Multiresolution Registration Network (MRN) Hierarchy with Prior Knowledge Learning

Dongdong Gu[1,2], Xiaohuan Cao[2], Guocai Liu[1(✉)], Dinggang Shen[3,2], and Zhong Xue[2(✉)]

[1] College of Electrical and Information Engineering, Hunan University, Changsha, China
lgc630819@hnu.edu.cn
[2] Shanghai United Imaging Intelligence Co., Ltd., Shanghai, China
zhong.xue@uii-ai.com
[3] School of Biomedical Engineering, ShanghaiTech University, Shanghai, China

Abstract. Deep learning has been extensively used in unsupervised deformable image registration. U-Net structures are often used to infer deformation fields from concatenated input images, and training is achieved by minimizing losses derived from image similarity and field regularization terms. However, the mechanism of multiresolution encoding and decoding with skip connections tends to mix up the spatial relationship between corresponding voxels or features. This paper proposes a multiresolution registration network (MRN) based on simple convolution layers at each resolution level and forms a framework mimicking the ideas of well-accepted traditional image registration algorithms, wherein deformations are solved at the lowest resolution and further refined level-by-level. Multiresolution image features can be directly fed into the network, and wavelet decomposition is employed to maintain rich features at low resolution. In addition, prior knowledge of deformations at the lowest resolution is modeled by kernel-PCA when the template image is fixed, and such a prior loss is employed for training at that level to better tolerate shape variability. The proposed algorithm can be directly used for group analysis or image labeling and potentially applied for registering any image pairs. We compared the performance of MRN with different settings, *i.e.*, w/wo wavelet features, w/wo kernel-PCA losses, using brain magnetic resonance (MR) images, and the results showed better performance for the multiresolution representation and prior knowledge learning.

Keywords: Medical image registration · Multiresolution representation · Prior knowledge · Wavelet decomposition · Convolutional neural network

1 Introduction

In recent years, deep neural networks have been extensively used in deformable image registration, among which VoxelMorph [1] has been widely adopted due to easy training, fast inferring speed, and comparable results to traditional state-of-the-art methods. They have also been applied to align multi-modality images by disentangling the problem to a mono-modal one [2]. In learning-based registration, many deep registration methods

C. Lian et al. (Eds.): MLMI 2021, LNCS 12966, pp. 61–69, 2021.
https://doi.org/10.1007/978-3-030-87589-3_7

employ U-Net structures using multiresolution encoding and decoding paths with skip connections. Basically, a 3D deformation can be inferred by the networks after concatenating the input images, and unsupervised training is achieved by minimizing the losses derived from image similarity and deformation regularization terms.

However, the mechanism of multiresolution encoding and decoding architecture, especially with the skip connection, tends to mix up the spatial relationship between corresponding voxels of the images under registration or the feature maps at different resolutions. This may leave the task of solving deformations entirely to the convolutional layers and lead to more convolutional layers or complex network structures.

Recently, cascaded registration models have been proposed to improve the performance of registration [3–5]. Zhao *et al.* presented a volume tweening network (VTN) [3], which gradually registers a pair of images by using cascaded registration subnetworks, and each time the deformation field between the intermediately warped moving image and the fixed image are refined. Lately, they proposed a recursive cascaded network in [4], so the entire system can be jointly trained. During testing, one cascade network may be iteratively applied multiple times. But clearly, the networks should be different at different levels, especially the first network and the subsequent ones. de Vos *et al.* trained each cascade network one by one by fixing the weights of previous networks [5]. Despite using multiple networks, the effort seeks for more convolutional layers if one looks the cascade networks from end-to-end as a whole. In a typical U-shape network architecture, since the misalignment between feature maps caused by a residual or skip connection is complicated, a simple bilinear up-sampling and concatenating operation may break the symmetry between the down-sampling encoder and up-sampling decoder. Inspired by the Flow-Net [6], an semantic flow estimator layer was adopted in [7], so that the motion between two feature maps in different convolution layers can be compensated.

This paper uses convolutional neural network (CNN) for image registration while adopting the well-accepted idea from traditional image registration algorithms. We incorporate multiresolution representation of images, multilevel or hierarchical registration, and prior knowledge constraints in the training stage. Specifically, we apply elegant CNNs with less layers and parameters to solve the deformation field at the lowest resolution or to refine the fields at higher resolutions. The method is explicitly structured in multiresolution and allows for corresponding image features be directly fed into the subnetwork as inputs, thereby no skip-connection concatenation is needed. Various multiresolution image representations could be used, and herein wavelet decomposition is employed to maintain rich features. CNN at each level can be trained separately, and any image pairs can be well registered using MRN. In the case of group analysis or image labeling, where the fixed image (or template) does not change, a prior knowledge-based loss can be applied to the deformation field at the lowest resolution. Therefore, from a group of valid deformations, the kernel-PCA statistics can be defined to form a new prior knowledge loss.

In experiments, using brain magnetic resonance (MR) images, we compared the performance of MRN with different settings, *i.e.*, w/wo wavelet features, w/wo prior knowledge losses. We used Dice of brain tissues and number of deformation folding to evaluate the effectiveness of the proposed method. The experimental results showed better performance for the multiresolution representation and prior knowledge learning in terms of registration accuracy and topological correctness.

2 Method

2.1 Multiresolution Image Registration Hierarchy

The proposed network addresses several key components for multiresolution registration, including formulating a multi-level structure to first solve the deformations at the lowest resolution and then gradually refine the deformation fields, using multiresolution image features to help better match anatomical structures, and applying prior knowledge or statistics of the deformation fields to regularize the deformations for more robust registration. The CNN-based registration can be summarized in Fig. 1.

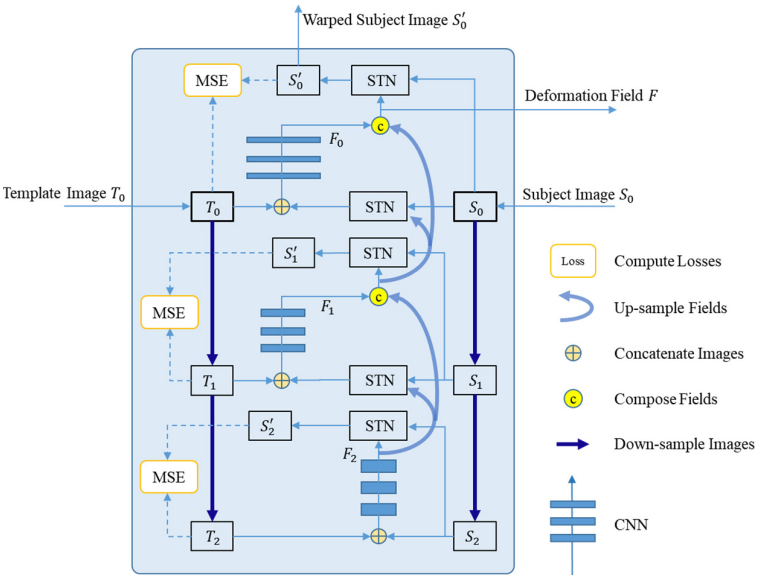

Fig. 1. The structure of MRN. The input template and subject images T_0 and S_0 are fed into the network, which are down-sampled (or transformed using wavelet packet transformation) into different resolution levels. Registration is performed hierarchically at different resolution levels, and the outputs include a deformation field and a warped subject.

As shown in Fig. 1, the input template and subject images (T_0 and S_0) are first down-sampled (half the size) to lower resolution levels (T_1, T_2 and S_1, S_2). F_2 represents the output deformation field at the lowest resolution (level 2), which aligns images T_2 and S_2. The deformed image $S_2^{'}$ can be generated using spatial transformation network (STN) based on F_2. Mean square error (MSE) between the deformed image and template image, and deformation smoothness losses can be applied to train the CNN at this level. The output deformation field F_2 is then up-sampled to level 1 and used to deform the subject image S_1. To train the CNN at level 1, MSE and smoothness losses can be used similarly, and the training can be performed after the lower level network is properly trained. Using the same way, a refinement field F_0 can be solved by training the CNN at level 0. Therefore, the network yields the final deformation field F by solving the deformation F_2 and the refinement fields F_1 and F_0 and then composing them.

It can be seen that the structure of the proposed MRN mimics the traditional registration and gradually refines the deformation fields. The network can be trained level-by-level or even trained independently when available deformations can be generated and properly down-sampled. As the outputs of each CNN are deformation fields, the training and performance of the network can be easily monitored. Moreover, the inputs at different levels are flexible, as long as they reflect multiresolution representations of the images to be registered. Notice that the deformation fields at the lowest resolution reflect the major anatomical variability, and hence different priors can be used at this level. Herein, as an example, we use a kernel-PCA loss to regularize the training at the lowest resolution by assuming the template image is fixed. For the higher two levels, only smoothness losses are used for regularization since the fields only reflect refinements. Finally, to solve the deformations at each level, we use simple CNN (Sect. 2.4) rather than the U-Net structure to eliminate its shortcomings mentioned in Introduction.

2.2 Multiresolution Representation and Prior Knowledge Learning

Multiresolution Representation of Input Images

As mentioned above, MRN allows for using different image representations as inputs. To maintain abundant image information in both low and high frequency, we adopt wavelet packet transform (WPT) [8] to generate images at resolution levels 1 and 2. Figure 2 shows an example of images after WPT. Three selected high-pass bands in the red boxes of the original image are combined to form the high-frequency image (mean absolute values), so the inputs for each image includes an image channel and a high-frequency channel, and the number of input channels of CNNs is 4 for all the levels.

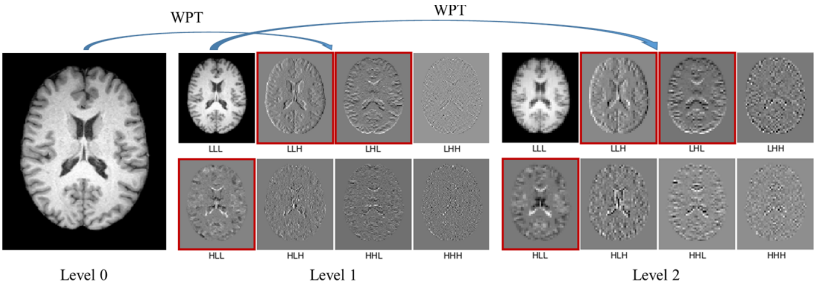

Fig. 2. Multiresolution representation of images with WPT. From left to right: original image as the input at level 0; low pass and combined bands in red boxes as the input of level 1; low pass and combined red bands as the input of level 2. (Color figure online)

Statistical Priors for Learning the CNN at the Lowest Resolution Level

Denoting the deformations at level 2 as $\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_N$, where N is the number of samples available. The prior distribution of the deformations can be modeled by kernel-PCA [9] with Gaussian kernel functions. The samples can be obtained from a traditional method

such as SyN [10] after proper down-sampling. According to kernel-PCA, the kernel matrix of N samples can be calculated by:

$$\kappa\left(\mathbf{f}_i, \mathbf{f}_j\right) = \exp\left\{-\gamma E\left(\mathbf{f}_i, \mathbf{f}_j\right)\right\},\ i = 1, \ldots, N; j = 1, \ldots, N, \tag{1}$$

where $E(\mathbf{f}_i, \mathbf{f}_j)$ is the Euclidean distance between two deformation fields $\mathbf{f}_i$ and $\mathbf{f}_j$ normalized by the number of voxels, $E\left(\mathbf{f}_i, \mathbf{f}_j\right) = \frac{1}{|\mathrm{M}|}\sum_{v\in\mathrm{M}}\|\mathbf{f}_i(v) - \mathbf{f}_j(v)\|_2$, and γ is a constant. M represents the image domain, and $|\mathrm{M}|$ is the number of voxels. Then, the centered kernel matrix can be calculated by:

$$\kappa_c = \kappa - 1\kappa - \kappa 1 + 1\kappa 1, \tag{2}$$

where 1 is a square matrix with element of $1/N$. Finally, the eigenvectors and eigenvalues of κ_c form the projection matrix ϕ and the variances λ along principal components. The projection of any new deformation field $\mathbf{f}$ can be calculated by:

$$\boldsymbol{\mu} = \phi^T \boldsymbol{k}, \tag{3}$$

where $\boldsymbol{k} = \left[\kappa\left(\mathbf{f}, \mathbf{f}_j\right), j = 1, \ldots, N\right]$. The shape of the input field $\mathbf{f}$ can be regularized by enforcing vector $\boldsymbol{\mu}$ within the range defined by the variance as:

$$L_{\mathrm{kpca}}(\mathbf{f}) = \sum\nolimits_{i=1,\ldots K} \boldsymbol{\mu}_i^2/\lambda_i, \tag{4}$$

where K is the number of principal eigenvectors. Traditionally, least squares estimation needs to be performed when reconstructing the constrained deformation field. Herein, with tensor programming of PyTorch, the above loss function can be automatically optimized through the gradient graphs of the software package.

2.3 Network Training Strategies

As seen from Fig. 1, MRN can be trained level-by-level. In addition to the prior loss defined in Eq. (4), we also use the MSE and smoothness loss functions. MSE is defined as the similarity between template images and deformed subject images:

$$L_{\mathrm{MSE},l} = \frac{1}{|\mathrm{M}|}\sum\nolimits_{v\in\mathrm{M}} \|T_l(v) - S_l(\mathbf{f}(v) + v)\|^2,\ l = 0, 1, 2. \tag{5}$$

where l represents resolution level. The smoothness loss is defined as:

$$L_{\mathrm{grad}} = \frac{1}{|\mathrm{M}|}\sum\nolimits_{v\in\mathrm{M}} \|\nabla\mathbf{f}(v)\|^2. \tag{6}$$

In summary, the total loss for each level l is defined as:

$$L_l = \alpha L_{\mathrm{MSE},l} + \beta L_{\mathrm{Grad},l} + \eta L_{\mathrm{kpca}},\ l = 0,\ 1,\ 2. \tag{7}$$

η is set to zero for level 0 and 1, so the prior loss does not apply for higher resolutions.

2.4 Algorithm Implementation

The CNNs at each level consist of 6 blocks, and each block includes a $3 \times 3 \times 3$ convolutional layer with no padding followed by ReLU, and the last layer does not have ReLU. The size of output at each level is smaller than that of the input (12 voxel difference) because no-padding operation is used. We cropped input images into patches with size $140 \times 140 \times 140$ and stitched back to the image space to save GPU memory. The output deformation field at the lower level should meet the size of the output of the current level after up-sampling. This gives an exact effective size of $56 \times 56 \times 56$ for each patch. So, we crop overlapping patches by skipping 56 voxels. Another consideration is that the kernel-PCA model is computed from the entire field at level 2, but for patch implementation, we can only use partial deformation field. Thus, the kernel vector $\boldsymbol{k}$ in Eq. (3) is calculated by using the deformation within the corresponding regions of each patch, *i.e.*, $\boldsymbol{\kappa} = \left[\kappa\left(P(\mathbf{f}), P(\mathbf{f}_j)\right), j = 1, \ldots, N\right]$, and $P()$ stands for only picking the deformation within the patch. With bigger patches covering a large area of the brain, distance calculated between two whole deformation fields is approximated by computing via partial fields, avoiding computing the statistics for different patch locations.

The network was implemented using PytTorch with Adam optimization. NVIDIA Geforce RTX 2080 Ti was used for training and testing. We trained the network up to1000 epochs (about 118,000 iterations) with batch size 1. The learning rate was set to 1e-5. The weights of loss terms in Eq. (7) were chosen as $\alpha = 1$, $\beta = 0.01$. η was set to 0.01 for the first half epochs and 0.1 for the rest.

3 Results

3.1 Datasets and Experiment Setting

We evaluated the performance of the proposed method using 150 T1 brain MR images from ADNI [11]. We randomly chose 120 images for training and 30 images for testing. One of the training images was selected as the template image (see Fig. 3 left). The goal was to train MRN by registering 119 sample images onto the template image in an unsupervised fashion. All the images were preprocessed by first applying N3 bias correction and then aligning them globally onto the template image space using affine registration. Skull stripping was performed using BET [12]. The images used for training and testing are with size $180 \times 216 \times 180$ and in isotropic $1 \times 1 \times 1\text{mm}^3$ spacing. For CNN inputs, the two images or two sets of WPT data are normalized to the range between 0 and 1 before cropping patches so that the intensities across different patch locations remain consistent for each image.

3.2 Algorithm Comparison

Figure 3 shows an example of the registration results of MRN using WPT inputs and kernel-PCA priors. It can be seen that the warped image at level 2 deforms a lot toward the template image, and the deformation can be further refined in the following levels. The proposed method also yields smooth deformation fields with less folding (will be

addressed below). Figure 4 plots more results of different testing subjects for visual assessment. The registration of all the 30 testing samples are successful, and the warped images look similar to the template image.

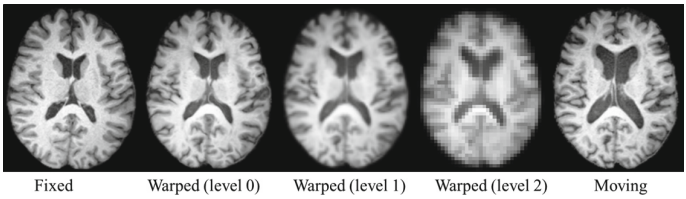

Fig. 3. Illustration of the (intermediate) results of the proposed algorithm. From left to right: the fixed image, the deformed images at level 0, 1, and 2, and the moving image.

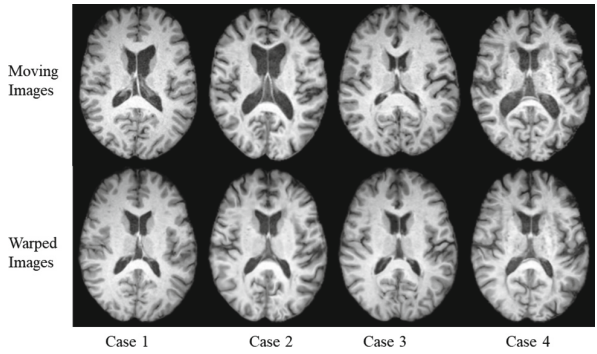

Fig. 4. Registration results of different subjects. Top: input moving images; bottom: warped images on the same template space. The fixed image is the same as Fig. 3.

In order to evaluate the performance of the proposed algorithm, we compared different training strategies and methods, including 1) using original images and their down-sampled versions as the network input (MRN-Image); 2) using original images and gradient magnitudes as input at level 0 and WPT data as input at level 1 and 2 (MRN-WPT); 3) with kernel-PCA prior loss based on 2) (MRN-WPT-KPCA); 4) with kernel-PCA prior loss based on 1) (MRN-KPCA); 5) a deep registration network similar to the U-Net architecture in [1] and the baseline used in [13] (Deep Registration); and 6) SyN [10]. The network structures and parameters are the same for all the MRN-based methods. The parameters of deep registration network and SyN are carefully tuned. We tested the methods on 30 testing images respectively.

Dice similarity coefficients (DSC) of white matter (WM), gray matter (GM) and Cerebrospinal Fluid (CSF) with ventricle between the masks on the template image and the ones warped from each subject image are used for quantitative evaluation. We also calculated the smoothness metric, *i.e.* permillage of folding in the deformation fields.

Figure 5 shows results of 30 testing data in terms of DSC for different settings of the compared methods. It can be seen that compared to the Deep Registration method, the DSC of GM, WM and CSF of our proposed MRN were all improved. The methods using

WPT features slightly outperformed others, and in terms of kernel-PCA constraints, they yield similar DSCs, but overall the standard deviations are at least 10% smaller, indicating more consistent and robust results over the testing datasets, and statistical models could generate more variability during network training.

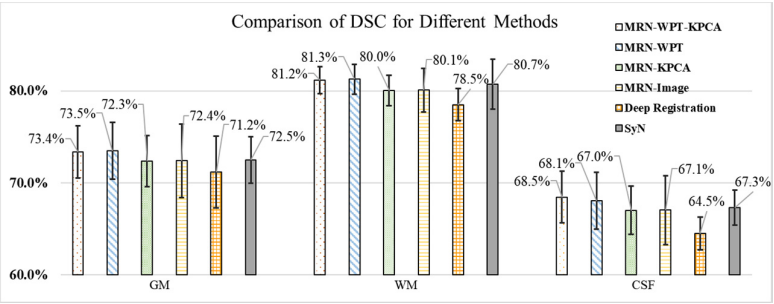

Fig. 5. Comparison of DSC for different methods.

We also counted the number of voxels with incorrect topology (*i.e.*, with negative Jacobian determinants), and the mean and standard deviation values in MRN-Image, MRN-WPT, MRN-KPCA, MRN-WPT-KPCA are 0.49 ± 0.37, 0.53 ± 0.36, 0.46 ± 0.38, and 0.32 ± 0.28, respectively (unit in permillage). The number of folding decreases greatly in MRN-WPT-KPCA, indicting our proposed method can yield reliable deformations, as less folding means better morphologically allowable fields. Moreover, the median of folding in MRN-WPT-KPCA is only 0.2‰, which is significantly smaller than other methods (all > 0.35‰).

One of the drawbacks of the comparison is that as the statistical model losses are only applied in the lowest resolution, the differences might be overwhelmed by the processing at the higher two levels. Additionally, we believe that the statistical losses may help improve the robustness of the registration networks and plan to validate and test on atlas-based applications on bigger datasets with extensive data augmentation.

4 Conclusion

We proposed an MRN using simple convolution layers at each resolution level by mimicking the ideas of well-accepted traditional image registration algorithms. Simple CNNs were used to solve the deformations at each resolution and gradually refine the deformation level-by-level. MRN allows for image features in different resolutions be directly fed into the network, and WPT images are naturally fit into the structure. Additionally, for registering with template image, the prior knowledge of deformations at the lowest resolution is modeled by kernel-PCA, which can be easily embedded into the network training. Experiments using brain MR images showed the advantages of MRN compared with different settings, *i.e.*, w/wo wavelet features, w/wo prior knowledge losses. The results showed better performance for the multiresolution representation and prior knowledge learning.

Acknowledgement. This work was partially supported by the National Key Research and Development Program of China (2018YFC0116400) and the National Natural Science Foundation of China (62071176).

References

1. Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V.: VoxelMorph: a learning framework for deformable medical image registration. IEEE Trans. Med. Imaging **38**, 1788–1800 (2019)
2. Qin, C., Shi, B., Liao, R., Mansi, T., Rueckert, D., Kamen, A.: Unsupervised deformable registration for multi-modal images via disentangled representations. In: Chung, A.C.S., Gee, J.C., Yushkevich, P.A., Bao, S. (eds.) IPMI 2019. LNCS, vol. 11492, pp. 249–261. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-20351-1_19
3. Zhao, S., et al.: Unsupervised 3D end-to-end medical image registration with volume tweening network. IEEE J. Biomed. Health Inform. **24**, 1394–1404 (2019)
4. Zhao, S., Dong, Y., Chang, E.I., Xu, Y.: Recursive cascaded networks for unsupervised medical image registration. In: International Conference on Computer Vision (ICCV), 2019, pp. 10600–10610. IEEE (2019)
5. de Vos, B.D., Berendsen, F.F., Viergever, M.A., Sokooti, H., Staring, M., Išgum, I.: A deep learning framework for unsupervised affine and deformable image registration. Med. Image Anal. **52**, 128–143 (2019)
6. Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., Brox, T.: Flownet 2.0: evolution of optical flow estimation with deep networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2462–2470. IEEE (2017)
7. Li, X., et al.: Semantic flow for fast and accurate scene parsing. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12346, pp. 775–793. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58452-8_45
8. Xue, Z., Shen, D., Davatzikos, C.: Statistical representation of high-dimensional deformation fields with application to statistically constrained 3D warping. Med. Image Anal. **10**, 740–751 (2006)
9. Rosipal, R., Girolami, M., Trejo, L.J., Cichocki, A.: Applications: kernel PCA for feature extraction and de-noising in nonlinear regression. Neural Comput. Appl. **10**, 231–243 (2001)
10. 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. Med. Image Anal. **12**, 26–41 (2008)
11. Mueller, S.G., et al.: Ways toward an early diagnosis in Alzheimer's disease: the Alzheimer's disease neuroimaging initiative (ADNI). Alzheimer's Dement. **1**, 55–66 (2005)
12. Smith, S.M.: Fast robust automated brain extraction. Hum. Brain Mapp. **17**, 143–155 (2002)
13. Gu, D., et al.: Pair-wise and group-wise deformation consistency in deep registration network. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 171–180. Springer, Cham (2020) https://doi.org/10.1007/978-3-030-59716-0_17