Deep learning for automated analysis of cardiac imaging: applications in Cine and 4D flow MRI

Sun, X.

Chapter 5 Deep learning based automated left ventricle segmentation and flow quantification in 4D flow cardiac MRI

This chapter was adapted from:

Xiaowu Sun, Li-Hsin Cheng, Sven Plein, Pankaj Garg, Rob J. van der Geest. **Deep learning-based automated left ventricle segmentation and flow quantification in 4D flow cardiac MRI.** Journal of Cardiovascular Magnetic Resonance.(under review)

Abstract

Background: 4D flow MRI enables assessment of cardiac function and intra-cardiac blood flow dynamics from a single acquisition. However, due to the poor contrast between the chambers and surrounding tissue, quantitative analysis relies on the segmentation derived from a registered cine MRI acquisition. This requires an additional acquisition and is prone to imperfect spatial and temporal inter-scan alignment. Therefore, in this work we developed and evaluated deep learning-based methods to segment the left ventricle from 4D flow MRI directly.

Methods: We compared five deep learning-based approaches with different network structures, data pre-processing and feature fusion methods. For the data pre-processing, the 4D flow MRI was reformatted into a stack of short-axis view slices. Two feature fusion approaches were proposed to integrate the features from magnitude and velocity images. The networks were trained and evaluated on an in-house dataset of 103 subjects with 69,619 2D images and 3090 3D volumes. The performance was evaluated using various metrics including Dice, average surface distance (ASD), end-diastolic volume (EDV), end-systolic volume (ESV), left ventricular ejection fraction (LVEF), kinetic energy (KE) and flow components. The Monte Carlo dropout method was used to assess the confidence and to describe the uncertainty area in the segmentation results.

Results: Among the five models, the model combining 2D U-Net with late fusion method operating on short-axis reformatted 4D flow volumes achieved the best results with Dice of 84.51% and ASD of 3.13 mm. The averaged absolute error between manual and automated segmentation for EDV, ESV, LVEF and normalized KE was 20.27 ml, 17.21 ml, 7.41% and 0.54 μJ/ml, respectively. Flow component results derived from automated segmentation showed high correlation and small average error compared to results derived from manual segmentation.

Conclusions: Deep learning-based methods can achieve accurate automated LV segmentation and subsequent quantification of volumetric and hemodynamic LV parameters from 4D flow MRI without requiring an additional cine MRI acquisition.

5.1 Background

Four-dimensional flow magnetic resonance imaging (4D flow MRI) provides time-resolved three-dimensional imaging of cardiac geometry and multi-directional intra-cardiac blood flow velocity from a single acquisition [1]. Several quantitative left ventricular (LV) hemodynamic parameters can be derived from the acquired data, including intra-cardiac kinetic energy (KE), vorticity and functional flow components [2, 3]. Quantitative assessment of these parameters relies on accurate segmentation of the LV cavity. However, the contrast between the blood pool and the surrounding tissue is typically extremely poor in the acquired magnitude images of a 4D flow acquisition. For this reason, the segmentation is usually performed using the images of an additionally acquired balanced Steady State Free Precession (b-SSFP) cine MR acquisition [4, 5]. Based on the known spatial relation between the two acquisitions, the obtained segmentation can be transferred to the domain of the 4D flow acquisition. Unfortunately, due to breath-hold inconsistency and differences in heart rate, the cine MR images are prone to a spatial and temporal misalignment resulting in sub-optimal segmentation of the 4D flow acquisition. Therefore, it would be advantageous when the segmentation could be performed directly from the 4D flow acquisition, not requiring any additional acquisition.

Bustamante proposed a multi-atlas registration method to automatically generate a segmentation of the entire thoracic cardiovascular system using eight 3D phase-contrast MR angiogram volumes as atlases [6]. A disadvantage of this approach is the high computational cost of the required image registration. In recent years, deep learning-based segmentation methods have been proposed and achieved immense success in medical image segmentation tasks. U-Net, consisting of a contracting and expanding path, has demonstrated excellent performance in segmentation of MR imaging data of the heart, brain and various other organs [7]. Benefiting from these convolutional neural networks (CNNs), a few studies reported the use of deep learning for the segmentation of 4D flow MRI. Berhane et al. developed a 3D U-Net with DenseNet-based dense blocks to segment the aortic arch from 4D flow MRI [8]. Based on U-Net and attention gate mechanism, Wu demonstrated that incorporating the information from the combination of magnitude and velocity images results in improved performance in LV myocardium segmentation in 4D myocardial velocity mapping MRI [9]. Corrado et al. applied a fine-tuned CNN model trained on cine b-SSFP MRI data and used registration to derive segmentation of the 4D flow MRI [10]. However, this approach relies on the availability of a cine MRI acquisition. Bustamante et al. recently reported a 3D U-Net based method for segmentation of the cardiac chambers and great thoracic vessels directly from 4D flow MRI magnitude images, ignoring the velocity images [11]. An excellent geometric agreement with manual segmentation results was reported (DICE $\geq$ 0.9) and also good agreement of the derived quantitative results, such as end-diastolic (ED) and

and-systolic (ES) volumes and blood flow kinetic energy. However, since the employed 4D flow acquisition was acquired directly after gadolinium contrast administration it remains unknown whether the presented method performs equally well on non-contrast-enhanced imaging data. A CNN-based segmentation method that takes both magnitude and velocity images as input may yield better segmentation performance than one only using magnitude images.

Our main contributions are summarized as follows: (1) We evaluated multiple strategies to take advantage of the magnitude and velocity images of the 4D flow MRI acquisition. (2) We compared the performance of five different U-Net-based networks. (3) We used a Monte Carlo dropout method to evaluate the segmentation uncertainty of the implemented CNN models.

5.2 Methods

5.2.1 Study cohort and imaging protocol

The dataset used in the study included 103 subjects, including 75 post-myocardial infarction (MI) patients (15 females, 60 males; mean age 69±12 years, range 40-94) and 28 healthy volunteers (11 females, 17 males; mean age 48±17, range 23-80). The study was approved by the local medical ethical committee of the University of Leeds, UK, and all participants provided written informed consent. All subjects underwent a comprehensive cardiac MR imaging protocol on a 1.5T MR system (Philips Healthcare), including cine MR imaging in standard cardiac views and 4D flow MR with whole-heart coverage.

A short-axis cine stack was acquired with a slice thickness of 8-10 mm and an inter-slice gap of 2 mm using 10-17 slices to cover the LV from the apex to the base. Imaging was performed during breath-holding in end-expiration. Other imaging parameters were a field of view (FOV) 300×300 mm^2 to 470×470 mm^2, pixel spacing 0.83-1.19 mm, echo time (TE) 1.27-1.62 ms, repetition time (TR) 2.55-3.25 ms. Using retrospective gating 30 phases were reconstructed to cover a full cardiac cycle. 4D flow MRI was acquired using an echo-planar imaging (EPI) accelerated sequence with retrospective electrocardiogram gating during free-breathing without using respiratory motion compensation. The 3D volume of the acquisition was planned in an oblique orientation with a voxel size of 3×3×3 mm^3 , a field of view of 370-400 × 370-400 mm^2 and 33-52 reconstructed slices to cover the whole heart. The orientation of the acquired 3D volume varied from subject to subject and was adjusted such as to encompass the complete heart and proximal aorta using a minimal number of slices. The number of reconstructed cardiac phases was 30. Other scan parameters of the 4D flow MRI acquisition were TE 1.9-3.8 ms, TR 4.8-13.9 ms, flip angle 10˚ and velocity encoding (VENC) 150 cm/s. A more detailed description of the scan parameters can be found in previous work [12]. In patients, the 4D flow

acquisition was added to a regular clinical scan protocol, including late-gadolinium enhancement (LGE) imaging. Typically, the 4D flow acquisition was obtained post-contrast (Magnevist, 0.2 mmol/kg) in the waiting period between contrast administration and LGE imaging.

5.2.2 Ground truth generation

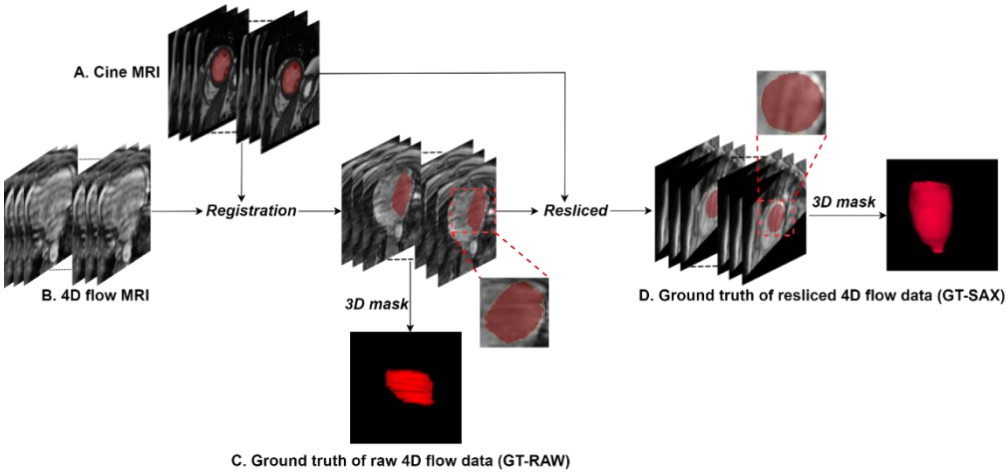

Figure.5.1. The procedure of ground truth generation. A: The mask of left ventricle was first annotated in the short-axis cine MRI. B,C: it was propagated to original 4D flow MRI using rigid registration method. D: Given the orientation of short-axis cine MRI, the raw 4D flow MRI was resliced into short-axis view.

One experienced observer semi-automatically defined the LV endocardial contours in all slices and phases of the short-axis cine stack using in-house developed Mass research software (Version V2017-EXP; Leiden University Medical Center, Leiden, the Netherlands). Following SCMR recommendations, papillary muscles and trabeculations were included within the defined contours in order to derive a consistent and time-continuous segmentation of the LV geometry. Correction for spatial misalignment, resulting from patient movement between the cine MR and 4D flow acquisition, was performed using rigid registration using Elastix software as previously described [13, 14]. Subsequently, we generated two types of LV blood pool masks for the 4D flow MRI acquisition. The first type of mask, further labelled as RAW, was generated by labelling the pixels of the original slices of the 4D flow acquisition as either blood pool or background according to the nearest labelled pixel in the short-axis cine acquisition. Due to the relatively low through-plane resolution of the short-axis stack and the varying orientation of the acquired 4D flow volumes, the resulting RAW masks frequently suffer from jagged boundaries and are less smooth compared to the original contours as defined in the short-axis stack. Therefore a second type of mask, further labelled as SAX, was generated by reformatting the volume of the 4D flow acquisition into a stack of short-axis slices.

Given the known short-axis orientation, the original 4D flow acquisition was resliced into a short-axis view using a slice spacing of 3 mm and a fixed number of 41 slices. The in-plane resolution was chosen to be equal to that of the cine short-axis stack and ranged from 0.83×0.83 mm2 to 1.19×1.19 mm2. Subsequently, the SAX mask was generated by labelling the pixels in the reformatted 4D flow images as either blood pool or background, following the same approach as for the RAW mask. The resulting blood pool regions are more smooth compared to the RAW mask regions and vary less in shape since all masks are defined in short-axis orientation. Accordingly, two ground truths are available for training and testing: GT-RAW for the original 4D flow data and GT-SAX for the resliced 4D flow data. Figure.5.1 describes the procedure of the ground truth generation, illustrating the more irregular GT-RAW masks compared to the GT-SAX masks. After excluding the images without LV, the dataset contained 90,313 SAX 2D image pairs, 69,619 RAW 2D image pairs and 3,090 (103×30) 3D volumes.

5.2.3 Networks

Table.5.1. Different methods with different networks and inputs. SAX indicates that the resliced data in the short-axis view was used as input to the network. RAW indicates that the raw 4D flow data was used as input.

Method	Input orientation	Ground truth	Network	Input Size	Output Size
SAX2D	SAX	GT-SAX	2D U-Net	(256,256,4)	(256,256,1)
RAW2D	RAW	GT-RAW	2D U-Net	(256,256,4)	(256,256,1)
SAX3D	SAX	GT-SAX	3D U-Net	(256,256,40,4)	(256,256,40,1)
RAW3D	RAW	GT-RAW	3D U-Net	(256,256,32,4)	(256,256, 32,1)
SAX2DF	SAX	GT-SAX	2D Fusion Network	(256,256,1), (256,256,3)	(256,256,1)

We compare five deep learning models to investigate the effect of data preprocessing, information fusion strategies and network structures on the segmentation performance. The five proposed methods are summarized in Table.5.1. RAW and SAX represent the two different input orientations. RAW used the original 4D flow data to train the network, either as a 3D volume, or as individual 2D slices and SAX used 4D flow data resliced into the short-axis orientation. Each 2D slice was center-cropped to a fixed size of 256×256. The number of slices in the 3D volume of the original 4D flow data varies from 33 to 52 The middle 32 slices in the original 4D flow data are stacked as the input of RAW3D. In the resliced dataset with fixed number of 41 slices, the last 40 slices are stacked as the input of SAX3D after excluding the first slice, resulting in an even number of spatial input dimension, which is convenient for the repeated down-sampling operations with a factor of 2.

SAX2D and RAW2D models are adapted from 2D U-Net, an encoder-decoder CNN model with long-skip connections. The architecture includes five-scaled resolutions. Each level contains two convolutional blocks composed of a convolution layer with kernel size of 3×3 followed by a instance normalization (IN) layer, a rectified linear unit (ReLU) and one dropout layer. In the encoder feature maps are down-sampled by a max-pooling layer with kernel size of 2×2, while in the decoder transposed convolution layers are used to increase the resolution to its original scale. The long-skip connections are used to concatenate the features from fine to coarse scales at each level. Finally, a convolution layer with kernel size of 1×1, followed by a Sigmoid function, is used to generate the probability map. The final segmentation results are determined by choosing the class with the highest probability at each pixel.

RAW3D and SAX3D models employ a 3D U-Net architecture, which is used to investigate the performance of varying volumetric inputs. The 3D volume generated from each phase is considered as an independent input of 3D U-Net. Compared to 2D U-Net, the kernel size of all convolution layers in 3D U-Net is set to 3×3×3. The 3D U-Net introduced four max-pooling layers for the down-sampling operations. The kernel size of all pooling layers in RAW3D are set to 2×2×2. Whereas in SAX3D the first three pooling layers are set to 2×2×2 and last pooling layer is set to 2×2×1 because the spatial dimension will be reduced to 5 after three down-sampling operations.

Magnitude and velocity images can be considered as different modalities providing different information for the segmentation. To fuse the information from these two modalities, we introduce two approaches named early fusion and late fusion, respectively. SAX2D, SAX3D, RAW2D and RAW3D use the early fusion method where the magnitude and velocity images are concatenated along the channel dimension as the input. Whereas SAX2DF uses the late fusion method. As illustrated in Figure.5.2, separate encoders are used to extract the features from these two modalities. Thereafter, features in the same level from two encoders are added together. The aggregated features in the bottleneck are up-sampled to the original resolution. The other multi-scale aggregated feature maps are then concatenated with the features up-sampled from the lower level. The structure of decoder used in SAX2DF is the same as that in 2D U-Net.

Dice loss and cross-entropy were jointly used as the loss function to train the models. All the experiments were implemented using Pytorch with the following parameters: batch size=50; learning rate=0.0001; optimizer=Adam. Five-fold cross-validation was applied to assess the performance and the averaged values are reported. All the experiments were implemented on a machine equipped with an NVIDIA Quadro RTX 6000 GPU with 24 GB internal memory.

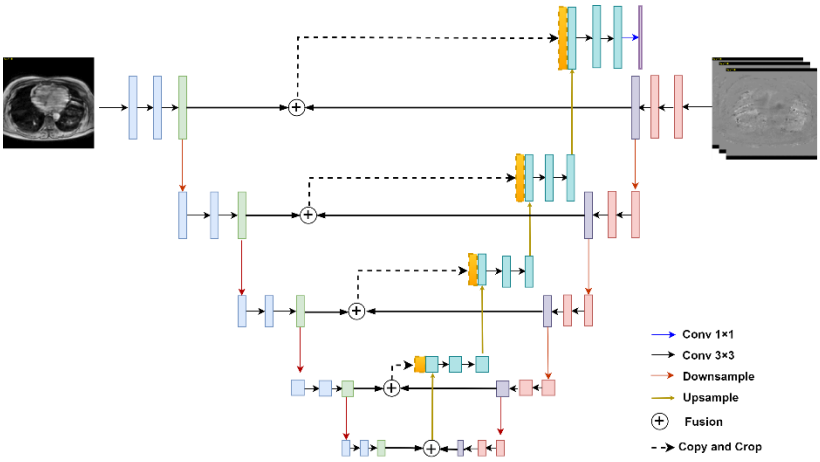

Figure.5.2. The network architecture of SAX2DF. SAX2DF separates magnitude and the three velocity images as two inputs and uses two encoders to extract the features from each input. The late fusion method us used to integrate those features.

5.2.4 Evaluation metrics

The performance of the automated methods was evaluated using segmentation accuracy, uncertainty score and volumetric and flow related clinical metrics.

Segmentation Accuracy. Dice and average surface distance (ASD) were used to assess the segmentation accuracy. Dice measures the overlap between the prediction and the ground truth. ASD is the average of all the distances from all surface points on the boundary of the predicted region to the boundary of the ground truth, which can be described as formula (5.1)

$$ASD = \frac{1}{n_S + n_{S'}} \left(\sum_{p=1}^{n_S} d(p, S') + \sum_{p'=1}^{n_{S'}} d(p', S) \right) \tag{5.1}$$

where $d(p, S') = \min_{p' \in S'} \| p - p' \|_2$ is the minimum of the Euclidean distance between a point p on surface S and the surface S'. Dice and ASD reported in this work are computed based on each 3D volume and averaged over all phases.

Clinical metrics. End-diastolic volume (EDV), end-systolic volume (ESV), LV ejection fraction (LVEF) and kinetic energy (KE) were derived as clinical metrics. The KE was computed as formula 5.2 with ρ_{blood} being the density of the blood (1.06g/cm^3), V_{voxel} the voxel volume and v the velocity magnitude of one voxel. The

total KE is the summation of the KE of each voxel within the LV region. The total KE values were indexed for LV EDV and averaged over the complete cardiac cycle.

$$KE = \frac{1}{2}\rho_{blood} \cdot V_{voxel} \cdot v^2 \tag{5.2}$$

Additionally, three phasic KE parameters were derived: peak systolic, peak E-wave and peak A-wave KE. The result of LV segmentation was also used for LV flow component analysis. Based on previously described methods by Eriksson *et al* the segmented LV blood pool at the ED moment was used to define seeding particles of size 3×3×3 mm^3 and particle pathlines were derived using particle tracing in forward (until the next ES moment) and backward (until the previous ES moment) direction [4]. The particle position at the two ES moments was then used to classify the defined pathlines as either direct flow (DIR), delayed ejection flow (DEL), retained inflow (RET) or residual volume (RES). The relative size of each flow component was expressed as a percentage of the ED volume. The clinical parameters derived from automated LV segmentation were compared to the results derived of manual segmentation.

Uncertainty Score. Segmentation of anatomical structures is inherently ambiguous especially near an object border which is not clearly defined due to the poor contrast or restriction imposed by the image acquisition. The uncertainty score can give some insights into the confidence of a model in its predicted segmentation results [15]. In case of a high uncertainty score, it is more likely that the segmentation result is inaccurate. Usually, a CNN model only produces a single segmentation map without any information to explain its confidence in its prediction. A high probability value in a segmentation map doesn't imply a high confidence score. A model also can be uncertain in pixels with high probability. In order to investigate the segmentation uncertainty of the different models we applied the Monte Carlo (MC) dropout method [16] to quantify the model's confidence in the segmentation result.

Generally during the testing phase, the dropout layers in the network are removed. The uncertainty score can be derived by preserving the dropout layers during testing while executing multiple inference runs. In our experiments the drop rate in the middle-level dropout layers was set to 0.5 and the testing was repeated 20 times resulting in 20 predictions denoted as $P_i(i = 1, 2, \cdots 20)$. The uncertainty score can be derived using equation 5.3 where $P = \frac{1}{20}\sum_{i=1}^{20} P_i$.

$$UQ = -P \times \log_2 P - (1-P) \times \log_2 (1-P) \tag{5.3}$$

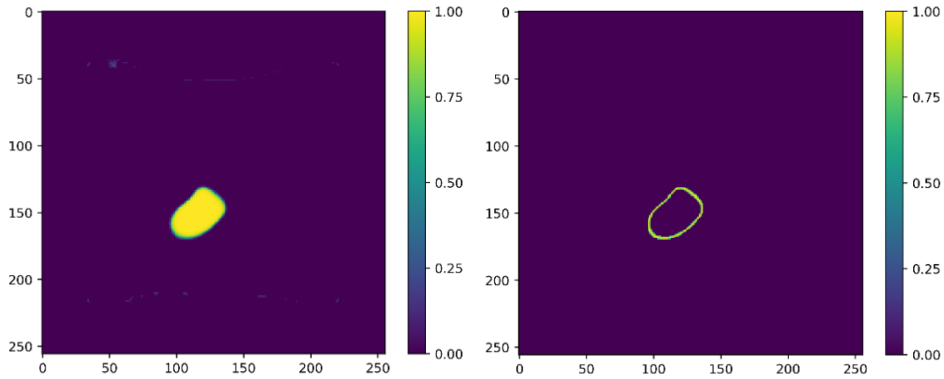

Figure.5.3. An example of segmentation probability and its corresponding uncertainty map. **Left**: Probability map derived from the last layer of RAW2D. **Right**: Corresponding uncertainty map derived from MC method.

Figure.5.3 shows an example of a segmentation probability map and its corresponding uncertainty map. The uncertainty score for the pixels within the LV chamber is low, implying a high confidence of the models' prediction, but due to the poor contrast between the heart chamber and myocardium the uncertainty near the ambiguous LV border with a corresponding probability varying from 0.4 to 0.6 is substantially higher. To compute the mean of uncertainty and to quantify the segmentation quality, we first computed the uncertainty score for the whole LV chamber where each pixel's prediction probability is larger than 0.5. Then to highlight the higher uncertainty in the boundary region, we further computed the score for this area with a prediction probability ranging from 0.4 to 0.6.

Statistical analysis. The correlation of the clinical metrics derived from the manual and predicted segmentation results were assessed using the Pearson correlation coefficient (PCC). Additionally, bias and limits of agreement (LOA, 1.96×standard deviation) were used to describe the agreement of prediction and ground truth.

5.3 Results

First, we compared the results derived by the five models on various evaluation metrics. Second, we explored the impact of the fusion methods on the uncertainty score. Lastly, we investigated the performance of the best model on the KE and flow components.

5.3.1 Segmentation results

Table.5.2 summarizes the segmentation performance derived from different models. SAX2DF achieved the best results with Dice of 84.51%, ASD of 3.13 mm and absolute error of ESV and LVEF of 17.21 ml, 7.41%, respectively. The best results in an absolute error of EDV and KE were obtained using the model of SAX2D, yielding an error of 19.96 ml and 0.41 μJ/ml, respectively.

Table.5.2. Segmentation performance derived from different methods. The reported clinical results are the error between the ground truth and prediction. The best results are shown in bold. KE was normalized to the EDV. SAX2D-M and SAX3D-M represent models that only use the magnitude images as the input. The other methods use both magnitude and velocity images as the input

Method	Dice(%)	ASD(mm)	EDV(ml)	ESV(ml)	LVEF(%)	KE(μJ/ml)
SAX2D	84.33±6.28	3.30±1.89	**19.96±22.08**	19.35±16.04	9.54±7.17	**0.41±0.39**
RAW2D	79.97±7.38	4.01±1.86	29.32±22.91	32.35±24.33	10.84±8.58	1.16±1.18
SAX3D	82.84±6.65	4.41±4.68	22.47±19.19	23.21±22.12	10.07±8.89	0.84±1.02
RAW3D	79.77±7.56	3.67±1.64	24.25±19.56	26.91±24.04	10.58±8.01	0.91±0.85
SAX2DF	**84.51±6.58**	**3.13±2.33**	20.27±20.31	**17.21±16.03**	**7.41±6.07**	0.54±0.51
SAX2D-M	81.46±7.39	4.10±2.91	26.20±23.93	32.65±22.04	21.43±11.03	0.88±2.17
SAX3D-M	80.61±0.09	5.48±5.23	24.53±18.75	38.47±31.27	18.97±16.48	1.01±0.93

Due to the different ground truth masks used, a direct comparison of the performance using Dice and ASD derived from RAW and SAX data is not easily possible. Therefore, also clinical parameters were used to compare the performance of the models. Table.5.2 shows that SAX2D outperformed RAW2D and SAX3D performed better than RAW3D in all clinical metrics including EDV, ESV, LVEF and KE, demonstrating the models using images in short-axis view orientation can generate a better prediction.

We further compared the results derived from only using magnitude images (SAX2D-M and SAX3D-M) and combining magnitude and velocity images. The comparison was restricted to the models using the short-axis view data, since these models provided the best performance. Table.5.2 reveals that the Dice derived from models using the combination of magnitude and velocity data is 3% higher compared to the models using the magnitude images only. Adding velocity images as input to the model is clearly shown to be beneficial. The variation in Dice and ASD over the cardiac phase for each model is illustrated in Figure.5.4. All models achieved the best Dice and ASD in phases 1, 2 and 30 which is around the ED phase. The lowest performance is observed in the phases varying between phase 11-13, which is around the ES phase. These results demonstrate that LV segmentation from 4D flow data is more accurate in the ED phase than in the ES phase.

The PCC, bias and LOA of clinical evaluation metrics comparing manual with automatic segmentation results are reported in Table 5 3. Figure.5.5 and Figure.5.6 show the scatter plots, including PCC and Bland-Altman plots of four clinical metrics. SAX2DF achieved the highest correlation of 90.34%, 92.09% and 75.83% for EDV, ESV and LVEF, respectively. The best PCC for KE was achieved using SAX2D method. Although the PCC in LVEF derived from all five methods are lower than 80%, the results in the other three metrics demonstrate a good linear correlation with the results derived from manual segmentation. Notably, all five models achieved a PCC for KE higher than 90%. Although there is a significant variation in the performance of EDV and ESV estimation derived from the different methods, the biases for those two metrics derived from SAX2D, RAW3D and SAX2DF are smaller than 10 ml. The smallest biases in EDV and ESV are 2.03 ml and 3.35 ml derived from SAX3D and SAX2DF, respectively. RAW2D achieved the worst performance, with a bias of 19.19 ml and 20.52 ml in EDV and ESV, respectively. RAW3D and SAX2DF achieved the smallest bias in LVEF and normalized KE with 3.09% and 0.02 μJ/ml, respectively.

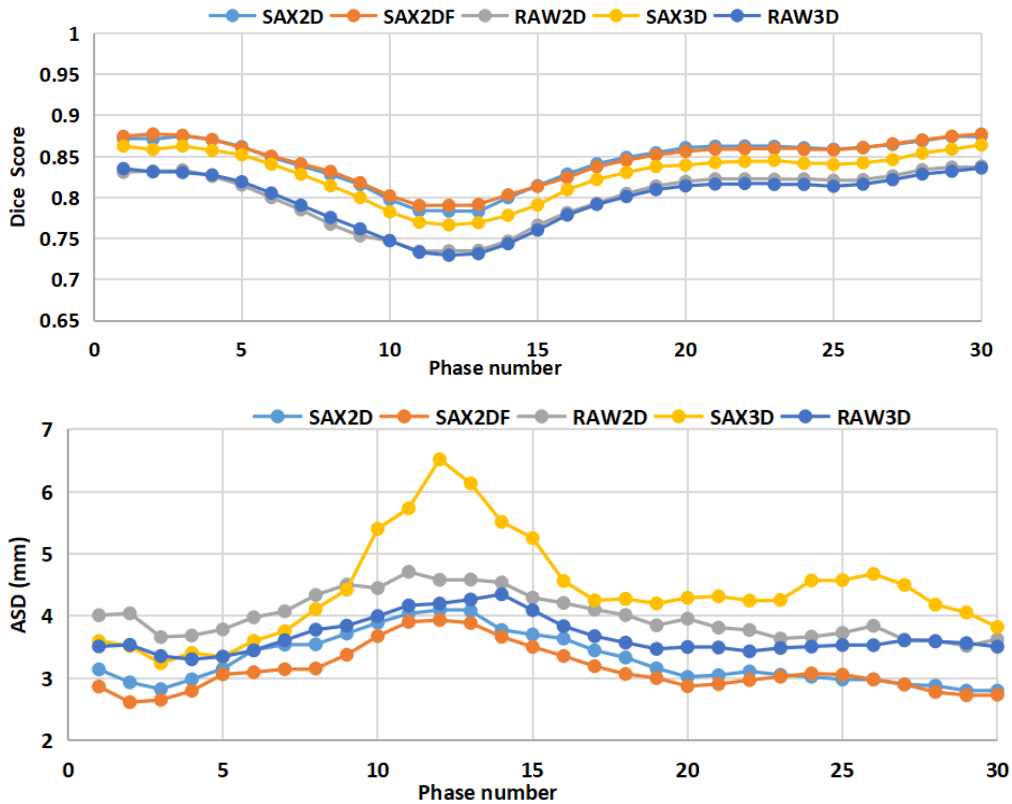

Figure.5.4. Average Dice and ASD results plotted over time (averaged over all subjects). The x-axis is the phase number, y-axis is the averaged Dice (upper) and ASD (bottom) derived from different models.

Table 5-3. PCC and Bias of clinical metrics from the prediction against the reference. The best results are shown in bold.

		SAX2D	RAW2D	SAX3D	RAW3D	SAX2DF
PCC	EDV(%)	89.44	86.51	88.69	87.97	**90.34**
	ESV(%)	91.47	80.61	85.51	79.48	**92.09**
	LVEF(%)	70.29	50.08	50.12	47.77	**75.83**
	KE(%)	**98.63**	92.58	92.82	93.51	94.32
Bias ± LOA	EDV (ml)	7.24±56.71	19.19±62.64	**-2.03±57.94**	-7.24±59.56	7.96±54.15
	ESV (ml)	8.31±46.62	-20.52±68.58	-11.39±58.88	-7.24±69.46	**-3.35±45.75**
	LVEF (%)	7.73±17.86	6.55±23.92	6.89±23.92	**3.09±25.37**	5.11±15.92
	KE(μJ/ml)	0.23±1.02	0.75±2.50	0.44±2.44	0.28±2.37	**0.02±1.47**

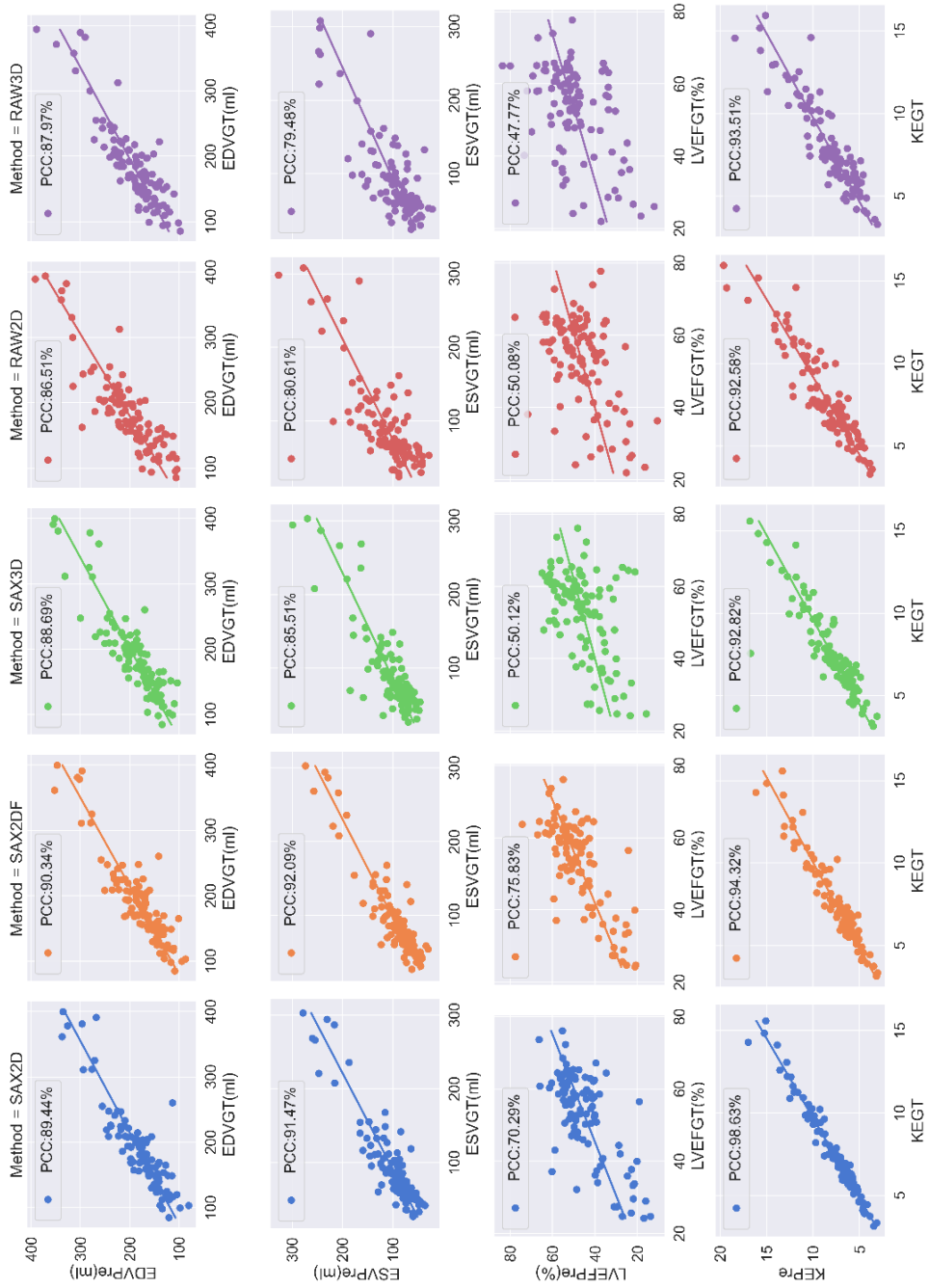

Figure.5.5. Correlation of clinical metrics derived from manual and automated segmentation. Each column represents one CNN model. The four rows denote four clinical metrics including EDV, ESV, LVEF and KE. For each plot, the x-axis is the measure derived from the manual segmentation and y-axis represents the results derived from automated prediction.

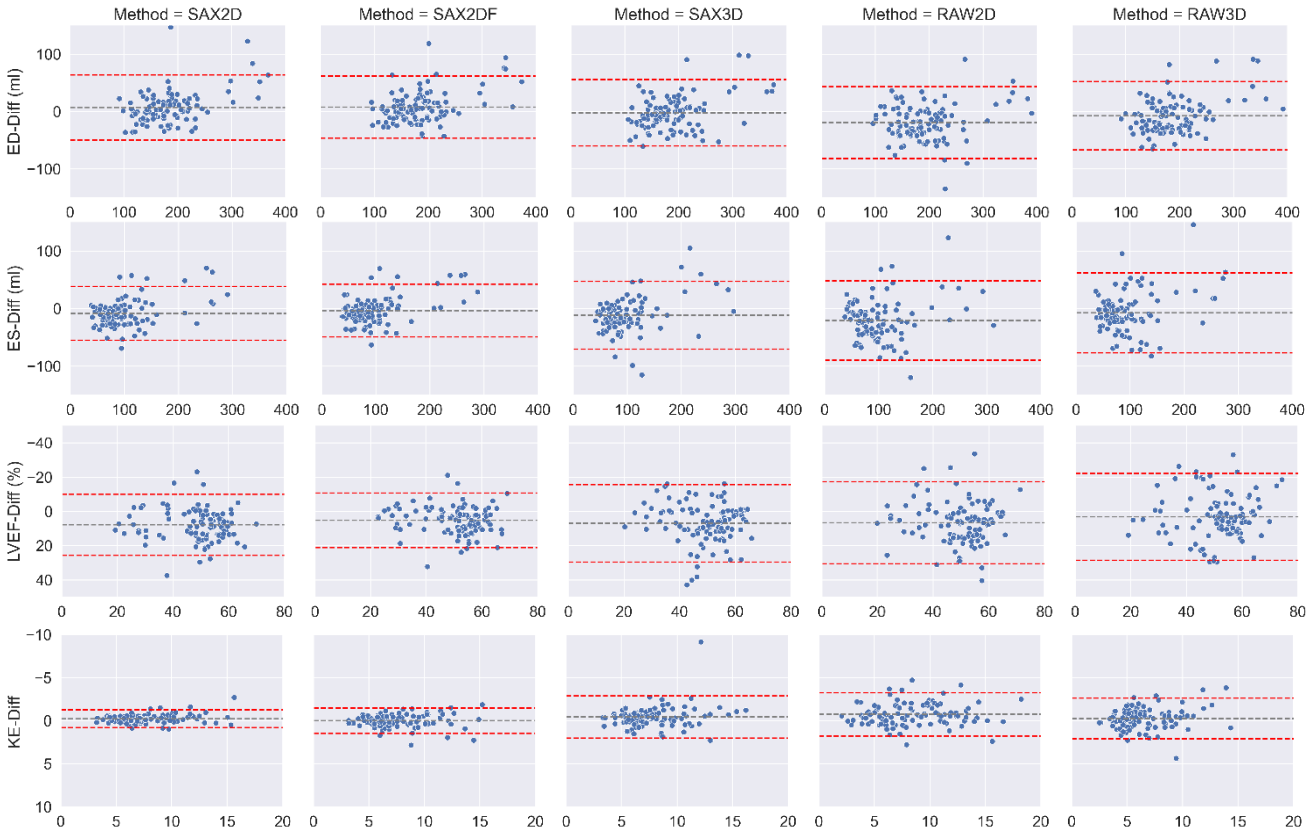

Figure.5.6. Bland-Altman plots of clinical metrics comparing automated and manual segmentation. The columns represent five models and the rows show the results of EDV, ESV, LVEF and KE, respectively. In each Bland-Altman plot, the x-axis denotes the average of two measurements and the y-axis represents the difference between two measurements. The black line represents the bias and the two red lines denote the LOA.

Examples of 2D and 3D segmentation masks derived from the five models are shown in Figure.5.7.

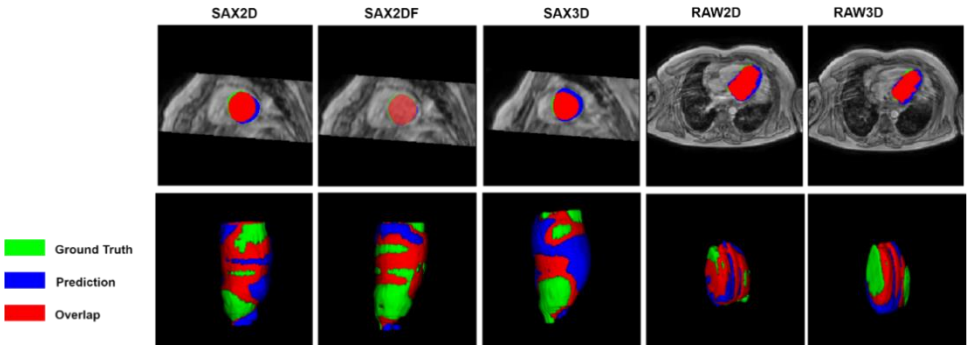

Figure.5.7. Examples of automated LV segmentation results in 2D and 3D. The first two rows are the results of 2D and 3D segmentation results. Green color represents the ground truth, blue color is the prediction, and red parts are the overlap between the prediction and ground truth.

5.3.2 Uncertainty results

Table.5.4 reports the averaged uncertainty scores both in the LV blood pool and the defined boundary area over 3090 phases (30 phases per subject, 103 subjects in total) derived from the five proposed models. SAX2DF achieved the lowest uncertainty scores with 0.12 and 0.75 in the whole LV and boundary area. SAX3D has lower uncertainty than SAX2D (0.13 vs. 0.15, 0.76 vs. 0.83). Similarly, RAW3D has a lower uncertainty than RAW2D (0.13 vs. 0.20, 0.77 vs. 0.82). The 3D models are shown to be more confident in its predictions than the 2D models. When comparing SAX2D and SAX2DF, it can be concluded that the late fusion method resulted in a lower uncertainty score.

Table.5.4. The averaged uncertainty value derived from different defined areas. The LV chamber refers to the area with a probability larger than 0.5. Boundary area refers to the area with probability ranging from 0.4 to 0.6.

Area	SAX2D	SAX3D	RAW2D	RAW3D	SAX2DF
LV chamber	0.15±0.32	0.13±0.36	0.20±0.88	0.13±0.41	**0.12±0.44**
Boundary area	0.83±0.29	0.76±0.54	0.82±0.66	0.77±0.28	**0.75±0.17**

5.3.3 Flow quantitative analysis

SAX2DF is the best segmentation model among the proposed five models, according to the performance on segmentation accuracy, clinical metrics and uncertainty score. Therefore, we further investigated the performance of SAX2DF in quantifying KE and flow components. The low averaged error of indexed KE ranging from -0.03 mJ to 0.04 mJ and flow components varying from -4.58% to 3%, as reported in Table.5.5, shows a good agreement between the prediction and ground

truth. A detailed summary of the PCC of KE and flow components derived from the automatic and manual methods is illustrated in Figure.5.8.

Table.5.5. The averaged difference and PCC between the automatic and manual methods in flow components quantification and KE. RES: Residual volume; RET: retained inflow; DEL: delayed ejection flow; DIR: direct flow. KE parameters were indexed to LV end-diastolic volume.

	flow components (%)				KE (mJ)			
	RES	RET	DEL	DIR	Max RR	Max Systole	Max E-wave	Max A-wave
Mean ± SD	3.00±5.96	0.78±25.9	0.80±2.68	-4.58±4.34	-0.03±0.38	0.04±0.52	0.02±0.45	0.02±0.3
PCC (%)	86.7	91.1	89.1	94.2	97.3	88.2	96.8	95.5

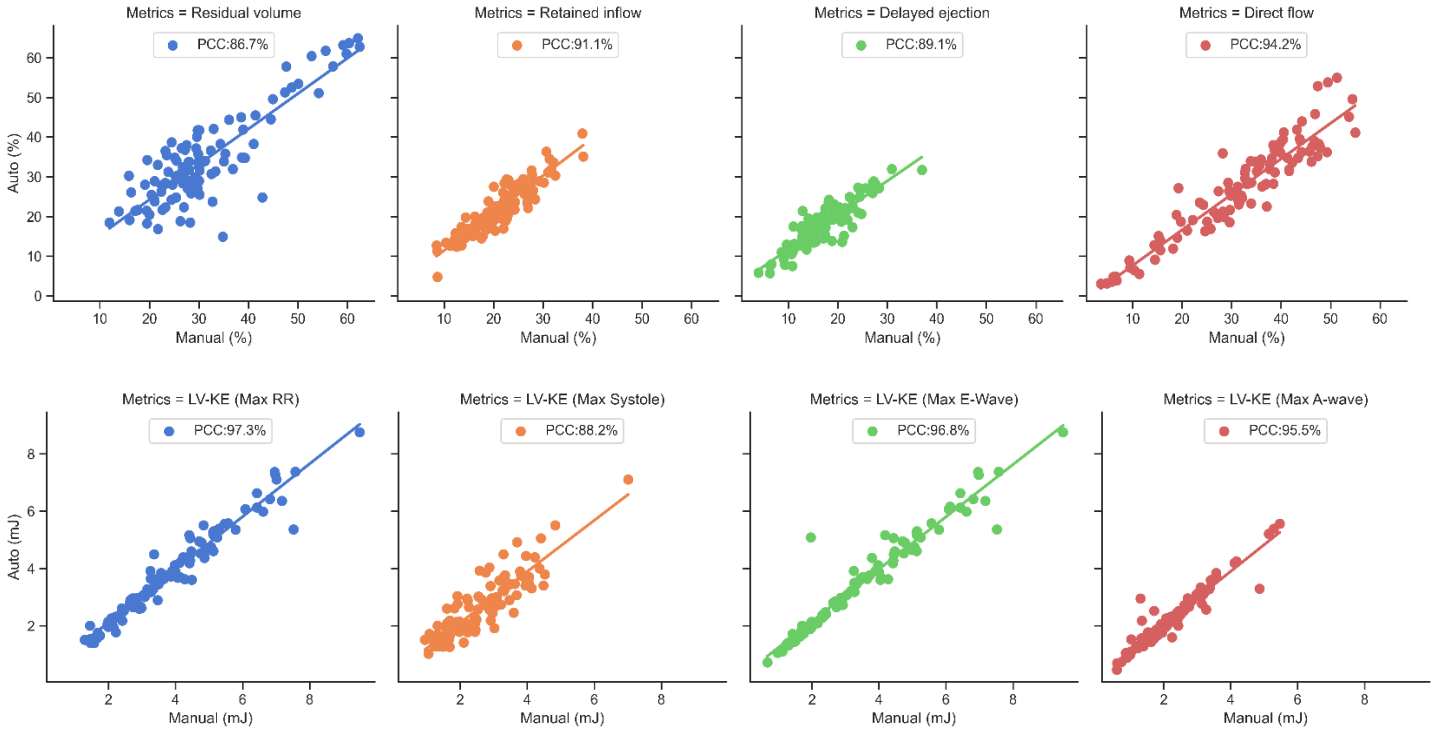

Figure.5.8. Correlation of left ventricle kinetic energy and four flow components derived from the SAX2DF and manual method. **First row:** flow components including residual volume, retained inflow, delayed ejection and direct flow. **Second row:** kinetic energy including max RR, systole, E-wave and A-wave KE

Figure.5.9 visualizes the result of LV flow component analysis derived from manual and CNN based segmentation in five subjects. The results demonstrate a good agreement between those two segmentation methods. More flow components visualization videos and segmentation result videos can be found in https://github.com/xsunn/4DflowLVSegmentation.

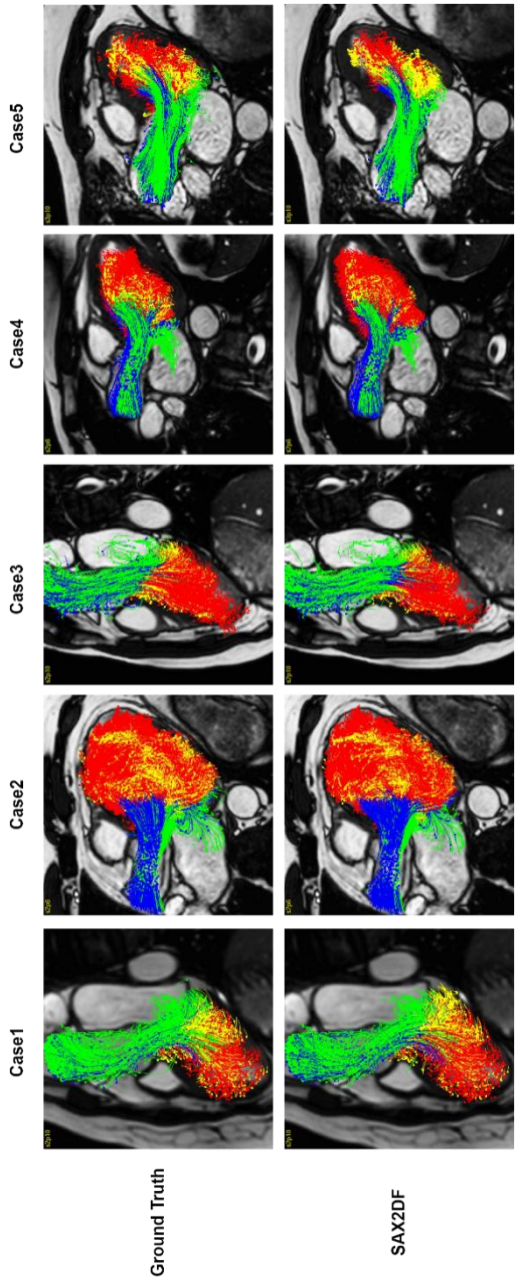

Figure.5.9. Visualization of the different ventricular flow components by track particle derived from the manual segmentation and the method of SAX2DF. Green: direct flow. Yellow: retained inflow. Blue: delayer ejection flow. Red: residual flow.

5.4 Discussion

In this work, we developed and evaluated CNN-based methods for automatic segmentation and LV flow assessment from 4D flow cardiac MRI. The main findings of our study were (1) CNN models showed good performance in LV segmentation with an average Dice of 84.5% across 103 subjects with 90,313 resliced 2D image pairs; (2) Data preprocessing has an impact on the segmentation results; (3) Combining the features from magnitude and velocity images together can benefit the segmentation performance in 4D flow MRI; (4) High correlation and low bias of EDV, ESV, KE and flow components analysis demonstrate CNN-based segmentation can provide reliable quantification of LV flow in 4D flow data.

Segmentation in 4D flow cardiac MRI is challenging due to the poor contrast between the heart chamber and its surrounding tissue. Few approaches have been proposed to overcome this challenge. Atlas-based methods [4] and registration-based methods [10] are two prevailing traditional approaches. The atlas-based method relies on image registration to generate accurate transformation between a labelled atlas and the images. Registration-based segmentation methods rely on the registration between labelled cine MRI data and 4D flow data. Both of these methods require additional data and high computational costs due to the registration. Bustamente *et al.* [11] employed a 3D U-Net architecture for LV segmentation, but in their proposed method, only the magnitude images were used as input and information from velocity images were ignored. In this work, we compared five models named SAX2D, SAX3D, RAW2D, RAW3D and SAX2DF to segment the LV from 4D flow MRI without any additional cine MRI and we also investigated the impact of different data pre-processing approaches, feature fusion methods and model structure on the segmentation results.

The performance derived from our proposed method is not as good as that of Bustamente's. The data cohort used in their work is much larger than ours; in our work, 2472 3D volumes are employed for training, which is significantly smaller than Bustamente's 5760 3D volumes. Meanwhile, our results are averaged over 3090 3D volumes using five-fold cross-validation, whereas their results were directly derived from 1640 3D volumes. Furthermore, simply using the magnitude images as input allows them to introduce various data augmentation techniques to enlarge the training data. However, the conventional data augmentation methods such as rotation, Gaussian noise and transformation cannot directly work on our proposed approach because the velocity images are more complicated than the magnitude images. Therefore, compared to their work, we trained the model with fewer data but evaluated the performance on a larger data set. Since in Bustamente's work all 4D flow acquisitions were obtained post contrast injection in both patients and volunteers and navigator gating breathing motion was applied, it is expected that the image quality of the obtained magnitude images was higher in that study.

For data preparation, given the known orientation, the original 4D flow MRI acquisition volume was resliced into short-axis slices. The raw data and resliced short-axis data served as two independent training data sets to train the networks. Improved segmentation results were derived when using the resliced short-axis data as the training data, demonstrating resliced short-axis data provided more accurate information for the segmentation, which could be explained by the more various shapes and ambiguous borders in the raw data when compared to the more consistent convex left ventricular shape in the short-axis view.

Considering magnitude and velocity images as two different modalities in 4D flow MRI, we proposed two approaches named early fusion and late fusion to fuse the information from these modalities. SAX2D, SAX3D, RAW2D and RAW3D employed early fusion by concatenating two modalities along the channel dimension as the input. While for the late fusion, SAX2DF employed two encoders to extract features from two modalities and then concatenated the features along the spatial dimension. A modestly improved performance was observed in SAX2DF when compared to the other methods, revealing that late fusion works better. We also compared the segmentation performance between 2D and 3D U-Net based methods. The results show that compared to SAX3D, SAX2D achieved better performance in all evaluation metrics. Constrained to the input spatial dimension, in SAX3D the kernel size of the final pooling layer was set to 2×2×1, resulting in the spatial features not being extracted completely. Moreover, a total of 3420 resliced 3D samples (104 subjects, 30 phases in each subject) were used to train and test the 3D U-Net, which is much less than 91,182 2D samples. As a result, the smaller training data size may be the primary reason why SAX3D did not outperform the SAX2D model.

CNN models produce a pixel-level prediction without any knowledge about the confidence of the model in its predictions. In this work, we introduced the Monte Carlo dropout method to estimate the uncertainty of the model in its segmentation results. The uncertainty score assesses segmentation reliability and offers the quantification of error to increase trust into CNN models. The results showed that the most uncertain area in the prediction is near the LV endocardial boundary, which can be explained by the poor contrast in the magnitude images and also because of the low blood flow velocity near the LV wall. Segmenting the myocardium in addition to the LV blood pool may reduce the uncertainty but cannot eliminate the uncertainty. When analyzing the uncertainty scores derived from different models, it reveals that the 3D models (SAX3D and RAW3D) performed better than the 2D models (SAX2D and RAW2D). Because 3D models are able to extract more spatial information from the input than the 2D models. It can be observed that although SAX2DF is a 2D model, benefiting from the late fusion method, SAX2DF achieved the lowest uncertainty score among all five models. A further evaluation of the

results derived from the best model, SAX2DF, was performed by comparing the KE and flow components. The results shows a good agreement between the ground truth and prediction.

There are several limitations in our work. The major limitation is the lack of generalization of the proposed models. The data used in this study was acquired from one vendor and one center. Meanwhile, there is no publicly available 4D flow MRI dataset currently. Therefore the model might not generalize well to the other datasets from different vendors or centers. As Bai [17] pointed out, a CNN model can perform well in other datasets using fine-tuning or transfer learning. Additionally, exploiting advanced data augmentation methods utilizing domain knowledge is also crucial for model generalization and robustness [18]. However, due to the complicated structure of velocity images, commonly used data augmentation methods are not suitable for 4D flow data. A novel efficient late fusion based feature fusion method also needs to be investigated.

5.5 Conclusions

In conclusion, we developed multiple deep learning-based 4D flow MRI LV segmentation models that do not require additional cine MRI. The proposed CNN models were evaluated on a large in-house dataset, achieving good performance on several metrics. The results demonstrate that a model employing late fusion and trained on resliced short-axis view data generates the best performance for left ventricular segmentation in 4D flow MRI.

References

1. Stankovic Z, Allen BD, Garcia J, Jarvis KB, Markl M. 4D flow imaging with MRI. Cardiovascular diagnosis and therapy. 2014;4(2):173.
2. Rizk J. 4D flow MRI applications in congenital heart disease. European Radiology. 2021;31:1160-74.
3. Gupta AN, Avery R, Soulat G, Allen BD, Collins JD, Choudhury L, Bonow RO, Carr J, Markl M, Elbaz MS. Direct mitral regurgitation quantification in hypertrophic cardiomyopathy using 4D flow CMR jet tracking: evaluation in comparison to conventional CMR. Journal of Cardiovascular Magnetic Resonance. 2021;23(1):1-3.
4. Eriksson J, Carlhäll CJ, Dyverfeldt P, Engvall J, Bolger AF, Ebbers T. Semi-automatic quantification of 4D left ventricular blood flow. Journal of Cardiovascular Magnetic Resonance. 2010;12:1-0.
5. Kanski M, Arvidsson PM, Töger J, Borgquist R, Heiberg E, Carlsson M, Arheden H. Left ventricular fluid kinetic energy time curves in heart failure from cardiovascular magnetic resonance 4D flow data. Journal of Cardiovascular Magnetic Resonance. 2015;17:1-0.
6. Bustamante M, Gupta V, Forsberg D, Carlhäll CJ, Engvall J, Ebbers T. Automated multi-atlas segmentation of cardiac 4D flow MRI. Medical image analysis. 2018;49:128-40.
7. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. InMedical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18 2015 (pp. 234-241). Springer International Publishing.
8. Berhane H, Scott M, Elbaz M, Jarvis K, McCarthy P, Carr J, Malaisrie C, Avery R, Barker AJ, Robinson JD, Rigsby CK. Fully automated 3D aortic segmentation of 4D flow MRI for hemodynamic analysis using deep learning. Magnetic resonance in medicine. 2020;84(4):2204-18.
9. Wu Y, Hatipoglu S, Alonso-Álvarez D, Gatehouse P, Firmin D, Keegan J, Yang G. Automated multi-channel segmentation for the 4D myocardial velocity mapping cardiac MR. InMedical Imaging 2021: Computer-Aided Diagnosis 2021 (Vol. 11597, pp. 169-175). SPIE..
10. Corrado PA, Wentland AL, Starekova J, Dhyani A, Goss KN, Wieben O. Fully automated intracardiac 4D flow MRI post-processing using deep learning for biventricular segmentation. European Radiology. 2022 Aug;32(8):5669-78.
11. Bustamante M, Viola F, Engvall J, Carlhäll CJ, Ebbers T. Automatic Time-Resolved Cardiovascular Segmentation of 4D Flow MRI Using Deep Learning. Journal of Magnetic Resonance Imaging. 2023;57(1):191-203.

12. Garg P, Westenberg JJ, van den Boogaard PJ, Swoboda PP, Aziz R, Foley JR, Fent GJ, Tyl FG, Coratella L, ElBaz MS, Van Der Geest RJ. Comparison of fast acquisition strategies in whole-heart four-dimensional flow cardiac MR: Two-center, 1.5 Tesla, phantom and in vivo validation study. Journal of Magnetic Resonance Imaging. 2018;47(1):272-81.
13. Elbaz MS, van der Geest RJ, Calkoen EE, de Roos A, Lelieveldt BP, Roest AA, Westenberg JJ. Assessment of viscous energy loss and the association with three-dimensional vortex ring formation in left ventricular inflow: In vivo evaluation using four-dimensional flow MRI. Magnetic resonance in medicine. 2017;77(2):794-805.
14. Klein S, Staring M, Murphy K, Viergever MA, Pluim JP. Elastix: a toolbox for intensity-based medical image registration. IEEE transactions on medical imaging. 2009;29(1):196-205.
15. Baumgartner CF, Tezcan KC, Chaitanya K, Hötker AM, Muehlematter UJ, Schawkat K, Becker AS, Donati O, Konukoglu E. Phiseg: Capturing uncertainty in medical image segmentation. InMedical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22 2019 (pp. 119-127). Springer International Publishing.
16. Gal Y, Ghahramani Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. Ininternational conference on machine learning 2016 (pp. 1050-1059). PMLR.
17. Bai W, Sinclair M, Tarroni G, Oktay O, Rajchl M, Vaillant G, Lee AM, Aung N, Lukaschuk E, Sanghvi MM, Zemrak F. Automated cardiovascular magnetic resonance image analysis with fully convolutional networks. Journal of Cardiovascular Magnetic Resonance. 2018;20(1):1-2.
18. Chen C, Bai W, Davies RH, Bhuva AN, Manisty CH, Augusto JB, Moon JC, Aung N, Lee AM, Sanghvi MM, Fung K. Improving the generalizability of convolutional neural network-based segmentation on CMR images. Frontiers in cardiovascular medicine. 2020;7:105.

Availability of data and materials

The datasets used in this study will not be made publicly available. However, the code for the network is available:https://github.com/xsunn/4DflowLVSegmentation.

Competing interests

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author Contributions

XS designed and implemented the method, performed data analysis and wrote the manuscript. RG designed this study, prepared the dataset and revised the manuscript. LC designed the network and revised the manuscript. SP and PG provided support on the clinical aspects and they also provided the data used in the study. All authors read and approved the manuscript.

Acknowledgements

XS is supported by the China Scholarship Council No. 201808110201. LC is supported by the RISE-WELL project under H2020 Marie Skłodowska-Curie Actions. Prof. Sven Plein from the University of Leeds is acknowledged for granting access to the image data used in this work.