NeuroImage 279 (2023) 120303

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Affine image registration of arterial spin labeling MRI using deep learning networks

Zongpai Zhang [a], Huiyuan Yang [a], Yanchen Guo [a], Nicolas R. Bolo [b], Matcheri Keshavan [b], Eve DeRosa [c], Adam K. Anderson [c], David C. Alsop [d], Lijun Yin [a], Weiying Dai [a,*]

[a] Department of Computer Science, State University of New York at Binghamton, Binghamton, NY 13902, USA
[b] Department of Psychiatry, Beth Israel Deaconess Medical Center & Harvard Medical School, Boston, MA 02215, USA
[c] Department of Psychology, Cornell University, Ithaca, NY 14850, USA
[d] Department of Radiology, Beth Israel Deaconess Medical Center & Harvard Medical School, Boston, MA 02215, USA

ARTICLE INFO

Keywords:
Image registration
Affine registration
Arterial spin labeling
Deep learning

ABSTRACT

Convolutional neural networks (CNN) have demonstrated good accuracy and speed in spatially registering high signal-to-noise ratio (SNR) structural magnetic resonance imaging (sMRI) images. However, some functional magnetic resonance imaging (fMRI) images, e.g., those acquired from arterial spin labeling (ASL) perfusion fMRI, are of intrinsically low SNR and therefore the quality of registering ASL images using CNN is not clear. In this work, we aimed to explore the feasibility of a CNN-based affine registration network (ARN) for registration of low-SNR three-dimensional ASL perfusion image time series and compare its performance with that from the state-of-the-art statistical parametric mapping (SPM) algorithm. The six affine parameters were learned from the ARN using both simulated motion and real acquisitions from ASL perfusion fMRI data and the registered images were generated by applying the transformation derived from the affine parameters. The speed and registration accuracy were compared between ARN and SPM. Several independent datasets, including meditation study (10 subjects × 2), bipolar disorder study (26 controls, 19 bipolar disorder subjects), and aging study (27 young subjects, 33 older subjects), were used to validate the generality of the trained ARN model. The ARN method achieves superior image affine registration accuracy (total translation/total rotation errors of ARN vs. SPM: 1.17 mm/1.23° vs. 6.09 mm/12.90° for simulated images and reduced MSE/L1/DSSIM/Total errors of 18.07% / 19.02% / 0.04% / 29.59% for real ASL test images) and 4.4 times (ARN vs. SPM: 0.50 s vs. 2.21 s) faster speed compared to SPM. The trained ARN can be generalized to align ASL perfusion image time series acquired with different scanners, and from different image resolutions, and from healthy or diseased populations. The results demonstrated that our ARN markedly outperforms the iteration-based SPM both for simulated motion and real acquisitions in terms of registration accuracy, speed, and generalization.

1. Introduction

Arterial spin labeling (ASL) is a noninvasive magnetic resonance imaging (MRI) technique to measure cerebral blood flow (CBF) (Detre et al., 1992; Williams et al., 1992) with naturally existing arterial blood water as an endogenous tracer. ASL imaging acquires pairs of images: a labeled image and a control image. Labeled images are obtained by magnetically labeling arterial blood water with radiofrequency pulses from the MRI scanner, while control images are without labeling of blood water. Labeled and control images are acquired in a temporally interleaved fashion. Subtraction of labeled images from controls images is a relative measure of perfusion proportional to CBF (Detre et al., 1992; Williams et al., 1992). The ASL signal-to-noise ratio (SNR) is inherently low because the signal from labeled blood is only about 1% of the full tissue signal.

To improve SNR, a series of labeled-control image pairs are normally acquired. They are averaged to generate CBF maps or used to produce functional connectivity maps. However, subject head movements and physiological motions (such as cardiac pulsation and respiratory motion) can cause misalignment of ASL time series (Ye et al., 2000), which can severely affect the quality of further measurements. Therefore, accuracy of image realignment (motion correction across ASL time series)

* Corresponding author.
E-mail address: wdai@binghamton.edu (W. Dai).

https://doi.org/10.1016/j.neuroimage.2023.120303
Received 15 April 2023; Received in revised form 27 July 2023; Accepted 31 July 2023
Available online 1 August 2023

is crucial for robust CBF and functional connectivity measurements. For real time applications, the speed of image realignment across a time series of images is very important.

Many conventional image realignment algorithms, such as those from Statistical Parametric Mapping (SPM) (Friston et al., 1995), FMRIB's Linear Image Registration Tool (FLIRT) (Jenkinson et al., 2002; Jenkinson and Smith, 2001), and Automated Image Registration (AIR) (Woods et al., 1998; Woods et al., 1998), have been developed and applied to register ASL time series images. The image realignment algorithms search for an affine transformation with a set of parameters to optimize pixel correspondence between a pair of fixed (target) and moving (source) images by maximizing a similarity measure of spatial correspondence between images. Conventional ASL image realignment optimization algorithms are often inaccurate and computationally expensive because the ASL time series has limited SNR and these conventional algorithms are typically solved using iterative algorithms.

Recently, deep learning techniques have been used for 3D medical image realignments (or affine image registration) and achieved comparable performance to iterative algorithms when applied to anatomical images, such as chest CT images, cardiac cine MRI images, prostate ultrasound images, and T2-weighted MRI images (Balakrishnan et al., 2018; de Vos et al., 2019). The deep learning-based realignment methods are mostly supervised because they rely on the known ground-truth (or affine) transformation information (Liao et al., 2017; Miao et al., 2016; Chee and Wu, 2018; Hu et al., 2018). For instance, one obtained training examples using conventional image registration methods and used convolutional neural networks (CNN) and reinforcement learning to predict small steps towards optimal realignment via affine registration (Liao et al., 2017); another synthesized training examples by applying combinations of rotation, translation, and scaling (affine transformation parameters) to the moving image and trained the CNN to regress affine parameters hierarchically (Miao et al., 2016) or at the same time directly (Chee and Wu, 2018); or used training images with manually annotated anatomical labels and trained the CNN to predict displacement fields to align multiple labeled corresponding structures (Hu et al., 2018). Deep learning has been employed in supervised 3D affine multi-modal image registration in different ways: predict affine transformation parameters by adopting pretrained 2D VGG-19 networks for feature extraction and fully connected regression networks (Kori and Krishnamurthi, 2019), using a pretrained VGG-type CNN network, training it first on a large number of synthetic images, and then refining using a small number of real images (Zheng et al., 2017), training a model to predict the image of one modality from that of another modality of the same subject and registering the predicted image and fixed image from the same modality (Liu et al., 2019), using a fully connected network to determine the control points and then CNN for feature detection (Zou et al., 2019). However, conventional image alignment methods are typically unsupervised because ground-truth transformations are not available. Unsupervised medical image registration methods have been explored but they are mainly applied to deformable 3D medical image registration (Balakrishnan et al., 2018; de Vos et al., 2019). Although deep learning-based affine image registration is sometimes used as the first step, its efficacy is not evaluated as a separate metric. In addition, the performance of deep learning-based affine image registration has not been evaluated in ASL images, or low-SNR functional images in general, which is a stronger test of their potential utility. Here, we aim to assess the feasibility and efficacy of an unsupervised deep learning method for realigning ASL time series by simulating affine registration, training a model based on a real ASL dataset and investigating generalization of the trained model to different ASL datasets.

2. Methods

We modeled a CNN-based linear registration of 3D ASL difference images, which we named an affine registration network (ARN). The ARN has a pair of 3D moving and 3D fixed images as input and aims to output affine registration parameters between the two. It aims to minimize the difference between the fixed image and the moved image after applying the affine transformation to the moving image. The ARN contains several CNN layers with gradually reduced image resolutions to derive useful low-resolution features and two fully connected layers.

2.1. ARN architecture

The ARN is composed of an encoder CNN followed by a fully connected network, as shown in Fig. 1. The network has the 3D moving image and the fixed image (each as grayscale ASL difference image) as input by concatenating them as a two-channel image. In this study, the input size is 128×128×40×2. The encoder CNN contains one convolution layer and four CNN blocks. The convolution layer has a kernel size of 4 × 4 × 4 and stride size of 1 × 1 × 1. Each of the four CNN blocks consists of a Rectified Linear Unit (ReLU) activation layer, a 3D convolution layer, and a batch normalization layer. For the four CNN blocks, the stride sizes of the first three convolution layer are 2 × 2 × 2, while the last convolution layer is 2 × 2 × 1. Twenty channels are applied in each convolutional layer. Hierarchical features are captured by these CNN layers with different spatial resolutions. The output of the encoder CNN is flattened and is passed as input to one fully connected layer (linear layer and ReLU layer) with 250 nodes. Another fully connected layer outputs 6–12 nodes (linear layer only). We used 6 nodes in the study because the non-aligned MRI images (specifically ASL perfusion images here) are from subject motion, involving only translations (3 parameters) and rotations (3 parameters). The moved image is generated by applying the affine transformation (see Affine Transformation section) from the learned 6 parameters to the moving image and interpolating the 3D mesh grid via trilinear interpolation based on 8 neighboring voxels.

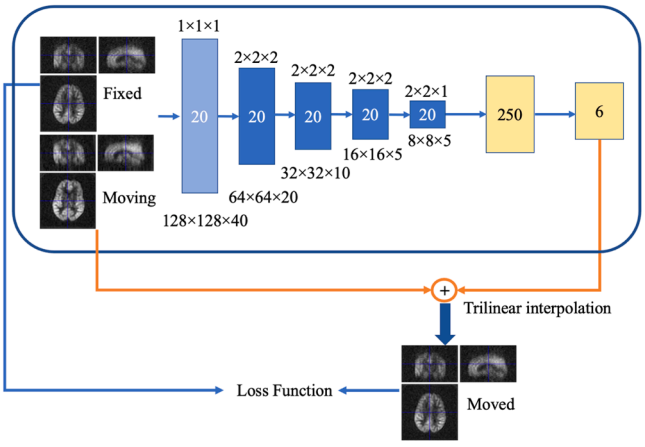

Fig. 1. ARN network architecture for affine image registration. The input layer has both the 3D fixed image and moving image (each as 128 × 128 × 40 matrix) by concatenating them as a two-channel image (input size as 128 × 128 × 40 × 2). The ARN network has a encoder CNN and a fully-connected neural network (FNN). The CNN contains one convolution layer and four CNN blocks. These convolution layers have kernel size of 4 × 4 × 4. Each of four CNN block consists of a Rectified Linear Unit (ReLU) activation layer, a 3D convolution layer, and a batch normalization layer. The output size and stride size of each CNN layer are listed on the bottom and top of the layer (shown as a rectangle). Twenty channels are applied in each convolutional layer. The output of the encoder CNN is flattened and is passed into FNN. The FNN contains two fully connected layers, in which they have 250 nodes (linear layer and ReLU layer) and 6 nodes (linear layer only), respectively.

2.2. Affine transformationThe application order of translation and rotation matrices affects values of the affine transformation matrix. To compare with the state-of-the-art linear registration method — SPM, we used the same application order of translation and rotation matrices. Let us call 6 parameters from the ARN output p(1), p(2), …, p(6). Specifically, the affine transformation matrix M is defined as the translation matrix T multiplied by rotation matrix R:

$$M = T^{*}R \quad (1)$$

$$T = \begin{bmatrix} 1 & 0 & 0 & p(4) \\ 0 & 1 & 0 & p(5) \\ 0 & 0 & 1 & p(6) \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (2)$$

$$R = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & cos(p(3)) & sin(p(3)) & 0 \\ 0 & -sin(p(3)) & cos(p(3)) & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} cos(p(2)) & 0 & sin(p(2)) & 0 \\ 0 & 1 & 0 & 0 \\ -sin(p(2)) & 0 & cos(p(2)) & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \cos(p(1)) & sin(p(1)) & 0 & 0 \\ -sin(p(1)) & cos(p(1)) & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (3)$$

2.3. Loss function

The ARN was trained in an unsupervised way by minimizing the loss between moved images and fixed images. The total loss was calculated from three loss functions between the fixed and moved image. The relative weights of the three loss functions were determined empirically. Their definitions are as follows.

2.4. MSE loss

Mean squared error measures the average squared difference between the predicted image (moved image) and the actual image (fixed image).

$$MSE\ loss = \frac{1}{n}\sum_{i=1}^{n}(F - M)^2 \quad (4)$$

where n is the total number of voxels of input images F and M, F is the fixed image and M is the moved image.

2.5. Pixel-wise L1 loss

The pixel-wise L_1 loss function describes the pixel level difference between the fixed image F and the moved image M.

$$L_1\ loss = \sum_{i=1}^{n}|F - M| \quad (5)$$

2.6. Pixel-wise structural dissimilarity loss

The dissimilarity loss utilizes the structure similarity index (SSIM) (Woods et al., 1998) to calculate the dissimilarity of two volumes which will help to generate clearer motion boundaries.

$$DSSIM\ loss = \frac{1}{n}\sum_{i=1}^{n}(1 - SSIM(F,\ M)) \quad (6)$$

2.7. Total loss

The total loss is calculated as

$$L_{total} = MSE + \lambda_1 \cdot L_1 + \lambda_2 \cdot DSSIM \quad (7)$$

λ_1 and λ_2 are the loss weights which control the relative importance of L_1 loss and dissimilarity loss. We set the range of λ_1 and λ_2 from 0 to 2000 with an increment of 100 and compare the total loss on the test data for each pair of λ_1 and λ_2 after training has converged. The best results were achieved with $\lambda_1 = 100$ and $\lambda_2 = 1000$.

2.8. Experiments

2.8.1. Dataset

We evaluated the performance of the ARN method using a simulated ASL dataset and a real brain ASL dataset and tested the generality of the ARN method with several ASL datasets from our previous projects. All experiments were performed in Python using Tensor Flow on an Nvidia RTX 2080Ti GPU and an Intel® Core™ i7–8700 K 3.7 GHz CPU with 6 cores and 64GB of internal memory.

2.8.2. Real imaging data

Dynamic pseudo-continuous arterial spin labeling (PCASL) perfusion (Dai et al., 2008) MRI images, with a 2 s of labeling duration and 1.8 s of post-labeling delay, from 20 subjects (33.3 ± 4.6 yrs old, 8 females) were obtained from our previous study (Dai et al., 2016). Each subject was acquired with a time series of 39 3D PCASL perfusion images (temporal resolution of 30 s, total acquisition time of 20 min). All 3D perfusion images were acquired with a 3D stack of spirals rapid acquisition with refocused echoes (RARE) imaging sequence. Each 3D ASL control or label image was acquired with three interleaved/segmented spirals (in-plane spatial resolution of 3.64 mm). To reduce the effect of physiological noises and head motion, we applied heavy background suppression to suppress gray white matter, fat, and CSF signals to less than 0.3% of the fully relaxed signal by using the algorithm in Maleki et al. (2012) and Dai et al. (2011). We determined the pulse timings of background suppression pulses by minimizing the sum of squared differences between theoretical magnetization and target magnetization (specifically, zero for CSF and fat and 0.3% for gray matter and white matter). This heavy background suppression causes both control and label images to have relatively low signals. Therefore, we choose to register 3D ASL difference images, control images minus adjacent label images, (each with 6 TRs, a temporal resolution of 30 s), instead of registering control and label images separately. 3D ASL difference images were reconstructed and interpolated into a 128 × 128 matrix for each of 40 slices with a nominal spatial resolution of 1.88 × 1.88 × 4 mm^3. For each subject, any 3D ASL difference image was randomly chosen from 39 ASL time series as the fixed image, each of the other 38 ASL difference images (from other time points) was used as a moving image and formed 38 pairs of images together with the fixed image. We used 5-fold cross validation in order to evaluate the performance of the ARN model for unseen subjects. All 20 subjects were randomly divided into 5 folds, in which each fold has 4 subjects. For the ith ($1 \le i \le 5$) partition, the ith fold (4 subjects, 152 pairs of images) served as the test set and the remaining 4 folds served as the training set (16 subjects, 608 pairs of images).

2.8.3. Simulated motion

The ground-truth translation and rotation parameters are unknown for any pair of real MRI images. To quantify the accuracy of these six parameters derived from the ARN model, we simulated subjects' motion by applying random translations and rotations to the fixed image. We used the same above-mentioned ASL dataset (Dai et al., 2008) with 20 subjects and 39 3D ASL images from each subject. For each subject, 34 3D ASL images were randomly chosen from 39 time points as fixed images. Twenty subjects have 680 fixed images in total. For each fixed image, a moving image was generated by applying an affine transformation with 6 parameters (x, y, and z translations in a range of $[-2, 2]$ voxels, x, y, and z rotations in a range of $[-5°, 5°]$ with a uniform distribution for each of six parameters). From 680 simulated pairs of images, 510 pairs of images were randomly chosen as training data, and the other 170 pairs of images were used for testing.

2.8.4. Datasets for testing generalizability of the ARN model

Three datasets with ASL time series in our previous projects were used to test generalizability of the ARN model. The datasets were acquired to evaluate the changes of ASL functional connectivity before and after meditation (meditation dataset) (Zhang et al., 2021), the deficits of ASL functional connectivity and low frequency fluctuation in bipolar disorder compared to normal controls (bipolar disorder dataset) (Dai et al., 2020), the changes of perfusion and functional connectivity in older adults compared to young adults (aging dataset) (Li et al., 2020; Zhang et al., 2022). For the meditation dataset (Zhang et al., 2021), 10 subjects (19.20 ± 0.28 yrs old, 4 females) were assessed using GE 3T MR750 scanner at the Cornell University MRI facility before and after meditation training with a time series of 49 3D PCASL perfusion images (temporal resolution of 20 s, total acquisition time of 17 min), in which each 3D image was a $128 \times 128 \times 40$ matrix with an interpolated spatial resolution of $1.88 \times 1.88 \times 4$ mm^3. For the bipolar disorder dataset (Dai et al., 2020), 19 bipolar disorder subjects (33.53 ± 9.60 yrs old, 9 females) and 26 normal controls (31.23 ± 10.80 yrs old, 14 females) were assessed using GE Signa HDxt scanner at the MRI research center of Beth Israel Deaconess Medical Center with a time series of 26 3D PCASL perfusion images (temporal resolution of 20 s, total acquisition time of 9 min), in which each 3D image was with the same interpolated spatial resolution as the meditation dataset. For the aging dataset (Zhang et al., 2022), 33 young adults (30.82 ± 11.56 yrs old, 23 females) and 27 older adults (68.63 ± 4.84 yrs old, 15 females) were assessed using GE 3T MR750 scanner at the Cornell University MRI facility with a time series of 29 3D PCASL perfusion images (temporal resolution of 20 s, total acquisition time of 10 min), in which each 3D image was with the same interpolated spatial resolution as the meditation dataset. No training was performed on the ARN model. Using the trained ARN model, 600 pairs of moving and fixed images (approximately same number of pairs randomly chosen from each subject) for each group of every dataset were used to test the performance of the ARN model.

2.8.5. Comparison with SPM affine registration

Statistical parametric mapping (SPM12) is one of the most popular neuroimage registration tools (Friston et al., 1995). Realignment is the affine image registration method in SPM based on iterative algorithms. It was used to compare with the performance of the ARN method in terms of accuracy and speed. To evaluate whether the accuracy of the ARN method is dependent on the degree of head motion, we calculated the relative reduction of total loss from ARN with respect to SPM across three ranges of head motion. Because no ground truth of head motion in real datasets and more accuracy in the ARN method, we use six motion parameters from the ARN method to estimate the degree of head motion. Specifically, framewise displacement (FD) was calculated from these six parameters as an index to represent head motion (Power et al., 2012). Based on FD values from the subgroups of each dataset, we divided small, medium, and large head motions with equal number of FD values.

3. Results

3.1. Performance of ARN and SPM on simulated images

3.1.1. Comparison of loss between ARN and SPM

For the simulated motion, the MSE loss, L1 loss, and DSSIM loss between the moved images and the fixed images from the SPM affine registration and the proposed ARN are shown in Fig. 2a–c. Specifically,

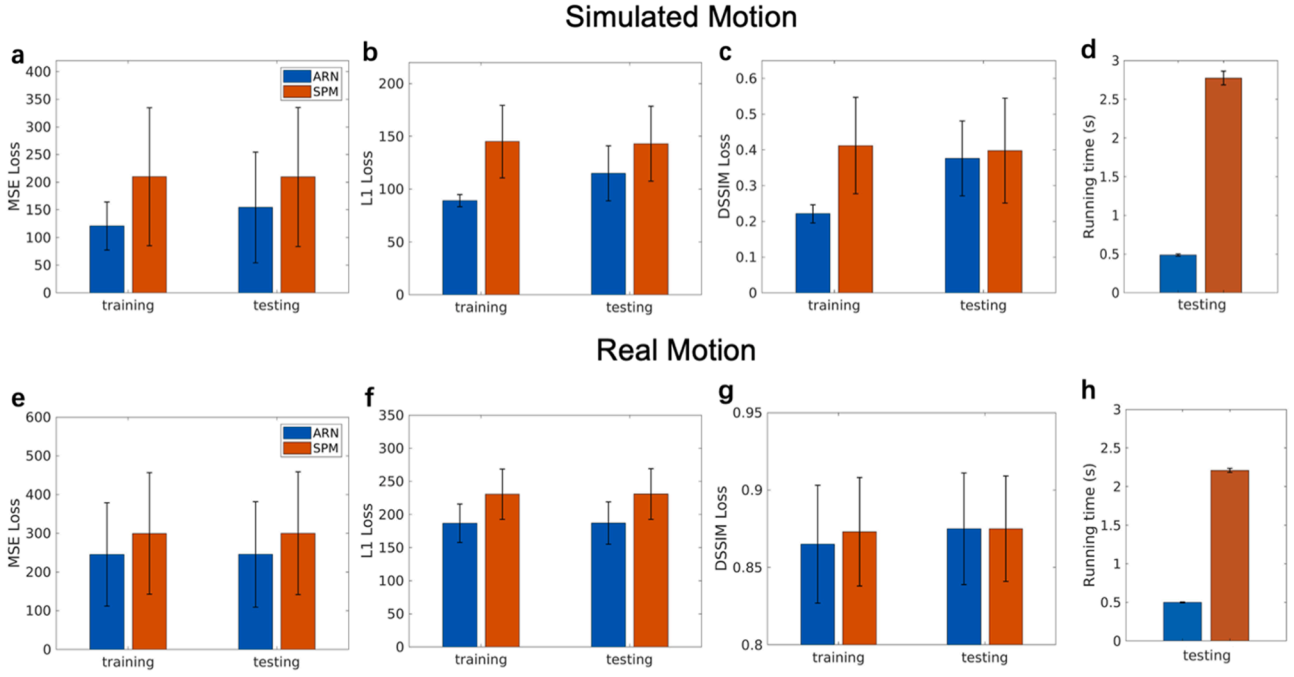

Fig. 2. Comparisons of three types of loss from ARN and SPM affine registration for simulated motion and real 3D perfusion imaging acquisition. (a) MSE loss (shown as its square root), (b) pixel-wise L1 loss, (c) DSSIM loss, and (d) running time using simulated 3D perfusion data; and (e) MSE loss (shown as its square root), (f) pixel-wise L1 loss, and (g) DSSIM loss, (h) running time using real perfusion data were compared between ARN and SPM.

Table 1
Absolute errors of six affine parameters derived from ARN and SPM compared to the corresponding ground truth parameters for simulated data.

	Parameter errors from training		Parameter errors from testing	
	ARN	SPM	ARN	SPM
x-translation	0.51±0.37	3.35±2.41	0.50±0.36	2.89±2.23
y-translation	0.43±0.37	3.42±2.31	0.46±0.37	2.97±2.21
z-translation	0.17±0.14	0.22±0.18	0.21±0.14	0.23±0.17
total-translation	1.11±0.88	6.99±4.9	1.17±0.87	6.09±4.61
x-rotation	0.18±0.12	3.78±2.58	0.49±0.36	3.58±2.56
y-rotation	0.17±0.14	3.94±2.67	0.39±0.30	3.75±2.67
z-rotation	0.12±0.08	5.40±3.09	0.35±0.31	5.57±3.13
total-rotation	0.47±0.34	13.12±8.34	1.23±0.97	12.90±8.36

*The absolute error of each parameter is defined as the absolute value of the difference between the derived parameter and ground-truth parameter. The translation parameters are in millimeters and the rotation parameters are in degrees.

the MSE loss, L1 loss, DSSIM loss, and total loss from the ARN affine registration was 28.03 ± 10.90%, 19.77 ± 18.17%, 6.20 ± 61.94%, 39.74 ± 17.13% less than the corresponding loss from the SPM affine registration using the test data, which are significantly smaller ($p < 10^{-6}$, $p < 10^{-6}$, $p = 0.0015$, $p < 10^{-6}$), respectively. These results show that ARN can provide more accurate affine image registration on simulated imaging data.

3.1.2. Comparison of the errors in affine parameters derived from ARN and SPM

The predicted 6 affine parameters generated from the ARN and SPM are compared with the parameters that are used to generate moving images from the ground-truth parameters. Compared with the ground-truth parameters, the errors of the 6 parameters generated from the ARN are significantly smaller than those from the SPM ($p < 10^{-4}$) with a mean error of 9.5 times smaller in the test data (Table 1). This implies that ARN can predict motion parameters much closer to the ground truth.

3.1.3. Comparison of registration speeds between ARN and SPM

The running time for registering a pair of test images from ARN and SPM is also listed in Fig. 2d. The average running time of ARN is 5.7 times less than that of SPM. It takes ARN less than half a second to register a pair of unseen images. Although the training time of ARN is about 11 h on the current computing environment, it can be performed prior to the registration task.

3.2. Performance of ARN and SPM on real ASL perfusion MRI images

For the real ASL perfusion MRI imaging acquisition, the MSE loss, L1 loss, and DSSIM loss between the moved images and the fixed images from the SPM affine registration and the proposed ARN are also shown in Fig. 2e–g. Specifically, the MSE loss, L1 loss, and DSSIM loss from the ARN affine registration were 18.07 ± 2.48%, 19.02 ± 1.77%, 0.04 ± 1.58% less than the corresponding loss from the SPM affine registration using the test data, which are significantly smaller ($p < 10^{-6}$, $p < 10^{-6}$) for the MSE loss and L1 loss, respectively. This demonstrates that ARN can provide more accurate affine image registration on real ASL imaging acquisition, which is consistent with its performance on simulated motion.

The running time for registering a pair of real ASL perfusion MRI images on the test data from ARN and SPM is also listed in Fig. 2h. On average, ARN achieves 4.4 times faster speedup compared to SPM. It takes ARN half a second to register a pair of unseen images. Comparison of ARN and SPM affine registration for an example image pair is shown in Fig. 3.

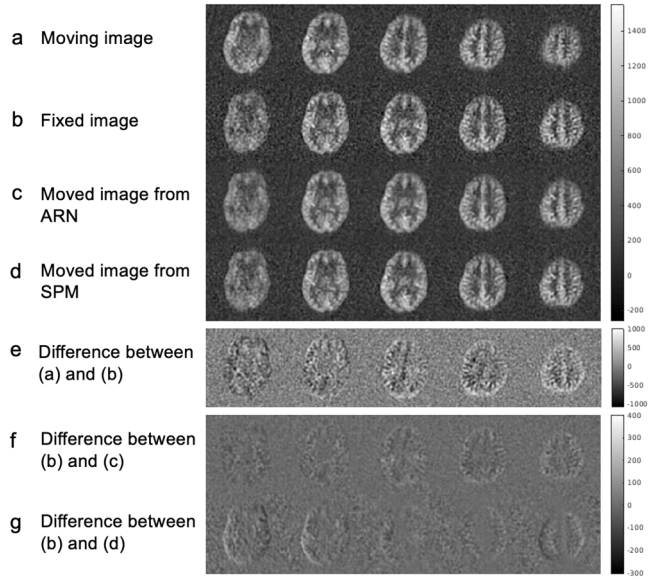

Fig. 3. Representative real ASL images from inferior to superior slices (column 1 to 5) from ARN and SPM affine registration. (a) moving (source) image, (b) fixed (target) image, (c) moved (registered) image from ARN, (d) moved (registered) image from SPM, (e) difference image between fixed and moving image, (f) difference between fixed image and moved image from ARN, and (g) difference between fixed image and moved image from SPM. (g) and (f) were smoothed with 4 mm full width half maximum (FWHM) Gaussian kernel for clearer comparison. The scale bars are shown on the right.

3.3. Performance of ARN and SPM on three ASL MRI validation datasets

The three ASL validation datasets further corroborate markedly improved accuracy of the ARN model in affine ASL image registration compared to the current state-of-the-art SPM method. The MSE, L1, DSSIM, and total losses between moved and fixed images from both SPM and ARN and the statistical differences between these two methods are shown in Table 2. For the meditation dataset ($n = 10$ with measurements twice), the MSE loss, L1 loss, DSSIM loss, and total loss from ARN are 16.58 ± 3.62%, 15.16 ± 2.21%, 1.23 ± 1.31%, 23.44 ± 4.49% less compared to SPM. For the bipolar disorder dataset ($n = 45$), the MSE, L1, DSSIM, and total losses from ARN are 13.61 ± 1.91%, 14.58 ± 1.46%, 0.33 ± 0.71%, 21.15 ± 2.55% less in the normal control group ($n = 26$) and 14.46 ± 1.98%, 15.25 ± 1.44%, 0.38 ± 0.68%, 22.32 ± 2.62% less in the bipolar disorder group ($n = 19$) compared to SPM, respectively. For the aging dataset ($n = 60$), the MSE loss, L1 loss, DSSIM loss, and total loss from ARN are 28.81 ± 8.27%, 20.34 ± 4.33%, 0.32 ± 1.00, 35.77 ± 7.99% less in the young age group ($n = 33$) and 34.79 ± 10.46%, 24.13 ± 5.45%, 0.11 ± 1.12%, 40.39 ± 9.51% less in the older age group ($n = 27$) compared to SPM, respectively. For all the sub-groups, ARN exhibited significantly smaller MSE loss, L1 loss, and total loss compared to SPM (Table 2). Significantly smaller DSSIM loss was only observed in the meditation dataset. Relative reduction of total loss in ARN (to SPM) in the older age group is significantly higher than that in the young age group ($p < 0.0001$) and relative reduction in the bipolar disorder patient group is slightly higher than that in the control group ($p < 0.0001$). The running time of ARN for the affine image registration of a pair of ASL images is similar to that reported in Fig. 2 for three validation datasets (not shown).

These results demonstrate that the markedly improved performance of the ARN method relative to SPM can be generalized to different ASL image datasets, which were acquired with different scanners, image resolutions, healthy aging and diseased populations and different magnitudes of movement and realignment.

The summary of six motion parameters (three translation and three

Table 2
Affine registration comparisons from ARN and SPM for the four loss values between the fixed and moved images on the test data of the model dataset and three ASL validation datasets.

Datasets (subgroup)	Performance	Fixed and Moved from ARN	Fixed and Moved from SPM	P values
Model	$\sqrt{MSE}$	242.80±43.09	296.04±49.70	< 0.0001
	L1	187.07±31.91	230.89±38.35	< 0.0001
	DSSIM	0.87±0.033	0.87±0.034	0.86
	Total loss (× 10^4)	8.03±2.17	11.41±2.89	< 0.0001
Meditation	$\sqrt{MSE}$	99.12±66.54	118.79±78.70	< 0.0001
	L1	70.98±13.22	83.61±15.03	< 0.0001
	DSSIM	0.83±0.025	0.84±0.030	< 0.0001
	Total loss (× 10^4)	1.78±0.57	2.33±0.77	< 0.0001
Bipolar disorder (control group)	$\sqrt{MSE}$	135.96±67.30	157.15±75.55	< 0.0001
	L1	104.05±13.89	121.75±15.77	< 0.0001
	DSSIM	0.89±0.036	0.90±0.036	0.23
	Total loss (× 10^4)	2.98±0.59	3.78±0.73	< 0.0001
Bipolar disorder (patient group)	$\sqrt{MSE}$	136.70±69.80	159.51±78.65	< 0.0001
	L1	104.82±15.20	123.62±17.52	< 0.0001
	DSSIM	0.90±0.040	0.91±0.042	0.23
	Total loss (× 10^4)	3.08±0.64	3.87±0.79	< 0.0001
Aging (young age group)	$\sqrt{MSE}$	84.30±74.08	114.88±88.79	< 0.0001
	L1	58.71±17.69	73.36±20.70	< 0.0001
	DSSIM	0.83±0.037	0.83±0.040	0.34
	Total loss (× 10^4)	1.38±0.71	2.14±0.99	< 0.0001
Aging (older age group)	$\sqrt{MSE}$	82.58±104.46	114.13±124.67	< 0.0001
	L1	49.79±23.79	64.91±28.40	< 0.0001
	DSSIM	0.89±0.039	0.89±0.041	0.71
	Total loss(× 10^4)	1.27±1.31	2.04±1.82	< 0.0001

*Total loss = MSE + 100 L1+1000 DSSIM.

rotation parameters) derived from ARN for the subgroups of each dataset is shown in Table 3. The demographic information for each subgroup is also shown in the Table. Significantly larger head motion ($p < 10^{-4}$) was observed on the meditation dataset, which may be caused by longer duration compared to the other two datasets. We also observed significantly larger head motion in older adults compared to young adults ($p < 10^{-3}$) and slightly larger head motion in bipolar disorder patients compared to controls although it was not statistically significant ($p = 0.12$). FD values from ARN and SPM showed significant correlation ($p < 10^{-6}$). The accuracy of ARN relative to SPM for small, medium, and large range of head motions is shown in Fig. 4. In general, the improvement of ARN accuracy from SPM is smaller for large head motion compared to small motion. However, the larger improvement of ARN accuracy relative to SPM in older adults (vs. young adults) benefits from the motion range of the trained ARN model (FD of 4.05 ± 1.91 mm) closer to that of the older age group (FD of 4.13 ± 1.35 mm) than of the young age group (FD of 3.86 ± 0.80 mm) despite degraded performance of ARN from the large range of head motion (Fig. 4e). The slightly larger improvement of ARN accuracy relative to SPM in bipolar disorder patients (vs. controls) also benefits from closer motion range with the trained ARN model.

4. Discussion

We have presented a CNN-based framework, ARN, for unsupervised learning of 3D affine ASL image registration. The ARN exploits a combined loss function based on three types of loss functions (MSE loss, L1 loss, and DSSIM loss) between fixed and moved image pairs to derive six affine parameters. In this case, six parameters that are required for supervised training, are not needed for unsupervised training. Also, in the real ASL MRI image registration, six parameters are not available for any pairs of acquired images. The proposed ARN can achieve more accurate affine image registration results for real ASL MRI datasets. We have also demonstrated that the improved affine image registration accuracy from

Table 3
The demographic information, absolute value of six affine parameters derived from ARN, and framewise displacement for different validation datasets and the dataset for building the ARN model. The translation parameters are in millimeters and the rotation parameters are in degrees.

	Parameters from ARN					
	Meditation	BD* control	BD patient	Young	Older	ARN Model
# of subjects	$10 \times 2^{\perp}$	26	19	33	27	20
Female/Male	4/6	14/12	9/10	23/10	15/12	8/12
Age	19.20±0.28 yrs	31.23±10.89 yrs	33.53±9.68 yrs	30.82±11.56 yrs	68.63±4.84 yrs	33.3 ± 4.6 yrs
x-translation	0.46±0.15	0.22±0.12	0.23±0.11	0.29±0.14	0.35±0.27	0.44±0.34
y-translation	0.80±0.29	0.39±0.22	0.41±0.20	0.56±0.29	0.67±0.54	0.51±0.35
z-translation	2.62±0.24	2.26±0.19	2.27±0.17	2.42±0.25	2.51±0.44	2.68±1.61
x-rotation	0.13±0.05	0.06±0.04	0.07±0.03	0.09±0.05	0.11±0.09	0.11±0.12
y-rotation	0.23±0.06	0.15±0.04	0.16±0.04	0.19±0.06	0.21±0.10	0.10±0.10
z-rotation	0.40±0.05	0.47±0.05	0.47±0.03	0.39±0.08	0.38±0.09	0.27±0.21
FD^{+}	4.54±0.73	3.46±0.56	3.52±0.52	3.86±0.80	4.13±1.35	4.05±1.91

* BD stands for bipolar disorder. $^{\perp}\times$ 2 means that each subject had two measurements (before and after 2-month meditation practice). $^{+}$FD represents framewise displacement in millimeters. BOLD fonts represent significance differences in FD. FD values in the Meditation participants are significantly larger than those in other groups. FD values in the older adults are significantly larger than young adults. FD values in the BD patients are slightly larger than BD controls.

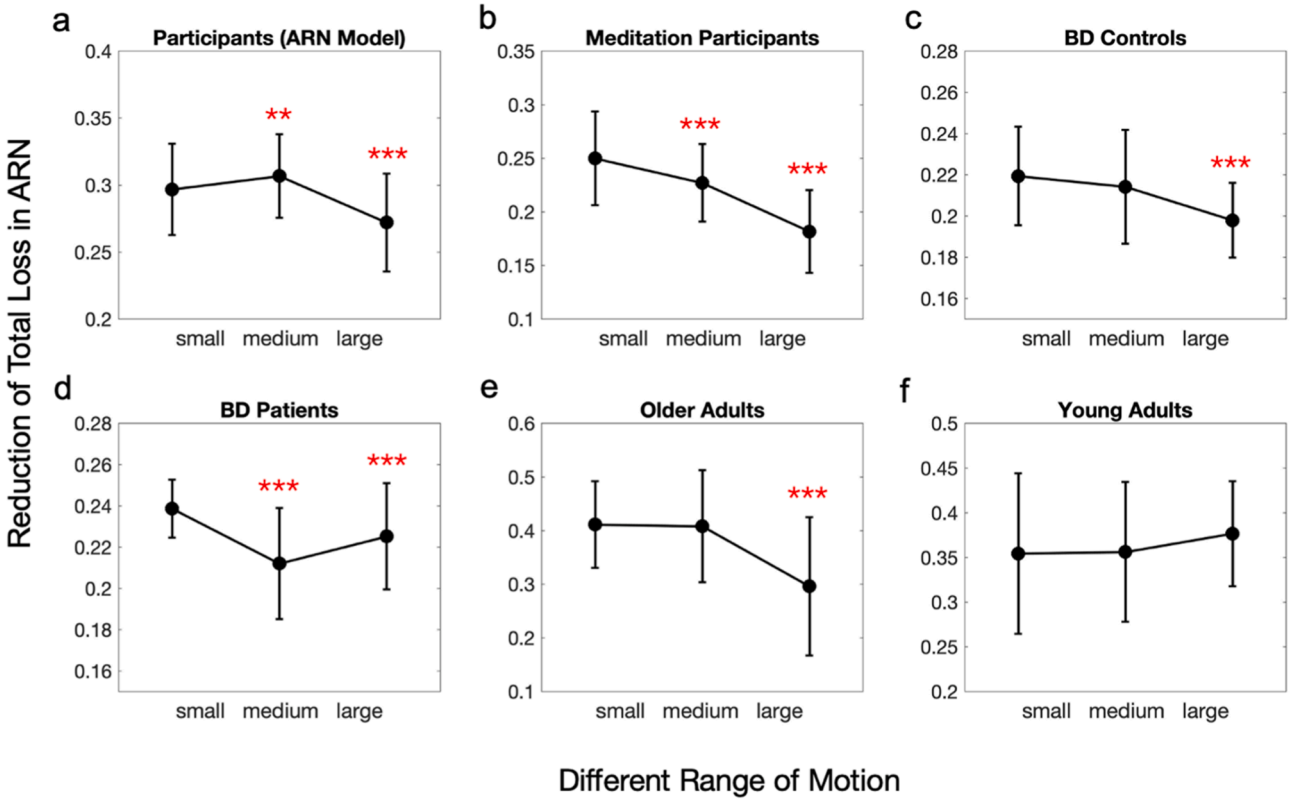

Fig. 4. Reduction of total loss in ARN relative to SPM for small, medium, and large head motions from all subgroups of real datasets: (a) participants from all 5 test folds, (b) meditation participants, (c) BD controls, (d) BD patients, (e) older adults, and (f) young adults. * stands for $0.01 < P < 0.05$, ** stands for $0.001 < P < 0.01$, *** stands for $P < 0.001$, compared to the small range of motion. BD represents bipolar disorder.

ARN can be generalized (without need of further training) to ASL datasets with different image resolutions, from different MRI scanners, from different age groups, with different degrees of head motion, and with patient and healthy control groups.

We also used simulated 3D ASL imaging data to verify the source of more accurate image registration results from ARN because the ground-truth 6-parameter affine registration parameters are known in this case. Results from the simulated data demonstrated that the outperformed affine image registration results from ARN emerge from more accurate estimation of six affine parameters. We observed relatively less improvement of affine registration quality from ARN compared to SPM in real 3D ASL MRI images than in simulated ASL images. These are expected because many noise sources, such as MRI instrumental noises, physiological noises, and fluctuations from brain networks, appeared in the real MRI images inevitably. Not only marked improving accuracy, ARN also outperforms the registration speed by 4 to 5 times compared to SPM.

The ARN can be generalized to 12 parameter affine image registration. If an application has scaling and skew between a pair of images, 6 additional parameters (3 parameters for scaling on x, y, z directions and 3 skew parameters for skewing on xy, xz, and yz directions) can be added into the output of ARN (Fig. 1); the transformation matrix (Equation 1) can be revised by incorporating the scaling and skew matrix. We constrained the model to 6-parameter affine image registration in ASL image alignment because the brain images that are acquired from the MRI scanner at different time points primarily involve just translations and rotations.

This study is not without limitations. First, all the ASL imaging data were acquired with multiple-segment 3D imaging and heavy suppression of background tissue signals. For the ASL images without or with less background suppression, motion between label and control may be a bigger registration issue than that with our background suppressed time series. Second, all the ASL image time series did not have severely altered perfusion patterns, such as those from tumor or stroke imaging. Third, we did not test the longitudinal ASL image registration at different time points. Perfusion patterns, such as those during the stroke recovery, may have changed over time and the affected perfusion patterns would pose challenges for the ARN method. Therefore, studies on image registration of ASL image time series acquired with single-shot 2D or 3D image acquisitions and/or with less background suppression and ASL image time series/longitudinal images with abnormal perfusion patterns are warranted to validate the ARN method.

Conclusion

We present a deep learning-based ASL image registration method for unsupervised affine image registration, that requires no ground-truth transformation parameters and anatomical landmarks. The method achieves superior image affine registration accuracy and 4 to 5 times faster speed compared to the state-of-the-art 3D affine image registration.

Data and code availability statement

Raw data were generated from MRI scanner. Reconstruction software is vendor's proprietary product. Sharing of derived data will be supported by direct request to the PIs for different data sets. Before sharing data the PIs will make sure that all data are free of identifiers that could directly or indirectly link information to an individual or vulnerable group and that all sharing is compliant with institutional and IRB policies.

CRediT authorship contribution statement

Zongpai Zhang: Methodology, Formal analysis, Data curation, Software, Visualization, Writing – original draft. **Huiyuan Yang:** Data curation, Software, Writing – review & editing. **Yanchen Guo:** Data curation, Software. **Nicolas R. Bolo:** Resources, Writing – review & editing. **Matcheri Keshavan:** Resources, Writing – review & editing. **Eve DeRosa:** Resources, Writing – review & editing. **Adam K. Anderson:** Resources, Writing – review & editing. **David C. Alsop:** Resources, Writing – review & editing. **Lijun Yin:** Resources, Supervision, Writing – review & editing. **Weiying Dai:** Methodology, Conceptualization, Funding acquisition, Project administration, Supervision, Writing – original draft.

Declaration of Competing Interest

The authors have no conflict of interest to report.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported by National Institute on Aging R01AG066430, National Science Foundation CMMI-2123061, National Institute of Mental Health R21MH126260, and Transdisciplinary Areas of Excellence (TAE) seed grant number 1154428 at Binghamton University.

References

Balakrishnan, G., et al., 2018. An unsupervised learning model for deformable medical image registration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 9252–9260.
Chee E. and Z. Wu, 2018. AIRNet: self-supervised affine registration for 3D medical images using neural networks. arXiv:1810.02583.
Dai, W., et al., 2008. Continuous flow-driven inversion for arterial spin labeling using pulsed radio frequency and gradient fields. Magn. Reson. Med. 60 (6), 1488–1497.
Dai, W., et al., 2016. Quantifying fluctuations of resting state networks using arterial spin labeling perfusion MRI. J. Cereb. Blood Flow Metab. 36 (3), 463–473.
Dai, W., et al., 2020. Abnormal perfusion fluctuation and perfusion connectivity in bipolar disorder measured by dynamic arterial spin labeling. Bipolar Disord. 22 (4), 401–410.
Dai, W., et al., 2011. Sensitivity calibration with a uniform magnetization image to improve arterial spin labeling perfusion quantification. Magn. Reson. Med. 66 (6), 1590–1600.
de Vos, B.D., et al., 2019. A deep learning framework for unsupervised affine and deformable image registration. Med. Image Anal. 52, 128–143.
Detre, J.A., et al., 1992. Perfusion imaging. Magn. Reson. Med. 23, 37–45.
Friston, K.J., et al., 1995. Characterizing dynamic brain responses with fMRI: a multivariate approach. Neuroimage 2 (2), 166–172.
Hu, Y., et al., 2018. Weakly-supervised convolutional neural networks for multimodal image registration. Med. Image Anal. 49, 1–13.
Jenkinson, M., Smith, S., 2001. A global optimisation method for robust affine registration of brain images. Med. Image Anal. 5 (2), 143–156.
Jenkinson, M., et al., 2002. Improved optimization for the robust and accurate linear registration and motion correction of brain images. Neuroimage 17 (2), 825–841.
Kori A. and G. Krishnamurthi, 2019. Zero shot learning for multi-modal real time image registration. arXiv:1908.06213.
Li, Z., et al., 2020. Age-associated changes in cerebral blood flow-related measures using arterial spin labeling. Proc. Int. Soc. Magn. Reson. Med. 28, 3984.
Liao, R., et al., 2017. An articial agent for robust image registration. In: Proceedings of the ThirtyFirst AAAI Conference on Articial Intelligence.
Liu, X., et al., 2019. Image synthesis-based multi-modal image registration framework by using deep fully convolutional networks. Med. Biol. Eng. Comput. 57, 1037–1048.
Maleki, N., Dai, W., Alsop, D.C., 2012. Optimization of background suppression for arterial spin labeling perfusion imaging. MAGMA 25 (2), 127–133.
Miao, S., Wang, Z.J., Liao, R., 2016. A CNN regression approach for real time 2D/3D registration. IEEE Trans. Med. Imaging 35 (5), 1352–1363.
Power, J.D., et al., 2012. Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. Neuroimage 59 (3), 2142–2154.
Williams, D.S., et al., 1992. Magnetic resonance imaging of perfusion using spin inversion of arterial water. Proc. Natl. Acad. Sci. U. S. A. 89, 212–216.
Woods, R.P., et al., 1998a. Automated image registration: I. general methods and intrasubject, intramodality validation. J. Comput. Assist. Tomogr. 22, 139–152.
Woods, R.P., et al., 1998b. Automated image registration: II. intersubject validation of linear and nonlinear models. J. Comput. Assist. Tomogr. 22 (1), 153–165.
Ye, F.Q., et al., 2000. Noise reduction in 3D perfusion imaging by attenuating the static signal in arterial spin tagging (ASSIST). Magn. Reson. Med. 44 (1), 92–100.
Zhang, Z., et al., 2021. The longitudinal effect of meditation on resting-state functional connectivity using dynamic arterial spin labeling: a feasibility study. Brain Sci. 11 (10), 1263.
Zhang, Z., et al., 2022. Potential regulation of cerebral blood flow by the basal forebrain. Proc. Int. Soc. Magn. Reson. Med. 29, 5469.
Zheng, J., Miao, S., Liao, R., 2017. Learning CNNs with pairwise domain adaption for real-time 6dof ultrasound transducer detection and tracking from x-ray images. Med. Image Comput. Comput. Assist. Interv. - MICCAI. 10434, 646–654.
Zou, M., et al., 2019. Rigid medical image registration using learning-based interest points and features. Comput. Mater. Contin. 60, 511–525.