2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)
April 4-7, 2018, Washington, D.C., USA

DEFORMABLE MEDICAL IMAGE REGISTRATION USING GENERATIVE ADVERSARIAL NETWORKS

Dwarikanath Mahapatra, Bhavna Antony, Suman Sedai, Rahil Garnavi

IBM Research - Australia, Melbourne

ABSTRACT

Conventional approaches to image registration consist of time consuming iterative methods. Most current deep learning (DL) based registration methods extract deep features to use in an iterative setting. We propose an end-to-end DL method for registering multimodal images. Our approach uses generative adversarial networks (GANs) that eliminates the need for time consuming iterative methods, and directly generates the registered image with the deformation field. Appropriate constraints in the GAN cost function produce accurately registered images in less than a second. Experiments demonstrate their accuracy for multimodal retinal and cardiac MR image registration.

Index Terms— GANs, deformable registration, displacement field

1. INTRODUCTION

Image registration is a fundamental step in most medical image analysis problems, and a comprehensive review of algorithms can be found in [1]. Conventional registration methods use iterative gradient descent based optimization using cost functions such as mean square error (MSE), normalized mutual information, etc. Such methods tend to be time consuming, especially for volumetric images. We propose a fully end-to-end deep learning (DL) approach that does not employ iterative methods, but uses generative adversarial networks (GANs) for obtaining registered images and the corresponding deformation field.

Wu et al. [2] use convolutional stacked autoencoders (CAE) to extract features from fixed and moving images, and use it in a conventional iterative deformable registration framework. Miao et al [3] use convolutional neural network (CNN) regressors in rigid registration of synthetic images. Liao et al [4] employ CNNs and reinforcement learning for iterative registration of CT to cone-beam CT in cardiac and abdominal images. DL based regression methods still require conventional methods to generate the transformed image.

Jaderberg et al. [5] introduced spatial transformer networks (STN) to align input images in a larger task-specific network. STNs, however, need many labeled training examples and have not been used for medical image analysis. Sokooti et. al. [6] propose RegNet that uses CNNs trained on simulated deformations to generate displacement vector fields for a pair of unimodal images. Vos et. al. [7] propose the deformable image registration network (DIR-Net) which takes pairs of fixed and moving images as input, and outputs a transformed image non-iteratively. Training is completely unsupervised and unlike previous methods it is not trained with known registration transformations.

While RegNet and DIRNet are among the first methods to achieve registration in a single pass, they have some limitations such as: 1) using spatially corresponding patches to predict transformations. Finding corresponding patches is challenging in low contrast medical images and can adversely affect the registration task; 2) Multimodal registration is challenging with their approach due to the inherent problems of finding spatially corresponding patches; 3) DIRNet uses B-splines for spatial transformations which limits the extent of recovering a deformation field; 4) Use of intensity based cost functions limits the benefits that can be derived from a DL based image registration framework.

To overcome the above limitations we make the following contributions: 1) we use GANs for multimodal medical image registration, which can recover more complex range of deformations ; 2) novel constraints in the cost function, such as VGG, SSIM loss and deformation field reversibility, ensure that the trained network can easily generate images that are realistic with a plausible deformation field. We can choose any image as the reference image and registration is achieved in a single pass.

2. METHODS

GANs are generative DL models trained to output many image types. Training is performed in an adversarial setting where a discriminator outputs a probability of the generated image matching the training data distribution. GANs have been used in various applications such as image super resolution [8, 9], image synthesis and image translation using conditional GANs (cGANs) [10] and cyclic GANs (cycGANs) [11].

In cGANs the output is conditioned on the input image and a random noise vector, and requires training image pairs. On the other hand cycGANs do not require training image

978-1-5386-3636-7/18/$31.00 ©2018 IEEE

pairs but enforce consistency of deformation field. We leverage the advantage of both methods to register multimodal images. For multimodal registration we use cGANs to ensure the generated output image (i.e., the transformed floating image) has the same characteristic as the floating image (in terms of intensity distribution) while being similar to the reference image (of a different modality) in terms of landmark locations. This is achieved by incorporating appropriate terms in the loss function for image generation. Additionally, we enforce deformation consistency to obtain realistic deformation fields. This prevents unrealistic registrations and allows any image to be the reference or floating image. A new test image pair from modalities not part of the training set can be registered without the need for re-training the network.

2.1. Generating Registered Images

Let us denote the registered (or transformed) image as I^{Trans}, obtained from the input floating image I^{Flt}, and is to be registered to the fixed reference image I^{Ref}. For training we have pairs of multimodal images where the corresponding landmarks are perfectly aligned (e.g., retinal fundus and fluoroscein angiography (FA) images). Any one of the modalities (say fundus) is I^{Ref}. I^{Flt} is generated by applying a known elastic deformation field to the other image modality (in this case FA). The goal of registration is to obtain I^{Trans} from I^{Flt} such that I^{Trans} is aligned with I^{Ref}. Applying synthetic deformations allows us to: 1) accurately quantify the registration error in terms of deformation field recovery; and 2) determine the similarity between I^{Trans} and FA images.

The generator network that outputs I^{Trans} from I^{Flt} is a feed-forward CNN whose parameters θ_G are,

$$\widehat{\theta} = \arg\min_{\theta_G} \frac{1}{N} \sum_{n=1}^{N} l^{SR}\left(G_{\theta_G}(I^{Flt}), I^{Ref}, I^{Flt}\right), \quad (1)$$

where the loss function l^{SR} combines content loss (to ensure that I^{Trans} has desired characteristics) and adversarial loss, and $G_{\theta_G}(I^{Flt}) = I^{Trans}$. The content loss is,

$$l_{content} = NMI(I^{Trans}, I^{Ref}) + SSIM(I^{Trans}, I^{Ref}) + VGG(I^{Trans}, I^{Ref}). \quad (2)$$

I^{Trans} should: 1) have identical intensity distribution as I^{Flt} and; 2) have similar structural information content as I^{Ref}. $NMI(I^{Trans}, I^{Ref})$ denotes the normalized mutual information (NMI) between I^{Ref} and I^{Trans}. NMI is a widely used cost function for multimodal deformable registration [12] since it matches the joint intensity distribution of two images. $SSIM(I^{Trans}, I^{Ref})$ denotes the structural similarity index metric (SSIM) [13] and calculates image similarity based on edge distribution and other landmarks. Since it is not based on intensity values it accurately quantifies landmark correspondence between different images.

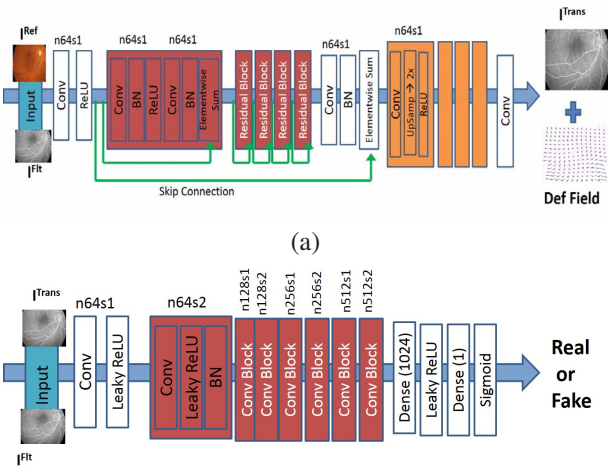


Fig. 1. (a) Generator Network; (b) Discriminator network. $n64s1$ denotes 64 feature maps (n) and stride (s) 1 for each convolutional layer.

$VGG(I^{Trans}, I^{Ref})$ is the $L2$ distance between two images using all 512 feature maps of Relu $4-1$ layer of a pre-trained $VGG-16$ network [14]. VGG loss improves robustness since the cost function takes into account multiple feature maps that capture information at different scales.

The adversarial loss of Eqn. 3 ensures that I^{Trans} has an identical intensity distribution as I^{Flt}. We could realize this condition by having an extra $NMI(I^{Trans}, I^{Flt})$ term in Eqn. 2. But this does not lead to much improvement in results.

The generator network G (Figure 1(a)) employs residual blocks, each block having two convolutional layers with 3×3 filters and 64 feature maps, followed by batch normalization and ReLU activation. In addition to generating the registered image G also outputs a deformation field. The discriminator D (Figure 1 (b)) has eight convolutional layers with the kernels increasing by a factor of 2 from 64 to 512 . Leaky ReLU is used and strided convolutions reduce the image dimension when the number of features is doubled. The resulting 512 feature maps are followed by two dense layers and a final sigmoid activation to obtain a probability map. D evaluates similarity of intensity distribution between I^{Trans} and I^{Ref}, and the error between generated and reference deformation fields.

2.2. Deformation Field Consistency

CycGANs learn mapping functions between two domains X and Y given training samples $x_{i=1}^{N} \in X$ and $y_{j=1}^{M} \in Y$. It has two transformations $G : X \rightarrow Y$ and $F : Y \rightarrow X$, and two adversarial discriminators D_X and D_Y , where D_X differentiates between images x and registered images $F(y)$ and D_Y distinguishes between y and $G(x)$. Here $X = I^{Flt}$ and $Y = I^{Ref}$. G registers I^{Flt} to I^{Ref} while F registers

I^{Ref} to I^{Flt}. In addition to the content loss (Eqn 2) we have: 1) an adversarial loss to match I^{Trans}'s distribution to I^{Flt}; and 2) a cycle consistency loss to ensure transformations G, F do not contradict each other.

2.2.1. Adversarial Loss

The adversarial loss function for G is given by:

$$L_{cycGAN}(G, D_Y, X, Y) = E_{y \in p_{data}(y)} \left[\log D_Y(y)\right] + E_{x \in p_{data}(x)} \left[\log\left(1 - D_Y(G(x))\right)\right], \quad (3)$$

We retain notations X, Y for conciseness. There also exists $L_{cycGAN}(F, D_X, Y, X)$ the corresponding adversarial loss for F and D_X.

2.2.2. Cycle Consistency Loss

A network may arbitrarily transform the input image to match the distribution of the target domain. Cycle consistency loss ensures that for each image $x \in X$ the reverse deformation should bring x back to the original image, i.e. $x \rightarrow G(x) \rightarrow F(G(x)) \approx x$. Similar constraints also apply for mapping F and y. This is achieved using,

$$L_{cyc}(G, F) = E_x \left\|F(G(x)) - x\right\|_1 + E_y \left\|G(F(y)) - y\right\|_1, \quad (4)$$

The full objective function is

$$L(G, F, D_{I^{Flt}}, D_{I^{Ref}}) = L_{cycGAN}(G, D_{I^{Ref}}, I^{Flt}, I^{Ref}) + L_{cycGAN}(F, D_{I^{Flt}}, I^{Ref}, I^{Flt}) + \lambda L_{cyc}(G, F) \quad (5)$$

where $\lambda = 10$ controls the contribution of the two objectives. The optimal parameters are given by:

$$G^*, F^* = \arg\min_{F,G} \max_{D_{I^{Flt}}, D_{I^{Ref}}} L(G, F, D_{I^{Flt}}, D_{I^{Ref}}) \quad (6)$$

The above formulation ensures I^{Trans} to be similar to I^{Flt} and also match I^{Ref}. We do not need to explicitly condition I^{Trans} on I^{Ref} or I^{Flt} as that is implicit in the cost function (Eqns 2,3), which allows any pair of multimodal images to be registered even if the modality was not part of the training set.

3. EXPERIMENTS AND RESULTS

We demonstrate the effectiveness of our approach on retinal and cardiac images. Details on dataset and experimental set up are provided later. Our method was implemented with Python and TensorFlow (for GANs). For GAN optimization we use Adam [15] with $\beta_1 = 0.93$ and batch normalization. The ResNet was trained with a learning rate of 0.001 and 10^5 update iterations. MSE based ResNet was used to initialize G. The final GAN was trained with 10^5 update iterations at learning rate 10^{-3}. Training and test was performed on a NVIDIA Tesla K40 GPU with 12 GB RAM.

3.1. Retinal Image Registration Results

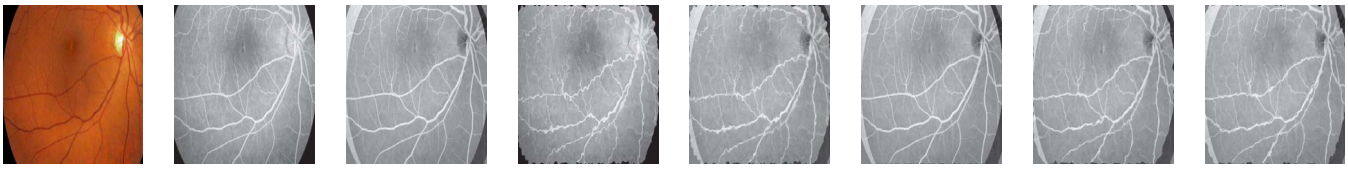
The data consists of retinal colour fundus images and fluorescein angiography (FA) images obtained from 30 normal subjects. Both images are 576×720 pixels and fovea centred [16]. Registration ground truth was developed using the Insight Toolkit (ITK). The Frangi vesselness[17] feature was utilised to find the vasculature, and the maps were aligned using sum of squared differences (SSD). Three out of 30 images could not be aligned due to poor contrast and one FA image was missing, leaving us with a final set of 26 registered pairs. We use the fundus images as I^{Ref} and generate floating images from the FA images by simulating different deformations (using SimpleITK) such as rigid, affine and elastic deformations(maximum displacement of a pixel was ± 10 mm. 1500 sets of deformations were generated for each image pair giving a total of 39000 image pairs.

Our algorithm's performance was evaluated using average registration error (Err_{Def}) between the applied deformation field and the recovered deformation field. Before applying simulated deformation the mean Dice overlap of the vasculature between the fundus and FA images across all 26 patients is 99.2, which indicates highly accurate alignment. After simulating deformations the individual Dice overlap reduces considerably depending upon the extent of deformation. The Dice value after successful registration is expected to be higher than before registration. We also calculate the 95 percentile Hausdorff Distance (HD_{95}) and the mean absolute surface distance (MAD) before and after registration. We calculate the mean square error (MSE) between the registered FA image and the original undeformed FA image to quantify their similarity. The intensity of both images was normalized to lie in $[0, 1]$. Higher values of Dice and lower values of other metrics indicate better registration. The average training time for the augmented dataset of 39000 images was 14 hours.

Table 1 shows the registration performance for GAN_{Reg}, our proposed method, and compared with the following methods: $DIRNet$ - the CNN based registration method of [7]; $Elastix$ - an iterative NMI based registration method [18]; and $GAN_{Reg_{nCyc}}$ - GAN_{Reg} without deformation consistency constraints. GAN_{Reg} has the best performance across all metrics. Figure 2 shows registration results for retinal images. GAN_{Reg} registers the images closest to the original and is able to recover most deformations to the blood vessels, followed by $DIRNet$, $GAN_{Reg-nCyc}$, and $Elastix$. It is obvious that deformation reversibility constraints significantly improve registration performance. Note that the fundus images are color while the FA images are grayscale. The reference image is a grayscale version of the fundus image.

3.2. Cardiac Image Registration Results

The second dataset is the Sunybrook cardiac dataset [19] with 45 cardiac cine MRI scans acquired on a single MRI-scanner. They consist of short-axis cardiac image slices each contain-



(a) (b) (c) (d) (e) (f) (g) (h)

Fig. 2. Example results for retinal fundus and FA registration. (a) Color fundus image; (b) Original FA image; (c) ground truth difference image before simulated deformation; (d) Deformed FA image or the floating image; Difference image (e) before registration; after registration using (f) GAN_{Reg}; (g) $DIRNet$; (f) Elastix .

	Bef. Reg.	After Registration			
		GAN_{Reg}	DIRNet [7]	Elastix [18]	$GAN_{Reg_{nCyc}}$
Dice	0.843	0.946	0.911	0.874	0.887
Err_{Def}	14.3	5.7	7.3	12.1	9.1
HD_{95}	11.4	4.2	5.9	9.7	8.0
MAD	9.1	3.1	5.0	8.7	7.2
MSE	0.84	0.09	0.23	0.54	0.37
$Time$ (s)		0.7	0.9	15.1	0.7

Table 1. Comparative average performance of different methods before and after retinal image registration. $Time$ denotes time in seconds taken to register a test image pair.

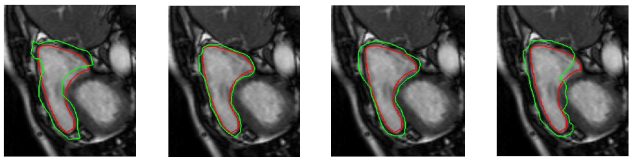
ing 20 timepoints that encompass the entire cardiac cycle. Slice thickness and spacing is 8 mm, and slice dimensions are 256×256 with a pixel size of 1.28×1.28 mm. The data is equally divided in 15 training scans (183 slices), 15 validation scans (168 slices), and 15 test scans (176 slices). An expert annotated the right ventricle (RV) and left ventricle myocardium at end-diastolic (ED) and end-systolic (ES) time points. Annotations were made in the test scans and only used for final quantitative evaluation.

We calculate Dice values before and after registration, HD_{95}, and MAD. We do not simulate deformations on this dataset and hence do not calculate Err_{Def}, MSE. Being a public dataset our results can be benchmarked against other methods. While the retinal dataset demonstrates our method's performance for multimodal registration, the cardiac dataset highlights the performance in registering unimodal dynamic images. The network trained on retinal images was used for registering cardiac data without re-training. The first frame of the sequence was used as the reference image I^{Ref} and all other images were floating images.

Table 2 summarizes the performance of different methods, and Figure 3 shows superimposed manual contour of the RV (red) and the deformed contour of the registered image (green). Better registration is reflected by closer alignment of the two contours. Once again it is obvious that GAN_{Reg} has the best performance amongst all competing methods, and its advantages over $GAN_{Reg-nCyc}$ when including deformation consistency.

	Bef. Reg.	After Registration			
		GAN_{Reg}	DIRNet [7]	Elastix [18]	$GAN_{Reg_{nCyc}}$
Dice	0.62	0.85	0.80	0.77	0.79
HD_{95}	7.79	3.9	5.03	5.21	5.12
MAD	2.89	1.3	1.83	2.12	1.98
$Time$ (s)		0.8	0.8	11.1	0.8

Table 2. Comparative average performance of different methods before and after cardiac image registration. $Time$ denotes time in seconds taken to register a test image pair..



(a) (b) (c) (d)

Fig. 3. Example results for cardiac RV registration. Superimposed contours of the ground truth (red) and deformed segmentation mask of moving image (green): (a) before registration; after registration using (b) GAN_{Reg}; (c) $DIRNet$; (d) Elastix.

4. CONCLUSION

We have proposed a GAN based method for multimodal medical image registration. Our proposed method allows fast and accurate registration and is independent of the choice of reference or floating image. Our primary contribution is in using GAN for medical image registration, and combining conditional and cyclic constraints to obtain realistic and smooth registration. Experimental results demonstrate that we perform better than traditional iterative registration methods and other DL based methods that use conventional transformation approaches such as B-splines.

5. REFERENCES

[1] M.A. Viergever, J.B.A Maintz, S. Klein, K. Murphy, M. Staring, and J.P.W. Pluim, "A survey of medical im-

age registration," *Med. Imag. Anal*, vol. 33, pp. 140–144, 2016.

[2] G. Wu, M. Kim, Q. Wang, B. C. Munsell, , and D. Shen., "Scalable high performance image registration framework by unsupervised deep feature representations learning.," *IEEE Trans. Biomed. Engg.*, vol. 63, no. 7, pp. 1505–1516, 2016.

[3] S. Miao, Y. Zheng Z.J. Wang, and R. Liao, "Real-time 2d/3d registration via cnn regression," in *IEEE ISBI*, 2016, pp. 1430–1434.

[4] R. Liao, S. Miao, P. de Tournemire, S. Grbic, A. Kamen, T. Mansi, and D. Comaniciu, "An artificial agent for robust image registration," in *AAAI*, 2017, pp. 4168–4175.

[5] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," in *NIPS*, 2015, pp. –.

[6] H. Sokooti, B. de Vos, F. Berendsen, B.P.F. Lelieveldt, I. Isgum, and M. Staring, "Nonrigid image registration using multiscale 3d convolutional neural networks," in *MICCAI*, 2017, pp. 232–239.

[7] B. de Vos, F. Berendsen, M.A. Viergever, M. Staring, and I. Isgum, "End-to-end unsupervised deformable image registration with a convolutional neural network," in *arXiv preprint arXiv:1704.06065*, 2017.

[8] C. Ledig and et. al., "Photo-realistic single image super-resolution using a generative adversarial network," *CoRR*, vol. abs/1609.04802, 2016.

[9] D Mahapatra, B Bozorgtabar, S Hewavitharanage, and R Garnavi, "Image super resolution using generative adversarial networks and local saliency maps for retinal image analysis," in *MICCAI*, 2017, pp. 382–390.

[10] P. Isola, J.Y. Zhu, T. Zhou, and A.A. Efros, "Image-to-image translation with conditional adversarial networks," in *CVPR*, 2017.

[11] J.Y. Zhu, T.park, P. Isola, and A.A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *arXiv preprint arXiv:1703.10593*, 2017.

[12] D. Rueckert, L.I Sonoda, C. Hayes, D.L.G Hill, M.O Leach, and D.J Hawkes., "Nonrigid registration using free-form deformations: application to breast mr images.," *IEEE Trans. Med. Imag.*, vol. 18, no. 8, pp. 712–721, 1999.

[13] Z. Wang and et. al., "Image quality assessment: from error visibility to structural similarity.," *IEEE Trans. Imag. Proc.*, vol. 13, no. 4, pp. 600–612, 2004.

[14] K. Simonyan and A. Zisserman., "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[15] D.P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *arXiv preprint arXiv:1412.6980,*, 2014.

[16] S.A.M Hajeb, H. Rabbani, and M.R. Akhlaghi., "Diabetic retinopathy grading by digital curvelet transform.," *Comput Math Methods Med.*, pp. 7619–01, 2012.

[17] A.F. Frangi, W.J. Niessen, K.L. Vincken, and M.A. Viergever, "Multiscale vessel enhancement filtering," in *MICCAI*, 1998, pp. 130–137.

[18] S. Klein, M. Staring, K. Murphy, M.A. Viergever, and J.P.W. Pluim., "Elastix: a toolbox for intensity based medical image registration.," *IEEE Trans. Med. Imag.*, vol. 29, no. 1, pp. 196–205, 2010.

[19] P. Radau, Y. Lu, K. Connelly, G. Paul, and et. al ., "valuation framework for algorithms segmenting short axis cardiac MRI," in *The MIDAS Journal-Cardiac MR Left Ventricle Segmentation Challenge*, 2009.