Self- and Cross-attention based Transformer for left ventricle segmentation in 4D flow MRI

Xiaowu Sun[1] X.SUN@LUMC.NL
Li-Hsin Cheng[1] L.CHENG@LUMC.NL
Rob J. van der Geest[*1] R.J.VAN_DER_GEEST@LUMC.NL
[1] *Division of Image Processing, Department of Radiology, Leiden University Medical Center, The Netherlands*

Abstract

The conventional quantitative analysis of 4D flow MRI relies on the co-registered cine MRI. In this work, we proposed a self- and cross-attention based Transformer to segment the left ventricle directly from the 4D flow MRI and evaluated our method on a large dataset using various metrics. The results demonstrate that self- and cross-attention improve the segmentation performance, achieving a mean Dice of 82.41%, ASD of 4.51 mm, left ventricle ejection fraction (LVEF) error of 7.96% and kinetic energy (KE) error of 1.34 μJ/ml.

Keywords: 4D flow MRI, segmentation, transformer, self-attention, cross-attention.

1. Introduction

Recently four-dimensional (4D) flow MRI has been introduced to visualize the inter-cardiac blood flow in the clinical practice. The 4D flow data provides both the structural (magnitude images) and functional (velocity images) information. Although 4D flow MRI provides more detail over the conventional cine MRI, the contrast between the cardiac cavities and the surrounding tissues is extremely poor. Therefore, the quantitative analysis of 4D flow MRI still relies on the segmentation results derived from the co-registered cine MRI. However, the breathing-related motion and heart rate difference may introduce the spatial and temporal misalignment between those two acquisitions. Additionally, magnitude and velocity image can be considered two different modalities. How to integrate those two modalities remains a challenge. Therefore, in this work, we aim to explore the Transformer-based method to segment the left ventricle (LV) directly from the 4D data without any additional cine MRI.

Our contributions are two-fold: (1) Self- and cross-attention mechanisms were introduced to explore and fuse the intra- and inter-relationship between two acquisitions. (2) The proposed model was trained and evaluated on a large dataset using various metrics including Dice, Average Surface Distance (ASD) and four clinical metrics.

2. Methods

The self-attention module, as expressed in equation (1), uses Q (Query), K (Key), V (Value) which are generated from the same input to explore the intra-relationship. While Q, K, V in cross-attention module are generated from different inputs (Chen et al., 2021). If we concatenated two modalities vertically along the spatial dimension, as illustrated in Figure 1, the input could be represented as $X = \begin{pmatrix} M \\ V_e \end{pmatrix}$ where M and V_e are the magnitude and velocity image, respectively. Then Q can be computed using a learnable linear projection as described in equation (2), where Q_m and Q_{V_e} are the Queries generated from the magnitude and velocity respectively.

$$\mathrm{Atten}\left(Q,K,V\right)=\mathrm{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{1}$$

$$Q=XW^q=\begin{pmatrix} M \\ V_e \end{pmatrix}W^q=\begin{pmatrix} MW^q \\ V_eW^q \end{pmatrix}=\begin{pmatrix} Q_m \\ Q_{V_e} \end{pmatrix} \tag{2}$$

Similarly, K and V can be represented as $K=\begin{pmatrix} K_m \\ K_{V_e} \end{pmatrix}$, $V=\begin{pmatrix} V_m \\ V_{V_e} \end{pmatrix}$. Removing the softmax and scaled function, the simplified attention mechanism in equation (1) can be expanded as follows:

$$F=QK^TV=\begin{pmatrix} Q_m \\ Q_{V_e} \end{pmatrix}\begin{pmatrix} K_m^T & K_{V_e}^T \end{pmatrix}\begin{pmatrix} V_m \\ V_{V_e} \end{pmatrix}=\begin{pmatrix} Q_mK_m^TV_m+Q_mK_{V_e}^TV_{V_e} \\ Q_{V_e}K_m^TV_m+Q_{V_e}K_{V_e}^TV_{V_e} \end{pmatrix}=\begin{pmatrix} F_m \\ F_{V_e} \end{pmatrix} \tag{3}$$

where $Q_mK_m^TV_m$ and $Q_{V_e}K_{V_e}^TV_{V_e}$ are derived from self-attention, while $Q_mK_{V_e}^TV_{V_e}$ and $Q_{V_e}K_m^TV_m$ are generated from cross-attention, F_m, F_{V_e} represent the final fused features for magnitude and velocity images. Hence, concatenating the two modalities along the vertical dimension allows the model to take into consideration the intra- and inter-modality relationships.

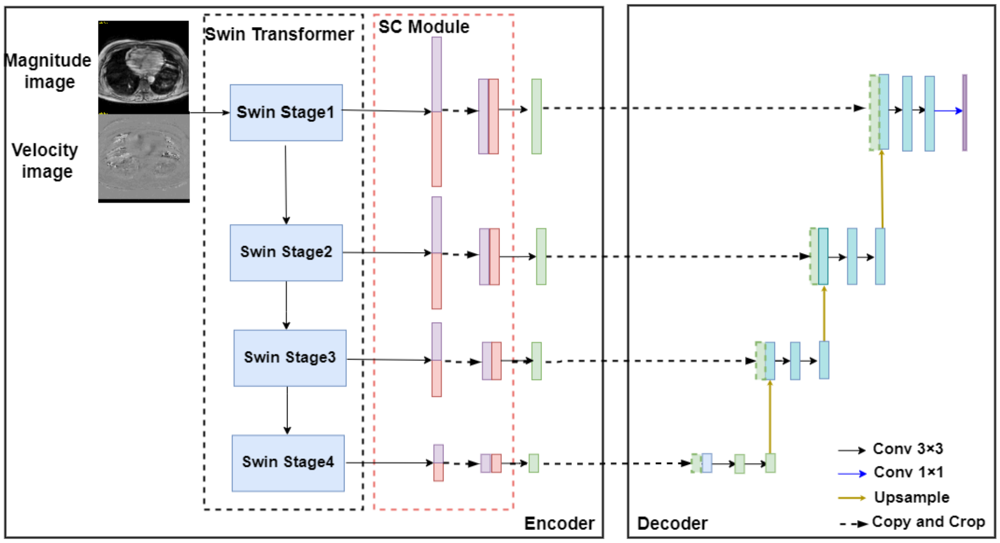

Figure 1: Proposed segmentation network. Left:Encoder part. Right:Decoder part. SC Module: Slip-and-Concatenate the features.

Figure 1 illustrates the proposed segmentation network. In the encoder part, Swin Transformer (Liu et al., 2021) was introduced as the backbone for the Encoder part. The SC module is introduced to split the features F generated from each hierarchical stage into two parts F_m and F_{V_e}, then those two feature maps were concatenated followed by a convolution layer as the output of the Encoder part. The Decoder part was kept the same as in U-Net. The sum of Dice and cross-entropy was used as the loss function.

3. Results and Conclusion

The dataset is from University of Leeds, UK. 28 healthy volunteers and 76 post-myocardial infarction patients were scanned on a 1.5T MR system (Philips Healthcare). More details about this dataset can be found here(Garg et al., 2018). We randomly split the dataset into three parts with 64, 20 and 20 cases (total number of 54 869, 17 694 and 18 619) for training, validation and testing respectively. Beside Dice and ASD, the clinical metrics including end-diastolic volume (EDV), end-systolic volume (ESV), left ventricle ejection fraction (LVEF) and kinetic energy (KE) are introduced to evaluate the performance of our proposed method. KE was normalized to EDV.

Ablation. We compared the performance of our model using the other three input integration including OM, OV and MVC respectively. OM and OV implies only the magnitude or velocity images were used as the input. MVC represents that two modalites are concatenated along the channel dimension. For these three inputs, SC Module will be removed from the Encoder. U-Net with using MVC as the input was introduced as the baseline.

The results in Table 1 show that although the baseline achieved the best performance in ASD, our proposed method with MVV as the input performed the best on the other five metrics. OM, OV and MVC introduced only the self-attention mechanism to fuse the information, while the MVV introduced both self- and cross-attention to explore all possible relationships within and across modalities.

In conclusion, self- and cross-attention is beneficial for fusing the information from both velocity and magnitude, and improving the segmentation of the 4D flow MRI.

Table 1: Comparison of the mean and standard deviation predicted by different methods.

Model	Dice (%)	ASD (mm)	EDV-Err (ml)	ESV-Err (ml)	LVEF-Err (ml)	KE-Err (μJ/ml)
U-Net	80.21±7.30	**3.84±1.66**	21.52±20.71	21.54±15.90	9.16±6.09	1.56±3.31
OM	76.79±7.92	6.15±7.82	31.23±30.40	45.32± 22.58	31.01±11.65	3.27±2.27
OV	78.78±7.92	3.97±1.60	27.54±24.74	21.72±27.60	8.44±6.90	1.87±2.24
MVC	78.17±7.43	5.65±5.36	29.06±33.10	41.74±21.39	27.69± 11.43	5.06±7.09
MVV	**82.41±6.78**	4.51±3.33	**17.59±16.46**	**18.83±14.32**	**7.96±4.02**	**1.34±2.35**

References

C. Chen, Q. Fan, and R. Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *IEEE/CVF ICCV*, pages 357–366, 2021.

P. Garg, J. Westenberg, van den Boogaard, et al. Comparison of fast acquisition strategies in whole-heart four-dimensional flow cardiac mr: Two-center, 1.5 tesla, phantom and in vivo validation study. *J. Magn. Reson. Imaging*, 47(1):272–281, 2018.

Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, et al. Swin transformer: Hierarchical vision transformer using shifted windows. In *IEEE/CVF ICCV*, pages 10012–10022, 2021.