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

Sun, X.

Chapter 6 Transformer based feature fusion for left ventricle segmentation in 4D flow MRI

This chapter was adapted from:

Xiaowu Sun, Li-Hsin Cheng, Sven Plein, Pankaj Garg, Rob J. van der Geest. **Transformer based feature fusion for left ventricle segmentation in 4D flow MRI.** International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). Springer, Cham, 2022.

Abstract

Four-dimensional flow magnetic resonance imaging (4D Flow MRI) enables visualization of intra-cardiac blood flow and quantification of cardiac function using time-resolved three directional velocity data. Segmentation of cardiac 4D flow data is a big challenge due to the extremely poor contrast between the blood pool and myocardium. The magnitude and velocity images from a 4D flow acquisition provide complementary information, but how to extract and fuse these features efficiently is unknown. Automated cardiac segmentation methods from 4D flow MRI have not been fully investigated yet. In this paper, we take the velocity and magnitude image as the inputs of two branches separately, then pro-pose a Transformer based cross- and self-fusion layer to explore the inter-relationship from two modalities and model the intra-relationship in the same modality. A large in-house dataset of 104 subjects (91,182 2D images) was used to train and evaluate our model using several metrics including the Dice, Average Surface Distance (ASD), end-diastolic volume (EDV), end-systolic volume (ESV), Left Ventricle Ejection Fraction (LVEF) and Kinetic Energy (KE). Our method achieved a mean Dice of 86.52%, and ASD of 2.51 mm. Evaluation on the clinical parameters demonstrated competitive results, yielding a Pearson correlation coefficient of 83.26%, 97.4%, 96.97% and 98.92% for LVEF, EDV, ESV and KE respectively. Code is available at github.com/xsunn/4DFlowLVSeg.

6.1 Introduction

Quantitative assessment of left ventricular (LV) function from magnetic resonance imaging (MRI) is typically based on the use of short-axis multi-slice cine MRI due to its excellent image quality [1,2]. Recently, four-dimensional (4D) Flow MRI has been introduced, encoding blood flow velocity in all three spatial directions and time dimension. 4D Flow MRI can be used for detailed analysis of intra-cardiac blood flow hemodynamics, providing additional information over conventional cine MRI. The segmentation of the cardiac cavities is an important step to derive quantitative blood flow results, such as the total LV kinetic energy (KE) [3]. 4D Flow MRI generates four image volumes including a magnitude image and three velocity images, one for each spatial dimension. Figure.6.1 shows an example of magnitude and velocity images from one slice out of a 4D Flow MRI data set. The example highlights the extremely poor contrast between the heart chambers and the myocardium in the 4D Flow data. Therefore, most authors have used segmentations derived from co-registered short-axis cine MR in order to quantify ventricular blood flow parameters from the 4D Flow data. However, this relies on accurate spatial and temporal registration of the two MR sequences. Inconsistent breath-hold positioning may introduce spatial misalignment while heart rate differences will result in temporal mismatch between the acquisitions. The aim of the current work was therefore to develop an automated method for LV segmentation from 4D Flow MRI data, not requiring additional cine MRI data.

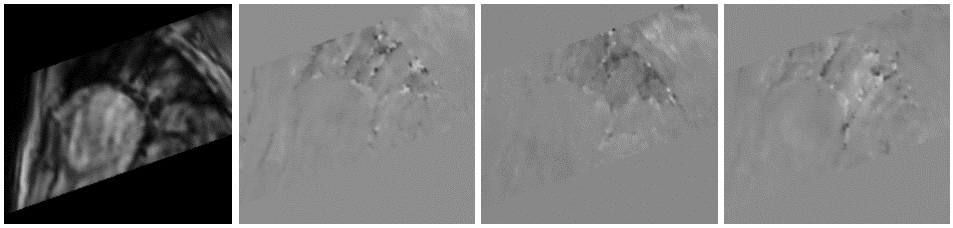

Figure.6.1. A sample of cardiac 4D Flow data in short-axis view. The first image is the magnitude image, and the last three images are the velocities in x, y and z dimensions respectively.

Since U-Net [4] was proposed, convolutional neural networks (CNNs) have been predominant in the task of medical image segmentation. Many variants of U-Net have been proposed further improving the performance. For instance, nnU-Net [5] introducing automated self-configuring outperformed most existing approaches on 23 diverse public datasets. Although those CNN based networks have achieved an excellent performance, restricted by the locality of convolutional kernels, they cannot capture long-distance relations [6,7].

Transformer is considered as an alternative model using its self-attention mechanism to overcome the limitation of CNN. Transformer was designed firstly for

natural language processing (NLP) tasks such as machine translation and document classification. More recently, Transformer-based approaches were introduced in medical image processing. TransUnet [8] applied a CNN-Transformer hybrid encoder and pure CNN decoder for segmentation. However, TransUnet still uses convolutional layers as the main building blocks. Inspired by the Swin Transformer [9], Cao proposed a U-Net-like pure Transformer based segmentation model which uses hierarchical Swin Transformer as the encoder and a symmetric Swin Transformer with patch expanding layer as the decoder [10]. Other Transformer-based networks [11,12,13] also mark the success of Transformer in medical image segmentation and reconstruction.

Although numerous deep learning-based segmentation methods have been proposed in various modalities, the automatic segmentation of the LV directly from 4D Flow data has not been explored yet. A specific challenge is that the magnitude and velocity images of a 4D Flow acquisition have different information content and should be considered as different modalities. Moreover due to velocity noise, a careful fusion method is needed to avoid redundancy or insufficient feature integration [14,15].

In this paper, we present, to the best of our knowledge, the first study to segment the LV directly from 4D Flow MRI data. Our main contributions are: (1) we propose two self- and cross-attention-based methods to fuse the information from different modalities in 4D Flow data; (2) we evaluate our method in a large 4D Flow dataset using multiple segmentation and clinical evaluation metrics.

6.2 Method

6.2.1 Attention mechanism

Attention mechanism, mapping the queries and a set of keys-value pairs to an output, is the fundamental component in Transformer. In this section, we first introduce how the self-attention module models the intra-relationship of features from the same image modality. Then we explain how cross-attention explores the inter-relationship of features from two different modalities. The two attention modules are illustrated in Figure.6.2.

Self-Attention module. In self-attention module [6], the **Q** (queries), **K** (keys) and **V** (values) are generated from the same modality. **Q** and **K** determine a weight matrix after the scaled dot product which is used to compute the weighted sum of **V** as the output. The computing process can be described as in equation (6.1):

$$Atten(Q_a, K_a, V_a) = softmax(\frac{Q_a K_a^{\ T}}{\sqrt{d}})V_a \tag{6.1}$$

where *d* is the key dimensionality, and *a* denotes modality *a*.

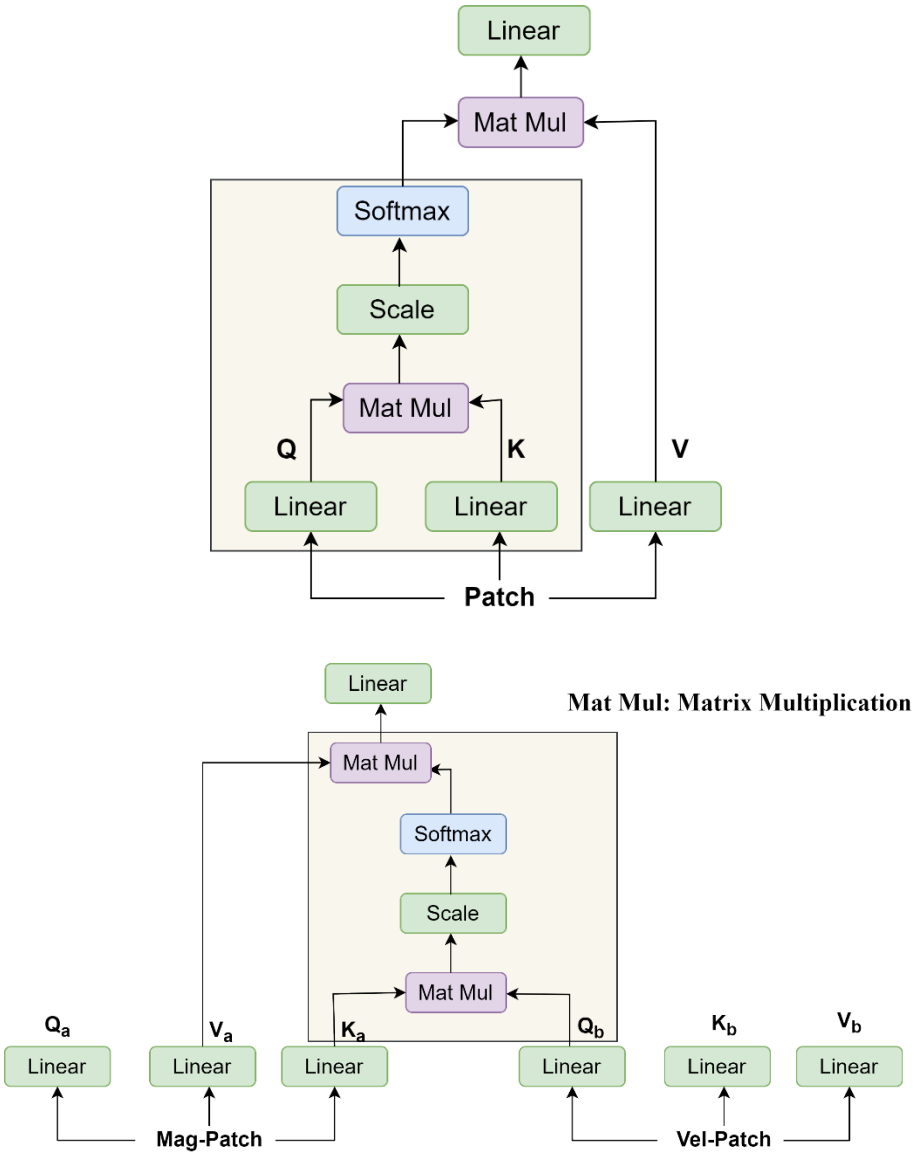

Figure.6.2. The structure of self-attention (upper) and cross-attention (bottom) modules.

Cross-Attention module. Although self-attention explores the intra-modality relationship, the inter-modality relationship, such as the relationship between pixels in the magnitude image and velocity image is not explored. The cross-attention module takes two patches as the input to generate the **Q**, **K**, and **V**. **V** and **K** are generated from the same modality, while **Q** is derived from another modality. The

other operations are kept the same as in self-attention. It can be expressed as equation (6.2). Hence, cross-attention can be adopted to fuse the information from different modalities.

$$Atten(Q_b, K_a, V_a) = softmax(\frac{Q_b K_a^T}{\sqrt{d}})V_a \tag{6.2}$$

Multi-head self(cross)-attention module. To consider various attention distributions and multiple aspects of features, the multi-head attention mechanism [6] is introduced. The multi-head attention is the concatenation of *h* single attentions along the channel dimension followed by a linear projection. Thus, the multi-head attention can be formulated as equation (6.3, 6.4)

$$MultiHead(Q, K, V) = Concat(H_1, H_2, \cdots, H_h)W^0 \tag{6.3}$$

$$H_i = Atten(Q_i, K_{i,} V_i) \tag{6.4}$$

where *Atten* is self-attention or cross-attention, Q_i, K_i, V_i are the *i*-th vector of Q, K ,V . In each single attention head, the channel dimension $d^{'} = d/h$.

6.2.2 Feature Fusion Layer

To fuse the features generated from the magnitude and velocity images, we proposed a feature fusion layer (FFL). The structure of FFL shown in Figure.6.3 contains two branches, each branch has one cross-fusion layer and one self-fusion layer.

Cross-Fusion Layer (CFL). CFL is proposed to fuse the features from different modalities. The structure of CFL is illustrated in the upper dash box in Figure.6.3. Given **Q**, **K** and **V** generated from two modalities, the Multi-head Cross-Attention (MCA) module followed by a linear projection firstly integrate those information. Then the fused features are added to the original input. Subsequently, another two linear projections and one residual connection followed by a normalization layer are used to enhance the fused information.

Self-Fusion Layer (SFL). The lower dash box in Figure.6.3 shows the structure of SFL. SFL is a simple stack of Multi-head Self-Attention (MSA), linear projection, residual and normalization layer. Different from CFL, the SFL only uses one input to generate the values for the MSA. CFL aims to fuse the features from different image modalities, SFL further enhances the fused features using self-attention.

Having two feature maps from the magnitude and velocity images respectively, we first transform the feature maps into sequence data using the patch embedding. Specifically, the feature $\mathrm{f} \in \mathrm{R}^{H \times W \times C}$ is divided into $N = HW/P^2$ patches, where the

patch size P is set to 16. The patches are flattened and embedded into a latent D-dimension, obtaining a embedding sequence $e \in \mathrm{R}^{N \times D}$. However, dividing feature maps into patches leads to loss of spatial information. Therefore, a learnable positional encoding sequence is added to the embedding sequence to address this issue. Then the sequence data is passed into the FFL. In this work, we used a stack of 4 FFLs as the feature fusion network (FFN).

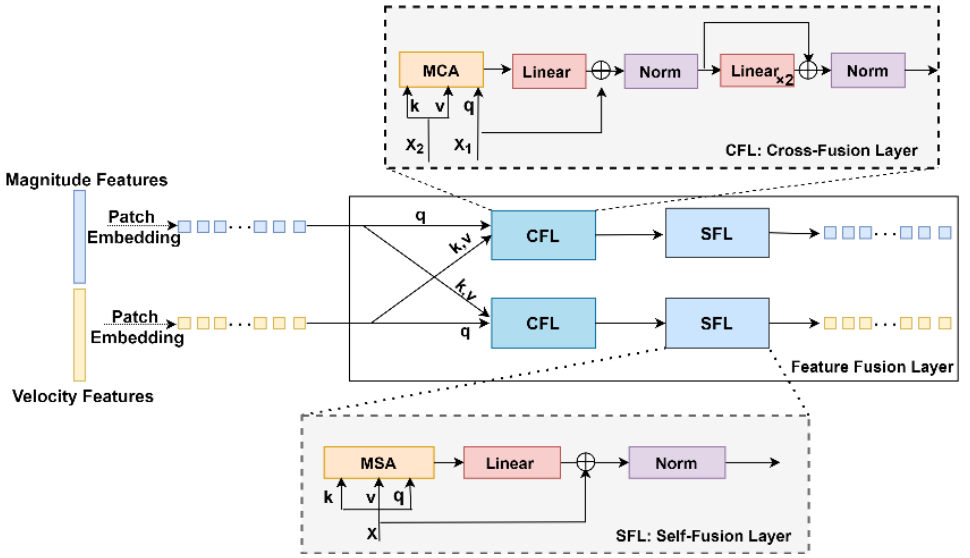

Figure.6.3. Structure of feature fusion layer (FFL). The input of the feature fusion layer is two features derived from magnitude and velocity images respectively. The upper box is the structure of cross-fusion layer (CFL) and the lower one is the structure of self-fusion layer (SFL).

6.2.3 Network Structure

Figure.6.4 illustrates the proposed segmentation network, which takes the U-Net as the backbone. The encoder uses two parallel branches to extract features from magnitude and velocity image separately. The features at the same level are integrated using the feature fusion network. By doing so, the size of integrated features reduces due to the patch embedding. Hence, the fused features are up-sampled first, then added to the original features as the final aggregated features.

The four-level paired aggregated features derived from the encoder are taken as the inputs to the decoder part. The fused features at the same level generated from the magnitude and velocity branch in the encoder are concatenated followed by a convolutional layer to reduce the number of feature maps. The remaining decoder parts including the up-sampling, convolutional and softmax layers are the same as in U-Net.

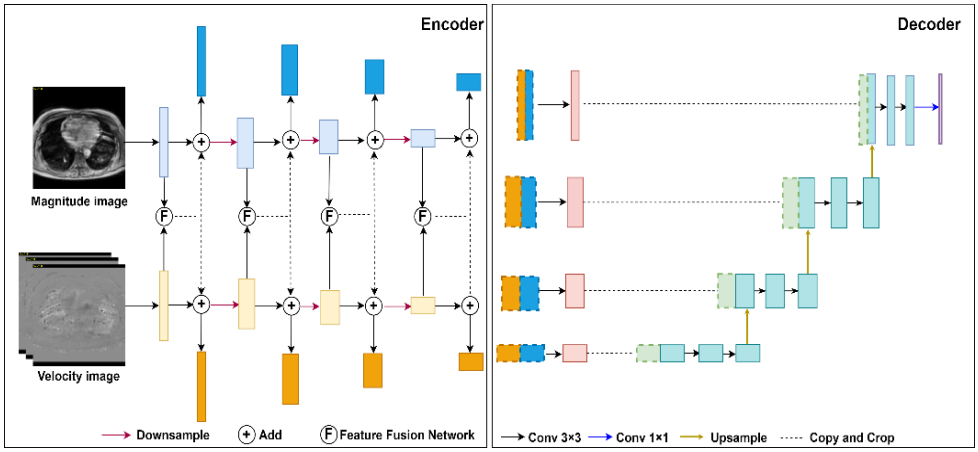

Figure.6.4. The architecture of our proposed segmentation network structure. The feature fusion network is a stack of 4 FFLs.

6.3 Materials

6.3.1 Dataset

4D flow MRI was performed in 28 healthy volunteers and 76 post-myocardial infarction patients on a 1.5T MR system (Philips Healthcare). The 4D flow acquisition covered the complete LV and was acquired in axial orientation with a voxel size of 3×3×3 mm^3 and reconstructed into 30 cardiac phases. The other imaging parameters are as follows: flip angle=10°, velocity encoding (VENC) of 150 cm/s, FOV= 370-400×370-400 mm^2, echo time (TE)=1.88-3.75 ms, repetition time (TR)= 4.78-13.95 ms. In addition, standard cine-MRI was performed in multiple short-axis slices covering the LV from base to apex. More details about the MR acquisition protocol can be found here [16].The short-axis cine acquisition was used to segment the LV endocardial boundaries in all slices and phases. After rigid registration with the 4D flow acquisition, the defined segmentation served as ground truth segmentation of the 4D flow acquisition. Based on the known short-axis orientation, the 4D flow data was resliced into the short-axis orientation using a slice spacing of 3 mm and a fixed number of 41 slices. The spatial in-plane resolution was defined equal to the available short-axis cine acquisition and varied from 0.83×0.83 mm^2 to 1.19×1.19 mm^2.

Excluding the images without any objects this resulted in 91 182 annotated pairs of 2D images, each pair has one 2D magnitude image and three-directional velocity images. The subjects were randomly split into three parts with 64, 20, 20 (total number of images: 55 825, 17 335 and 18 022) for training, validation and testing respectively. We normalized the magnitude image into [0,1] using min-max method. The images were cropped into 256×256.

6.3.2 Evaluation metrics

Segmentation metrics. To quantitatively evaluate the segmentation performance, Dice and Average Surface Distance (ASD) were measured.

Clinical metrics. The clinical metrics, including the end-diastolic volume (EDV), end-systolic volume (ESV), left ventricle ejection fraction (LVEF) and kinetic energy (KE) [3] were measured. The formula of LVEF and KE are defined as:

$$LVEF = \frac{EDV - ESV}{EDV} \times 100\% \qquad KE = \sum_{i=1}^{N} \frac{1}{2} \rho_{blood} \cdot V_i \cdot v_i^2 \qquad (6.5)$$

where N means the number of voxels in the LV, ρ_{blood} represents the density of blood (1.06g/cm^3), V is the voxel volume and v is the velocity magnitude. For each phase, the total KE is the summation of the KE of every voxel within LV. KE was normalized to EDV as recommended by other researchers [3].

Statistical Analysis. The results are expressed as mean ± standard deviation. Pearson correlation coefficient (PCC) was introduced to measure the correlation of the clinical metrics between the manual and automatic segmentation approaches. Paired evaluation metrics were compared using Wilcoxon-signed-rank test with $P < 0.05$ indicating a significant difference.

The Dice, ASD and KE reported in this work are the mean values as computed over 30 phases per subject.

6.4 Experiment and results

All the models were implemented in Pytorch and trained with a NVIDIA Quadro RTX 6000 GPU with 24 GB memory from scratch. We employed Adam as the optimizer with 0.0001 as the learning rate. All of the models were trained for 1000 epochs with a batch size of 15. The sum of Dice loss and cross-entropy loss was used as the loss function. Additionally, due to the complexity of the velocity images, we did not employ any data augmentation methods to enlarge the dataset.

We first evaluated our model against the U-Net, TransUnet [8], and U-NetCon. TransUnet added the self-attention module to the last layer of the encoder. The structure of U-NetCon (shown in the Supplementary) is similar to our proposed network. After removing the feature fusion network, the U-NetCon introduces two U-Net encoders which extract the features from two modalities separately and subsequently, the features from the same level in the encoder are concatenated as the input of the decoder. The input of U-Net and TransUnet is a four-channel stack of one magnitude and three velocity images. Whereas, in our method and U-NetCon, the magnitude and velocity images are taken as two separate input branches.

Table.6.1. Segmentation performance of different methods. Err means the absolute error between the manual and automatic segmentation methods.

Model	Dice (%)	ASD (mm)	EDV-Err (ml)	ESV-Err (ml)	LVEF-Err (%)	KE-Err (μJ/ml)
U-Net	84.62±5.91	2.99±1.66	20.35±31.53	16.01±19.76	7.60±7.10	1.50±1.64
U-NetCon	84.57±6.15	3.19±1.74	22.57±29.46	17.08±24.46	6.11±5.43	0.95±1.94
TransUnet	84.27±5.35	3.09±1.33	18.09±22.91	23.92±16.06	11.79±7.64	0.51±0.48
Ours	**86.52±5.54**	**2.51±1.14**	**9.02±10.03**	**11.86±10.55**	**5.10±4.55**	**0.36±0.34**

Table.6.1 reports the evaluation results of various metrics. It shows our method achieved the best performance for all of the six metrics. In Table.6.2 the PCC of the clinical metrics derived from different models are presented. Our method performs the best on all clinical metrics demonstrating a high correlation. Comparing the results of U-Net and TransUnet, the Dice and ASD only showed marginal improvement, but the performance decreased in LVEF with a low PCC of 48.7%**.** In order to evaluate the effectiveness of feature fusion network, we further compared our method to U-NetCon. As compared to U-NetCon, our method improves the Dice by 2% and the PPC by 3%, 9%, 7% and 16% for LVEF, EDV, ESV and KE, respectively, confirming that the proposed feature fusion network efficiently aggregates the features from magnitude and velocity images. More results about the boxplot and correlation comparing the Dice and four clinical parameters derived from our method and U-NetCon can be found in the supplementary.

Table.6.2. PCC of the clinical metrics derived from manual and automatic segmentation results.

Model	LVEF	EDV	ESV	KE
U-Net	70.65%	84.09%	91.50%	83.76%
U-NetCon	80.61%	88.46%	89.49%	82.46%
TransUnet	48.70%	91.36%	90.33%	97.86%
Ours	**83.26%**	**97.40%**	**96.97%**	**98.92%**

The P-value of Wilcoxon test results between the ground truth and our method in LVEF, EDV, ESV, KE are 0.13, 0.43, 0.35 and 0.43, as shown in Figure.6.5. All of those P-values are larger than 0.05, which confirmed that there is no significant different between the clinical parameters derived from the manual and our automatic segmentation.

6.5 Conclusion

In this paper, we proposed a Transformer based feature fusion network to aggregate the features from different modalities for LV segmentation in 4D flow MRI data. In

the feature fusion network, we introduced a self- and a cross-fusion layer to investigate the inter- and intra- relationship for the features from two different modalities. The proposed method was trained and evaluated in a large in-house dataset and the results of the segmentation accuracy and clinical parameters demonstrate superiority of our method against state-of-arts. We expect that the use of carefully designed data augmentation methods for the velocity images may result in further improvement of the performance of the proposed method.

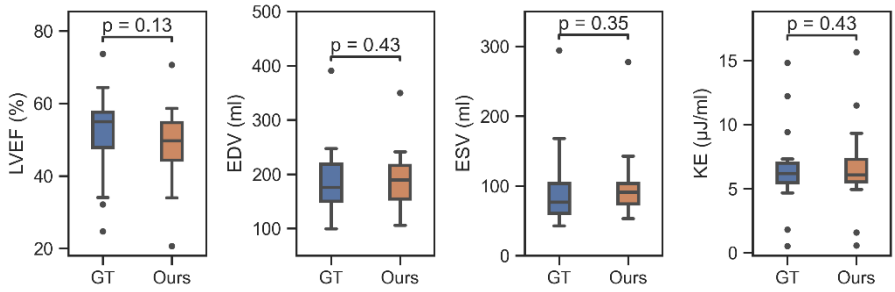

Figure.6.5. Box plots comparing four clinical evaluation metrics including EDV, ESV, LVEF and KE derived from the manual segmentation and our prediction. GT represents the ground truth. P-value was computed using Wilcoxon-signed-rank test. P<0.05 indicate a significant difference between two variables.

References

1. Tao, Q., et al.: Deep learning-based method for fully automatic quantification of left ventricle function from cine MR images: a multivendor, multicenter study. Radiology 290(1), 81–88 (2019)
2. Bai, Wenjia, et al. "Automated cardiovascular magnetic resonance image analysis with fully convolutional networks." Journal of Cardiovascular Magnetic Resonance 20.1 (2018): 1-12.
3. Garg, Pankaj, et al. "Left ventricular blood flow kinetic energy after myocardial infarction-insights from 4D flow cardiovascular magnetic resonance." Journal of Cardiovascular Magnetic Resonance 20.1 (2018): 1-15.
4. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedi- cal image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
5. Isensee, Fabian, et al. "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation." Nature methods 18.2 (2021): 203-211.
6. Vaswani, A., et al. "Attention is all you need. " Adv. Neural. Inf. Process. Syst. 30, 5998–6008 (2017)
7. Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint arXiv:2010.11929 (2020).
8. Chen, J., Lu, Y., Yu, Q., Luo, X., et al. Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021).
9. Liu, Ze, et al. "Swin transformer: Hierarchical vision transformer using shifted windows." arXiv preprint arXiv:2103.14030 (2021).
10. Cao, Hu, et al. "Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation." arXiv preprint arXiv:2105.05537 (2021).
11. Li, Hang, et al. "DT-MIL: Deformable Transformer for Multi-instance Learning on Histopathological Image." International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, 2021.
12. Luo, Yanmei, et al. "3D Transformer-GAN for High-Quality PET Reconstruction." International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, 2021.
13. Ji, Yuanfeng, et al. "Multi-Compound Transformer for Accurate Biomedical Image Segmentation." International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, 2021.
14. Berhane, Haben, et al. "Fully automated 3D aortic segmentation of 4D flow MRI for hemodynamic analysis using deep learning." Magnetic resonance in medicine 84.4 (2020): 2204-2218.

15. Wu, Yinzhe, et al. "Automated multi-channel segmentation for the 4D myocardial velocity mapping cardiac MR." Medical Imaging 2021: Computer-Aided Diagnosis. Vol. 11597. International Society for Optics and Photonics, 2021.
16. Garg, Pankaj, et al. "Left ventricular thrombus formation in myocardial infarction is associated with altered left ventricular blood flow energetics." European Heart Journal-Cardiovascular Imaging 20.1 (2019): 108-117.

Supplementary

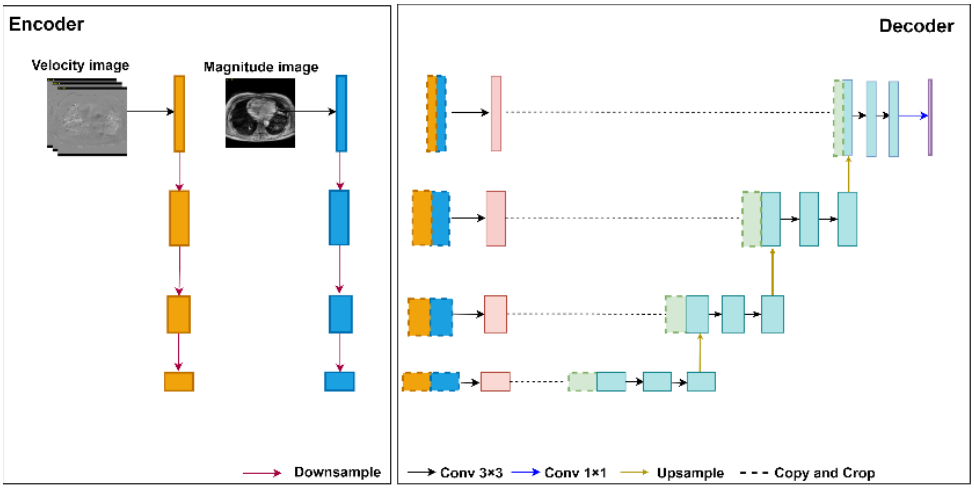

Figure. S6.1. The structure of U-NetCon.

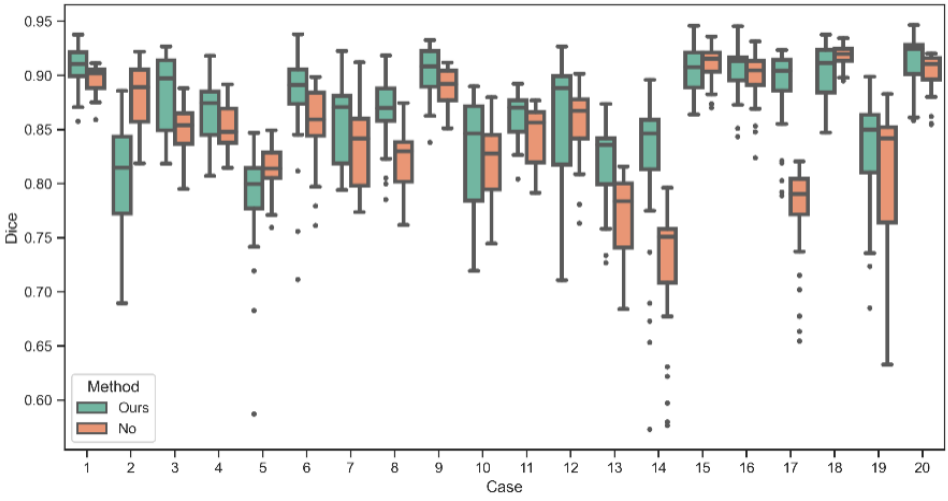

Figure. S6.2. Box plot comparing the Dice derived from our method and U-NetCon on 20 testing cases. The Dice was computed based on each phase, and each box contains 30 phases.

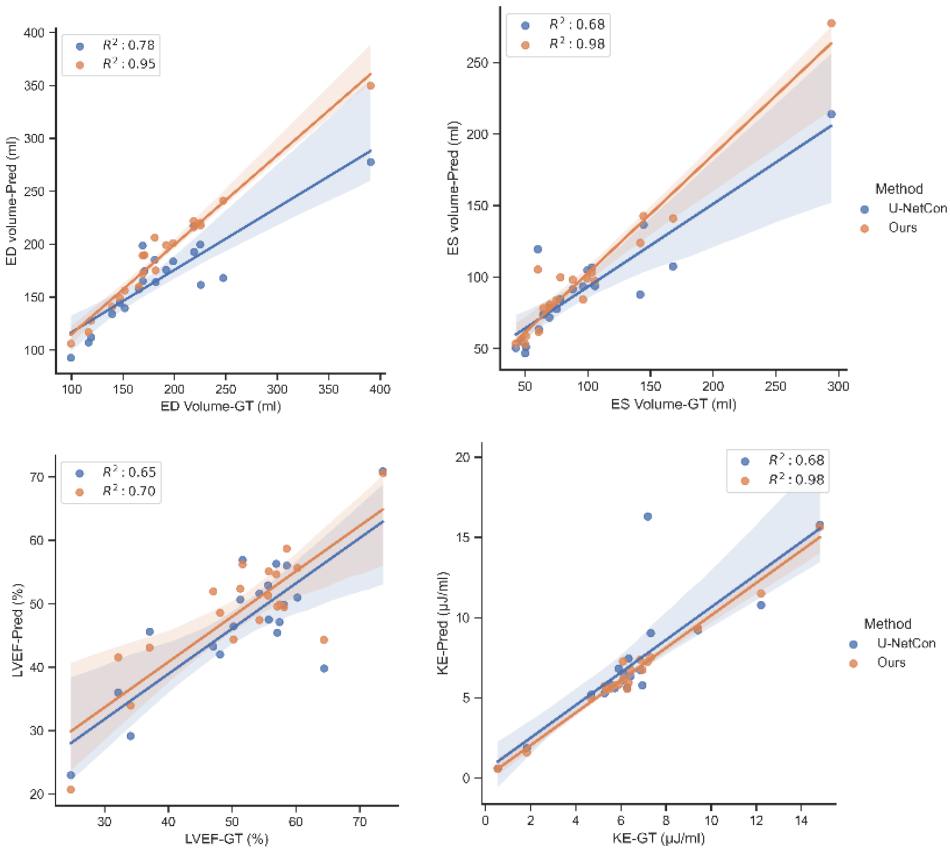

Figure. S6.3. Correlation comparing EDV, ESV, LVEF and KE derived from our method and U-NetCon. GT in the X-axis represents the parameters derived from the ground truth, the Y-axis represents the parameters derived from the prediction of different models.

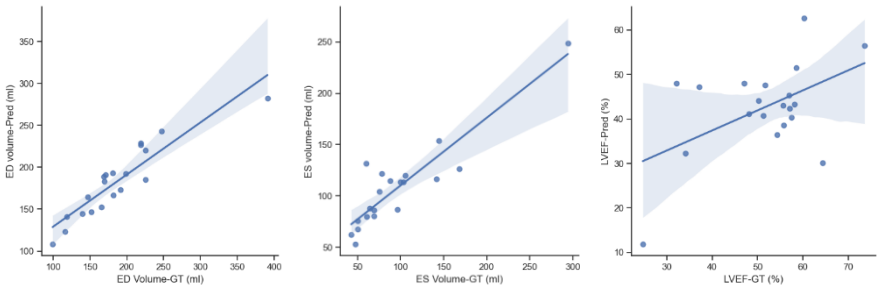

Figure. S6.4. Correlation of EDV, ESV and LVEF derived from TransUnet and ground truth.

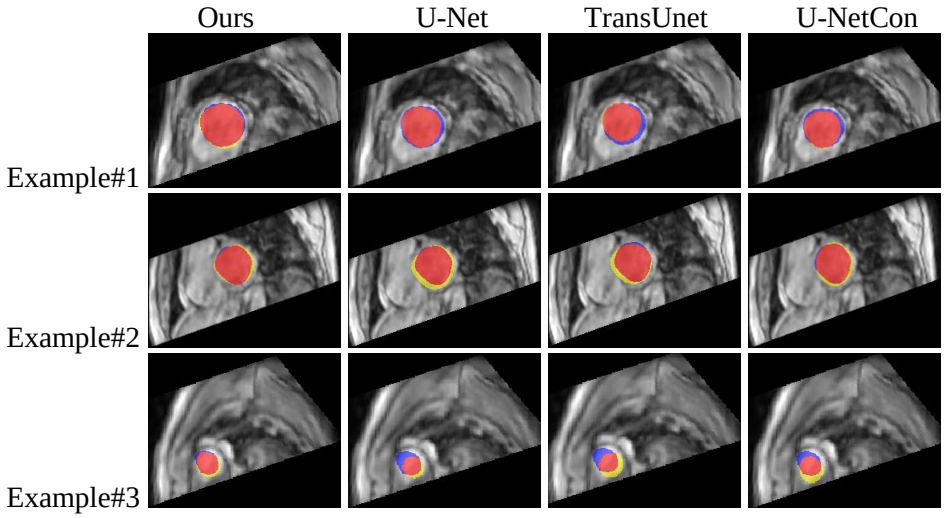

Figure. S6.5. Examples of segmentation results from our method, U-Net, TransUnet and U-NetCon. The blue represents the ground truth, the yellow is the prediction, and the red is the overlap between the prediction and ground truth

.