Unsupervised Similarity Learning for Image Registration with Energy-Based Models

Daniel Grzech[1], Loïc Le Folgoc[1,5], Mohammad Farid Azampour[1,2,3], Athanasios Vlontzos[1,4], Ben Glocker[1], Nassir Navab[3,7], Julia Schnabel[3,6], and Bernhard Kainz[1,8]

[1] Imperial College London, [2] Munich Center for Machine Learning, [3] Technische Universität München, [4] Spotify Technology SA, [5] Télécom Paris, [6] King's College London, [7] Johns Hopkins University, [8] Friedrich-Alexander-Universität Erlangen-Nürnberg
d.grzech17@imperial.ac.uk

Abstract. We present a new model for deformable image registration, which learns in an unsupervised way a data-specific similarity metric. The proposed method consists of two neural networks, one that maps pairs of input images to transformations which align them, and one that provides the similarity metric whose maximisation guides the image alignment. We parametrise the similarity metric as an energy-based model, which is simple to train and allows us to improve the accuracy of image registration compared to other models with learnt similarity metrics by taking advantage of a more general mathematical formulation, as well as larger datasets. We also achieve substantial improvement in the accuracy of inter-patient image registration on MRI scans from the OASIS dataset compared to models that rely on traditional functions.

Keywords: energy-based models · image registration

1 Introduction

Image registration seeks to determine a transformation between a source and a target image, ensuring that corresponding regions in the transformed source and target images align and contain identical anatomical or physiological data. The main limitation of deformable image registration in medical imaging pipelines is accuracy, which remains unsatisfactory in case of many challenging problems, *e.g.* inter-patient registration of large organs or multi-modal registration. State-of-the-art image registration models, *e.g.* DLIR [35] and VoxelMorph [8, 9, 2, 3], use deep learning to learn a function that maps a pair of input images to a transformation which aligns them. Neural networks offer a speed up of many orders of magnitude over traditional iterative methods for image registration and can benefit from training on labelled data, *e.g.* organ segmentations [23]. However, most deep learning-based image registration models rely on traditional similarity metrics, *i.e.* sum of squared differences (SSD), local cross-correlation (LCC), and mutual information (MI), which means that without labelled training data, they show only moderate improvement in terms of accuracy over traditional methods.

We propose a novel methodology to improve the accuracy of image registration by learning a similarity metric in an unsupervised way. In previous literature, methods for unsupervised similarity learning in medical image registration usually rely on ad-hoc adversarial training of generative adversarial networks (GANs) [12, 11, 27]. Methods which focus on mono-modal image registration do not take advantage of their full potential to increase accuracy due to constraints on the model size and training data imposed by the mathematical formulation [15]. In this work, we adopt a fast, general, and mathematically rigorous approach. We formulate the similarity metric as an energy-based model (EBM) and make use of recent advancements in maximum likelihood (ML) training in order to learn a data-driven function from raw images.

1.1 Contributions

1. We propose a new formulation for unsupervised similarity learning in deformable medical image registration based on an energy-based model (EBM);
2. The similarity metric is trained using stochastic gradient Langevin dynamics (SGLD) and leads to improved image registration accuracy over neural networks trained with traditional metrics, with few additional parameters compared to other image registration models based on deep learning;
3. Unlike previous approaches to unsupervised similarity learning, we use full-sized images rather than image patches and enable similarity learning both in case of atlas-based and pairwise image registration.

1.2 Related work

EBMs use concepts developed in order to study the statistical properties of gases in thermal equilibrium. They are an unsupervised learning model that makes it possible to learn the underlying probability distribution of the data and draw new samples based on the trained model [29, 22].

In machine learning, EBMs were originally used for the efficient learning of representations with sparse auto-encoders [29, 22]. The first work to propose training an EBM parametrised by a deep neural network via Langevin sampling was used for image generation [37]. EBMs trained in this way were also used *e.g.* for point cloud [38] and video generation [39, 40]. Some works reinterpreted other types of machine learning models as EBMs, *e.g.* classifiers [14] and GANs [6]. However, the success of generative modelling in computer vision with variational autoencoders (VAEs) [20], GANs [13], and most recently diffusion models [30], limited the adoption of EBMs. These types of generative models were also applied in the context of medical image registration, *e.g.* VoxelMorph which used the ideas behind VAEs [9], the unsupervised similarity learning methods based on GANs [12, 11, 27], and DiffuseMorph which was based on diffusion models [19].

Non-standard similarity metrics in image registration tend to be either tailored to the problem through the insight of human experts [16] or learnt in a supervised way [4, 33, 5, 7]. There is limited existing literature on unsupervised similarity learning for deformable image registration. Most such works are based

on GANs [12, 11, 27] and rely on image patches generated from the dataset in an ad-hoc way. The pairs of input image patches are labelled as "well-registered" (positive examples x^+) or "unregistered" (negative examples x^-). x^+ do not involve the model output—instead they are defined as patches of the target image and of the weighted voxel-wise sum of the source and target images—while x^- are defined as patches of the target and the warped source images.

The unsupervised similarity learning method most similar to ours uses a variational Bayesian formulation and contrastive divergence in order to train a neural network which parameterises the similarity metric [15]. However, it suffers from slow speed due to the iterative nature of the registration algorithm. Moreover, the capacity of the neural network is constrained by intricate initialisation and memory, which makes it complicated to take advantage of large datasets and limits the improvement in accuracy.

Despite their relative lack of popularity, EBMs trained using SGLD continue to attract interest because they are simpler than other generative models [37, 40] and generate samples without the use of an explicit generator [10, 26]. To the best of our knowledge, this is the first time that EBMs are successfully used in the context of medical image registration. The code and the model weights will be made available in a public repository after the review period.

2 Method

2.1 Background

Let $\mathcal{I} \coloneqq \{I_k \mid k \in \{1, \dots, N\}\}$ be a dataset of images, where $I_k \colon \Omega_k \to [0, 1]$. We denote the dataset of image pairs by $\mathcal{D} \coloneqq \mathcal{I} \times \mathcal{I} = \{(I_t, I_s) \mid t, s \in \{1, \dots, N\}\}$, where I_t and I_s are respectively the target and the source image. In case of mono-modal image registration, we wish to align the image domains Ω_t and Ω_s using a transformation $\varphi\left(w_{ts}\right) : \Omega_s \to \Omega_t$ parameterised by w_{ts}, *i.e.*, to find parameters such that $I_t \simeq I_s\left(w_{ts}\right) \coloneqq I_s \circ \varphi^{-1}\left(w_{ts}\right)$.

Diffeomorphic transformations preserve topology and are invertible, which are important properties in medical imaging applications. In order to guarantee diffeomorphic properties and to parameterise the transformations in a memory-efficient way, we use stationary velocity free-form deformations based on cubic B-splines [32, 25, 28]. The use of spline functions implicitly ensures smoothness of the output. The numerical integration of spatially smooth stationary velocity fields via scaling-and-squaring produces diffeomorphisms [1].

Unsupervised image registration models based on deep learning compute a function $f_\theta \colon I_t \times I_s \mapsto \varphi^{-1}\left(w_{ts}\right)$, which is implemented as a neural network with parameters θ. The loss function is similar to that used in traditional iterative image registration methods:

$$\hat{\theta} = \arg\min_{\theta} \mathcal{L}_{\text{sim}}\Big(I_t, I_s\left(w_{ts}\right)\Big) + \lambda_{\text{reg}} \mathcal{L}_{\text{reg}}\left(w_{ts}\right) \tag{1}$$

where $\mathcal{L}_{\text{sim}}$ is a similarity metric, λ_{reg} is a scalar regularisation weight, and $\mathcal{L}_{\text{reg}}$ ensures that the output transformations are spatially smooth. The usual choices

of similarity metrics include SSD, LCC, and MI, *i.e.*, aside from improved generalisation and a regularisation effect, unsupervised deep learning image registration models are comparable to traditional algorithms in terms of accuracy.

2.2 Energy-based model

To maximise accuracy, we propose not only to fit a function which aligns the images in the dataset, but also to learn a data-specific similarity metric $\mathcal{L}_{\text{data}}$ to use as part of the loss function in Eq. (1).

In most deep learning-based image registration models, the function $f_\theta \coloneqq f^{(1)}_{\theta_D} \circ f^{(2)}_{\theta_E}$ is a convolutional neural network (CNN) which consists of an encoder $f^{(2)}_{\theta_E}$ with parameters θ_E and a decoder $f^{(1)}_{\theta_D}$ with parameters θ_D. The encoder $f^{(2)}_{\theta_E}$ maps input images to a latent space where similar images are close with respect to some distance [21]. Similarly, a suitable similarity metric maps similar image pairs to low energy values and disparate image pairs to high energy values.

We propose to use a learnable similarity metric implemented as an EBM. Let $x \in \mathcal{D}$ be a pair of input images and $f_{\theta_S}(x) \in \mathbb{R}_+$ an energy function, implemented as a neural network with weights θ_S. The energy function defines a Boltzmann probability distribution:

$$p_{\theta_S}(x) = \frac{1}{Z(\theta_S)} \cdot \exp\big(-f_{\theta_S}(x)\big) \tag{2}$$

$$Z(\theta_S) = \int_{\mathcal{D}} \exp\big(-f_{\theta_S}(x)\big)\mathrm{d}x \tag{3}$$

where $Z(\theta_S)$ is a normalisation constant. This energy function is the equivalent of the similarity term inside the loss function used by traditional image registration via iterative optimisation of transformation parameters individually for every pair of images in the dataset.

It is difficult to draw samples from this distribution with Markov chain Monte Carlo (MCMC) methods, *e.g.* random walk and Gibbs sampling, which suffer from long mixing times, especially for high-dimensional data like images [10]. To sample from Eq. (3), we adopt SGLD, which makes use of the energy function gradient to decrease the mixing time [36]. Eq. (4) draws samples from the distribution defined by the energy function as $t \to \infty$ and $\tau \to 0$ [36]:

$$\tilde{x}^{(t)} \leftarrow \tilde{x}^{(t-1)} - \frac{\tau}{2}\nabla_x f_{\theta_S}\big(\tilde{x}^{(t-1)}\big) + \omega^{(t)}, \omega^{(t)} \sim \mathcal{N}(0, \tau) \tag{4}$$

We wish to find a similarity metric which maximises the likelihood of images in the dataset when registering them with the image registration network f_θ trained using Eq. (1). This is done by minimising the negative log-likelihood of the image pairs aligned with the image registration model. The loss function and

its gradient are given by:

$$\begin{aligned}\mathcal{L}_{\mathrm{ML}}(\theta_S) &= \mathbb{E}_{x\sim p_{\mathcal{D}}}\left[-\log p_{\theta_S}(x) \right] \\ &= \mathbb{E}_{x\sim p_{\mathcal{D}}}\left[f_{\theta_S}(x) - \log Z(\theta_S) \right] \qquad (5)\\ \nabla_{\theta_S}\mathcal{L}_{\mathrm{ML}} &\approx \mathbb{E}_{x^+\sim p_{\mathcal{D}}}\left[\nabla_{\theta_S} f_{\theta_S}\left(x^+\right) \right] - \mathbb{E}_{x^-\sim q_{\theta_S}}\left[\nabla_{\theta_S} f_{\theta_S}\left(x^-\right) \right] \qquad (6)\end{aligned}$$

where q_{θ_S} is defined by the sampling in Eq. (4) [34]. This gradient decreases the energy of the registered image pairs x^+ and increases the energy of unregistered image pairs x^- [10]. Because the training data does not include ground truth transformations, the registered image pairs x^+ are sampled using Eq. (4), with ${x^+}^{(0)}$ set to the fixed and warped moving image, and the unregistered image pairs x^- are the fixed and the unwarped moving image, with the two images that comprise both image pairs concatenated along the channel dimension.

Algorithm 1: Model training

Data: Images $\mathcal{D} = \mathcal{I} \times \mathcal{I}$
Initialisation similarity metric f_{init}, *e.g.* SSD, LCC, or MI
Result: Image registration network $f_\theta = f^{(1)}_{\theta_D} \circ f^{(2)}_{\theta_E}$
Data-specific similarity metric f_{θ_S}

$\mathcal{L}_{\mathrm{sim}} \leftarrow f_{\mathrm{init}}$ `/* Pre-training */`
for *Epoch* **do**
  for $(I_t, I_s) \in \mathcal{D}$ **do**
    Update θ using Eq. (1) and the Adam optimiser, learning rate η_θ
    Update θ_S using Eq. (7) and the Adam optimiser, learning rate $\eta^{\mathrm{pre}}_{\theta_S}$
  end
end

$\mathcal{L}_{\mathrm{sim}} \leftarrow f_{\theta_S}$ `/* Training */`
for *Epoch* **do**
  for $(I_t, I_s) \in \mathcal{D}$ **do**
    Update θ using Eq. (1) and the Adam optimiser, learning rate η_θ
    Sample images using Eq. (4)
  end
  Update θ_S using Eq. (6) and the Adam optimiser, learning rate η_{θ_S}
end

2.3 Model training

To successfully train the model and avoid issues with training stability, we need a suitable similarity function from the start. However, unlike [15] we avoid the need to manually initialise the neural network weights, which is not feasible *e.g.* in case of MI, and train the model in two phases. In the 1st phase, we pre-train

the similarity metric f_{θ_S} to match the value of a similarity metric of choice f_{init} suitable for the task, *e.g.* SSD, LCC, or MI, by minimising the squared difference between the two values:

$$\hat{\theta}_S = \arg\min_{\theta_S} \left(f_{\theta_S}\left(I_t, I_s\right) - f_{\text{init}}\left(I_t, I_s\right) \right)^2 \tag{7}$$

The initialisation similarity metric f_{init} is also used to pre-train the image registration network. In the 2$^{\text{nd}}$ phase, we begin to use the data-specific similarity metric within the loss function in Eq. (1) by setting $\mathcal{L}_{\text{sim}} \coloneqq f_{\theta_S}$, and train the EBM using Eq. (6). The training loops are summarised in Algorithm 1.

2.4 Implementation

Figure 1 shows the neural network architecture used in the experiments. The image registration network is based on the official implementation of VoxelMorph [8, 9, 2, 3, 18], which in turn is based on the 3D UNet [31], with the transformation control point spacing of two voxels along each dimension [28]. The similarity metric f_{θ_S} is pre-trained to match SSD or MI. When initialised to SSD, it takes as input the voxel-wise squared differences between the intensities of the fixed and warped moving images. When initialised to MI, it takes as input the raw fixed and warped moving images.

We calibrate the model hyperparameters to get zero voxels with a non-positive Jacobian determinant of the transformation on the validation image pairs. The regularisation weight is set to $\lambda_{reg} = 1.5 \times 10^3$ and $\lambda_{reg} = 3.0 \times 10^5$ in case of initialisation to SSD and MI respectively. To ensure fair comparison, we train the baseline models with the regularisation weights set to the same values. We use the Adam optimiser with different learning rates for each part of the model. The similarity metric f_{θ_S} is pre-trained with a learning rate of $\eta_{\theta_S}^{\text{pre}} = 1 \times 10^{-4}$ and trained with learning rates of $\eta_{\theta_S} = 1 \times 10^{-5}$ and $\eta_{\theta_S} = 1 \times 10^{-6}$ in case of initialisation to SSD and MI respectively. The registration network is trained with a learning rate of $\eta_\theta = 2 \times 10^{-4}$. We use a step size of $\tau = 2 \times 10^{-3}$ to sample images with SGLD and stop after 2,500 iterations. We train the model with mini-batches of a single image pair, with 100 updates to the image registration network parameters θ per update to the similarity metric network parameters θ_S. The similarity metric f_{θ_S} is pre-trained to match the initialisation function f_{init} for 500 epochs and then trained for 1,000 epochs.

Training requires at least 16 GB GPU memory and on a system with a single GeForce RTX 3090 GPU takes approximately two and a half, and five days in case of the models initialised to SSD and MI respectively.

3 Results

In order to facilitate comparison to other image registration models, we train the model on data from the publicly available OASIS dataset [24], and in particular the images used in the mono-modal task at the MICCAI 2021 learn2reg

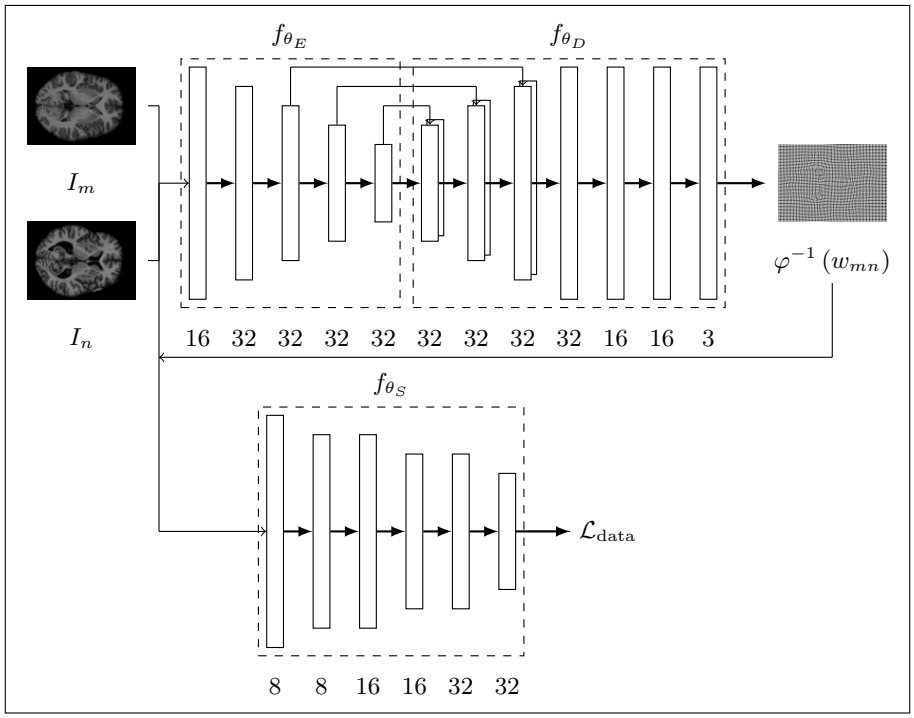

Fig. 1: Image registration network $f_{\theta_D} \circ f_{\theta_E}$ and similarity metric f_{θ_S} with f_{init} set to SSD. The arrows indicate 3D convolutional layers, with the rectangle height representing the output dimensions and the number of output channels displayed underneath. Each layer is followed by a leaky ReLU activation. When f_{init} is set to MI, the similarity metric sub-network f_{θ_S} has an encoder-decoder architecture similar to the image registration network, where the encoder has one additional layer with 64 channels and is followed by five decoder layers.

challenge [17]. The dataset consists of 447 T1-weighted magnetic resonance imaging (MRI) brain scans and includes the anatomical labels of 35 subcortical structures. 408 scans are used for training and validation, with 10 validation and 9 test image pairs. We train and test the framework on full-resolution volumes of dimensions $160 \times 192 \times 224$ voxels and 1 mm voxel length along each dimension.

To evaluate the proposed method when initialised to SSD, we compare it with VoxelMorph based on the same transformation model and trained with a traditional SSD similarity metric [28], as well as with a learnt similarity metric trained using SGLD but manually initialised to SSD [15]. To evaluate the method when initialised to MI, we compare it with VoxelMorph based on the same transformation model and trained with a traditional MI similarity metric. The model selected for evaluation has the best average accuracy among those that output zero voxels with a non-positive Jacobian determinant of the trans-

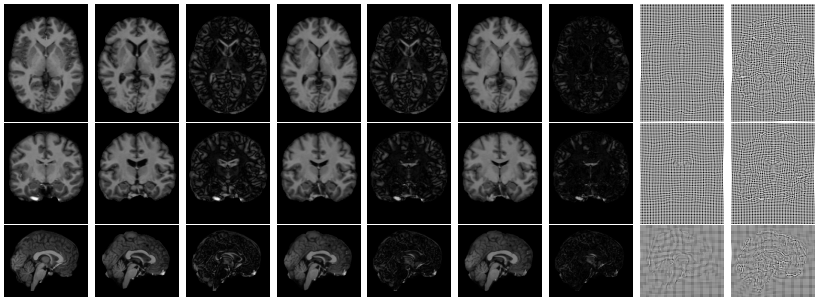

Fig. 2: The model output on a test image pair when the similarity metric is initialised to MI. Image registration models trained with the learnt similarity metrics output transformations that better align the source and the target image without decreasing the transformation smoothness.

formation on the validation image pairs. When initialised to SSD, the proposed method achieves 0.7537 average Dice score (DSC) on the test data, compared to 0.6292 in case of the image registration network trained with MI and 0.7361 in case of the learnt similarity metric manually initialised to SSD. When initialised to MI, the proposed method achieves 0.7278 average DSC on the test data, compared to 0.6831 in case of the image registration network trained with MI. These results are summarised in Figure 3. Importantly, use of the learnt similarity metrics does not have a negative impact on the output transformations smoothness. The transformations output on the test data by each of the models have zero voxels where the transformation Jacobian determinant is non-positive.

4 Conclusion

We present a new framework for unsupervised similarity learning for deformable medical image registration. We show on the examples of SSD and MI that it can significantly improve the accuracy of image registration without a negative impact on the transformation smoothness. Because the models do not require manual initialisation of neural network weights, they can be used for a wider range of image registration tasks than previous unsupervised similarity learning models. The simplicity of our similarity metric formulation as an EBM allows us to train the models on full-resolution images rather than image patches and to make use of large datasets for pairwise image registration.

4.1 Acknowledgments

We received support from the EPSRC Centre for Doctoral Training in Smart Medical Imaging, the ERC project MIA-NORMAL 101083647, the State of

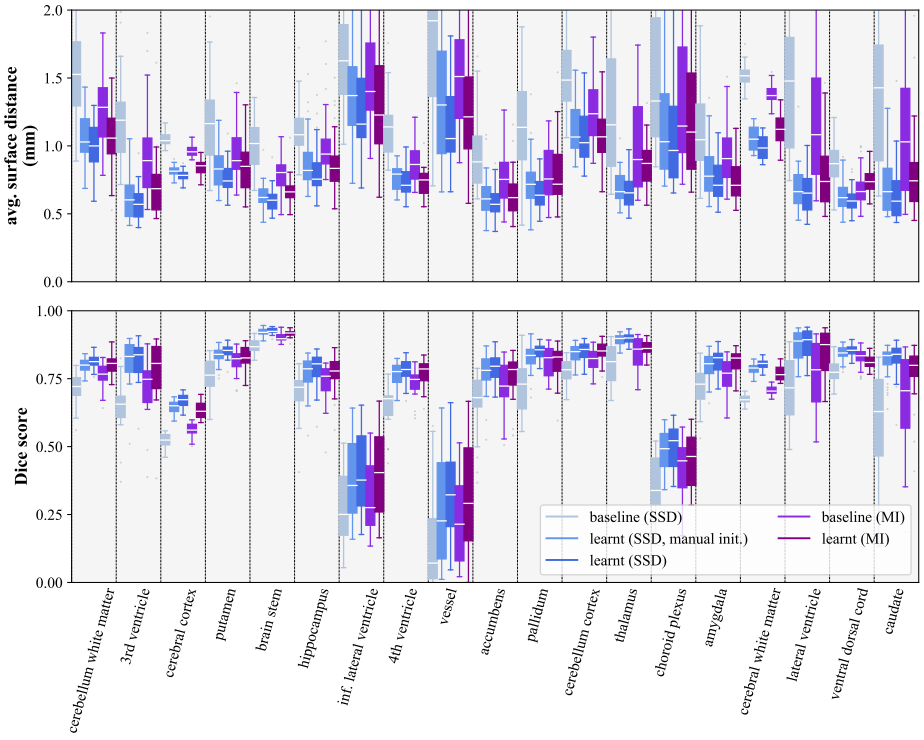

Fig. 3: Average surface distances (ASDs) and DSCs on subcortical structure segmentations when aligning test image pairs using the baseline model trained with SSD, the proposed method with the EBM manually initialised to SSD, and the proposed method, as well as the baseline model trained with MI and the proposed method. To test statistical significance we use a one-tailed Welch's t-tests at the 0.05 significance level. We found that improvement over the model trained with SSD is statistically significant for all 19/19 subcortical structure segmentations for ASD and DSC. In case of the model with manual initialisation to SSD, the improvement is statistically significant for 7/19 structures in terms of ASD and 6/19 structures in terms of DSC. For MI, the improvement is statistically significant for 16/19 structures in terms of ASD and 15/19 in terms of DSC.

Bavaria (HTA), and DFG 512819079. HPC resources were provided by NHR@FAU of FAU Erlangen-Nürnberg under the NHR projects b180dc and b143dc. NHR@FAU hardware is partially funded by DFG - 440719683.

References

1. Arsigny, V., Commowick, O., Pennec, X., Ayache, N.: A Log-Euclidean Framework for Statistics on Diffeomorphisms. In: MICCAI (2006)

2. Balakrishnan, G., Zhao, A., Sabuncu, M., Guttag, J., Dalca, A.: An Unsupervised Learning Model for Deformable Medical Image Registration. In: CVPR (2018)
3. Balakrishnan, G., Zhao, A., Sabuncu, M., Guttag, J., Dalca, A.: VoxelMorph: A Learning Framework for Deformable Medical Image Registration. IEEE TMI (2019)
4. Bronstein, M.M., Bronstein, A.M., Michel, F., Paragios, N.: Data fusion through cross-modality metric learning using similarity-sensitive hashing. In: CVPR. pp. 3594–3601 (2010)
5. Cao, X., Yang, J., Wang, L., Xue, Z., Wang, Q., Shen, D.: Deep learning based inter-modality image registration supervised by intra-modality similarity. In: MLMI MICCAI (2018)
6. Che, T., Zhang, R., Sohl-Dickstein, J., Larochelle, H., Paull, L., Cao, Y., Bengio, Y.: Your gan is secretly an energy-based model and you should use discriminator driven latent sampling. In: NeurIPS (2020)
7. Cheng, X., Zhang, L., Zheng, Y.: Deep similarity learning for multimodal medical images. In: Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization. vol. 6, pp. 248–252 (2018)
8. Dalca, A., Balakrishnan, G., Guttag, J., Sabuncu, M.: Unsupervised Learning for Fast Probabilistic Diffeomorphic Registration. In: MICCAI (2018)
9. Dalca, A., Balakrishnan, G., Guttag, J., Sabuncu, M.: Unsupervised Learning of Probabilistic Diffeomorphic Registration for Images and Surfaces. MedIA **57** (2019)
10. Du, Y., Mordatch, I.: Implicit Generation and Generalization in Energy-Based Models. In: NeurIPS (2019)
11. Fan, J., Cao, X., Wang, Q., Yap, P.T., Shen, D.: Adversarial learning for mono- or multi-modal registration. MedIA **58** (2019)
12. Fan, J., Cao, X., Xue, Z., Yap, P.T., Shen, D.: Adversarial Similarity Network for Evaluating Image Alignment in Deep Learning based Registration. In: MICCAI (2018)
13. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D.: Generative Adversarial Nets. In: NIPS (2014)
14. Grathwohl, W., Wang, K.C., Jacobsen, J.H., Duvenaud, D., Norouzi, M., Swersky, K.: Your classifier is secretly an energy based model and you should treat it like one. In: ICLR (2020)
15. Grzech, D., Azampour, M.F., Glocker, B., Schnabel, J., Navab, N., Kainz, B., Folgoc, L.L.: A variational Bayesian method for similarity learning in non-rigid image registration. In: CVPR (2022)
16. Heinrich, M.P., Jenkinson, M., Bhushan, M., Matin, T., Gleeson, F.V., Brady, S.M., Schnabel, J.A.: MIND: Modality independent neighbourhood descriptor for multi-modal deformable registration. MedIA **16**, 1423–1435 (2012)
17. Hering, A., et al.: Learn2Reg: comprehensive multi-task medical image registration challenge, dataset and evaluation in the era of deep learning. IEEE TMI (2022)
18. Hoffmann, M., Billot, B., Greve, D.N., Iglesias, J.E., Fischl, B., Dalca, A.V.: SynthMorph: Learning contrast-invariant registration without acquired images. IEEE TMI **41**, 543–558 (2022)
19. Kim, B., Han, I., Ye, J.C.: Diffusemorph: Unsupervised deformable image registration using diffusion model. In: ECCV (2022)
20. Kingma, D.P., Welling, M.: Auto-Encoding Variational Bayes. In: ICLR (2014)
21. Krebs, J., Delingette, H., Mailhé, B., Ayache, N., Mansi, T.: Learning a Probabilistic Model for Diffeomorphic Registration. IEEE TMI **38**(9) (2019)
22. LeCun, Y., Chopra, S., Hadsell, R., Ranzato, M.A., Huang, F.J.: A Tutorial on Energy-Based Learning. Predicting Structured Data pp. 1–59 (2006)

23. Lee, M.C.H., Oktay, O., Schuh, A., Schaap, M., Glocker, B.: Image-and-Spatial Transformer Networks for Structure-Guided Image Registration. In: MICCAI (2019)
24. Marcus, D.S., Fotenos, A.F., Csernansky, J.G., Morris, J.C., Buckner, R.L.: Open access series of imaging studies: Longitudinal MRI data in nondemented and demented older adults. Journal of Cognitive Neuroscience **22** (2010)
25. Modat, M., Daga, P., Cardoso, M.J., Ourselin, S., Ridgway, G.R., Ashburner, J.: Parametric non-rigid registration using a stationary velocity field. In: MMBIA. IEEE (2012)
26. Nijkamp, E., Hill, M., Zhu, S.C., Wu, Y.N.: Learning Non-Convergent Non-Persistent Short-Run MCMC Toward Energy-Based Model. In: NeurIPS (2019)
27. Qin, C., Shi, B., Liao, R., Mansi, T., Rueckert, D., Kamen, A.: Unsupervised Deformable Registration for Multi-Modal Images via Disentangled Representations. In: IPMI (2019)
28. Qiu, H., Qin, C., Schuh, A., Hammernik, K., Rueckert, D.: Learning Diffeomorphic and Modality-invariant Registration using B-splines. In: MIDL (2021)
29. Ranzato, M., Poultney, C., Chopra, S., LeCun, Y.: Efficient Learning of Sparse Representations with an Energy-Based Model. In: NIPS (2006)
30. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022)
31. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI (2015)
32. Rueckert, D., Sonoda, L.I., Hayes, C., Hill, D.L.G., Leach, M.O., Hawkes, D.J.: Nonrigid registration using free-form deformations: Application to breast MR images. IEEE TMI **18**(8), 712–721 (1999)
33. Simonovsky, M., Gutiérrez-Becker, B., Mateus, D., Navab, N., Komodakis, N.: A deep metric for multimodal registration. In: MICCAI. pp. 10–18 (2016)
34. Turner, R.: CD notes (2005)
35. de Vos, B.D., Berendsen, F.F., Viergever, M., Sokooti, H., Staring, M., Išgum, I.: A deep learning framework for unsupervised affine and deformable image registration. MedIA **52** (2019)
36. Welling, M., Teh, Y.W.: Bayesian Learning via Stochastic Gradient Langevin Dynamics. In: ICML. pp. 681–688 (2011)
37. Xie, J., Lu, Y., Zhu, S.C., Wu, Y.N.: A Theory of Generative ConvNet. In: ICML (2016)
38. Xie, J., Xu, Y., Zheng, Z., Zhu, S.C., Wu, Y.N.: Generative pointnet: Deep energy-based learning on unordered point sets for 3d generation, reconstruction and classification. In: CVPR (2021)
39. Xie, J., Zhu, S.C., Wu, Y.N.: Synthesizing dynamic patterns by spatial-temporal generative convnet. In: CVPR (2017)
40. Xie, J., Zhu, S.C., Wu, Y.N.: Learning Energy-Based Spatial-Temporal Generative ConvNets for Dynamic Patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence **43**(2), 516–531 (2021)