the paired endocardial (Endo) and epicardial (Epi) edge points. To maintain a realistic shape, they imposed a thickness constraint on the segmented myocardium for the 3D deformation. Liu et al. (2017) proposed the multi-component Gaussian mixture model to deal with the intensity heterogeneity of myocardium in the LGE CMR images. To maintain a realistic shape of the resulting segmentation, they further employed a coupled level set to model the paired Endo and Epi surfaces, and regularize the shape of segmented myocardium to resemble a flat donut on 2D slices. Recently, Yue et al. (2019) proposed a deep neural network (DNN) based method to segment LGE CMR, where the shape prior and spatial prior networks, respectively referred to as shape reconstruction neural network and spatial constraint network, were employed to improve the segmentation results.

Integrating the information from MS-CMR sequences, particularly from the same patient, can assist the segmentation of LGE CMR (Zhuang, 2019). This is because the appearance of the myocardium could be clearer in other CMR sequences. For example, both T2-weighted CMR and bSSFP cine sequence present clearer boundaries of myocardium. Fig. 2 gives an example of the three sequences from our challenge. A straightforward solution is to propagate the segmentation of the auxiliary sequence(s), such as bSSFP CMR acquired from the same subject and the same session, directly to the target LGE CMR (Dikici et al., 2004; Lu et al., 2013; Tao et al., 2015). This requires accurate segmentation from bSSFP CMR, generally done by manual work, and a robust registration, to avoid accumulated errors.

Recently, more advanced strategies have been adopted to maintain an accurate and shape-realistic segmentation on LGE CMR. The strategies include (1) transferring the shape prior knowledge from bSSFP CMR to the segmentation of LGE CMR and (2) simultaneous segmentation of MS-CMR images. For transferring shape prior, Liu et al. (2018) built a sparse shape model and target-specific dictionary from a set of bSSFP CMR images. The cardiac segmentation of an unseen LGE CMR image was then sparsely recomposed from the shape dictionary. For the simultaneous segmentation combining MS-CMR, Liu et al. (2014) and Liu et al. (2019) combine the information of LGE and T2 images from the same subject to ob-

Table 1: Summary of the previous challenges related to cardiac segmentation from MICCAI/ ISBI society. LV: left ventricle; Myo: myocardium; RV: right ventricle; LA: left atrium; WHBP: whole heart blood pool; WH: whole heart.

Reference	Year	Data	Target	Pathologies
Radau et al. (2009)	2009	45 bSSFP MRI	LV, Myo	hypertrophy, infarction
Suinesiaputra et al. (2011)	2011	200 bSSFP MRI	LV, Myo	myocardial infarction
Petitjean et al. (2015)	2012	48 bSSFP MRI	RV	congenital heart disease
Karim et al. (2016)	2012	30 LGE MRI	LV scars	myocardial infarction
Karim et al. (2013)	2013	60 LGE MRI	LA scar	atrial fibrillation
Tobon-Gomez et al. (2015)	2013	30 CT, 30 bSSFP MRI	LA	atrial fibrillation
Karim et al. (2018)	2016	10 CT, 10 black-blood MRI	LA wall	atrial fibrillation
Moghari et al. (2016)	2016	20 bSSFP MRI	WHBP, Myo	congenital heart disease
Bernard et al. (2018)	2017	150 bSSFP MRI	LV, Myo, RV	infarction, dilated/ hypertrophic cardiomyopathy, abnormal RV
Zhuang et al. (2019)	2017	60 CT, 60 bSSFP MRI	WH	atrial fibrillation, congenital heart disease, coronary heart disease
Zhao and Xiong (2018)	2018	150 LGE MRI	LA	atrial fibrillation

tain a shape-constrained segmentation result of the LGE CMR. Zhuang (2016, 2019) proposed a multivariate mixture model (MvMM) to combine the information of three-sequence CMR from the same subject. The three-source images were then simultaneously registered and segmented within a unified framework. These works have illustrated the great potential of combining multi-source images for the automatic segmentation of LGE CMR.

1.2. Motivation

For LGE CMR, not only the papers focusing on automated segmentation are scarce, but also the open data are rarely released. Table 1 presents the recent challenges and public datasets for cardiac segmentation. One can see that only Karim et al. (2016), Karim et al. (2013) and Zhao and Xiong (2018) provided the LGE CMR images for benchmark studies, of which two were for left atrial segmentation and one for ventricles. None of them includes multi-source images to assist the segmentation of LGE CMR.

We therefore organized the MS-CMR Segmentation challenge 2019, consisting of 45 MS-CMR images, in conjunction with MICCAI 2019. The challenge particularly provided three-sequence CMR, including the LGE CMR, and was aimed to encourage the development of new segmentation algorithms, as well as validating existing ones. Twenty-three submissions were evaluated before the deadline, and fourteen teams presented their work at the conference event. In this paper, we introduce the related information of the challenge, elaborate on the methodologies of the representative methods, and analyze their results in details.

The rest of this paper is organized as follows. Section 2 provides details of the materials and evaluation framework from the challenge. Section 3 introduces the evaluated methods for benchmarking. Section 4 presents the results, followed by discussions in Section 5. Finally, we conclude this work in Section 6.

2. Materials and setup

2.1. MS-CMR images

MS-CMR segmentation challenge provided image sets from 45 subjects who underwent cardiomyopathy, of which each consists of three CMR sequences, i.e. LGE CMR, bSSFP CMR and T2 CMR (Zhuang, 2016, 2019). The data had been collected with institutional ethics approval and had been anonymized.

The three CMR sequences were all breath-hold, multi-slice, acquired from the cardiac short-axis views. The LGE CMR was a T1-weighted, inversion-recovery, gradient-echo sequence, consisting of 10 to 18 slices and covering the main body of the ventricles. The typical parameters are as follows, TR/TE: 3.6/1.8 ms; slice thickness: 5 mm; in-plane resolution: reconstructed into 0.75×0.75 mm.

The T2 CMR was a T2-weighted, black blood Spectral Presaturation Attenuated Inversion-Recovery (SPAIR) sequence, generally consisting of a small number of slices. For example, among the 35 cases, 13 have only three slices, and the others have five (13 subjects), six (8 subjects) or seven (one subject) slices. The typical parameters are as follows, TR/TE: 2000/90 ms; slice thickness: 12-20 mm; in-plane resolution: reconstructed into 1.35×1.35 mm.

The bSSFP CMR was a balanced-steady state free precession cine sequence. Since both the LGE and T2 CMR were scanned at the end-diastolic phase, the same cardiac phase of the bSSFP cine data was selected for this study. The bSSFP images generally consist of 8 to 12 contiguous slices, covering the full ventricles from the apex to the basal plane of the mitral valve, with some cases having several slices beyond the ventricles. The typical parameters are as follows, TR/TE: 2.7/1.4 ms; slice thickness: 8-13 mm; inplane resolution: reconstructed into 1.25×1.25 mm.

The data were split into two sets, i.e., the training set and the test set, as Table 2 shows. For the training data, both the LGE CMR images and the corresponding gold standard were released to the participants for building, training and cross-validating their models. Besides, the

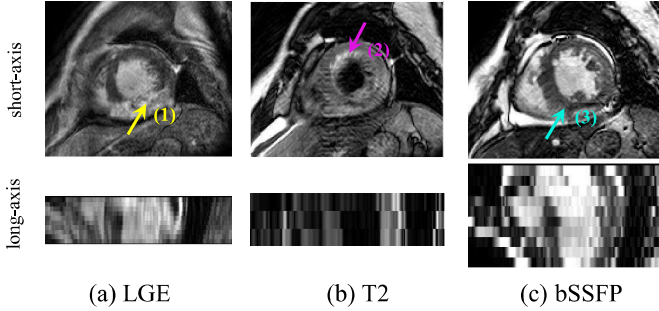


Figure 2: Illustration of multi-sequence CMR images acquired from the same patient: (a) LGE CMR visualizes the scars as brighter texture, in contrast to the dark healthy myocardium, as arrow (1) donates; (b) T2 CMR presents the myocardial edema in higher intensity values, as arrow (2) shows, and can visualize the trabeculations and papillary muscles in great detail; (c) bSSFP CMR provides good boundary information between the myocardium and blood pools which are generally indistinct in LGE CMR containing scars, as arrows (1) and (3) indicate.

challenge provides images and gold standard labels from the other two sequences, to assist the segmentation of LGE CMR. For the test data, only LGE CMR images were released. Once the participants developed their algorithms, they could submit their segmentation results of the test data to the challenge moderators for an independent evaluation. To avoid parameter tuning, the challenge only allowed a maximum of two evaluations for one algorithm. *Note that the MS-CMR data and gold standard segmentation of all images are now available upon request and registration via the challenge homepage¹.*

2.2. Gold standard segmentation and evaluation metrics

Three observers were employed to manually label the left ventricular blood pool (LV), right ventricular blood pool (RV) and left ventricular myocardium (Myo). The observers were well trained raters who were post-graduate students either in biomedical engineering or medical imaging field. The manual labeling was performed slice-by-slice using a brush tool in the software ITK-SNAP (Yushkevich et al., 2006). On average, each of the segmentation took 20 minutes. All the manual segmentation results were validated by senior experts in cardiac anatomy before used in the construction of gold standard segmentation, which was achieved by averaging the three manual delineations using the shape-based approach (Rohlfing and Maurer, 2006).

To evaluate the accuracy of a segmentation result, we employed two widely used metrics, i.e., the Dice score and Hausdorff Distance (HD) (Zhuang, 2013), which are respectively defined as,

$$\text{Dice}(V_{\text{seg}}, V_{\text{GD}}) = \frac{2|V_{\text{seg}} \cap V_{\text{GD}}|}{|V_{\text{seg}}| + |V_{\text{GD}}|}, \quad (1)$$

¹www.sdspeople.fudan.edu.cn/zhuangxiahai/0/mscmrseg19/

Table 2: The distribution of training and test data. I^{lab} : images with gold standard labels; I^{unl} : images without labels.

Dataset	LGE	T2	bSSFP
Training	5 I^{lab}	35 I^{lab} , 10 I^{unl}	35 I^{lab} , 10 I^{unl}
Test	40 I^{unl}	None	None

and

$$\text{HD}(X, Y) = \max \left[\sup_{x \in X} \inf_{y \in Y} d(x, y), \sup_{y \in Y} \inf_{x \in X} d(x, y) \right], \quad (2)$$

where V_{GD} and V_{seq} denote the gold standard and automatic segmentation, respectively; X and Y represent two sets of contour points, and $d(x, y)$ indicates the distance between the two points x and y .

2.3. Participants

As an ongoing event, the challenge had received eighty-seven requests of registration before the submission of this manuscript, among which sixty-five teams participated the event before the date of the workshop (Oct 13th, 2019). Twenty-three submitted results were evaluated before the submission deadline, and nine algorithms were selected for this benchmark work. All these selected teams agreed to include their methods and results for publication in this paper. The selection criterion was based on the novelty of the proposed methodologies and performance of the algorithms. *Note that the team abbreviations in the remaining of this paper refer both to the teams and their corresponding methods.*

3. Evaluated Methods

In this section, we elaborate on the nine benchmarked algorithms. Three of them, i.e., ICL, XMU and INRIA, are based on unsupervised learning, without using labeled LGE images. The other six teams, i.e., SCU, UB, FAU, NVIDIA, HIT and SUSTech, trained their models using 3 to 5 labeled LGE images, and thus are considered as weakly supervised methods. Table 3 summarizes these algorithms and their training data.

3.1. Imperial College London (ICL)

ICL formulates an unsupervised learning algorithm, by decomposing the training process into two stages, i.e. image translation for style transfer from bSSFP to LGE and image segmentation (Chen et al., 2020a).

As Fig. 3 shows, the translation network builds on previous work of multi-modal unsupervised image-to-image translation (UNIT) network (Huang et al., 2018). Here, the representations from shape and appearance of the two modalities are disentangled, swapped and combined to synthesize LGE-stylized images, which are then fed as training samples to the downstream segmentation network. Apart from weighted cross-entropy loss, L_2 distance loss between edge maps is also employed for contour constraint of cardiac structures, especially the myocardium.

Table 3: Summary of the benchmark methods. The abbreviations are as follows, Dim: the segmentation is performed on 2D slices or 3D images; DAug: whether use data augmentation strategy; Img Syn: whether use image synthesis or reconstruction strategy; Pre/ Post: whether use pre-processing or post-processing strategy. For pre-processing, R: image resample or resize; C: image cropping; IN: image intensity normalization.

Teams	Dim	DAug	Img Syn	Network	Pre/ Post	Training data		
						LGE	bSSFP	T2
ICL	2D	Y	Y	UNIT + cascaded U-Net	R, C, IN/ Y	40 I^{unl}	20 I^{lab} , 20 I^{unl}	0
XMU	2D	N	N	Attention U-Net with GFRM	R, C/ N	40 I^{unl}	35 I^{lab}	35 I^{lab}
INRIA	3D	Y	Y	TCL-Net	R, IN/ N	5 I^{unl}	35 I^{lab}	35 I^{lab}
SCU	2D	Y	N	SK-U-net	C, IN/ Y	5 I^{lab}	35 I^{lab}	35 I^{lab}
UB	2D	Y	Y	CycleGAN + U-Net	R, C, IN/ N	5 I^{lab} , 40 I^{unl}	35 I^{lab} , 10 I^{unl}	35 I^{lab}
FAU	2D	Y	N	U-Net	IN, C/ Y	4 I^{lab}	35 I^{lab}	35 I^{lab}
NVIDIA	3D	Y	N	MAS + AH-Net	R/ Y	5 I^{lab} , 40 I^{unl}	35 I^{lab}	35 I^{lab}
HIT	2D	Y	Y	Res-U-Net	R, C/ N	4 I^{lab} , 30 I^{unl}	35 I^{lab}	0
SUSTech	2D	Y	N	U-Net with discriminator	C, IN/ N	3 I^{lab}	35 I^{lab}	35 I^{lab}

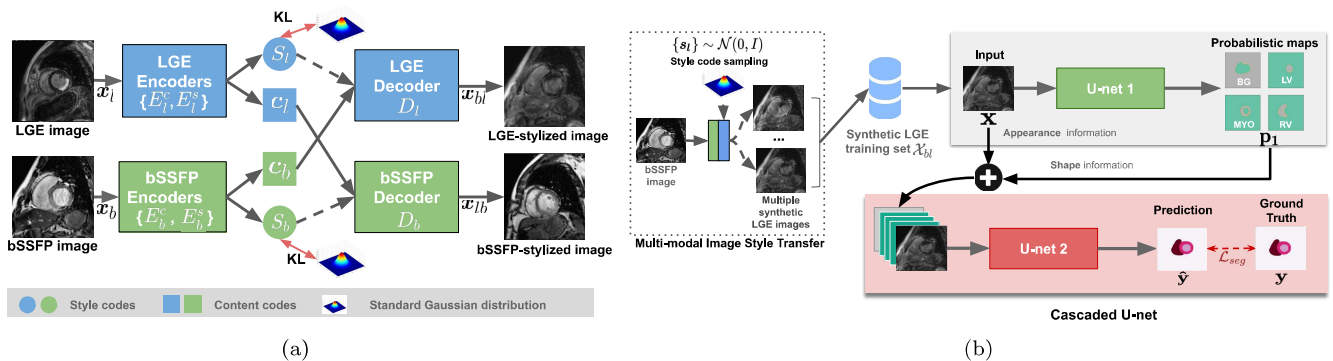


Figure 3: Overview of the two-stage framework proposed by ICL: (a) the multi-modal image UNIT translation network, which consists of two encoder-decoder pairs for the two domains: bSSFP and LGE, respectively; (b) the two-stage cascaded segmentation network.

3.2. Xiamen University (XMU)

XMU focuses on the alignment of features that are extracted from source (bSSFP and T2) and target (LGE) by two parameter-sharing segmentation networks (Wang et al., 2020a). The networks are built upon a baseline model that exploits techniques of pyramid pooling and attention modules to achieve learning of fine-grained features. At the same time, two discriminator networks, minimizing discrepancy in both the feature and output space, are designated for domain adaption from source to target via adversarial training. The networks are optimized based on a hybrid loss function that incorporates cross entropy and Jaccard loss for segmentation in the source domain, and an adversarial loss for feature and mask alignment in the target domain.

Particularly, as Fig. 4 shows, images from the source domain and target domain are fed into a shared modified U-Net for segmentation. The output is the corresponding segmentation. Within the segmentation network, feature maps of different sizes are combined by upsampling and concatenating, which are then sent to the group-wise feature re-calibration module (GFRM), where a feature-level discriminator measures the discrepancy in the feature space. The predicted masks from source and target domains are also compared by a mask-level discriminator as an additional constraint.

3.3. INRIA Sophia Antipolis (INRIA)

INRIA proposes to use style data augmentation to prevent the model from over-fitting to any specific contrast and to focus the optimization on the fundamental geometry features of the target (Ly et al., 2020). To increase the diversity of training images from the source domain (T2 and bSSFP), multiple image processing functions are applied to the source images, including adaptive histogram equalization, Laplacian transformation, Sobel edge detection, intensity inversion and histogram matching towards the target modality (LGE).

The dual U-Net strategy is adopted (Jia et al., 2018), where the two networks are trained independently. These two networks, consecutively applied in the segmentation pipeline, are responsible for localization of the epicardium and refinement of the segmentation, respectively. Additional improvement on the segmentation accuracy is attained with the proposed thresholded connection layer network (TCL-Net), which is able to eliminate non-target pixels and to highlight the target pixels of the input image.

3.4. Sichuan University (SCU)

SCU uses a supervised method and is based on a neural network primarily modified from a 2D U-Net, where an input sample consists of three channel 2D images extracted from three neighbored slices in an image. The segmentation network is equipped with the squeeze-and-excitation

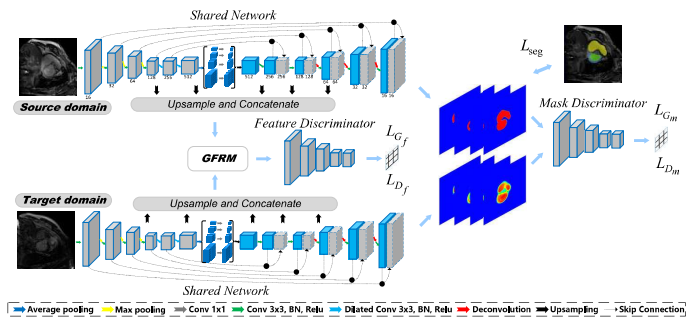


Figure 4: Overview of the multi-modal image translation network proposed by XMU.

residual (SE-Res) module and the selective kernel (SK) module, inserted in the encoding and decoding stages, respectively (Wang et al., 2020b). The SE-Res module adaptively recalibrates channel-wise feature responses by explicitly modelling interdependencies between channels (Hu et al., 2018), while the SK module is supposed to adaptively adjust the size of its receptive field based on multiple scales of input information (Li et al., 2019). A series of techniques, including hole filling and connected component analysis, are employed to post-process the output of the segmentation network.

3.5. University of Barcelona (UB)

UB proposes to address the limited availability of training samples for LGE CMR segmentation by enriching the CNN models using two complementary methods (Campello et al., 2020). First, an image-to-image translation generative adversarial network is trained to convert bSSFP images into LGE-like ones using the CycleGAN strategy (Zhu et al., 2017). Second, an LGE-specific data argumentation scheme, i.e., partially region rotation of scars, is proposed based on the provided labeled LGE images, as shown in Fig. 5. The image synthesis and scar tissue augmentation strategies can account for both the global appearance of LGE image and the local appearance of scar tissues. Based on the original sequences and synthetic LGE images from image translation, a modified U-Net network is then trained for LGE segmentation by integrating a deep supervision term in the upsampling path of the U-Net.

3.6. Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU)

FAU employs a transfer learning strategy, by training an encoder-decoder segmentation network using labeled T2 and bSSFP images, and then fine-tuning the network with the labeled LGE images for supervised training (Vesal et al., 2020). The encoder-decoder network, called Dilated-Residual U-Net (Vesal et al., 2018), can capture both global and local context information using dilated convolutional layers in the bottleneck. Moreover, before

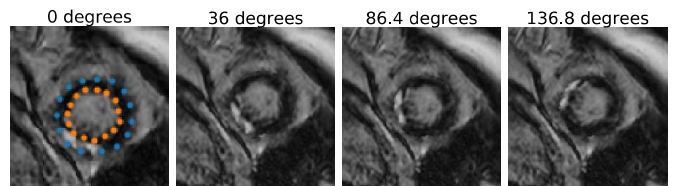


Figure 5: LGE-specific data argumentation strategy: examples of three rotations of myocardium by landmarks shown in the first column, proposed by UB.

the images are fed to the network, the image contrast is enhanced slice-by-slice, using contrast limited adaptive histogram equalization. Besides, connected component analysis as a post-processing step is employed to separately remove the small misclassified areas in the softmax layer output for Myo, LV, and RV.

3.7. NVIDIA

NVIDIA combines classical methods of multi-atlas label fusion with deep learning by creating the so-called noisy labels for unlabeled LGE images using registration techniques (Roth et al., 2020).

The noisy labels are derived from two sets of registrations. One is the inter-patient and intra-modality registration from the annotated 5 LGE images to the remaining 40 ones. The other is the intra-patient and inter-modality registration from the corresponding annotated bSSFP/T2 to the 40 unlabeled LGE images. The registration process contains affine transformation followed by free-form deformation in a coarse-to-fine manner, where normalized mutual information is employed as the similarity measure. For deformation regularization, the bending energy on the deformation fields is used to encourage smoothness and a penalty term to encourage inverse consistency is also included.

Multi-atlas label fusion is conducted before the fused pseudo labels are used to supervise an ensemble of segmentation networks. The networks are constructed using convolutional kernels from the pretrained *ImageNet* weights. Besides, the weights are initialized in three orthogonal planes to encourage view discrepancy.

The final prediction is achieved by fusing probability maps from each model using a median operator followed by a post-processing step of 3D connected component analysis to reduce outliers.

3.8. Harbin Institute of Technology (HIT)

HIT adopts an intensity histogram match technology to achieve data augmentation. The method is developed as follows: First, the provided bSSFP images, with gold standard segmentation, are intensity-translated to generate fake LGE CMR images, by referring to the real LGE CMR and using histogram matching techniques. The resulting LGE images have the same shape of the heart, as the original bSSFP data, and thus the gold standard segmentation of the bSSFP can be used for the fake LGE

Table 4: Results of the evaluated algorithms on LGE CMR segmentation. Teams updating their results after the challenge deadline are indicated with an asterisk (*), and teams using the unlabeled LGE images (I^{unl}) for training are indicated with a dag (†).

Teams	Volumetric Dice			Volumetric HD (mm)			Training
	Myo	LV	RV	LV Endo	LV Epi	RV Endo	
ICL†	0.826 ± 0.035	0.919 ± 0.026	0.875 ± 0.050	10.28 ± 3.376	12.45 ± 3.142	15.38 ± 6.942	Unsupervised
XMU†	0.796 ± 0.059	0.896 ± 0.047	0.846 ± 0.086	13.59 ± 5.206	15.70 ± 5.814	15.21 ± 6.327	Unsupervised
INRIA*	0.705 ± 0.115	0.870 ± 0.051	0.762 ± 0.150	41.74 ± 7.696	42.79 ± 13.26	34.38 ± 8.065	Unsupervised
SCU*	0.843 ± 0.048	0.926 ± 0.028	0.890 ± 0.044	9.748 ± 3.280	11.65 ± 4.002	13.34 ± 4.615	Supervised
UB†	0.810 ± 0.061	0.898 ± 0.045	0.866 ± 0.050	10.78 ± 4.066	11.96 ± 3.620	15.91 ± 6.895	Supervised
FAU	0.789 ± 0.073	0.912 ± 0.034	0.833 ± 0.084	11.29 ± 4.559	12.54 ± 3.379	17.11 ± 6.141	Supervised
NVIDIA†	0.780 ± 0.047	0.890 ± 0.043	0.844 ± 0.063	11.58 ± 7.524	16.25 ± 6.336	18.12 ± 9.262	Supervised
HIT†	0.751 ± 0.119	0.884 ± 0.070	0.791 ± 0.165	14.30 ± 8.170	14.75 ± 7.823	17.87 ± 9.322	Supervised
SUSTech	0.610 ± 0.102	0.824 ± 0.068	0.710 ± 0.135	23.69 ± 14.66	24.62 ± 12.66	23.46 ± 7.596	Supervised
Average	0.775 ± 0.093	0.895 ± 0.047	0.828 ± 0.114	21.87 ± 15.23	23.65 ± 16.07	21.66 ± 11.49	Unsupervised
	0.764 ± 0.109	0.889 ± 0.060	0.822 ± 0.117	13.56 ± 9.316	15.30 ± 8.389	17.64 ± 8.092	Supervised
	0.766 ± 0.104	0.891 ± 0.056	0.822 ± 0.116	16.37 ± 12.27	18.06 ± 12.18	19.35 ± 9.587	All
Inter-Ob	0.764 ± 0.069	0.881 ± 0.064	0.816 ± 0.084	12.03 ± 4.443	14.32 ± 5.164	21.53 ± 9.460	Inter-Ob

Table 5: Details on the average run time and computer systems used for the evaluated methods. T: average run time in seconds (including pre- and post-processing of images).

Teams	T(s)	GPU	CPU and RAM	Platform
ICL	0.92	GeForce RTX 2080Ti; 11GB	Intel Core i9-9900K; 64GB (segmentation) Intel Core i7 8700; 32GB (translation)	Pytorch
XMU	15	GeForce GTX 1080Ti; 12GB	Intel i5-7640X; 16GB	Pytorch
INRIA	0.958	GeForce GTX 1080Ti; 12GB	Intel Xeon Silver 4110; 8GB	Keras
SCU	12	Tesla P40; 24GB	Intel(R) Xeon(R) E5-2680; 128GB	Pytorch
UB	6.3	GeForce 1080; 8GB	Intel Core i7-8700X; 32GB	Tensorflow/ Pytorch
FAU	2.5	Titan X-Pascal; 12GB;	Intel(R) Xeon(R) 4114; 32GB	TensorFlow/ Keras
NVIDIA	6.64	Tesla V100; 16GB (training) Titan Xp; 12GB (testing)	Intel(R) Core(TM) i7-7800X; 32GB	TensorFlow
HIT	10	GeForce RTX 2080Ti; 12GB	Intel(R) Xeon(R) Silver 4114; 32GB	Keras
SUSTech	60	GTX Titan X; 12GB	Intel Core i7-6700; 64GB	TensorFlow/ Keras

images. Then, the real labeled LGE CMR images and the fake LGE with gold standard are used to train a ResUNet in the supervised learning fashion (He et al., 2016). To improve the segmentation accuracy, the output of the segmentation network is further post-processed using the label-vote strategy and connected component analysis (Liu et al., 2020).

3.9. Southern University of Science and Technology (SUSTech)

SUSTech proposes an adversarial segmentation framework, where a discriminator is used to drive the generator to produce good-quality masks similar to the ground truth segmentation (Chen et al., 2020b). Specifically, the method transfers the label masks of bSSFP and T2 CMR images to LGE CMR by using a normalized index which identifies the correspondence between axial slices from different modalities. Then, the generator model, formed by a segmentation network, is trained to output realistic LGE labels with the combination of supervised segmentation loss and discriminator loss.

4. Results

4.1. Performance of the evaluated methods

Table 4 presents the quantitative results of the nine evaluated algorithms on the LGE CMR segmentation, where the inter-observer variations over the manual delineation are given. The mean volumetric Dice scores of the evaluated algorithms are respectively 0.766 ± 0.104 (Myo), 0.891 ± 0.056 (LV), and 0.822 ± 0.116 (RV), and the mean volumetric HD values are respectively 16.37 ± 12.27 mm (LV Endo), 18.06 ± 12.18 mm (LV Epi), and 19.35 ± 9.587 mm (RV Endo). The mean Dice scores are comparable to that of the inter-observer Dice scores. Compared with inter-observer HD values, the average HD is however evidently worse, though the top-ranking methods could perform much better. Interestingly, the unsupervised methods performed comparably to supervised ones, particularly if we exclude INRIA and SUSTech, each respectively from the two categories.

Fig. 6 visualizes the segmentation results of case 20 and 45, which are the median (case 20) and worst (case 45) cases in terms of average Dice of the nine methods. Note that the presented slices are the middle one of these two cases. Most of the evaluated algorithms could segment case 20 well, though some performed bad on RV segmentation. By contrast, case 45 was much more challenging.

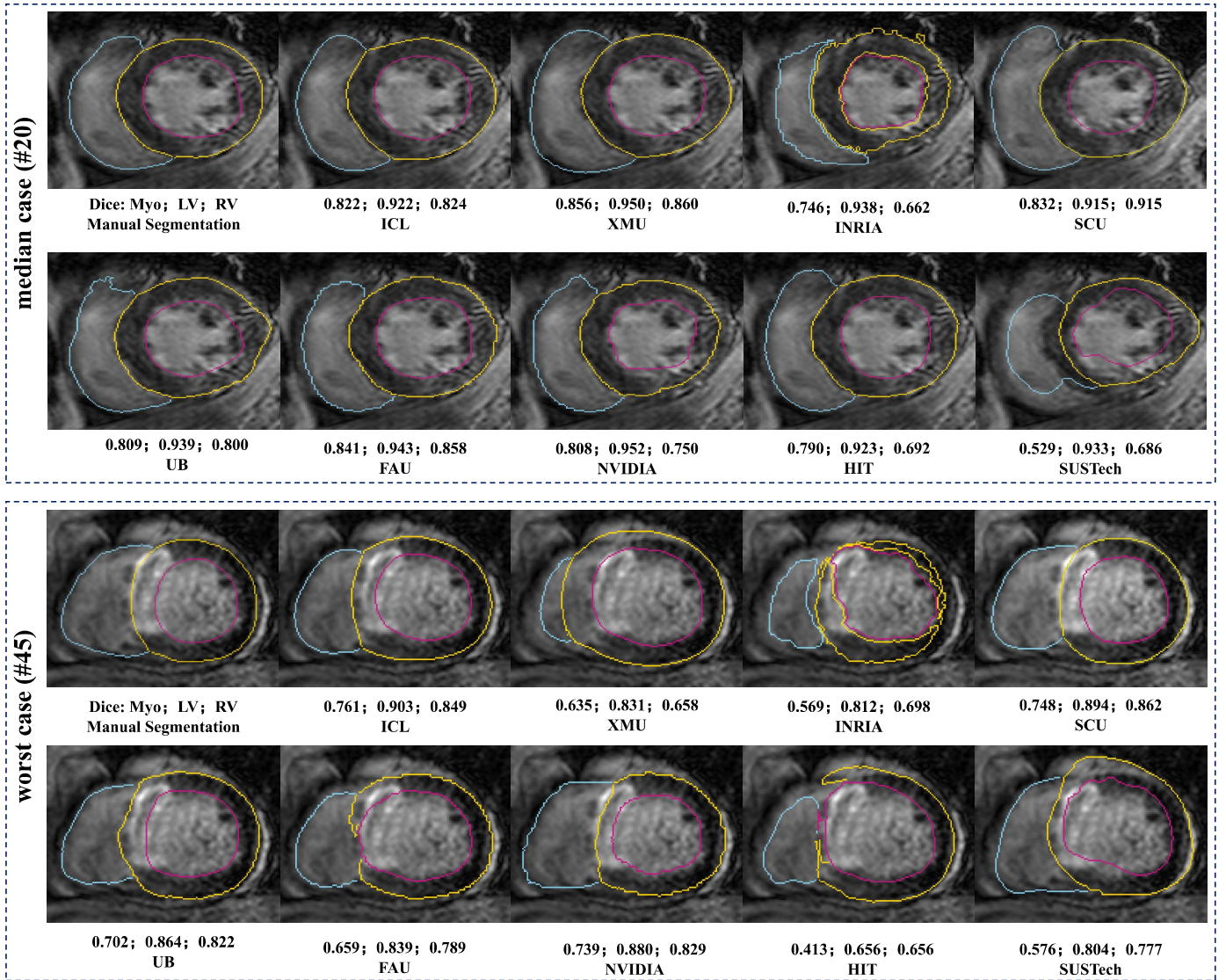


Figure 6: 2D visualization of the segmentation results of the median and worst cases by the nine evaluated methods. The median and worst cases were from the test set in terms of mean Dice of Myo by the nine methods. The contours of LV Endo, LV Epi and RV are respectively in colors of magenta, yellow and blue.

Over one-fourth of the myocardium in the visualized slice was scarring tissues, which could result in failure for most of the methods due to their similar intensity distribution to the blood pool. One can see that seven out of the nine methods misclassified a great portion of the scars into blood pool.

Table 5 summarizes the information of implementation platforms and average run time of the evaluated methods. Most of them can complete a segmentation case within seconds on average for those deep neural network-based methods, thanks to the parallel computing of GPU. Note that the run time includes the pre- and post-processing of the methods, and it may not indicate the true computation complexity of the algorithms due to the different implementation platforms and hardware.

4.2. Accuracies of different substructures

From Table 4, one can see that the Dice scores of LV are evidently better than that of RV and Myo. The HD values of myocardium, including LV Endo and Epi, are generally better than that of RV Endo. This indicates the different levels of challenges for different substructures, namely RV presents more difficulties. This could be owing to the fact that LV has a relatively regular shape, in contrast to RV which has large shape variations as well as poor contrast. This is confirmed by the inter-observer variation study, where the mean HD between different raters could reach over 20 mm and mean Dice only 0.816 for RV, due to the challenge of achieving consistent segmentation on certain areas. Also, Fig. 6 confirms that the results of RV segmentation by different methods could vary more than the other substructures.

Table 6: Slice-wise accuracies of the apical (Apex), mid-ventricular (Mid), and basal (Base) slices.

Teams	Slice-wise Dice			Slice-wise HD (mm)			Training
	Apex	Middle	Base	Apex	Middle	Base	
ICL [†]	0.706 ± 0.203	0.883 ± 0.078	0.857 ± 0.140	11.24 ± 10.73	6.140 ± 5.144	8.229 ± 7.306	Unsupervised
XMU [†]	0.502 ± 0.292	0.844 ± 0.136	0.828 ± 0.165	20.76 ± 17.64	7.604 ± 8.775	8.992 ± 7.758	Unsupervised
INRIA*	0.612 ± 0.230	0.795 ± 0.142	0.724 ± 0.221	14.45 ± 15.52	12.74 ± 13.57	16.99 ± 15.93	Unsupervised
SCU*	0.705 ± 0.203	0.895 ± 0.076	0.894 ± 0.072	10.12 ± 9.639	5.300 ± 4.823	6.304 ± 3.875	Supervised
UB [†]	0.694 ± 0.218	0.855 ± 0.103	0.822 ± 0.170	11.45 ± 11.30	7.331 ± 6.670	10.45 ± 10.68	Supervised
FAU	0.622 ± 0.268	0.831 ± 0.148	0.864 ± 0.109	17.35 ± 15.87	9.753 ± 10.51	8.572 ± 6.966	Supervised
NVIDIA [†]	0.727 ± 0.158	0.852 ± 0.085	0.821 ± 0.110	10.64 ± 8.371	8.775 ± 9.029	11.09 ± 8.482	Supervised
HIT [†]	0.582 ± 0.275	0.807 ± 0.179	0.771 ± 0.240	17.81 ± 15.90	9.864 ± 11.35	14.66 ± 16.00	Supervised
SUSTech	0.594 ± 0.202	0.734 ± 0.155	0.736 ± 0.150	14.96 ± 9.106	12.54 ± 9.612	16.30 ± 11.24	Supervised
Average	0.607 ± 0.258	0.841 ± 0.127	0.803 ± 0.187	15.49 ± 15.43	8.829 ± 10.19	11.40 ± 11.75	Unsupervised
	0.654 ± 0.231	0.829 ± 0.140	0.818 ± 0.161	13.72 ± 12.50	8.927 ± 9.231	11.23 ± 10.81	Supervised
	0.638 ± 0.242	0.833 ± 0.136	0.813 ± 0.170	14.31 ± 13.57	8.894 ± 9.563	11.29 ± 11.13	All
Inter-Ob	0.789 ± 0.102	0.858 ± 0.074	0.851 ± 0.089	8.975 ± 7.511	7.444 ± 5.076	14.35 ± 13.38	Inter-Ob

4.3. Slice-wise accuracies from different positions

Table 6 presents the slice-wise Dice and HD values for the slices from different positions of the ventricles, including the apical slices (denoted as Apex), mid-ventricular slices (as Middle) and basal slices (as Base). Both of the Dice and HD metrics were computed by averaging all the numbers from the three substructures, i.e., Myo, LV, RV for Dice, and LV Endo, LV Epi and RV Endo for HD. One can see that the different methods could have different superiority/ inferiority regarding segmentation for the slices from different positions. For example, XMU and FAU had particularly poor Dice and HD for apical slices, though they performed well for the Middle and Base; UB achieved superior HD and Dice on Apex and Middle compared with most of the algorithms, but its segmentation on Base was evidently poor.

Table 6 also provides the inter-observer variations. The inter-observer Dice and HD values of Apex are evidently poorer than that of Middle and Base; but for HD the worst slice comes from the Base. Both of the two metrics show good values for middle slices, indicating less difficulties and more reliability for manual delineation, which is consistent with the automatic segmentation.

Compared to the inter-observer variations, none of the evaluated methods could achieve satisfactory accuracies on Apex. In fact, their mean Dice or HD values are all much worse. This is unlike the results in Table 4 where majority of the evaluated methods could achieve better Dice and HD values than the inter-observer variations. Such poor performance of the automatic segmentation on Apex is well illustrated in Fig. 7, which colorizes the mean accuracies of the three substructures evaluated on the nine methods, for all slices of the test dataset. One can see from the figure that the apical slices are generally with much worse accuracies, particularly for segmentation of LV Myo and RV.

4.4. Myocardial segmentation in presence of scars

Besides the inferior image quality of LGE CMR compared to the bSSFP CMR, the inhomogeneous intensity

of myocardium with the presence of enhanced scars could particularly challenge an automatic segmentation algorithm. This was illustrated by Case 45, Fig. 6, where the scar area resulted in the failure of most of the evaluated algorithms. In this study, we investigate the correlation of myocardial segmentation accuracies with the presence and amount of scars in the images.

Fig. 8 (a) visualizes the scar percentage for each slice of the test images, i.e., the area percentage computed as the ratio of scar area to the total area of myocardium in a slice. One can compare this figure to the visualizations in Fig. 7, to study the correlation between slice positions and scar percentage, which nevertheless is not evident due to the complex patterns of scars. We further plotted the scatter points describing the accuracies of myocardial segmentation against the scar percentage for each slice containing myocardial infarcts, in Fig. 8 (b) and (c). One can see that the Dice accuracies of myocardial segmentation and scar percentage are generally negatively correlated, though this correlation is very weak for basal slices. This is reasonable as the scar percentage of basal slices is generally small, and the challenge of segmentation in Base could be mainly attributed to the other factors such as irregular shape of the ventricles. The HD errors in Fig. 8 (c) illustrate positive correlation to scar percentage, which is similar to inverse Dice metric. However, the correlations of HD are generally weaker than that of Dice.

5. Discussion

5.1. Supervised versus unsupervised learning

An interesting observation from this challenge is that the average performance of the supervised methods was not evidently better than that of the unsupervised ones, particularly when the outlier approaches were excluded for comparisons. One may argue that this was due to the fact that ICL and XMU, the two well-performing unsupervised approaches, employed the 40 unlabeled test images as target images in domain adaptation, for the training of their target segmentation networks. By contrast, INRIA solely

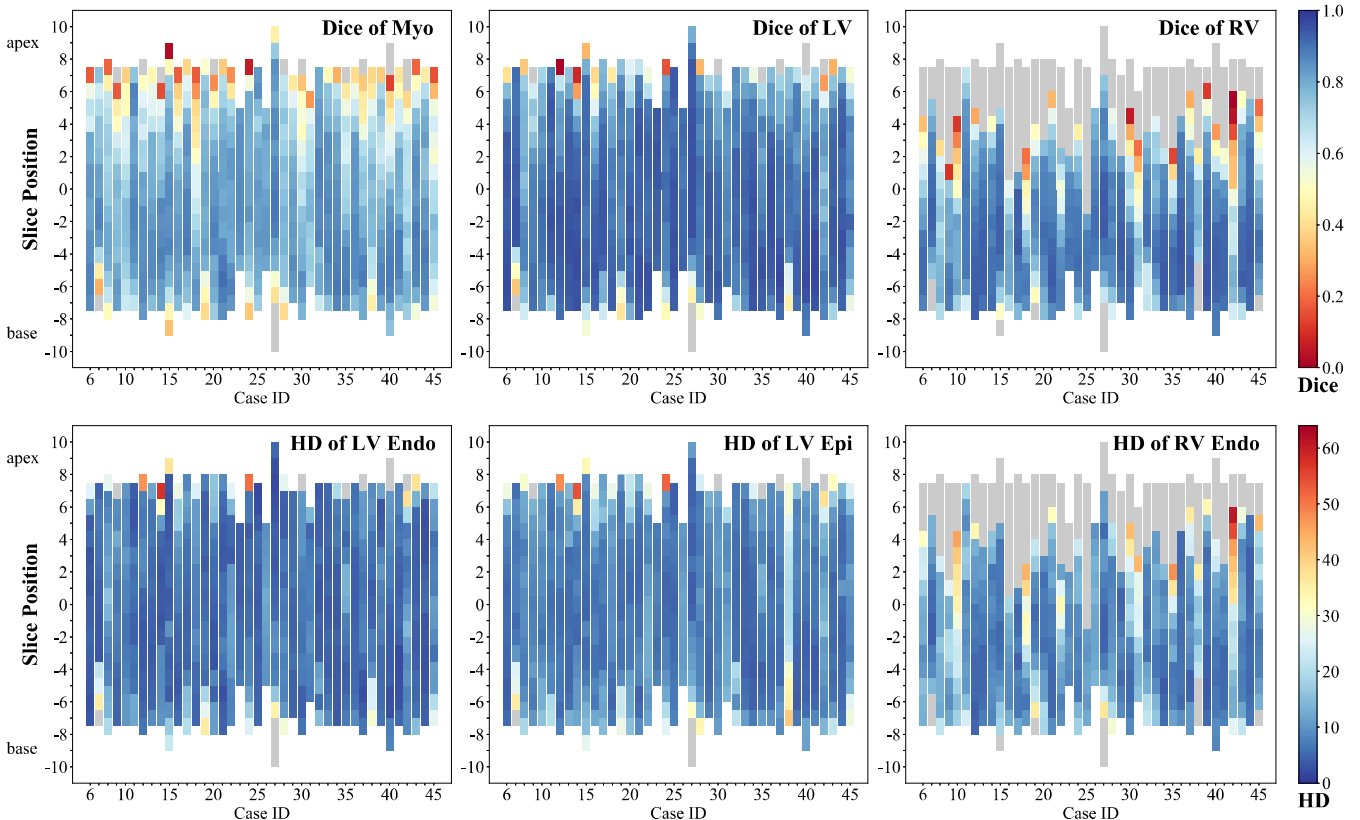


Figure 7: The average Dice and HD values of the nine evaluated methods on each slice of the 40 test LGE CMR images. Note that grey color indicates that the slices do not contain the manual segmentation of the substructure, and thus no accuracy metric is computed.

used the 5 unlabeled LGE CMR images from the training set, and its accuracies were much poorer. However, three of the six supervised methods, i.e., UB, NVIDIA and HIT, also used the unlabeled test images for self-learning, and one can see from Table 4 that their average segmentation accuracies were not better than that of the other three (SCU, FAU and SUSTech), nor better than ICL or XMU.

The other reason could be attributed to the fact that the image appearance of LGE CMR and bSSFP CMR is similar, and thus the domain shift between them could be small for the unsupervised domain adaptation methods when they considered the LGE CMR as target domain and used the labeled bSSFP CMR as the source domain. Particularly, we realized that for the supervised methods their models were all first trained by bSSFP (and T2) CMR and then fine-tuned using the 3 to 5 labeled LGE CMR images. Considering the small domain shift between the two sequences, we conclude that the major performance difference between the supervised and unsupervised approaches could be mainly caused by the 3-5 extra labeled training data on top of the 20-35 ones, more than by the difference of methodologies.

5.2. Segmentation/ training with multi-sequence CMR

Previous discussion indicates the other CMR sequences, particularly the bSSFP CMR, could be critical to the success of LGE CMR segmentation. For the usage of multi-

sequence CMR, three schemes were adopted: (1) propagate the labels from the other sequences to the target LGE CMR; (2) directly train the segmentation models using the labeled (or/ and unlabeled) data from the other sequences; (3) make use of the complementary information of the paired multi-sequence CMR of the same subject for segmentation of LGE images.

Among the evaluated methods, only NVIDIA adopted scheme (1), SCU and FAU used scheme (2), and remained teams employed scheme (3). Specifically, NVIDIA utilized registration-based methods to propagate the label of other sequences to the LGE CMR. Though the registration between LGE and other sequences can be challenging, it is effective to obtain results with realistic shapes. The readers are referred to the segmentation results of Case 45 in Fig. 6, where NVIDIA maintained a reasonably accurate result with realistic shape for this particular challenging case. SCU and FAU directly trained their segmentation networks using all the labeled data from three sequences. HIT utilized bSSFP CMR labeled images to train a localization network, which can offer an ROI as prior knowledge for LGE segmentation. XMU, INRIA and HIT adopted histogram matching, by mapping intensity of bSSFP and T2 CMR images into that of LGE CMR, to generate fake LGE images for supervised training of their models. ICL and UB used a style transfer network to synthesize LGE images, and XMU and SUSTech generated pseudo-LGE

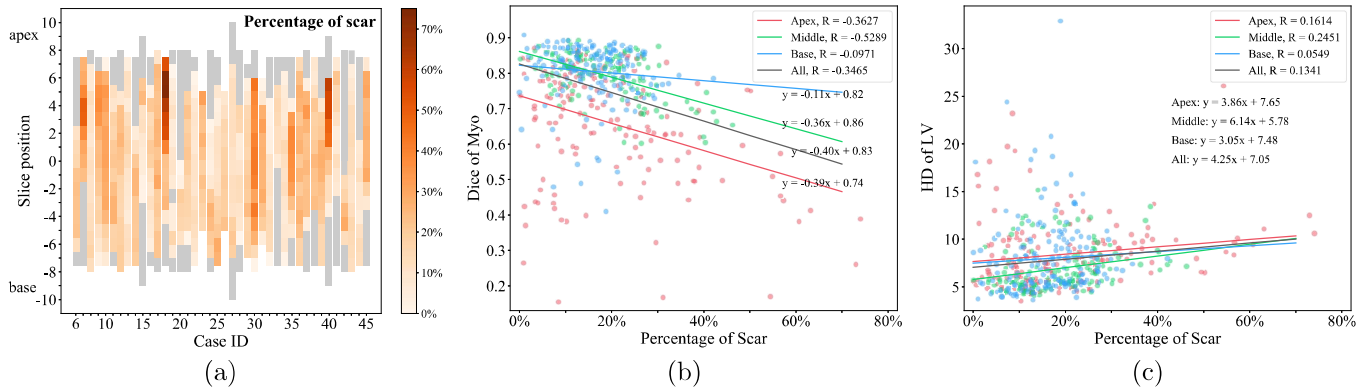


Figure 8: Results of scar presence study: (a) The color map of scar percentage (to myocardium) in each slice of the test images, where grey color indicates no scar is presented. (b) The correlation between the Dice of Myo and the scar percentage for slices from different positions. (c) The correlation between the HD of LV (mean of HD of LV Endo and Epi) and the scar percentage. Apex, Mid, and Base respectively represent the apical, mid-ventricular and basal slices, and All indicates all slices are considered, whose points were omitted in (b) and (c).

labels from other sequences. One can see that the paired corresponding information among different sequences can be learned implicitly by directly adopting other sequences of the same patient as extra training data. To generate pseudo-LGE images/ labels, style transfer demonstrated to be more effective than the conventional methods such as histogram matching.

5.3. Discussion of the literature

In the preceding discussion, we summarized the three usages of multi-sequence CMR for LGE image segmentation. In the literature, the development of automated segmentation on LGE CMR follows the similar three schemes:

- (1) Propagate the prior labels from another sequence (typically bSSFP CMR) from the same patient to the target LGE CMR (Wei et al., 2013; Lu et al., 2013; Tao et al., 2015).
- (2) Directly apply the segmentation method on the LGE CMR images (Liu et al., 2017, 2018; Yue et al., 2019). Images from other CMR sequences, commonly not paired images from the same subject of the target LGE CMR, could be used to assist the segmentation or training of the neural networks.
- (3) Combine the complementary information of the paired multi-sequence CMR of the same subject for segmentation of LGE CMR (Liu et al., 2014, 2019; Zhuang, 2019).

Table 7 summarizes the works of these three categories from recent literature.

One can see that most of the reported works adopted multi-sequence CMR for the segmentation of LGE CMR, except the two works from Liu et al. (2017); Yue et al. (2019). For the auxiliary sequences, images from bSSFP CMR are generally useful. This is because bSSFP images have similar appearance to LGE. This is particularly useful to the first category of methods, where the prior labels are propagated to LGE CMR via image registration. Another reason is the wide availability of the bSSFP images,

which can be used for pre-training or for shape regularization of a deep neural network-based algorithm. Finally, for the combined segmentation using the paired images from the same patient, both the T2 and bSSFP could be employed. For this category of methods, the registration between LGE and bSSFP/ T2 needs to be simultaneously implemented with the procedure of segmentation, to maintain a consistent segmentation among all sequences of the same patient (Zhuang, 2019).

For accuracy, all the reported works of the prior label propagation methods achieved over 0.80 Dice accuracy on myocardium segmentation. For the other methods, only the recent work by (Yue et al., 2019) obtained over 0.80 Dice of Myo. The label propagation methods require prior segmentation of the bSSFP images, which are generally manually performed. Hence, they are not considered as fully automatic schemes. Furthermore, considering the ratio of the Myo Dice to the Myo inter-observer Dice, one can see that neither of the two reported label propagation methods achieved a ratio over 1. By contrast, the ratio of Yue et al. (2019), i.e., 0.812 versus 0.757, is much better, and is comparable to the top-ranking methods in this benchmark study. Finally, the combined segmentation incorporating multi-sequence CMR reported relatively lower accuracies. However, these methods did not require any manual segmentation or labeled CMR images for training (Zhuang, 2019). In the future, we expect more research on the topic of the combined segmentation embedded with deep neural networks for automatic segmentation of LGE CMR.

6. Conclusion

Cardiac segmentation from LGE CMR is important in clinics for diagnosis and treatment management of patients. However, fully automatic segmentation is still arduous. Particularly, the training data with manual labels are limited in the research community of medical image analysis. This manuscript describes the results from the MS-CMR segmentation challenge, which provides 45 sets

Table 7: Results of *LGE* CMR segmentation from the literature. InterOb: inter-observer variations; min: minutes; sec: seconds.

Reference	Data	Methodologies	Runtime	Results of LGE seg in Dice and HD (mm)
Wei et al. (2013)	21 (LGE+ bSSFP)	Propagate prior segmentation in bSSFP to LGE via 2D translation	N/A	Myo: 0.828 ± 0.025 LV: 0.948 ± 0.009 InterOb-Myo: 0.835 ± 0.025 InterOb-LV: 0.951 ± 0.009
Lu et al. (2013)	10 (LGE+ bSSFP)	Propagate segmentation in bSSFP to LGE via deformable registration	N/A	Myo: 0.85 ± 0.05 LV: 0.95 ± 0.02
Tao et al. (2015)	50 (LGE+ bSSFP)	Propagate segmentation in bSSFP to LGE via contour fitting refined with scar pattern	N/A	Myo: 0.805 ± 0.083 InterOb-Myo: 0.81 ± 0.06
Liu et al. (2017)	33 LGE	Coupled level set and multi-component Gaussian mixture model	7.29 min	Myo: 0.736 ± 0.056 LV: 0.905 ± 0.032 InterOb-Myo: 0.731 ± 0.081 InterOb-LV: 0.886 ± 0.051
Liu et al. (2018)	30 LGE, 51 bSSFP	Guided random walks and sparse shape representation from manual segmentation of bSSFP	9 min	Myo: 0.746 ± 0.078 LV: 0.851 ± 0.058 InterOb-Myo: 0.739 ± 0.051
Yue et al. (2019)	45 LGE	Deep neural network with shape reconstruction trained from segmentation labels and spatial constraints	N/A	Myo: 0.812 ± 0.105 , Epi: 11.04 ± 5.818 mm LV: 0.915 ± 0.052 , 12.25 ± 6.455 mm RV: 0.882 ± 0.084 , 18.07 ± 14.17 mm InterOb-Myo: 0.757 ± 0.083
Liu et al. (2014)	6 (LGE+T2)	Combined segmentation of LGE and T2 based on multi-component Gaussian mixture model	N/A	Myo: 0.643 ± 0.084
Liu et al. (2019)	32 (LGE+T2)	Combined segmentation of LGE and T2 with cross constraint and discrepancy compensation of shape	15.2 sec	Myo: 0.781 ± 0.062 LV: 0.879 ± 0.044
Zhuang (2019)	35 (LGE+ T2+bSSFP)	Unified segmentation of three-sequence CMR based on MvMM	9.17 min	Myo: 0.717 ± 0.076 , Epi: 11.2 ± 4.05 mm LV: 0.866 ± 0.063 , 10.6 ± 4.67 mm InterOb-Myo: 0.757 ± 0.083 , Epi: 12.5 ± 5.83 mm InterOb-LV: 0.876 ± 0.069 , 10.6 ± 4.65 mm

of MS-CMR images. Nine representative methods were selected for evaluation and comparisons, and their methodologies and segmentation performance were then analyzed and discussed.

The challenge originally only provided five labeled LGE CMR images for training and validation, and forty for test, to resemble the reality that only very limited labeled training images of LGE CMR are available. In addition, the challenge provided the paired bSSFP and T2 CMR images for all the LGE CMR images, and a much larger number of bSSFP and T2 images were with gold standard labels for training. The evaluation was performed by the organizers, blinded to the participants for a fair comparison in the challenge. Note that the gold standard segmentation of the test data as well as the training data had been open to researchers after the workshop of the challenge.

As a result, one third of the evaluated methods adopted the domain adaptation strategy, meaning none of the target domain images, i.e., LGE CMR, has label information in the training stage. The label information came from

the source domain, i.e., the bSSFP CMR images with gold standard segmentation. The other two thirds of methods employed the labeled bSSFP (and T2 images) for pre-training and then fine-tuned the networks with the 3 to 5 LGE CMR images with gold standard. Interestingly, we did not find evident difference of segmentation performance between these two groups of methods. We concluded this was probably due to the fact that the image appearance of bSSFP CMR and LGE CMR is similar, thus the domain shift is marginal for the domain adaptation-based algorithms.

All the evaluated methods made full use of the MS-CMR images for the development of automatic segmentation of LGE CMR. This is similar to the development of LGE CMR segmentation from the literature, where most of the reported works used bSSFP and (or) T2 CMR to assist the segmentation of LGE CMR or facilitate the training of a LGE CMR segmentation neural network. We concluded that the auxiliary sequences from MS-CMR could be helpful and critical for a successful and robust segmen-

tation of LGE CMR.

Finally, several works in the literature employed the paired MS-CMR for combined and simultaneous segmentation of all the CMR sequences, including LGE CMR. However, none of the evaluated methods in this challenge adopted this strategy. We concluded this should be attributed to the fact that consistent segmentation among MS-CMR requires simultaneous registration between sequences in the procedure of combined segmentation, which can be still arduous for deep neural network-based algorithms. We therefore expect more research on novel methodologies to achieve combined computing with simultaneous registration and segmentation of the multi-source images in the future.

Authors Contributions

XZ initialized the challenge and provided all the resources; XZ, LL, JX and XL collected the materials and composed the manuscript. CC, CO, DR, VC, KL, SV, NR, YL, GL, JC, HL, BL, MS, HR, WZ, JW, XD, XW, SY were participants of the MS-CMR segmentation challenge. The participants provided the description of their algorithms and segmentation results for evaluation, and contributed equally to this paper. All the authors have read and approved the publication of this work.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (61971142).

References

- Amado, L.C., Gerber, B.L., Gupta, S.N., Rettmann, D.W., Szarf, G., Schock, R., Nasir, K., Kraitchman, D.L., Lima, J.A., 2004. Accurate and objective infarct sizing by contrast-enhanced magnetic resonance imaging in a canine myocardial infarction model. *Journal of the American College of Cardiology* 44, 2383–2389.
- Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.A., Cetin, I., Lekadir, K., Camara, O., Ballester, M.A.G., et al., 2018. Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE transactions on medical imaging* 37, 2514–2525.
- Campello, V.M., Martín-Isla, C., Izquierdo, C., Petersen, S.E., Ballester, M.A.G., Lekadir, K., 2020. Combining multi-sequence and synthetic images for improved segmentation of late gadolinium enhancement cardiac MRI, in: Pop, M., Sermesant, M., Camara, O., Zhuang, X., Li, S., Young, A., Mansi, T., Suinesiaputra, A. (Eds.), *Statistical Atlases and Computational Models of the Heart. Multi-Sequence CMR Segmentation, CRT-EPiggy and LV Full Quantification Challenges*, Springer International Publishing, Cham. pp. 290–299.
- Chen, C., Ouyang, C., Tarroni, G., Schlemper, J., Qiu, H., Bai, W., Rueckert, D., 2020a. Unsupervised multi-modal style transfer for cardiac mr segmentation, in: Pop, M., Sermesant, M., Camara, O., Zhuang, X., Li, S., Young, A., Mansi, T., Suinesiaputra, A. (Eds.), *Statistical Atlases and Computational Models of the Heart. Multi-Sequence CMR Segmentation, CRT-EPiggy and LV Full Quantification Challenges*, Springer International Publishing, Cham. pp. 209–219.
- Chen, J., Li, H., Zhang, J., Menze, B., 2020b. Adversarial convolutional networks with weak domain-transfer for multi-sequence cardiac mr images segmentation, in: Pop, M., Sermesant, M., Camara, O., Zhuang, X., Li, S., Young, A., Mansi, T., Suinesiaputra, A. (Eds.), *Statistical Atlases and Computational Models of the Heart. Multi-Sequence CMR Segmentation, CRT-EPiggy and LV Full Quantification Challenges*, Springer International Publishing, Cham. pp. 317–325.
- Dikici, E., O'Donnell, T., Setser, R., White, R., 2004. Quantification of delayed enhancement MR images, in: *Medical Image Computing and Computer-Assisted Intervention*, pp. 250–257.
- Flett, A.S., Hasleton, J., Cook, C., Hausenloy, D., Quarta, G., Ariti, C., Muthurangu, V., Moon, J.C., 2011. Evaluation of techniques for the quantification of myocardial scar of differing etiology using cardiac magnetic resonance. *JACC: cardiovascular imaging* 4, 150–156.
- He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.
- Hu, J., Shen, L., Sun, G., 2018. Squeeze-and-excitation networks, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141.
- Huang, X., Liu, M.Y., Belongie, S., Kautz, J., 2018. Multimodal unsupervised image-to-image translation, in: *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 172–189.
- Jia, S., Despinasse, A., Wang, Z., Delingette, H., Pennec, X., Jais, P., Cochet, H., Sermesant, M., 2018. Automatically segmenting the left atrium from cardiac images using successive 3D u-nets and a contour loss, in: *International Workshop on Statistical Atlases and Computational Models of the Heart*, Springer. pp. 221–229.
- Karim, R., Bhagirath, P., Claus, P., Housden, R.J., Chen, Z., Karimaghloo, Z., Sohn, H.M., Rodríguez, L.L., Vera, S., Albà, X., et al., 2016. Evaluation of state-of-the-art segmentation algorithms for left ventricle infarct from late gadolinium enhancement MR images. *Medical image analysis* 30, 95–107.
- Karim, R., Blake, L.E., Inoue, J., Tao, Q., Jia, S., Housden, R.J., Bhagirath, P., Duval, J.L., Varela, M., Behar, J., et al., 2018. Algorithms for left atrial wall segmentation and thickness-evaluation on an open-source CT and MRI image database. *Medical image analysis* 50, 36–53.
- Karim, R., Housden, R.J., Balasubramaniam, M., Chen, Z., Perry, D., Uddin, A., Al-Beyatti, Y., Palkhi, E., Acheampong, P., Obom, S., et al., 2013. Evaluation of current algorithms for segmentation of scar tissue from late gadolinium enhancement cardiovascular magnetic resonance of the left atrium: an open-access grand challenge. *Journal of Cardiovascular Magnetic Resonance* 15, 105.
- Li, X., Wang, W., Hu, X., Yang, J., 2019. Selective kernel networks, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 510–519.
- Liu, J., Xie, H., Zhang, S., Gu, L., 2019. Multi-sequence myocardium segmentation with cross-constrained shape and neural network-based initialization. *Computerized Medical Imaging and Graphics* 71, 49–57.
- Liu, J., Zhuang, X., Liu, J., Zhang, S., Wang, G., Wu, L., Xu, J., Gu, L., 2014. Myocardium segmentation combining T2 and DE MRI using multi-component bivariate gaussian mixture model, in: *2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI)*, IEEE. pp. 886–889.
- Liu, J., Zhuang, X., Wu, L., An, D., Xu, J., Peters, T., Gu, L., 2017. Myocardium segmentation from DE MRI using multicomponent gaussian mixture model and coupled level set. *IEEE Transactions on Biomedical Engineering* 64, 2650–2661.
- Liu, J., Zhuang, X., Xie, H., Zhang, S., Gu, L., 2018. Myocardium segmentation from DE MRI with guided random walks and sparse shape representation. *International journal of computer assisted radiology and surgery* 13, 1579–1590.
- Liu, Y., Wang, W., Wang, K., Ye, C., Luo, G., 2020. An automatic cardiac segmentation framework based on multi-sequence MR image, in: Pop, M., Sermesant, M., Camara, O., Zhuang, X., Li, S., Young, A., Mansi, T., Suinesiaputra, A. (Eds.), *Statistical Atlases and Computational Models of the Heart. Multi-*

- Sequence CMR Segmentation, CRT-EPiggy and LV Full Quantification Challenges, Springer International Publishing, Cham. pp. 220–227.
- Lu, Y.L., Wright, G., Radau, P.E., 2013. Automatic myocardium segmentation of LGE MRI by deformable models with prior shape data. *Journal of Cardiovascular Magnetic Resonance* 15, Please correct the reference.
- Ly, B., Cochet, H., Sermesant, M., 2020. Style data augmentation for robust segmentation of multi-modality cardiac MRI, in: Pop, M., Sermesant, M., Camara, O., Zhuang, X., Li, S., Young, A., Mansi, T., Suinesiaputra, A. (Eds.), *Statistical Atlases and Computational Models of the Heart. Multi-Sequence CMR Segmentation, CRT-EPiggy and LV Full Quantification Challenges*, Springer International Publishing, Cham. pp. 197–208.
- Moghari, M.H., Pace, D.F., Akhondi-Asl, A., Powell, A.J., 2016. HVSMR 2016: MICCAI workshop on whole-heart and great vessel segmentation from 3D cardiovascular MRI in congenital heart disease. <http://segchd.csail.mit.edu/index.html>.
- Petitjean, C., Zuluaga, M.A., Bai, W., Dacher, J.N., Grosgeorge, D., Caudron, J., Ruan, S., Ayed, I.B., Cardoso, M.J., Chen, H.C., et al., 2015. Right ventricle segmentation from cardiac MRI: a collation study. *Medical image analysis* 19, 187–202.
- Radau, P., Lu, Y., Connelly, K., Paul, G., Dick, A., Wright, G., 2009. Evaluation framework for algorithms segmenting short axis cardiac MRI. *The MIDAS Journal-Cardiac MR Left Ventricle Segmentation Challenge* 49.
- Rohlfing, T., Maurer, C.R., 2006. Shape-based averaging. *IEEE Transactions on Image Processing* 16, 153–161.
- Roth, H., Zhu, W., Yang, D., Xu, Z., Xu, D., 2020. Cardiac segmentation of LGE MRI with noisy labels, in: Pop, M., Sermesant, M., Camara, O., Zhuang, X., Li, S., Young, A., Mansi, T., Suinesiaputra, A. (Eds.), *Statistical Atlases and Computational Models of the Heart. Multi-Sequence CMR Segmentation, CRT-EPiggy and LV Full Quantification Challenges*, Springer International Publishing, Cham. pp. 228–236.
- Suinesiaputra, A., Cowan, B.R., Finn, J.P., Fonseca, C.G., Kadish, A.H., Lee, D.C., Medrano-Gracia, P., Warfield, S.K., Tao, W., Young, A.A., 2011. Left ventricular segmentation challenge from cardiac MRI: a collation study, in: *International Workshop on Statistical Atlases and Computational Models of the Heart*, pp. 88–97.
- Tao, Q., Piers, S.R., Lamb, H.J., Rj, V.D.G., 2015. Automated left ventricle segmentation in late gadolinium-enhanced MRI for objective myocardial scar assessment. *Journal of Magnetic Resonance Imaging* 42, 390–399.
- Thygesen, K., Alpert, J.S., White, H.D., 2008. Universal definition of myocardial infarction. *European Heart Journal* 29, 1209.
- Tobon-Gomez, C., Geers, A.J., Peters, J., Weese, J., Pinto, K., Karim, R., Ammar, M., Daoudi, A., Margeta, J., Sandoval, Z., et al., 2015. Benchmark for algorithms segmenting the left atrium from 3D CT and MRI datasets. *IEEE transactions on medical imaging* 34, 1460–1473.
- Vesal, S., Ravikumar, N., Maier, A., 2018. Dilated convolutions in neural networks for left atrial segmentation in 3D gadolinium enhanced-MRI, in: *International Workshop on Statistical Atlases and Computational Models of the Heart*, Springer. pp. 319–328.
- Vesal, S., Ravikumar, N., Maier, A., 2020. Automated multi-sequence cardiac mri segmentation using supervised domain adaptation, in: Pop, M., Sermesant, M., Camara, O., Zhuang, X., Li, S., Young, A., Mansi, T., Suinesiaputra, A. (Eds.), *Statistical Atlases and Computational Models of the Heart. Multi-Sequence CMR Segmentation, CRT-EPiggy and LV Full Quantification Challenges*, Springer International Publishing, Cham. pp. 300–308.
- Wagner, A., Mahrholdt, H., Holly, T.A., Elliott, M.D., Regenfus, M., Parker, M., Klocke, F.J., Bonow, R.O., Kim, R.J., Judd, R.M., 2003. Contrast-enhanced MRI and routine single photon emission computed tomography (SPECT) perfusion imaging for detection of subendocardial myocardial infarcts: an imaging study. *The Lancet* 361, 374–379.
- Wang, J., Huang, H., Chen, C., Ma, W., Huang, Y., Ding, X., 2020a. Multi-sequence cardiac MR segmentation with adversarial domain adaptation network, in: Pop, M., Sermesant, M., Camara, O., Zhuang, X., Li, S., Young, A., Mansi, T., Suinesiaputra, A. (Eds.), *Statistical Atlases and Computational Models of the Heart. Multi-Sequence CMR Segmentation, CRT-EPiggy and LV Full Quantification Challenges*, Springer International Publishing, Cham. pp. 254–262.
- Wang, X., Yang, S., Tang, M., Wei, Y., Han, X., He, L., Zhang, J., 2020b. Sk-unet: An improved u-net model with selective kernel for the segmentation of multi-sequence cardiac MR, in: Pop, M., Sermesant, M., Camara, O., Zhuang, X., Li, S., Young, A., Mansi, T., Suinesiaputra, A. (Eds.), *Statistical Atlases and Computational Models of the Heart. Multi-Sequence CMR Segmentation, CRT-EPiggy and LV Full Quantification Challenges*, Springer International Publishing, Cham. pp. 246–253.
- Wei, D., Sun, Y., Ong, S.H., Chai, P., Teo, L.L., Low, A.F., 2013. Three-dimensional segmentation of the left ventricle in late gadolinium enhanced MR images of chronic infarction combining long- and short-axis information. *Medical Image Analysis* 17, 685–697.
- Yue, Q., Luo, X., Ye, Q., Xu, L., Zhuang, X., 2019. Cardiac segmentation from LGE MRI using deep neural network incorporating shape and spatial priors. Please correct the reference .
- Yushkevich, P.A., Piven, J., Hazlett, H.C., Smith, R.G., Ho, S., Gee, J.C., Gerig, G., 2006. User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. *Neuroimage* 31, 1116–1128.
- Zhao, J., Xiong, Z., 2018. 2018 atrial segmentation challenge. <http://atriaseg2018.cardiacatlas.org/>.
- Zhu, J.Y., Park, T., Isola, P., Efros, A.A., 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks, in: *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232.
- Zhuang, X., 2013. Challenges and methodologies of fully automatic whole heart segmentation: A review. *Journal of Healthcare Engineering* 4, 371–407.
- Zhuang, X., 2016. Multivariate mixture model for cardiac segmentation from multi-sequence MRI, in: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer. pp. 581–588.
- Zhuang, X., 2019. Multivariate mixture model for myocardial segmentation combining multi-source images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 2933 – 2946.
- Zhuang, X., Li, L., Payer, C., Štern, D., Urschler, M., Heinrich, M.P., Oster, J., Wang, C., Smedby, Ö., Bian, C., et al., 2019. Evaluation of algorithms for multi-modality whole heart segmentation: An open-access grand challenge. *Medical image analysis* 58, 101537.