# GroupRegNet: A Groupwise One-shot Deep Learning-based 4D Image Registration Method

 $^1\mathrm{Yunlu}$ Zhang,  $^1\mathrm{Xue}$  Wu<br/>, $^{1,2,3}\mathrm{H}.$  Michael Gach,  $^{1,2}\mathrm{Harold}$  Li,  $^{1,2}\mathrm{Deshan}$  Yang

Departments of <sup>1</sup>Radiation Oncology, <sup>2</sup>Biomedical Engineering, and <sup>3</sup>Radiology, Washington University in Saint Louis, St. Louis, MO, USA, 63110

E-mail: yangdeshan@wustl.edu

September 2020

Abstract. Accurate deformable 4-dimensional (4D) (3-dimensional in space and time) medical images registration is essential in a variety of medical applications. Deep learning-based methods have recently gained popularity in this area for the significantly lower inference time. However, they suffer from drawbacks of nonoptimal accuracy and the requirement of a large amount of training data. A new method named GroupRegNet is proposed to address both limitations. The deformation fields to warp all images in the group into a common template is obtained through one-shot learning. The use of the implicit template reduces bias and accumulated error associated with the specified reference image. The one-shot learning strategy is similar to the conventional iterative optimization method but the motion model and parameters are replaced with a convolutional neural network (CNN) and the weights of the network. GroupRegNet also features a simpler network design and a more straightforward registration process, which eliminates the need to break up the input image into patches. The proposed method was quantitatively evaluated on two public respiratory-binned 4D-CT datasets. The results suggest that GroupRegNet outperforms the latest published deep learning-based methods and is comparable to the top conventional method pTVreg. To facilitate future research, the source code is available at https://github.com/vincentme/GroupRegNet.

Submitted to: Physics in Medicine & Biology

## 1. Introduction

4-dimensional (4D) (3-dimensional in space and time) medical images have been used in a variety of medical applications. For instance, 4D computed tomography (4D-CT) images have been used to determine patient-specific tumor motion patterns through deformable image registration (DIR), which is a critical step in the planning and delivery of radiation therapy of lung cancer. The subsequent tumor response and anatomy change during treatment can then be studied by registering newly acquired longitudinal volumetric scans.

Numerous research studies have been devoted to developing accurate DIR algorithms. However, as limited by image noise, the lack of features, the use of multiple imaging modalities, and often irregular patient respiratory motion patterns, a generalized, accurate, robust, and computational efficient DIR algorithm has yet to be developed.

DIR algorithms can be roughly categorized into two types: conventional and learning-based methods. The conventional methods [5, 10, 14] formulate the registration problem as an iterative optimization problem while the learning-based methods generate a regression model from the training data. Recently, deep learning-based registration methods [7, 4, 8, 11] have gained popularity due to their low inference time. Like many subareas in computer vision, the features learned through training have shown to be more robust and more general than handcrafted features. In brief, the learning-based methods can be classified as supervised or unsupervised. The former requires a large amount of annotated segmentation data or artificial deformation data to train the network. However, generating such data is time-consuming and thus often impractical, limited to a specific problem, and error-prone. Therefore, most of the recently proposed learning-based methods adopted the unsupervised approach that is guided by the similarity loss. However, these unsupervised learning methods have not yet achieved the accuracy of a few conventional methods [7].

In addition to accuracy, most deep learning-based methods require a large amount of high-quality training data. Even for unsupervised methods, a sufficient number of training images of the same modality is required for training; this is however often not possible for many medical applications. The recent developed one-shot learning strategy [4] eliminated this constraint while achieving excellent results. The one-shot learning strategy is similar to classical registration methods but replaces the conventional motion model and its parameters with a convolutional neural network (CNN) and its weights where the weights are trained from scratch only using the images to be registered.

Another strategy that can be beneficial is via groupwise registration that registers multiple images to a common space instead of in pairs, and is especially suitable for 4D-CT registration. Three variations of groupwise registration exist in the literature: reference-based, sum-of-pairs [10], and implicit template approach [14]. In specific, the reference-based approach requires the selection of one particular image as reference, the sum-of-pairs approach attempts to reduce the losses among all pairs of images, and the implicit template approach is able to avoid the bias caused by selecting one particular image as reference while being computationally efficient comparing to the sum-of-pairs approach.

In this study, an unsupervised deep learning-based DIR method that employs both groupwise registration and one-shot strategy, GroupRegNet, is proposed to register 4D medical images and then to determine all pairwise deformation vector fields (DVFs). The major contributions of this work are summarized as follows. First, groupwise registration with implicitly determined template image strategy is implemented using a neural network. The periodic motion in the 4D image group is also utilized through

cyclic loss. Second, the one-shot unsupervised learning approach eliminates the need for abundant training data. Third, the proposed method features a simpler network design, a minimal prepossessing, and a straightforward registration process compared to other learning-based DIR methods. In terms of accuracy, the GroupRegNet method outperforms the latest published deep learning-based methods and is comparable to a top conventional method pTVreg.

#### 2. Methods

## 2.1. Problem formulation

Let  $I^N$  denotes a group of gray scale images  $I^N = \{I_n | n = 1, ..., N\}$ .  $I_n : \Omega \to \mathbb{R}, \Omega \subset \mathbb{R}^d$  represents each image in the group. The proposed method applies for  $I_n$  as 2D or 3D images, but throughout the rest of the paper, we assume they are 3D images representing one phase in time in a 4D-CT dataset. The objective of GroupRegNet is to find a set of dense transformations that map the same anatomical locations between any two individual images in the group.

The optimization problem to be solved by GroupRegNet is formulated as:

$$\underset{T_{\text{tem}}^{N}}{\operatorname{argmin}}(L_{\text{simi}}(T_{\text{tem}}^{N} \circ I^{N}, I_{\text{tem}}) + \lambda_{0}L_{\text{smo}}(T_{\text{tem}}^{N})) + \lambda_{1}L_{\text{cyc}}(T_{\text{tem}}^{N})), \tag{1}$$

where  $L_{\text{simi}}$ ,  $L_{\text{smo}}$ , and  $L_{\text{cyc}}$  are the similarity, smoothness, and cyclic regularization losses,  $T_{\text{tem}}^N$  is a set of transformations  $\{T_{\text{tem}}^n|n=1,\ldots,N\}$  that maps anatomical locations in the template to the corresponding locations in the input images,  $T_{\text{tem}}^n \circ I_n$  and  $T_{\text{tem}}^N \circ I^N$  represent the warped nth input image and all warped input images, respectively,  $I_{\text{tem}} = \frac{1}{N} \sum_n (T_{\text{tem}}^n \circ I_n)$  is the implicit template by averaging warped input images [12],  $\lambda_0$  and  $\lambda_1$  are the weights for smoothness and cyclic regularization, respectively. The cyclic regularization term will only be present if the relative motion in the image group is periodic or symmetric. The objective of the iterative optimization then becomes finding the optimal transformation  $T_{\text{tem}}^n$  that aligns every image in the group to a template image while keeping the deformation field smooth and cyclically consistent. The inverse transformation  $T_n^{\text{tem}}$  that maps the same anatomical locations in the input image to the implicit template is determined from a fixed-point method [3]. The transformation mapping between the nth and mth image  $T_m^n$  can be calculated using the composition of the deformation field:  $T_m^n(x) = T_{\text{tem}}^n(T_m^{\text{tem}}(x))$ .

Figure 1 illustrates the components and data flowing of GroupRegNet. As compared to the common structure of a learning-based method VoxelMorph [1], GroupRegNet uses similar components including a CNN (to be explained in the later subsections), a spatial transformer (implemented as a 3D linear interpolation), a similarity loss, a cyclic loss, and a smoothness loss. The input images are processed by the CNN to directly estimate the displacement fields. Existing methods in the literature explicitly select the reference and moving images to form a pair and then warp the moving image to the reference image. By contrast, in GroupRegNet, the input images in the group are

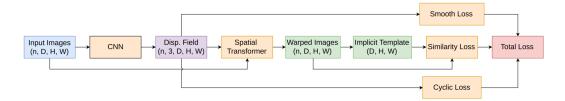


Figure 1: Flowchart of GroupRegNet. The expression (n, D, H, W) represents the number of images in the group and the spatial dimensions of the image.

first stacked in the channel dimension before feeding into the neural network, and the computed transformation then aims to warp the input image into the common space of the template image. It should be noted that CNN's output is the displacement field  $D_{\text{tem}}^n(x)$  instead of the transformation field  $T_{\text{tem}}^n(x)$ , which are related through  $T_{\text{tem}}^n(x) = D_{\text{tem}}^n(x) + x$ . The details of the components in this flowchart are further elaborated in the next subsections.

#### 2.2. Network design

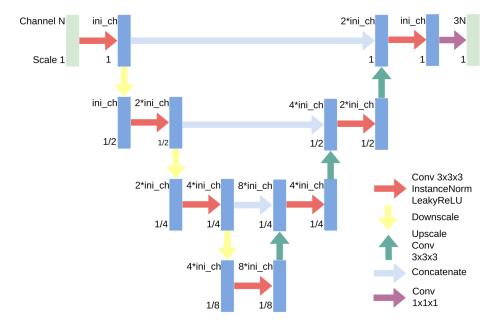


Figure 2: Detailed structure of the CNN sub-network. The overall design is similar to U-Net with modifications. The rectangle blocks represent the feature maps with denoted number of channels (top) and image scale (bottom).

The CNN model consists of convolution, downscale/upscale, and skip connection. The detailed structure of the CNN is shown in figure 2. The overall structure is the same as U-net, which is used by most medical image registration networks. However, several changes have been made to meet the need of the one-shot groupwise registration.

- (i) In the original U-net, the downscale and upscale layers are implemented by maxpooling and transposed convolution. They are replaced by a more straightforward interpolation layer to convert the scales between feature maps.
- (ii) The number of batches will always be one since only one group of images will be fed into the network during the optimization process. Therefore, the batch normalization is replaced by the instance normalization.
- (iii) The two consecutive sets of convolution-normalization-activation operations are reduced to one. This change increases efficiency without impairing the performance. The leaky rectified activation layer is used instead of the original rectified linear activation (ReLU).
- (iv) Due to the size limitation of the common video memory, the input image is downscaled to a lower resolution before being fed into the CNN. The output displacement field  $D_{\text{tem}}^n$  is then upscaled to the original resolution to warp the input images. The scale used in this work is 0.5.

## 2.3. Loss functions

The local normalized cross-correlation (NCC) coefficient is adopted to measure the similarity loss  $L_{\text{simi}}$  between the template and warped input images for its robustness against noise and intensity shift. Let  $\bar{f}(x) = \sum_{x_i} f(x_i)/n^3$  and  $\hat{f}(x) = \sum_{x_i} (f(x_i) - \bar{f}(x))^2$  denote the local mean and variance images, respectively, where  $x_i$  loops over a cubic volume with a size  $n^3$  around the voxel x, with n=5 in the current implementation. The NCC coefficient between the two images is calculated using

$$NCC(f,g) = \frac{1}{|\Omega|} \sum_{x \in \Omega} \frac{\sum_{x_i} (f(x_i) - \bar{f}(x))(g(x_i) - \bar{g}(x))}{\sqrt{\hat{f}(x)\hat{g}(x)}}.$$
 (2)

Accordingly, the similarity loss  $L_{\text{simi}}$  is the average negative NCC coefficient between an individual warped input image and the template image

$$L_{\text{simi}}(T_{\text{tem}}^N \circ I^N, I_{tem}) = -\frac{1}{N} \sum_n NCC(T_{\text{tem}}^n \circ I_n, I_{tem}). \tag{3}$$

 $L_{\text{simi}}$  is in the range of [-1,1] for which a lower value indicates a higher similarity.

The smoothness regularization loss  $L_{\rm smo}$  encourages a smooth and realistic transformation, which accounts the displacement field gradient and the gradient of the image [9]:

$$L_{\text{smo}}(D_{\text{tem}}^{N}, I_{tem}) = \frac{1}{3N|\Omega|} \sum_{n, x \in \Omega, i \in X, Y, Z} (\|\nabla_{i} D_{\text{tem}}^{n}(x)\|_{1} \exp(-|\nabla_{i} I_{tem}(x)|)).(4)$$

Here  $\nabla_i D_{\text{tem}}^n(x)$  is the partial derivative of the displacement field with respect to axis i, which is approximated by a forward difference. Our initial choice of the smoothness term was the isotropic total variation [13]. The current term in equation 4 slightly increases the accuracy and efficiency by promoting the consistency between the gradient of displacement and the gradient of image intensity to preserve the edges.

An optional cyclic consistent regularization loss is used if deformation fields in the group are periodic or symmetric, such as those present in a respiratory-binned 4D-CT. This loss reduces the sum of displacements from one location in the template to all corresponding locations in the input images so that the estimated template is at the center of all input images in the image manifold. An alternative cyclic loss is to reduce the composition of transformations through the cycle of motion. However, in practice, it is computationally expensive to implement. The smoothness and cyclic consistent loss are both positive values, and a lower loss represents the higher smoothness or consistency of the deformation field, respectively.

$$L_{\text{cyc}}(T_{\text{tem}}^N) = \sqrt{\frac{1}{3|\Omega|}} \sum_{x \in \Omega, i \in X, Y, Z} (\sum_n T_{\text{tem}, i}^n(x))^2.$$
 (5)

# 2.4. One-shot learning and convergence criterion

The one-shot learning strategy is used in GroupRegNet to eliminate the requirement of abundant training data. The input images in the group are stacked in the channel dimension, then fed into the neural network to derive the current total loss and to update the weights iteratively through backpropagation. The weights in CNN are independently initialized at the beginning of each iterative registration process. In this sense, the one-shot strategy is similar to the iterative optimization in the variational registration.

After each iteration, a set of convergence criteria is evaluated to determine whether the iterative process should be terminated. The main criterion is the standard deviation of the recent similarity losses. A list of  $N_{\text{stop}}$  latest similarity losses is maintained. A lower standard deviation of this list indicates that a more stable solution has been reached. More specifically, the optimization will stop if

- (i) The standard deviation  $\sigma$  of  $N_{\text{stop}}$  latest similarity losses is less than the threshold  $\sigma_{\text{stop}}$ .
- (ii) Current similarity loss is not smaller than the previous minimum similarity loss and not larger than the previous minimum plus  $\sigma_{\text{stop}}/3$ .
- (iii) The number of computed iterations should be larger than a predefined value  $N_{\rm iter}$ .

The parameter  $N_{\text{stop}}$ ,  $\sigma_{\text{stop}}$ , and  $N_{\text{iter}}$  are empirically determined to be 100, 0.0007, and 200, and they are kept the same for all experiments. The determined displacement field  $D_{\text{tem}}^{N}$  is the output from the CNN of the last iteration. For all evaluated cases, this set of criteria and parameters have proved to be able to overcome the local minimum while avoiding prolonged computation. One example of the convergence curve of different losses vs. the number of iterations is shown in figure 3.

## 2.5. Implementation details

The proposed algorithm is implemented in PyTorch. The Adam optimizer with the learning rate of 0.01 is used for optimization. The number of downscales in CNN is set

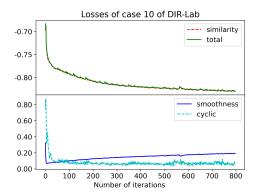


Figure 3: Example of a convergence curve: similarity  $L_{\text{simi}}$ , smoothness  $L_{\text{smo}}$ , cyclic  $L_{\text{cyc}}$ , and total loss  $L_{\text{total}}$  vs. number of iteration for case 10 of the DIR-Lab dataset.

to 3 and the initial number of channels is 32. In this setting, the total number of trainable parameters in the network is 2.4 million. The default Kaiming initialization method is used for all convolutional layers. The regularization terms  $\lambda_0$  and  $\lambda_1$  are empirically set to  $1 \times 10^{-3}$  and  $1 \times 10^{-2}$ , respectively. Computations are conducted on an 8-core CPU AMD Ryzen 3700X with a Nvidia 2080Ti GPU. To facilitate future research, the source code is available at https://github.com/vincentme/GroupRegNet.

# 3. Experiments

#### 3.1. Datasets

To quantitatively evaluate the accuracy of GroupRegNet, the publicly available 4D-CT dataset DIR-Lab [2] was used. This dataset provides 10 thorax 4D-CT scans, each consisting of 10 respiratory-binned phases. Three hundred pairs of corresponding landmarks in the lung were manually delineated by an expert at phases of end-inhalation (EI) and end-exhalation (EE). Two additional observers annotated part of the landmarks with the reported inter-observer variance ranged from  $0.70\pm0.99\,\mathrm{mm}$  to  $1.03\pm2.19\,\mathrm{mm}$ . In addition, 75 sets of landmarks were delineated in all expiratory phase images, i.e. T00, T10, to T50.

The registration accuracy was evaluated by comparing the Euclidean distance, i.e., target registration error (TRE), between the deformed landmarks using the determined deformation fields and annotated landmarks. Note that the 300 pairs of landmarks provided by DIR-Lab suffer from two limitations. First, the number and density of landmarks are limited. Second, the accuracy of landmarks is only at the voxel level. Fu et al. [6] recently proposed an automatic method that can generate a large amount of matching landmarks (1886 pairs on average) evenly distributed in the lung region with subvoxel-level accuracy (average TRE of  $0.47 \pm 0.45 \,\mathrm{mm}$ ). Therefore, these dense matching landmarks were also used in this study. The landmarks provided by DIR-Lab and by Fu et al. [6] are denoted by Landmark300 and LandmarkDense, respectively.

Another dataset, the point-validated pixel-based breathing thorax (POPI) from [12]

was also used to quantitatively evaluate the registration algorithm. This dataset consists of six respiratory phase-binned 4D-CT. About 100 pairs of corresponding landmarks per case at EI and EE phases were created by a semi-automatic approach.

In addition, 4D-CT scans of three lung cancer patients were obtained in the authors' department to quantitatively evaluate tumor tracking using GroupRegNet. The 4D-CT dataset for each patient consists of 10 3D CT volumes representing ten respiratory phases (T00, T10, ..., T90), from the end-inhalation phase (T00) to the end-exhalation phase (T50) then back (T90). The image voxel size was  $1.18 \times 1.18 \times 2 \,\mathrm{mm}^3$ . The tumors were manually contoured by a trained medical physicist on every respiratory phase for each patient. The sizes, shapes, and locations of the tumor targets varied among patients as shown in figure 4. For case 2, the tumor was in the center of the right lung. For cases 1 and 3, the tumors were next to the chest wall and the spine, respectively. To track the tumor motion, the 4D-CT images were first registered using GroupRegNet, then the segmentation mask of the EI phase was warped to other phases with the computed DVFs. The GroupRegNet tumor tracking accuracy was evaluated by computing the average and standard deviation of the Dice coefficients, the distances between the centers of mass, and the 95% Hausdorff distances using the manually contoured tumor target masks as reference.

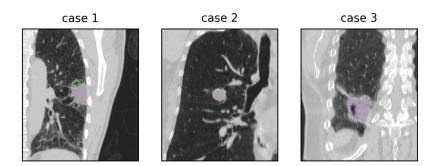


Figure 4: Sizes, shapes, and locations of the contoured tumor targets, shown in violet shade, in coronal views of the EI phases of three patient cases.

## 3.2. Prepossessing

To reduce computation time and improve convergence, the input images were cropped to the bounding box that encompassed the landmarks in all phases plus an 8-voxel margin in all directions. In the cases of tumor target tracking, a 50-voxel margin was added to all directions of the boundingbox of the tumor target in the EI phase. The CT image intensity was approximately normalized to the range of [-1,1] after dividing by 1000. The input images were not spatially resampled, segmented, or vessel enhanced before being fed into GroupRegNet.

#### 4. Results

## 4.1. Accuracy evaluated on landmarks

Table 1: Comparison of TREs (mean±std in mm): GroupRegNet vs. other learning-based and conventional DIR methods using the DIR-Lab dataset evaluated by (a) Landmark300 and (b) LandmarkDense.

case	before reg.	GroupRegNet	LungRegNet[7]	Fechter[4]	MJ-CNN[8]	GDL-FIRE[11]	Fu[5]	Bartlomiej[10]	pTVreg[13]
1	$3.89 \pm 2.78$	$1.02 \pm 0.51$	$0.98 \pm 0.54$	$1.21 \pm 0.88$	$1.20 \pm 0.63$	$1.20 \pm 0.60$	$1.06 \pm 0.50$	$0.90 \pm 1.0$	$0.80 \pm 0.89$
2	$4.34 \pm 3.90$	$1.04 \pm 0.49$	$0.98 \pm 0.52$	$1.13 \pm 0.65$	$1.13 \pm 0.56$	$1.19 \pm 0.63$	$1.09 \pm 0.57$	$0.94 \pm 1.0$	$0.77 \pm 0.90$
3	$6.94 \pm 4.05$	$1.24 \pm 0.71$	$1.14 \pm 0.64$	$1.32 \pm 0.82$	$1.30 \pm 0.70$	$1.67 \pm 0.90$	$1.51 \pm 1.00$	$1.06\pm1.1$	$0.92 \pm 1.07$
4	$9.83 \pm 4.86$	$1.43 \pm 0.97$	$1.39 \pm 0.99$	$1.84 \pm 1.76$	$1.55\pm0.96$	$2.53 \pm 2.01$	$1.73\pm1.55$	$2.53 \pm 3.2$	$1.30\pm1.27$
5	$7.48 \pm 5.51$	$1.41\pm1.22$	$1.43\pm1.31$	$1.80\pm1.60$	$1.72\pm1.28$	$2.06\pm1.56$	$1.80\pm1.63$	$1.31\pm1.5$	$1.13\pm1.42$
6	$10.89 \pm 6.96$	$1.31 \pm 0.72$	$2.26 \pm 2.93$	$2.30 \pm 3.78$	$2.02\pm1.70$	$2.90 \pm 1.70$	$2.25\pm2.61$	$1.89 \pm 1.9$	$0.78 \pm 0.92$
7	$11.02 \pm 7.42$	$1.28 \pm 0.65$	$1.42 \pm 1.16$	$1.91 \pm 1.65$	$1.70\pm1.03$	$3.60 \pm 2.99$	$1.41 \pm 0.98$	$1.52 \pm 1.4$	$0.79 \pm 0.91$
8	$14.99 \pm 9.00$	$1.33 \pm 1.08$	$3.13 \pm 3.77$	$3.47 \pm 5.00$	$2.64 \pm 2.78$	$5.29 \pm 5.52$	$3.53 \pm 5.70$	$1.87 \pm 2.3$	$1.00\pm1.29$
9	$7.92 \pm 3.97$	$1.30 \pm 0.69$	$1.27 \pm 0.94$	$1.47 \pm 0.85$	$1.51 \pm 0.94$	$2.38 \pm 1.46$	$2.31 \pm 1.88$	$1.37 \pm 1.1$	$0.91 \pm 0.95$
10	$7.30 \pm 6.34$	$1.22 \pm 0.63$	$1.93 \pm 3.06$	$1.79 \pm 2.24$	$1.79 \pm 1.61$	$2.13 \pm 1.88$	$1.18\pm1.97$	$1.27\pm1.4$	$0.82 \pm 0.97$
ave.	$8.46 \pm 5.48$	$1.26 \pm 0.77$	$1.59 \pm 1.58$	$1.83 \pm 2.35$	$1.66\pm1.44$	$2.50 \pm 1.16$	$1.78\pm1.83$	$1.47\pm1.6$	$0.92 \pm 1.06$
ave. RMSE	10.08	1.48	2.24	2.98	2.20	2.76	2.55	2.2	1.41

(a) Landmark300								
case	before reg.	GroupRegNet	pTVreg[13]					
1	$3.43 \pm 2.86$	$0.59 \pm 0.33$	$0.32 \pm 0.17$					
2	$4.67 \pm 4.23$	$0.56 \pm 0.36$	$0.38 \pm 0.22$					
3	$5.55 \pm 4.08$	$0.71 \pm 0.37$	$0.41 \pm 0.23$					
4	$7.55 \pm 5.11$	$0.70 \pm 0.35$	$0.56 \pm 0.55$					
5	$4.91 \pm 4.84$	$0.65 \pm 0.36$	$0.47 \pm 0.30$					
6	$9.30 \pm 7.46$	$0.96 \pm 0.57$	$0.70 \pm 1.63$					
7	$8.18 \pm 6.73$	$0.78 \pm 0.40$	$0.48 \pm 0.25$					
8	$8.58 \pm 6.71$	$0.81 \pm 0.43$	$0.61 \pm 1.68$					
9	$5.81 \pm 3.77$	$0.83 \pm 0.45$	$0.48 \pm 0.25$					
10	$6.12 \pm 5.31$	$0.77 \pm 0.43$	$0.44 \pm 0.25$					
ave.	$6.41 \pm 5.11$	$0.74 \pm 0.41$	$0.49 \pm 0.55$					
ave. RMSE	8.20	0.85	0.74					

(b) LandmarkDense

The accuracy of GroupRegNet was compared with seven recently published methods on the DIR-Lab dataset, as shown in table 1. The landmarks in EI phase (phase T00) were deformed to EE phase (phase T50) according to the calculated DVFs, and then compared to the annotated landmarks in EE phase to derive the TREs. GroupRegNet and pTVreg were evaluated on both LandmarkDense and Landmark300, while other methods only reported results on Landmark300.

The average TRE of GroupRegNet was  $1.26\pm0.77\,\mathrm{mm}$ , evaluated on Landmark300, which was lower than most of the surveyed methods, and comparable to pTVreg [13], which is the top method listed on the DIR-Lab website. The average root mean square error (RMSE) of GroupRegNet and pTVreg were at least 30% smaller than other methods. GroupRegNet performed particularly better for cases with large deformations (e.g., cases 6, 7 and 8). It should also be noted that the variance of the TREs using GroupRegNet was even less or at least equal to the inter-observer variance, suggesting that its accuracy was superior to that of manual annotations in most regions.

When evaluated using LandmarkDense, the average TRE and RMSE of GroupRegNet were  $0.74 \pm 0.41 \,\mathrm{mm}$  and  $0.85 \,\mathrm{mm}$ , respectively, demonstrating a submillimeter accuracy. The average RMSEs were similar comparing GroupRegNet vs. pTVreg while the former usually yielded smaller standard deviations but slightly larger average TREs. Note that the standard deviations of pTVreg in cases 6 and 8 were unexpectedly large, which was not observed in GroupRegNet.

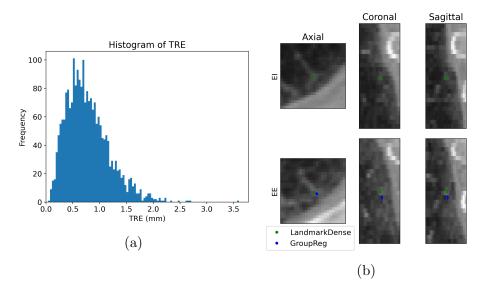


Figure 5: Accuracy of GroupRegNet evaluated on LandmarkDense in case 7 of DIR-Lab. (a) histogram of TREs, (b) the location of the worst point determined by GroupRegNet in phases EI and EE.

The TRE histogram for GroupRegNet in case 7 is shown in figure 5(a) where the percentage of the TREs below 1 mm, 1.5 mm, and 2 mm are 75%, 94%, and 99%, respectively. The worst point with a TRE of 3.6 mm is shown in figure 5(b). The relatively large error was likely caused by the low signal-to-noise ratio and the rapidly changing displacement in this region. Figure 6 provides a typical example of the DIR results. Most structures align well as shown in the red/cyan superimposed image post registration.

The results evaluated on POPI are shown in table 2. GroupRegNet reduced the original TRE from  $8.12 \pm 4.77 \,\text{mm}$  to  $1.03 \pm 0.64 \,\text{mm}$ . Comparing to the results from Fechter and Baltas [4] and GDL-FIRE [11], the average RMSE was reduced by 44%.

All previous evaluations were carried out between phases EI and EE. The 75 landmarks annotated on the expiratory phases of the DIR-Lab dataset were utilized to test whether there are large variations among different phases. The landmarks in phase T00 were deformed to other phases and then compared to manual annotations, as shown in table 3. The TREs of phases T10 and T50 were usually smaller than those of other phases, which could be attributed to the former having smaller deformations and the latter being more stable than the intermediate phases. In addition, the intensity

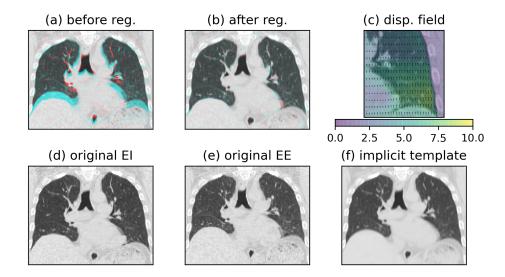


Figure 6: Example of the registration result of a coronal slice of DIR-Lab case 7. The red/cyan superimposed images of phases EI and EE (a) before and (b) after registration. (c) The colormap and vector plot show the magnitude and direction of a portion of the determined displacement field from phase EI to EE. (d), (e), and (f) are the images of phases EI, EE, and determined implicit template, respectively.

Table 2: Comparison of TREs (mean±std in mm) between GroupRegNet and other learning-based or conventional methods on POPI dataset.

case	dimensions	before reg.	${\tt GroupRegNet}$	Fechter[4]	GDL-FIRE[11]
1	512x512x141	$5.90 \pm 2.73$	$1.10 \pm 0.59$	$1.09 \pm 0.68$	$1.34 \pm 0.74$
2	512x512x169	$14.04\pm7.20$	$1.27 \pm 0.93$	$2.71 \pm 3.28$	$2.98 \pm 2.38$
3	512x512x170	$7.67 \pm 5.05$	$0.92 \pm 0.51$	$1.40\pm1.54$	$1.57 \pm 1.01$
4	512x512x187	$7.33 \pm 4.89$	$0.88 \pm 0.47$	$1.17\pm1.83$	$1.64 \pm 1.62$
5	512x512x139	$7.09 \pm 5.08$	$1.01 \pm 0.81$	$1.30 \pm 0.97$	$1.62 \pm 1.09$
6	512x512x161	$6.68 \pm 3.68$	$0.97 \pm 0.51$	$1.27\pm0.95$	$1.26 \pm 0.73$
ave.		$8.12 \pm 4.77$	$1.03 \pm 0.64$	$1.49 \pm 1.54$	$1.74 \pm 1.26$
ave. RMSE		9.42	1.21	2.14	2.15

difference maps between each phase and the warped template image via the reverse DVF  $T_n^{\text{tem}}$  are shown in figure 7. There was not a single intensity-difference map that was obviously better or worse than its counterpart, suggesting similar GroupRegNet performance regardless of phases.

The percentage of the negative determinant of the Jacobian matrix was calculated to evaluate the regularity of the deformation fields. For most DIR-Lab cases, the percentage was zero, except for case 8, which was 0.03%. For the POPI dataset cases, the percentage was  $0.06 \pm 0.13\%$ . These results indicate the determined deformation field was well regularized with a minimal percentage of the negative determinant of the Jacobian matrix.

Table 3: Comparison of TREs (mean±std in mm) of GroupRegNet on different target phase images from the DIR-Lab dataset using 75 landmarks. The result of pTVreg on phase T50 is included for reference.

case	T10	T20	T30	T40	T50	pTVreg T50
1	$0.38 \pm 0.29$	$0.95 \pm 0.65$	$1.27 \pm 0.60$	$1.22 \pm 0.64$	$1.15 \pm 0.57$	$0.92 \pm 0.49$
2	$0.97 \pm 0.77$	$0.94 \pm 0.58$	$0.93 \pm 0.56$	$0.94 \pm 0.52$	$1.00 \pm 0.53$	$0.92 \pm 0.49$
3	$1.17 \pm 0.79$	$1.12 \pm 0.58$	$1.10 \pm 0.59$	$1.23\pm0.61$	$1.21 \pm 0.62$	$1.01 \pm 0.50$
4	$1.10 \pm 0.60$	$1.32 \pm 0.80$	$1.41 \pm 0.83$	$1.55\pm1.24$	$1.39 \pm 0.98$	$1.28 \pm 0.91$
5	$1.47 \pm 1.08$	$1.20\pm0.57$	$1.23\pm0.91$	$1.21\pm0.62$	$1.57\pm1.86$	$1.34\pm1.78$
6	$1.07 \pm 0.89$	$1.79 \pm 1.78$	$1.64 \pm 1.70$	$1.54 \pm 1.36$	$1.31 \pm 0.83$	$1.04 \pm 0.76$
7	$0.99 \pm 0.79$	$1.42\pm1.16$	$1.58\pm1.07$	$1.27\pm0.81$	$1.35 \pm 0.65$	$0.94 \pm 0.49$
8	$1.04 \pm 0.52$	$1.42\pm1.22$	$1.32 \pm 1.09$	$1.73 \pm 2.13$	$1.50 \pm 1.78$	$1.22 \pm 1.74$
9	$1.14 \pm 0.62$	$1.17 \pm 0.70$	$1.22 \pm 0.67$	$1.37 \pm 0.70$	$1.31 \pm 0.77$	$1.09 \pm 0.78$
10	$1.16 \pm 0.93$	$1.38 \pm 1.18$	$1.64 \pm 1.36$	$1.23 \pm 0.63$	$1.16 \pm 0.56$	$0.91 \pm 0.43$

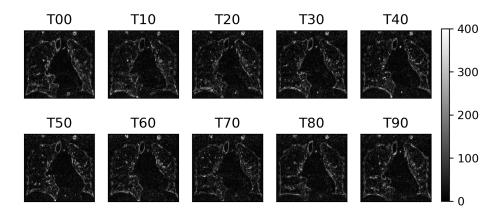


Figure 7: Intensity-difference map between each phase and warped template in coronal view of DIR-Lab case 10.

## 4.2. Accuracy evaluated on segmentation contours

As shown in table 4, the average and standard deviation of the Dice coefficients, distances between the centers of mass, and 95% Hausdorff distances of the warped masks were computed with the manual contours as the reference. After registration, the average Dice coefficient was increased from 0.8 to 0.9, the distance between the centers of mass was reduced by 81% to  $\sim 1\,\mathrm{mm}$ , and the 95% Hausdorff distance was reduced by 50% to  $< 3\,\mathrm{mm}$ . Figure 8 provides a visual example of the tracked target in different phases of case 3.

#### 4.3. Computation variance and speed

Due to the stochastic nature of weights initialization in the neural network, concerns may arise with regard to optimization convergence and variance among multiple runs. In addition, the computation speed is important in practical applications. Two cases

Table 4: Comparison of the Dice coefficients, distances between the centers of mass, and 95% Hausdorff distances of segmentation contours of the tumor targets before and after registration.

case	Dice coefficient		distance between the centers of mass 95% Hausdorff distance					
	mean	$\pm std$	$mean \pm std (mm)$					
	before	after	before	after	before	after		
1	$0.789 \pm 0.070$	$0.903 \pm 0.006$	$4.58 \pm 2.14$	$1.10 \pm 0.32$	$4.73 \pm 1.48$	$2.32 \pm 0.01$		
2	$0.807 \pm 0.080$	$0.913 \pm 0.015$	$1.82 \pm 0.81$	$0.37 \pm 0.16$	$2.17 \pm 0.63$	$1.33 \pm 0.38$		
3	$0.781 \pm 0.130$	$0.887\pm0.043$	$5.74 \pm 3.60$	$0.78 \pm 0.31$	$6.08 \pm 3.22$	$2.84 \pm 0.87$		

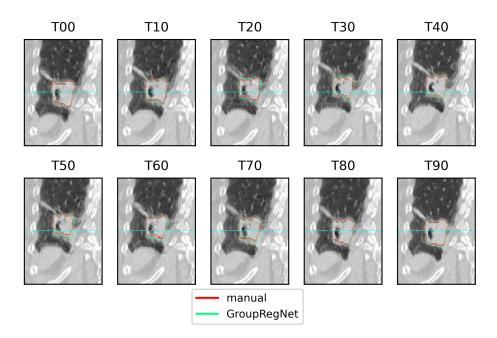


Figure 8: Comparison of the tracked targets in ten phases by GroupRegNet and manual contouring of case 3. The images are shown in coronal views, and the horizontal line in each figure is at the same height for visual reference.

Table 5: Comparison of variance of repeatability error, TRE, and computation speed from five repeated runs by GroupRegNet on selected cases. The input images were cropped, so the dimensions were smaller than the originals.

	repeatability error	mean of TREs	std of TREs	cropped dimensions	num. of iter.	computation time	time per iter.
	me	ean±std (mm)				S	
DIR-Lab case 1	$0.21 \pm 0.12$	$0.59 \pm 0.02$	$0.34 \pm 0.03$	$240 \times 157 \times 83$	$317\pm24$	$265\pm18$	0.8
DIR-Lab case 6	$0.41 \pm 0.21$	$0.94 \pm 0.03$	$0.55 \pm 0.03$	$294 \times 184 \times 97$	$764 \pm 55$	$973 \pm 255$	1.3
POPI case 2	$0.47 \pm 0.23$	$1.29 \pm 0.02$	$0.95 \pm 0.02$	$271 \times 196 \times 116$	$1073 \pm 151$	$1792 \pm 250$	1.7
POPI case 5	$0.37 \pm 0.17$	$1.02 \pm 0.01$	$0.84 \pm 0.02$	$169 \times 128 \times 99$	$712\pm78$	$412 \pm 44$	0.6

with relatively small and large motions from both datasets were repeatedly registered five times using GroupRegNet. The variance of the registration accuracy, number of iteration, and computation time are summarized in table 5. The variance of the registration accuracy was evaluated in terms of repeatability errors and statistics of TREs. The former was calculated as the distance between the displaced landmarks

and their average locations over five runs. Then the average and standard deviation of the repeatability error were computed over all landmarks and runs. The determined repeatability errors ranged from 0.2 mm to 0.5 mm. Although the variances were not minimum, the standard deviation of the statistics of the TREs was at the level of 0.03 mm, indicating similar accuracies of repeated runs. Furthermore, all registrations were completed without convergence issues.

Computation time per iteration ranged from 0.6s to 1.7s, and varied with image size and motion magnitude. The overall computation time was in the range of few minutes to 30 minutes, which is not slow considering that all 10 phases were registered and all pairwise DVFs determined.

#### 5. Discussion

A new DIR method GroupRegNet is presented to register 4D medical images and to determine all pairs of dense DVFs. The results on two respiratory 4D-CT datasets suggest that it is able to achieve state-of-the-art performance. This study is unique in that it has successfully combined and implemented implicit template groupwise registration and one-shot unsupervised learning approach. Although many components have been introduced in the literature, in this work they are strategically integrated, and the method outperforms many other complex and dedicated methods. For instance, figure 6(c) shows the DVF transition around the chest wall where the sliding motion was successfully revealed without additional dedicated steps such as segmentation or DVF decomposition[5]. The implicit template shown in figure 6(f) was successfully revealed by averaging the warped input images, which showed less noise compared to the original images. This is also an advantage of the implicit template groupwise registration method over the pairwise registration method; for the latter both the reference and moving images are inevitably corrupted by noise.

From a broader perspective, GroupRegNet can be viewed as a mixture of conventional and learning-based methods. It follows the same iterative optimization process of the conventional approach and only uses the images to be registered as input. Furthermore, segmentation images, annotated landmarks, or deformation fields do not need to be provided to the neural network. GroupRegNet utilizes CNN as the motion model whose weights are learned through optimization. The performance improvement over the conventional approach can be attributed to the more expressive power of the deep neural network and to fewer assumptions in the DVF. Comparing to a typical training and inference procedure of learning-based methods, the one-shot learning strategy eliminates the requirement of abundant training images and annotations, thus improves the accuracy. The problem of over-fitting for the one-shot training strategy was not presented due to the well regularized total loss.

GroupRegNet utilizes several concepts from previously published registration methods that include group-wise implicit template registration and one-shot unsupervised learning. Fechter and Baltas [4] proposed the one-shot unsupervised

learning approach but the deformation field computation between timely adjacent 3D images required segmented inputs and a complex coarse-to-fine patch processing. Wu et al. [14] introduced the implicit template into the paradigm of the classical registration approach that consists of explicitly defined key points and the Gaussian mixture model for motion modeling. By contrast, GroupRegNet organically integrates these components, and features a simpler network design, a minimal prepossessing step, and a straightforward registration process compared to other methods. Comparing to other learning-based methods, Fu et al. [7] designed separate coarse and fine nets for large and small deformations, and also included a discriminator and a corresponding adversarial loss to regularize DVF. Moreover, both studies [7, 4] ran into limited GPU memory issues so that they had to be trained using a patch-based approach, which was time-consuming and hard to learn the global relationship. Fechter and Baltas [4] paid special attention to the smoothness regularization of the boundary voxels while Fu et al. [7] excluded these voxels in the loss. GroupRegNet tackles this limitation differently where the input images to CNN were downscaled to reduce the size of the feature maps so that the model can run on a typical GPU. Furthermore, the output DVFs are upscaled to the original resolution to warp the input images and then to compute the similarity loss. The global representations are learned instead of using local features in the patch. The output DVFs and similarity regularization computation are computed at the original resolution. This approach is also better than conducting all computation on a lower resolution, which loses the fine details of input images and reduces the accuracy of the DIR.

#### 6. Conclusion

In this paper, a groupwise one-shot learning neural network for 4D image registration was presented. The implicit template strategy was first integrated with the learning-based approach. The utilization of one-shot learning strategy eliminated the need for abundant training data. The simple network structure made the registration at the original resolution without breaking up the input images into patches. The accuracy of GroupRegNet in terms of average RMSE was better than that of the latest learning-based methods and comparable to the top conventional method. The performance of GroupRegNet is expected to be further improved with the addition of more complex networks and strategies, such as generative adversarial network and attention mechanism.

# Acknowledgments

This research was partially supported by the Agency for Healthcare Research and Quality (AHRQ) grant number R01-HS022888, National Institute of Biomedical Imaging and Bioengineering (NIBIB) grant R03-EB028427 and National Heart, Lung, and Blood Institute (NHLBI) grant R01-HL148210.

REFERENCES 16

#### References

[1] Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca. Voxelmorph: a learning framework for deformable medical image registration. *IEEE transactions on medical imaging*, 38(8):1788–1800, 2019.

- [2] Richard Castillo, Edward Castillo, Rudy Guerra, Valen E Johnson, Travis McPhail, Amit K Garg, and Thomas Guerrero. A framework for evaluation of deformable image registration spatial accuracy using large landmark point sets. *Physics in Medicine & Biology*, 54(7):1849, 2009.
- [3] Mingli Chen, Weiguo Lu, Quan Chen, Kenneth J Ruchala, and Gustavo H Olivera. A simple fixed-point approach to invert a deformation field a. *Medical physics*, 35 (1):81–88, 2008.
- [4] Tobias Fechter and Dimos Baltas. One shot learning for deformable medical image registration and periodic motion tracking. *IEEE Transactions on Medical Imaging*, 2020.
- [5] Yabo Fu, Shi Liu, H Harold Li, Hua Li, and Deshan Yang. An adaptive motion regularization technique to support sliding motion in deformable image registration. *Medical physics*, 45(2):735–747, 2018.
- [6] Yabo Fu, Xue Wu, Allan M Thomas, Harold H Li, and Deshan Yang. Automatic large quantity landmark pairs detection in 4dct lung images. *Medical physics*, 46 (10):4490–4501, 2019.
- [7] Yabo Fu, Yang Lei, Tonghe Wang, Kristin Higgins, Jeffrey D Bradley, Walter J Curran, Tian Liu, and Xiaofeng Yang. Lungregnet: an unsupervised deformable image registration method for 4d-ct lung. *Medical Physics*, 47(4):1763–1774, 2020.
- [8] Zhuoran Jiang, Fang-Fang Yin, Yun Ge, and Lei Ren. A multi-scale framework with unsupervised joint training of convolutional neural networks for pulmonary deformable image registration. *Physics in Medicine & Biology*, 65(1):015011, 2020.
- [9] Liang Liu, Jiangning Zhang, Ruifei He, Yong Liu, Yabiao Wang, Ying Tai, Donghao Luo, Chengjie Wang, Jilin Li, and Feiyue Huang. Learning by analogy: Reliable supervision from transformations for unsupervised optical flow estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6489–6498, 2020.
- [10] Bartłomiej W Papież, Daniel R McGowan, Michael Skwarski, Geoff S Higgins, Julia A Schnabel, and Michael Brady. Fast groupwise 4d deformable image registration for irregular breathing motion estimation. In *International Workshop* on Biomedical Image Registration, pages 37–46. Springer, 2018.
- [11] Thilo Sentker, Frederic Madesta, and René Werner. Gdl-fire4d: Deep learning-based fast 4d ct image registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 765–773. Springer, 2018.

REFERENCES 17

[12] Jef Vandemeulebroucke, Simon Rit, Jan Kybic, Patrick Clarysse, and David Sarrut. Spatiotemporal motion estimation for respiratory-correlated imaging of the lungs. *Medical physics*, 38(1):166–178, 2011.

- [13] Valery Vishnevskiy, Tobias Gass, Gabor Szekely, Christine Tanner, and Orcun Goksel. Isotropic total variation regularization of displacements in parametric image registration. *IEEE transactions on medical imaging*, 36(2):385–395, 2016.
- [14] Guorong Wu, Qian Wang, Jun Lian, and Dinggang Shen. Estimating the 4d respiratory lung motion by spatiotemporal registration and super-resolution image reconstruction. *Medical physics*, 40(3):031710, 2013.