Journal of Modern Optics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tmop20

Multimodal MR image registration using weakly supervised constrained affine network

Xiaoyan Wang, Lizhao Mao, Xiaojie Huang, Ming Xia & Zheng Gu

To cite this article: Xiaoyan Wang, Lizhao Mao, Xiaojie Huang, Ming Xia & Zheng Gu (2021) Multimodal MR image registration using weakly supervised constrained affine network, Journal of Modern Optics, 68:13, 679-688, DOI: 10.1080/09500340.2021.1939897

To link to this article: https://doi.org/10.1080/09500340.2021.1939897

Published online: 21 Jun 2021.

JOURNAL OF MODERN OPTICS
2021, VOL. 68, NO. 13, 679–688
https://doi.org/10.1080/09500340.2021.1939897

Taylor & Francis
Taylor & Francis Group

Multimodal MR image registration using weakly supervised constrained affine network

Xiaoyan Wang[a,c], Lizhao Mao[a], Xiaojie Huang[b], Ming Xia[a] and Zheng Gu[b]

[a]School of Computer Science and Technology, Zhejiang University of Technology, Hangzhou, People's Republic of China; [b]The Second Affiliated Hospital, School of Medicine, Zhejiang University, Hangzhou, People's Republic of China; [c]Key Laboratory of Visual Media Intelligent Processing Technology of Zhejiang Province, Zhejiang University of Technology, Hangzhou, People's Republic of China

ABSTRACT

Multimodal image registration is an important technique for many clinical applications. However, it is particularly challenging to obtain good spatial alignment. This paper introduces a novel architecture named the constrained affine network, which combines deformable image registration with affine transformation for multimodal MR image registration. A weakly supervised manner is adapted to train the network and anatomical labels are used in training. The network directly learns to predict a displacement vector field (DVF) between pairs of input images. Different from the existing deformable image registration methods based on the convolutional neural network (CNN), the method proposes a global constrained affine module, which can predict an affine transformation by pre-computing the range of affine parameters, and the model can be combined with a deformable registration network. We evaluated the proposed method on 3D multimodal medical images. Experimental results indicate that the proposed method has better performance.

ARTICLE HISTORY
Received 18 September 2020
Accepted 24 May 2021

KEYWORDS
Medical image registration; constrained affine; convolutional neural network (CNN); MRI

1. Introduction

Medical image registration is a prerequisite for many clinical applications, such as image-guided intervention, image fusion [1], medical diagnosis and so on. The aim of the registration algorithm is to find an optimum spatial transformation between a pair of images, and thus to ensure the complementary information of multi-sequence magnetic resonance imaging (MRI). However, most traditional registration algorithms are based on iterative optimization to search for the optimal parameters, which results in intensive computation, and so they do not scale well to clinical practice. Recently, registration approaches based on deep learning have been actively explored and they can be completed in a shorter time than the traditional optimization-based registration methods [2].

Image registration methods based on deep learning can generally be divided into deformable and rigid registration. Deformable registration is used to find the association through an optimal nonlinear transformation, and is a fundamental method widely used in medical image analysis [3]. Fan et al. [4] proposed a BIR-Net to perform brain image registration using a dual supervision learning strategy, they employed two ground truths generated by both ANTs [5] and LCC-Demons [6]. Unlike conventional methods to obtain ground truths described above, synthetic random transformations are used to train CNN. This method does not require manually annotated dataset and the output of the network is the displacement vector field on a thin plate spline transform grid [7,8]. The above-mentioned registration methods are all supervised. However, it is difficult to generate many matching clinical medical image pairs with known transformations.

For unsupervised learning, the registration methods are designed to overcome the shortcomings of supervised registration methods. Jaderberg et al. [9] proposed a spatial transformer network (STN), which can align the input images as a component within the network. Inspired by STN, an unsupervised 3D medical image registration framework was developed to predict a dense spatial transformation in medical image datasets [10–12]. The article used a U-net architecture as the registration modal and named it VoxelMorph. Later, Zhao et al. [13] proposed a volume tweening network (VTK) for deformable image registration, which significantly improved the accuracy of unsupervised registration algorithms. The final prediction can be disintegrated into

CONTACT Xiaojie Huang caicaitu@zju.edu.cn The Second Affiliated Hospital, School of Medicine, Zhejiang University, Hangzhou 310009, People's Republic of China

simple, slight changes with small displacements, thereby to resolve the difficulty of input images in the presence of large displacement. Meanwhile, Zhao et al. [14] proposed a similar idea and it also a recursive cascade pattern. However, the difference is that this method calculates the loss function at the final subnetwork component, but VTK calculates the loss in each subnetwork. These approaches do not require any ground truths and thus have an advantage over supervised registration methods. However, unsupervised registration methods usually ignore the inherent inverse-consistent property of transformation between each pair of images, thus failing to ensure the transformations are inverse of each other [15]. In addition, Hu et al. [16,17] developed a weakly supervised registration framework for multimodal image registration, which predicts a dense correspondence using labels that represent the same spatial correspondences of anatomical structures. They proposed a composite-net that combines a Global-Net with a Local-Net, and the Global-Net is sensitive to initialization. In order to control the magnitude of the initial transformations, they used a summand node to allow random initialization with a zero mean and a small variation in the affine transformation parameters.

Although these image registration methods show promising registration accuracy and efficiency, they still have inherent limitations. Firstly, since affine parameters are sensitive to network initialization, most methods obtain random initialization through network iterative optimization, but affine initialization is very susceptible to the influence of deformation parameters, thus increasing the amount of calculation for network training. Secondly, the affine and the deformable components are independently designed to limit the computational memory, which increase the difficulty of training.

In this work, we focus on fusing the affine transformation of prior information into CNN in a weakly supervised learning method. CNN can be trained to generate coarse-grained spatial transformations by maximizing the similarity of its anatomical labels to register image pairs without obtaining the ground-truth data in training. Our contributions are as follows:

1. We propose a novel multimodal carotid medical image registration based on strongly constrained affine transformation. Displacement vector field (DVF) and affine transformation are integrated into the same network to prevent the limited computational memory. It turns out that the method is accurate and faster compared to competing methods. Through the fusion of affine and deformable methods, we can obtain a reasonable dense spatial transformation.
2. In particular, the proposed scheme optimizes deformable and affine transformation by uniformly constraining the prior information of the parameter range to normalize and solves the problem of scale imbalance in the initialization of training affine parameters.
3. Extensive experiments based on the diverse CNN model and multiple datasets (carotid MRI and brain MRI) demonstrate that the proposed method further improves the registration performance. In addition, the registration results using and without using the anatomical centre point distance as the registration similarity measure are compared to verify the sensitivity of the distance to the experimental dataset.

2. Methods and experiments

2.1. Registration architecture

Medical image registration aims to find an optimal spatial transformation that establishes the spatial correspondences between different anatomical structures. Through the training of the convolution network, we can predict the DVF corresponding to each voxel in the fixed image space, and the DVF can transform or spatially resample the moving image to match the fixed image. To obtain DVF, we give N pairs fixed images I_f and moving images I_m over an n-D spatial domain $\Omega \subset R^n$. We set a function $u = f_\theta(I_f, I_m)$ as a displacement field, where f is the convolutional network and θ is the parameter to be trained.

Figure 1 shows our training model for image registration, let If and Im be inputs to predict the function parameters θ and to obtain the displacement field. The linear image resampler is then used to obtain warped image I_w using warp $u(I_m)$. In this work, the parameter θ is updated by calculating the label similarity between the fixed image label Lf and the moving image label Lm. During training, we use the adaptive gradient descent optimization algorithm to find the optimal parameters $\overset{\frown}{\theta}$ by minimizing the loss function as follows:

$$\overset{\frown}{\theta} = \arg\min_{\theta} L(L_f, L_m, u) \tag{1}$$

where Lf and Lm are used to quantify the label similarity, and on this basis, we add the centre distance. Due to the sparsity of Lf and Lm, it is essential to utilize u to compute regularization losses for the smoothness of the entire field. The architecture model and the loss function details will be described in the next sections.

The application stage is shown in Figure 2, only a pair of fixed and moving images is needed as the input of the trained network, and the DVF can be predicted by the

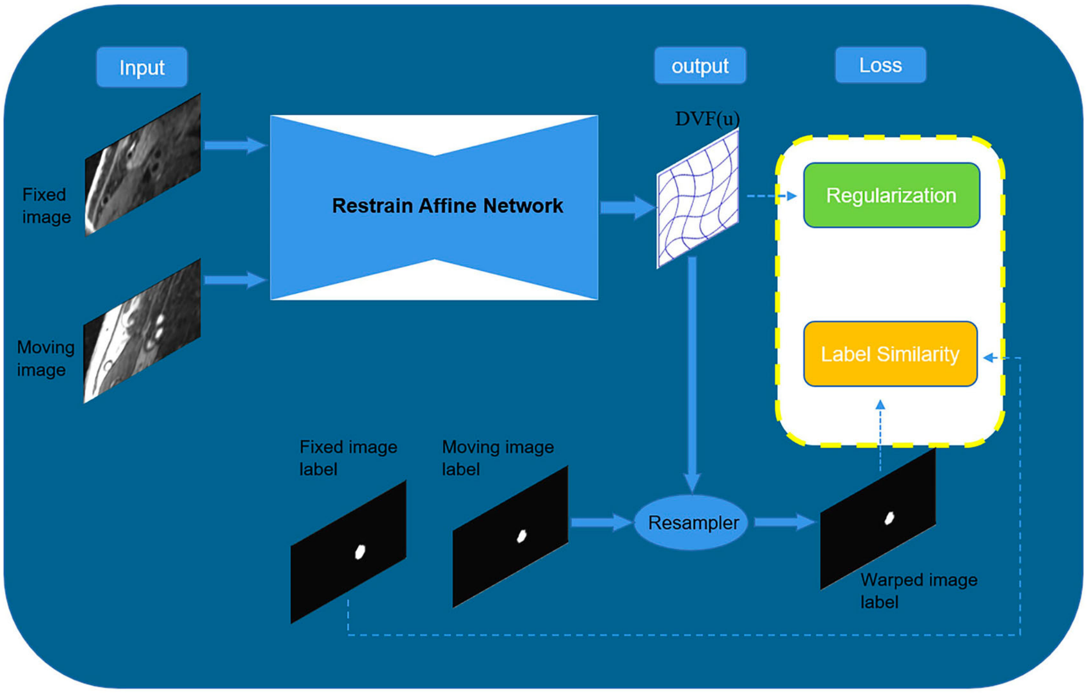

Figure 1. Illustration of the overall training process of the weakly supervised registration framework without ground truth, and single anatomical structure labels (L_f and L_m) are available during training. Dotted box indicates what the loss is computed. The framework can be trained by optimizing a loss function integrated by the label similarity and the deformation regularization.

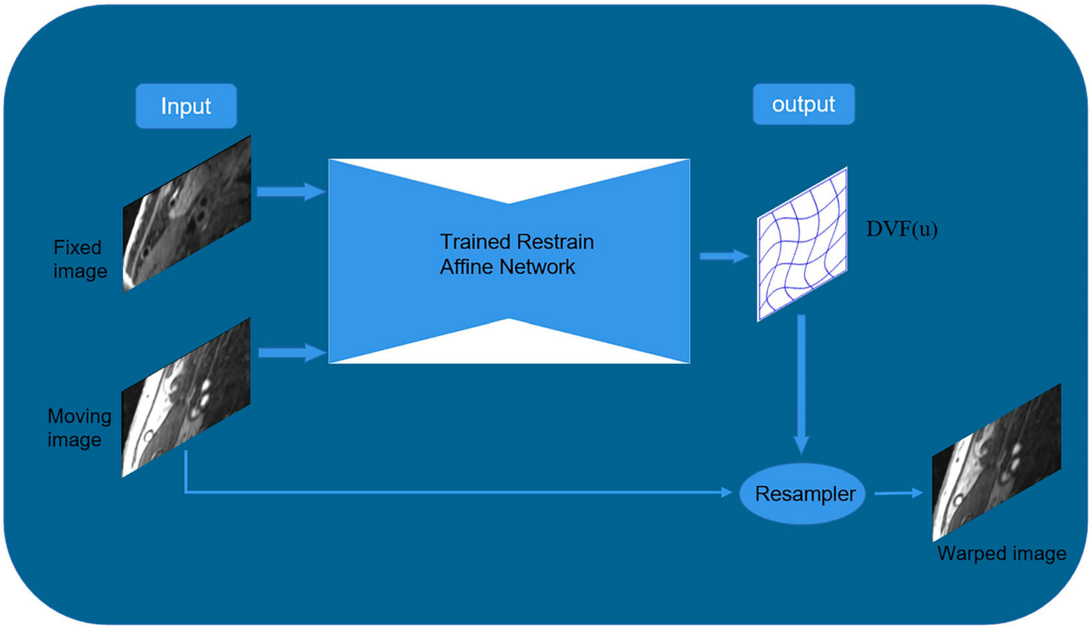

Figure 2. Illustration of the registration application process of the weakly supervised registration framework without any labels.

network. As a result, labels are not required in the application stage, and the fully automatic image registration method is an effective way to predict DVF without any kind of segmentation to assist alignment.

2.2. Constrained affine network

In order to learn the complex affine and deformable displacement field, we propose a constrained affine network (CAN) to facilitate global feature matching. Figure 3 depicts the structure of CAN. The network uses a single input which concatenated the information from I_f and I_m channels related to the two images. Since the U-net contains symmetrical contractive [18,19] and supplements feature information via a skip connection [20], the U-net has shown good results in localized learning and feature learning from a small number of datasets, so we utilize

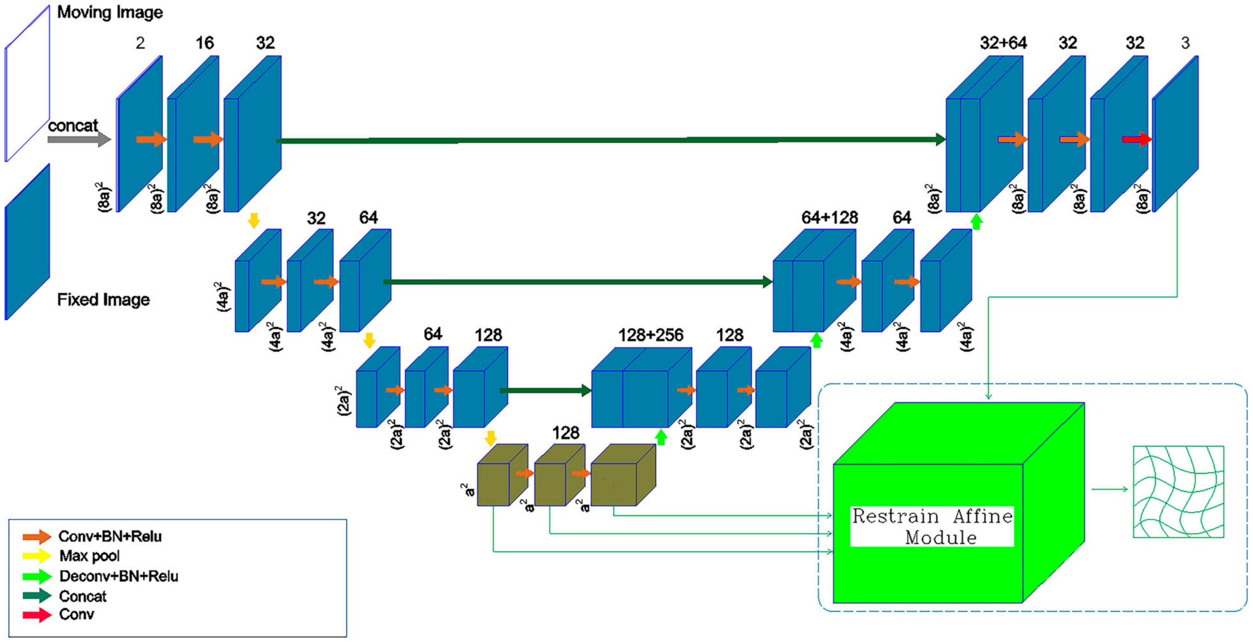

Figure 3. Illustration of the architecture of the constrained affine network, and the UNet framework is implemented in our architecture by combing the deformable and the affine registration components.

the U-net model to predict the whole DVF. The network takes two channels as the input and the number of the final output channels is three. In the figure, the brownness convolution displays the input site of the CAN. The constrained affine part has four inputs to synthesize all the predictions of the transformation parameters. The only output of the constrained affine module warps the moving image to align with the fixed image. A kernel size of 7×7×7 is used in the first 3D convolution layer to ensure sufficient receptive field. The registration network consists of convolutional encoding and decoding stages. The encoding stage follows the typical structure of a convolution, which includes repeated applications using two 3×3×3 convolutions followed by a relu activation layer and a batch normalization layer [21] to double the number of feature channels. In order to achieve down-sampling, we use the max-pooling layer with stride size of two to reduce the spatial dimension in half. In the decoding stage, each step includes up-sampling through transposed convolution and a concatenating skip connections, and also uses two 3×3×3 convolutions to capture hierarchical features and predict the spatial correspondence of the input images. The final layer of the network uses a 1×1×1 convolution to directly output a dense DVF in three directions without activation and the batch normalization layer.

The registration network predicts displacement fields comprised of two spatial transformations: deformation transformation and affine transformation. Deformation transformation is used to resolve the resource of mismatching between each voxel to correct local deformation, but unlike nonlinear deformable registration, affine transformation is one of the global transformations performed by physical parameters [22]. Due to the sensitivity of affine parameters to network initialization, in our model, a constrained affine model is designed to resolve this problem.

Between the down-sampling and up-sampling of the network, there is a convolution of cascaded structures, which can obtain different abstract features by connected convolution layers as the multiple feature input of the affine transformation model. Figure 4 shows the details of the internal structure of the constrained affine module. There are three convolution layers between down-sampling and up-sampling of the U-net as a multiple feature input. Traditional affine transformation as a prior information is implemented to initialize the sensitive affine parameters and the constrained values are pre-computed. We assign three appropriate ranges by enlarging and narrowing the pre-computed constrained values. Optimization is performed through network training to refine the previous assign ranges and the predicted affine grid. The affine model is a shared down-sampling structure with deformation transformation. In this way, we can make full use of the characteristic parameters of different layers and project them linearly into the affine matrix to predict more accurate affine deformation. In the process of initializing and iterative optimization, we limit the affine parameters to an optional range to optimize the adjustment of sensitive affine parameters. In order to

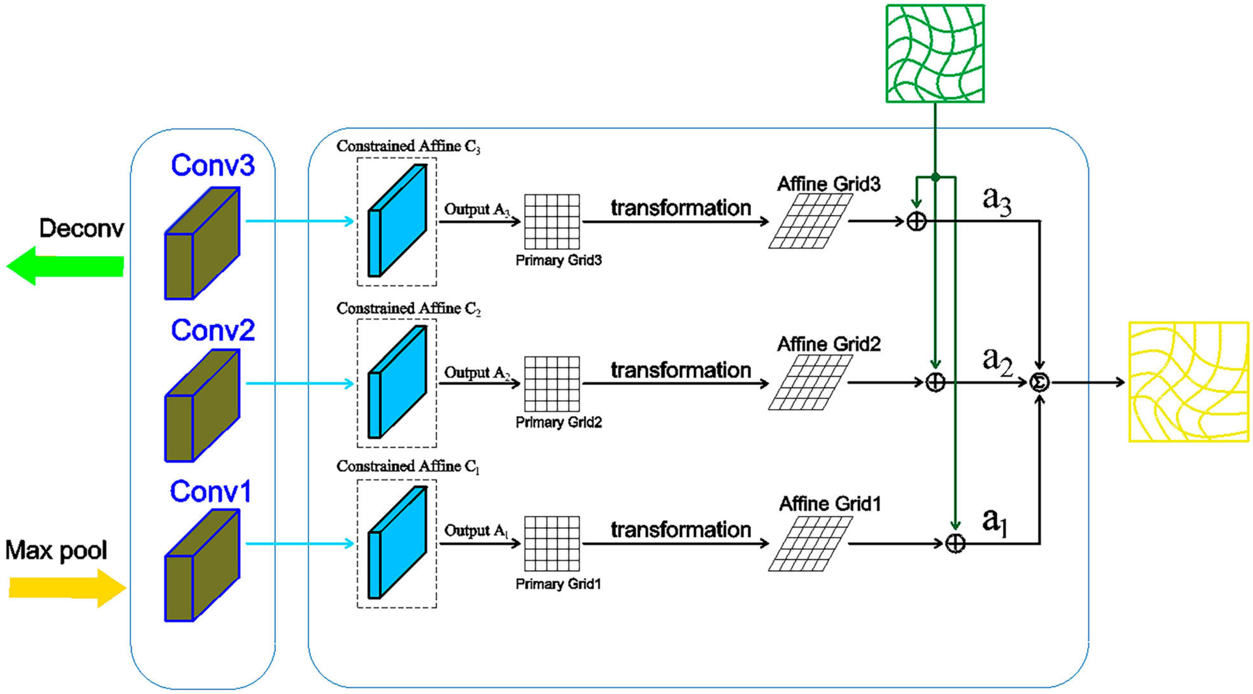

Figure 4. Illustration of the multiple feature input module for the output deformation transformation summation used in the proposed constrained affine network. Specially, the hyper-parameters are added up to one ($a_1 = 0.2$, $a_2 = 0.3$, $a_3 = 0.5$).

directly predict affine transformation, a neural network with a single 12 neurons layer is used, which represents the affine parameters for translation, rotation, scaling and shearing. We set the input of the affine layer to x and then output the optimized affine parameters for the global transformation, the affine transformation layer can be represented by the following equation:

$$RM(x) = \theta_{rw}x + \theta_b \qquad (2)$$

where RM represents the affine transformation matrix with 12 degrees of freedom, and the affine map be further optimized by θ_{rw} and θ_b, and the allowable range of $\theta_{rw}x$ is constrained. Where $R \le \theta_{rw}x \le R$, R is a constrained value. Thus, the parameter of the matrix can be strongly constrained, and each affine layer has a different constrained value to increase the variable range of affine prediction.

2.3. Loss function

In this section, we first describe the loss function used in our model. The loss function consists of two components. The first component $l_{ls}(L_f, u(L_m))$ is a similarity loss, which uses the dice-coefficient to measure the similarity of anatomical labels between the fixed and deformed moving labels, and the other is a deformation regularization $l_{smooth}(u)$ to penalize the non-smooth displacements. Thus, to train the inter-model registration network, the following loss can be applied:

$$l(L_f, L_m, u) = 1 - l_{ls}(L_f, u(L_m)) + \alpha l_{smmoth}(u) \qquad (3)$$

where α is a regularization parameter. In our experiments, the labels are only used to evaluate l_{ls} and not used as the network input. In addition, we use centre point distance dt of the anatomical label as a part of l_{ls}, and we add a sigmoid function to map the distance value to the range of 0 and 1 to prevent the value from oscillating too much and affecting training, l_{ls} is given as follows:

$$L_{ls} = dice(L_f, u(L_m)) + \frac{1}{1 + e^{-dt}} \qquad (4)$$

Additionally, penalizing the transformation field to ensure the smoothness of u:

$$\begin{aligned} l_{smooth} = \frac{1}{V} \iiint_0^V & [(\partial^2 T/\partial x^2)^2 \\ & + (\partial^2 T/\partial y^2)^2 + 2(\partial^2 T/\partial xy)^2 \\ & + 2(\partial^2 T/\partial xz)^2 + 2(\partial^2 T/\partial yz)^2] d_x d_y d_z \qquad (5) \end{aligned}$$

where V denotes the volume of the image domain, T is the deformation fields, and the regularization weight is set between the bending energy and the similarity loss. Note that for affine transformation, this regularization is zero, so only the deformable transformation is penalized [23].

2.4. Datasets

For evaluating the presented approach, a carotid artery dataset collected from our cooperative hospital was used. The patients gave informed written consent, and our study protocol has been approved by the local regional ethics committee. Additionally, in order to verify the generalizability of the algorithm, we also evaluate the registration model on the Brain tumour Segmentation (BraTS) dataset [24–26].

The clinical carotid dataset contains 3D carotid MRI images collected from 11 patients with carotid atherosclerosis (including the 30–69% of stenosis), and each patient has several difference sequences to analyse the development of CVD. In our experiment, T1 with the Gd-based contrast agent (T1Gd) and T1-weighted (T1) CUBE sequences are respectively selected as the fixed and moving images. The MR images were collected by a bilateral four-channel phased-array carotid surface coil, which applied fast spin-echo for carotid arteries on a 1.5 T MR system. In addition, CUBE imaging was performed before and after the intravenous administration of 0.1 ml/kg of a 1.0 M Gd-based contrast agent. The imaging parameters of T1 and T1Gd CUBE sequences were as follows: each slice size is 512×512, thickness: 1.2 mm (interpolated to 0.6 mm), and 64—72 slices in the coronal plane. N4-biased field correction [27] is applied to correct the intensity inhomogeneity of the selected sequence images. Then, we resampled all sequence images in the same image spacing. The corresponding carotid structures in the image were manually labelled by experts. All the carotid images are resized and cropped automatically into a uniform size by removing the outer border of the images according to the carotid artery position of each patient, and the cropping area of different patients is consistent. After that, we cut the 3D carotid MRI images of 11 patients with carotid atherosclerosis into two halves and remove two abnormal right carotid artery data. Finally, we obtained a total of 20 pairs of labelled MR images with the resolution of 112×64×64 pixels and 0.6 mm slice spacing. A four-fold cross-validation procedure was used to ensure the balance of training data. Each training and test set therefore consisted of 15 and 5 samples, respectively. In addition, there are many manually landmark annotations in the carotid lumen bifurcation and plaque utilized to evaluate the registration algorithms. Figure 5 shows examples of landmarks in the carotid dataset.

We obtain publicly available brain MRI scans from The BraTS 2020 challenge. The training set contains 373 cases of brain MRIs with the patients' tumour segmented manually, and four contrasts (T1Gd, T1c, T2, T2-FLAIR) are used in the study. The size of the four contrasts images is about 240×240×155 with 1 mm isotropic resolution in a standardized axial orientation. We utilize T1c and T2 to form 80 pairs of random fixed images and moving image for experiments, 60 pairs for training and 20 pairs for testing. All raw scans are down-sampled to the size of 96×96×64. Specially, in order to make the registration algorithm focus on the important components of the tumour structures, we remove the oedema portion of the tumour segmentation. As the raw scans are well aligned, we apply elastic deformations followed by Gaussian smoothing to generate random synthetic deformation fields to transform the T2 images [28]; thus, in this way we can produce reasonable misalignments. Then, the synthetic deformations can be used as ground truth for evaluations.

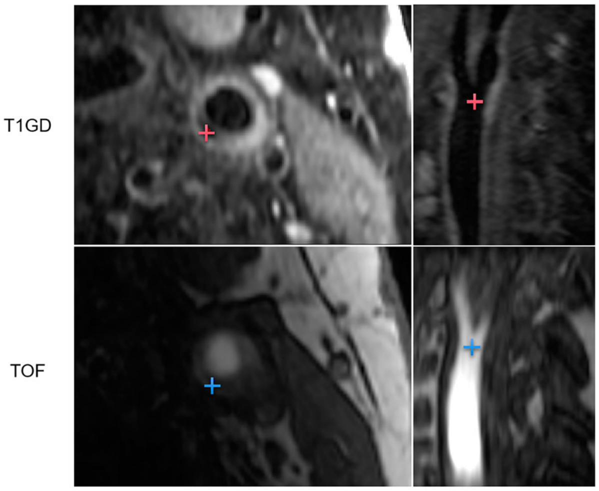

Figure 5. Example of the carotid lumen bifurcation and plaque. Crosses display the position of landmarks.

2.5. Experimental set-up and evaluation

2.5.1. Training details

We build our constrained affine network based on the network architecture of the 3D U-net [29]. Unlike the original network, we halve the number of channels and kernel size of 7×7×7 used in the first convolution layer to greatly increase the receptive field. Furthermore, for the constrained affine module, the bias parameters θ_b is initialized to [1., 0., 0., 0., 0., 1., 0., 0., 0., 0., 1., 0.] (3×4 matrix). Meanwhile, we use the traditional affine method to obtain affine parameters as prior information to precompute the constrained value R, which ensures the value is limited within reasonable scope.

In our implementation, some data augmentation methods are added to improve the generalization ability of the model. We apply a random affine transformation augmentation during training. In addition, we flip all

Table 1. Quantitative carotid MRI and BraTS MRI registration performance with different training networks.

	Carotid			BraTS		
Method	DSC (%)	Lm.Dist (mm)	Time(s) CPU/GPU	DSC (%)	CPD (mm)	Time(s) CPU/GPU
U-net	0.767	0.954	4.711/0.267	0.848	2.045	12.569/0.627
MultiResUnet	0.762	1.418	4.263/0.284	0.794	4.856	11.418/0.581
AttentionUnet	0.823	1.262	4.634/0.232	0.858	1.966	12.234/0.632
U-net + CAN	0.833	0.757	4.272/0.241	0.869	1.812	11.368/0.612
MultiResUnet + CAN	0.811	1.421	4.562/0.217	0.821	1.070	13.570/0.679
AttentionUnet + CAN	0.839	0.692	4.417/0.184	0.862	1.886	11.087/0.570

patches in the x direction to expand the training data. Our method is implemented and trained using TensorFlow [30] with a single Nvidia Geforce GTX 1070Ti. The Adam Optimizer [31] is used for training with an initial learning rate of 1e−4. Each model is trained in a batch size of 4 through a gradient accumulation method with an iteration count of 20,000, and Xavier initializer is utilized as the initialization for the other network parameters. The choice of hyper-parameters is important for training models. As the sensitivity of the dataset to hyper-parameters is different during the experiments, a is set respectively to 0.7 and 0.01 on the carotid datasets and BraTS datasets based on the performance of the test set.

2.5.2. Evaluation metric

We quantify the accuracy of the algorithm by comparing results of three methods. Dice scores (DSC) between the fixed mask and the warped moving mask are computed by warping the segmentation of the registration structure using the predicted displacement fields. In addition, each of the two datasets has a different evaluation methodology. For each carotid artery scan, there are many manually landmarks including identified anatomical distribution of carotid vascular bifurcation and plaque. We compute the average distance between the landmarks (Lm.Dist) between the fixed landmarks and the warped landmarks. Similar to the Dice compute style using mask, for each BraTS scan, we calculated the centre point distance (CPD) of the anatomical labels between the fixed mask and the warped moving mask as a part of the evaluation metric. In terms of the runtime performance, the average running time to register each pair of images in seconds is provided as a comprehensive evaluation of registration performance.

2.5.3. Comparison among networks

In order to accurately evaluate the performance of the constrained affine network. In our experiment, AttentionUnet [32] and MultiResUnet [33], two extensions to the U-net model, are used to make a more obvious comparison. We compare against these original networks and the variant networks together with the CAN. Specially, we train Attention Unet and MultiRes Unet model with the same parameters in the U-net. Meanwhile, in order to evaluate the impact of the centre point distance strategy in the training, we use two loss functions in the experiment to improve the credibility. The only difference between these two functions is whether there is dt. Finally, we obtained six different models to reveal the effect of our proposed registration algorithm.

3. Results and discussion

3.1. Model performance

As shown in Table 1, our proposed method is evaluated by the average DSC, Lm.Dist and CPD of the anatomical label between the fixed and the warped moving mask. For the carotid artery dataset, DSC measures the overlap of anatomical labels, and the Lm.Dist in millimeter (mm) is measured on the corresponding annotated spatial landmarks. In addition, CPD evaluates the average centre point distance among all the anatomical labels in the BraTS dataset. It is obvious that our method achieves a better structural alignment. We obtain

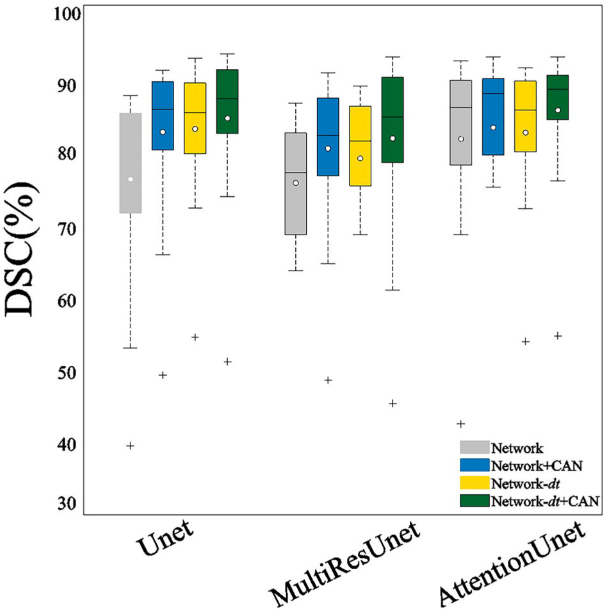

Figure 6. Boxplots of DSC values of the performance from carotid dataset. The quantitative results are also summarized in Table 1.

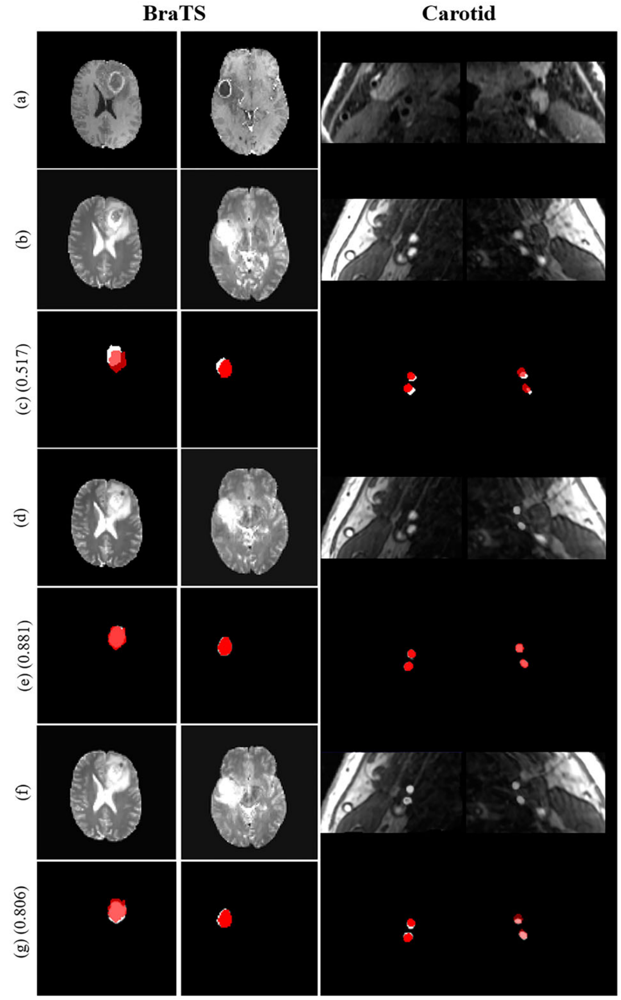

Figure 7. Example of registration results comparison among Unet+CAN (d/e) and Unet (f/g). Each row refers to an example registration case (a) a MRI scan of the fixed image; (b) a MRI scan of the moving image; (c) the original overlap of the fixed anatomical labels and the moving anatomical labels (red areas represent the labels of the fixed image, and white areas represent the labels of the moving images); (d/f) a MRI scan of the warped image; (e/g) the overlap after registration of the fixed anatomical labels and the warped anatomical labels. The average DSC values of the four examples are shown on the picture.

an average improvement from 78.4% to 82.8% across all three training networks in terms of DSC on the carotid dataset, while reducing the Lm.Dist effectively from 1.211 to 0.957 mm. For ease of visualization, Boxplots for the DSC of carotid anatomical labels are shown in Figure 6. For the BraTS dataset, the quantitative results are also summarized in Table 1. We compared the registration performance with a different metric in CPD. Our method yields the shortest distance of the centre point among these registration networks. In Figure 7, we compare U-net and U-net + CAN by visualizing four example registration results. We intercept the slice along the y axis to display the overlap of the anatomical label and there may be other cases in other slice locations. This shows that the constrained affine network has a better structural alignment. At the same time, experiments show that our method can achieve high registration accuracy efficiently on GPU. The running time of the registration process depends not only on the network structure but also on the degree of misalignment between the input images. Compared to the carotid artery images, it is more time consuming on high-resolution BraTS images.

Table 2 presents the correlation of the loss function with and without centre point distance of anatomical labels. As in Table 1, Table 2 has the same evaluation indicators. As described in Section 2.5.2, the only difference is that the centre point distance with the sigmoid function being added during all the training processes. Considering the network was trained with the loss function based on the centre point distance; interestingly, adding centre point distance to dice has almost similar DSC for BraTS data and a better DSC result for the carotid artery data. The results indicate clearly that the carotid artery data are more sensitive to the loss function of distance than the BraTS data. Furthermore, based on visual and quantitative analysis of the results, AttentionUnet have shown effectiveness in the performance of each group.

Because medical image folding is anatomically impossible, especially for registration between images of the same patient, Jacobian determinants can measure the ability of the DVF to keep the image topology unchanged [22]. If the value of the Jacobian determinant is smaller than zero, it indicates that the voxel positions in the DVF show folding. Setting the regularization weight $a = 0.7$ and $a = 0.01$, respectively, in Equation (3) will lead to no negative Jacobian determinants in any of the DVF. In order to display the effect of the hyper-parameter a, we set

Table 2. Summary of registration results when using the *dt* of the anatomical label as a part of loss function.

	Carotid			BraTS		
Method	DSC (%)	Lm.Dist (mm)	Time(s) CPU/GPU	DSC (%)	CPD (mm)	Time(s) CPU/GPU
U-net-*dt*	0.837	0.900	4.634/0.213	0.854	2.213	14.140/0.717
MultiResUnet-*dt*	0.795	1.578	4.473/0.297	0.795	3.142	11.847/0.594
AttentionUnet-*dt*	0.843	0.704	4.726/0.187	0.841	2.453	13.778/0.694
U-net-*dt* + CAN	0.852	0.985	4.712/0.209	0.868	2.328	13.088/0.794
MultiResUnet-*dt* + CAN	0.824	1.352	4.581/0.317	0.830	2.258	12.816/0.643
AttentionUnet-*dt* + CAN	0.863	0.628	4.592/0.211	0.862	2.093	11.581/0.617

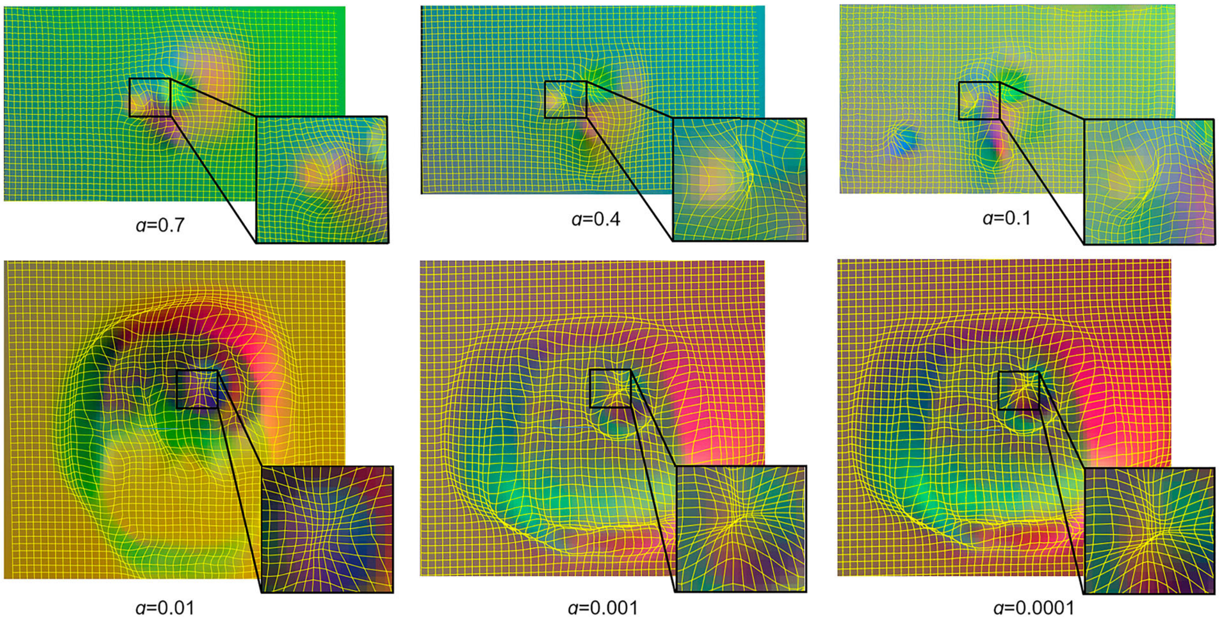

Figure 8. Typical visual DVF for carotid and BraTS datasets, displacement in three spatial dimensions is mapped to RGB color. Inspection of the network predicted DVF with small regularisation weight *a*.

different parameters to compare the quality of the DVF, results are show in Figure 8.

3.2. Limitations

Although our proposed method can achieve good results in experiments, it still has a few limitations. First, it highly depends on the segmentation precision of fixed and moving images to achieve effective performance, and so this method may struggle when encountering imperfect or limited data labels. The second limitation is that it is difficult to ensure the accuracy of the constrained value *R*. Usually, we pre-compute *R* using the traditional affine method without considering each individual instance, and the existence of this error will cause some data to exceed the range of *R*. Additionally, the relatively small datasets used for training will face the risk of overfitting and uncertainty of results. Although the evaluation conducted on our experimental data set demonstrates the advantage of the proposed network in overfitting avoidance, its performance in real application scenarios has to be further investigated.

4. Conclusion

In this paper, we have presented a constrained affine network architecture for 3D multimodal medical image registration. The proposed network consists of two structures. The first one is a deformable registration structure, which generates nonlinear transformation, and the second is a constrained affine structure to synthesize transformation parameters for global alignment. The constrained affine module can be easily transferred to various CNNs. Experimental results indicate that the proposed method exhibits higher accuracy and less computation time. In summary, we have developed a new medical image registration method which has immense potential for multi-model clinical applications.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by Natural Science Foundation of Zhejiang Province under grants LQ20H160052, LY18F030019, LY18F020030 and the National Natural Science Foundation of China under grant 11302195, U1908210, 61401397ïijŇResearch Program of De-partment of Science and Technology of Zhejiang Province under grant LGF19H180019, Research Program of Science and Technology Plan of health commission of Zhejiang province under grant 2019326619.

References

[1] Ali FE, El-Dokany IM, Saad AA, et al. A curvelet transform approach for the fusion of MR and CT images. J Mod Optic. 2010;57(4):273–286.

[2] Fu Y, Lei Y, Wang T, et al. Deep learning in medical image registration: a review. Phys Med Biol. 2020;65(20):20TR01.

[3] Sotiras A, Davatzikos C, Paragios N. Deformable medical image registration: a survey. IEEE T Med Imaging. 2013;32(7):1153–1190.

[4] Fan J, Cao X, Yap PT, et al. BIRNet: brain image registration using dual-supervised fully convolutional networks. Med Image Anal. 2019;54:193–206.

[5] Avants BB, Epstein CL, Grossman M, et al. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. Med Image Anal. 2008;12(1):26–41.

[6] Lorenzi M, Ayache N, Frisoni GB, et al. LCC-Demons: a robust and accurate symmetric diffeomorphic registration algorithm. NeuroImage. 2013;81:470–483.

[7] Eppenhof KAJ, Pluim JPW. Pulmonary. CT registration through supervised learning with convolutional neural networks. IEEE T Med Imaging. 2019;38(5):1097–1105.

[8] Eppenhof KAJ, Lafarge MW, Moeskops P, et al. Deforma ble image registration using convolutional neural networks. Med Imag: Image Process. 2018;10574:105740s.

[9] Jaderberg M, Simonyan K, Zisserman A, et al. Spatial transformer networks. 2015; arxiv Prepr arXiv:1506.02025.

[10] Balakrishnan G, Zhao A, Sabuncu MR, et al. Voxelmorph: a learning framework for deformable medical image registration. IEEE T Med Imaging. 2019;38(8):1788–1800.

[11] Balakrishnan G, Zhao A, Sabuncu MR, et al. An unsupervised learning model for deformable medical image registration. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2018 Jun 19–27; Salt Lake, UT. p. 9252–9260.

[12] Dalca AV, Balakrishnan G, Guttag J, et al. Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces. Med Image Anal. 2019;57:226–236.

[13] Zhao S, Lau T, Luo J, et al. Unsupervised 3D end-to-end medical image registration with volume tweening network. IEEE J Biomed Health Inform. 2020;24(5):1394–1404.

[14] Zhao S, Dong Y, Chang EI, et al. Recursive cascaded networks for unsupervised medical image registration. Proceedings of the IEEE International Conference on Computer Vision; 2019 Jun 16–20; Seoul, Korea. p. 10600–10610.

[15] Isola P, Zhu J, Zhou T, et al. Image-to-image translation with conditional adversarial networks. Proceedings of the IEEE International Conference on Computer Vision; 2019 June 16–20; California. p. 1125–1134.

[16] Hu Y, Modat M, Gibson E, et al. Label-driven weakly-supervised learning for multimodal deformable image registration. Proceedings of the IEEE International Symposium on Biomedical Imaging; 2018 April 4–7; Washington, DC, p. 1070–1074.

[17] Hu Y, Modat M, Gibson E, et al. Weakly-supervised convolutional neural networks for multimodal image registration. Med Image Anal. 2018;49:1–13.

[18] Isola P, Zhu J, Zhou T, et al. Image-to-image translation with conditional adversarial networks. Proceedings of the IEEE International Conference on Computer Vision; 2017 July 21–26; Honolulu, Hawaii. p. 1125–1134.

[19] Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. Proceedings of the Medical Image Computing and Computer-Assisted Intervention; 2015 Oct 5–9; Munich: Spring Press; 2017. p. 224–241.

[20] Drozdzal M, Vorontsov E, Chartrand G, et al. The importance of skip connections in biomedical image segmentation. 2016; arxiv Prepr arXiv:1608.04117.

[21] Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. 2016; arXiv Prepr arXiv:1502.03167.

[22] Jenkinson M, Smith SJM. A global optimisation method for robust affine registration of brain images. Med Image Anal. 2001;5(2):143–156.

[23] Rueckert D, Sonoda LI, Hayes C, et al. Nonrigid registration using free-form deformations: application to breast MR images. IEEE T Med Imaging. 1999;18(8):712–721.

[24] Menze BH, Jakab A, Bauer S, et al. The multimodal brain tumor image segmentation benchmark (BRATS). IEEE T Med Imaging. 2014;34(10):1993–2024.

[25] Bakas S, Akbari H, Sotiras A, et al. Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. Sci Data. 2017;4:170117.

[26] Bakas S, Reyes M, Jakab A, et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. 2018; arXiv Prepr arXiv:1811.02629.

[27] Tustison NJ, Avants BB, Cook PA, et al. N4TK: improved N3 bias correction. IEEE T Med Imaging. 2010;29(6):1310–1320.

[28] Qin C, Shi B, Liao R, et al. Unsupervised deformable registration for multi modal images via disentangled representations. Inf Process Med Imaging. 2019;11492:249–261.

[29] Çiçek Ö, Abdulkadir A, Lienkamp SS, et al. 3D U-Net: learning dense volumetric segmentation from sparse annotation. Proceedings of the Medical Image Computing and Computer-Assisted Intervention; 2016 Oct 17–21; Athens: Spring Press; 2016. p. 424–432.

[30] Abadi M, Agarwal A, Barham P, et al. Tensorflow: large-scale machine learning on heterogeneous distributed systems. 2018; arXiv Prepr arXiv:1603.04467.

[31] Kingma DP, Ba J. Adam: a method for stochastic optimization. 2014; arXiv Prepr arXiv:1412.6980.

[32] Oktay O, Schlempe J, Folgoc LL, et al. Attention u-net: learning where to look for the pancreas. 2018; arXiv arXiv:1804.03999.

[33] Ibtehaz N, Rahman MS. MultiResUNet: rethinking the U-net architecture for multimodal biomedical image segmentation. Neural Netw. 2020;121:74–87.