TLRN: Temporal Latent Residual Networks For Large Deformation Image Registration

Nian Wu[1], Jiarui Xing[1], and Miaomiao Zhang[1,2]

[1] Department of Electrical and Computer Engineering, University of Virginia, USA
[2] Department of Computer Science, University of Virginia, USA

Abstract. This paper presents a novel approach, termed *Temporal Latent Residual Network (TLRN)*, to predict a sequence of deformation fields in time-series image registration. The challenge of registering time-series images often lies in the occurrence of large motions, especially when images differ significantly from a reference (e.g., the start of a cardiac cycle compared to the peak stretching phase). To achieve accurate and robust registration results, we leverage the nature of motion continuity and exploit the temporal smoothness in consecutive image frames. Our proposed TLRN highlights a temporal residual network with residual blocks carefully designed in latent deformation spaces, which are parameterized by time-sequential initial velocity fields. We treat a sequence of residual blocks over time as a dynamic training system, where each block is designed to learn the residual function between desired deformation features and current input accumulated from previous time frames. We validate the effectivenss of TLRN on both synthetic data and real-world cine cardiac magnetic resonance (CMR) image videos. Our experimental results shows that TLRN is able to achieve substantially improved registration accuracy compared to the state-of-the-art. Our code is publicly available at https://github.com/nellie689/TLRN.

1 Introduction

Temporal/time-series image registration plays an important role in various domains, including medical imaging [29, 20], video analysis [10], and remote sensing [34]. This process involves aligning images captured at a sequence of time points; hence enabling the analysis and tracking of dynamic changing process [23, 25, 33, 9]. In the domain of medical applications, it is desirable that the estimated deformations being diffeomorphisms (i.e., one-to-one smooth and invertible smooth mappings between images [4]). Such properties are essential to preserve the local and global topological structure of studied subjects, ensuring biologically meaningful results [13, 24].

A major challenge of temporal image registration is the large deformation propagated over time, for example, the start of a cardiac cycle compared to the peak stretching phase in CMR videos [33, 23, 25]. Traditional pairwise registration methods, which align images to a single reference image, often suffer from accumulating errors over time and fail to capture the complex and continuous nature of large deformations [28, 5, 12, 3]. To alleviate this issue, existing approaches

have been proposed to concatenate transformations estimated from consecutive frames with relatively small deformations [28, 6]. However, these methods often overlook the temporal dynamics of deformation fields or motion in the image data, resulting in compromised registration accuracy. Recognizing this limitation, several research focused on leveraging temporal information to enforce smoothness and continuity over time [20]. Other related studies introduced spatio-temporal parametric models based on B-splines to enforce the temporal consistency of estimated deformation fields [19, 22, 7]. With the advancement of deep learning, predictive registration methods have gained popularity due to their faster inference compared to traditional optimization-based approaches [3, 5, 31, 12]. Recent works have harnessed the power of deep networks to generate deformation fields from a sequence of image videos while considering general temporal smoothness [17, 18, 26, 27]. Despite achieving impressive results, these methods do not explicitly model the relationships between transformations propagated over time, resulting in limited capability of capturing the underlying representations of large and complex deformations in dynamic processes.

In this paper, we introduce a novel approach, termed TLRN, to effectively capture fine details and complex patterns of large deformations by incrementally refining learned latent deformation features across time. To achieve this, we carefully design a multi-level structure of residual blocks in the temporal latent velocity space to model the hierarchical relationships between deformations. Each level builds upon and refines the feature pace learned by the previous levels. The main contributions of our proposed method are twofold:

(i) Develop a temporal residual network in the latent space of velocity fields to explicitly model the hierarchical relationship between time-sequential deformations.
(ii) Effectively incorporate the temporal smoothness and continuity of deformations/motions inherent in image videos, leading to improved registration accuracy.

We validate the proposed model, TRNL, using both synthetic data and real CMR image videos. Experimental results demonstrate that TRNL achieves superior registration accuracy, producing more robust and better regularized transformation fields compared to state-of-the-art deep learning-based registration networks. [3, 12, 5, 15].

2 Background: Diffeomorphic Image Registration

In this section, we briefly review diffeomorphic image registration in the context of stationary velocity fields (SVF) [30]. Note that our model is general to other registration models, such as large deformation diffeomorphic metric mapping [4].

Given a source image, S^A, and a target image, S^B, defined on a d-dimensional torus domain $\Omega = \mathbb{R}^d/\mathbb{Z}^d$ $(S^A(x), S^B(x) : \Omega \to \mathbb{R})$. The problem of diffeomorphic image registration is typically formulated as an energy minimization over a

time-dependent deformation fields, $\{\phi_t : t \in [0,1]\}$, i.e.,

$$E(\phi_1) = \frac{1}{2\sigma^2}\,\mathrm{Dist}(S^A \circ \phi_1^{-1}, S^B) + \mathrm{Reg}(\phi_1). \tag{1}$$

Here, $\circ$ denotes an interpolation operator that deforms a source image to match the target. The $\mathrm{Dist}(\cdot,\cdot)$ is a distance function that measures the dissimilarity between images weighted by a positive parameter σ, and $\mathrm{Reg}(\cdot)$ is a regularization term to enforce the smoothness of transformation fields. In this paper, we use a commonly used sum-of-squared intensity differences (L_2-norm) [4, 32] as the distance function.

In the setting of SVF [30], the diffeomorphic transformation fields, ϕ_t, is parameterized by a constant velocity field v over time, i.e.,

$$\frac{d\phi_t}{dt} = v(\phi_t), \text{ s.t. } \phi_0 = x. \tag{2}$$

The solution of Eq. (2) is identified with a group exponential map, which is numerically computed through a scaling and squaring method [30]. For a simplified notation, we will drop the time index in following sections, i.e., $\phi_1 \triangleq \phi$.

3 Our Method: TLRN

In this section, we introduce a novel temporal latent residual networks, TLRN, that can effectively predict a sequence of deformation fields in time-series image registration. Our proposed TLRN consists of two key components: (i) an unsupervised registration network that learns the velocity fields for the input time-series images, and (ii) a temporal residual learning submodule designed in the learned latent velocity spaces to effectively adjust integrated deformation features from current and previous time steps. An overview of our proposed network is shown in Fig. 1.

3.1 Network Architecture

Learning latent velocity fields via unsupervised registration. Given a training dataset of N image sequences, where each sequence includes $T+1$ time frames, denoted as $\{I_i^\tau\}, i \in \{1, \cdots, N\}, \tau \in \{0, \cdots, T\}$. That is to say, for the ith training data, we have a sequence of image frames $\{I_i^0, \cdots, I_i^T\}$. By setting the first frame $\{I_i^0\}$ as a reference (source) image, there exists a number of T pairwise images, $\{(I_i^0, I_i^1), (I_i^0, I_i^2) \cdots, (I_i^0, I_i^T)\}$, to be aligned by their associated velocity fields $\{v_i^1, v_i^2 \cdots, v_i^T\}$. Similar to [3, 12], we employ U-Net architecture as the backbone of our registration encoder, $\mathcal{E}_{\theta_v}$, and decoder, $\mathcal{D}_{\theta_v}$, parameterized by θ_v. The encoder $\mathcal{E}_{\theta_v}$ projects the input image sequences into a latent velocity space $\mathcal{Z}$, which embeds a sequence of latent representations of the velocity fields, denoted as $\{z^1, z^2, \cdots, z^T\} \in \mathcal{Z}$. The decoder $\mathcal{D}_{\theta_v}$ is then used to project the latent features back to the input image space.

Temporal latent residual learning. To leverage the temporal smoothness and continuity of time-series images, we develop a temporal residual learning scheme in the latent velocity space. Specifically, we introduce a latent recurrent unit with integrated residual blocks to perform two key tasks: (i) fusing the latent velocity features from current and previous time points, and (ii) re-adjusting these features through learned residual functions with reference to an optimal solution to further reduce error. Given the adjusted latent feature $\hat{z}^{\tau-1}$ at previous time point and z^{τ} at current time point, the output feature $\hat{z}^{\tau}$ of our proposed residual block can be represented as

$$\hat{z}^{\tau} := \mathcal{F}_{\theta_r}(\hat{z}^{\tau-1} \oplus z^{\tau}) + W_{\theta_s}(\hat{z}^{\tau-1} \oplus z^{\tau}), \tag{3}$$

where $\mathcal{F}_{\theta_r}$ represents a residual function parameterized by θ_r and $\oplus$ denotes vector concatenation. Following the principles in [11], we perform a learnable linear projection, W_{θ_s}, by the shortcut connections to match the dimensions of $\mathcal{F}_{\theta_r}$ and $(\hat{z}^{\tau-1} \oplus z^{\tau})$. Note that the choice of residual function, $\mathcal{F}_{\theta_r}$, is flexible. In this paper, we employ a composition of two convolutional layers and Leaky Rectified Linear Unit (LeakyReLU) [21].

Network loss. By defining $\Theta = (\theta_v, \theta_r, \theta_s)$ for all network parameters, we finally formulate the loss function of TLRN as

$$l(\Theta) = \sum_{i=1}^{N}\sum_{\tau=1}^{T} \lambda \left\| (I_i^0 \circ \phi_i^{\tau}\,(v_i^{\tau}(\hat{z}_i^{\tau};\Theta)) - I_i^{\tau} \right\|_2^2 + \left\| \nabla v_i^{\tau}(\hat{z}_i^{\tau};\Theta) \right\|^2 + \mathrm{reg}(\Theta). \tag{4}$$

Here, ϕ_i^{τ} is the transformation fields between (I_i^0, I_i^{τ}), λ is a positive weighting parameter, and $\mathrm{reg}(\cdot)$ represents network regularity. We jointly optimize all parameters till convergence.

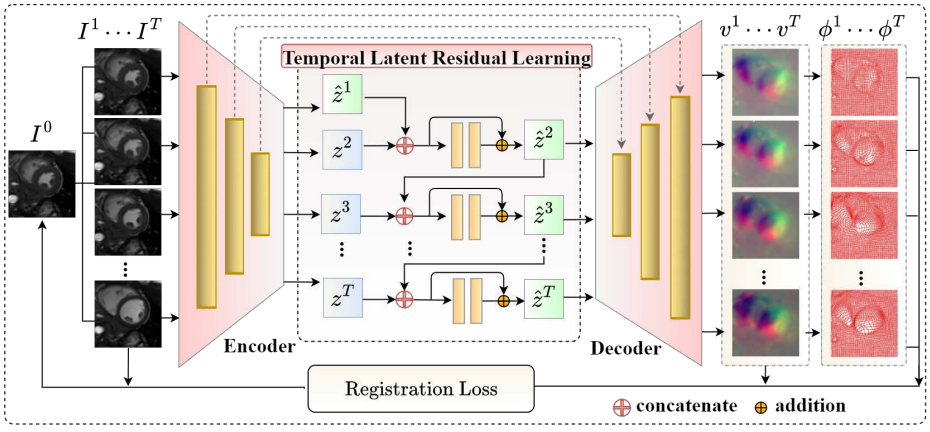

Fig. 1. An overview of our proposed network TLRN.

4 Experimental Evaluation

We validate our model, TLRN, on both synthetic data and real CMR image videos. All our experiments are trained on servers with AMD EPYC 7502 CPU of 126GB memory and Nvidia GTX 4090Ti GPUs. We use the Adam optimizer [16] with 20000 epochs, learning rate of $1e^{-4}$, and a batch size of 32.

4.1 Datasets

2D Lemniscate Sequence. We simulate a number of 1400 "lemniscate" series ($64^2 \times T$), where $T + 1 = 12$ (examples are shown in Fig. 2). For each reference image I_i^0, we first generate a 2D contour with its coordinates computed by $\begin{cases} x = a \cdot \cos(\alpha)/(\sin^2(\alpha)+1) \\ y = a \cdot \sin(\alpha)\cos(\alpha)/(\sin^2(\alpha)+1) \end{cases}$, where a is a scaling factor, and α is uniformly sampled from $[0, 2\pi]$. We then determine the contour thickness by varying the coordinates with parameters σ_x and σ_y, i.e., $[x \pm \sigma_x, y \pm \sigma_y]$. The subsequent time frames $\{I_i^1, \cdots, I_i^T\}$ are deformed versions of the reference image, simulated by applying affine transformations such as scaling, rotation, and translation. We split the dataset into 1000/200/200 sequences for training, validation, and testing.
Cardiac MRI Series. We include 600 cine CMR videos collected from sixty subjects. Each video sequence covers half of the cardiac motion cycle (with $T + 1 = 7$), spanning from the peak stretching phase to the peak contracting phase. We crop the original image to the size of 64×64, focusing on the structure of left ventricles. The LV myocardium of all video sequences are manually annotated by clinical experts. We randomly choose 400 sequences from 40 subjects for training, 100 sequences from 10 subjects for validation, and the rest for testing.

4.2 Experimental Design

We compare the performance of TLRN with four state-of-the-art deep learning-based diffeomorphic registration algorithms in the context of both SVF [1] and LDDMM [4]. These approaches include Voxelmorph (VM) [2], TransMorph (TM) [5], SVF-R2Net [15], and Lagomorph (LM) [12]. All methods are trained on the same dataset with their best performance reported.

Note that we consider the comparison with VM [2] as an ablation study for the proposed temporal latent residual learning block. Both methods use U-Net as the network backbone with the same number of convolutional layers. We maintain consistent network parameters, including regularity weights on velocity fields and the timesteps of integration, to ensure a fair comparison.

Evaluation Metric. We first evaluate the registration accuracy of TLRN on synthetic dataset by computing the mean squared error (MSE) between deformed sequence images and target sequence images over the time and comparing the results with all baselines. We report the percentage of negative Jacobian determinants of predicted transformation fields from our method vs. all baselines.

We assess the registration accuracy on CMR videos by performing registration-based segmentation of the left ventricular (LV) myocardium. The propagated

segmentations by deforming manually segmented myocardium on the reference image using predicted transformation fields from all methods are then compared. To evaluate volume overlap between the propagated segmentation A and the manual segmentation B, we compute the dice similarity coefficient (DSC) [8] by $\text{DSC}(A, B) = 2(|A| \cap |B|)/(|A| + |B|)$, where $\cap$ denotes an intersection of two regions. Additionally, we measure the maximum discrepancy between boundaries of the propagated segmentation and the manual segmentation by Hausdorff distance (HD) [14]. Given two sets of boundary points $X \in A$ and $Y \in B$ for LV myocardium, we compute the distance by $\text{HD}(X, Y) = max\{h(X, Y), h(Y, X)\}$, where $h(X, Y) = max_{x_i \in X}\, min_{y_i \in Y}\, \|x_i - y_i\|$, and vice versa.

4.3 Results

Fig. 2 visualizes the predicted time-series of both deformed images and deformation fields (from I^0 to the remaining frames $I^1 \sim I^T$) from all methods. It shows that our method, TRNL, consistently achieves improved quality of deformed images with well regularized transformation fields along the time points. Moreover, our model obtains the lowest percentage of negative Jacobian determinants (TLRN: $0.175\% \pm 0.373\%$; LM: $0.474\% \pm 0.969\%$, SVF-R2Net: $1.607\% \pm 1.917\%$, VM: $0.398\% \pm 0.591\%$, and TM: $0.176\% \pm 0.471\%$), which indicates better quality and more realistic registration results.

Fig. 3 compares MSE between the deformed images and the target images across all methods over multiple time steps. It shows that TLRN achieves substantially lower MSE compared to all baselines. The improved registration accuracy and smoothness of deformation fields (especially on later time frames when large deformations occur) indicate that our model effectively leverages the spatiotemporal continuity within the cardiac motion sequence.

The left panel of Fig. 4 presents visual examples of the reference/source frame, target sequence, and deformed sequence over all methods. The right panel of Fig. 4 displays a comparison between the manually labeled segmentation on LV myocardium and propagated segmentation labels deformed by deformations fields predicted from all methods.

Fig. 5 displays a quantitative comparison between the manually delineated vs. propagated segmentation labels deformed by deformations fields predicted from all methods. The left panel reports the statistics of dice scores (the higher the better), and the right panel reports the statistics of Hausdorff distance (the lower the better) on the LV myocardium. All metrics are computed throughout the cardiac motion cycle, from the peak contraction phase to the peak stretching phase. It shows that TLRN is able to produce superior quality of final dice and HD scores in comparison with other baselines.

5 Conclusions & Discussion

This paper presents a novel temporal latent residual network, TLRN, to predict a sequence of deformation fields in time-series image registration. To the best of

our knowledge, our model is the first to develop a spatio-temporal network with residual blocks in the latent deformation space, parameterized by velocity fields. The learned residual functions over time are well utilized to effectively adjust the continuous deformation features from current and previous time frames. Experimental results on both synthetic sequences and real-world cine CMR videos show that the proposed TLRN is able to achieve an improved registration accuracy with better regularized deformation fields.

Fig. 2. Left to right: examples of deformed reference/source image across time. Top to bottom: a comparison of deformed images and transformation fields predicted from our model TRLN and baselines.

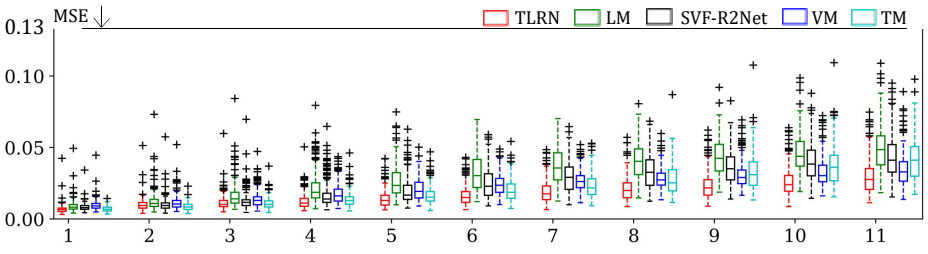

Fig. 3. A comparison of MSE between deformed and target time-series images.

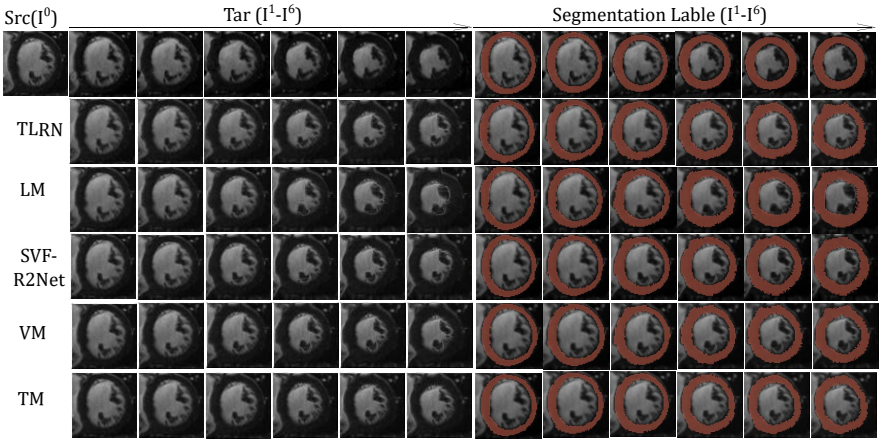

Fig. 4. Left to right: examples of CMR image videos and overlaid LV myocardium segmentation maps. Top to bottom: a comparison of manually delineated segmentation lables vs. propagated segmentation from all methods.

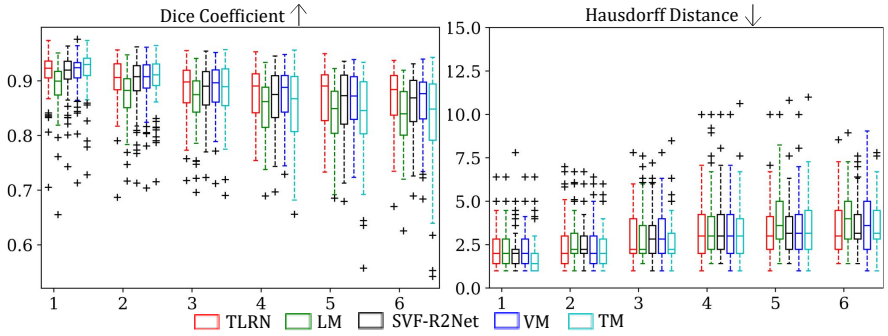

Fig. 5. Left to right: a comparison of dice vs. Hausdorff distance score on predicted LV myocardium segmentation labels from all methods over the time frames $\tau(1 \sim 6)$.

Acknowledgments. This work was supported by NSF CAREER Grant 2239977 and NIH 1R21EB032597.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Arsigny, V., Commowick, O., Pennec, X., Ayache, N.: A log-euclidean framework for statistics on diffeomorphisms. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 924–931. Springer (2006)
2. 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)
3. Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V.: Voxelmorph: a learning framework for deformable medical image registration. IEEE transactions on medical imaging **38**(8), 1788–1800 (2019)
4. Beg, M.F., Miller, M.I., Trouvé, A., Younes, L.: Computing large deformation metric mappings via geodesic flows of diffeomorphisms. International journal of computer vision **61**(2), 139–157 (2005)
5. Chen, J., Frey, E.C., He, Y., Segars, W.P., Li, Y., Du, Y.: Transmorph: Transformer for unsupervised medical image registration. Medical image analysis **82**, 102615 (2022)
6. Csapo, I., Holland, C.M., Guttmann, C.R.: Image registration framework for large-scale longitudinal mri data sets: strategy and validation. Magnetic Resonance Imaging **25**(6), 889–893 (2007)
7. De Craene, M., Piella, G., Camara, O., Duchateau, N., Silva, E., Doltra, A., D'hooge, J., Brugada, J., Sitges, M., Frangi, A.F.: Temporal diffeomorphic free-form deformation: Application to motion and strain estimation from 3d echocardiography. Medical image analysis **16**(2), 427–450 (2012)
8. Dice, L.R.: Measures of the amount of ecologic association between species. Ecology **26**(3), 297–302 (1945)
9. Geng, X., Christensen, G.E., Gu, H., Ross, T.J., Yang, Y.: Implicit reference-based group-wise image registration and its application to structural and functional mri. Neuroimage **47**(4), 1341–1351 (2009)
10. Ghanem, B., Zhang, T., Ahuja, N.: Robust video registration applied to field-sports video analysis. In: IEEE International conference on acoustics, speech, and signal processing (ICASSP). vol. 2 (2012)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
12. Hinkle, J., Womble, D., Yoon, H.J.: Diffeomorphic autoencoders for lddmm atlas building (2018)
13. Hong, Y., Golland, P., Zhang, M.: Fast geodesic regression for population-based image analysis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 317–325. Springer (2017)
14. Huttenlocher, D.P., Klanderman, G.A., Rucklidge, W.J.: Comparing images using the hausdorff distance. IEEE Transactions on pattern analysis and machine intelligence **15**(9), 850–863 (1993)
15. Joshi, A., Hong, Y.: Diffeomorphic image registration using lipschitz continuous residual networks. In: International Conference on Medical Imaging with Deep Learning. pp. 605–617. PMLR (2022)
16. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
17. Krebs, J., Delingette, H., Ayache, N., Mansi, T.: Learning a generative motion model from image sequences based on a latent motion matrix. IEEE Transactions on Medical Imaging **40**(5), 1405–1416 (2021)
18. Krebs, J., Mansi, T., Ayache, N., Delingette, H.: Probabilistic motion modeling from medical image sequences: application to cardiac cine-mri. In: Statistical Atlases and Computational Models of the Heart. Multi-Sequence CMR Segmentation, CRT-EPiggy and LV Full Quantification Challenges: 10th International Workshop, STACOM 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13, 2019, Revised Selected Papers 10. pp. 176–185. Springer (2020)

19. Ledesma-Carbayo, M.J., Kybic, J., Desco, M., Santos, A., Suhling, M., Hunziker, P., Unser, M.: Spatio-temporal nonrigid registration for ultrasound cardiac motion estimation. IEEE transactions on medical imaging **24**(9), 1113–1126 (2005)
20. Liao, R., Turk, E.A., Zhang, M., Luo, J., Grant, P.E., Adalsteinsson, E., Golland, P.: Temporal registration in in-utero volumetric mri time series. In: Medical Image Computing and Computer-Assisted Intervention-MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part III 19. pp. 54–62. Springer (2016)
21. Maas, A.L., Hannun, A.Y., Ng, A.Y., et al.: Rectifier nonlinearities improve neural network acoustic models. In: Proc. icml. vol. 30, p. 3. Atlanta, GA (2013)
22. Metz, C.T., Klein, S., Schaap, M., van Walsum, T., Niessen, W.J.: Nonrigid registration of dynamic medical imaging data using nd+ t b-splines and a groupwise optimization approach. Medical image analysis **15**(2), 238–249 (2011)
23. Morais, P., Heyde, B., Barbosa, D., Queirós, S., Claus, P., D'hooge, J.: Cardiac motion and deformation estimation from tagged mri sequences using a temporal coherent image registration framework. In: Functional Imaging and Modeling of the Heart: 7th International Conference, FIMH 2013, London, UK, June 20-22, 2013. Proceedings 7. pp. 316–324. Springer (2013)
24. Niethammer, M., Huang, Y., Vialard, F.X.: Geodesic regression for image time-series. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2011: 14th International Conference, Toronto, Canada, September 18-22, 2011, Proceedings, Part II 14. pp. 655–662. Springer (2011)
25. Perperidis, D., Mohiaddin, R.H., Rueckert, D.: Spatio-temporal free-form registration of cardiac mr image sequences. Medical image analysis **9**(5), 441–456 (2005)
26. Qiao, M., Wang, S., Qiu, H., De Marvao, A., O'Regan, D.P., Rueckert, D., Bai, W.: Cheart: A conditional spatio-temporal generative model for cardiac anatomy. IEEE transactions on medical imaging (2023)
27. Qin, C., Wang, S., Chen, C., Bai, W., Rueckert, D.: Generative myocardial motion tracking via latent space exploration with biomechanics-informed prior. Medical Image Analysis **83**, 102682 (2023)
28. Reinhardt, J.M., Ding, K., Cao, K., Christensen, G.E., Hoffman, E.A., Bodas, S.V.: Registration-based estimates of local lung tissue expansion compared to xenon ct measures of specific ventilation. Medical image analysis **12**(6), 752–763 (2008)
29. Singh, M., Thompson, R., Basu, A., Rieger, J., Mandal, M.: Image based temporal registration of mri data for medical visualization. In: 2006 International Conference on Image Processing. pp. 1169–1172. IEEE (2006)
30. Vercauteren, T., Pennec, X., Perchant, A., Ayache, N.: Symmetric log-domain diffeomorphic registration: A demons-based approach. In: International conference on medical image computing and computer-assisted intervention. pp. 754–761. Springer (2008)
31. Wang, J., Zhang, M.: Deepflash: An efficient network for learning-based medical image registration. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4444–4452 (2020)
32. Wu, N., Zhang, M.: Neurepdiff: Neural operators to predict geodesics in deformation spaces. In: International Conference on Information Processing in Medical Imaging. pp. 588–600. Springer (2023)
33. Xing, J., Wu, N., Bilchick, K., Epstein, F., Zhang, M.: Multimodal learning to improve cardiac late mechanical activation detection from cine mr images. arXiv preprint arXiv:2402.18507 (2024)
34. Yang, Z., Dan, T., Yang, Y.: Multi-temporal remote sensing image registration using deep convolutional features. Ieee Access **6**, 38544–38555 (2018)