Unsupervised Learning of Diffeomorphic Image Registration via TransMorph

Junyu Chen(✉), Eric C. Frey, and Yong Du

Russell H. Morgan Department of Radiology and Radiological Science,
Johns Hopkins Medical Institutes, Baltimore, MD, USA

Abstract. In this work, we propose a learning-based framework for unsupervised and end-to-end learning of diffeomorphic image registration. Specifically, the proposed network learns to produce and integrate time-dependent velocity fields in an LDDMM setting. The proposed method guarantees a diffeomorphic transformation and allows the transformation to be easily and accurately inverted. We also showed that, without explicitly imposing a diffeomorphism, the proposed network can provide a significant performance gain while preserving the spatial smoothness in the deformation. The proposed method outperforms the state-of-the-art registration methods on two widely used publicly available datasets, indicating its effectiveness for image registration. The source code of this work is available at: `https://bit.ly/3EtYUFN`.

Keywords: Image registration · Transformer · Deep neural networks.

1 Introduction

Deformable image registration functions by establishing the spatial correspondence between the moving and the fixed images. Traditionally, image registration has been accomplished by optimizing a pair-wise objective function iteratively [3, 5, 9, 18]. Over the last decade, deep learning has emerged as a major area of research in the field of medical image analysis, including registration [4, 7, 8, 12, 15, 16, 20]. Learning-based registration models optimize a global functional for a dataset during training, thereby obviating the time-consuming and computationally expensive per-image optimization during inference.

Diffeomorphic image registration is appealing in many medical imaging applications, owing to its properties like topology preservation and transformation invertibility. A diffeomorphic transformation can be achieved via the time integration of sufficiently smooth time-stationary [1, 2, 11] or time-dependent velocity fields [3,5]. Almost all existing *end-to-end* learning-based registration models adopt stationary velocity fields because of their ease of implementation and relatively low computational cost [8,15,16]. In this work, however, we demonstrate how time-dependent velocity fields can be efficiently incorporated into an *end-to-end* deep neural network framework, which results in diffeomorphisms (an illustrative example is shown in Fig. 1) and improved registration performance.

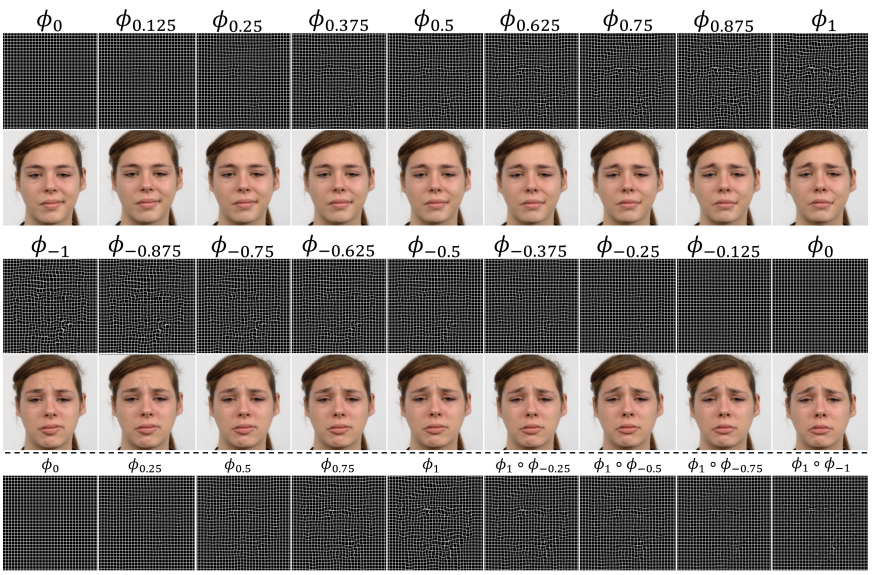

Fig. 1. Inversion and composition of the deformation fields using the proposed method. A neural network learns to generate time-dependent velocity fields for 8 time-steps.

2 Background on LDDMM

In the LDDMM setting [5], the transformation ϕ_t is computed as the flow of a time-dependent velocity field v_t, specified by the ODE: $\frac{d\phi}{dt} = v_t(\phi_t)$ with $t \in [0, 1]$. The final transformation at $t = 1$ is gained by integrating the velocity fields in time: $\phi_1 = \phi_0 + \int_0^1 v_t(\phi_t)dt$ with $\phi_0 = Id$. Then, the optimal transformation is formulated as a variational problem of the form:

$$v^* = \arg\min_v \left(\lambda \int_0^1 \|v_t\|_V^2 dt + \|I_0 \circ \phi_1 - I_1\|_{L^2}^2 \right), \tag{1}$$

where $\|\cdot\|_{L^2}$ denotes the standard L_2-norm, $\|f\|_V = \|Lf\|_{L^2}$ and L is a differential operator of the type $(-\alpha\Delta + \gamma)^\beta Id$ with $\beta > 1.5$, and I_0 and I_1 are the moving and fixed images, respectively. With sufficiently smooth v, a dffieomorphism is guaranteed in this setting.

3 Methods

In this work, a neural network was used to generate velocity fields with a predetermined discretized number of time-steps, specified by N (as shown in Fig. 2). Then, the field integration layer integrates the generated velocity fields to form the transformation at the end-point, i.e., $\phi_1 \approx Id + \sum_{t=1}^N v_t \circ \phi_t$, and the inverse transformation ϕ_{-1} is computed as $Id - \sum_{t=1}^N v_t \circ \phi_t$. The proposed

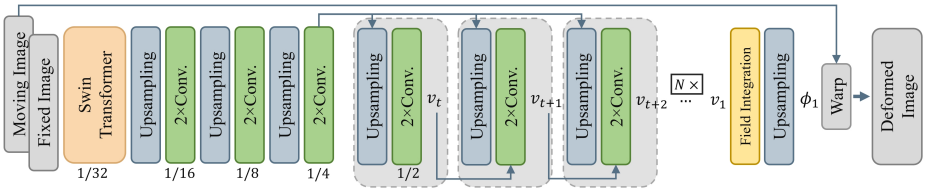

Fig. 2. Network architecture. The network integrates N time-steps of velocity fields to form a final deformation field. Note that skip connections and activation functions were omitted for visualization.

network may be trained self-supervisedly, end-to-end, using moving and fixed image pairs. We chose our previously developed `TransMorph` [6] (denoted as TM) as the base network since it showed state-of-the-art performance on several datasets. However, we underline that the proposed method is not architecture-specific and can readily be integrated into any architecture. The loss function was derived from Eqn. 1 with an additional term to account the available label map information:

$$\mathcal{L}(v, I_0, I_1) = \sum_t \|Lv_t\|_{L^2}^2 + \|I_0 \circ \phi_1 - I_1\|_{L^2}^2 + \frac{1}{M}\sum_m \|S_0^m \circ \phi_1 - S_1^m\|_{L^2}^2, \quad (2)$$

where S_0 and S_1 denote the M-channel label maps of the moving and fixed images, respectively, where each channel corresponds to the label map of an anatomical structure. We denote the model trained using this loss function as TM-TVF$_{LDDMM}$.

As a consequence of imposing a diffeomorphic transformation, excessive regularization may lead to a suboptimal registration accuracy measured by image similarity or segmentation overlap. Here, we demonstrate that by integrating time-dependent velocity fields, we could implicitly enforce transformation smoothness and improve performance without explicitly imposing a diffeomorphism. In this setting, we used a diffusion regularizer to regularize *only* the velocity field at the end-point:

$$\mathcal{L}(v, I_0, I_1) = \|\nabla v_1\|_{L^2}^2 + NCC(I_0 \circ \phi_1, I_1) + Dice(S_0 \circ \phi_1, S_1), \quad (3)$$

where ∇v is the spatial gradient operator applied to v, $NCC(\cdot)$ denotes normalized cross-correlation, and $Dice(\cdot)$ denotes Dice loss. We denote the model trained using this loss function as TM-TVF.

4 Experiments and Results

We validated the proposed method using two publicly available datasets, one in 2D and one in 3D. The 2D dataset is the Radboud Faces Database (RaFD) [13], and it comprises eight distinct facial expression images for each of 67 subjects. We randomly divided the subjects into 53, 7, and 7 subjects, and used face

	VM-2 [4]	VM-diff [8]	CycleMorph [12]	TM [6]	TM-TVF$_{LDDMM}$	TM-TVF
SSIM	0.858±0.038	0.805±0.044	0.875±0.038	0.899±0.035	0.829±0.049	**0.910±0.028**
FSIM	0.669±0.039	0.613±0.041	0.687±0.042	0.716±0.043	0.620±0.053	**0.734±0.033**
% of $\lvert J_\phi \rvert \leq 0$	0.008±0.008	**<0.001**	0.001±0.002	0.002±0.002	**<0.001**	**<0.001**

Table 1. SSIM [19] and FSIM [21] comparisons between the proposed method and the others on the RaFD dataset.

	Validation			Test		
	Dice	SDlogJ	HdDist95	Dice	SDlogJ	HdDist95
ConvexAdam [17]	0.846±0.016	**0.067±0.005**	1.500±0.304	0.81	**0.07**	1.63
LapIRN [16]	0.861±0.015	0.072±0.007	1.514±0.337	0.82	**0.07**	1.67
TM [6]	0.862±0.014	0.128±0.021	1.431±0.282	0.820	0.124	1.656
TM-TVF$_{LDDMM}$	0.833±0.016	0.090±0.005	1.630±0.353	-	-	-
TM-TVF	**0.869±0.014**	0.094±0.018	**1.396±0.297**	**0.824**	0.090	**1.633**

Table 2. Validation and test results for the OASIS dataset from the 2021 Learn2Reg challenge [10]. The validation results came from the challenge's leaderboard, whereas the test results came directly from the challenge's organizers.

images of subjects glancing in the direction of the camera. A total of 2968, 392, and 392 image pairs were used for training, validation, and testing. The images were cropped then resized into 256×256. The 3D dataset is the OASIS dataset [14] obtained from the 2021 Learn2Reg challenge [10]. This dataset comprises a total of 451 brain T2 MRI images, with 394, 19, and 38 images being used for training, validation, and testing, respectively. We trained the proposed method for 500 epochs using a learning rate of $1e^{-4}$. The number of time-steps, N, was empirically set to 8. We set $\alpha = 0.01$, $\gamma = 0.01$, and $\beta = 2$ for RaFD dataset, and $\alpha = 0.01$, $\gamma = 0.001$, and $\beta = 2$ for OASIS dataset. Note that due to the absence of segmentation in the RaFD dataset, the segmentation losses in Eqns. 1 and 2 were omitted.

Table 1 and 2 show quantitative results of the proposed models on the RaFD and OASIS datasets. On both datasets, the proposed TM-TVF yielded the highest performance against all other methods, including the first-ranking method (LapIRN [16]) from the Learn2Reg challenge. Specifically, TM-TVF outperformed its base network TM in image similarity and segmentation overlap on the two datasets, with p values < 0.0001 from paired t-tests. Although, a diffeomorphism was not explicitly guaranteed in TM-TVF, it still produced much smoother transformations than TM and VM measured by SDlogJ and the percentage of non-positive Jacobian determinant. On the other hand, although TM-TVF$_{LDDMM}$ guarantees a diffeomorphic transformation (as shown in Fig. 1, 3, and 4), it results in relatively poor registration performance, which is most likely owing to the excessive regularization imposed on the transformation.

5 Conclusion

In conclusion, we have proposed a learning-based framework for learning to generate time-dependent velocity fields in the LDDMM setting. The quantitative results show that the framework outperformed state-of-the-art registration models, indicating the effectiveness of the proposed method. Moreover, the proposed method is not architecture-specific and may be easily incorporated to improve registration performance in any network architecture.

Acknowledgment

This work was supported by a grant from the National Cancer Institute, U01-CA140204.

Appendix A. Additional qualitative Results

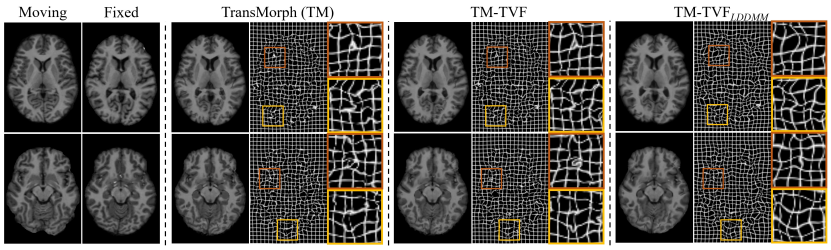

Fig. 3. Qualitative comparisons of the deformation field smoothness. TM yielded a deformation field with noticeable folded voxels, but TM-TVF generated a smoother field with state-of-the-art registration accuracy (as seen in Tables 1 and 2). TM-TVF$_{LDDMM}$ generated a highly regularized deformation field with nearly no visible folded voxels.

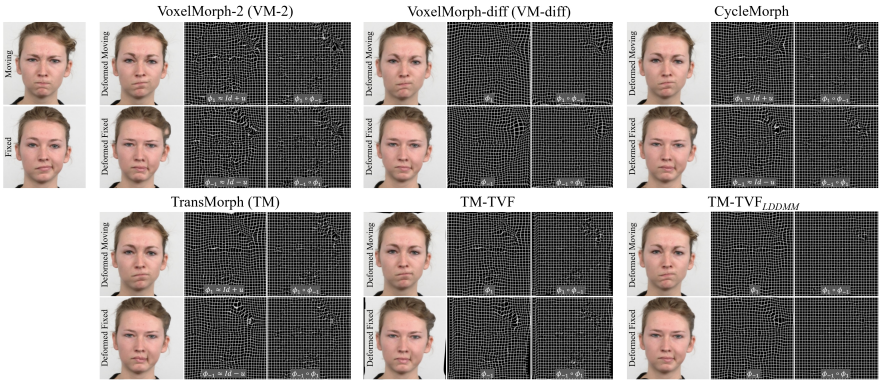

Fig. 4. Qualitative comparisons of facial expression registration. TM-TVF$_{LDDMM}$ produced a smooth and invertible transformation, but all other transformations were not. Additionally, TM-TVF yielded the best qualitative results for both forward and backward registration. Note that the transformation inversions for VM-2, CycleMorph, and TM were approximated using $Id - u$, where u denotes the displacement field.

References

1. Arsigny, V., Commowick, O., Pennec, X., Ayache, N.: A log-euclidean framework for statistics on diffeomorphisms. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 924–931. Springer (2006)
2. Ashburner, J.: A fast diffeomorphic image registration algorithm. Neuroimage **38**(1), 95–113 (2007)
3. Avants, B.B., Epstein, C.L., Grossman, M., Gee, J.C.: Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. Medical image analysis **12**(1), 26–41 (2008)
4. Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V.: Voxelmorph: a learning framework for deformable medical image registration. IEEE transactions on medical imaging **38**(8), 1788–1800 (2019)
5. Beg, M.F., Miller, M.I., Trouvé, A., Younes, L.: Computing large deformation metric mappings via geodesic flows of diffeomorphisms. International journal of computer vision **61**(2), 139–157 (2005)
6. Chen, J., Frey, E.C., He, Y., Segars, W.P., Li, Y., Du, Y.: Transmorph: Transformer for unsupervised medical image registration (2021), `https://arxiv.org/abs/2111.10480`
7. Chen, J., He, Y., Frey, E.C., Li, Y., Du, Y.: Vit-v-net: Vision transformer for unsupervised volumetric medical image registration. In: Medical Imaging with Deep Learning (2021)
8. Dalca, A.V., Balakrishnan, G., Guttag, J., Sabuncu, M.R.: Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces. Medical image analysis **57**, 226–236 (2019)
9. Heinrich, M.P., Jenkinson, M., Brady, M., Schnabel, J.A.: Mrf-based deformable registration and ventilation estimation of lung ct. IEEE transactions on medical imaging **32**(7), 1239–1248 (2013)
10. Hering, A., Hansen, L., Mok, T.C., Chung, A., Siebert, H., Häger, S., Lange, A., Kuckertz, S., Heldmann, S., Shao, W., et al.: Learn2reg: comprehensive multi-task medical image registration challenge, dataset and evaluation in the era of deep learning. arXiv preprint arXiv:2112.04489 (2021)
11. Hernandez, M., Bossa, M.N., Olmos, S.: Registration of anatomical images using paths of diffeomorphisms parameterized with stationary vector field flows. International Journal of Computer Vision **85**(3), 291–306 (2009)
12. Kim, B., Kim, D.H., Park, S.H., Kim, J., Lee, J.G., Ye, J.C.: Cyclemorph: cycle consistent unsupervised deformable image registration. Medical Image Analysis **71**, 102036 (2021)
13. Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D.H., Hawk, S.T., Van Knippenberg, A.: Presentation and validation of the radboud faces database. Cognition and emotion **24**(8), 1377–1388 (2010)
14. Marcus, D.S., Wang, T.H., Parker, J., Csernansky, J.G., Morris, J.C., Buckner, R.L.: Open access series of imaging studies (oasis): cross-sectional mri data in young, middle aged, nondemented, and demented older adults. Journal of cognitive neuroscience **19**(9), 1498–1507 (2007)
15. Mok, T.C., Chung, A.: Fast symmetric diffeomorphic image registration with convolutional neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4644–4653 (2020)
16. Mok, T.C., Chung, A.: Conditional deformable image registration with convolutional neural network. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 35–45. Springer (2021)

17. Siebert, H., Hansen, L., Heinrich, M.P.: Fast 3d registration with accurate optimisation and little learning for learn2reg 2021. arXiv preprint arXiv:2112.03053 (2021)
18. Vercauteren, T., Pennec, X., Perchant, A., Ayache, N.: Diffeomorphic demons: Efficient non-parametric image registration. NeuroImage **45**(1), S61–S72 (2009)
19. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
20. Yang, X., Kwitt, R., Styner, M., Niethammer, M.: Quicksilver: Fast predictive image registration–a deep learning approach. NeuroImage **158**, 378–396 (2017)
21. Zhang, L., Zhang, L., Mou, X., Zhang, D.: Fsim: A feature similarity index for image quality assessment. IEEE transactions on Image Processing **20**(8), 2378–2386 (2011)