


Deep Learning for Cardiac Motion Estimation: Supervised vs. Unsupervised Training

Huaqi Qiu() , Chen Qin, Loic Le Folgoc, Benjamin Hou,
Jo Schlemper, Daniel Rueckert

Biomedical Image Analysis Group, Imperial College London, London, UK
`huaqi.qiu15@imperial.ac.uk`

Abstract. Deep learning based registration methods have emerged as alternatives to traditional registration methods, with competitive accuracy and significantly less runtime. Two different strategies have been proposed to train such deep learning registration networks: *supervised* training strategy where the model is trained to regress to generated ground truth deformation; and *unsupervised* training strategy where the model directly optimises the similarity between the registered images. In this work, we directly compare the performance of these two training strategies for cardiac motion estimation on cardiac cine MR sequences. Testing on real cardiac MRI data shows that while the *supervised* training yields more regular deformation, the *unsupervised* more accurately captures the deformation of anatomical structures in cardiac motion.

1 Introduction

Cardiac motion analysis assesses regional deformation parameters such as volume output, strain and torsion, which are indicative for the diagnosis and treatment for patients with cardiovascular diseases [8,9]. The deformation parameters can be derived from displacement field estimated from cardiac magnetic resonance (MR) images. Traditionally cardiac motion estimation is cast as a series of pairwise registration tasks. Shen et al. [8] extended a hierarchical attribute-matching based registration method to simultaneously estimate cardiac motion of all frames in a sequence by formulating cardiac motion as spatial-temporal 4D registration. Shi et al. [9] applied B-spline free-form deformation (FFD) registration [7] on both cine and tagged cardiac MR images by spatially weighting the complementary information from the two modalities.

Deep learning methods have been successfully applied to deformable registration, demonstrating competitive performances with significantly superior speed. Several methods that train deep convolutional neural networks (ConvNets) to perform one-shot prediction of the deformation between two images have been proposed. A critical difference in the proposed methods is the supervision signal used during training. On the one hand, networks are trained to perform a regression task using ground truth deformation that are acquired either via random simulation [3,10] or traditional registration algorithms [2,12]. These methods are

termed *supervised* methods since the ground truth of the deformation is used in training. On the other hand, several recent *unsupervised* methods opt to directly optimise the parameters of the network to maximise intensity-based similarity for all image pairs in a training dataset [1,11]. Most related to this work, [6] incorporated unsupervised registration method to provide complementary motion information for cardiac segmentation. Despite the advances of both *supervised* and *unsupervised* methods, it remains unclear which training strategy is more suitable for cardiac motion estimation.

In this work, we trained a deep learning registration network to perform cardiac motion estimation using both *supervised* and *unsupervised* training strategy, and compared the performances on both the accuracy and the regularity of the estimated motion. We show that the *unsupervised* model was able to extract motion that describes the deformation of anatomical structure more accurately, while the *supervised* model produced spatially smoother and more topology-preserving deformation.

2 Background

The objective of cardiac motion estimation is to determine the spatial transformation of cardiac structures over time. Let $\{I_t\}_{t=0,1,2,\dots,N_T}$ represent a sequence of cardiac cine MR images where N_T is the total number of frames and let $\mathbf{p}_0 \in \mathbb{R}^2$ denotes the position of a point on the first frame ($t = 0$). We can determine the spatial transformation $\mathcal{T}(\cdot)$ using image registration such that $I_0(\mathbf{p}_0)$ and $I_t(\mathcal{T}_t(\mathbf{p}_0))$ represent the same anatomical structure. The transformation can be described by a dense displacement field (DDF), denoted by \mathbf{u}_t where $\mathbf{u}_t(\mathbf{p}_0) = \mathbf{p}_t - \mathbf{p}_0$.

Deep learning has been used to perform the registration with one-step prediction by modelling a complex function $f_\theta(I_0, I_t) = \mathbf{u}_t$ that maps a pair of images to the optimal displacement field using convolutional neural network (ConvNet), where θ is the parameters of the network. The parameters θ in the registration network can be trained using two different supervision signals: ground truth DDF \mathbf{u}_{GT} (*supervised*), or the similarity between the pairs of images after registration (*unsupervised*).

3 Method

This paper adapts and compares two training strategies, *supervised* and *unsupervised*, for a deep learning based cardiac motion estimation in cine MR image sequences. The registration networks and the training strategies were set up in a comparable manner for a fair comparison. An overview of both the *supervised* and *unsupervised* registration frameworks is illustrated in Fig 1.

3.1 Supervised Training

Ground truth deformation The ground truth deformation is required for *supervised* training of the registration network. Existing deep learning methods for

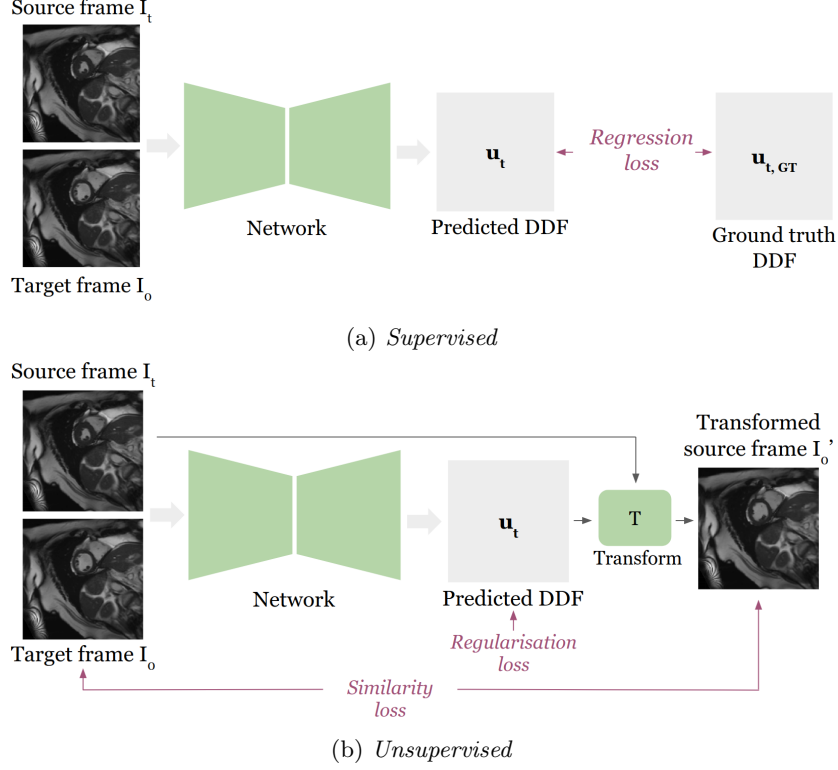


Fig. 1: *Supervised* and *unsupervised* registration framework

deformable registration usually generate the ground truth displacement \mathbf{u}_{GT} using traditional registration methods [2], and use the original image pair $\{(I_0, I_t)\}$ as input to the network. The network sees the real image pairs during training but the ground truth deformation does not completely capture the transformation due to residual errors from the traditional registration methods used to estimate the ground truth deformation. Alternatively, the deformation field acquired from traditional registration can be used to deform image I_t to generate a pseudo-target image $I'_0 = I_t \circ \mathcal{T}_{\mathbf{u}_t}$. We then use the image pairs $\{(I'_0, I_t)\}$ as input to the network. The ground truth in this setting fully captures the deformation between the input image pair (I'_0, I_t) and thus is not limited by residual registration errors. These two variants of supervised training are compared in Section 4.2. B-spline FFD [7] is used for traditional registration.

Training As shown in Figure 1(a), the network predicts the DDF \mathbf{u}_t from each pair of input images. For cardiac motion estimation, a sequence of image pairs $\{(I_0, I_t)\}_{t=1,2,3,\dots,N_T}$ is given as input to the network in one batch such that each training iteration optimises the group registration of the sequence [6,8].

The end-diastolic (ED) frame is used as the first frame (or the *target* frame) and is repeated in each pair in the batch. To train the model, we use Mean Square Error (MSE) between the predicted and ground truth DDF as the regression loss:

$$\mathcal{L}_{supervised} = \frac{1}{N_T} \sum_{t=1}^{N_T} \left(\frac{1}{|\Omega|} \sum_{\mathbf{p} \in \Omega} (\mathbf{u}_t(\mathbf{p}) - \mathbf{u}_{GT}(\mathbf{p}))^2 \right) \quad (1)$$

where N_T is the number of frames in one batch/sequence and Ω is the spatial domain of the images.

3.2 Unsupervised Training

As illustrated in Figure 1(b), we use image intensity-based similarity as loss function with an additional regularisation on the predicted displacements. The loss function that the training minimises at each iteration is:

$$\mathcal{L} = \mathcal{L}_{MSE} + \lambda \mathcal{L}_{smooth} \quad (2)$$

The first term in the loss function measures the pixel-wise difference between the target image and the registered source image:

$$\mathcal{L}_{MSE} = \frac{1}{N_T} \sum_{t=1}^{N_T} \left(\frac{1}{|\Omega|} \sum_{\mathbf{p} \in \Omega} (I'_0(\mathbf{p}) - I_0(\mathbf{p}))^2 \right) \quad (3)$$

Here I_t is transformed to I'_0 using differentiable bi-linear sampling in the spatial transformation network [4], enabling backpropagation for training. The second term in the loss function encourages spatially smooth deformation by minimising the variation of displacements using approximated Huber loss [6] on first-order spatial derivatives of \mathbf{u}_t ,

$$\mathcal{L}_{smooth} = \frac{1}{N_T} \sum_{t=1}^{N_T} \left(\frac{1}{|\Omega|} \sum_{\mathbf{p} \in \Omega} \sqrt{\left| \frac{\partial \mathbf{u}_t(\mathbf{p})}{\partial x} \right|^2 + \left| \frac{\partial \mathbf{u}_t(\mathbf{p})}{\partial y} \right|^2} \right) \quad (4)$$

Similar to the *supervised* training, one sequence of image pairs from one cardiac sequence is used in each input batch. The weight λ of the smoothness regularisation loss is set to 10^{-4} which is selected based on the performance on the validation dataset.

3.3 Network Architecture

A schematic of the network is shown in Figure 2. The same network architecture is used in both training strategies and is adapted from the motion estimation branch of the joint segmentation and motion estimation framework proposed in [6]. The network employs two encoder branches with 3×3 convolutional

kernels to extract features from the images. A stride of 2 is used every two convolutional layers to reduce the resolution of feature maps by 2 and increase the size of receptive field [1]. The features from all levels of the two encoders are concatenated before a convolution layer and upsampling to full resolution. Further convolutional layers are applied to fuse information from different scales before making the final prediction.

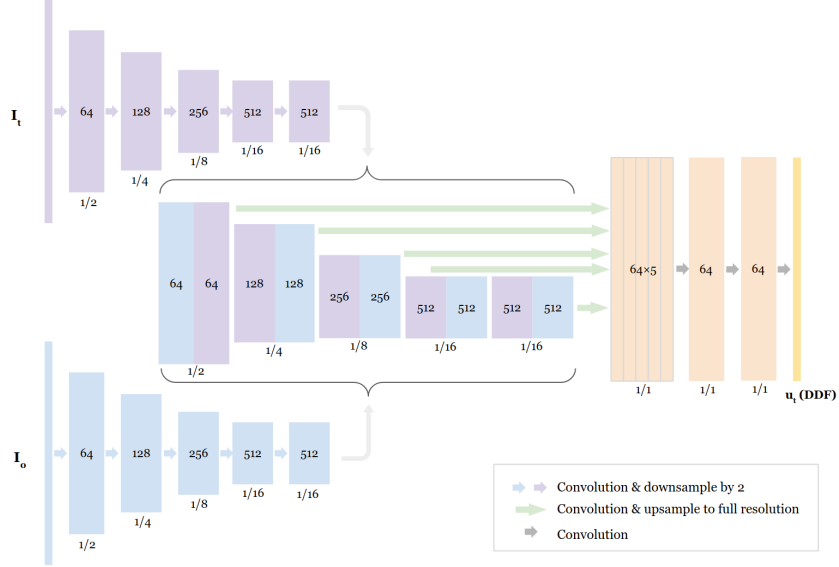


Fig. 2: Architecture of the registration network. The coloured blocks represent images or feature maps with the number of channels written inside. The resolution of the feature maps with respect to the input is written underneath the blocks. The final output has 2 channels encoding the displacement in 2 directions.

4 Experiments

4.1 Set up

Data The two training strategies are evaluated using short-axis view cardiac MR images of healthy subjects from the UK Biobank study¹. Randomly selected image sequences of 120 subjects were used for training and validation with another 100 subjects used for testing. Each sequence contains temporally pre-aligned 2D stacks of images of 50 consecutive time points in a complete cardiac cycle. In-plane resolution of the images is $1.8 \times 1.8mm$ per pixel while through-plane

¹ UK Biobank Imaging Study. <http://imaging.ukbiobank.ac.uk>

resolution is $10mm$ per pixel. The low resolution between planes could lead to physically implausible displacements of anatomical structure in 3D registration, which is the reason that our motion estimation is performed in 2D plane. The segmentation of the left-ventricular cavity (LV), myocardium wall (MYO) and right-ventricular cavity (RV) on the ED frame and the end-systolic (ES) frame is used to evaluate the accuracy of the estimated motion.

Metrics The estimated cardiac motion is evaluated on both accuracy and smoothness. To evaluate the accuracy, we first estimate the motion between the ED frame and the ES frame. Then we apply the estimated motion to deform the segmentation mask of the ES frame towards the ED frame, and measure its overlap with the ground true ED frame segmentation using the Dice score and Hausdorff Distance (HD). HD is measured on the outer contours of the anatomical structures. To evaluate regularity of deformation, we calculate the determinant of the Jacobian matrix $J_\phi(\mathbf{p}) = \nabla\phi(\mathbf{p})$, or simply *the Jacobian*, where ϕ denotes the transformation. We compute the percentage of points that exhibit non-diffeomorphic deformation, indicated by $|J_\phi(\mathbf{p})| \leq 0$. We also calculate magnitudes of the gradient of *the Jacobian*, i.e. $|\nabla|J_\phi||$ which is a second-order metric measuring the spatial smoothness of deformation [5].

Comparison To ensure fairness of the comparison, the *supervised* and *unsupervised* model use exactly the same network architecture described in Section 3.3. Both models are trained using the same amount of data for the same number of iterations and tested on data of the same testing subjects. As a reference of performance, the traditional B-spline FFD registration algorithm is also evaluated on the same testing data. The FFD algorithm is set to use the sum squared difference (SSD) as dissimilarity measure and Bending Energy (BE) as regularisation[7]. A 3-level hierarchical multi-resolution approach is used where the spacing of B-spline control points on the highest resolution is set to $8mm$. The same setting of FFD was used to generate the ground truth deformation for *supervised* training. The regularisation weights in the unsupervised method and FFD introduce a trade-off between accuracy and deformation regularity, making the selection of these parameters for fair comparison non-trivial. In this paper, both regularisation weights were selected to maximise the accuracy performance on the validation dataset.

Implementation details Input images are pre-processed by cropping to the size of 160×160 so that the registration is focused on the region of interest. The intensity value of input images is normalised to $[0, 1]$. The deep learning registration networks were implemented in Pytorch and trained for 500 epochs on NVIDIA® GeForce® Titan Xp GPUs. The B-spline FFD registration was performed using the implementation in MIRTk². The runtime of FFD is measured on an Intel® Core™ i7-8700 CPU.

² <https://mirtk.github.io/>

4.2 Results

Table 1 shows the results of the accuracy and regularity of different methods. When comparing the results of different methods, the Wilcoxon signed-rank test is performed to assess the statistical significance. It can be observed that the *unsupervised* training outperforms ($p \ll 0.001$) *supervised* training in terms of accuracy especially on left ventricle and right ventricle, and on-par with B-spline FFD on most metrics. Between *supervised* models, the one trained using the $\{(I'_0, I_t)\}$ image pair (“sup+warp.”) performs similar to the one trained using the $\{(I_0, I_t)\}$ image pair (“sup+orig.”) except better on myocardium measurements. In terms of regularity, the supervised methods produce deformations that are spatially smoother (lower $|\nabla|J_\phi||$) and significantly less topology-altering (lower $\%|J| \leq 0$).

Table 1: Accuracy and regularity of cardiac motion estimated by different methods. The accuracy metrics are also evaluated on unregistered input images (“Unreg”) as a reference. The mean and standard deviation over 100 testing subjects are presented. The best results with statistical significant advantage ($p \ll 0.001$) are highlighted in bold.

Method	Dice			HD			$ \nabla J_\phi $	$\% J \leq 0$
	LV	Myo	RV	LV	Myo	RV		
Unreg	0.641(0.058)	0.322(0.086)	0.551(0.077)	11.40(1.40)	8.90(2.00)	11.50(1.90)	-	-
FFD	0.941(0.049)	0.754(0.084)	0.671(0.109)	4.52(2.33)	4.73(1.44)	8.93(2.16)	0.021(0.023)	0.081(0.119)
DL(unsup)	0.943(0.046)	0.740(0.077)	0.709(0.087)	4.05(1.47)	4.62(1.25)	9.34(2.13)	0.047(0.014)	0.375(0.162)
DL(sup+warp.)	0.920(0.049)	0.735(0.080)	0.668(0.101)	4.61(1.11)	4.84(1.48)	9.06(1.91)	0.040(0.007)	0.025(0.038)
DL(sup+orig.)	0.926(0.048)	0.702(0.0801)	0.657(0.089)	4.41(1.35)	5.29(1.22)	9.22(1.88)	0.019(0.004)	0.030(0.040)

Figure 3 visually demonstrates the difference amongst different motion estimation methods on one exemplar subject. The deep learning model trained using *supervised* strategy performs inferior to its *unsupervised* counterpart. It can be observed, from the ED frame image reconstructed by deforming the ES frame image, that the supervised method significantly underestimates the deformation and produces some artefacts in the middle of the LV blood pool. The unsupervised method captures the deformation better but violates some topological structure especially around epicardial contour, as illustrated by the folding that can be observed on the deformed meshgrid.

Runtime advantage Despite not achieving significant performance advantage over the traditional method, the deep learning models are able to register a sequence of 50 2D frames in 80.05 milliseconds whereas FFD takes 23.18 seconds. The runtimes are measured and averaged over 100 test subjects.

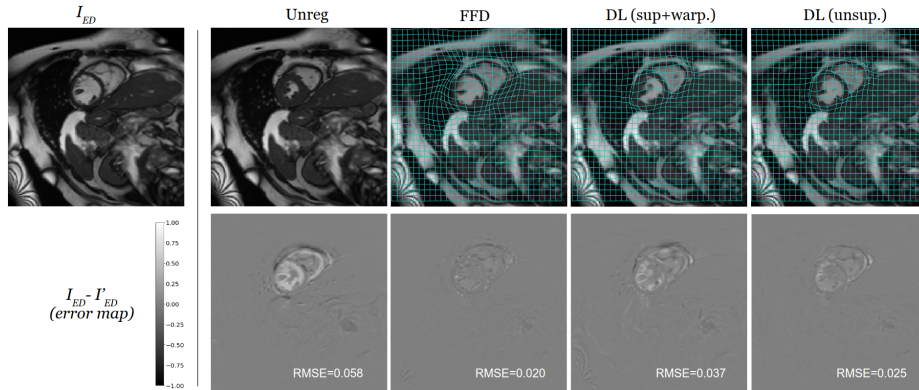


Fig. 3: Visualisation of motion estimation results. The target ED frame image I_{ED} is shown on the top left. The rest of the first row shows the ED frame reconstructed by deforming the ES frame image using deformation estimated by the various methods (I'_{ED}), overlaid by the meshgrid deformed using the same estimated deformation. The second row shows the error maps (with RMSE values) between the reconstructed image and the ED frame.

5 Conclusion and Discussion

In this work, we evaluated and compared the effect of different training strategy on the performance of deep learning registration network on the task of cardiac motion estimation. In terms of accuracy, we found that *unsupervised* training, which uses only image similarity, outperforms the *supervised* training strategies. This could be attributed to the fact that the unsupervised learning optimise directly on the image intensity difference, while the *supervised* training is either restricted by the registration error from FFD (“sup+orig.”) or the difference between the testing target images and training target images (“sup+warp”). Although performing inferior on accuracy, the *supervised* methods produce spatially smoother and more topology-preserving deformation.

The superior regularity of the supervised methods could be a result of inheriting spatial smoothness property from the B-spline basis functions in FFD and further regularisation in the regression. It is also possible that the better regularity can only be achieved while under-estimating deformation. This will be further investigated in the future. Future studies should also include a supervised model trained using randomly generated or permuted deformation ground truth in the comparison. This will help to understand the need for realistic ground truth deformation for the supervised method. Another limitation of the paper is that only two representative *supervised* and *unsupervised* designs are experimented whereas a study of more existing methods under the same experimental setting would be able to draw a more general conclusion.

Acknowledgements. This work was supported by the EPSRC Programme Grant EP/P001009/1 and EP/R005982/1. The cardiac image dataset has been provided under UK Biobank Access Application 40119.

References

1. Balakrishnan, G., Zhao, A., Sabuncu, M.R., et al.: Voxelmorph: a learning framework for deformable medical image registration. *IEEE transactions on medical imaging* (2019)
2. Cao, X., Yang, J., Zhang, J., et al.: Deformable image registration based on similarity-steered cnn regression. In: MICCAI. pp. 300–308. Springer (2017)
3. Eppenhof, K.A., Lafarge, M.W., Moeskops, P., et al.: Deformable image registration using convolutional neural networks. In: *Medical Imaging 2018: Image Processing*. vol. 10574, p. 105740S (2018)
4. Jaderberg, M., Simonyan, K., Zisserman, A., et al.: Spatial transformer networks. In: *Advances in neural information processing systems*. pp. 2017–2025 (2015)
5. Krebs, J., e Delingette, H., Mailhé, B., et al.: Learning a probabilistic model for diffeomorphic registration. *IEEE transactions on medical imaging* (2019)
6. Qin, C., Bai, W., Schlemper, J., et al.: Joint learning of motion estimation and segmentation for cardiac mr image sequences. In: MICCAI. pp. 472–480. Springer (2018)
7. Rueckert, D., Sonoda, L.I., Hayes, C., et al.: Nonrigid registration using free-form deformations: application to breast mr images. *IEEE transactions on medical imaging* **18**(8), 712–721 (1999)
8. Shen, D., Sundar, H., Xue, Z., et al.: Consistent estimation of cardiac motions by 4d image registration. In: MICCAI. pp. 902–910. Springer (2005)
9. Shi, W., Zhuang, X., Wang, H., et al.: A comprehensive cardiac motion estimation framework using both untagged and 3-d tagged mr images based on nonrigid registration. *IEEE transactions on medical imaging* **31**(6), 1263–1275 (2012)
10. Sokooti, H., de Vos, B., Berendsen, F., et al.: Nonrigid image registration using multi-scale 3d convolutional neural networks. In: MICCAI. pp. 232–239. Springer (2017)
11. de Vos, B.D., Berendsen, F.F., Viergever, M.A., et al.: A deep learning framework for unsupervised affine and deformable image registration. *Medical image analysis* **52**, 128–143 (2019)
12. Yang, X., Kwitt, R., Styner, M., Niethammer, M.: Quicksilver: Fast predictive image registration—a deep learning approach. *NeuroImage* **158**, 378–396 (2017)