Recursive Deformable Image Registration Network with Mutual Attention

Jian-Qing Zheng[1,2][0000−0002−1823−1419], Ziyang Wang[3][0000−0003−1605−0873], Baoru Huang[4][0000−0002−4421−652X], Tonia Vincent[1][0000−0002−3412−5712], Ngee Han Lim[1][0000−0002−8520−7857], and Bartłomiej W. Papież[2,5][0000−0002−8432−2511]

[1] The Kennedy Institute of Rheumatology, University of Oxford, U.K.
[2] Big Data Institute, University of Oxford, U.K.
[3] Department of Computer Science, University of Oxford, U.K.
[4] Department of Surgery and Cancer, Imperial College London
[5] Nuffield Department of Population Health, University of Oxford, UK
{jianqing.zheng@kennedy,bartlomiej.papiez@bdi}.ox.ac.uk

Abstract. Deformable image registration, estimating the spatial transformation between different images, is an important task in medical imaging. Many previous studies have used learning-based methods for multi-stage registration to perform 3D image registration to improve performance. The performance of the multi-stage approach, however, is limited by the size of the receptive field where complex motion does not occur at a single spatial scale. We propose a new registration network combining recursive network architecture and mutual attention mechanism to overcome these limitations. Compared with the state-of-the-art deep learning methods, our network based on the recursive structure achieves the highest accuracy in lung Computed Tomography (CT) data set (Dice score of 92% and average surface distance of 3.8mm for lungs) and one of the most accurate results in abdominal CT data set with 9 organs of various sizes (Dice score of 55% and average surface distance of 7.8mm). We also showed that adding 3 recursive networks is sufficient to achieve the state-of-the-art results without a significant increase in the inference time.

Keywords: Deformable Image Registration · Recursive Network · Mutual Attention

1 Introduction

Deformable image registration (DIR) is an essential computer vision task which has been widely studied [18]. In medical imaging, DIR enables the estimation of the non-linear correspondence between different acquisitions over time to monitor progress of treatment, or between different types of scanners (e.g. multi-modal image fusion) to provide complementary disease information. The classical registration algorithms have been developed as continuous optimization [2, 15, 20], or discrete optimization problems [8]. Their computational performance, however, is limited due to highly dimensional, non-convex problem, and low capability to capture complex, global and local deformations [16]. Recently, researchers have shifted interest to deep-learning-based

unsupervised learning methods in deformable image registration, because data-driven methods benefit significantly from a large amount of given paired/unpaired images compared with classical methods [12, 22, 1]. A fast learning-based approach, VoxelMorph, is presented in [4], where convolutional neural networks (CNN) and spatial transformer layers [11] are used to register two images by regressing directly dense displacement field. Other deep learning approaches investigated different representations of the transformation e.g. diffeomorphism [14], which preserve the topology of the transformation. The direct regression of the spatial transformation via neural networks however, only gives one prediction on registration without any progressive refinement.

Multi-stage architecture is one of the solutions that is beneficial to CNN [22, 10, 25]. A weakly supervised multi-model registration method [10], utilizing an end-to-end convolution based network, aims to predict displacement fields to align multiple labelled corresponding structures for individual image pairs. Alternatively, an end-to-end multi-stage networks [24] are proposed for a deep recursive cascade architecture that allows unlimited number of cascades that can be built on the backbone networks. All of these multi-stage cascaded network structures, however, are still potentially suffering from the limited capture range of the receptive field.

The attention mechanism [21] addresses the limited receptive field of CNNs and has been widely utilized in transformer networks. Optimal correspondence matching was studied in [13] for a stereo matching task, where self-attention-based transformer is proposed to relax the limitation of a fixed disparity range. Local feature matching can also benefit from self and cross attention, because transformer networks are proved to obtain feature descriptors that are conditioned on both images [19]. The attention-based mechanism was applied to registration [23, 17, 5] previously, however is computationally expensive, and thus has not been explored in recursive deformable image registration.

In this paper, we propose a Recursive Mutual Attention Network (RMAn), combining the Mutual Attention (MA) based module with a recursive architecture to increase the size of the receptive field. The recursive architecture provides the progressive refinement to 3D image deformable registration so that MA module can expand the global receptive field on a pair of low-resolution feature maps without extra cost of computation. Our contributions in this paper are as follows.

1. A Mutual Attention based Recursive Network (RMAn) is proposed for deformable image registration, combining the mutual attention [21] into recursive networks [24];
2. The proposed network achieves the best performance against the state-of-the-art deep learning network structures respectively in lung Computed Tomography (CT) data set (Dice similarity coefficient of 92% and average surface distance for of 3.8mm lung) and comparable performance in abdomen (9 organs) CT data set.

2 Methods

2.1 Image Registration

Image registration can be defined as estimation of the spatial transformation $\phi : \mathbb{R}^n \rightarrow \mathbb{R}^n$, represented by the corresponding parameters or a series of displacements denoted

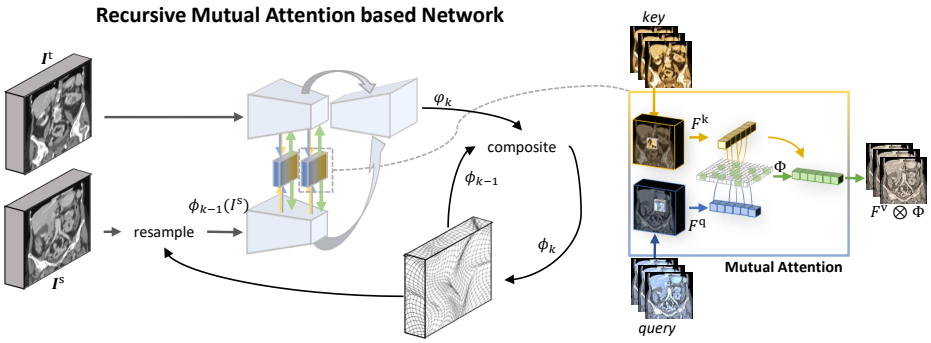

Fig. 1. Proposed framework of Recursive Mutual Attention based Network, including a Siamese Encoder-Decoder structure with Mutual Attention interconnected, and the network structure detailed in Fig. 2, where $k \in [1, K] \cap \mathbb{Z}$ denotes the recursive index and $K \in \mathbb{Z}_+$ denotes the total recurrent number.

by $\phi[\boldsymbol{x}] \in \mathbb{R}^d$ at the spatial position $\boldsymbol{x} \in \mathbb{Z}^d$ of a target image $\boldsymbol{I}^{\mathrm{t}} \in \mathbb{R}^n$ from a source image $\boldsymbol{I}^{\mathrm{s}} \in \mathbb{R}^n$, where n is the size of a 3D image defined as $n = H \times W \times T$, and d, T, H, W denoting the image dimension, thickness, height, and width, respectively. Originally, image registration was solved as an optimization problem by minimization of a dissimilarity metric $\mathcal{D}$ and a regularization term $\mathcal{S}$:

$$\hat{\phi} = \underset{\phi}{\operatorname{argmin}}\big(\mathcal{D}(\phi(\boldsymbol{I}^{\mathrm{s}}),\, \boldsymbol{I}^{\mathrm{t}}) + \lambda \mathcal{S}(\phi, \boldsymbol{I}^{\mathrm{t}})\big) \tag{1}$$

where $\hat{\phi}$ denotes the estimated spatial transformation, λ denotes the weight of the regularization. More recently, the registration is performed via CNN $\mathcal{R}$ directly regressing the spatial transformation e.g. using the Dense Displacement Field (DDF) [3, 14]:

$$\phi = \mathcal{R}(\boldsymbol{I}^{\mathrm{s}}, \boldsymbol{I}^{\mathrm{t}}; w) \tag{2}$$

with the training process based on minimizing the loss function (e.g. given in Eq. (1)) with the trainable weights w (w is omitted in the following part of the paper to simplify the formula). However the direct regression of spatial transformations via convolution neural networks could suffer due to limited capture range of the receptive field of convolutional layers when dealing with large motion.

2.2 Recursive Registration Networks

Inspired by [24], we proposed a recursive network structure for coarse-to-fine registration of a pair of images as shown in Fig. 1. In coarse-to-fine approach, the residual transformation φ_k between the target image $\boldsymbol{I}^{\mathrm{t}}$ and the warped source feature map based on previous level $k-1$ registration $\phi_{k-1}(\boldsymbol{I}^{\mathrm{s}})$ is estimated via $\mathcal{R}$ and accumulated via composition:

$$\begin{cases} \phi_k = \phi_{k-1} \circ \varphi_k \\ \varphi_k = \mathcal{R}(\phi_{k-1}(\boldsymbol{I}^{\mathrm{s}}), \boldsymbol{I}^{\mathrm{t}}) \end{cases} \tag{3}$$

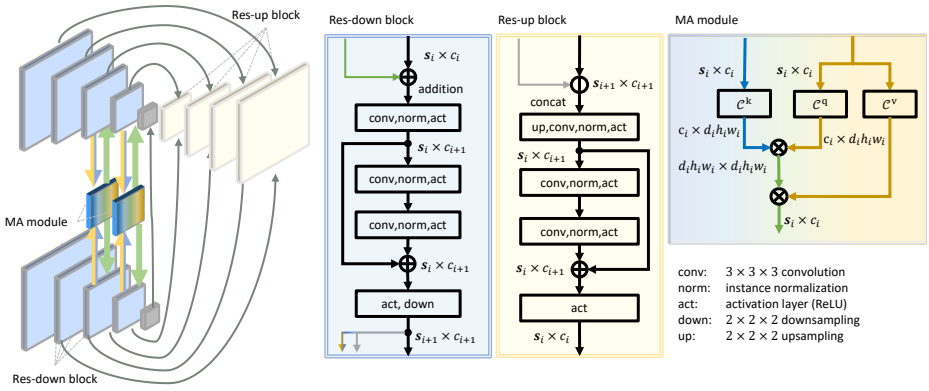

Fig. 2. The subnetwork in Fig. 1 including three main components, a Siamese Encoder consists of four pairs of Residual Downsampling (Res-down) blocks, Residual Upsampling (Res-up) block, and two Mutual Attention (MA) modules.

where $\circ$ denotes the composition of two spatial transformations, and ϕ_0 is initialized as the identity transform. The subnetwork used in Fig. 1 including a weight-sharing two-branch Siamese encoder interconnected with a Mutual Attention module to extract and retrieve the common features, and a decoder to estimate the DDF φ_k, where each component of the network structure is shown in Fig. 2, and where the convolution layers in each Res-down and Res-up blocks are set with the kernel size of 3, and atrous rate (1,1,3) following the theoretical optimization of receptive field size in [27].

2.3 Mutual Attention

Similar to the idea from [13, 19, 7, 26], Mutual Attention (MA) mechanism [21] is used in the RMAn to obtain the global receptive field and use so-called indicator matrices to quantify the relationship between each pair of pixels from two images, and the usage of multiple indicator matrices is called multi-head. The feature maps $\boldsymbol{F}^{\mathrm{k}}, \boldsymbol{F}^{\mathrm{q}} \in \mathbb{R}^{c \times n}$ are extracted from two stream of the two images $\boldsymbol{I}^{\mathrm{s}}, \boldsymbol{I}^{\mathrm{t}}$ via the encoder part as shown in Fig. 1, where c denotes the feature channel number. Each element of $\boldsymbol{F}^{\mathrm{k}}$ (yellow arrow in Fig. 1 and Fig. 2) as a key vector is retrieved in the query vectors via correlation from $\boldsymbol{F}^{\mathrm{q}}$ (blue arrow) in an indicator matrix $\Phi \in \mathbb{R}^{n \times n}$ which can be formulated as:

$$\Phi = \mathrm{softmax}(\mathcal{C}^{\mathrm{q}}(\boldsymbol{F}^{\mathrm{k}})^{\top} \mathcal{C}^{\mathrm{k}}(\boldsymbol{F}^{\mathrm{q}})) \tag{4}$$

Then the vector from $\boldsymbol{F}^{\mathrm{q}}$ is passed through the corresponding linear mapping to the other stream via the Φ:

$$\begin{cases} \boldsymbol{F}^{\mathrm{v}} = \mathcal{C}^{\mathrm{v}}(\boldsymbol{F}^{\mathrm{k}}) \\ \boldsymbol{F}^{\mathrm{k}\to\mathrm{q}} = \boldsymbol{F}^{\mathrm{v}} \otimes \Phi \end{cases} \tag{5}$$

where $\boldsymbol{F}^{\mathrm{k}\to\mathrm{q}}$ denotes the feature maps passed from one stream to the other (green arrow in Fig. 1 and Fig. 2), $\mathcal{C}^{\mathrm{q}}$, $\mathcal{C}^{\mathrm{k}}$ and $\mathcal{C}^{\mathrm{v}}$ denote the linear transformation for query, key

and value feature vectors, respectively. Because the MA module is used bi-directly, the feature forwarded as both the key and query features are denoted as half blue half yellow arrows in Fig. 1 and Fig. 2, and the corresponding green arrow always point to the branch of the query stream.

3 Experiments

3.1 Datasets

We evaluated the proposed RMAn for unsupervised deformable registration problem using two publicly available data sets with the ground truth annotations for 9 organs in abdomen CT data set and lung volumes annotations in lung CT data set.

Unpaired Abdomen CTs are selected from [6]. The ground truth segmentation of spleen, right kidney, left kidney, esophagus, liver, aorta, inferior vena cava, portal, splenic vein, and pancreas are annotated for all CT scans. The inter-subject registration of the abdominal CT scans is challenging due to large inter-subject variations and great variability in organ volume, from 10 milliliters (esophagus) to 1.6 liters (liver). Following the previously presented works, each volume is resized to $2 \times 2 \times 2mm^3$ in the pre-processing step. From totally 30 subjects, 23 and 7 are respectively used for training and testing, forming 506 and 42 different pairs of images.

Unpaired Chest (Lung) CTs are selected from [9]. The CT scans are all acquired at the same time point of the breathing cycle with a slice thickness of 1.00 mm and slice spacing of 0.70 mm. Pixel spacing in the X-Y plane varies from 0.63 to 0.77 mm with an average value of 0.70 mm. The ground truth annotations of lungs for all scans are provided. Following the previously presented works, each volume is resized to $1 \times 1 \times 1mm^3$ in the pre-processing step. We perform inter-subject registration from the total of 20 subjects, 12 and 8 are respectively used for training and testing, forming 132 and 56 different pairs of images.

3.2 Training Details

We normalize the input image into 0-1 range and augment the training data by randomly cropping input images during training. For the experiments on inter-subject registration of abdomen and lung CT, the models are first pre-trained for 50k iteration on synthetic DDF, with the loss function set as:

$$\mathcal{L}_{\text{syn}} = \sum \|\phi - \tilde{\phi}\|_2^2 + \lambda \sum \|\nabla \phi\|_2^2 \tag{6}$$

Then the models are trained on real data for 100k iterations with the loss function:

$$\mathcal{L} = \mathcal{D}(\boldsymbol{I}^{\text{t}} - \phi(\boldsymbol{I}^{\text{s}})) + \lambda \|\nabla \phi \odot \mathrm{e}^{-\|\nabla \boldsymbol{I}^{\text{t}}\|_2^2}\|_2^2 \tag{7}$$

where normalized cross correlation and mean squared error are used in abdomen and lung CT respectively for $\mathcal{D}$ following [4]. The whole training takes one week, including the data transfer, pretraining and fine-tuning. With a training batch size of 3, The model was end-to-end trained with Adam optimizer with the initial learning rate set as 0.001.

3.3 Implementation and Evaluation

Implementation: The code for inter-subject image registration tasks was developed based on the framework of [3] in Python using Tensorflow and Keras. It was run on Nvidia Tesla P100-SXM2 GPU with 16GB memory, and Intel(R) Xeon(R) Gold 6126 CPU @ 2.60GHz.

Baselines: We compared RMAn with the relevant state-of-the-art networks. The Voxelmorph [4] is adopted as the representative state-of-the-art, deep learning method of direct regression (DR). The composite network combing CNN (Global-net) and U-net (Local-net) following to [10], recursive cascaded network (RCN) [24] were also adopted into the framework as the relevant baselines representing multi-stage (MS) networks, as well as D-net [26] was adopted for DIR based on the MA mechanism.

Evaluation Criterion: Following [22], we calculated the Dice Coefficient Similarity (DSC), Hausdorff Distance (HD), and Average Surface Distance (ASD) on annotated organs for the performance evaluation of nine organs in abdomen CT and one organ (lung) in chest CT. We additionally calculated the negative number of Jacobian determinant in tissues' region (detJ) for rationality evaluation on prediction. The model size and running time for comparison with the previous methods on inter-subject registration of lung and abdomen are shown in Tab. 1.

4 Results

Table 1. Average of Dice Similarity Coefficient (DSC), Average Surface Distance (ASD), Hausdorff Distance (HD) and negative number of Jacobian determinant in tissues' region (detJ) for unsupervised inter-subject registration of abdomen and chest CT using the Voxelmorph (VM1) [4] and its enhanced version with double number of feature channels (VM2), D-net [26] adopted for deformable registration, convolution networks cascaded with U-net (Cn+Un) [10], 5-recursive cascaded network based on the structure of the Voxelmorph (RCn) [24], and our RMAn network, with different registration (reg.) types and varying Parameter Number (#Par), and Time cost per Pair of Images (TPI).

model	reg. type	abdomen (9 organs)				chest (lung)				efficiency	
		DSC↑ (%)	HD↓ (mm)	ASD↓ (mm)	detJ↓ (e3)	DSC↑ (%)	HD↓ (mm)	ASD↓ (mm)	detJ↓ (e3)	#Par↓ (e6)	TPI↓ (sec)
Initial	–	30.9	49.5	16.04	–	61.9	41.6	15.86	–	–	–
VM1	DR	44.7	**43.8**	9.24	2.23	84.0	32.9	6.38	5.94	0.36	0.23
VM2	DR	51.9	45.0	8.40	4.03	88.8	32.0	5.02	15.58	1.42	0.25
Dnet	MA	47.4	47.6	8.72	5.28	88.3	33.2	5.01	10.38	0.40	0.41
Cn+Un	MS	53.6	44.6	7.84	4.13	91.1	**29.7**	3.84	4.23	2.11	0.36
RCn	MS	**55.6**	44.9	7.79	2.91	89.8	33.1	4.68	5.68	0.36	0.44
RMAn	MS+MA	55.2	45.1	**7.78**	4.32	**92.0**	31.8	**3.83**	4.53	0.40	0.67

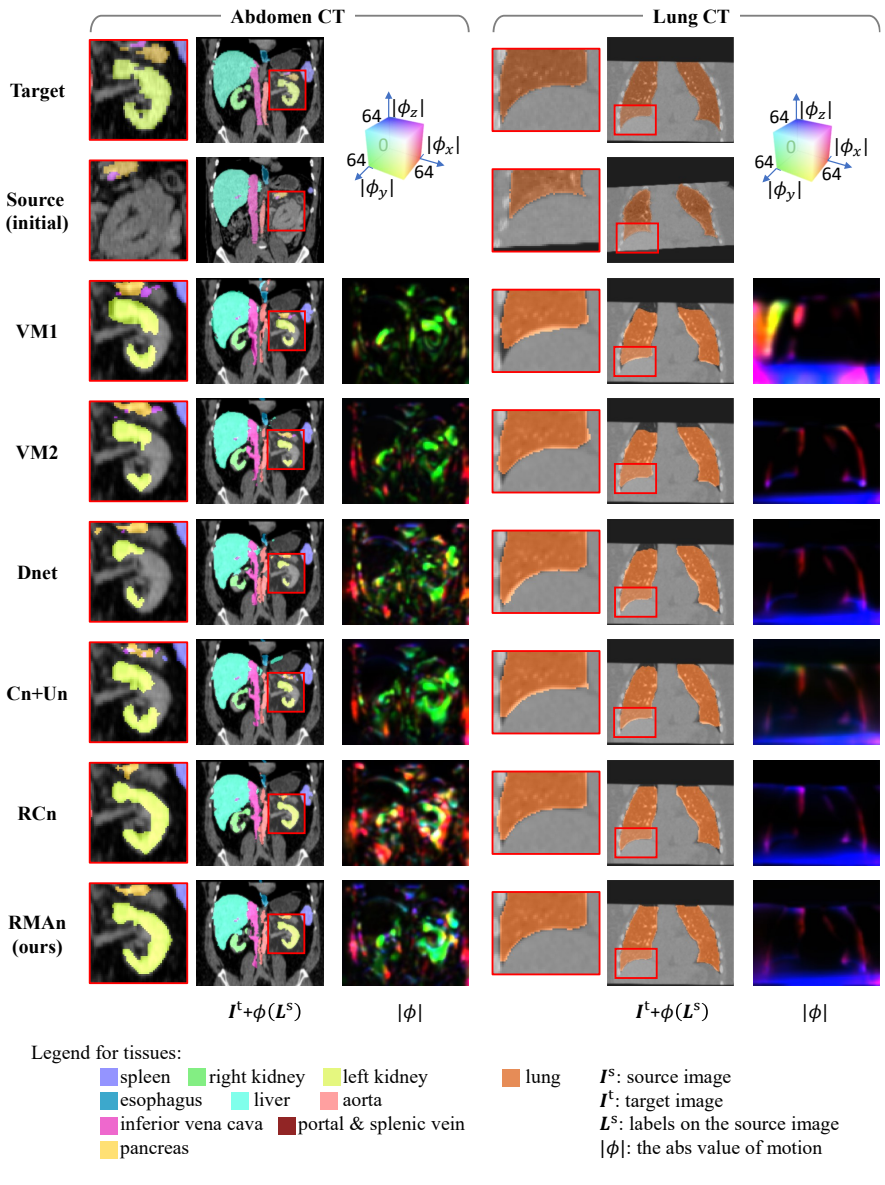

Fig. 3. Qualitative example in chest CT shows our network achieves plausible registration, with a significant improvement, especially at the edge area of the left kidney and the lung.

Comparison with the state-of-the-art Networks: Our proposed RMAn is compared with other methods on unsupervised DIR of abdomen and chest CT using all 10 organs. With an intuitive qualitative results shown in Fig. 3, RMAn achieves better performance on registration with an improvement in the area of lung boundaries (as depicted by the red box) and a plausible registration on the nine organs in the abdomen CT scans. The

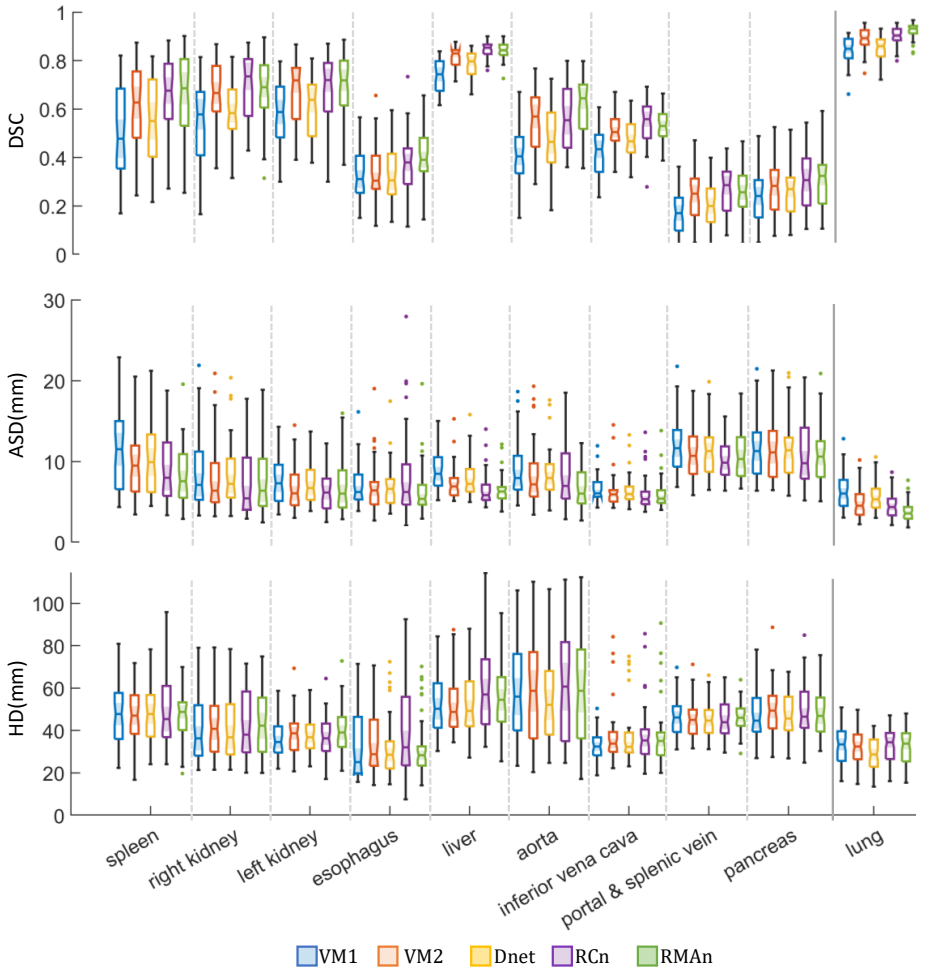

Fig. 4. RMANs achieve the best registration of the lung in chest CT scans as well as one of the best in the abdomen CT scans.

Table 2. Ablation study on recursive structure by inter-subject image registration of abdomen CT and lung CT using, with varying setting of recursive number (Rec. No.) for training and testing.

model	Rec. No.		abdomen (9 organs)				chest (lung)				efficiency	
	K_{train}	K_{infer}	DSC↑ (%)	HD↓ (mm)	ASD↓ (mm)	detJ↓ (e3)	DSC↑ (%)	HD↓ (mm)	ASD↓ (mm)	detJ↓ (e3)	#Par↓ (e6)	TPI↓ (sec)
MAn	1	1	47.4	47.6	8.72	5.28	88.3	33.2	5.01	10.38	0.40	0.41
RMAn	2	2	52.2	45.5	8.35	5.20	91.2	**31.6**	4.16	6.74	0.40	0.64
RMAn	3	3	54.4	44.9	7.91	5.01	91.4	32.6	4.01	5.36	0.40	0.65
RMAn	3	5	**55.2**	**45.1**	**7.78**	**4.32**	**92.0**	31.8	**3.83**	**4.53**	0.40	0.65

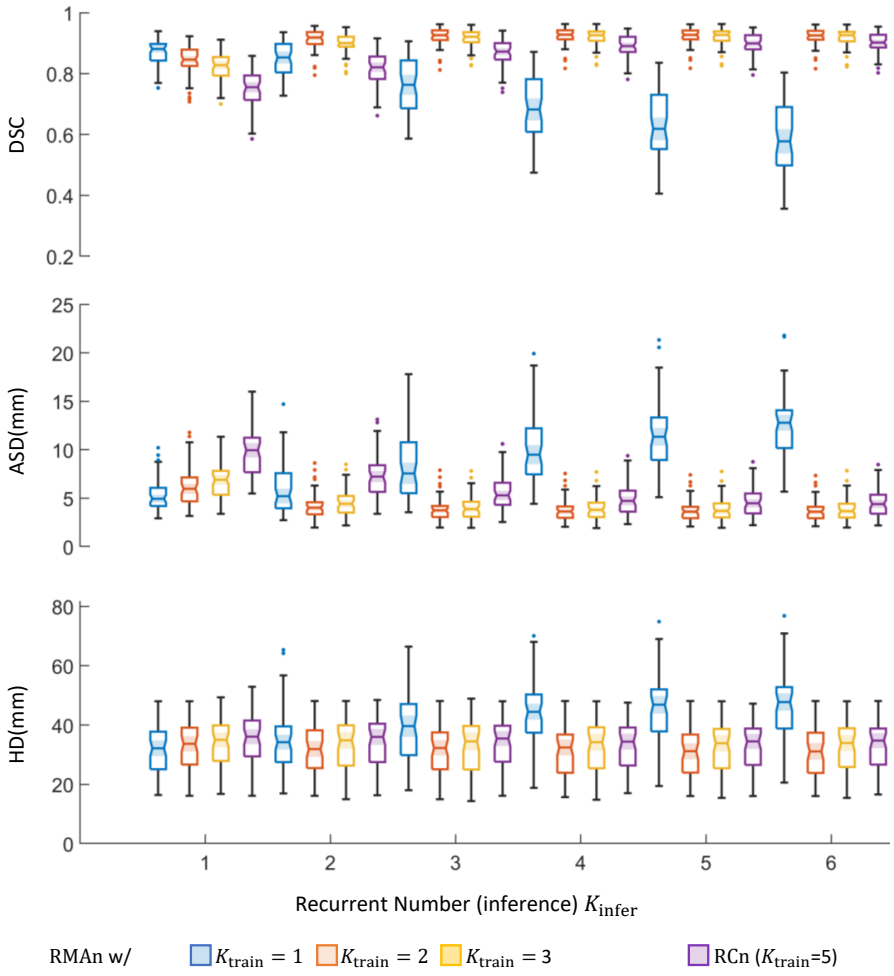

Fig. 5. The registration results on chest CT using our RMANs and the baseline RCn, with varying recursive number both for training and inference, shows that, with the increase of recursive number (inference), the model with recursive number (training) 2 and 3 achieve higher accuracy and converge closely, while it get worse with recursive number (training) 1, and RMAn outperform RCN with each K_{infer} in terms of DSC and ASD.

quantitative results shown in Fig. 4 illustrate that our RMAn achieves the best on lung and one of the best on the other nine abdominal organs. More numerical results are shown in Tab. 1, which demonstrates our network achieved comparable performance in this task with lower computational cost.

Ablation Study: Comparing VM1 and D-net in Tab. 1, the MA based architecture outperforms the pure encoder-decoder structure in two dataset with comparable network scale. To validate the effect of recursive architecture, we also tried several combina-

tion on varying recursive number for training and testing stage respectively on experiments of abdomen and lung CT as shown in Tab. 2 and Fig. 5. Comparing RMAn ($K_{\text{train}} = 1, K_{\text{infer}} = 1$) with others, the results show recursive architecture used in both training and testing phase results in the improved accuracy both in chest and abdomen CT scans, and the larger recurrent number for training could bring significant improvement. In addition, architecture reduces the negative number of Jacobian determinant, which thus improves the rationality of registration.

Number of Recurrent Stages: Furthermore, RMAn is tested with varying recurrent number for both training and inference as shown in Fig. 5. Surprisingly, the performance of RMAn with recurrent number $K_{\text{train}} = 1$ and $K_{\text{infer}} > 1$ for training and inference is even worse than MAn ($K_{\text{train}} = 1$ and $K_{\text{infer}} = 1$). This is probably due to the lack of recursive pattern during training for $K_{\text{train}} = 1$. As shown in Fig. 5, the RMAn with $K_{\text{train}} = 2$ and $K_{\text{train}} = 3$ as well as the RCn achieve improvement with more K_{infer}. We also compare our RMAn with baseline RCn based on varying K_{infer} as shown in Fig. 5. It shows RMAn outperform RCn for varying $K_{\text{infer}} \in [1, 8]$ in terms of DSC and ASD.

5 Discussion and Conclusion

The novel RMAn design is proposed based on the MA structure incorporated in a recursive architecture. It achieves the best registration results in the inter-subject lung CT registration and one of the best on other 9 organs in abdominal CT scans compared with the state of the art networks. The recursive architectures for registration are also investigated via varying training and inference recurrent number. The results show that larger inference recurrent number can improve the registration results, and on the other hand, also implies a small influence of the training recurrent number as long as the subnetwork is able to learn the pattern of recursive registration. The comparison of RMAn with RCn also proves the accuracy improvement stemming from the MA. In future, the proposed RMAn will be also applied to multi-modal image registration.

Acknowledgements

This work was supported by a Kennedy Trust for Rheumatology Research Studentship, the Centre for OA Pathogenesis Versus Arthritits (Versus Arthritis grant 21621). B. W. Papież acknowledges Rutherford Fund at Health Data Research UK (MR/S004092/1)

References

1. Aggarwal, H.K., Mani, M.P., Jacob, M.: MoDL: Model-based deep learning architecture for inverse problems. IEEE Transactions on Medical Imaging **38**(2), 394–405 (2018)
2. Avants, B.B., Epstein, C.L., Grossman, M., Gee, J.C.: Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. Medical Image Analysis **12**(1), 26–41 (2008)

3. 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)
4. 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)
5. Chen, J., Du, Y., He, Y., Segars, W.P., Li, Y., Frey, E.C.: Transmorph: Transformer for unsupervised medical image registration. arXiv preprint arXiv:2111.10480 (2021)
6. Dalca, A., Hu, Y., Vercauteren, T., Heinrich, M., Hansen, L., Modat, M., De Vos, B., Xiao, Y., Rivaz, H., Chabanas, M., et al.: Learn2reg-the challenge (2020)
7. Heinrich, M.P.: Closing the gap between deep and conventional image registration using probabilistic dense displacement networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 50–58. Springer (2019)
8. Heinrich, M.P., Jenkinson, M., Papież, B.W., Brady, S.M., Schnabel, J.A.: Towards Real-time Multimodal Fusion for Image-Guided Interventions using Self-similarities. In: International conference on medical image computing and computer-assisted intervention. pp. 187–194. Springer (2013)
9. Hering, A., Murphy, K., van Ginneken, B.: Learn2Reg Challenge: CT Lung Registration - Training Data (May 2020). https://doi.org/10.5281/zenodo.3835682, https://doi.org/10.5281/zenodo.3835682
10. Hu, Y., Modat, M., Gibson, E., Li, W., Ghavami, N., Bonmati, E., Wang, G., Bandula, S., Moore, C.M., Emberton, M., et al.: Weakly-supervised convolutional neural networks for multimodal image registration. Medical image analysis **49**, 1–13 (2018)
11. Jaderberg, M., Simonyan, K., Zisserman, A., et al.: Spatial transformer networks. Advances in neural information processing systems **28**, 2017–2025 (2015)
12. Jia, X., Thorley, A., Chen, W., Qiu, H., Shen, L., Styles, I.B., Chang, H.J., Leonardis, A., De Marvao, A., O'Regan, D.P., et al.: Learning a Model-Driven Variational Network for Deformable Image Registration. IEEE Transactions on Medical Imaging **41**(1), 199–212 (2021)
13. Li, Z., Liu, X., Drenkow, N., Ding, A., Creighton, F.X., Taylor, R.H., Unberath, M.: Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6197–6206 (2021)
14. Mok, T.C., Chung, A.: Fast symmetric diffeomorphic image registration with convolutional neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4644–4653 (2020)
15. Rueckert, D., Sonoda, L.I., Hayes, C., Hill, D.L., Leach, M.O., Hawkes, D.J.: Nonrigid registration using free-form deformations: application to breast mr images. IEEE transactions on medical imaging **18**(8), 712–721 (1999)
16. Schnabel, J.A., Heinrich, M.P., Papież, B.W., Brady, J.M.: Advances and challenges in deformable image registration: from image fusion to complex motion modelling. Medical Image Analysis **33**, 145–148 (2016)
17. Song, X., Guo, H., Xu, X., Chao, H., Xu, S., Turkbey, B., Wood, B.J., Wang, G., Yan, P.: Cross-Modal Attention for MRI and Ultrasound Volume Registration. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 66–75. Springer (2021)
18. Sotiras, A., Davatzikos, C., Paragios, N.: Deformable medical image registration: A survey. IEEE Transactions on Medical Imaging **32**(7), 1153–1190 (2013)
19. Sun, J., Shen, Z., Wang, Y., Bao, H., Zhou, X.: Loftr: Detector-free local feature matching with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8922–8931 (2021)

20. Thirion, J.P.: Image matching as a diffusion process: an analogy with maxwell's demons. Medical Image Analysis **2**(3), 243–260 (1998)
21. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems. pp. 5998–6008 (2017)
22. de Vos, B.D., Berendsen, F.F., Viergever, M.A., Sokooti, H., Staring, M., Išgum, I.: A deep learning framework for unsupervised affine and deformable image registration. Medical image analysis **52**, 128–143 (2019)
23. Zhang, Y., Pei, Y., Zha, H.: Learning dual transformer network for diffeomorphic registration. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 129–138. Springer (2021)
24. 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)
25. Zhao, S., Lau, T., Luo, J., Eric, I., Chang, C., Xu, Y.: Unsupervised 3D end-to-end medical image registration with volume tweening network. IEEE journal of biomedical and health informatics **24**(5), 1394–1404 (2019)
26. Zheng, J.Q., Lim, N.H., Papież, B.W.: D-net: Siamese based Network for Arbitrarily Oriented Volume Alignment. In: International Workshop on Shape in Medical Imaging. pp. 73–84. Springer (2020)
27. Zhou, X.Y., Zheng, J.Q., Li, P., Yang, G.Z.: ACNN: A Full Resolution DCNN for Medical Image Segmentation. In: 2020 IEEE International Conference on Robotics and Automation (ICRA). pp. 8455–8461. IEEE (2020)