Attention-Based Neural Network for Cardiac MRI Segmentation: Application to Strain and Volume Computation

Nicolas Portal, Catherine Achard, Saud Khan, Vincent Nguyen, Mikael Prigent, Mohamed Zarai, Khaoula Bouazizi, Johanne Sylvain, Alban Redheuil, Gilles Montalescot, et al.

IRBM 45 (2024) 100850

Contents lists available at ScienceDirect

IRBM

journal homepage: www.elsevier.com/locate/irbm

Elsevier Masson France
EM|consulte
www.em-consulte.com

Original Article

Attention-Based Neural Network for Cardiac MRI Segmentation: Application to Strain and Volume Computation

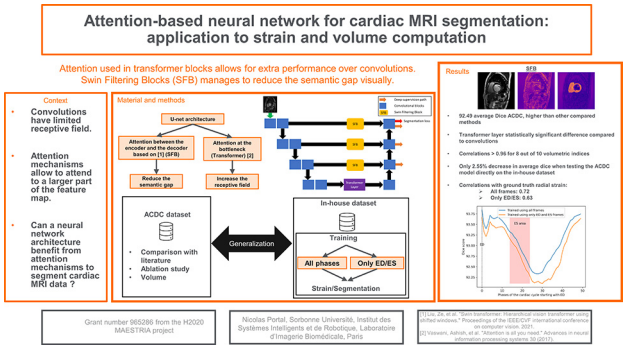
Nicolas Portal [a,b,*], Catherine Achard [b], Saud Khan [a], Vincent Nguyen [a], Mikael Prigent [c], Mohamed Zarai [c], Khaoula Bouazizi [a,c], Johanne Sylvain [d], Alban Redheuil [a,c,e], Gilles Montalescot [d], Nadjia Kachenoura [a,c,1], Thomas Dietenbeck [a,c,1]

[a] Sorbonne Université, CNRS, INSERM, Laboratoire d'Imagerie Biomédicale, LIB, F-75006, Paris, France
[b] Sorbonne Université, CNRS, INSERM, Institut des systèmes intelligents et de robotique, ISIR, Paris, France
[c] Institut de Cardiométabolisme et Nutrition (ICAN), F-75013 Paris, France
[d] Sorbonne Université, ACTION group, Pitié-Salpêtrière Hospital (AP-HP), F-75013 Paris, France
[e] Imagerie Cardio-Thoracique (ICT), Sorbonne Université, AP-HP, Groupe Hospitalier Pitié-Salpêtrière, F-75013 Paris, France

HIGHLIGHTS

- Neural network architecture relying on attention to bridge the semantic gap.
- Tested on both ACDC and an in-house dataset containing 271 patients.
- Generalization capabilities assessed by testing on out-of-distribution data.
- Comparison between training on all cardiac phases or only on ED and ES frames.
- State of the art performance and accurate volumetric indices are obtained.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 14 February 2024
Received in revised form 14 June 2024
Accepted 19 July 2024
Available online 23 July 2024

Keywords:
Segmentation
Deep learning
Transformers
Cardiac
MRI

ABSTRACT

Context: Deep learning algorithms have been widely used for cardiac image segmentation. However, most of these architectures rely on convolutions that hardly model long-range dependencies, limiting their ability to extract contextual information. Moreover, the traditional U-net architecture suffers from the difference of semantic information between feature maps of the encoder and decoder (also known as the semantic gap).

Material and method: To address this issue, a new network architecture relying on attention mechanism was introduced. Swin Filtering Blocks (SFB), that use Swin Transformer blocks in a cross-attention manner, were added between the encoder and the decoder to filter information coming from the encoder based on the feature map from the decoder. Attention was also employed at the lowest resolution in the form of a transformer layer to increase the receptive field of the network.

We conducted experiments to assess both generalization capability and to evaluate how training on all frames of the cardiac cycle rather than only the end-diastole and end-systole impacts strain and segmentation performances.

Results and conclusion: Visual inspection of feature maps suggested that Swin Filtering Blocks contribute to the reduction of the semantic gap. Performing attention between all patches using a transformer layer brought higher performance than convolutions. Training the model with all phases of the cardiac cycle

* Corresponding author at: Institut des systèmes intelligents et de robotique, Campus Pierre et Marie Currie, Pyramide - T55, 4 Pl. Jussieu 65, 75005 Paris, France.
E-mail address: portal@isir.upmc.fr (N. Portal).
[1] These authors have equally contributed.

https://doi.org/10.1016/j.irbm.2024.100850

resulted in slightly more accurate segmentations while leading to a more noticeable improvement for strain estimation. A limited decrease in performance was observed when testing on out-of-distribution data, but the gap widens for the most apical slices.

1. Introduction

According to the World Health Organization, an estimated 17.9 million people died from CVDs in 2019, representing 32% of all global deaths. Of these deaths, 85% were due to heart attack and stroke.[2] To more easily diagnose these pathologies, segmentation of cardiac structures on cardiovascular images gives crucial information as it allows to accurately delineate heart structures targeted by the disease and assess their remodeling. Magnetic resonance imaging (MRI) is a useful non-invasive and non-radiating imaging modality to detect such diseases throughout cine and tissue characterization images with their high contrast, resolution, and anatomical coverage [1]. Nowadays, many segmentation algorithms rely on deep learning methods as they achieved good results on computer vision tasks. The U-net architecture [2] is often used in practice as it proved to be an effective design to perform semantic segmentation. This architecture consists of an encoder to aggregate contextual information, a symmetric decoder to enable precise localization, and skip connections between the encoder and the decoder to exploit local information contained in high-resolution feature maps. This design effectively fuses the encoder high-resolution information with the decoder semantically-rich features. However, a simple concatenation of feature maps originating from the encoder and the decoder has been shown to be suboptimal [3,4]. Indeed, the encoder feature maps, though containing very fine-grained local details, carry less semantically rich features than the decoder feature maps. This phenomenon is known as the *semantic gap*. Traditionally, convolutions have been used to try to bridge such semantic gap [5] as they allow to gradually increase the receptive field. Some gating mechanisms have also been implemented to filter features being passed to the decoder, based on features originating from all levels [6] of the encoder.

More recently, attention mechanisms have been shown to be effective in improving neural network performances. Following a previous work [7] transformer blocks were used at the bottom of the U-net networks either in a 2D or a 3D configuration, improving the performance of medical image segmentation algorithms [8,9]. Convolutions were used to reduce the spatial resolution of feature maps allowing to subsequently incorporate self-attention at higher levels in the encoder [10]. A window and shifted window attention mechanisms were introduced to reduce the computational complexity of traditional transformer blocks while improving performances [11]. As a result, these blocks have been applied successfully to both 2D [12] and 3D [13] medical image segmentation. Attention has also been used to address the semantic gap issue.

However, most methods that specifically attempted to reduce the *semantic gap* either still use convolutions or rely on self-attention. The few methods that use cross-attention either rely on entire transformer blocks [14] or use full spatial attention [15], both of which are computationally expensive.

Accordingly, this work focuses on the design of a new network architecture capable of better managing the semantic gap between encoder and decoder data. Recent work [14,15] has shown that cross-attention mechanisms appear to be the ideal solution for determining, from semantically rich decoder data, where to pay attention for precise localization information in encoder data. However, these methods are very expensive in terms of number of parameters, and are therefore prone to poor generalization when the number of data is limited. The proposed architecture overcomes this problem and improves the segmentation performance evaluated on public and private datasets by performing attention in smaller windows. In order to assess the strengths and weaknesses of the method, we have carried out a detailed study of segmentation results. On the clinical side, we show that the proposed method also improves the accuracy of volumetric quantitative indices necessary to diagnose cardiovascular diseases. We also studied the generalizability of the algorithm on new databases (out-of-distribution data) to verify that the method could be used in different centers, without requiring re-learning.

The first part of this study presents the network architecture and includes a detailed explanation of the Swin Filtering Block (SFB) used to bridge the semantic gap. This section also outlines the data characteristics and the protocol followed for the experiments. Then, results obtained with this network architecture are presented, with an ablation study, comparison with literature as well as generalization performance and influence of training only with ED and ES phases. Finally, we discuss our findings in a third section before concluding.

2. Materials and methods

This section details the databases used in this article and presents the proposed network architecture. Our main contributions, the used dataset and the performed validations are summarized in Fig. 1 as a block diagram.

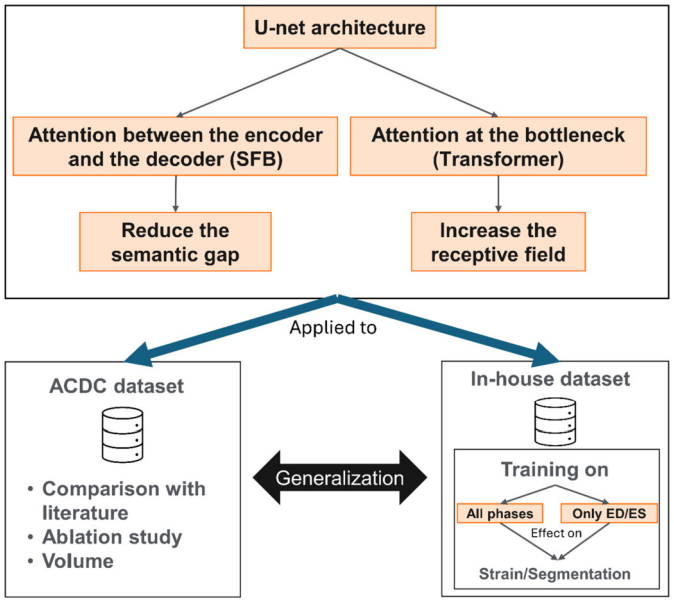


Fig. 1. Block Diagram of the proposed approach. Our network is tested on two datasets and we assess the consequences of training only with the ED and ES frames compared to all frames of the cardiac cycle. SFB: Swin Filtering Block.

[2] https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).

2.1. Study population and MRI data

Study data comprised the public Automated Cardiac Diagnosis Challenge (ACDC) dataset [16], which was fully anonymized and handled within the regulations set by the local ethics committee, as well as an in-house dataset (Quorum study, NCT03715998 [17]) where all participants gave their written informed consent for the initial protocol and for ancillary use of their data. The study protocol was approved by the local ethics committee. Procedures carried out on these subjects were in accordance with the declaration of Helsinki.

The ACDC dataset [16] comprised 100 subjects divided into 5 equal-sized groups according to specific heart conditions (healthy, hypertrophic cardiomyopathy, dilated cardiomyopathy, abnormal right ventricle and myocardial infarction). Short-axis images covering the heart from its base to its apex were acquired using either a 1.5 or 3 Tesla Siemens MRI scanner with a pixel size between 0.7 and 1.92 mm^2 (mean pixel size 1.51 mm^2) and a slice thickness between 5 and 10 mm (mean slice thickness 9.34 mm). 1902 labeled 2D slices are present in the dataset. Ground truth segmentation annotations were drawn manually by two cardiologists with more than 10 years of experience each. The left ventricle (LV), myocardium (MYO) and right ventricle (RV) were manually labelled at end-systole (ES) and end-diastole (ED) and from base to apex. More details regarding the annotation process can be found in [16]

The in-house dataset is a multi-center and multi-vendor dataset which includes patients who had an acute myocardial infarction (MI) with varying degree of severity. Images were acquired in 24 different centers and using MRI scanners from 3 different manufacturers (Siemens, General Electric and Philips). Short-axis images covering the heart from its base to its apex were acquired during the acute phase of the MI using either a 1.5 or 3 Tesla MRI scanner with a pixel size ranging between 0.68 and 2.34 mm^2 (mean pixel size 1.43 mm^2) and a slice thickness between 6 and 8 mm (mean slice thickness 7.4 mm).

The in-house dataset is used both for assessing the impact of training on all phases of the cardiac cycle and for evaluating the generalization performance of our model when applied to a different dataset.

In the first case, ground truth segmentation labels of 271 patients were obtained for all phases of the cardiac cycle using the CardioTrack software ([18], Laboratoire d'Imagerie Biomédicale, Sorbonne Université, INSERM, CNRS). This software uses a feature tracking algorithm to automatically track a manually initialized contour on the ES frame of the sequence. Segmentation labels are generated for each frame based on these contour points. Most annotations were provided only for 3 slices in the volume, representing a total of 34452 2D slices. For these slices, the software was also used to obtain ground truth radial and circumferential strain. ED and ES volumes of 195 of these 271 patients were also manually annotated using a commercial software (QMass, Medis, Leiden, The Netherlands, version 4.0.24.4). Annotations are generated automatically by the software and modified, if necessary, by a clinical expert. All slices of the volumes were annotated, representing 4072 2D slices. Annotated cardiac structures in the in-house dataset are the same as those in the ACDC dataset.

2.2. Model architecture

The proposed network called SFB-net (Swin Filtering Block network), is based on the U-net architecture [2] with an encoder, a decoder and skip-connections in-between, as illustrated in Fig. 2. Convolutional blocks, depicted in blue, were used throughout the network and doubled in the encoder as compared to the decoder to improve the model encoding ability [19]. These blocks contained 2 convolutions, each followed by a batch normalization layer and a Gaussian Error Linear Unit (gelu) [20] activation. We use Gelu rather than Relu throughout the network to stay consistent with the transformer layer which is generally implemented with the former activation function. Moreover, Gelu does not suffer from the "dying Relu" effect where the activation always yields zero outputs for negative inputs, preventing the network from adjusting its weights [21,22]. The number of filters was doubled at each encoder layer and halved for the corresponding decoder layers. Moreover, we propose to use strided convolutions instead of pooling layers to down-sample feature maps as it allows the network to more flexibly perform this operation using learnable parameters. The number of down-sampling was limited resulting in a feature map at the bottleneck 8 times smaller than the input image size. Up-sampling was carried out using 2D transposed convolutions. To compensate for the resulting shallowness of the network, which may reduce its receptive field, we introduce a conventional transformer layer at the bottleneck (depicted in purple). This enabled the network to take advantage of global contextual information. Note that transformer blocks were not used in the encoder and decoder since keeping convolutions at higher resolutions was shown to result in higher performances [10,23,24]. Indeed, convolutions generalize better to unseen images than transformers and extract local information found at higher resolutions more effectively. A Summary of our network architecture can be found in supplementary materials.

Deep supervision was applied at each stage of the decoder. More precisely, ground truth segmentations were down-sampled to match the size of the network's outputs. The loss weights, $\alpha_i \forall\ i\ \in \{0,1,2\}$ for each resolution, were halved when the image size is reduced and normalized so that the sum of weights equals 1. The final loss was the weighted sum of successive stages losses and was defined as:

$$\mathcal{L} = \alpha_1 \times \mathcal{L}_{\{H,W\}} + \alpha_2 \times \mathcal{L}_{\left\{\frac{H}{2},\frac{W}{2}\right\}} + \alpha_3 \times \mathcal{L}_{\left\{\frac{H}{4},\frac{W}{4}\right\}} \quad (1)$$

Where H and W represent the height and width of the input image respectively.

With:

$$\alpha_i = \frac{\frac{1}{2^i}}{\sum_{j=0}^{2} \alpha_j} \forall i \in \{0,1,2\} \quad (2)$$

For fair comparison with previous studies [13,25], a combination of cross-entropy and Dice loss was used to compute each individual $\mathcal{L}$.

2.3. Swin filtering blocks (SFB)

We also propose to introduce a filtering mechanism in the skip connection paths between the encoder and the decoder, as illustrated by the yellow blocks in Fig. 2 and described with more details in Fig. 3. The goal was to enable the decoder to filter out irrelevant information originating from the encoder. More precisely, the encoder's feature maps contained noise that should be discarded before concatenation with the decoder feature maps. To do so, local information contained in the encoder high-resolution feature maps and located in areas underlined by semantically-rich feature maps of the decoder were emphasized, while response in the remaining noisy areas were toned down. Multi-Head-Self-Attention (MHSA), [26] was used in this process, where window and shifted window versions of MHSA [11] were preferred for their lower computational load, reducing training time and memory consumption. Windowed Multi Head Self Attention (W-MHSA) performed attention in windows of M by M sized patches. When performing Shifted Window Multi-Head Self Attention (SW-MHSA), windows were shifted by $\lfloor \frac{M}{2} \rfloor$ both in the x and y direction so that

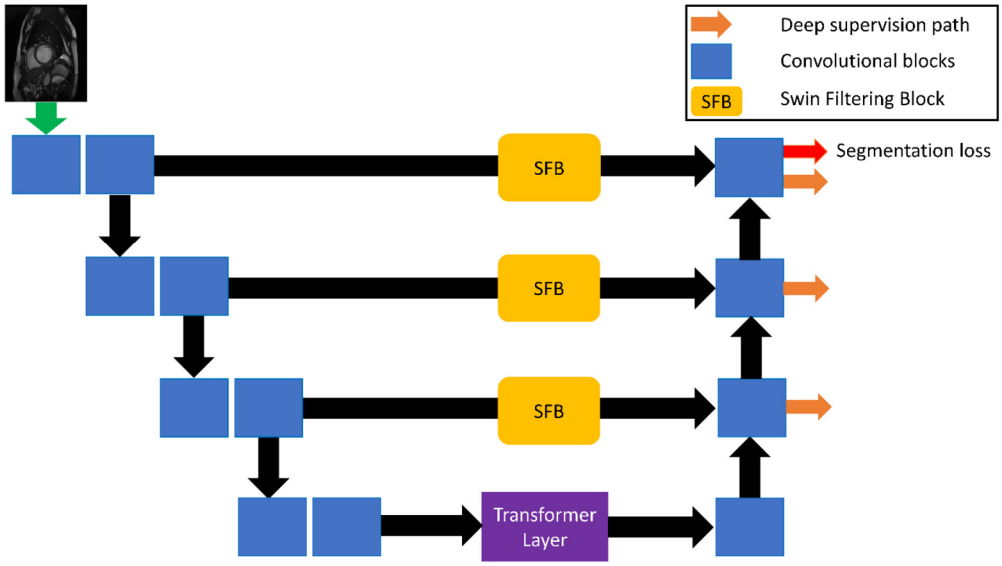


Fig. 2. Representation of SFB-net. Deep supervision is used in the decoder. Convolutional blocks are represented in blue, Swin Filtering Blocks (SFB) in yellow, and the conventional transformer layer in purple.

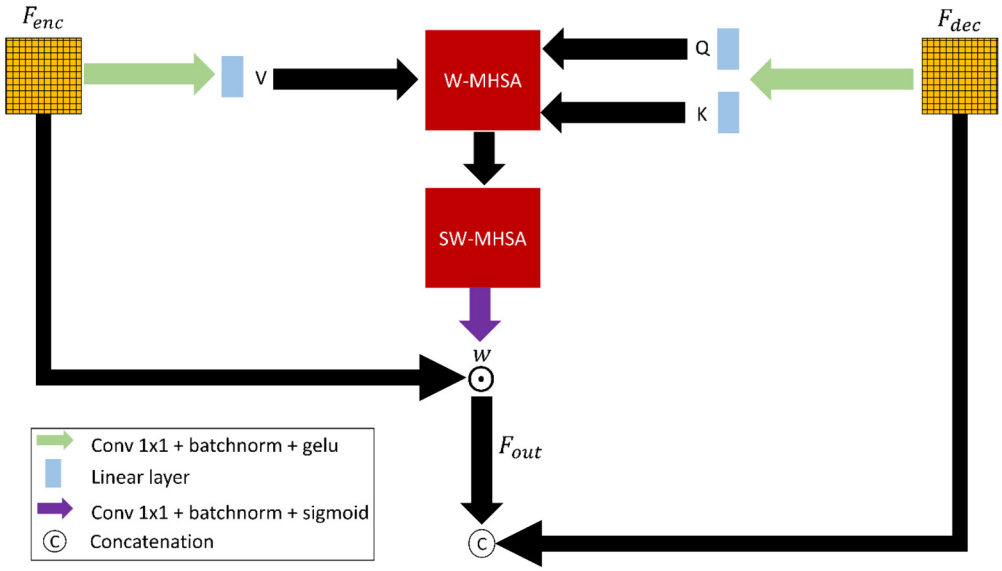


Fig. 3. Schematic representation of the Swin Filtering Blocks (SFB) used between the encoder and the decoder. W is the set of computed weights used to rescale the encoder feature map. W-MHSA: Window Multi Head Self Attention, SW-MHSA: Shifted Window Multi Head Self Attention.

attention can be conducted between patches belonging to different windows. W-MHSA was described as:

$$\mathrm{W-MHSA}(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d}} + B\right)V \tag{3}$$

where Q, K and $V \in \mathbb{R}^{M^2 \times d}$ are the query, key and value respectively, d is the query, key and value dimension and $B \in \mathbb{R}^{M^2 \times M^2}$ the learnable relative position bias added to each head which encode the relative position between patches. Q, K and V are tensors generated using separate linear layers from a feature map F as:

$$\begin{aligned} Q_F &= FA_Q^T + b_Q \\ K_F &= FA_K^T + b_K \\ V_F &= FA_V^T + b_V \end{aligned} \tag{4}$$

with $b_i \in \mathbb{R}^d$ the bias and $A_i \in \mathbb{R}^{d \times d}$ a weight matrix $\forall i \in \{Q, K, V\}$. Q_F, K_F and V_F are the query, key and value generated from feature map F. W-MHSA and SW-MHSA blocks were favorably used to perform cross-attention between the encoder and the decoder's feature maps. Cross-attention used the same process as self-attention but with key, query and value originating from different feature maps. Since feature maps coming from the encoder were rescaled based on those of the decoder, values were chosen to come from the encoder while both query and key should come from the decoder:

$$CA_{out} = \mathrm{SW-MHSA}\left(\mathrm{W-MHSA}\left(Q_{F_{dec}}, K_{F_{dec}}, V_{F_{enc}}\right)\right) \tag{5}$$

Where F_{dec} and F_{enc} are the feature maps coming from the decoder and encoder respectively. The result was passed to a sigmoid function to generate weights w ranging between 0 and 1 used to rescale the encoder feature map.

$$w = \sigma\left(conv\left(CA_{out}\right)\right) \tag{6}$$

Table 1
Number of patients and images in the training, validation and test set for ACDC and the in-house dataset. For ACDC the average number of 2D slices per fold is reported.

	Train (# 2D slices)	Validation (# 2D slices)	Test (# 2D slices)
ACDC (1 fold)	80 (1521.6)	.	20 (380.4)
In-house (195 patients)	124 (2629)	32 (663)	39 (780)
In-house (271 patients)	174 (22743)	43 (5334)	54 (6375)

where $\sigma(x) = \frac{1}{1+e^{-x}}$ is the sigmoid function and $conv$ is a standard convolution with 1 by 1 kernel followed by batch normalization. Finally, the rescaled encoder feature map F_{out} was obtained by applying the Hadamard product $\odot$ between computed weights w and the original encoder feature map:

$$F_{out} = F_{enc} \odot w \tag{7}$$

2.4. Implementation details

For the ACDC dataset, a 5-folds cross-validation was used to evaluate the model's performances. For the in-house dataset, we use 80% of patients for training and 20% for testing. 20% of patients of the training set are used for validation. The number of patients in the training and test set is provided in Table 1.

SFB-net was implemented with Pytorch and trained using a 16GB Tesla v100 SXM2. The nnU-net [25] framework was used as a starting point for this work. AdamW optimizer and cosine annealing scheduler were used for training. The initial learning rate and weight decay were both set to 0.0001. For fair comparison with other studies [13,25], the number of training epochs was set to 1000. Each epoch was made up of 250 iterations (random sampling of training images to form a batch. 250 is the default value of the nnU-net framework). Batch size was 10 for ACDC and 6 for the in-house dataset. Weights of the model after the last epoch are selected for inference.

We kept nnU-net pre-processing steps: before training, all volumes were resampled to have a voxel size equal to the median voxel size of the dataset. Since we use a 2D network, in our case, no resampling is performed along the z axis of the volume.

2D Images are center-cropped to size 224x224 for ACDC and 288x288 for the in-house dataset. Before passing them to the network we normalize them so that they have a mean of 0 and standard deviation of 1.

A wide range of data augmentations was applied on the fly during training for both the ACDC and in-house dataset.: rotation, scaling, gamma adjustment, brightness adjustment, mirroring, contrast modification, low-resolution simulation, noise, and blur. More information on data augmentation can be found in the supplementary materials. Test-time data augmentation was used as it is the default setting of nnU-net. The maximum number of filters at the bottleneck of the network was 512. Training and validation losses as well as validation accuracy in terms of Dice score for each class are available in supplementary materials.

3. Results

The proposed network yields segmentation performances higher on average than state-of-the-art networks, volume prediction highly correlated to ground truth measurements (r > 0.9) and satisfying generalization capabilities (performances drop < 5% in Dice score on cross-dataset evaluation). The transformer layer used at the lowest resolution brought statistically significant performance improvements. Generalization performance on out-of-distribution

Table 2
Ablation study performed on the ACDC dataset. Dice scores, Intersection over Union (IOU) scores, Average Symmetric Surface Distance ($ASSD$) and Hausdorff Distance (HD) are obtained before post-processing on 5 folds and reported as mean ± standard deviation. Bold values correspond to the highest performance for the heart structure. RV = right ventricular cavity, MYO = myocardium, LV = left ventricular cavity, SFB = Swin Filtering Block, d-s = deep supervision. Reported p-values refer to the comparison of the baseline SBF-net and the method of the current line. They are computed using the Wilcoxon signed rank test.

	Models	Average	RV	MYO	LV
Dice	SFB-net	**92.45 ± 3.15**	91.50 ± 6.45	**90.83 ± 2.99**	**95.04 ± 4.29**
	SFB-net w/o SFBs	92.42 ± 3.31 p=0.1140	91.63 ± 6.47 p=0.3008	90.77 ± 2.90 p=0.2114	94.86 ± 4.65 p=0.1224
	SFB-net w/o transformer	92.26 ± 3.35 *p=0.0159*	91.41 ± 6.47 p=0.3870	90.62 ± 3.20 *p=0.0011*	94.75 ± 4.62 p=0.0825
	SFB-net w/o d-s	92.41 ± 3.22 p=0.3236	**91.75 ± 5.86** p=0.8544	90.73 ± 3.12 p=0.8079	94.74 ± 4.86 p=0.2544
IOU	SFB-net	86.36 ± 4.98	84.91 ± 9.80	83.33 ± 4.86	90.83 ± 6.96
	SFB-net w/o SFBs	86.29 ± 5.16 p=0.1093	85.12 ± 9.70 p=0.3031	83.22 ± 4.72 p=0.1966	90.54 ± 7.37 p=0.1195
	SFB-net w/o transformer	86.04 ± 5.28 p=0.0147	84.77 ± 9.91 p=0.3783	82.99 ± 5.17 p=0.0012	90.35 ± 7.34 p=0.0845
	SFB-net w/o d-s	86.26 ± 5.11 p=0.3165	85.25 ± 9.09 p=0.8649	83.18 ± 5.06 p=0.8041	90.35 ± 7.69 p=0.2565
ASSD (mm)	SFB-net	**0.55 ± 0.49**	0.76 ± 1.00	**0.43 ± 0.30**	**0.46 ± 0.75**
	SFB-net w/o SFBs	0.56 ± 0.55 *p=0.1357*	0.72 ± 0.99 *p=0.5231*	0.45 ± 0.35 *p=0.0986*	0.51 ± 0.84 *p=0.4581*
	SFB-net w/o transformer	0.60 ± 0.51 *p=0.0128*	0.76 ± 0.98 *p=0.3964*	0.50 ± 0.47 *p=0.0006*	0.54 ± 0.72 *p=0.0718*
	SFB-net w/o d-s	0.58 ± 0.56 *p=0.0928*	**0.70 ± 0.85** *p=0.2721*	0.48 ± 0.46 *p=0.3147*	0.55 ± 0.93 *p=0.3230*
Hausdorff distance (mm)	SFB-net	**9.24 ± 6.29**	**12.99 ± 11.04**	**7.99 ± 8.02**	**6.73 ± 7.48**
	SFB-net w/o SFBs	10.22 ± 8.13 *p=0.1037*	14.23 ± 13.96 *p=0.4159*	8.66 ± 9.44 *p=0.2332*	7.78 ± 9.78 *p=0.2465*
	SFB-net w/o transformer	11.49 ± 10.96 *p=0.0350*	13.78 ± 11.63 *p=0.0755*	11.51 ± 17.66 *p=0.0176*	9.18 ± 14.15 *p=0.0419*
	SFB-net w/o d-s	10.55 ± 8.91 *p=0.0586*	14.30 ± 13.93 *p=0.0209*	9.21 ± 11.32 *p=0.3231*	8.13 ± 10.66 *p=0.0945*

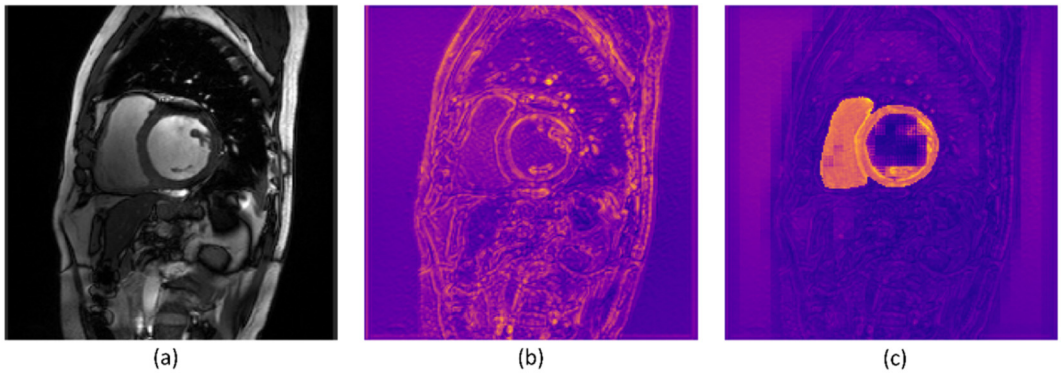


Fig. 4. (a) Image passed as input to the network. (b) and (c) feature maps originating from the encoder before and after being rescaled by the SFB respectively. The feature map was extracted at the highest resolution and averaged over the filter dimension.

data was solid, except for the most apical slices. Using all frames of the cardiac cycle during training resulted in a notable gain in strain estimation compared to using only the ED and ES frames.

3.1. Results and ablation study

Results in terms of Dice score, Intersection over Union (IOU), Hausdorff Distance (HD) and Average Symmetric Surface Distance (ASSD) for the ACDC dataset are presented in Table 2. Results on this specific dataset are obtained using 5-fold cross-validation. We also conducted an ablation study to further assess the effectiveness of the different components of the SFB-net approach. SFB-net was compared to its three variants:

- SFB-net w/o SFBs: SFB-net without SFBs.
- SFB-net w/o transformer: SFB-net where the transformer layer at the bottleneck is replaced by a convolution.
- SFB-net w/o d-s: SFB-net without deep supervision.

All the three ablation variants have around 21 million of parameters. Such ablation study showed that for most variants, small decline in performances was found although they did not reach statistical significance, as revealed by the slight increase in ASSD and HD and small decrease in Dice score and IOU when compared to SFB-Net. Fig. 4 illustrates an example of a feature map coming from the encoder before and after being rescaled by the weights computed by the SFB. The feature map was taken at the highest resolution in the network and averaged along the feature dimension. Finally, when replacing the transformer at the bottleneck by a single convolution the deterioration was significant as revealed by a more noticeable increase in ASSD and HD and notable decrease in Dice scores and IOU when compared to SFB-Net. Fig. 5 illustrates segmentation results of SFB-net and SFB-net with the bottleneck transformer layer replaced by one convolution before post processing.

Fig. 6 shows ASSD distribution through slice level within the heart volume for each heart structure. The number of slices with an ASSD > 5 mm was low and similar for the LV and myocardium but higher for the RV especially in the most basal slices. The worst segmented slices on ACDC are also illustrated in Fig. 6. These slices represent the most basal sections of the volume and exhibit common patterns. The network struggles to determine whether the right ventricular cavity should be segmented.

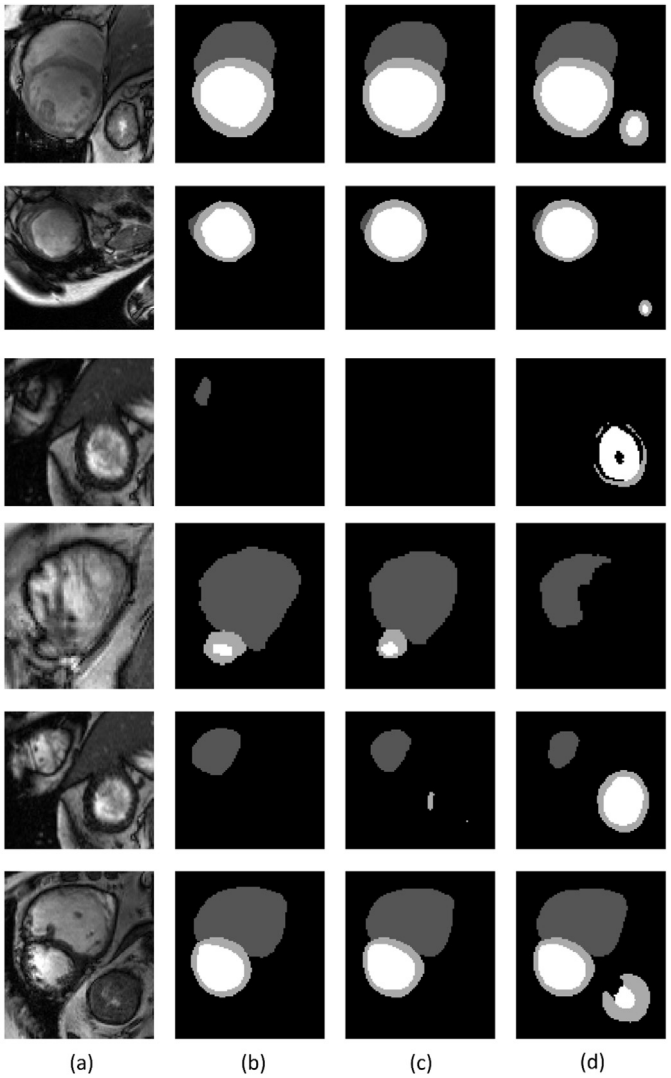


Fig. 5. Comparison of segmentation results between SFB-net and SFB-net w/o transformer on ACDC. (a) native image to be segmented, (b) ground truth segmentation, (c) SFB-net predictions, (d) SFB-net w/o transformer predictions.

3.2. Comparison with literature

Table 3 provides the results of the comparison on the ACDC dataset between our SFB-net approach and 6 recent methods of the literature (namely nnU-net, 2021 [25]; Ω-net, 2018 [27]; TransUnet, 2021 [8]; SwinUnet, 2022 [12]; Unetr, 2022 [9]; nnFormer, 2021 [13]).

Reported results came from respective literature manuscripts. Such comparisons revealed that the SFB-net approach achieved the highest overall Dice score (92.49%), as well as the highest Dice score for the myocardium (90.85%), while Dice scores of the LV (95.08%, 5th / 7) and RV (91.53%, 2nd / 7) cavities were slightly lower than literature findings.

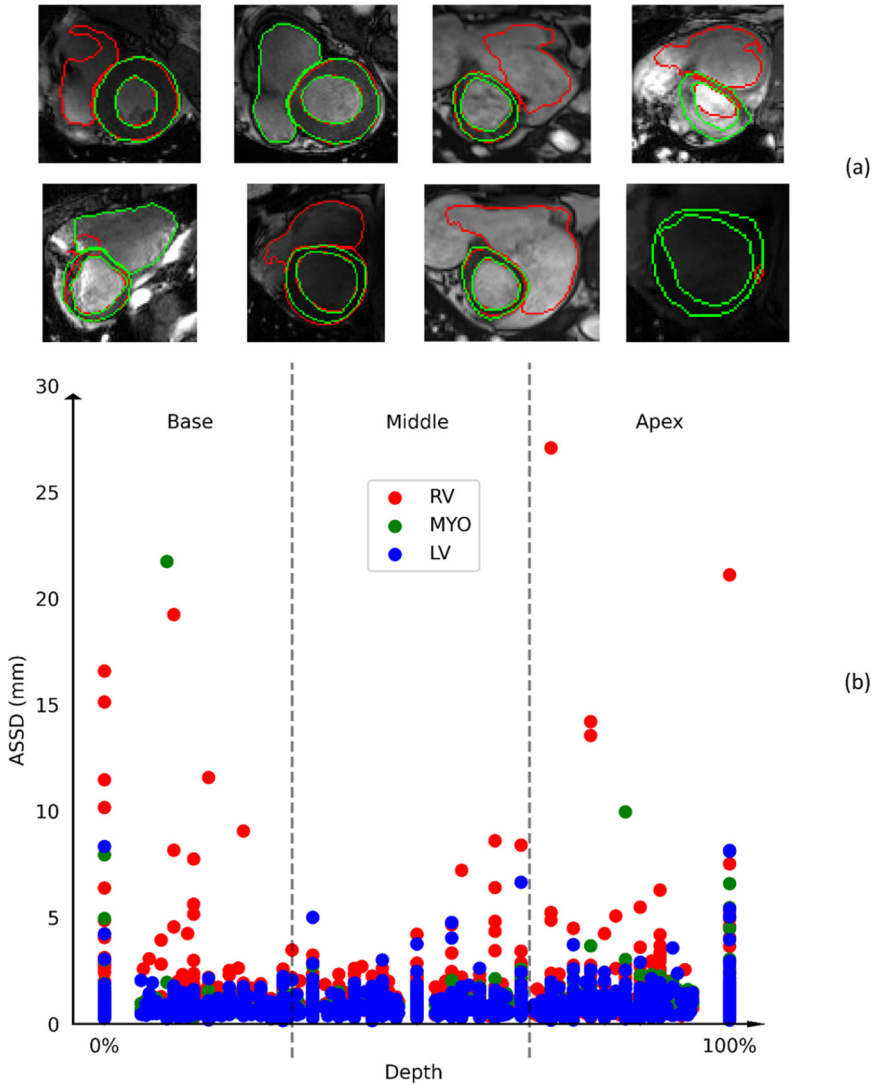


Fig. 6. ASSD per slice location within the heart volume on ACDC data. (a): illustration of suboptimal segmentations identified by their high ASSD, where ground truth contours are in green, predictions in red. (b): ASSD against slice level in the heart volume expressed in percentage (0% are most basal slice, 100% most apical slice).

Table 3
Comparison with literature on the ACDC dataset. The metric used is the Dice score. Only nnU-net and SFB-net results were re-computed. Other results are taken from the author's manuscript. Bold values correspond to the highest performance for the considered heart structure. RV = right ventricular cavity, MYO = myocardium, LV = left ventricular cavity.

Methods	Average	RV	MYO	LV
nnU-net [25]	91.75	90.67	90.18	94.40
Ω-net [27]	92.16	**92.00**	89.10	95.40
TransUnet [8]	89.71	88.86	84.54	95.73
SwinUnet [12]	90.00	88.55	85.62	**95.83**
Unetr [9]	88.61	85.29	86.52	94.02
nnFormer [13]	92.06	90.94	89.58	95.65
SFB-net	**92.49**	91.53	**90.85**	95.08

3.3. Volumetric quantitative indices

The associations between predicted and ground truth volumetric indices for each cardiac structure of the ACDC dataset are summarized in Table 4, along with Bland-Altman statistics. Ground truth values were computed from segmentation labels. Correlations between the predicted and ground truth measures were high ($\rho >$ 0.9) for all measurements and low Bland-Altman biases (< 1% except for the RV stroke volume: bias = -2.27%) and narrow limits of agreements were found.

3.4. Generalization performance

Table 5 compares performances of our model trained and tested either on the in-house dataset or ACDC and some qualitative examples are shown in Fig. 7.

Note that, the ACDC dataset contains basal slices where the RV appears in two parts as can be seen in Fig. 8. These slices do not exist in our in-house dataset. Therefore, in order to avoid abnormally low RV results on these slices, RV metrics when testing on ACDC were computed without considering the 2 most basal slices. Models tested on the same dataset they were trained on exhibited better results than other models. The reduction in Dice score was more pronounced for the RV than other structures (the model trained on ACDC achieved a Dice score 5.12, 4.25 and 1.44 points lower on the RV, MYO and LV respectively when testing on the in-house dataset than when testing on the ACDC dataset). Fig. 9 (a) shows the cumulative frequency plot of each model. Around 20% of volumes had a Dice score below 92 for models trained and tested on the same dataset, against around 60% for others. The main reason for this gap in performance between cross-dataset

Table 4
Associations between predicted and ground truth volumetric quantitative indices on ACDC. Correlation coefficients and Bland-Altman mean bias and limits of agreement estimated between predicted and ground truth quantitative indices on the 40 patients of the testing set of the in-house dataset. RV = right ventricular cavity, MYO = myocardium, LV = left ventricular cavity. Relative mean biases are reported as a percent of the corresponding ground truth value. P-values are computed using the Wilcoxon Signed Rank test between predicted and ground truth values.

parameter	ρ	Relative mean bias (%)	Absolute mean bias [Loa]	p-value
LV end diastole volume (ml)	1.0	-0.96	-1.585 [-11.295, 8.126]	0.0034
LV end systole volume (ml)	1.0	-1.26	-1.249 [-14.675, 12.177]	0.0908
LV Ejection Fraction (%)	0.98	0.26	0.122 [-7.439, 7.683]	0.9347
LV Stroke Volume (ml)	0.96	-0.51	-0.335 [-14.101, 13.430]	0.2471
MYO end diastole mass (g)	0.99	1.04	1.358 [-15.034, 17.751]	0.2542
MYO end systole mass (g)	0.99	0.44	0.657 [-17.546, 18.859]	0.4023
RV end diastole volume (ml)	0.98	-1.05	-1.607 [-20.574, 17.361]	0.0709
RV end systole volume (ml)	0.97	-0.10	-0.083 [-26.639, 26.473]	0.3153
RV Ejection Fraction (%)	0.90	-0.59	-0.276 [-15.597, 15.045]	0.5274
RV Stroke Volume (ml)	0.89	-2.27	-1.523 [-28.548, 25.501]	0.2726

Table 5
Generalization performance. Dice scores, IOUs, ASSDs and HDs are reported as mean ± standard deviation (in mm). RV = right ventricular cavity, MYO = myocardium, LV = left ventricular cavity. P-values are computed between the model trained on the in-house dataset and tested on ACDC, and the model both trained and tested on ACDC (third and fourth line for each metric). The Wilcoxon Signed Rank test is used to compute these p-values. When testing on ACDC, RV results for the 2 most basal slices were not considered.

	train	test	Average	RV	MYO	LV
Dice	ACDC	In-house	88.99 ± 4.44	86.75 ± 6.72	86.59 ± 5.25	93.64 ± 3.56
	In-house	In-house	92.14 ± 3.69	92.08 ± 3.95	89.22 ± 5.33	95.12 ± 3.10
	ACDC	ACDC	92.60 ± 2.94	91.87 ± 5.89	90.84 ± 2.98	95.08 ± 4.17
	In-house	ACDC	89.66 ± 4.03	87.90 ± 8.10	87.33 ± 4.23	93.76 ± 4.87
			<0.0001	*<0.0001*	*<0.0001*	*<0.0001*
IOU	ACDC	In-house	*80.71 ± 6.62*	*77.19 ± 9.96*	*76.69 ± 7.40*	*88.24 ± 5.94*
	In-house	In-house	*85.76 ± 5.77*	*85.55 ± 6.47*	*80.90 ± 7.71*	*90.84 ± 5.31*
	ACDC	ACDC	*86.57 ± 4.62*	*85.44 ± 8.82*	*83.36 ± 4.85*	*90.90 ± 6.78*
	In-house	ACDC	*81.87 ± 5.96*	*79.24 ± 11.53*	*77.75 ± 6.29*	*88.61 ± 7.82*
			<0.0001	*<0.0001*	*<0.0001*	*<0.0001*
ASSD (mm)	ACDC	In-house	0.71 ± 0.42	1.03 ± 0.88	0.60 ± 0.31	0.51 ± 0.33
	In-house	In-house	0.42 ± 0.25	0.44 ± 0.27	0.44 ± 0.28	0.37 ± 0.29
	ACDC	ACDC	0.43 ± 0.38	0.45 ± 0.71	0.42 ± 0.25	0.42 ± 0.58
	In-house	ACDC	0.70 ± 0.57	0.74 ± 0.86	0.75 ± 0.78	0.62 ± 0.73
			<0.0001	<0.0001	<0.0001	<0.0001
HD (mm)	ACDC	In-house	9.79 ± 3.25	14.43 ± 7.12	8.75 ± 3.69	6.02 ± 2.01
	In-house	In-house	7.08 ± 2.17	9.49 ± 4.75	6.73 ± 2.95	5.00 ± 1.91
	ACDC	ACDC	8.20 ± 4.19	11.42 ± 8.44	7.14 ± 4.81	6.04 ± 4.39
	In-house	ACDC	10.37 ± 5.21	12.41 ± 7.09	11.44 ± 9.74	7.26 ± 5.18
			<0.0001	0.0288	<0.0001	<0.0001

models and other models is presented in Fig. 9 (b) where Dice score of 2D slices relative to their depth in volumes is shown. It can be seen that the gap in performance was within 5 Dice points for basal and mid slices in volumes. However, the difference in performance was more pronounced for apical slices, especially when testing on the in-house dataset. Indeed, for this test set, the model trained on ACDC had a Dice score 30 points lower than the model trained on the in-house dataset for the most apical slices.

3.5. Effect of training only on the end-diastole and end-systole frames

In Table 6 results of our model when trained on all frames of the in-house dataset are compared with the performance of the same model but trained only on the ED and ES frame. The model trained only on ED and ES frames exhibited a small deterioration of performance for all segmentation metrics (average Dice: -0.23, average ASSD: +0,01 mm, average HD: +0,15 mm). However, training only with ED and ES frames had a more pronounced impact on LV radial and circumferential strain metrics as well as RV circumferential strain metrics. Indeed, correlations were respectively 0.63, 0.76 and 0.12 against 0.72, 0.84 and 0.57 for the model trained on all phases. For the strain peak index, the impact was less noticeable with the same LV radial correlation and a LV (RV) circumferential correlation of 0.88 (0.87) against 0.90 (0.82) when training on all phases. Regarding the RV circumferential strain peak values, training only with ED and ES frames resulted in more outliers as indicated by the wider limits of agreement ([-34.91; 33.41] against [-13.03; 11.54] for the model trained with all frames). Fig. 10 displays Dice scores of both algorithms against phase number in the cardiac cycle. Frames were sorted starting with the ED one and results were interpolated over the maximum number of phases in the testing set. It can be seen that, on average, there was a gap of 0.2 Dice points for frames near ED or ES, while the drop was around 0.5 Dice points for frames that were the most distant from ED or ES (for example around phases number 35 to 38).

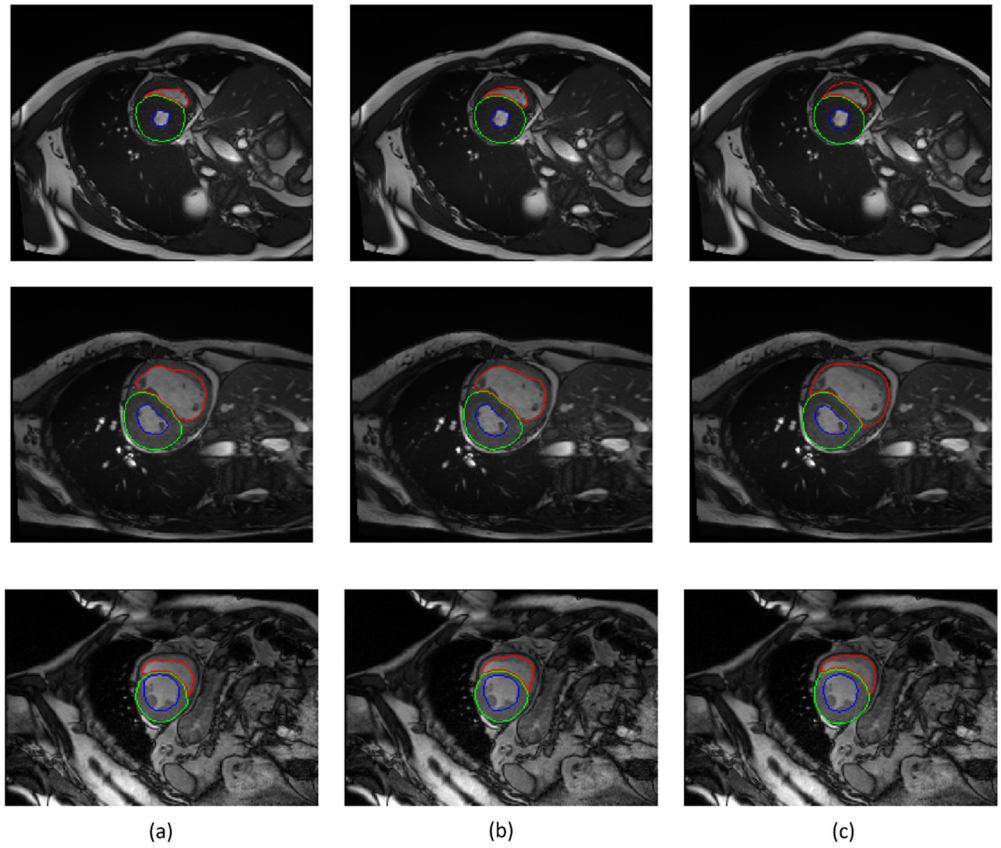


Fig. 7. Example of segmented slices from the 3 worst performing volumes in terms of Dice score on the ACDC dataset. RV contours are in red, myocardium contours in green and left ventricular cavity in blue. (a) Ground truth annotations, (b) predictions of the model trained on ACDC and (c) predictions of the model trained on the in-house dataset.

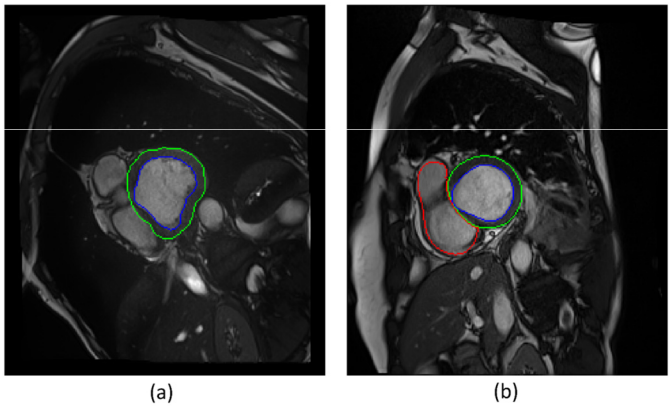


Fig. 8. Problematic basal slice in ACDC (Red: RV annotation, blue: LV, green: myocardium). (a): ACDC volumes contain basal slices with the RV appearing broken into two distinct parts. (b) In the in-house dataset, these slices did not exist. Instead, the RV in the most basal slices exhibited a less visible separation which did not result in the removal of the annotation.

4. Discussion

Although our Swin Filtering Blocks seem to be able to reduce the semantic gap visually, it did not materialize in a significant increase in segmentation performance. However, the transformer layer at the bottleneck, which, as SFB, also relies on attention, brought a significant increase in performance compared to using a convolution, demonstrating that attention mechanisms can indeed contribute to improving segmentation performance. Moreover, the proposed network architecture can be used to predict accurate clinical volumetric quantitative indices and generalizes well to data coming from different centers and manufacturers. Finally, training only on the ED and ES frames led to a small decrease in segmentation performance, but the impact on radial and circumferential strain was more noticeable.

When using the transformer layer at the bottleneck rather than a convolution, the network seemed less likely to provide incoherent segmentation such as a heart structures at two different positions. This may result from the larger receptive field of these layers which allowed to benefit from larger contextual information. This finding was also in line with a previous study [28] which showed that transformer architectures better preserved input spatial information throughout the network than Convolutional Neural Network (CNN). The ability to segment the heart structures as one single connected component also confirmed that transformers, as humans, relied more on shapes to make decisions, unlike convolutions which mainly used textures [29]. Looking at the SFB-rescaled feature maps, it can be noticed that noisy responses in non-cardiac structures were reduced while important areas were highlighted, suggesting that SFBs helped to reduce the semantic gap. Interestingly, in these maps, the area of the left ventricular cavity was not as bright as other areas of the heart. This may come from the fact that, since the left ventricular cavity is enclosed in the myocardium, the network only needed to learn to delineate the myocardium and could then infer the shape of the left ventricular cavity.

The ASSD graph against slice level indicated that performance dropped for slices near the heart base or apex, especially for the right ventricular cavity. Although results for this structure were lower throughout the volume, the increase in ASSD was more pronounced and appeared earlier towards the apex, which may stem from the structure size reducing quicker and its complex and individually-variable geometry, thus making it more difficult

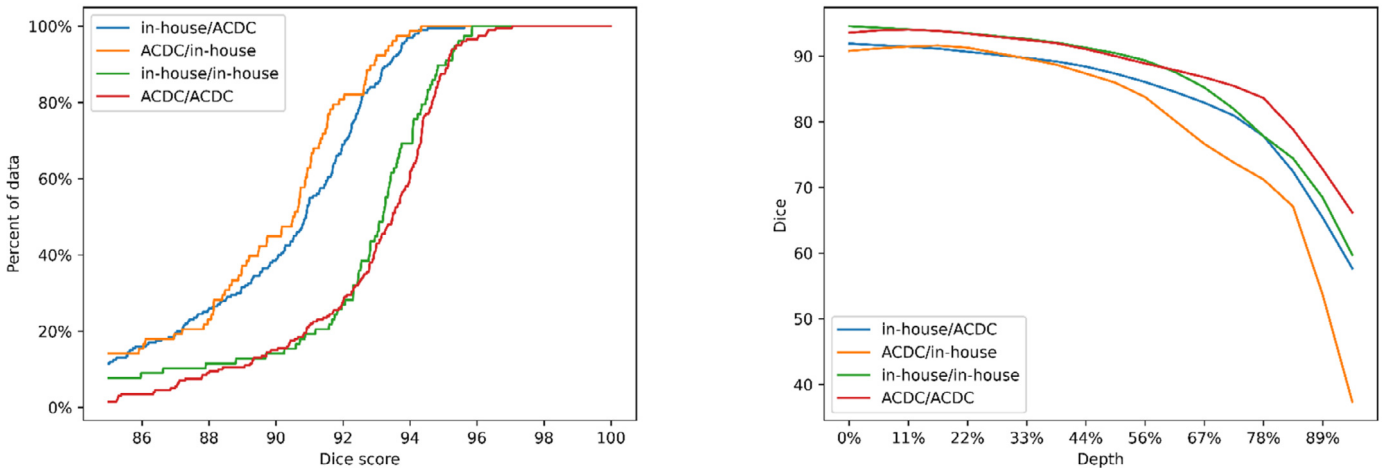


Fig. 9. Generalization performance analysis for models trained/tested on ACDC or the in-house dataset. (a) Cumulative frequency plot of Dice scores; (b) Dice scores of individual slices against relative depth in volumes. Results were interpolated over the maximum number of slices in a volume. 0% is the most basal slice and 100% the most apical one.

Table 6
Effect of training only on the end-diastole and end-systole frame. The model trained on all phases of the cardiac cycle of the in-house dataset is compared with the same model but trained only on the ED and ES frames. Dice scores, ASSDs and HDs are reported as mean ± standard deviation (in mm). RV = right ventricular cavity, MYO = *myocardium*, *LV* = left ventricular cavity.

	Metric	All Phases	Only ED and ES
Segmentation	Mean Dice	93.21 ± 1.58	92.98 ± 1.70
	Mean IOU	87.46 ± 2.63	87.08 ± 2.81
	Mean ASSD (mm)	0.14 ± 0.08	0.15 ± 0.09
	Mean HD (mm)	4.12 ± 1.20	4.27 ± 1.21
LV Radial strain	ES peak value correlation	0.72	0.63
	ES peak value mean bias [LoA]	1.22 [-28.75; 31.16]	6.317 [-29.35; 41.99]
	ES peak value relative mean bias (%)	1.97	10.19
	ES peak index correlation	0.84	0.84
	ES peak index mean bias [LoA]	0.07 [-3.49; 3.64]	0.21 [-3.41; 3.82]
	ES peak index relative mean bias (%)	0.54	1.52
LV Circumferential strain	ES peak value correlation	0.84	0.76
	ES peak value mean bias [LoA]	-0.273 [-4.00; 3.45]	-0.663 [-5.14; 3.81]
	ES peak value relative mean bias (%)	1.45	3.52
	ES peak index correlation	0.90	0.88
	ES peak index mean bias [LoA]	-0.10 [-3.01; 2.82]	-0.01 [-3.16; 3.14]
	ES peak index relative mean bias (%)	-0.70	-0.09
RV Circumferential strain	ES peak value correlation	0.57	0.12
	ES peak value mean bias [LoA]	-0.74 [-13.03; 11.54]	-0.75 [-34.91; 33.41]
	ES peak value relative mean bias (%)	4.92	4.97
	ES peak index correlation	0.82	0.87
	ES peak index mean bias [LoA]	-0.17 [-6.18; 5.84]	0.34 [-4.93; 5.62]
	ES peak index relative mean bias (%)	-1.17	2.34

to segment for the network. This result aligns with the findings of [16] who identified similar challenges with slices located at both ends of volumes, noting that this issue also occurs for clinical experts. When it came to basal slices, segmentation errors were again mainly present for the right ventricular cavity, where the annotations for the right ventricular cavity may seem inconsistent to the network, because of the presence of other structures such as out-flow tract. Acquisitions, being conventionally aligned on the LV axis, slice obliquity for the RV and the presence of the pulmonary artery may lead to a right ventricle in two parts fusing in lower slices. This may have led to difficulties for the network to identify the first slice in the stack from which the right ventricular cavity should be segmented. This was also confirmed by visual results of worst segmented slices since most of them were located near the base with a right ventricular cavity often appearing half broken and as a result not annotated by the expert for being too basal. This was also reflected in computed correlations for the RV ejection fraction and stroke volume which were lower than for other structures. Other clinical quantitative indices were in the same range as those previously reported in the literature [30].

Generalization performance was satisfying with a limited decrease in performance when testing on a different dataset the model was trained on. Indeed, the model trained on ACDC showed a small decrease in performance when tested on the in-house dataset with an increase in the average and maximum distance between predicted and ground truth contours of less than a pixel (increase in ASSD and HD of 0.22 mm and 1.19 mm respectively

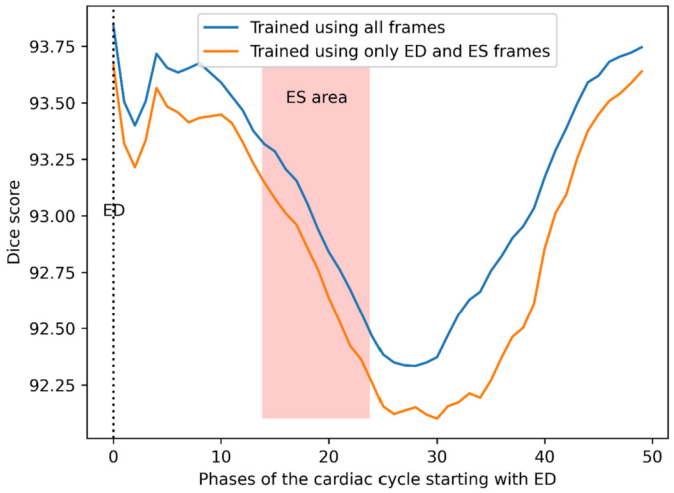


Fig. 10. Average Dice score against the phase in the cardiac cycle for our model trained on all phases and our model trained only with the ED and ES frames. Results are interpolated over the maximum length of a sequence in the testing set.

with average pixel size of 1.43 mm2). The model trained on the in-house dataset demonstrated satisfactory segmentation performance on the ACDC dataset. However, for this model, the predicted label for the RV was larger than both the ground truth and the prediction from the model trained on ACDC. This discrepancy may result from different annotation conventions for this cavity between the two datasets. Performance on out-of-distribution data was robust for most slices in volumes but decreased in the most apical slices, which is consistent with [31,32]. This can also be explained by the difference in annotation conventions since the ACDC dataset contains apical slices with no annotation for all classes while our in-house dataset always contains at least one class.

When it comes to performance across the cardiac cycle, our model trained only on the ED and ES frames performed only slightly worse than the model trained on all frames. However, the gap in Dice scores between the two models seems to widen as the distance relative to the ED and ES frame increases. This shows that while the drop in performance is limited, only training with the ED and ES frames yields lower results than training with all frames, especially as the distance relative to these two frames increases. It is worth noting, however, that the drop in performance affects all frames of the cardiac cycle, suggesting that training with more frames improves performance for all phases of the cardiac cycle. This likely results from the increased diversity in the training data. As manually annotating each slice of the cardiac cycle is time-consuming, our findings encourage the use of automatic or semi-automatic software able to provide ground truth segmentation across the cardiac cycle. Besides, the most noticeable consequences on strain metrics indicate that average segmentation metrics are limited in the ability to convey the practical clinical useability of an algorithm.

4.1. Limitations

Our method has some limitations. First, the dataset used to compare performances of the model trained on all frames and the model trained only with the ED and ES frames contains segmentation labels only for 3 slices in the volume. As a result, it was not possible to compute volumetric indices for the whole cardiac cycle. Moreover, these two models were not trained with any temporal information and were applied to each frame of the cardiac cycle independently at inference. Therefore, there is likely room for improvement, especially for strain metrics.

Another limitation revolved around the segmentation of the most basal or apical slices. For the most basal slices, this might be linked to the right ventricle shape which differs strongly from its shape in intermediate slices, while in the most apical slices, some structures might be absent leading to important distance values if the network predicted their presence. Of note, in the ACDC dataset, the heart was completely absent in some slices and would probably have been excluded from the analysis in a clinical setting.

Finally, although the network showed satisfying results when generalizing to other datasets, no explicit domain adaptation technique was used. This could be explored in future work.

5. Conclusion

A new deep learning network architecture relying on spatial attention was introduced to segment the cardiac structures from short-axis cine-MRI on two different datasets. The model showed satisfying generalization ability although there is still room for improvement for the most apical slices. Computed volumetric indices were close to ground truth indices and in line with literature showing the algorithm could be used in a medical context for assisted diagnosis. Using all phases of the cardiac cycle rather than only the ED and ES ones leads to an important jump in strain accuracy and slight gain in segmentation performance, encouraging the use of tools able to provide ground truth annotations for the whole cardiac cycles.

Human and animal rights

The authors declare that the work described has been carried out in accordance with the Declaration of Helsinki of the World Medical Association revised in 2013 for experiments involving humans as well as in accordance with the EU Directive 2010/63/EU for animal experiments.

Informed consent and patient details

The authors declare that they obtained a written informed consent from the patients and/or volunteers included in the article and that this report does not contain any personal information that could lead to their identification.

Funding

This work was funded by the grant number 965286 from the H2020 MAESTRIA project.

This work was performed using HPC resources from GENCI-IDRIS (Grant 20XX-AD011013634R1).

Author contributions

All authors attest that they meet the current International Committee of Medical Journal Editors (ICMJE) criteria for Authorship.

Declaration of competing interest

The authors declare that they have no known competing financial or personal relationships that could be viewed as influencing the work reported in this paper.

Acknowledgement

This work was funded by the grant number 965286 from the H2020 MAESTRIA project.

This work was performed using HPC resources from GENCI-IDRIS (Grant 20XX-AD011013634R1).

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.irbm.2024.100850.

References

[1] Rajiah PS, François CJ, Leiner T. Cardiac MRI: state of the art. Radiology 2023;307:e223008. https://doi.org/10.1148/radiol.223008.

[2] Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: Int. conf. med. image comput. comput.-assist. interv. Springer; 2015. p. 234–41.

[3] Wang X, Wang L, Zhong X, Bai C, Huang X, Zhao R, et al. Pal-net: a modified u-net of reducing semantic gap for surgical instrument segmentation. IET Image Process 2021;15:2959–69. https://doi.org/10.1049/ipr2.12283.

[4] Cao Z, Ma C, Wang Q, Zhang H. Relay-UNet: reduce semantic gap for glomerular image segmentation. In: Shi Z, Jin Y, Zhang X, editors. Intell. sci. IV. Cham: Springer International Publishing; 2022. p. 378–85.

[5] Wang H, Chen X, Yu R, Wei Z, Yao T, Gao C, et al. E-DU: deep neural network for multimodal medical image segmentation based on semantic gap compensation. Comput Biol Med 2022;151:106206. https://doi.org/10.1016/j.compbiomed.2022.106206.

[6] Li X, Zhao H, Han L, Tong Y, Tan S, Gated Yang K. Fully fusion for semantic segmentation. Proc AAAI Conf Artif Intell 2020;34:11418–25. https://doi.org/10.1609/aaai.v34i07.6805.

[7] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: transformers for image recognition at scale. In: 9th int. conf. learn. represent.; 2021.

[8] Chen J, Lu Y, Yu Q, Luo X, Adeli E, Wang Y, et al. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. 2021.

[9] Hatamizadeh A, Tang Y, Nath V, Yang D, Myronenko A, Landman B, et al. UNETR: Transformers for 3D Medical Image Segmentation. 2021.

[10] Wang H, Xie S, Lin L, Iwamoto Y, Han X-H, Chen Y-W, et al. Mixed transformer u-net for medical image segmentation. ICASSP 2022:2390–4. https://doi.org/10.1109/ICASSP43922.2022.9746172.

[11] Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin transformer: hierarchical vision transformer using shifted windows. In: 2021 IEEECVF int. conf. comput. vis. ICCV; 2021. p. 9992–10002.

[12] Cao H, Chen J, Wang Y, Jiang D, Zhang X, Tian Q, et al. Swin-unet: unet-like pure transformer for medical image segmentation. In: Karlinsky L, Michaeli T, Nishino K, editors. Comput. vis. – ECCV 2022 workshop. Cham: Springer Nature Switzerland; 2023. p. 205–18.

[13] Zhou H-Y, Guo J, Zhang Y, Yu L, Wang L, Yu Y. nnFormer: interleaved transformer for volumetric segmentation. arXiv:2109.03201 [cs.CV], 2022.

[14] Wang H, Cao P, Wang J, Zaiane OR. UCTransNet: rethinking the skip connections in U-Net from a channel-wise perspective with transformer. arXiv:2109.04335 [cs.CV], 2021.

[15] Petit O, Thome N, Rambour C, Themyr L, Collins T, Soler L. U-net transformer: self and cross attention for medical image segmentation. In: Lian C, Cao X, Rekik I, Xu X, Yan P, editors. Mach. learn. med. imaging. Cham: Springer International Publishing; 2021. p. 267–76.

[16] Bernard O, Lalande A, Zotti C, Cervenansky F, Yang X, Heng P-A, et al. Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? IEEE Trans Med Imaging 2018;37:2514–25. https://doi.org/10.1109/TMI.2018.2837502.

[17] Montalescot G, Alexander JH, Cequier-Fillat A, Solomon SD, Redheuil A, Hudec M, et al. Firibastat versus ramipril after acute mechanical reperfusion of anterior myocardial infarction: a phase 2 study. Am J Cardiovasc Drugs Drugs Devices Interv 2023;23:207–17. https://doi.org/10.1007/s40256-023-00567-8.

[18] Lamy J, Soulat G, Evin M, Huber A, de Cesare A, Giron A, et al. Scan-rescan reproducibility of ventricular and atrial MRI feature tracking strain. Comput Biol Med 2018;92:197–203. https://doi.org/10.1016/j.compbiomed.2017.11.015.

[19] Myronenko A. 3D MRI brain tumor segmentation using autoencoder regularization. In: Crimi A, Bakas S, Kuijf H, Keyvan F, Reyes M, van Walsum T, editors. Brainlesion glioma mult. scler. stroke trauma. brain inj. Cham: Springer International Publishing; 2019. p. 311–20.

[20] Hendrycks D, Gimpel K. Bridging nonlinearities and stochastic regularizers with Gaussian error linear units. CoRR. arXiv:1606.08415 [abs], 2016.

[21] Lu L, Shin Y, Su Y, Karniadakis GE. Dying ReLU and initialization: theory and numerical examples. Commun Comput Phys 2020;28:1671–706. https://doi.org/10.4208/cicp.OA-2020-0165.

[22] Arnekvist I, Carvalho JF, Kragic D, Stork JA. The effect of Target Normalization and Momentum on Dying ReLU. 2020.

[23] Xiao T, Singh M, Mintun E, Darrell T, Dollar P, Girshick R. Early convolutions help transformers see better. In: Ranzato M, Beygelzimer A, Dauphin Y, Liang PS, Vaughan JW, editors. Adv. neural inf. process. syst., vol. 34. Curran Associates Inc.; 2021. p. 30392–400.

[24] Dai Z, Liu H, Le QV, Tan M. CoAtNet: marrying convolution and attention for all data sizes. In: Ranzato M, Beygelzimer A, Dauphin Y, Liang PS, Vaughan JW, editors. Adv. neural inf. process. syst., vol. 34. Curran Associates Inc.; 2021. p. 3965–77.

[25] Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods 2021;18:203–11. https://doi.org/10.1038/s41592-020-01008-z.

[26] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al., editors, Adv. neural inf. process. syst., vol. 30. Curran Associates Inc.; 2017.

[27] Vigneault DM, Xie W, Ho CY, Bluemke DA, Noble JA. Ω-net (omega-net): fully automatic, multi-view cardiac MR detection, orientation, and segmentation with deep neural networks. Med Image Anal 2018;48:95–106. https://doi.org/10.1016/j.media.2018.05.008.

[28] Raghu M, Unterthiner T, Kornblith S, Zhang C, Dosovitskiy A. Do vision transformers see like convolutional neural networks? In: Ranzato M, Beygelzimer A, Dauphin Y, Liang PS, Vaughan JW, editors. Adv. neural inf. process. syst., vol. 34. Curran Associates Inc.; 2021. p. 12116–28.

[29] Tuli S, Dasgupta I, Grant E, Griffiths TL. Are convolutional neural networks or transformers more like human vision? CoRR, arXiv:2105.07197 [abs], 2021.

[30] Kawel-Boehm N, Hetzel SJ, Ambale-Venkatesh B, Captur G, Francois CJ, Jerosch-Herold M, et al. Reference ranges ("normal values") for cardiovascular magnetic resonance (CMR) in adults and children: 2020 update. J Cardiovasc Magn Reson 2020;22:87. https://doi.org/10.1186/s12968-020-00683-3.

[31] Campello VM, Gkontra P, Izquierdo C, Martín-Isla C, Sojoudi A, Full PM, et al. Multi-centre, multi-vendor and multi-disease cardiac segmentation: the M&Ms challenge. IEEE Trans Med Imaging 2021;40:3543–54. https://doi.org/10.1109/TMI.2021.3090082.

[32] Martín-Isla C, Campello VM, Izquierdo C, Kushibar K, Sendra-Balcells C, Gkontra P, et al. Deep learning segmentation of the right ventricle in cardiac MRI: the M&Ms challenge. IEEE J Biomed Health Inform 2023;27:3302–13. https://doi.org/10.1109/JBHI.2023.3267857.