Deformable Image Registration Using Vision Transformers for Cardiac Motion Estimation from Cine Cardiac MRI Images

Roshan Reddy Upendra[1(✉)], Richard Simon[2], Suzanne M. Shontz[3,4,5], and Cristian A. Linte[1,2]

[1] Center for Imaging Science, Rochester Institute of Technology, Rochester, NY, USA
ru6928@rit.edu
[2] Biomedical Engineering, Rochester Institute of Technology, Rochester, NY, USA
[3] Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS, USA
[4] Bioengineering Program, University of Kansas, Lawrence, KS, USA
[5] Institute for Information Sciences, University of Kansas, Lawrence, KS, USA

Abstract. Accurate cardiac motion estimation is a crucial step in assessing the kinematic and contractile properties of the cardiac chambers, thereby directly quantifying the regional cardiac function, which plays an important role in understanding myocardial diseases and planning their treatment. Since the cine cardiac magnetic resonance imaging (MRI) provides dynamic, high-resolution 3D images of the heart that depict cardiac motion throughout the cardiac cycle, cardiac motion can be estimated by finding the optical flow representation between the consecutive 3D volumes from a 4D cine cardiac MRI dataset, thereby formulating it as an image registration problem. Therefore, we propose a hybrid convolutional neural network (CNN) and Vision Transformer (ViT) architecture for deformable image registration of 3D cine cardiac MRI images for consistent cardiac motion estimation. We compare the image registration results of our proposed method with those of the VoxelMorph CNN model and conventional B-spline free form deformation (FFD) non-rigid image registration algorithm. We conduct all our experiments on the open-source Automated Cardiac Diagnosis Challenge (ACDC) dataset. Our experiments show that the deformable image registration results obtained using the proposed method outperform the CNN model and the traditional FFD image registration method.

Keywords: Vision Transformer · Cardiac MRI · Cardiac Motion Estimation · Medical Image Registration · Deep Learning

1 Introduction

The assessment of regional myocardial function such as myocardial wall deformation, strain, torsion and wall thickness, plays a crucial role in understanding,

O. Bernard et al. (Eds.): FIMH 2023, LNCS 13958, pp. 375–383, 2023.
https://doi.org/10.1007/978-3-031-35302-4_39

diagnosis, risk stratification and planning treatment of several myocardial disorders. Therefore, accurate cardiac motion estimation is an important step in assessing the kinematic and contractile properties of the myocardium, thereby, quantifying dynamic regional heart function.

Cine cardiac MRI provides high-resolution, dynamic 3D images of the cardiac chambers, which depict cardiac motion throughout the cardiac cycle. Thus, cardiac motion estimation can be formulated as an image registration problem, which involves finding an optical flow representation between the consecutive 3D frames of a 4D cine cardiac MRI dataset [19].

In the past decade, deep learning models have gained increased popularity in medical image registration [6]. A number of researchers leveraged these deep learning-based 4D deformable registration methods to estimate cardiac motion from cine cardiac MRI images [11,12,21]. In our earlier work [16], we presented a deep learning-based 4D deformable registration method for cardiac motion estimation from cine cardiac MRI dataset by leveraging the VoxelMorph framework [1]. Additionally, we demonstrated the application of the VoxelMorph-based cardiac motion estimation method to build dynamic patient-specific left ventricle (LV) myocardial models across subjects with different pathologies [17]. Although the convolutional neural network (CNN)-based cardiac motion estimation presented in our previous work [16,17] showed promising performance, the CNN-based approaches usually exhibit limitations in modeling explicit long-range spatial relations due to the limited receptive fields of convolution operations [3]. Therefore, the large variations in shape and size of the cardiac chambers can affect the registration performance of the CNN-based cardiac motion estimation methods.

In recent years, self-attention-based architectures (Transformer-based), due to their great success in sequence-to-sequence prediction in natural language processing have gained increasing interests in computer vision tasks [5], including medical image segmentation [3] and registration [4]. These current research studies show that fusing the self-attention mechanism with the CNN models overcome the limitation of the convolution operation in learning global semantic information, which is critical for the image registration task in cardiac motion estimation from the cine cardiac MRI images.

In this work, we propose a hybrid CNN-ViT architecture (Fig. 1) for consistent cardiac motion estimation from 4D cine cardiac MRI images. Here, we leverage the VIT-V-Net [4] to register the moving and fixed frame of the cardiac MRI volumes. We evaluate the proposed method by training the models on the ACDC dataset [2].

2 Methodology

2.1 Cardiac MRI Dataset

In this study, we use the 2017 ACDC dataset [2], consisting of short-axis cine cardiac MRI images from 150 subjects, divided into five equally-distributed subgroups: normal, dilated cardiomyopathy (DCM), hypertrophic cardiomyopathy

(HCM), prior myocardial infarction (MINF) and abnormal right ventricle (RV). These images were acquired as part of clinical diagnostic exams conducted on two different MRI scanners of 1.5 T and 3.0 T magnetic strength. These series of short axis MRI slices cover the LV from base to apex with a through-plane resolution of 5 mm to 10 mm and a spatial resolution of 1.37 mm^2/pixel to 1.68 mm^2/pixel.

2.2 Vision Transformer-Based Deformable Image Registration

Here, the aim of the deep learning model is to find an optical flow representation between a sequence of image volume pairs $\{(I_{ED}, I_{ED+t})\}_{t=1,2,3,\ldots,N_T-1}$ where I_{ED} is the image volume frame at end-diastole (ED) and N_T is the total number of frames for a particular subject. That is, for the given image volume pair (I_{ED}, I_{ED+t}), the deep learning model should predict a differentiable transformation function ϕ to warp the moving image volume I_{ED}, to produce a warped image volume $I_{ED} \circ \phi$. The similarity loss is computed between the fixed image volume I_{ED+t} and warped image volume $I_{ED} \circ \phi$, and this loss is used to back-propagate the deep learning network.

In this work, we employ the ViT-V-Net [4] architecture to estimate the differentiable optical flow representation (to ensure smoothness of the displacement field from ED to ES) between the image volume pairs (I_{ED}, I_{ED+t}). The direct application of ViT to the full-resolution cine cardiac MRI volume increases computational complexity. Similarly, splitting the image volume into 3D patches is not ideal, as it leads to the model not learning the local context information across the spatial and depth dimensions for volumetric registration [20]. Therefore, in ViT-V-Net [4], instead of feeding the whole high-resolution image volumes to the ViT, the image volumes are first encoded to low-resolution and high-level feature representations using a CNN encoder. Next, the high-level 3D context features are split into patches, which are then mapped onto a latent space using a trainable linear projection, i.e., patch embedding. These patch embeddings are added to the learnable position embeddings to retain the positional information of the patches, which are then fed into the ViT. The ViT consists of multiple alternating layers of multihead self-attention (MSA) and multi-layer

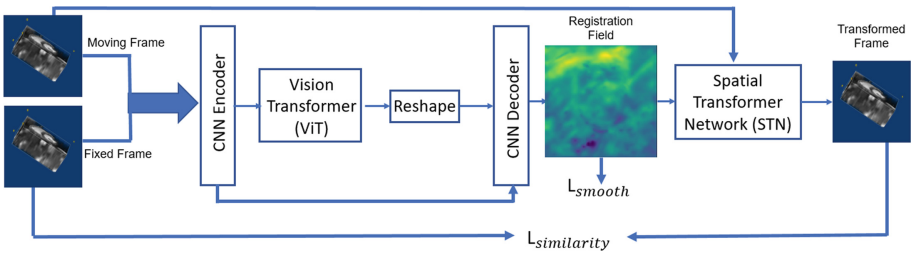


Fig. 1. Overview of the proposed hybrid CNN-ViT architecture for cardiac motion estimation.

perceptron (MLP) blocks. Finally, the output of ViT is fed into a CNN decoder to output the deformation field ϕ [4,5,18]. Also, long skip-connections are used between the V-Net [10] style encoder-decoder architecture. This deformation field is fed to the spatial transformer network (STN) [7] along with the moving image volume to produce a warped image volume (Fig. 1).

The loss function used to optimize the network described above is given by:

$$L = L_{\text{similarity}} + \lambda L_{\text{smooth}}, \tag{1}$$

where $L_{\text{similarity}}$ is the mean squared error (MSE) between the fixed image volume I_{ED+t} and the warped image volume $I_{ED} \circ \phi$. The smoothing loss L_{smooth} is the diffusion regularizer used in [4], on the spatial gradients of the deformation field ϕ, and λ is the regularization parameter.

2.3 Network Training

In order to rectify the inherent slice misalignments that occur during the cine cardiac MRI image acquisition, we train a variant of the U-Net model [13] to segment the cardiac chambers such as LV blood-pool, LV myocardium and RV blood-pool from 2D cine cardiac MRI images. We identify the centroid of these predicted segmentation maps as the LV blood-pool centers and stack the 2D MRI slices collinearly for all the frames of the cardiac cycle, resulting in slice misalignment corrected 3D images. These slice misalignment corrected 3D images were used to train all the registration algorithms reported in this work.

As mentioned earlier, we aim to find the optical flow representation between image pairs $\{(I_{ED}, I_{ED+t})\}_{t=1,2,3,\ldots,N_T-1}$. In order to do this, we employ 110 of the available 150 cardiac MRI dataset for training, 10 for validation and 30 for testing. The data-split in this work is consistent with our earlier work that involves VoxelMorph-based cardiac motion estimation [16,17], for comparison. We train our networks using an Adam optimizer with a learning rate of 10^{-4}, reduced by half every 10^{th} epoch for 50 epochs. Furthermore, all the deep learning models in this work were trained on a machine equipped with a NVIDIA RTX 2080 Ti GPU.

3 Experiments and Results

To evaluate the performance of our proposed framework for cardiac motion estimation, we compare it with the VoxelMorph [1] model, as well as the B-spline FFD non-rigid registration algorithm [14]. The VoxelMorph CNN model was trained using the same hyperparameters used for training the proposed hybrid CNN-ViT. The FFD algorithm was trained using the adaptive stochastic gradient descent optimizer, while sampling 2048 points per iteration for 500 iterations, with MSE as the similarity measure and binding energy as the smoothing loss. This FFD-based non-rigid image registration algorithm was implemented using SimpleElastix [8,9] on an Intel(R) Core(TM) i9-9900K CPU.

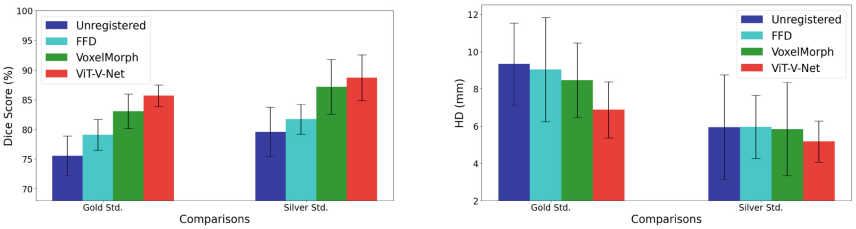
Table 1. Summary of registration evaluation on the test set (30 subjects): Unregistered (post slice misalignment correction), B-spline free form deformation (FFD), VoxelMorph and the proposed ViT-V-Net. Mean Dice score (std-dev) and Hausdorff distance (HD) for LV blood-pool (LV), LV myocardium (MC) and RV blood-pool (RV), for both "gold" and "silver" standard comparisons. Statistically significant differences between the registration metrics of VoxelMorph and ViT-V-Net registration were evaluated using the Student t-test and are reported using * for $p < 0.05$ and ** for $p < 0.01$. The best evaluation metrics achieved are labeled in **bold**.

		Dice (%)			HD (mm)		
		LV	MC	RV	LV	MC	RV
ED to ES frames: Gold standard	Unregistered	87.30 (3.20)	69.15 (2.99)	70.18 (3.85)	7.22 (1.64)	8.93 (2.72)	11.85 (2.47)
	FFD	88.94 (2.42)	74.93 (2.12)	73.38 (3.22)	6.35 (3.61)	8.87 (2.42)	11.89 (1.94)
	VoxelMorph	92.17 (4.21)	79.39 (3.22)	77.58 (1.30)	5.59 (1.21)	8.05 (2.94)	11.75 (2.11)
	ViT-V-Net	**93.31 (2.10)**	**82.24** (2.14)**	**81.27** (1.30)**	**5.11 (1.11)**	**6.50* (1.83)**	**9.02* (1.73)**
ED to all frames: Silver standard	Unregistered	81.29 (4.93)	80.15 (3.64)	77.32 (3.91)	3.13 (2.44)	6.08 (2.91)	8.61 (3.12)
	FFD	84.34 (2.34)	82.57 (1.03)	78.23 (4.10)	3.01 (1.03)	6.11 (2.89)	8.75 (1.41)
	VoxelMorph	**94.67 (5.96)**	84.08 (4.32)	82.73 (3.76)	**2.51 (1.31)**	6.07 (2.79)	8.96 (3.47)
	ViT-V-Net	93.67 (4.61)	**88.53** (3.78)**	**83.92* (3.17)**	2.66 (1.42)	**5.02* (0.93)**	**7.83* (1.08)**

We evaluate the performance of all our models by warping the segmentation map of the ED frame to the end-systole (ES) frame using the estimated registration field, and computing the Dice score and Hausdorff distance (HD) between ground truth ES segmentation map of the cardiac chambers, namely LV blood-pool, LV myocardium and RV blood-pool, and warped segmentation map of the ED frame. Since the segmentation maps of the ED and ES frame are manually annotated by experts, we refer to this comparison as the "gold" standard comparison.

Additionally, we warp the segmentation map of the ED frame to all the subsequent frames of the cardiac cycle, and compute the evaluation metrics between the warped segmentation map of the ED frame and segmentation maps predicted by the U-Net model (as described in Sect. 2.3). Since the segmentation maps used here were generated using techniques that were previously validated against expert annotations, we refer to it as "silver" standard comparison. These results are shown in Table 1.

In Table 1, we show that our proposed method achieved a mean Dice score of 85.67% and a mean HD of 6.87 *mm* for our "gold" standard comparison, and a mean Dice score of 88.71% and a mean HD of 5.17 *mm* for our "silver" standard comparison.



(a) (b)

Fig. 2. "Gold" and "silver" comparison of (a) mean Dice score, and (b) mean HD values before registration (post slice misalignment correction), B-spline free form deformation (FFD) registration, VoxelMorph and the proposed ViT-V-Net, on the test set (30 subjects).

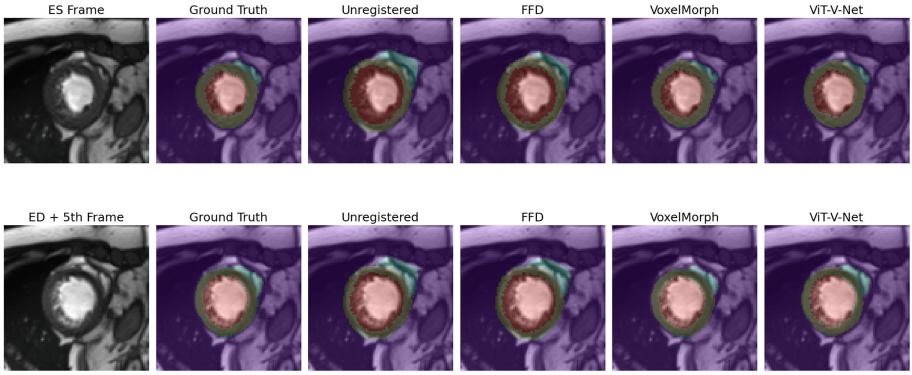


Fig. 3. Panel 1-1: End-systole (ES) frame; Panel 1-2: ground truth segmentation map of the cardiac chambers overlaid on the slice; Panel 1-3: segmentation map of the end-diastole (ED) frame overlaid on the ES frame without registration (Dice: 77.29%, HD: 9.09 mm); Panel 1-4: post registration contours using FFD algorithm (Dice: 79.02%, HD: 8.99 mm); Panel 1-5: post registration contours using VoxelMorph model (Dice: 83.11%, HD: 7.01 mm); Panel 1-6: post registration contours using ViT-V-Net framework (Dice: 85.57%, HD: 6.82 mm). Panel 2-1: ED + 5^{th} frame; Panel 2-2: U-Net predicted segmentation map of the cardiac chambers overlaid on the slice; Panel 2-3: segmentation map of the end-diastole (ED) frame overlaid on the ED + 5^{th} frame without registration (Dice: 78.07%, HD: 5.55 mm); Panel 2-4: post registration contours using FFD algorithm (Dice: 81.03%, HD: 4.92 mm); Panel 2-5: post registration contours using VoxelMorph model (Dice: 85.03%, HD: 3.97.01 mm); Panel 2-6: post registration contours using ViT-V-Net framework (Dice: 86.84%, HD: 3.28 mm). The Dice score and HD reported here are the average of the LV blood-pool, LV myocardium and RV blood-pool registration results.

We can observe that the proposed hybrid CNN-ViT model outperforms the CNN-only VoxelMorph model, as well as the FFD registration method (Fig. 2). In Fig. 3, we show an example of the cardiac chamber contours propagated using

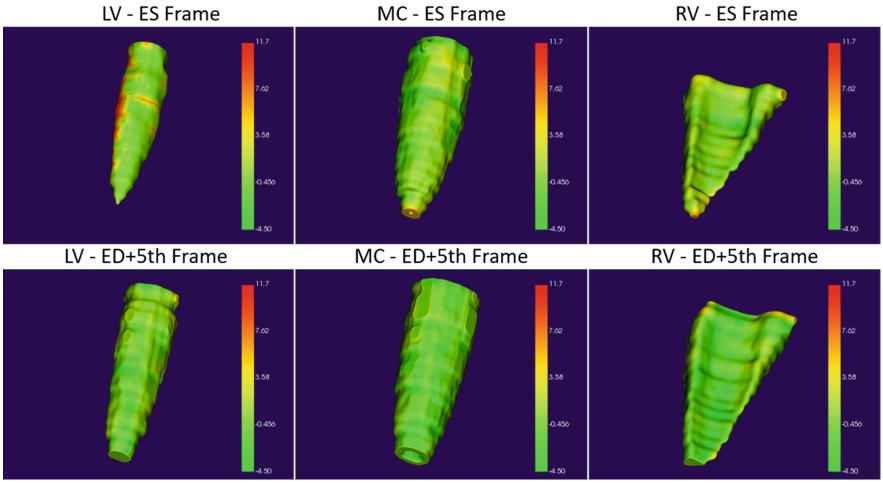


Fig. 4. Model-to-model distance between the isosurface mesh generated from VoxelMorph and ViT-V-Net propagation at end-systole (ES) frame (*top row*) and end-diastole (ED)+5^{th} frame (*bottom row*) for (*left to right*) left ventricle blood-pool (LV), left ventricle myocardium (MC) and right ventricle blood-pool (RV)

the registration methods from the ED frame to the ES frame, as well as the ED + 5^{th} frame. Additionally, in Fig. 4, we show an example of the model-to-model distance between the isosurface meshes of the cardiac chambers propagated using VoxelMorph framework and the proposed hybrid CNN-ViT framework from the ED frame to the ES frame, as well as the ED + 5^{th} frame. Here, we can observe that the two sets of isosurface meshes are in close agreement with each other.

4 Discussion and Conclusion

In this paper, we present a hybrid CNN-ViT deformable image registration method for consistent cardiac motion estimation from 3D cine cardiac MRI images. To the best of our knowledge, this is the first study to investigate the usage of ViT for cardiac motion estimation. In addition to the local context information learnt by the CNN encoder-decoder layers, the ViT encodes global context information by treating the CNN-encoded features as sequences.

We evaluate the performance of the proposed hybrid CNN-ViT framework by comparing it with the VoxelMorph framework, which is essentially a CNN encoder-decoder architecture without the ViT. We observe that the proposed hybrid framework outperforms the VoxelMorph framework for cardiac motion estimation from cine cardiac MRI images (Fig. 2).

In our earlier work [15,17], we showed that the VoxelMorph framework can be used to build patient-specific LV myocardial and RV models, respectively. However, thanks to the improved registration accuracy of the proposed method compared to the VoxelMorph model, this work will enable us to generate more

accurate patient-specific cardiac models featuring improved mesh (isosurface and volumetric) quality. As such, as part of our future work, we will demonstrate how the cardiac motion estimated using this proposed method may be used to build high quality, deformable patient-specific geometric models of cardiac chambers from cine cardiac MRI.

Acknowledgment. Research reported in this publication was supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award No. R35GM128877 and by the Office of Advanced Cyberinfrastructure of the National Science Foundation under Award No. 1808530 and Award No. 1808553.

References

1. Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V.: VoxelMorph: a learning framework for deformable medical image registration. IEEE Trans. Med. Imaging **38**(8), 1788–1800 (2019)
2. Bernard, O., et al.: Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? IEEE Trans. Med. Imaging **37**(11), 2514–2525 (2018)
3. Chen, J., et al.: TransUnet: transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
4. Chen, J., He, Y., Frey, E.C., Li, Y., Du, Y.: ViT-V-net: vision transformer for unsupervised volumetric medical image registration. arXiv preprint arXiv:2104.06468 (2021)
5. Dosovitskiy, A., et al.: An image is worth 16×16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
6. Haskins, G., Kruger, U., Yan, P.: Deep learning in medical image registration: a survey. Mach. Vision Appl. **31**(1), 1–18 (2020)
7. Jaderberg, M., Simonyan, K., Zisserman, A., et al.: Spatial transformer networks. Adv. Neural Inf. Process. Syst. **28** (2015)
8. Klein, S., Staring, M., Murphy, K., Viergever, M.A., Pluim, J.P.: Elastix: a toolbox for intensity-based medical image registration. IEEE Trans. Med. Imaging **29**(1), 196–205 (2009)
9. Marstal, K., Berendsen, F., Staring, M., Klein, S.: SimpleElastix: a user-friendly, multi-lingual library for medical image registration. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 134–142 (2016)
10. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV), pp. 565–571. IEEE (2016)
11. Qin, C., et al.: Joint motion estimation and segmentation from undersampled cardiac MR image. In: Knoll, F., Maier, A., Rueckert, D. (eds.) MLMIR 2018. LNCS, vol. 11074, pp. 55–63. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00129-2_7
12. Qiu, H., Qin, C., Le Folgoc, L., Hou, B., Schlemper, J., Rueckert, D.: Deep learning for cardiac motion estimation: supervised vs. unsupervised training. In: Pop, M., et al. (eds.) STACOM 2019. LNCS, vol. 12009, pp. 186–194. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-39074-7_20

13. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
14. Rueckert, D., Sonoda, L.I., Hayes, C., Hill, D.L., Leach, M.O., Hawkes, D.J.: Non-rigid registration using free-form deformations: application to breast MR images. IEEE Trans. Med. Imaging **18**(8), 712–721 (1999)
15. Upendra, R.R., et al.: Motion extraction of the right ventricle from 4D cardiac cine MRI using a deep learning-based deformable registration framework. In: 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pp. 3795–3799. IEEE (2021)
16. Upendra, R.R., Wentz, B.J., Shontz, S.M., Linte, C.A.: A convolutional neural network-based deformable image registration method for cardiac motion estimation from cine cardiac MR images. In: 2020 Computing in Cardiology, pp. 1–4. IEEE (2020)
17. Upendra, R.R., Wentz, B.J., Simon, R., Shontz, S.M., Linte, C.A.: CNN-based cardiac motion extraction to generate deformable geometric left ventricle myocardial models from cine MRI. In: Ennis, D.B., Perotti, L.E., Wang, V.Y. (eds.) FIMH 2021. LNCS, vol. 12738, pp. 253–263. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-78710-3_25
18. Vaswani, A., et al.: Attention is all you need. Adv. Neural Inf. Process. Syst. **30** (2017)
19. Wang, H., Amini, A.A.: Cardiac motion and deformation recovery from MRI: a review. IEEE Trans. Med. Imaging **31**(2), 487–503 (2011)
20. Wang, W., Chen, C., Ding, M., Yu, H., Zha, S., Li, J.: TransBTS: multimodal brain tumor segmentation using transformer. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12901, pp. 109–119. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87193-2_11
21. Zhang, X., You, C., Ahn, S., Zhuang, J., Staib, L., Duncan, J.: Learning correspondences of cardiac motion from images using biomechanics-informed modeling. arXiv preprint arXiv:2209.00726 (2022)