CAR-Net

Citation for published version (APA):
Chen, X., Xia, Y., Ravikumar, N., & Frangi, A. F. (2021). *CAR-Net: Unsupervised Co-Attention Guided Registration Network for Joint Registration and Structure Learning*.

Citing this paper
Please note that where the full-text provided on Manchester Research Explorer is the Author Accepted Manuscript or Proof version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version.

CAR-Net: Unsupervised Co-Attention Guided Registration Network for Joint Registration and Structure Learning

Xiang Chen[1], Yan Xia[1,2], Nishant Ravikumar[1,2], and Alejandro F Frangi[1,2,3,4]

[1] Center for Computational Imaging and Simulation Technologies in Biomedicine, School of Computing, University of Leeds, Leeds,UK

[2] Biomedical Imaging Department, Leeds Institute for Cardiovascular and Metabolic Medicine, School of Medicine University of Leeds, Leeds, UK

[3] Department of Cardiovascular Sciences, KU Leuven, Leuven, Belgium

[4] Department of Electrical Engineering, KU Leuven, Leuven, Belgium

Abstract. Image registration is a fundamental building block for various applications in medical image analysis. To better explore the correlation between the fixed and moving images and improve registration performance, we propose a novel deep learning network, Co-Attention guided Registration Network (CAR-Net). CAR-Net employs a co-attention block to learn a new representation of the inputs, which drives the registration of the fixed and moving images. Experiments on UK Biobank cardiac cine-magnetic resonance image data demonstrate that CAR-Net obtains higher registration accuracy and smoother deformation fields than state-of-the-art unsupervised registration methods, while achieving comparable or better registration performance than corresponding weakly-supervised variants. In addition, our approach can provide critical structural information of the input fixed and moving images simultaneously in a completely unsupervised manner.

Keywords: Medical Image Registration · Cardiac Image Registration · Co-Attention · Deep Learning.

1 Introduction

Image registration is an active research area in computer vision and medical image analysis. It is central to facilitating numerous downstream tasks (e.g. anatomical motion tracking/deformation analysis, disease diagnosis and monitoring progression, among others). Traditional methods view this task as an iterative process to optimise the parameters of the deformation model under the guidance of dissimilarity/similarity metrics. Therefore, despite achieving high accuracy and possessing favourable geometric attributes, traditional registration methods are generally time-consuming. To speed up image registration, recently, deep learning (DL)-based methods have been widely applied in this domain, where the inference is just one forward pass through the network.

DL-based registration methods could be roughly divided into three categories, supervised, unsupervised and weakly-supervised methods. Supervised methods [1,2,3,4] can achieve comparable registration performance to traditional methods, but require only a fraction of time (relative to the latter) to register images during inference. However, the ground-truth deformation fields are challenging to obtain, which makes it an obstacle for large-scale training of DL-based methods. This problem can be overcome through unsupervised methods [5,6,7], which generally comprise an end-to-end convolutional neural network (CNN) to predict the deformation fields and a spatial transformer network (STN) [8] to warp the moving image towards the fixed image. Dissimilarity/similarity between the warped moving image and the fixed image, and regularisation of the deformation fields constitute the loss function used to train networks. Although the ground-truth deformation fields are difficult to obtain, other types of auxiliary information (e.g. segmentation masks or landmarks) could also be used to guide registration. Weakly-supervised registration methods are based on introducing such 'weak' labels, and have been shown to achieve better registration performance than unsupervised methods [5,6,9,10].

However, most previous DL-based registration networks focus on building an end-to-end network from the pair of moving and fixed images to estimate deformation fields, seldom considering if the input pairs are sufficiently descriptive for the task. To tackle this issue, Lee et al. [11] proposed an Image-and-Spatial Transformer Network (ISTN), containing a dedicated image transformer network (ITN) to generate new representations of the inputs, and an STN to register the new representations of the inputs, achieving better performance than using STN to register the original inputs directly. However, ISTN uses a shared CNN to extract new representations of the input images, which requires annotations (e.g. segmentations, landmarks or centerlines) and thus does not exploit the intrinsic correlation between them in an unsupervised manner. Several recent weakly-supervised approaches have proposed to first register the segmentation masks [12] or semantic features [13] at a global level instead of directly aligning the original input images, which also highlights the importance of introducing additional structural information. However, sourcing such additional annotations can be expensive or infeasible for specific applications, limiting the generality of such weakly-supervised methods.

Image registration always involves two or more images as inputs (to be co-registered), where the contents in these images are highly-correlated but may comprise differences in spatial/shape characteristics. We argue that highlighting the correlated regions in the moving and fixed images can help learn coarse structural information from the input images automatically. To minimise the impact of irrelevant pixels and improve registration performance, a co-attention guided registration network is proposed in this paper. The co-attention block is used to generate new representations of the input images, where, correlated pixels are highlighted and irrelevant context is dismissed. With the learned new representations, the performance of DL-based registration networks can be further improved. Additionally, following training on a large-scale dataset, the repre-

sentations learned in an unsupervised manner, can be visualised as a coarse prediction of correlated structural information.

In summary, there are three main contributions of our proposed method: (1) a 3D co-attention block is designed to exploit the correlation of input image pairs to be registered, which can be easily incorporated into other networks. (2) Using the co-attention block, we build a new DL-based registration network, CAR-Net, which achieves state-of-the-art registration performance on cine-cardiac magnetic resonance (CMR) images from UK Biobank Imaging Study (UKBB). (3) Our co-attention block is shown to provide coarse structural information in an unsupervised manner, which is of potential value for developing explainable AI solutions to medical image registration.
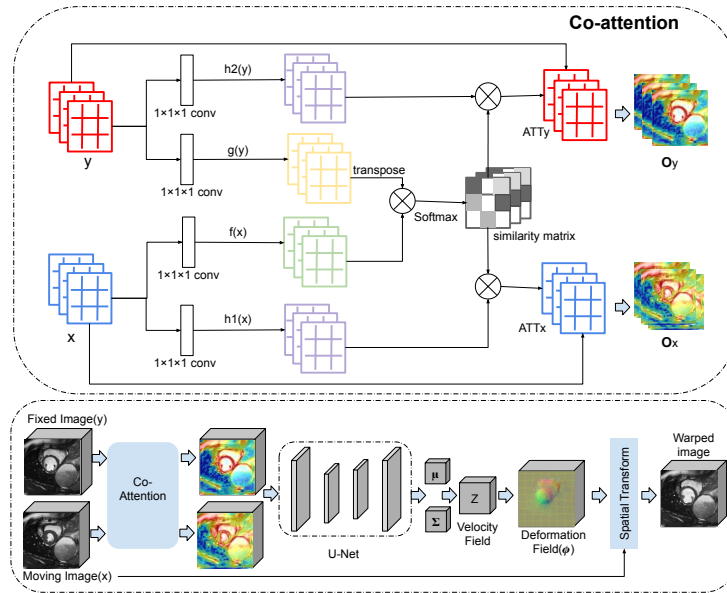
Fig. 1. Schema of our proposed CAR-Net. The whole registration network is displayed in the bottom row, with the details of the co-attention block in the top row. The co-attention block takes feature maps from the moving and fixed image as input (denoted as x and y). It predicts corresponding attention maps (denotes as ATT_x and ATT_y) and outputs ($\mathbf{O}_x$ and $\mathbf{O}_y$), with the same size as inputs. The CMR images were reproduced with the permission of UK Biobank©.

2   Methods

Given the moving image $\mathbf{I}_M$ and fixed image $\mathbf{I}_F$, image registration aims to find a coordinate mapping function $\phi(\circ)$ from $\mathbf{I}_M$ to $\mathbf{I}_F$, formulated as,

$$\mathbf{I}_F = \phi(\mathbf{I}_M). \tag{1}$$

Unsupervised DL-based image registration methods [5,6] initially concatenate $\mathbf{I}_M$ and $\mathbf{I}_F$, following which, an encoder-decoder is applied to predict the deformation fields. Finally, a STN block is used to deform the $\mathbf{I}_M$ to $\mathbf{I}_F$ based on the learned deformation fields. Networks are trained by optimising a loss function that computes a dissimilarity/similarity metric between the warped moving image and $\mathbf{I}_F$, and regularises the estimated deformation fields. Our approach, CAR-Net, utilises each of these fundamental building blocks of unsupervised registration networks, and incorporates a co-attention block that learns new image representations from the original inputs automatically.

Co-attention Block. To learn the transformation from the moving image to the fixed image, the key is to establish correspondence between these two images. Therefore, pixels across different spatial positions in these two images do not share the same importance for registration. In fact, individual pixels/regions in the images may contain confounding information that can adversely impact registration accuracy. To limit the impact of such pixels/regions on overall registration accuracy, it is intuitive to focus the attention of the registration algorithm on the key points/structures in the moving and fixed images. To this end, we propose a co-attention block to exploit the correlations between the input moving and fixed images, for registration (as shown in Fig. 1).

The co-attention block is inspired by [14,15,16], where it is employed for question-answering and video segmentation tasks. It is however, well-suited to the task of image registration as well (as there are two highly correlated inputs, the moving and fixed image). Our co-attention block takes both features from moving and fixed images as input, and predicts the corresponding co-attention feature maps, where the highly-correlated and motion-related pixels in both feature maps are highlighted. The moving and fixed image feature maps $\mathbf{F}_{mov}, \mathbf{F}_{fix} \in \mathcal{R}^{W \times H \times D \times C}$ (C, W, H and D are channels, width, height and depth of feature maps) are first transformed into two different feature spaces by two corresponding $1 \times 1 \times 1$ convolution layers (denoted as $f(\circ)$ and $g(\circ)$) and flattened (from $\mathcal{R}^{W \times H \times D \times C}$ to $\mathcal{R}^{N \times C}, N = W \times H \times D$) to calculate similarity matrix $\mathbf{S} \in \mathcal{R}^{N \times N}$. Based on similarity matrix $\mathbf{S}$ and feature maps learned with two additional $1 \times 1 \times 1$ convolution layers $h_1(\circ)$ and $h_2(\circ)$ from the input, the fixed attention maps $\mathbf{ATT}_{fix}$ and moving attention maps $\mathbf{ATT}_{mov}$ are

computed. The process of co-attention can be formulated as,

$$\begin{aligned}
\mathbf{S} &= f(\mathbf{F}_{mov}) \times g(\mathbf{F}_{fix})^T, \\
\mathbf{ATT}_{mov} &= Softmax(\mathbf{S}) \times h_2(\mathbf{F}_{fix}), \\
\mathbf{ATT}_{fix} &= Softmax(\mathbf{S}^T) \times h_1(\mathbf{F}_{mov}), \\
\mathbf{O}_{mov} &= h_1(\mathbf{F}_{mov}) + \alpha_1 \sigma(\mathbf{ATT}_{mov}) \cdot h_1(\mathbf{F}_{mov}), \\
\mathbf{O}_{fix} &= h_2(\mathbf{F}_{fix}) + \alpha_2 \sigma(\mathbf{ATT}_{fix}) \cdot h_1(\mathbf{F}_{fix}),
\end{aligned} \tag{2}$$

where, $\mathbf{O}_{mov}$ and $\mathbf{O}_{fix}$ are the output feature maps (learned new representations) of $\mathbf{F}_{mov}$ and $\mathbf{F}_{fix}$ after co-attention respectively. $Softmax(\circ)$ is the Softmax function, applied to the last channel of the similarity matrix $\mathbf{S}$. $\sigma(\circ)$ denotes the Sigmoid function, which is a $1 \times 1 \times 1$ convolution layer followed by a Sigmoid activation. The α_1 and α_2 are weights for the attention maps, learned automatically by back-propagation.

Network Architecture. CAR-Net can be seen as a combination of a co-attention block with a generic unsupervised registration network. The co-attention block learns new representations ($\mathbf{O}_{mov}$ and $\mathbf{O}_{fix}$) of the input moving and fixed images, which serve as inputs to the registration block that predicts deformation fields and yields warped moving images. The co-attention block can be directly applied to the input moving and fixed images, however, we apply two down-sampling layers before the co-attention block and an up-sampling block after it, due to computational constraints (GPU memory). The parameters of the co-attention block and registration block are optimised using the same loss function, allowing the co-attention block to learn the most relevant/correlated structural features from the input images in an unsupervised manner.

To ensure the generated deformation fields are smooth, we parameterise deformation fields ϕ by stationary velocity fields (SVF) z following previous research [7,6]. The registration part is a probabilistic model, comprising two main sub-blocks, a U-Net [17] like structure and a spatial transform block. The former is an encoder-decoder network (three down-sampling layers and three up-sampling layers) to estimate deformation fields ϕ_z, while the latter is used to deform the moving image to the fixed image based on the learned deformation fields from U-Net. The U-Net takes the concatenation of fixed and moving features $[\mathbf{F}_{mov}, \mathbf{F}_{fix}, \mathbf{O}_{mov}, \mathbf{O}_{fix}]$ as input and predicts feature maps at $64 \times 64 \times 16$, with which mean μ and variance Σ matrix are computed. Similar to [6], the velocity field z is sampled from a multivariate Gaussian distribution using the estimated μ and Σ, followed by an integration layer and up-sampling layer to predict a diffeomorphic deformation field ϕ_z. The predicted deformation field and moving image are fed into the spatial transform block (more details can be found in [8]) to predict the warped moving image.

Loss Functions. The loss function comprises two terms, the similarity loss and the regularisation of the deformation fields. In this paper, we choose normalised cross-correlation (NCC) as the similarity loss L_s, to evaluate the similarity between the warped moving image and the fixed image. Given warped

moving image x and fixed image y, L_s is computed as,

$$L_s = 1 - \frac{\sum_i (x_i - x_m)(y_i - y_m)}{\sqrt{\sum_i (x_i - x_m)^2}\sqrt{\sum_i (y_i - y_m)^2}}, \quad (3)$$

where, x_i, y_i are the intensity of the i-th pixel in the warped moving image and fixed image respectively, and x_m, y_m are the mean intensities in the corresponding images.

The deformation fields are regularised to improve the smoothness of predicted deformation fields. Following VM [6], we calculate the KL divergence between the approximate posterior $q_\psi(z|\mathbf{I}_F; \mathbf{I}_M)$ and the prior $p(z)$ $(p(z) = \mathcal{N}(z; 0, \Sigma_z))$ of velocity field z, formulated as,

$$L_r = KL(q_\psi(z|\mathbf{I}_F; \mathbf{I}_M)||p(z|\mathbf{I}_F; \mathbf{I}_M)), \quad (4)$$

where $q_\psi(z|\mathbf{I}_F; \mathbf{I}_M)$ is a multivariate normal,

$$q_\psi(z|\mathbf{I}_F; \mathbf{I}_M) = \mathcal{N}(z; \mu_{\mathbf{z}|\mathbf{I}_F, \mathbf{I}_M}, \Sigma_{\mathbf{z}|\mathbf{I}_F, \mathbf{I}_M}), \quad (5)$$

where $\mu_{\mathbf{z}|\mathbf{I}_F, \mathbf{I}_M}$ and $\Sigma_{\mathbf{z}|\mathbf{I}_F, \mathbf{I}_M}$ are the mean and variance of the distribution, learned by convolution layers. Therefore, the total loss function is formulated as,

$$L_{total} = \lambda_0 \times L_s + \lambda_1 \times L_r, \quad (6)$$

where, λ_0 and λ_1 are hyper-parameters that weight the relative influence of each loss term, which is to be defined in experiments.

3   Experiments and Results

Data and Implementation. To demonstrate the performance of CAR-Net, we perform intra-subject registration experiments on cine CMR images (spatial resolution at $\sim 1.8 \times 1.8 \times 10mm^3$) from UKBB. 2,000 subjects are randomly chosen from the UKBB dataset, each containing volumes at end-diastole (ED) and end-systole (ES), totally 4,000 volumes. These 2,000 subjects are then randomly split into a training set (1,600 subjects) and a testing set (400 subjects). We focus on intra-subject registration between ED and ES (i.e. ED-to-ES and ES-to-ED). A total of 3,200 image pairs were used for training, and 800 pairs for testing. Following previous cardiac registration methods [7], we first resample all original CMR images into a spatial resolution at $1.50 \times 1.50 \times 3.15mm^3$, then crop them into the same size of $128 \times 128 \times 32$ (using zero-padding if the slices are less than 32). To quantitatively evaluate the registration performance, we obtain segmentation masks for all volumes using the segmentation method proposed in [18]. CAR-Net was implemented in Python using Keras, and trained using a single Tesla M60 GPU. The Adam optimiser, with a learning rate of 1e-4 was used to train our network. We set the batch size to 2 due to limited GPU memory. Hyper-parameters λ_0 and λ_1 were determined empirically, and set to
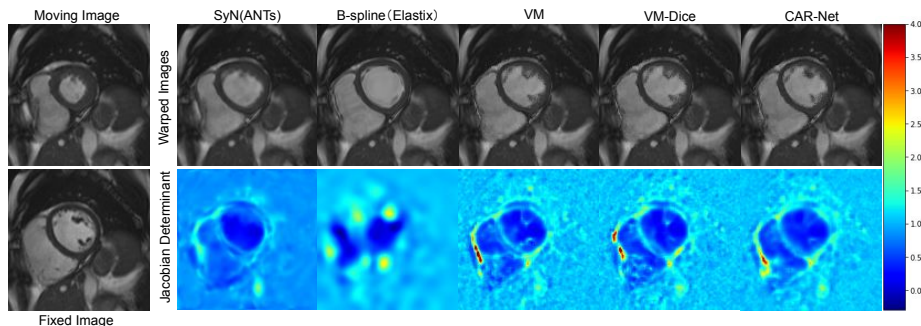
Fig. 2. Qualitative results for our CAR-Net and baseline networks viz. B-spline, SyN, VM and VM-Dice. The input moving and fixed images are on the left, while the corresponding warped moving images and Jacobian determinant of deformation fields (color bar is from -0.3 to 4) predicted by different approaches are on the right. The CMR images were reproduced with the permission of UK Biobank©.

20 and 0.1, respectively. The source code will be publicly available on Github following review and acceptance of the paper.

Comparison and Analysis. We compared the performance of CAR-Net with traditional and state-of-the-art DL-based registration methods, using data from UKBB. For traditional methods, we choose the B-spline registration method (max iteration step is 2000, sampling 6000 random points per iteration) available in SimpleElastix [19] and Symmetric Normalisation (SyN) [20] in ANTs (resolution level is 3 with 100 iterations in each sampling level) for comparison. For DL-based methods, we compare our CAR-Net with VM [6] (using NCC as similarity loss) and a weakly-supervised version of VM (named VM-Dice), which uses Dice score as an additional loss to guide the network training. Note that, the main difference between CAR-Net and VM is the incorporation of co-attention block, while the other hyper parameter settings of the networks (e.g. convolution kernels, channels in the registration block and learning rate) are the same.

Table 1. Quantitative comparison between CAR-Net and state-of-the-art methods using the Dice score of LVBP, LVM, RV, average Dice (denoted as Avg. Dice), HD and number of foldings (denoted as $|J_\phi| \leq 0$).

Methods	LVBP Dice (%)	LVM Dice (%)	RV Dice (%)	Avg. Dice (%)	HD (mm)	$|J_\phi| \leq 0$
before Reg	57.68 ± 6.21	30.88 ± 8.68	55.13 ± 7.51	47.90 ± 6.33	12.91 ± 2.48	–
B-spline	74.44 ± 11.50	68.06 ± 7.20	61.76 ± 12.05	68.09 ± 8.76	13.72 ± 3.57	1671.97 ± 1829.90
SyN	70.92 ± 9.36	57.88 ± 10.59	60.30 ± 8.35	63.03 ± 8.29	12.98 ± 2.68	0.00 ± 0.00
VM	81.73 ± 8.71	72.04 ± 4.65	65.73 ± 9.62	73.16 ± 6.26	12.96 ± 3.14	0.29 ± 2.32
VM-Dice	82.28 ± 8.75	72.53 ± 4.59	66.30 ± 9.67	73.70 ± 6.28	13.00 ± 3.24	0.22 ± 1.75
CAR-Net	$\mathbf{82.59 \pm 8.35}$	$\mathbf{72.87 \pm 4.60}$	$\mathbf{66.58 \pm 9.57}$	$\mathbf{74.01 \pm 6.18}$	$\mathbf{12.73 \pm 3.21}$	$\mathbf{0.00 \pm 0.00}$

Qualitative and Quantitative Results. The qualitative comparison between CAR-Net and state-of-the-art methods is shown in Fig. 2. We find that, the prediction by SyN and B-spline only captures coarse deformation from the moving image to the fixed image. DL-based methods achieve better registration performance than traditional methods, while warped images predicted by CAR-Net are more consistent with the fixed image, with smoother deformations.

Dice score for left ventricle blood pool (LVBP), left ventricle myocardium (LVM), right ventricle (RV) and average Dice), Hausdorff Distance (HD), and the number of foldings in the estimated deformation field (where Jacobian determinant is non-positive), are used to quantitatively evaluate the performance of our proposed method (as shown in Table1). Higher Dice score and lower HD stand for higher registration accuracy, while fewer foldings points to smoother deformation fields.

Quantitative results are consistent with the visual assessment of the image registration using each approach. DL-based methods were found to achieve much higher Dice scores than traditional registration methods (of 5-10% improvement). Our proposed CAR-Net consistently outperforms VM and other state-of-the-art methods in all metrics, achieving a higher Dice score ($\sim 1\%$ improvement than VM), lower HD and fewer foldings in the deformation fields. Note that, CAR-Net achieves completely smooth deformation, without foldings in the deformation fields. According to statistical analysis, the incorporation of a co-attention block (CAR-Net) provides statistically significant improvements over VM (i.e. without co-attention) in terms of the Dice scores for LVBP, LVM and the average (P-value<0.05). Although CAR-Net offers no significant improvement over the weakly-supervised version of VM (VM-Dice), CAR-Net is totally unsupervised and hence is not dependent on the availability of segmentation masks for structures/regions of interest.

Registration on Separate Regions. To capture the motion of specific regions, it is sometimes necessary to register them independently (e.g. LVBP, LVM, RV). Using the same trained models, we also compare the performance of CAR-Net and VM on the registration of separate LVBP, LVM and RV regions (generated based on the same testing set), as shown in Table2. Our CAR-Net achieves statistically significant improvements over both VM and VM-Dice in all metrics (Dice and HD) except the Dice score on LVBP. This further demonstrates the incorporation of co-attention block would bring more robust and accurate registration. A further visualisation and analysis of the learned co-attention maps is included in the supplementary material.

Table 2. Quantitative comparison between CAR-Net and VM on registration of sub regions (LVBP, LVM and RV) of cardiac images. Statistical significant improvements compared with other methods are highlighted with bold (P-value<0.05).

Methods	LVBP Dice (%)	LVBP HD (mm)	LVBP $|J_\phi| \leq 0$	LVM Dice (%)	LVM HD (mm)	LVM $|J_\phi| \leq 0$	RV Dice (%)	RV HD (mm)	RV $|J_\phi| \leq 0$
before Reg	57.68 ± 6.21	9.91 ± 1.98	–	30.88 ± 8.68	9.00 ± 1.25	–	55.13 ± 7.51	12.87 ± 2.37	–
VM	88.98 ± 4.93	5.11 ± 2.52	0.00 ± 0.11	61.97 ± 9.91	9.36 ± 1.74	0.00 ± 0.04	81.42 ± 6.80	9.92 ± 3.36	0.15 ± 1.91
VM-Dice	90.16 ± 4.71	4.73 ± 2.56	0.00 ± 0.05	64.23 ± 9.08	9.66 ± 1.84	0.00 ± 0.00	83.75 ± 5.85	9.25 ± 3.17	0.04 ± 0.68
CAR-Net	90.65 ± 3.91	$\mathbf{4.17 \pm 2.29}$	0.00 ± 0.00	$\mathbf{71.55 \pm 6.68}$	$\mathbf{8.13 \pm 2.03}$	0.00 ± 0.00	$\mathbf{86.31 \pm 5.02}$	$\mathbf{8.17 \pm 3.13}$	$\mathbf{0.00 \pm 0.00}$

4   Conclusion

This paper proposes a novel DL-based image registration network, CAR-Net, for 3D CMR image registration. The key idea is to incorporate 3D co-attention into a generic unsupervised registration framework, to learn highly-correlated structural information between the moving and fixed images automatically. Experiments on UKBB cardiac images demonstrate that CAR-Net achieves higher registration accuracy, completely smooth deformation fields and is more robust than state-of-the-art unsupervised methods and a weakly-supervised version of VM. CAR-Net employs co-attention to learn new representations of input images automatically, which could be seen as a coarse prediction of essential structural information and therefore, lends itself to interpreting network decisions.

References

1. Rohé, M.M., Datar, M., Heimann, T., Sermesant, M., Pennec, X.: SVF-Net: Learning Deformable Image Registration Using Shape Matching. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 266–274. Springer (2017)
2. Eppenhof, K.A., Pluim, J.P.: Pulmonary CT Registration Through Supervised Learning with convolutional Neural Networks. IEEE Transactions on Medical Imaging **38**(5), 1097–1105 (2018)
3. Eppenhof, K.A., Lafarge, M.W., Veta, M., Pluim, J.P.: Progressively Trained Convolutional Neural Networks for Deformable Image Registration. IEEE Transactions on Medical Imaging **39**(5), 1594–1604 (2019)
4. Cao, X., Yang, J., Zhang, J., Nie, D., Kim, M., Wang, Q., Shen, D.: Deformable Image Registration Based on Similarity-steered CNN Regression. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 300–308. Springer (2017)
5. 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)
6. Dalca, A.V., Balakrishnan, G., Guttag, J., Sabuncu, M.R.: Unsupervised Learning of Probabilistic Diffeomorphic Registration for Images and Surfaces. Medical Image Analysis **57**, 226–236 (2019)
7. Krebs, J., Delingette, H., Mailhé, B., Ayache, N., Mansi, T.: Learning A Probabilistic Model for Diffeomorphic Registration. IEEE Transactions on Medical Imaging **38**(9), 2165–2176 (2019)
8. Jaderberg, M., Simonyan, K., Zisserman, A., et al.: Spatial Transformer Networks. In: Advances in Neural Information Processing Systems. pp. 2017–2025 (2015)
9. Hu, Y., Modat, M., Gibson, E., Ghavami, N., Bonmati, E., Moore, C.M., Emberton, M., Noble, J.A., Barratt, D.C., Vercauteren, T.: Label-driven Weakly-supervised Learning for Multimodal Deformable Image Registration. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). pp. 1070–1074. IEEE (2018)
10. Mansilla, L., Milone, D.H., Ferrante, E.: Learning Deformable Registration of Medical Images with Anatomical Constraints. Neural Networks **124**, 269–279 (2020)

11. Lee, M.C., Oktay, O., Schuh, A., Schaap, M., Glocker, B.: Image-and-spatial Transformer Networks for Structure-guided Image Registration. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 337–345. Springer (2019)

12. Chen, L., Cao, X., Chen, L., Gao, Y., Shen, D., Wang, Q., Xue, Z.: Semantic Hierarchy Guided Registration Networks for Intra-subject Pulmonary CT Image Alignment. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 181–189. Springer (2020)

13. Ha, I.Y., Wilms, M., Heinrich, M.: Semantically Guided Large Deformation Estimation with Deep Networks. Sensors **20**(5), 1392 (2020)

14. Lu, X., Wang, W., Ma, C., Shen, J., Shao, L., Porikli, F.: See More, Know More: Unsupervised Video Object Segmentation with Co-attention Siamese Networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3623–3632 (2019)

15. Xiong, C., Zhong, V., Sochern, R.: Dynamic Coattention Networks For Question Answering. In: International Conference on Learning Representations (2017)

16. Lu, J., Yang, J., Batra, D., Parikh, D.: Hierarchical Question-image Co-attention for Visual Question Answering. In: Advances in Neural Information Processing Systems. pp. 289–297 (2016)

17. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 234–241. Springer (2015)

18. Bai, W., Sinclair, M., Tarroni, G., Oktay, O., Rajchl, M., Vaillant, G., Lee, A.M., Aung, N., Lukaschuk, E., Sanghvi, M.M., et al.: Automated Cardiovascular Magnetic Resonance Image Analysis with Fully Convolutional Networks. Journal of Cardiovascular Magnetic Resonance **20**(1), 65 (2018)

19. 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)

20. 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)