Importance Weighted Variational Cardiac MRI Registration Using Transformer and Implicit Prior

Kangrong Xu[1,2,3], Qirui Huang[1,2,3], and Xuan Yang[1,2,3(✉)]

[1] Shenzhen University, Shenzhen 518060, Guangdong, China
yangxuan@szu.edu.cn
[2] Guangdong Provincial Key Laboratory of Popular High Performance Computers, Shenzhen, Guangdong, China
[3] Shenzhen Key Laboratory of Service Computing and Application, Shenzhen, Guangdong, China

Abstract. The variational registration model takes advantage of explaining uncertainties of registration results. However, most existing variational registration models are based on convolutional neural networks (CNNs), which cannot capture distant information in images. Besides, the evidence lower bound (ELBO) and the commonly used standard prior cannot close the gap between the real posterior and the variational posterior in the vanilla variational registration model. This paper proposes a network in a variational image registration model for cardiac motion estimation to effectively capture the spatial correspondence of long-distance images and solve the shortcomings of CNNs. Our proposed network comprises a Transformer with a T2T module and the cross attention between the moving and the fixed images. To close the gap between the real posterior and the variational posterior, the importance-weighted evidence lower bound (iwELBO) is introduced into the variational registration model with an implicit prior. The coefficients of a parametric transformation using multi-supports CSRBFs are latent variables in our variational registration model, which improve registration accuracy significantly. Experimental results show that the proposed method outperforms state-of-arts research on public cardiac datasets.

Keywords: Variational inference · Transformer · Cross attention · Compact support radial basis function (CSRBF)

1 Introduction

Cardiac motion estimation is vital in evaluating cardiac function, detecting heart diseases, and understanding cardiac biomechanics. Deformable image registration (DIR) is the critical technique of cardiac motion estimation. It minimizes the

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43999-5_55.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14229, pp. 581–591, 2023.
https://doi.org/10.1007/978-3-031-43999-5_55

differences between the warped moving and fixed images to estimate a displacement vector field (DVF). Unsupervised deep-learning-based image registration has recently become mainstream due to the required non-annotation data [4,16] and rapid inference performance when the network is well-trained.

The probabilistic generative model shows potential in the unsupervised learning registration [11–13,17,18,22]. It allows the registration framework to be highly adaptable and can be applied to cases with a small amount of data and anatomical variability. Another advantage of the probabilistic generative registration model is that it can provide registration uncertainties [1], which plays a vital role in clinical application [14,21]. However, several issues exist with variational image registration approaches. The first is that traditional convolutions are limited in representing long-range relationships between image features. The second issue is a gap between the objective function ELBO and the log-likelihood of input image pairs, which deteriorates registration accuracy. Besides, non-parametric transformation is commonly used [5,20,27], which has the challenge of regularizing the DVF smooth and topology-preserving.

This paper proposed a novel variational image registration model to cope with the above issues by employing the Transformers with the cross-attention mechanism and introducing an importance-weighted ELBO (iwELBO) [7] with an implicit prior. Detailed contributions of our work include:

- A novel VAE network architecture is proposed, which employs the Transformer architecture to focus on cross-attention between the moving and fixed images. The predictive results of the transformation parameter distribution using our architecture are more accurate than traditional VAE architecture.
- We optimized the importance-weighted ELBO in the variational image registration model. We use approximated aggregated posterior as the prior to regularizing posterior. To our best knowledge, we are the first to combine the iwELBO and aggregated posterior to close the gap between the real and variational posterior.
- A parametric transformation based on multi-supports CSRBFs is embedded in our variational registration model. By imposing a sparse constraint, the coefficients of multi-CSRBFs are regularized to be sparse to select the optimal support for multi-support CSRBFs. The parametric transformation model improves registration accuracy significantly and makes it easy to regularize the smoothness of DVFs.

2 Proposed Method

2.1 Importance-Weighted Variational Image Registration Model

Given the moving and fixed images M, F, and n control points $\{p_i\}_{i=1}^n$, the parametric transformation based on multi-supports CSRBFs is $f_{\boldsymbol{z}}(\upsilon) = \upsilon + \sum_{i=1}^{n}\sum_{k=1}^{s} z_{i,k}\phi(\frac{\|\upsilon - p_i\|_2}{c_k})$, where $\phi(\frac{\|\cdot\|}{c})$ is a CSRBF with support c; $\|\upsilon - p_i\|_2$ is the Euclidean distance between the pixel υ and p_i. s different supports $\{c_k\}_{k=1}^s$

are provided for each CSRBF. $\boldsymbol{z} = \{\boldsymbol{z}_k\}_{k=1}^{s}$, $\boldsymbol{z}_k = \{z_{k,i}\}_{i=1}^{n}$ is the latent variable whose distribution is required to be estimated. The parametric transformation can control deformations using different supports. By imposing sparse constraints, selecting the optimal support from these given supports is possible, leading to more flexible deformation results.

The variational registration model aims to estimate the posterior of $p(\boldsymbol{z}|F, M)$. In the vanilla variational registration model, $q(\boldsymbol{z}|F, M)$ is estimated to approximate $p(\boldsymbol{z}|F, M)$ by optimizing $ELBO = \mathbb{E}_{\boldsymbol{z}\sim q(\boldsymbol{z}|F,M)}\left[\log \frac{p(F|\boldsymbol{z},M)p(\boldsymbol{z})}{q(\boldsymbol{z}|F,M)}\right]$, where $p(F|\boldsymbol{z}, M)$ is the probability of occurring F when the moving image M is deformed using a transformation $f_{\boldsymbol{z}}$ with latent variables $\boldsymbol{z}$. It can be expressed as Boltzmann distribution $p(F|\boldsymbol{z}, M) \propto e^{sim(F,M(f_{\boldsymbol{z}}))}$ using a similarity metric sim. $p(\boldsymbol{z})$ is the prior of $\boldsymbol{z}$.

The importance-weighted evidence lower bound (iwELBO) [7] is defined as,

$$iwELBO = \mathbb{E}_{\boldsymbol{z}^1,\ldots,\boldsymbol{z}^K\sim q(\boldsymbol{z}|F,M)}\left[\log \frac{1}{K}\sum_{k=1}^{K}\frac{p(F|\boldsymbol{z}^k, M)p(\boldsymbol{z}^k)}{q(\boldsymbol{z}^k|F,M)}\right] \tag{1}$$

where $\boldsymbol{z}^1, \ldots, \boldsymbol{z}^K$ are K-samples of latent variable $\boldsymbol{z}$ sampled from $q(\boldsymbol{z}|F, M)$. It is assumed that $q(\boldsymbol{z}|F, M) \sim \mathcal{N}(\boldsymbol{\mu}(F, M), \boldsymbol{\Sigma}(F, M))$, and $\boldsymbol{z}^1, \ldots, \boldsymbol{z}^K$ is sampled as $\boldsymbol{z}^k = \boldsymbol{\mu}(F, M) + \boldsymbol{\Sigma}(F, M)\boldsymbol{\epsilon}_k$, $\boldsymbol{\epsilon}_k \sim \mathcal{N}(0, \boldsymbol{I})$ using the reparametrization trick.

Denote $w_k = \frac{p(F|\boldsymbol{z}^k,M)p(\boldsymbol{z}^k)}{q(\boldsymbol{z}^k|F,M)}$, the gradient of iwELBO can be interpreted as normalized importance-weighted average gradients of each sample, which implies the sample with larger w_k contributes more to iwELBO. It is challenging to compute w_k directly due to the high dimensional latent variable $\boldsymbol{z}^k$. We tackle this problem by a trick.

$$\arg\max\ iwELBO = \arg\max \mathbb{E}_{\boldsymbol{z}^1,\ldots,\boldsymbol{z}^K\sim q(\boldsymbol{z}|F,M)}\left[\log \frac{1}{K}\sum_{k=1}^{K} e^{\log w_k \frac{1}{\lambda}}\right], \lambda > 0. \tag{2}$$

Because $\log w_k \propto sim(F, M(f_{\boldsymbol{z}^k})) + \log \frac{p(\boldsymbol{z}^k)}{q(\boldsymbol{z}^k|F,M)}$, our objective function is

$$\mathbb{E}_{\boldsymbol{z}^1,\ldots,\boldsymbol{z}^K\sim q(\boldsymbol{z}|F,M)}\left[\log \frac{1}{K}\sum_{k=1}^{K} e^{sim(F,M(f_{\boldsymbol{z}^k}))+\frac{1}{\lambda}\log\frac{p(\boldsymbol{z}^k)}{q(\boldsymbol{z}^k|F,M)}}\right]. \tag{3}$$

iwELBO is a tighter evidence lower bound of the log-likelihood of data; it approaches the log-likelihood $\log p(F|M)$ as $K \to \infty$. From the view of image registration, the iwELBO tends to converge to a transformation with the optimal w_k from K samples. When a large hyperparameter λ is used, $\frac{1}{\lambda}\log\frac{p(\boldsymbol{z}^k|M)}{q(\boldsymbol{z}^k|F,M)}$ is relative smaller compared with the similarity term. That implies the iwELBO prefers the sample $\boldsymbol{z}^k$ leading to the optimal similarity between the warped moving and fixed images, which is beneficial to push the network to predict more accurate $\boldsymbol{z}$. Besides, Huang *et al.* [15] pointed out that when only one sample corresponding to a high loss is drawn to estimate the iwELBO, iwELBO

tolerates this mistake due to the importance of weight. On the contrary, the sample with high loss will be highly penalized in traditional ELBO because the decoder treats the sample as real, observed data.

2.2 Aggregate Posterior as the Prior

A simple prior, such as the standard Normal in VAE, incurred over-regularization on the posterior and widened the gap between the variational posterior and the real posterior. Many researchers resolve this mismatch by proposing various priors [2,10,23–25]. Tomczak *et al.* stated that the optimal prior in VAE is the aggregated posterior of data. Takahashi *et al.* [23] introduced the density ratio trick to estimate this aggregated posterior implicitly. However, all these works are evaluated based on ELBO instead of iwELBO. We derive and approximate the optimal prior based on iwELBO, please refer to part 1 in the supplement.

The optimal prior should maximize the expectation of iwELBO:

$$\arg\max_{p(z)} \int p(F,M)\mathbb{E}_{\boldsymbol{z}^1,..,\boldsymbol{z}^K\sim q(\boldsymbol{z}|F,M)}\left[\log\frac{1}{K}\sum_{k=1}^{K} e^{sim(F,M(f_{\boldsymbol{z}^k}))+\frac{1}{\lambda}\log\frac{p(\boldsymbol{z}^k)}{q(\boldsymbol{z}^k|F,M)}}\right] d(F,M) \tag{4}$$

It can be derived that the optimal prior $p^*(z)$ is approximated as the aggregated posterior $\mathbb{E}_{p(F,M)}q(\boldsymbol{z}|F,M)$. Substituting the optimal prior $p^*(\boldsymbol{z})$ into Eq. (3), the objective function is rewritten as

$$\mathbb{E}_{\boldsymbol{z}^1,..,\boldsymbol{z}^K\sim q(\boldsymbol{z}|F,M)}\left[\log\frac{1}{K}\sum_{k=1}^{K} e^{sim(F,M(f_{\boldsymbol{z}^k}))+\frac{1}{\lambda}\log\frac{p_0(\boldsymbol{z}^k)}{q(\boldsymbol{z}^k|F,M)}+\frac{1}{\lambda}\log\frac{p^*(\boldsymbol{z}^k)}{p_0(\boldsymbol{z}^k)}}\right]. \tag{5}$$

where $p_0(\boldsymbol{z})$ is a simple given prior. To estimate the density ratio $\log\frac{p^*(\boldsymbol{z})}{p_0(\boldsymbol{z})}$, a binary discriminator $T(\mathbf{z})$ is trained by maximizing [23],

$$\max \mathbb{E}_{p^*(\boldsymbol{z})}[\sigma(T(\mathbf{z}))] + \mathbb{E}_{p_0(\boldsymbol{z})}[\log(1-\sigma(T(\mathbf{z})))] \tag{6}$$

where σ is the sigmoid function. The discriminator is a neural network composed of four fully connected layers, with the final layer outputting density ratio. A dropout layer is added before the output to prevent the discriminator network from overfitting. When $T(\mathbf{z})$ is well trained, $\log\frac{p^*(\boldsymbol{z}^k)}{p_0(\boldsymbol{z}^k)} \approx T(\mathbf{z}^k)$. The given prior $p_0(\boldsymbol{z})$ is defined as $\mathcal{N}(0,\boldsymbol{B}^{-1})$, where $\boldsymbol{B} = diag(B^1,\cdots,B^s)$, B^k is a $n\times n$ matirx with the entries $B^k_{ij} = \phi(\frac{\|p_i-p_j\|}{c_k})$, $k=1,\ldots,s$. Then,

$$\begin{aligned}\log\frac{p_0(\boldsymbol{z}^k)}{q(\boldsymbol{z}^k|F,M)} &= \log|\boldsymbol{\Sigma}(F,M)| + \log|\boldsymbol{B}| \\ &\quad - \frac{1}{2}\left[{\boldsymbol{z}^k}^T\boldsymbol{B}\boldsymbol{z}^k - (\boldsymbol{z}^k-\boldsymbol{\mu}(F,M))^T\boldsymbol{\Sigma}^{-1}(F,M)(\boldsymbol{z}^k-\boldsymbol{\mu}(F,M))\right].\end{aligned} \tag{7}$$

where ${\boldsymbol{z}^k}^T\boldsymbol{B}\boldsymbol{z}^k$ is the bending energy of DVF using multi-supports CSRBFs aiming to regularize DVF smooth. The sampling size for k is 5; λ is 110000.

We optimize our network by iterating a two-step procedure. The encoder is updated using Eq. (7) by fixing the discriminator. Next, the discriminator is updated using Eq. (6) by fixing the encoder. Above two steps are performed alternatively.

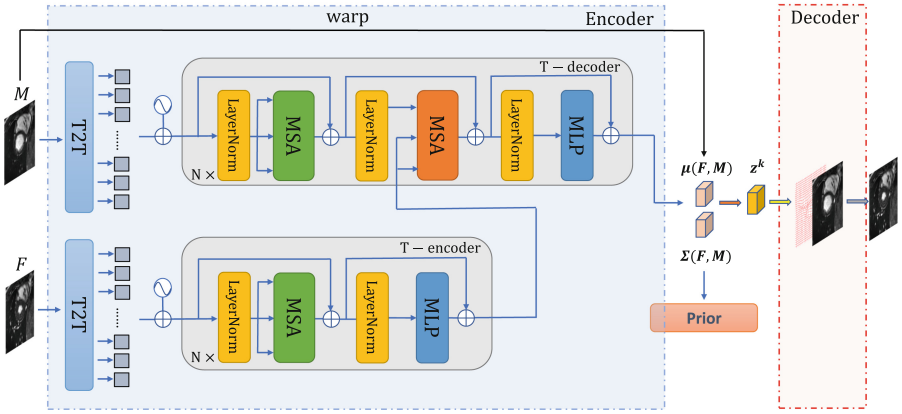

Fig. 1. The architecture of our network. The moving and fixed images M and F are preprocessed by T2T modules first and then input to our Transformer's T-encoder and T-decoder, respectively. Self-MSA and cross-MSA In our Transformer are marked by green and orange, respectively. (Color figure online)

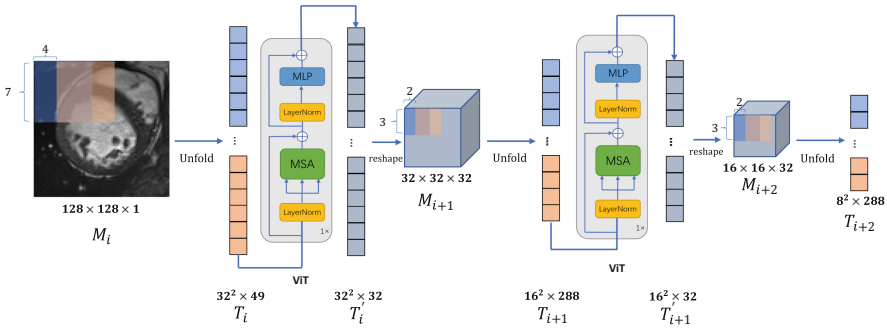

Fig. 2. T2T Module with three iterations.

2.3 Network

Our network architecture consists of an encoder and a decoder, as shown in Fig. 1. The encoder composes of T2T modules [26] and a Transformer to predict $\boldsymbol{\mu}(F, M)), \boldsymbol{\Sigma}(F, M)$. T2T modules preprocess M and F and then input to our Transformer's T-encoder and T-decoder, which pays self-attention and

cross-attention of M and F, respectively. Outputs of the cross-MSA (marked by orange) are fused information of M and F, weighted by similarity. More details of the Transformer Network can be seen in part 2 of the supplementary materials. The inherent ability to capture the correlation between two images makes Transformers easier to extract effective features for image registration. The decoder of our network generates the warped moving image using the DVF obtained by transformation based on multi-supports CSRBFs. The number of Transformer blocks N is 3.

T2T modules tackle the partitioning issue in Transformer by modeling the local structure information iteratively. As shown in Fig. 2, overlapped patches are encoded by an unfold operation as a vector token T_i and re-encoded as T_i' using a ViT. T_i' is reshaped as M_{i+1} with the size of overlapped patches. The above process is repeated three times to obtain the final vector T_{i+2}, which is the input of our Transformer.

The total loss function of our network $\mathcal{L}$ is combing the iwELBO and sparse constraints on $\boldsymbol{z}$ as $\mathcal{L} = iwELBO - \|\boldsymbol{z}\|_1$.

3 Experiments

Four public cardiac datasets are used to evaluate our method, including MICCAI2009 [19], York [3], ACDC [6], and M&Ms [8]. We combine MICCAI, York, and ACDC as a dataset denoted as ACDC+. Contours of the left ventricle (LV), the right ventricle (RV), or myocardium (MYO) at the end-diastolic (ED) phase or the end-systolic (ES) phase are provided by experts for different datasets. We take the ED and ES images as moving and fixed images, respectively. The training, validation, and testing slices are 1257, 130, 698 for ACDC+ and 1134, 266, 1030 for M&Ms. All images are cropped into the size of 128×128 containing the heart. Data augmentations such as flips and rotations are used. The LCC with the size of 9×9 is employed as the similarity metric. 64 global control points are evenly spaced on the 128×128 image, while 100 local control points are evenly spaced in an area of 64×64 in the center of the image that contains the heart. Our network is trained using PyTorch on a computer equipped with an Intel(R) Xeon(R) Silver 4210 CPU and Nvidia RTX 2080Ti GPU. The Adam optimizer with a learning rate of $5e^{-4}$ is employed. We use the estimated DVFs to map the contours of the moving image and compare the mapped contours with the contours of the fixed image using various metrics, including the Dice score, the average perpendicular distance (APD, in mm), and 95%-tile Hausdorff Distance (HD, in mm). Moreover, we count the number of anomalies to measure the topological property of the DVFs $|J_{f_z}| \leq 0$ and calculate the bending energy (BE, $\times 10^{-4}$) of DVFs to measure their smoothness.

3.1 Results

To compare the performance of our method with unsupervised registration networks KrebsDiff [17], DalcaDiff [11], NetGI [13], VoxelMorph [4], and Trans-

morph [9]. KrebsDiff, DalcaDiff, and NetGI are networks of variational registration. Transmorph is a network embedding Transformers. VoxelMorph is a vanilla unsupervised registration network. Registration results of two datasets using different networks are listed in Table 1. Our network outperforms other networks regarding Dice, HD, and APD. NetGI is the most similar registration model to our method, achieving the smoothest DVFs, while DalcaDiff preserved the topology of DVF best due to the diffeomorphism deformation it used. The bending energy and topology-preserving of DVFs using our network are close to that of NetGI and DalcaDiff, which implies that the transformation model based on multi-supports CSRBFs is good at generating smooth and topology-preserving DVFs. Visualization of the registration results using different networks is shown in Fig. 3. The myocardium of the fixed image is marked by green, while the warped moving image is marked by blue. The overlap of the myocardium is marked by red. Here, registration results of the basal, middle, and apical slices are provided. Note that objects in apical slices are small, while our network matches the small myocardium better than other networks.

Table 1. Comparison of registration results on ACDC+ and M&Ms datasets using different networks. Data format: mean (standard deviations)

Dataset	Method	BE	$\lvert J_{f_z}\rvert \leq 0$	Dice	HD	APD
ACDC+	KrebsDiff [17]	27.71 (15.97)	2.01 (14.67)	0.840 (0.063)	5.66 (1.88)	2.21 (0.77)
	DalcaDiff [11]	165.60 (40.34)	**0.01 (0.04)**	0.849 (0.064)	5.78 (1.93)	2.09 (0.81)
	VoxelMorph [4]	149.26 (32.05)	26.17 (21.08)	0.845 (0.066)	5.84 (1.95)	2.15 (0.84)
	NetGI [13]	**12.68 (5.91)**	4.96 (14.45)	0.854 (0.057)	5.46 (1.87)	2.00 (0.67)
	TransMorph [9]	104.78 (27.16)	20.12 (27.55)	0.850 (0.064)	5.71 (1.89)	2.08 (0.82)
	Ours	24.09 (6.04)	0.84 (4.28)	**0.867** (0.049)	**5.17** (1.57)	**1.87**(0.58)
M&Ms	KrebsDiff [17]	27.90 (15.97)	4.92 (10.09)	0.828 (0.054)	4.57 (2.24)	1.82 (0.93)
	DalcaDiff [11]	90.15 (55.86)	**0.01 (0.05)**	0.853 (0.050)	4.21 (2.07)	1.53 (0.82)
	VoxelMorph [4]	253.47 (59.40)	36.73 (25.40)	0.848 (0.059)	4.48 (2.34)	1.60 (0.92)
	NetGI [13]	**12.14 (4.47)**	0.18 (0.93)	0.847 (0.052)	4.32 (2.38)	1.63 (0.94)
	TransMorph [9]	169.36 (44.52)	33.08 (38.52)	0.860 (0.052)	4.13 (2.05)	1.48 (0.74)
	Ours	21.23 (5.64)	0.77 (1.56)	**0.869 (0.042)**	**3.84 (1.81)**	**1.41(0.65)**

3.2 Ablation Study

Ablation experiments are performed on the ACDC+ dataset to validate the influence of different components in our method. Table 2 lists evaluation results using the different combinations of components. Using ELBO as the objective function, the transformation based on multi-supports CSRBFs improves Dice 5%. Dice is improved 3% when the aggregated posterior is used as the prior. Whether the standard normal or the aggregated posterior as prior, the importance-weighted ELBO improves about 2–3% in Dice compared with ELBO. It is noted that when iwELBO is used, the aggregated posterior cannot improve registration compared

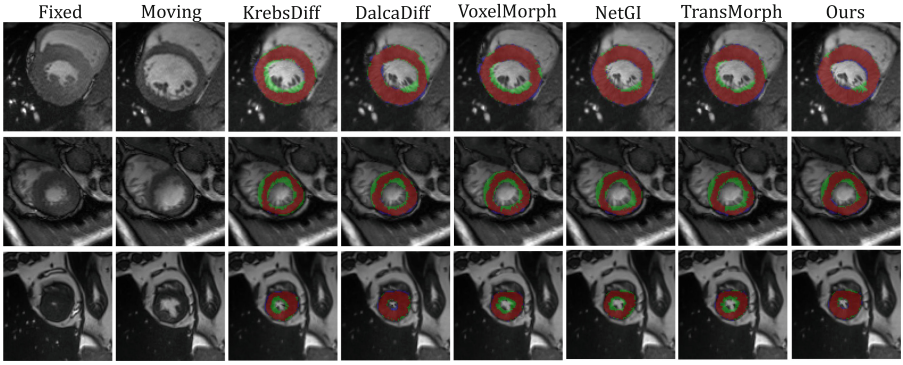

Fig. 3. Demonstration of registration results using different networks. (Color figure online)

with the standard Normal prior. The reason might be that the registration accuracy has a value close to its limit due to iwELBO; in this case, there is no space for the aggregated posterior as prior to improve registration accuracy further. Moreover, we find improvement in registration accuracy using iwELBO comes from the improvement of apical slices registration, details can be referred to in part 3 of the supplement.

Table 2. Comparison of the influence of different parts in our method on ACDC+ dataset. Data format: mean (stand deviation)

ELBO	iwELBO	Single support	Multi-supports	$\mathcal{N}(0, \boldsymbol{I})$ prior	Aggregated prior	Dice	BE	$\lvert J_{f_z} \rvert \leq 0$
✓		✓		✓		0.858 (0.053)	**7.48 (3.16)**	**0.36 (1.31)**
✓		✓			✓	0.861 (0.051)	8.48 (3.36)	0.50 (2.09)
✓			✓	✓		0.863 (0.053)	22.99 (5.88)	0.77 (3.52)
✓			✓		✓	0.865 (0.051)	21.16 (6.01)	0.62 (2.33)
	✓		✓	✓		0.866 (0.053)	20.63 (5.31)	0.69 (1.20)
	✓		✓		✓	**0.867 (0.049)**	24.09 (6.04)	0.84 (4.28)

Further, we replaced the encoder of our network using ViT. Since only the T-encoder existed in ViT, we concatenated the moving and fixed images as input of ViT. In this experiment, different loops in ViT and our Transformer, denoted as ViT-n and Ours-n (n is the number of loops), are employed to compare the performance of self-attention and cross-attention. Table 3 lists comparison results of ViT and our Transformer. It can be seen that cross-attention outperforms self-attention, and more loops are not beneficial in predicting the posterior parameters.

Table 3. Comparison of registration results on ACDC+ dataset using self-attention and cross-attention mechanisms, respectively. Data format: mean (standard deviations)

Method	BE	$\lvert J_{f_z}\rvert \leq 0$	Dice	HD	APD
ViT-3	26.28 (6.61)	0.68 (1.76)	0.865 (0.054)	5.23 (1.73)	**1.86 (0.63)**
ViT-6	19.58 (5.60)	0.70 (1.85)	0.862 (0.053)	5.30 (1.70)	1.91 (0.62)
Ours-3	24.09 (6.04)	0.84 (4.28)	**0.867 (0.049)**	**5.17 (1.57)**	1.87 (0.58)
Ours-6	**18.81 (5.26)**	**0.53 (1.32)**	0.862 (0.053)	5.33 (1.66)	1.92 (0.64)

4 Conclusion

In this paper, we proposed a novel variational registration model using Transformer to pay attention to cross-attention between images. The importance-weighted ELBO and the aggregated posterior as prior close the gap between the real posterior and the variational posterior. Our transformation using multi-supports CSRBFs generates flexible DVFs. Evaluation results on public cardiac datasets show that our method outperforms the state-of-art networks.

Acknowledgements. This paper is supported by the Shenzhen Fundamental Research Program (JCYJ20220531102407018).

References

1. Abdar, M., et al.: A review of uncertainty quantification in deep learning: techniques, applications and challenges. Inf. Fusion **76**, 243–297 (2021)
2. Akkari, N., Casenave, F., Daniel, T., Ryckelynck, D.: Data-targeted prior distribution for variational autoencoder. Fluids (2021)
3. Andreopoulos, A., Tsotsos, J.K.: Efficient and generalizable statistical models of shape and appearance for analysis of cardiac MRI. Med. Image Anal. **12**(3), 335–357 (2008)
4. Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V.: An unsupervised learning model for deformable medical image registration. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9252–9260 (2018)
5. 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)
6. 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)
7. Burda, Y., Grosse, R., Salakhutdinov, R.: Importance weighted autoencoders. arXiv preprint arXiv:1509.00519 (2015)
8. Campello, V.M., et al.: Multi-centre, multi-vendor and multi-disease cardiac segmentation: the M&Ms challenge. IEEE Trans. Med. Imaging 9458279 (2021)
9. Chen, J., Frey, E.C., He, Y., Segars, W.P., Li, Y., Du, Y.: TransMorph: transformer for unsupervised medical image registration. arXiv preprint arXiv:2111.10480 (2021)

10. Connor, M., Canal, G.H., Rozell, C.J.: Variational autoencoder with learned latent structure. ArXiv abs/2006.10597 (2021)
11. Dalca, A.V., Balakrishnan, G., Guttag, J., Sabuncu, M.R.: Unsupervised learning for fast probabilistic diffeomorphic registration. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11070, pp. 729–738. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00928-1_82
12. Fan, J., Cao, X., Xue, Z., Yap, P.-T., Shen, D.: Adversarial similarity network for evaluating image alignment in deep learning based registration. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11070, pp. 739–746. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00928-1_83
13. Gan, Z., Sun, W., Liao, K., Yang, X.: Probabilistic modeling for image registration using radial basis functions: Application to cardiac motion estimation. IEEE Trans. Neural Netw. Learn. Syst. (2022)
14. Gong, X., Khaidem, L., Zhu, W., Zhang, B., Doermann, D.: Uncertainty learning towards unsupervised deformable medical image registration. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 2484–2493 (2022)
15. Huang, C.W., Sankaran, K., Dhekane, E., Lacoste, A., Courville, A.: Hierarchical importance weighted autoencoders. In: International Conference on Machine Learning, pp. 2869–2878. PMLR (2019)
16. Kim, B., Kim, D.H., Park, S.H., Kim, J., Lee, J.G., Ye, J.C.: CycleMorph: cycle consistent unsupervised deformable image registration. Med. Image Anal. **71**, 102036 (2021)
17. Krebs, J., Mansi, T., Mailhé, B., Ayache, N., Delingette, H.: Unsupervised probabilistic deformation modeling for robust diffeomorphic registration. In: Stoyanov, D., et al. (eds.) DLMIA/ML-CDS -2018. LNCS, vol. 11045, pp. 101–109. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00889-5_12
18. Liu, R., Li, Z., Zhang, Y., Fan, X., Luo, Z.: Bi-level probabilistic feature learning for deformable image registration. In: Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence, pp. 723–730 (2021)
19. Radau, P., Lu, Y., Connelly, K., Paul, G., Dick, A., Wright, G.: Evaluation framework for algorithms segmenting short axis cardiac MRI. The MIDAS J.-Cardiac MR Left Ventricle Segment. Challenge **49** (2009)
20. Sandkühler, R., Andermatt, S., Bauman, G., Nyilas, S., Jud, C., Cattin, P.C.: Recurrent registration neural networks for deformable image registration. Adv. Neural Inf. Process. Syst. **32** (2019)
21. Sedghi, A., Kapur, T., Luo, J., Mousavi, P., Wells, W.M.: Probabilistic image registration via deep multi-class classification: characterizing uncertainty. In: Greenspan, H., et al. (eds.) CLIP/UNSURE -2019. LNCS, vol. 11840, pp. 12–22. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32689-0_2
22. Sheikhjafari, A., Noga, M., Punithakumar, K., Ray, N.: Unsupervised deformable image registration with fully connected generative neural network (2018)
23. Takahashi, H., Iwata, T., Yamanaka, Y., Yamada, M., Yagi, S.: Variational autoencoder with implicit optimal priors. In: AAAI (2019)
24. Tomczak, J.M., Welling, M.: VAE with a VampPrior. In: AISTATS (2018)
25. Xu, H., Chen, W., Lai, J., Li, Z., Zhao, Y., Pei, D.: On the necessity and effectiveness of learning the prior of variational auto-encoder. ArXiv abs/1905.13452 (2019)

26. Yuan, L., et al.: Tokens-to-token ViT: training vision transformers from scratch on ImageNet. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 558–567 (2021)
27. Zhao, S., Dong, Y., Chang, E.I., Xu, Y., et al.: Recursive cascaded networks for unsupervised medical image registration. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10600–10610 (2019)