Neural Computing and Applications (2022) 34:12663–12679
https://doi.org/10.1007/s00521-022-07113-8

ORIGINAL ARTICLE

GPLFR—Global perspective and local flow registration-for forward-looking sonar images

Peng Huang[1] · Chunsheng Guo[1] · Xingbing Fu[2] · Lingyun Xu[3] · Di Zhou[4]

Received: 8 March 2021 / Accepted: 17 February 2022 / Published online: 16 March 2022

Abstract

Forward-looking sonar (FLS) image registration is a key step in many underwater applications such as underwater target detection, ocean observation, and mapping. However, low resolution, low signal-to-noise ratio, and the complex nonlinear transformation relationship between FLS images from two different viewpoints have brought great challenges to register them. In order to better cope with this challenge, we propose a global perspective and local flow registration (GPLFR) method for FLS images. GPLFR consists of two networks, i.e., a regression correction network (RCNet) and a deformable network (IRRDNet) with the iterative refinement of the residual. For a given pair of FLS images, RCNet is used to estimate the global transformation parameters to achieve global registration, and then, IRRDNet is used to estimate the deformation field or flow field to realize local alignment. The experimental results on real FLS image and 2D face expression image registration tasks demonstrate the effectiveness and robustness of the proposed method.

Keywords Image registration · Forward-looking sonar · Regression correction network · Deformable network · Deep learning

1 Introduction

Image registration is the process of aligning two images acquired from different time, different angles, and different sensors in the same scene [1]. Image registration has a wide range of applications in image mosaic and fusion [2, 3] and image change detection [4, 5].

Image registration methods can be divided into two categories: One is sparse feature-based registration methods [6–9], and the other is dense deformation field-based registration methods [10–12].

The sparse feature-based registration method fits the global transformation model through the steps of feature extraction and feature matching. The method needs to extract enough salient features and select a suitable transformation model to achieve a good global registration of the image.

The dense deformation field-based registration method constructs a dense deformation field by optimizing the energy function, thereby establishing a dense nonlinear correspondence between image pairs to achieve local alignment. A large proportion of the registration method research is medical images. Usually, due to the local deformation of the patient (due to breathing, anatomical changes, etc.), the transformation between two medical images cannot be simply described by a homography matrix. More complex transformation models are needed. For example, Balakrishnan et al. [10] proposed VoxelMorph (VM), which uses the similarity function between the warped image and the fixed image to predict the displacement vector field (DVF), which can better deal with the local deformation between medical image pairs.

✉ Chunsheng Guo
gcs@hdu.edu.cn

[1] School of Communication Engineering, Hangzhou Dianzi University, Hangzhou 310018, China

[2] School of Cyberspace, Hangzhou Dianzi University, Hangzhou 310018, China

[3] Nanjing Institute of Metrological Supervision and Testing, Nanjing 210049, China

[4] Zhejiang Uniview Technologies Co. Ltd., Hangzhou 310051, China

FLS is one of the most important tools for ocean exploration, and it can play an important role in underwater operations in turbid environments.

In the FLS imaging system, the transmitting array transmits pulse signals forward or in the vertical direction in a sector, while the receiving array receives echo signals and obtains the sonar image according to the time and intensity of the echo arrival. FLS images have been applied to many underwater applications such as submarine target detection [13], target tracking [14, 15], path planning [16]. But most of these applications need to rely on a very important step: FLS image registration. However, due to the special characteristics of the FLS image, it brings severe challenges to the FLS image registration. On the one hand, there is a complex nonlinear transformation relationship between FLS images from two different viewpoints (details of the model certification can be found in Appendix); traditional FLS image registration methods [6, 17–20] usually only use an affine transformation model to realize sonar image registration.

However, this model exhibits multiple limitations. First, the affine transformation model is a global transformation model, which ignores the local deformation between the FLS image pairs, and cannot obtain the accurate correspondence between the FLS image pairs. When the FLS has a large elevation angle and large deformation between image pairs, the simple global transformation model will fail. Second, the affine transformation is a linear transformation, and the global registration performance may be limited. Then, Aykin et al. [21] and Sekkati et al. [22] tried to estimate the elevation angle information of each position of the image so that it can use an accurate transformation model for registration to improve the registration accuracy. But this depends on the accuracy of the elevation angle estimation, which can be a very complicated and difficult process [23]. On the other hand, due to the low resolution, low signal-to-noise ratio, and low pixel-level feature repeatability of FLS images [24, 25], it also brings great challenges to the traditional registration methods of the sparse feature-based and the dense deformation field-based. For the sparse feature-based registration methods [6–8], firstly, the method can extracts fewer salient features and cannot build a complex transformation model. Then the feature matching is unstable and it is easy to produce wrong matching relations [17]. When processing FLS images that are far away in time or space, the difficulty of feature extraction and matching will be further increased. Finally, this method needs to assume that the image pairs follow a global transformation. If there is a local deformation in the real transformation, the method will fail. For the dense deformation field-based registration methods, firstly, due to the strong noise of the FLS and the unobvious texture feature, it becomes difficult to construct the dense deformation field. Secondly, when the image pair has a large displacement and a large appearance change, the registration performance is poor [26]. Therefore, in order to better solve the difficulties in the registration of FLS images, it is natural to consider a hybrid approach which combining the benefits of these two methods.

In this paper, we propose GPLFR method, which consists of a RCNet and a IRRDNet. Among them, the RCNet is used to realize the global registration of the FLS image, which consists of two parts: the regression network and the correction network. The regression network is used to directly estimate model transformation parameters, without feature extraction and feature matching. The correction network includes a geometric transformation network and a comparison network. The geometric transformation network obtains the warped image through three-dimensional rotation and translation transformation, pinhole camera projection, and bilinear interpolation according to the transformation parameters obtained by the regression network. The comparison network updates the transformation model parameters according to the similarity between the warped image and the reference image to estimate the best transformation parameters. Since the regression correction transformation model is nonlinear, it can roughly fit the complex nonlinear transformation relationship between FLS images of two different viewpoints. Compared with the linear affine transformation, it can provide a more accurate global registration effect. The IRRDNet is used to finely fit the complex nonlinear transformation relationship between FLS images to achieve local alignment. This method is based on local flow registration method, combined with the idea of iterative residual refinement (IRR) [27], takes the output of the previous iteration as input, and residually refines the deformation field estimated in the previous stage by repetitively using the same network to yield better accuracy without increasing the network size.

All in all, in order to better deal with the challenge of FLS image registration, we propose the GPLFR method, which consists of two networks: RCNet and IRRDNet. Among them, RCNet is based on the regression correction transformation model to achieve global registration. Compared with the linear affine transformation model, RCNet can provide more accurate global registration performance. IRRDNet introduces the idea of IRR on the basis of the local flow registration network, which can further improve the performance of local alignment without increasing the network parameters.

2 Related works

Traditional image registration can be divided into feature-based registration methods and region-based registration methods [1, 28]. The feature-based registration method uses SIFT [29], SURF [30], ORB [31], and other methods to extract the salient features of a pair of images, including points, lines, edges, and contours. Next, select the appropriate similarity measure function to perform feature matching. Then use random sample consensus (RANSAC) [32] or its variants [33, 34] to eliminate mismatched points and estimate a more accurate transformation matrix. Finally, the moving image is warped to the reference image according to the transformation matrix. For example, Negahdaripour et al. [6] used Harris detectors to detect corner features of sonar images to achieve image registration. Tao et al. [35] used the improved SURF [30] algorithm to extract and match feature points to achieve sonar image matching. Li et al. [7] used SIFT [29] to perform feature matching on sonar images and introduced RANSAC [32] to eliminate mismatched points to improve the registration accuracy of sonar images. The region-based image registration method does not need to extract and match features, but iteratively optimizes the similarity between image pairs to perform registration. The commonly used similarity measurement methods are mean square error (MSE) [36],mutual information [37–39], cross-correlation [40–42], and phase correlation [17, 19]. For example, Gai et al. [36] achieved image registration by minimizing the MSE between the fixed image and the moving image to find the best transformation parameters. Sarvaiya et al. [41] and Briechle et al. [40] perform image registration by maximizing the normalized cross-correlation between image pairs. Song et al. [43] estimate the transformation parameters by maximizing mutual information for sonar image registration. Hurtós et al. [17, 19] propose a Fourier-based technique to perform 2D FLS image registration, because sonar images have the characteristics of low resolution, low signal-to-noise ratio, and obvious viewpoint changes. Using traditional methods to extract the features of sonar is very challenging [44].

With the rapid development of deep learning, breakthroughs have been made in the field of computer vision. Image registration methods based on deep learning are also emerging. Sarnel et al. [45] proposed the use of a radial basis neural networks to estimate transformation parameters to improve the accuracy of image registration. Valdenegro-Toro [44] uses CNN to learn a matching function to match sonar image blocks, and the matching effect is better than traditional feature point matching methods. Ot et al. [46] proposed SMNet to determine whether a pair of sonar images are in the same scene. Yang et al. [9] used the powerful feature extraction capabilities of CNN to obtain robust multi-scale feature descriptors to perform image registration. Cheng et al. [47] use deep learning methods to learn the similarity between image pairs, which is more accurate and robust than traditional similarity measures. DeTone et al. [48] used CNN to directly estimate homography transformation parameters to achieve image registration.

In the field of medical image registration, a large number of image registration methods based on deep learning have emerged, which can be divided into supervised and unsupervised image registration methods. In the supervised image registration method, the real transformation parameters need to be obtained as labels to train the network. Chee et al. [49] proposed AIRNet, which directly estimates the transformation parameters of the two input images to achieve affine image registration. In order to deal with complex nonlinear transformations, Sokooti et al. [50] proposed RegNet, which uses CNN to directly predict the DVF from a pair of input images, and uses a large number of artificially synthesized smooth DVF to train the network. In order to better predict DVF with large displacement, kooti et al. [51] proposed a multistage supervision deformable image registration method based on the RegNet [50]. Although the direct registration method with supervised learning has achieved some success, it is still very difficult to obtain labels with real transformation parameters [52].

Inspired by the STN proposed by Jaderberg et al. [53], researchers have proposed image registration method based on unsupervised learning. De Vos et al [12] first proposed unsupervised deep learning methods to achieve end-to-end deformable image registration. ConvNet regressor was used to generate local deformable parameters and STN was used to resample moving images to generate warped images. The network is trained by the similarity between the warped image and the fixed image. Balakrishnan et al. [10] proposed VM, which uses an encoder–decoder network structure similar to U-Net [54] to predict DVF. VM needs to assume that the input image pair has been pre-aligned, and when there is a large displacement between the images, the registration effect is not good [26]. Zhao et al . [11] achieved image registration by recursive cascading. The moving images are sequentially warped through each cascade network and finally aligned with the reference image. Zou et al [55] proposed a new unsupervised registration method, that is, based on the recursive cascaded registration network, introducing the global and local information of anatomical segmentation to improve the registration accuracy.

3 Methods

The GPLFR model proposed in this paper is shown in Fig. 1. GPLFR consists of RCNet and IRRDNet. The RCNet is used to achieve global registration, and the IRRDNet is used to achieve local alignment. The following introduces the RCNet and the IRRDNet.

3.1 RCNet

The RCNet consists of the regression network and the correction network, as shown in Fig. 2. The regression network takes the sonar image and the line feature and coordinate information of the image as input to obtain the transformation model parameters. The correction network includes geometric transformation network and comparison network. The geometric transformation network obtains the warped image through three-dimensional rotation and translation transformation, pinhole camera projection, and bilinear interpolation according to the transformation parameters obtained by the regression network. The comparison network updates the transformation model parameters according to the similarity between the warped image and the reference image to estimate the best transformation parameters. The regression network and the correction network will be described in detail below.

3.1.1 Regression network

The network structure of the regression network is shown on the left side of Fig. 2. The input of the regression network is 6 channels, including two channels of moving image and fixed image, two linear feature channels, and two coordinate channels. The output of the regression network is a 10-dimensional vector: three rotation and three translation parameters required for three-dimensional rotation and translation transformation, two translation parameters, and two scaling parameters for the projection transformation. The regression network is inspired from the VGG-16 network [56], which is composed of 6 convolutional blocks and 5 fully connected blocks. Each convolutional block is composed of a convolutional layer, a ReLU layer, and a max-pooling layer. The fully connected block is composed of a fully connected layer and a ReLU layer. The last fully connected block contains only one fully connected layer. The size of the convolution kernel in the convolution layer is 5, the stride of the convolutional kernel is 1, and the paddings on both sides are 2. The kernel size of the fourth pooling layer is 3, the kernel size of the remaining pooling layer is 2, and the stride of all pooling layer is 2.

Since the pooling layer processing causes the loss of spatial coordinate information, the network uses more fully connected layers. The sonar image has relatively weak texture due to high noise, but the line features are obvious. In order to strengthen the function of the line feature, the linear detection of the moving image and the fixed image was performed by the LSD [57] algorithm to obtain two line feature channels and input to the regression network. At the same time, in order for the network to learn complete translation invariance or a certain degree of translation dependence, the method of CoordConv [58] is used for

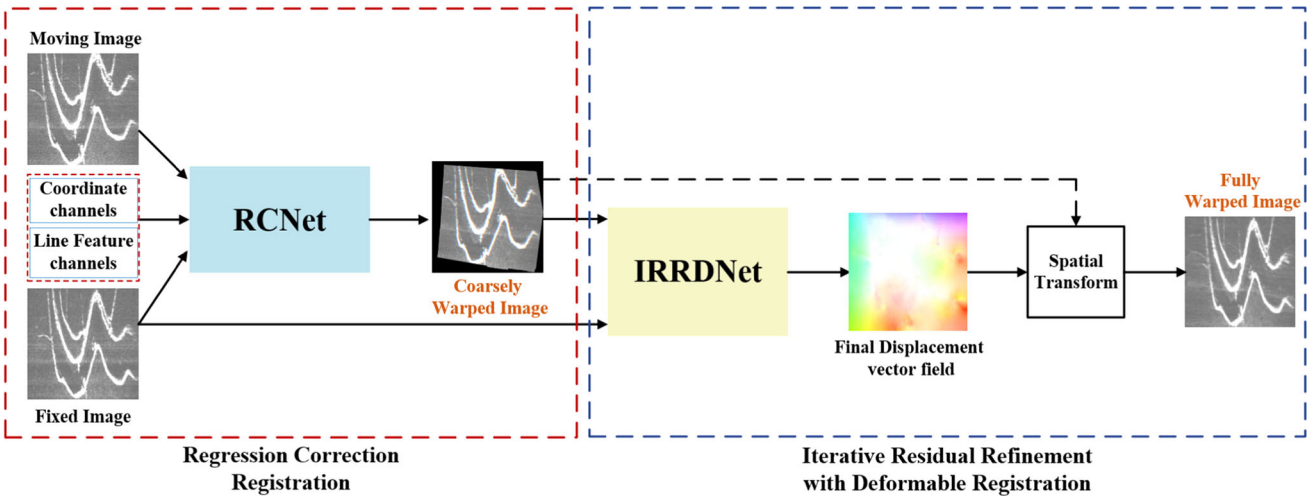

Fig. 1 The GPLFR model consists of regression-corrected registration (left) and iterative residual refinement with deformable registration (right). The input of the RCNet is 6 channels, including two channels of moving image and fixed image, two linear feature channels, and two coordinate channels, and the output is a coarsely warped image. Then feed the coarsely warped image and the fixed image into the IRRDNet to obtain the final DVF. The STN warps the coarsely warped image according to the estimated DVF to obtain the final warped image

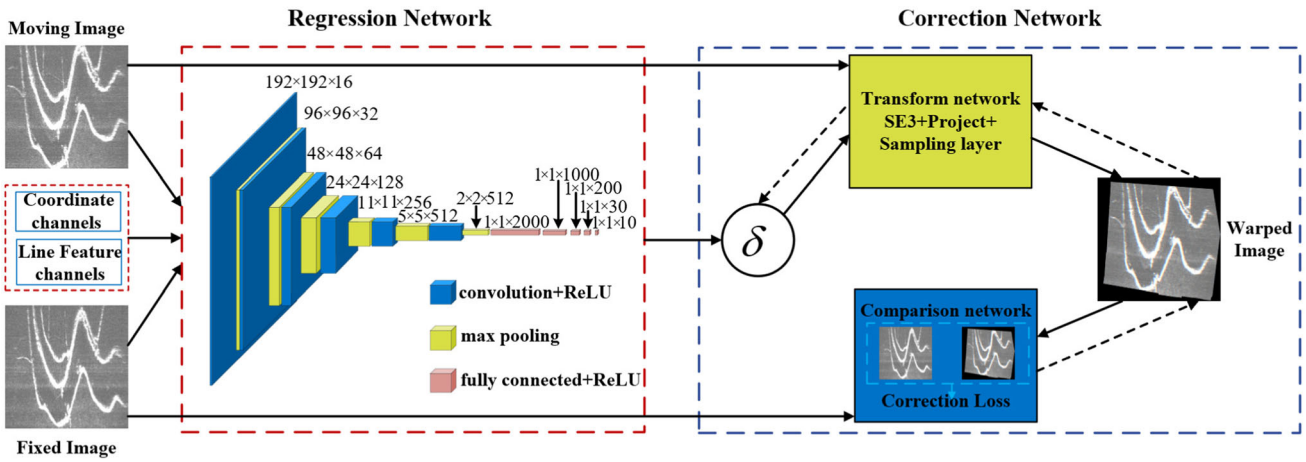

Fig. 2 Architecture of RCNet. First, the regression network estimates the transformation parameters; then, the correction network warps the moving image through the transform network and uses the similarity between the warped image and the fixed image to correct the parameters to obtain the final warped image

reference, and the x and y coordinate information of the reference image is fed into the regression network.

3.1.2 Correction network

The correction network structure is shown in Fig. 3, which mainly includes the geometric transformation network [59] and the comparison network. The transformation network includes SE3 layer, projection layer, and bilinear interpolation layer. Feed the transformation parameters obtained by the regression network into the SE3 layer to obtain the 3D rotation and translation matrix, and generate a 3D sampling grid through the 3D grid generator, then the 3D sampling grid is projected to the 2D plane through the projection layer, and the moving image is resampled into a warped image through bilinear interpolation. Finally, through the comparison network, the similarity between the warped image and the reference image is used as the loss function to update the transformation parameters to estimate the best transformation parameters. The SE3 layer and the projection layer will be described in detail below.

(1) *SE3 layer*:

The SE3 layer performs 3D rotation and translation processing, which can be expressed as a 4×4 matrix $\mathbf{T}$, $\mathbf{T} = \begin{bmatrix} \mathbf{R} & \mathbf{t} \\ 0 & 1 \end{bmatrix} \in SE3$. The 3D rotation is represented by a 3×3 matrix $\mathbf{R}$ with three degrees of freedom. The 3D translation is represented by a 3×1 vector $\mathbf{t} = [t_1, t_2, t_3]^T$. Based on the Euler's theorem, the rotation matrix can be described by the axis of rotation and the angle around it (called the angle axis representation). For a three-

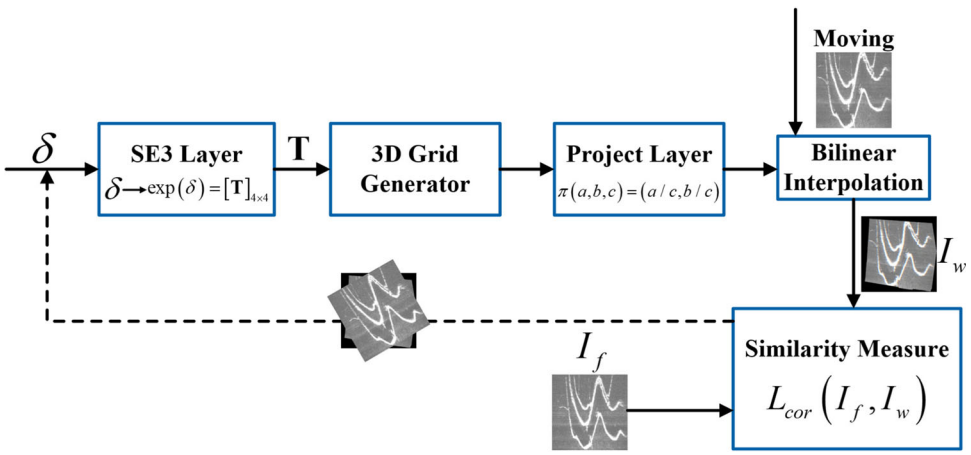

Fig. 3 Architecture of correction network. First, feed the transformation parameters δ generated by the regression network into the SE3 layer, and generate a 3D sampling grid through a 3D grid generator. Then it is projected onto a 2D plane through the projection layer, and the moving image is resampled into a warped image through bilinear interpolation. Finally, the transformation parameters are updated according to the loss of similarity between the warped image and the fixed image

dimensional vector $\mathbf{v}$, by defining unit vector $\hat{\mathbf{v}} = \frac{\mathbf{v}}{\|\mathbf{v}\|_2}$ as the axis of orientation and $\theta = \|\mathbf{v}\|_2$ as the angle of rotation (in radians). The counterclockwise rotation around the axis is defined as the positive direction. The rotation angle is always nonnegative and at most π, that is $\theta \in [0, \pi)$, the rotation matrix is calculated as :

$$\mathbf{R} = \exp\left(\theta[\hat{\mathbf{v}}]_{\mathrm{x}}\right), \tag{1}$$

where $[\hat{\mathbf{v}}]_{\mathrm{x}}$ is the skew-symmetric operator:

$$[\hat{\mathbf{v}}]_{\mathrm{x}} = \begin{bmatrix} 0 & -\hat{v}_3 & \hat{v}_2 \\ \hat{v}_3 & 0 & -\hat{v}_1 \\ -\hat{v}_2 & \hat{v}_1 & 0 \end{bmatrix}. \tag{2}$$

Using Rodriguez's rotation formula, Eq.1 can be simplified as [60]:

$$\mathbf{R} = \mathbf{I}_3 + \sin\theta[\hat{\mathbf{v}}]_{\mathrm{x}} + (1 - \cos\theta)[\hat{\mathbf{v}}]_{\mathrm{x}}^2, \tag{3}$$

where $\mathbf{I}_3$ is the 3×3 identity matrix.

To implement backpropagation of the network layer, it requires calculating the derivative of the network output to the network input. According to the reference [61], the derivative can be expressed as:

$$\begin{aligned} \frac{\partial \mathbf{T}}{\partial v_i} &= \begin{bmatrix} \frac{\partial \mathbf{R}}{\partial v_i} & 0 \\ 0 & 0 \end{bmatrix} \quad i = 1, 2, 3, \\ \frac{\partial \mathbf{T}}{\partial t_i} &= \begin{bmatrix} 0 & \frac{\partial \mathbf{t}}{\partial t_i} \\ 0 & 0 \end{bmatrix} \quad i = 1, 2, 3, \end{aligned} \tag{4}$$

where

$$\begin{aligned} \frac{\partial \mathbf{R}}{\partial v_i} &= \frac{v_i[\mathbf{v}]_{\mathrm{x}} + [\mathbf{v} \times (\mathbf{I}_3 - \mathbf{R})e_i]_{\mathrm{x}}}{\|\mathbf{v}\|_2}\mathbf{R}, \\ \frac{\partial \mathbf{t}}{\partial t_i} &= e_i. \end{aligned} \tag{5}$$

The same as the above $[\ ]_{\mathrm{x}}$ turns a 3×1 vector to a skew-symmetric matrix, e_i is the ith column of the identity matrix. However, since the derivative calculation needs to be divided by the vector norm, the threshold is set to check the vector norm to avoid the overflow of the derivative calculation.

(2) *Projection layer*

The projection layer maps the 3D point $[z_1, z_2, z_3]^T$ onto the 2D point $\mathbf{p} = [x, y]^T$ by the focal length and camera center point position:

$$\begin{bmatrix} x \\ y \end{bmatrix} = \frac{1}{z_3}\begin{bmatrix} f_x & 0 & u_0 \\ 0 & f_y & v_0 \end{bmatrix}\begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix}, \tag{6}$$

where f_x and f_y represent the focal length of the camera along the x and y axes, and u_0 and v_0 represent the camera center point. The derivative of $[z_I, z_2, z_3]^T$ can be expressed as :

$$\frac{\partial \mathbf{p}}{\partial z_i} = \begin{bmatrix} \frac{\partial x}{\partial z_i} \\ \frac{\partial y}{\partial z_i} \end{bmatrix}, \quad i = 1, 2, 3, \tag{7}$$

where

$$\frac{\partial \begin{bmatrix} x \\ y \end{bmatrix}}{\partial \begin{bmatrix} z_1 & z_2 & z_3 \end{bmatrix}^T} = \begin{bmatrix} f_x\frac{1}{z_3} & 0 & -f_x\frac{z_1}{z_3^2} \\ 0 & f_y\frac{1}{z_3} & -f_y\frac{z_2}{z_3^2} \end{bmatrix}. \tag{8}$$

However, since the derivative calculation needs to be divided by z_3, it is necessary to ensure that the value of z_3 is not too small to avoid overflow of the derivative calculation.

3.2 IRRDNet

The deformable registration network is to finely fit the complex nonlinear transformation relationship between the FLS images to achieve local alignment.

Let I_w and I_f denote the warped image obtained by the RCNet and the fixed image. We model a nonlinear function $g_\theta(I_f, I_w) = \phi$ using a CNN, where θ are learnable parameters of g and ϕ is a deformation field (also known as flow field). Then we warp I_w to $I'_{\mathrm{w}} = \phi \circ I_{\mathrm{w}}$ using a STN [53] and use the similarity between I'_w and I_f to update θ.

In order to better refine ϕ and estimate a more accurate correspondence, the deformation field of IRRDNet is calculated as:

$$\phi_k = g_\theta(I_f, \phi_{k-1} \circ I_w) + \phi_{k-1}, \quad k = 1, \ldots, N, \tag{9}$$

where ϕ_k represents the kth iterative and predicts a deformation field, and $I_w^{(k-1)} = \phi_{k-1} \circ I_w$ represents I_w warped by ϕ_{k-1}. N represents the total number of IRR steps. The final warped image is constructed by

$$I_w^{(N)} = \phi_N \circ I_w. \tag{10}$$

The IRRDNet structure is shown in Fig. 4. IRRDNet adopts the network structure in VM, which is similar to the encoding–decoding structure of U-Net. The encoding stage uses strided convolutional layers to extract image features of different scales. In the decoder stage, successive

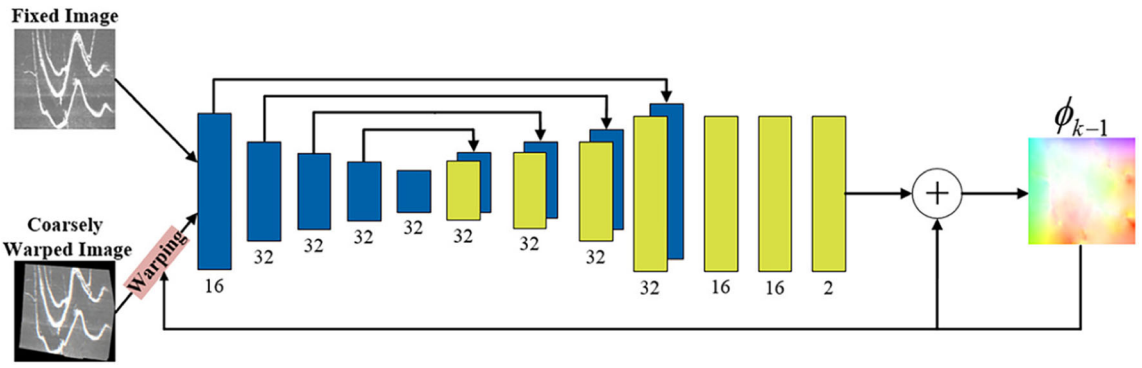

Fig. 4 Architecture of IRRDNet. The IRRDNet residually refines the deformation field estimated in the previous stage by repetitively uses the same network

deconvolution and convolutional layer stacking enable the network to restore low-resolution features to the same resolution as the input image. At the same time, in order to fuse feature information of different scales and reduce the loss of information during the upsampling process, skip connection is used to splice the feature maps obtained through the upsampling during the decoding process and the feature maps obtained through the strided convolutional layer during the encoding process for channel dimension splicing.

The network takes I_w and I_f as input, and the output is the deformation field ϕ_N. Since the input image is a 2D image, the network uses 2D convolution (followed by a LeakyReLU layer with parameter 0.2). The size of the convolution kernel in the encoding stage is 4×4, and the stride is 2, so that the feature map size of each layer output is halved. In the decoding stage, the size of the convolution kernel is 3×3, and the stride is 1.

3.3 Loss function

The following introduces the loss functions of the RCNet and the IRRDNet:

The loss function of the RCNet is as follows:

$$L_r\left(t,\hat{t};I_f,I_w\right) = L_{re}(t,\hat{t}) + L_{cor}\left(I_f,I_w\right), \tag{11}$$

where $L_{re}(t,\hat{t}) = \|t-\hat{t}\|_2$ represents the loss function of the regression network, t represents the ground truth label and $\hat{t}$ represents the predicted value. The term $L_{cor}\left(I_f,I_w\right)$ uses the similarity between the warped image and the fixed image as the loss function of the correction network, and takes MSE and negation of normalized cross-correlation (NCC) in the registration task of sonar image and facial expression image, respectively. MSE is given by:

$$\mathrm{MSE}\left(I_f,I_w\right) = \frac{1}{|\Omega|}\sum_{\mathbf{p}\in\Omega}\left(I_f(\mathbf{p}) - I_w(\mathbf{p})\right)^2 \tag{12}$$

where Ω is the image domain. $\mathbf{p}$ represents the pixel in the image domain. NCC is defined by:

$$NCC\left(I_f,I_w\right) = \sum_{\mathbf{p}\in\Omega}\frac{\sum_{\mathbf{p}_i}\left(I_f(\mathbf{p}_i)-\bar{I}_f(\mathbf{p})\right)\left(I_w(\mathbf{p}_i)-\bar{I}_w(\mathbf{p})\right)}{\sqrt{\sum_{\mathbf{p}_i}\left(I_f(\mathbf{p}_i)-\bar{I}_f(\mathbf{p})\right)^2\sum_{\mathbf{p}_i}\left(I_w(\mathbf{p}_i)-\bar{I}_w(\mathbf{p})\right)^2}} \tag{13}$$

where Ω is the image domain. $\mathbf{p}$ represents the pixel in the image domain, and $\bar{I}_f$ and $\bar{I}_w$ represent the average intensity value of I_f and I_w.

The loss function of the IRRDNet is defined as:

$$L_d\left(I_f,I_w^{(N)}\right) = L_{sin}\left(I_f,I_w^{(N)}\right) + \lambda R(\phi_N) \tag{14}$$

where L_{sim} represents the similarity between the warped image and the fixed image, and MSE and negation of NCC are, respectively, used in the registration task of the sonar image and the facial expression image. $R(\phi_N)$ is a regularization item to generate a smooth ϕ_N and λ is the regularization parameter. $R(\phi_N)$ is calculated as :

$$R(\phi_N) = \sum_{\mathbf{p}\in\Omega}\|\nabla\phi_N(\mathbf{p})\|^2 \tag{15}$$

where $\nabla\phi_N(\mathbf{p}) = \left(\frac{\partial\phi_N(\mathbf{p})}{\partial x},\frac{\partial\phi_N(\mathbf{p})}{\partial y}\right)$, $\frac{\partial\phi_N(\mathbf{p})}{\partial x}$ and $\frac{\partial\phi_N(\mathbf{p})}{\partial y}$ represent the gradients of $\phi_N(\mathbf{p})$ along the x-axis and y-axis, respectively.

4 Experiments and discussion

4.1 Data sets

In order to prove the effectiveness and robustness of our method, we conducted experiments on two data sets. First,

we conduct experiments on real FLS data sets to evaluate the feasibility and effectiveness of the algorithm proposed in this paper. Secondly, to verify the universality and robustness of the proposed method, we conduct experiments on public 2D facial expression image data sets. The two types of data sets are described below.

4.1.1 Realistic FLS image

The FLS data set was collected from Zhanghe Reservoir in Hubei Province and contains 16 pairs of FLS images. The size of the sonar images are 192×192. These images are acquired by the same FLS device at different times and at different angles. They have low resolution, low signal-to-noise ratio, and obvious appearance changes, which brings great difficulties and challenges to image registration.

4.1.2 2D Face expression image

2D facial expression data set comes from the Radboud Faces Database (RaFD) [62]. It contains facial expression images of 67 objects. Each object contains 8 different expressions: angry, contemptuous, disgusted, fearful, happy, neutral, sad, and surprised. And there are three different gaze directions for each facial expression, which are left-gazed, front-gazed, and right-gazed. There are a total of 1608 facial expression images, in which there is a nonlinear transformation relationship between different expression images of the same object. We use the expression images of 53 objects as the training set, 7 objects as the validation set, and the remaining 7 objects as the test set. We crop the image to 640×640, resize it to 128×128, and convert the RGB image to a grayscale image.

4.2 Training details

The training strategies for FLS image and facial expression image registration tasks are as follows:

4.2.1 FLS image registration

The training stage includes two parts:

(1) *Training of RCNet*: We use randomly generated transformation parameters to warp FLS image to generate a training data set. The training data set includes 50,000 pairs of sonar images and ground truth labels have a total of 10 parameters. These parameters are randomly generated and follow a uniform distribution. Among them, the range of six parameters of SE3 is $[-0.3, 0.3]$; the range of two translation parameters of projection transformation is $[-20, 20]$; the range of two scaling parameters of projection transformation is [0.8, 1.2]. We used Adam [63] with a learning rate of $1e^{-5}$ as the optimizer to train the network. Due to the limitation of GPU memory, the batch size during training is set to 1. When the epoch size reaches 80, the model starts to converge.

(2) *Training of IRRDNet*: Due to the small number of real FLS data sets, it is difficult to use a large amount of FLS image to train the network for image registration, so the network adopts the registration method of reference [64]. Image registration can be performed without training the network through a large amount of data. The algorithm of IRRDNet registration is shown in Algorithm 1. First, fix the parameters of the trained RCNet. Then I_w and I_f is fed into the initialized but not trained IRRDNet. The untrained network outputs ϕ_N and then obtains $I_w^{(N)}$. Finally, the loss $L_d\left(I_f, I_w^{(N)}\right)$ is backpropagated to update the parameters in g_θ. Repeat the above process until the maximum number of iterations is reached. We used Adam with learning rate of $1e^{-3}$ as the optimizer to update the network, the parameter λ is set to 1000, numbers of IRR steps N is set to 4, and the maximum number of iterations *iter* is set to 4000.

Algorithm 1 IRRDNet Registration

procedure CNNRGE(I_w, I_f) ▷ Input I_w and I_f
  g_θ =Initialize (IRRDNet)
  while $i <$*iter* **do** ▷ For *iter* number of iterations
    $\phi_k = g_\theta\left(I_f, \phi_{k-1} \circ I_w\right) + \phi_{k-1} \quad k = 1, \ldots, N$ ▷ Predict deformation,ϕ_k
    $I_w^{(N)} = \phi_N \circ I_w$ ▷ Deform Coarsely Warped Image,I_w
    $L_d\left(I_f, I_w^{(N)}\right) = L_{sin}\left(I_f, I_w^{(N)}\right) + \lambda R\left(\phi_N\right)$ ▷ Compute loss
    g_θ =BackPropagate(g_θ, L_d) ▷ Update IRRDNet
    $i = i + 1$
  return $I_w^{(N)}, \phi_N$

4.2.2 Face expression image registration

The training stage includes two parts:

(1) *Training of RCNet:* We use randomly generated transformation parameters to generate 50 synthetic transformed images for each image in the facial expression training set, with a total of 63600 pairs of facial expression images. Considering that the global deformation between the facial expression image pair is very small, the range of six parameters of SE3 is [−0.05,0.05]; the range of two translation parameters of projection transformation is [−1, 1]; two scaling parameters of projection transformation is set to 1. We used Adam [63] with a learning rate of $1e^{-5}$ as the optimizer to train the network; the batch size during training is set to 1; epoch is set to 60.

(2) *Training of IRRDNet:* We use the facial expression training set to train the model, in which the image pairs need to be globally aligned with the trained RCNet, and use Adam [63] with learning rate of $5e^{-6}$ as the optimizer to update the network; the batch size during training is set to 1; epoch is set to 40; the parameter λ is set to 1; numbers of IRR steps N is set to 4.

4.3 Evaluation metrics

On the real FLS image, we use MSE, Pearson's correlation coefficient (PCC) [65], and structural similarity (SSIM) [66] as the evaluation criteria for registration performance.On the face expression image, we use normalized mean square error (NMSE) and SSIM as the evaluation criteria for registration performance. PCC is given by:

$$PCC\left(I_f, I_w^{(N)}\right) = \frac{\sum_{\mathbf{p}\in\Omega}\left(I_f(\mathbf{p})-\bar{I}_f\right)\left(I_w^{(N)}(\mathbf{p})-\bar{I}_w^{(N)}\right)}{\sqrt{\sum_{\mathbf{p}\in\Omega}\left(I_f(\mathbf{p})-\bar{I}_f\right)}\sqrt{\sum_{\mathbf{p}\in\Omega}\left(I_w^{(N)}(\mathbf{p})-\bar{I}_w^{(N)}\right)}} \tag{16}$$

where Ω is the image domain. $\mathbf{p}$ represents the pixel in the image domain, and $\bar{I}_f$ and $\bar{I}_w^{(N)}$ represent the average intensity value of I_f and $I_w^{(N)}$. SSIM is defined by:

$$SSIM\left(I_f, I_w^{(N)}\right) = \frac{\left(2\mu_{I_w^{(N)}}\mu_{I_f}+C_1\right)\left(2\sigma_{I_w^{(N)}I_f}+C_2\right)}{\left(\mu_{I_f}^2+\mu_{I_w^{(N)}}^2+C_1\right)\left(\sigma_{I_f}^2+\sigma_{I_w^{(N)}}^2+C_2\right)}, \tag{17}$$

where $\mu_{I_w^{(N)}}$ and μ_{I_f} and $\sigma_{I_w^{(N)}}$ and σ_{I_f} are means and standard deviations of the images $I_w^{(N)}$ and I_f, respectively. $\sigma_{I_w^{(N)}I_f}$ is the covariance of $I_w^{(N)}$ and I_f. C_1 and C_2 are small constants needed to avoid instability.

MSE and NMSE represent the difference in pixel intensity between the reference image and the warped image. The smaller the value, the smaller the difference in pixel intensity between image pairs. PCC is used to measure the correlation between image pairs, and the value range is between −1 and 1. The value is positive, and the closer the value is to 1, the stronger the correlation between the image pairs. SSIM represents the structural similarity between image pairs, and its value range is between −1 and 1. The larger the value, the more similar the structure between image pairs.

4.4 Qualitative and quantitative evaluation

4.4.1 FLS image registration

In order to evaluate the performance of the algorithm proposed in this paper on the FLS image registration task, we compared two traditional image registration methods: Fourier Merlin Registration Method (FM) [67] and SIFT [29] registration methods and four latest deep learning methods: AIRNet [49], PSAT [68], VM [10], and LapIRN [69]. Among them, AIRNet and PSAT are an affine registration method based on deep learning, and VM and LapIRN is a deformable registration method. Since VM and LapIRN are used for deformable registration, the input image pairs need to be assumed to be globally aligned. To fairly compare, we use our proposed RCNet for global registration, and the registration process is also consistent with registration process of the IRRDNet in this paper.

FM and SIFT are run on a server with AMD Ryzen 5 2600 (6 cores @ 3.40GHz), while all other methods run on a single NVIDIA GTX 1080Ti.

(1) *Qualitative evaluation*: Fig. 5 shows the registration results obtained by different methods. The third to seventh columns, respectively, show the registration results of the RCNet and the other four global registration methods. In order to better judge the effect of global registration, the visualized image of the warped image and the fixed image overlapped is given. The results show that RCNet obtains a better global registration effect than the other four methods in most cases, especially when the global transformation is more complicated. The reason is that the convolutional network can extract more robust features and RCNet has a certain nonlinear fitting ability. FM achieves registration through phase correlation technology and logarithmic polar transformation technology. This method has good registration performance for small scaling factor and rotation angle transformation. However, the registration effect is poor when the scaling factor and rotation angle are large.

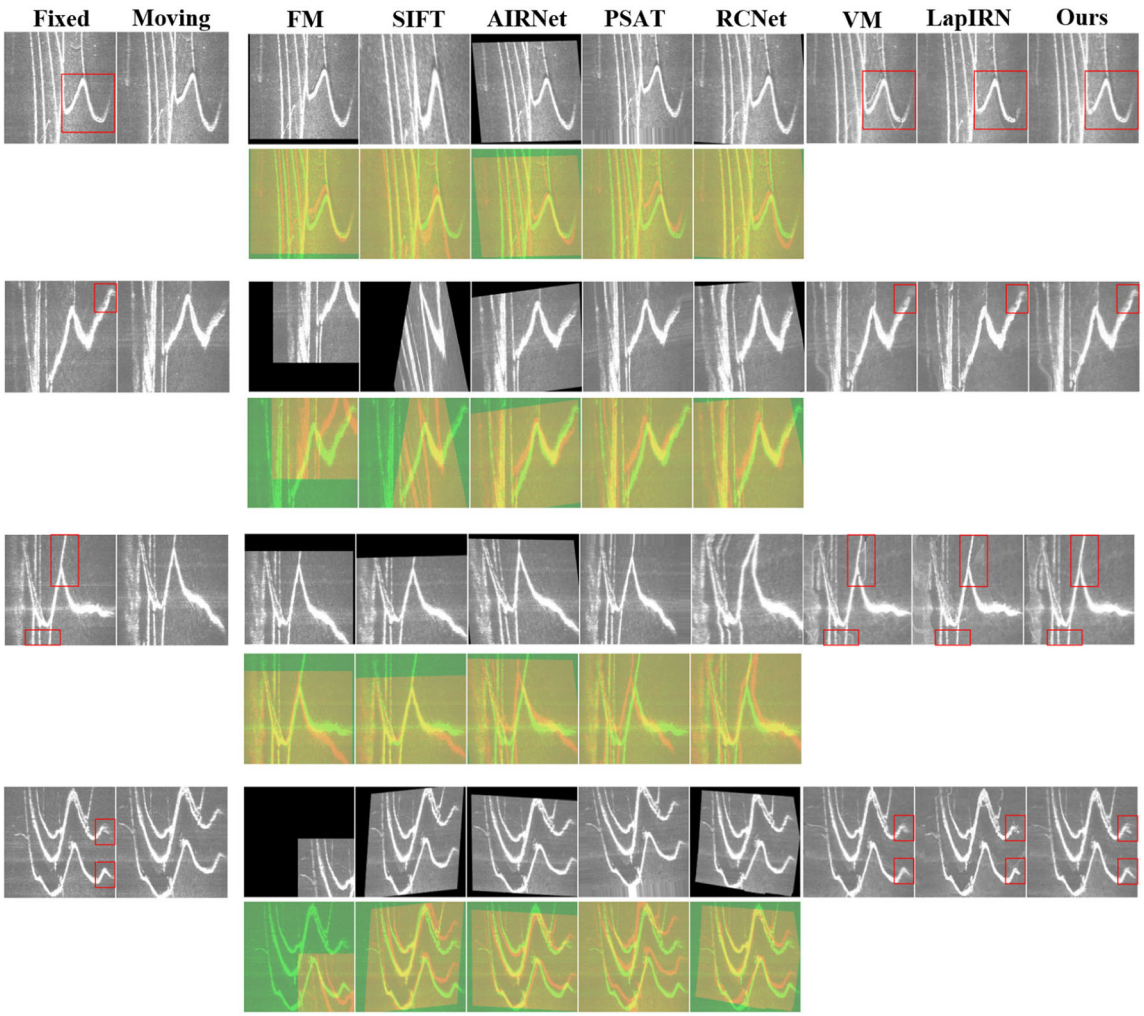

Fig. 5 Registration results of different methods on FLS image. Columns 3–7 use overlapping visualizations to judge the effect of image alignment. The first column and the 8th–10th column, respectively, use red rectangles to circle the areas where the difference between the fixed image and the warped image is obvious

The registration effect of the SIFT method is also poor in most cases. The reason is that although the SIFT method can achieve registration by extracting scale invariant features, it is very sensitive to mismatched points and easily affects registration results when there are outliers. AIRNet directly predicts the affine transformation matrix through the network to achieve affine transformation, while PAST formulates affine transformation by learning specific geometric transformation parameters (e.g., translations, rotation, scaling, and shearing). In most cases, the registration effect of these two registration methods are slightly better than the first two traditional image registration methods, but due to its linear transformation characteristics, the global registration performance is limited.

The eighth and tenth columns, respectively, represent the fine registration results obtained by the deformable registration method after coarse registration. Among them, VM directly estimates the deformation field to achieve deformable registration, and the LapIRN method uses Laplacian pyramid network to achieve deformable

Table 1 Quantitative results of different registration methods on FLS image

Method	Category	MSE	PCC	SSIM	Time(s)
Before registration	–	79.11	0.28	0.24	–
FM	Global	76.69	0.29	0.26	0.02
SIFT	Global	75.89	0.30	0.25	0.29
AIRNet	Global	74.66	0.32	0.27	0.04
PSAT	Global	77.43	0.33	0.29	**0.01**
RCNet	Global	74.63	0.38	0.31	0.03
VM	Local	61.07	0.90	0.62	42.58
LapIRN	Local	59.26	0.89	0.63	207.25
Ours	Local	**57.15**	**0.92**	**0.65**	113.79

registration. In order to better demonstrate the registration effect, the area where the difference between the fixed image and the warped image is obvious is circled with a red rectangular frame. It can be seen that VM, LapIRN, and the method proposed in this paper can register sonar images well. VM directly estimates the deformation field, and the registration effect is not good. LapIRN uses Laplace pyramid network to refine the deformation field, which can better improve the registration performance. However, its registration performance is limited because the flow field estimator in LapIRN is too simple. In this paper, we residually refine the deformation field estimated in the previous stage by repetitively using the same network, which can refine the deformation field better, and a better registration effect is obtained.

(2) *Quantitative evaluation*: In order to quantitatively compare the registration effects of different methods, Table 1, respectively, shows the average MSE, PCC, SSIM between the warped image and the fixed image, and the corresponding registration time. In terms of MSE, PCC, and SSIM, it can be seen that the method proposed in this paper is the best: The average MSE is about 4% smaller than LapIRN; the average PCC is about 3% higher than LapIRN; the average SSIM is about 3% higher than LapIRN; the average MSE is about 6% smaller than VM; the average PCC is about 2% higher than VM; the average SSIM is about 5% higher than VM. In the coarse registration stage, RCNet has the best registration effect: The average MSE is about 4% smaller than PSAT and the average MSE is about 3% smaller than FM and 2% smaller than SIFT and close to AIRNet; the average PCC is about 31% higher than FM and 27% higher than SIFT and 19% higher than AIRNet and 15% higher than PSAT; the average SSIM is about 19% higher than FM and 24% higher than SIFT and 15% higher than AIRNet and 7% higher than PSAT; the results demonstrate RCNet has a better effect in roughly fitting the nonlinear transformation relationship between the sonar image pairs, and obtains a better global registration effect. In terms of registration time, PSAT has the fastest registration speed in global registration. In local alignment, the registration speed is relatively slow. The reason is that we adopted the registration strategy of reference [64]. This strategy requires a certain number of iterations to achieve the best registration effect, and this process takes some time. In future research, we will study better registration strategies to improve registration speed.

To further evaluate the performance of the method proposed in this paper, Fig. 6 shows the histograms of the average MSE, PCC, and SSIM between the warped image and the fixed image using different registration methods. It can be seen that after the global registration, the deformable registration method can fit the nonlinear transformation relationship between the FLS image pairs well, and the registration performance has been greatly improved, which demonstrates the feasibility of the deformable registration method in the FLS image. In addition, in order to compare the impact of different numbers of IIR steps on the registration performance, Table 2 gives the registration results of different numbers of IRR steps. Note that registration performance continues to improve as numbers of IRR steps increase, because residually refine the deformation field estimated in the previous stage by repetitively uses the same network, a better registration effect is obtained without increasing the network size. When the number of IRR steps is 4, the registration performance begins to stabilize. Therefore, the number of IRR steps of the final IRRDNet of this paper is selected 4. Fig. 7 plots the registration results of different numbers of IRR steps, which can better illustrate the growth trend.

4.4.2 Face expression image registration

In order to evaluate the performance of our method in facial expression image registration task, we compare it with the two latest deep learning-based registration

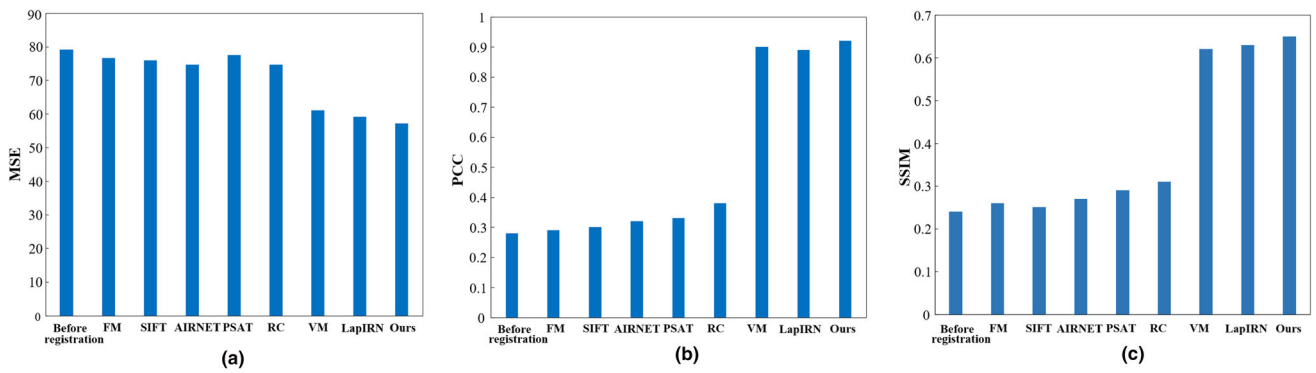

Fig. 6 Histograms of registration results of different registration methods on FLS image **a** MSE, **b** PCC, **c** SSIM. **a** Lower is better. **b** and **c** Higher is better

Table 2 Registration results of different numbers of IRR steps on FLS image

Iteration num	MSE	PCC	SSIM
N = 1	61.07	0.90	0.62
N = 2	58.68	0.91	0.64
N = 3	58.36	0.91	0.64
N = 4	57.15	0.92	0.65
N = 5	57.14	0.92	0.65
N = 6	57.06	0.92	0.65

methods: VM [10] and DiffuseMorph [70]. On the test set, we, respectively, deform different facial expression images of the same person gazing in the same direction.

(1) *Qualitative evaluation:* We respectively, deform different facial expression images of the same person gazing in the same direction on the test set. Figure 8 shows the registration results of the 2D facial expression images, in order to better demonstrate the registration effect. The area where the difference between the fixed image and the warped image is obvious is pointed out with a red arrow. It can be seen that the warped image obtained by the method in this paper is more similar to the fixed image, especially in the mouth area.

(2) *Quantitative evaluation:* In order to quantitatively analyze the registration results, we also deform different facial expression images of the same person gazing in the same direction on the test set. Table 3, respectively, shows the registration results of the 2D facial expression images of the three gaze directions and the average registration results of the three gaze directions. It can be seen that our method has achieved the best registration results, reducing the average NMSE by 0.0247 and increasing the average SSIM by 0.209, and the average NMSE is about 47% lower than the VM, about 41% lower than the DiffuseMorph, and the average SSIM is about 3% higher than the VM, and about 2% higher than the DiffuseMorph.

Fig. 8 Visualization of registration results of different methods in facial expression image. From top to bottom, the results are deformed from the left-gazed disgusted to the left-gazed happy images, from the front-gazed neutral to the front-gazed happy images, from the right-gazed surprised to the right-gazed sad images, from the front-gazed neutral to the front-gazed contemptuous images, from the left-gazed disgusted to the left-gazed neutral images, from the right-gazed fearful to the right-gazed angry images, from the left-gazed neutral to the left-gazed sad images ▶

4.5 Discussion

On the FLS data set, compared with the comparison algorithm, the proposed GPLFR method achieves the best registration effect on indicators such as MSE, PCC, and SSIM. Moreover, the proposed GPLFR can better solve the complex nonlinear transformation relationship between FLS images from different viewpoints and cleverly cope with the challenges of low resolution, low signal-to-noise ratio, and unobvious features of FLS images. Specifically, in the global registration, the RCNet model is nonlinear. Therefore, it can roughly fit the complex nonlinear transformation relationship between FLS images from different viewpoints. The results show that a better global registration effect can be obtained than the linear affine transformation model. In the local alignment, IRRDNet can finely fit the complex nonlinear transformation relationship between FLS images. On the one hand, the experiment results prove the feasibility of the local flow registration method in the FLS image. On the other hand, they also demonstrate that the local flow registration method combined with IRR can yield better accuracy without increasing the network size.

The limitation of this work is that in the local alignment, the IRRDNet registration time is slightly longer on FLS image registration task. As the numbers of IRR steps

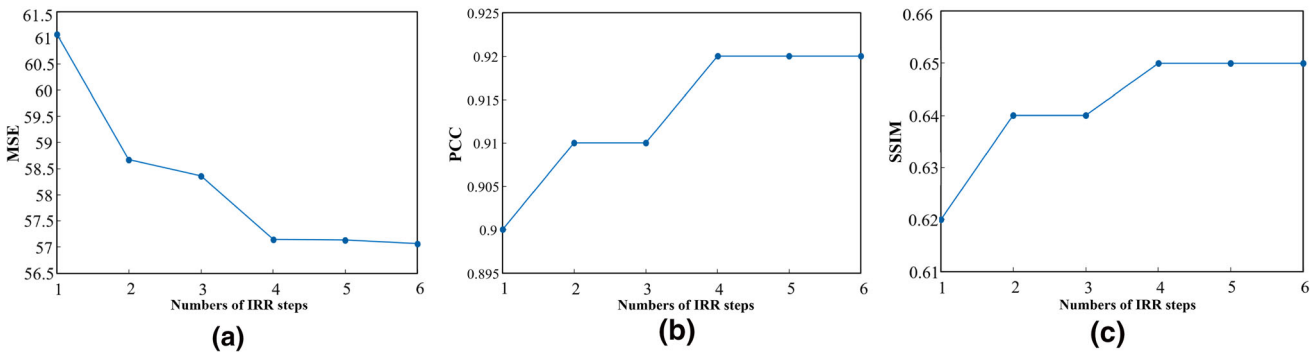

Fig. 7 The registration results of different numbers of IRR steps on FLS image **a** MSE, **b** PCC, **c** SSIM. **a** Lower is better. **b** and **c** Higher is better


Moving
VM
DiffuseMorph
Ours
Fixed

Table 3 Quantitative results of different registration methods on facial expression image

Method	Front-gazed		Left-gazed		Right-gazed		Average	
	NMSE $\times 10^{-1}$	SSIM	NMSE $\times 10^{-1}$	SSIM	NMSE $\times 10^{-1}$	SSIM	NMSE $\times 10^{-1}$	SSIM
Initial	0.250	0.709	0.266	0.707	0.275	0.701	0.264	0.706
VM	0.028	0.891	0.037	0.886	0.031	0.887	0.032	0.888
DiffuseMorph	0.027	0.898	0.030	0.894	0.029	0.894	0.029	0.895
Ours	0.016	0.917	0.018	0.914	0.017	0.914	**0.017**	**0.915**

Bold values indicate that our method achieves the best results on average NMSE and SSIM among the three methods

increase, the time required for image registration will become longer, and the performance requirements of the GPU will be higher. In future research, we will study better registration strategies to improve registration speed and explore better registration methods to enhance the accuracy and robustness of the registration on FLS image registration task.

5 Conclusion

We propose a GPLFR method for FLS image registration. The method consists of a RCNet for global registration and a IRRDNet for local registration. The results on the real FLS data set show the effectiveness the method, which can better deal with the challenges of FLS image registration.In addition, the experimental results on facial expression images verify the robustness of the method.

Appendix

In the sonar-based spherical coordinate system, the projection model of the sonar is shown in Fig. 9. The FLS projects the 3D point $\mathbf{P}$ in the scene onto the 2D image plane, and the projection point is $\mathbf{P}_s$. The point $\mathbf{P}$ is defined as:

$$\mathbf{P} = \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} R\cos\varphi\sin\theta \\ R\cos\varphi\cos\theta \\ R\sin\varphi \end{pmatrix}, \tag{18}$$

$$\begin{bmatrix} R \\ \theta \\ \varphi \end{bmatrix} = \begin{bmatrix} \sqrt{x^2+y^2+z^2} \\ \tan^{-1}(x/y) \\ \tan^{-1}\left(z/\sqrt{x^2+y^2}\right) \end{bmatrix}, \tag{19}$$

where $[x \quad y \quad z]^T$ is the Cartesian coordinates of point $\mathbf{P}$ and $[R \quad \theta \quad \varphi]^T$ represents the distance, azimuth, and elevation angle of point $\mathbf{P}$ in the spherical coordinate system.

The projection point $\mathbf{P}_s$ is defined as:

$$\mathbf{P}_s = \begin{pmatrix} x_s \\ y_s \end{pmatrix} = \begin{bmatrix} R\sin\theta \\ R\cos\theta \end{bmatrix} = \frac{1}{\cos\varphi}\begin{bmatrix} x \\ y \end{bmatrix}. \tag{20}$$

Now suppose that the sonar device follows rigid body motion, and then the coordinate points $\mathbf{P}$ and $\mathbf{P}'$ of different views of the same scene satisfy the following transformation relationship:

$$\mathbf{P}' = \mathbf{R}\mathbf{P} + \mathbf{t}, \tag{21}$$

where $\mathbf{R}$ is 3×3 3D rotation matrix and $\mathbf{t}$ is the 3D translation vector.

Let $\mathbf{n} = [n_x, n_y, n_z]^T$ be the scaled normal vector derived from the plane equation $Z = Z_o + \zeta_x X + \zeta_y Y$, and satisfying $\mathbf{n} \cdot \mathbf{P} = 1$, and then [22]

$$\mathbf{P}' = \left(\mathbf{R} + \mathbf{t}\mathbf{n}^T\right)\mathbf{P} = \mathbf{Q}\mathbf{P}, \tag{22}$$

where

$$\mathbf{Q} = \begin{bmatrix} q_{11} & q_{12} & q_{13} \\ q_{21} & q_{22} & q_{23} \\ q_{31} & q_{32} & q_{33} \end{bmatrix}. \tag{23}$$

Using Eq.18, we can get

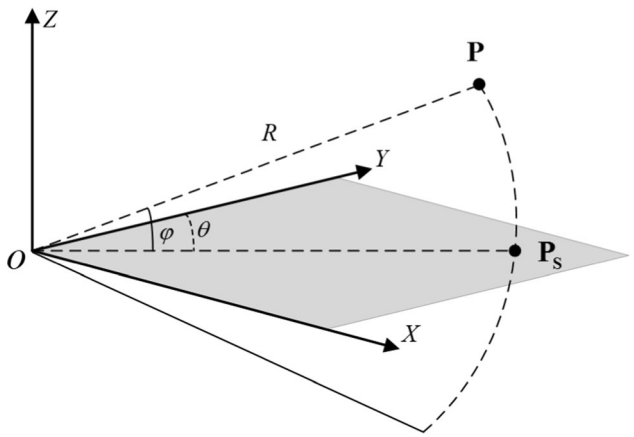

Fig. 9 Sonar coordinate system and projection model in mapping a 3D point onto zero-elevation plane

$$\begin{bmatrix} R'\cos\varphi'\sin\theta' \\ R'\cos\varphi'\cos\theta' \\ R'\sin\varphi' \end{bmatrix} = \mathbf{Q}\begin{bmatrix} R\cos\varphi\sin\theta \\ R\cos\varphi\cos\theta \\ R\sin\varphi \end{bmatrix}, \tag{24}$$

which can be rewritten:

$$\begin{bmatrix} R'\sin\theta' \\ R'\cos\theta' \\ R'\tan\varphi' \end{bmatrix} = \left(\frac{\cos\varphi}{\cos\varphi'}\right)\mathbf{Q}\begin{bmatrix} R\sin\theta \\ R\cos\theta \\ R\tan\varphi \end{bmatrix}. \tag{25}$$

Using Eq.20, we can get

$$\begin{bmatrix} x'_s \\ y'_s \\ R'\tan\varphi' \end{bmatrix} = \left(\frac{\cos\varphi}{\cos\varphi'}\right)\mathbf{Q}\begin{bmatrix} x_s \\ y_s \\ R\tan\varphi \end{bmatrix}, \tag{26}$$

and then

$$\begin{aligned} x'_s &= \left(\frac{\cos\varphi}{\cos\varphi'}\right)[q_{11}x_s + q_{12}y_s + q_{13}R\tan\varphi], \\ y'_s &= \left(\frac{\cos\varphi}{\cos\varphi'}\right)[q_{21}x_s + q_{22}y_s + q_{23}R\tan\varphi], \end{aligned} \tag{27}$$

where the projection point of point $\mathbf{P}'$ on the 2D image plane is:

$$\mathbf{P}'_s = \begin{pmatrix} x'_s \\ y'_s \end{pmatrix} = \begin{bmatrix} R'\sin\theta' \\ R'\cos\theta' \end{bmatrix} = \frac{1}{\cos\varphi'}\begin{bmatrix} x' \\ y' \end{bmatrix}, \tag{28}$$

Finally, sonar image points satisfy the transformation [6]:

$$\begin{bmatrix} x'_s \\ y'_s \\ 1 \end{bmatrix} = \mathbf{H}\begin{bmatrix} x_s \\ y_s \\ 1 \end{bmatrix}, \tag{29}$$

where

$$\mathbf{H} = \begin{bmatrix} \alpha q_{11} & \alpha q_{12} & \beta q_{13} \\ \alpha q_{21} & \alpha q_{22} & \beta q_{23} \\ 0 & 0 & 1 \end{bmatrix}; \alpha = \frac{\cos\varphi}{\cos\varphi'}, \beta = R\frac{\sin\varphi}{\cos\varphi'}. \tag{30}$$

Although it appears to be an affine model, the elements of $\mathbf{H}$ are different throughout the image due to the dependence on elevation angles φ and φ'. The sonar images from two different viewpoints show a complex nonlinear transformation relationship [22].

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Zitova B, Flusser J (2003) Image registration methods: a survey. Image Vis Comput 21(11):977–1000
2. Liu J, Gong J, Guo B, Zhang W (2017) A novel adjustment model for mosaicking low-overlap sweeping images. IEEE Trans Geosci Remot Sens 55(7):4089–4097
3. Goshtasby AA, Nikolov S (2007) Image fusion: advances in the state of the art. Infor fus 2(8):114–118
4. Zanetti M, Bruzzone L (2017) A theoretical framework for change detection based on a compound multiclass statistical model of the difference image. IEEE Trans Geosci Remot Sens 56(2):1129–1143
5. Vakalopoulou M, Karantzalos K, Komodakis N, Paragios N (2015) Simultaneous registration and change detection in multitemporal, very high resolution remote sensing data. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp 61–69
6. Negahdaripour S, Firoozfam P, Sabzmeydani P (2005) On processing and registration of forward-scan acoustic video imagery. In: The 2nd Canadian conference on computer and robot vision (CRV'05), IEEE, pp 452–459
7. Li H, Dong Y, He X, Xie S, Luo J (2014) A sonar image mosaicing algorithm based on improved sift for usv. In: 2014 IEEE International conference on mechatronics and automation, IEEE, pp 1839–1843
8. Negahdaripour S, Aykin M, Sinnarajah S (2011) Dynamic scene analysis and mosaicing of benthic habitats by fs sonar imaging-issues and complexities. In: OCEANS'11 MTS/IEEE KONA, IEEE, pp 1–7
9. Yang Z, Dan T, Yang Y (2018) Multi-temporal remote sensing image registration using deep convolutional features. IEEE Access 6:38544–38555
10. Balakrishnan G, Zhao A, Sabuncu MR, Guttag J, Dalca AV (2019) Voxelmorph: a learning framework for deformable medical image registration. IEEE Trans Medical Imag 38(8):1788–1800
11. Zhao S, Dong Y, Chang EI, Xu Y, et al. (2019) Recursive cascaded networks for unsupervised medical image registration. In: Proceedings of the IEEE international conference on computer vision, pp 10600–10610
12. de Vos BD, Berendsen FF, Viergever MA, Staring M, Išgum I (2017) End-to-end unsupervised deformable image registration with a convolutional neural network. In: Deep learning in medical image analysis and multimodal learning for clinical decision support, Springer, pp 204–212
13. Galceran E, Djapic V, Carreras M, Williams DP (2012) A real-time underwater object detection algorithm for multi-beam forward looking sonar. IFAC Proceed Vol 45(5):306–311
14. Quidu I, Jaulin L, Bertholom A, Dupas Y (2012) Robust multitarget tracking in forward-looking sonar image sequences using navigational data. IEEE J Ocean Eng 37(3):417–430
15. Clark DE, Bell J (2005) Bayesian multiple target tracking in forward scan sonar images using the phd filter. IEE Proceed-Radar, Sonar Navigat 152(5):327–334
16. Petillot Y, Ruiz IT, Lane DM (2001) Underwater vehicle obstacle avoidance and path planning using a multi-beam forward looking sonar. IEEE J Ocean Eng 26(2):240–251
17. Hurtos N, Ribas D, Cufí X, Petillot Y, Salvi J (2015) Fourier-based registration for robust forward-looking sonar mosaicing in low-visibility underwater environments. J Field Robot 32(1):123–151
18. Hurtós N, Nagappa S, Cufí X, Petillot Y, Salvi J (2013) Evaluation of registration methods on two-dimensional forward-

looking sonar imagery. In: 2013 MTS/IEEE OCEANS-Bergen, IEEE, pp 1–8
19. Hurtós N, Petillot Y, Salvi J, et al. (2012) Fourier-based registrations for two-dimensional forward-looking sonar image mosaicing. In: 2012 IEEE/RSJ International conference on intelligent robots and systems, IEEE, pp 5298–5305
20. Zhang J, Sohel F, Bian H, Bennamoun M, An S (2016) Forward-looking sonar image registration using polar transform. In: OCEANS 2016 MTS/IEEE Monterey, IEEE, pp 1–6
21. Aykin M, Negahdaripour S (2012) On feature extraction and region matching for forward scan sonar imaging. In: 2012 Oceans, IEEE, pp 1–9
22. Sekkati H, Negahdaripour S (2007) 3-d motion estimation for positioning from 2-d acoustic video imagery. In: Iberian conference on Pattern Recognition and Image Analysis, Springer, pp 80–88
23. Hurtós Vilarnau N, et al. (2014) Forward-looking sonar mosaicing for underwater environments
24. Hurtós N, Palomeras N, Nagappa S, Salvi J (2013) Automatic detection of underwater chain links using a forward-looking sonar. In: 2013 MTS/IEEE OCEANS-Bergen, IEEE, pp 1–7
25. Guo Y, Wei L, Xu X (2020) A sonar image segmentation algorithm based on quantum-inspired particle swarm optimization and fuzzy clustering. Neural Comput Appl 32(22):16775–16782
26. Zhao S, Lau T, Luo J, Eric I, Chang C, Xu Y (2019) Unsupervised 3d end-to-end medical image registration with volume tweening network. IEEE J Biomed Health Infor 24(5):1394–1404
27. Hur J, Roth S (2019) Iterative residual refinement for joint optical flow and occlusion estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5754–5763
28. Brown LG (1992) A survey of image registration techniques. ACM Comput Surveys (CSUR) 24(4):325–376
29. Lowe DG (1999) Object recognition from local scale-invariant features. Proceedings of the seventh IEEE International conference on computer vision, IEEE 2:1150–1157
30. Bay H, Tuytelaars T, Van Gool L (2006) Surf: Speeded up robust features. In: European conference on computer vision, Springer, pp 404–417
31. Rublee E, Rabaud V, Konolige K, Bradski G (2011) Orb: An efficient alternative to sift or surf. In: 2011 International conference on computer vision, IEEE, pp 2564–2571
32. Fischler MA, Bolles RC (1981) Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications ACM 24(6):381–395
33. Moisan L, Moulon P, Monasse P (2012) Automatic homographic registration of a pair of images, with a contrario elimination of outliers. Image Process Line 2:56–73
34. Raguram R, Chum O, Pollefeys M, Matas J, Frahm JM (2012) Usac: a universal framework for random sample consensus. IEEE Trans Patt Anal Mach Intell 35(8):2022–2038
35. Tao W, Zhao J, Liu J, Zhang H (2010) Study on the side-scan sonar image matching navigation based on surf. In: 2010 International conference on electrical and control engineering, IEEE, pp 2181–2184
36. Gai S, Xu X, Xiong B (2020) Paper currency defect detection algorithm using quaternion uniform strength. Neural computing and applications pp 1–18
37. Viola P, Wells WM III (1997) Alignment by maximization of mutual information. International J Comput Vis 24(2):137–154
38. Maes F, Collignon A, Vandermeulen D, Marchal G, Suetens P (1997) Multimodality image registration by maximization of mutual information. IEEE Transact Med Imag 16(2):187–198
39. Wang G, Xu X, Jiang X, Ding S (2016) Medical image registration based on self-adapting pulse-coupled neural networks and mutual information. Neur Comput Appl 27(7):1917–1926
40. Briechle K, Hanebeck UD (2001) Template matching using fast normalized cross correlation. Optical Pattern Recognition XII. Int Soci Optic Phot 4387:95–102
41. Sarvaiya JN, Patnaik S, Bombaywala S (2009) Image registration by template matching using normalized cross-correlation. In: 2009 International conference on advances in computing, control, and telecommunication technologies, IEEE, pp 819–822
42. Das A, Bhattacharya M (2011) Affine-based registration of CT and MR modality images of human brain using multiresolution approaches: comparative study on genetic algorithm and particle swarm optimization. Neural Comput Appl 20(2):223–237
43. Song S, Herrmann JM, Si B, Liu K, Feng X (2017) Two-dimensional forward-looking sonar image registration by maximization of peripheral mutual information. Int J Adv Robot Sys 14(6):1729881417746270
44. Valdenegro-Toro M (2017) Improving sonar image patch matching via deep learning. In: 2017 European conference on mobile robots (ECMR), IEEE, pp 1–6
45. Sarnel H, Senol Y (2011) Accurate and robust image registration based on radial basis neural networks. Neural Comput Appl 20(8):1255–1262
46. Ot P, dos Santos MM, Drews PLJ, da Costa Botelho SS, et al. (2017) Forward looking sonar scene matching using deep learning. In: 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA), IEEE, pp 574–579
47. Cheng X, Zhang L, Zheng Y (2018) Deep similarity learning for multimodal medical images. Comput Method Biomech Biomed Eng: Imag Visual 6(3):248–252
48. DeTone D, Malisiewicz T, Rabinovich A (2016) Deep image homography estimation. arXiv:1912.02942
49. Chee E, Wu Z (2018) Airnet: Self-supervised affine registration for 3d medical images using neural networks. arXiv:1810.02583
50. Sokooti H, De Vos B, Berendsen F, Lelieveldt BP, Išgum I, Staring M (2017) Nonrigid image registration using multi-scale 3d convolutional neural networks. In: International conference on medical image computing and computer-assisted intervention, Springer, pp 232–239
51. Sokooti H, de Vos B, Berendsen F, Ghafoorian M, Yousefi S, Lelieveldt BP, Isgum I, Staring M (2019) 3d convolutional neural networks image registration based on efficient supervised learning from artificial deformations. arXiv:1908.10235
52. Fu Y, Lei Y, Wang T, Curran WJ, Liu T, Yang X (2020) Deep learning in medical image registration: a review. Phys Medic Biol 65:20TR01
53. Jaderberg M, Simonyan K, Zisserman A, et al. (2015) Spatial transformer networks. In: Advances in neural information processing systems, pp 2017–2025
54. Ronneberger O, Fischer P, Brox T (2015) U-net: Convolutional networks for biomedical image segmentation. In: International conference on medical image computing and computer-assisted intervention, Springer, pp 234–241
55. Zou W, Luo Y, Cao W, He Z, He Z (2021) A cascaded registration network rcinet with segmentation mask. Neural Computing and Applications pp 1–17
56. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556
57. Von Gioi RG, Jakubowicz J, Morel JM, Randall G (2012) Lsd: a line segment detector. Image Process Line 2:35–55
58. Liu R, Lehman J, Molino P, Petroski Such F, Frank E, Sergeev A, Yosinski J (2018) An intriguing failing of convolutional neural networks and the coordconv solution. Adv Neural Infor Process Sys 31:9605–9616

59. Handa A, Bloesch M, Pătrăucean V, Stent S, McCormac J, Davison A (2016) gvnn: Neural network library for geometric computer vision. In: European conference on computer vision, Springer, pp 67–82
60. Cheng H, GUPTA K (1989) An historical note on finite rotations. J Appl Mech 56(1):139–145
61. Gallego G, Yezzi A (2015) A compact formula for the derivative of a 3-d rotation in exponential coordinates. J Math Imag Vis 51(3):378–384
62. Langner O, Dotsch R, Bijlstra G, Wigboldus DH, Hawk ST, Van Knippenberg A (2010) Presentation and validation of the radboud faces database. Cognit Emot 24(8):1377–1388
63. Kingma DP, Ba J (2014) Adam: A method for stochastic optimization. arXiv:1412.6980
64. Chen J, Li Y, Du Y, Frey EC (2020) Generating anthropomorphic phantoms using fully unsupervised deformable image registration with convolutional neural networks. Med Phys 47(12):6366–6380
65. Saad ZS, Glen DR, Chen G, Beauchamp MS, Desai R, Cox RW (2009) A new method for improving functional-to-structural MRI alignment using local pearson correlation. Neuroimage 44(3):839–848
66. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP (2004) Image quality assessment: from error visibility to structural similarity. IEEE Trans Image process 13(4):600–612
67. Guo X, Xu Z, Lu Y, Pang Y (2005) An application of fourier-mellin transform in image registration. In: The Fifth international conference on computer and information technology (CIT'05), IEEE, pp 619–623
68. Chen X, Meng Y, Zhao Y, Williams R, Vallabhaneni SR, Zheng Y (2021) Learning unsupervised parameter-specific affine transformation for medical images registration. In: International conference on medical image computing and computer-assisted intervention, Springer, pp 24–34
69. Mok TC, Chung AC (2020) Large deformation diffeomorphic image registration with laplacian pyramid networks. In: International conference on medical image computing and computer-assisted intervention, Springer, pp 211–221
70. Kim B, Han I, Ye JC (2021) Diffusemorph: Unsupervised deformable image registration along continuous trajectory using diffusion models. arXiv:2112.05149