Semantic similarity metrics for learned image registration

Steffen Czolbe   PER.SC@DI.KU.DK
Oswin Krause   OSWIN.KRAUSE@DI.KU.DK
Department of Computer Science, University of Copenhagen, Denmark

Aasa Feragen   AFHAR@DTU.DK
DTU Compute, Technical University of Denmark, Denmark

Abstract

We propose a semantic similarity metric for image registration. Existing metrics like Euclidean Distance or Normalized Cross-Correlation focus on aligning intensity values, giving difficulties with low intensity contrast or noise. Our approach learns dataset-specific features that drive the optimization of a learning-based registration model. We train both an unsupervised approach using an auto-encoder, and a semi-supervised approach using supplemental segmentation data to extract semantic features for image registration. Comparing to existing methods across multiple image modalities and applications, we achieve consistently high registration accuracy. A learned invariance to noise gives smoother transformations on low-quality images. Code and experiments are available at `github.com/SteffenCzolbe/DeepSimRegistration`.

Keywords: Image Registration, Deep Learning, Representation Learning

1. Introduction

Deformable registration, or nonlinear image alignment, is a fundamental tool in medical imaging. Registration models find correspondences between a set of images and derive a geometric transformation to align them. Most algorithmic and deep learning-based methods solve the registration problem by the minimization of a loss function, consisting of a similarity metric and a regularization term ensuring smoothness of the transformation. The similarity metric is essential to the optimization; it judges the quality of the match between registered images and has a strong influence on the result.

Pixel-based similarity metrics like euclidean distance and patch-wise cross-correlation are well explored within algorithmic (Avants et al., 2011, 2008; Faisal Beg et al., 2005; Rueckert et al., 1999; Thirion, 1998; Vercauteren et al., 2007) and deep learning based (Alvén et al., 2019; Balakrishnan et al., 2019; Dalca et al., 2018, 2019; de Vos et al., 2019; Liu et al., 2019; Lee et al., 2019; Hu et al., 2019a,b; Yang et al., 2017; Xu and Niethammer, 2019) image registration. These metrics assume that if the image intensities are aligned, or strongly correlated, the images are well aligned. Each choice of metric adds additional assumptions on the characteristics of the specific dataset. Thus, a common methodological approach is to trial registration models with multiple different pixel-based metrics, and choose the metric performing best on the dataset (Balakrishnan et al., 2019; Hu et al., 2019a).

The shortcomings of pixel-based similarity metrics, such as blur in the generated images, have been studied substantially in the image generation community (Hou et al., 2017; Zhang et al., 2018) and have been superseded by deep similarity metrics approximating human visual perception. Here, features are extracted from neural networks pre-trained on image-classification tasks (Deng et al., 2009). Performance can further be improved by fine-tuning the features to human perception (Czolbe et al., 2020; Zhang et al., 2018), leading to generative models that produce photo-realistic images. We propose to apply deep similarity metrics within image registration to achieve a similar increase of performance for registration models.

Contribution We propose a data-driven similarity metric for image registration based on the alignment of semantic features. We explore both unsupervised (using auto-encoders) and semi-supervised methods (using a segmentation model) to learn filters of semantic importance to the dataset. We use the learned features to construct a similarity metric used for training a registration model, and validate our approach on three biomedical datasets of different image modalities and applications. Across all datasets, our method achieves consistently high registration accuracy, outperforming even metrics utilizing supervised information on two out of three datasets, while yielding an inconclusive result on the third one. Our models learn to ignore noisy image patches, leading to smoother transformations on low-quality data.

2. Background & related work

Image registration Most image registration frameworks model the problem as finding a transformation $\Phi\colon \Omega \to \Omega$ that aligns a moving image $\mathbf{I}\colon \Omega \to \mathbb{R}$ to a fixed image $\mathbf{J}\colon \Omega \to \mathbb{R}$. The morphed source image, obtained by applying the transformation, is expressed by function composition as $\mathbf{I} \circ \Phi$. The domain Ω denotes the set of all coordinates $\mathbf{x} \in \mathbb{R}^d$ within the image[1]. Images record intensity at discrete pixel-coordinates $\mathbf{p}$ but can be viewed as a continuous function by interpolation.

The transformation is found through iterative algorithms (Avants et al., 2008; Faisal Beg et al., 2005; Rueckert et al., 1999; Thirion, 1998; Vercauteren et al., 2007), or predicted with learning based techniques (Balakrishnan et al., 2019; Dalca et al., 2018; de Vos et al., 2019; Liu et al., 2019; Lee et al., 2019; Hu et al., 2019b; Yang et al., 2017; Xu and Niethammer, 2019). In both cases, the optimal transformation is found by minimization of a similarity measure D and a λ-weighted regularizer R, expressed via the loss function

$$L(\mathbf{I}, \mathbf{J}, \Phi) = D(\mathbf{I} \circ \Phi, \mathbf{J}) + \lambda R(\Phi) \ . \tag{1}$$

As many non-linear transformation models are over-parametrized and have multiple minima, regularization is necessary. Smooth transformation fields, that avoid folds or gaps, are assumed to be physically plausible and encouraged by the regularizer. We use the *diffusion regularizer* throughout this paper, which is defined as

$$\Phi(\mathbf{p}) = \mathbf{p} + \mathbf{u}(\mathbf{p}), \ \ R(\Phi) = \sum_{\mathbf{p} \in \Omega} \|\nabla \mathbf{u}(\mathbf{p})\|^2 \ , \tag{2}$$

with the gradient of the displacement field $\nabla \mathbf{u}(\mathbf{p})$ approximated via finite differences.

Similarity metrics for image registration Denote by D the (dis-)similarity between the morphed moving image $\mathbf{I} \circ \Phi$ and the fixed image $\mathbf{J}$. Pixel-based metrics are well explored within algorithmic image registration, a comparative evaluation is given by Avants et al. (2011). We briefly recall the two most popular choices, *mean squared error* (MSE) and *normalized cross correlation* (NCC). The pixel-wise MSE is intuitive and computationally efficient. It is derived from maximizing the negative log-likelihood of a Gaussian normal distribution, making it an appropriate choice under the

---

1. While the domain Ω is continuous in $\mathbb{R}^d$, recorded images and computations thereon are discrete. For simplicity of notation, we denote both the continuous and discrete domain as Ω. We implement $\sum_{\mathbf{p} \in \Omega}$ as a vectorized operation over the discrete pixel/voxel-coordinates and calculate $|\Omega|$ as the total count of discrete pixels/voxels of the image. The transformation Φ is implemented as a map from a discrete domain to a continuous one, and the sampling of continuous points from a discrete image is implemented via bi-/tri-linear interpolation.

assumption of Gaussian noise. On a grid of discrete points $\mathbf{p}$ from domain Ω, the MSE is defined as $\mathrm{MSE}(\mathbf{I} \circ \Phi, \mathbf{J}) = \frac{1}{|\Omega|} \sum_{\mathbf{p} \in \Omega} \|\mathbf{I} \circ \Phi(\mathbf{p}) - \mathbf{J}(\mathbf{p})\|^2$.

Patch-wise normalized cross correlation is robust to variations in brightness and contrast, making it a popular choice for images recorded with different acquisition tools and protocols, or even across image modalities. For two image patches $\mathbf{A}, \mathbf{B}$, represented as column-vectors of length N with patch-wise means $\bar{\mathbf{A}}, \bar{\mathbf{B}}$ and variance $\sigma_{\mathbf{A}}^2, \sigma_{\mathbf{B}}^2$, it is defined as

$$\mathrm{NCC}_{\mathrm{patch}}(\mathbf{A}, \mathbf{B}) = \sum_{n=1}^{N} \frac{(\mathbf{A}_n - \bar{\mathbf{A}})(\mathbf{B}_n - \bar{\mathbf{B}})}{\sigma_{\mathbf{A}} \sigma_{\mathbf{B}}} \quad . \tag{3}$$

The Patch-wise similarities are then averaged over the image (Gee et al., 1993; Avants et al., 2008). Note that an alternative, computationally efficient variant of NCC is sometimes used in image registration (Avants et al., 2011). If annotations are available, these unsupervised similarity measures can be extended by a *supervised* component to measure both intensity differences and the alignment of annotated label maps (Balakrishnan et al., 2019).

Learned similarity metrics for image registration   While deep-learning-based image registration has received much interest recently, similarity metrics utilizing the compositional and data-driven advantages of neural networks remain under-explored. Some current works explore how to incorporate scale-space into learned registration models, but similarity metrics remain pixel-based (Hu et al., 2019a; Li and Fan, 2018). Learned similarity metrics are proposed by Haskins et al. (2019) and Krebs et al. (2017), but both approaches require ground truth registration maps, which are either synthetically generated or manually created by a medical expert. Lee et al. (2019) propose to learn annotated structures of interest as part of the registration model to aid alignment, but the method discards sub-regional and non-annotated structures.

Learned common data representations and similarity metrics are frequently used in multi-modal image registration (Heinrich et al., 2012; Chen et al., 2016; Simonovsky et al., 2016; Pielawski et al., 2020). While these approaches learn common representations from well-aligned images of multiple modalities, we aim to find a semantically augmented representation of images of a single modality.

3. Method

A discussion of NCC   Our design of a semantic similarity metric starts by examining the popular NCC metric. We see that NCC between image patches $\mathbf{A}$ and $\mathbf{B}$ is equivalent to the cosine-similarity between the corresponding mean-centered vectors $f(\mathbf{A}) = \mathbf{A} - \bar{\mathbf{A}}$ and $f(\mathbf{B}) = \mathbf{B} - \bar{\mathbf{B}}$:

$$\mathrm{NCC}_{\mathrm{patch}}(\mathbf{A}, \mathbf{B}) = \frac{\langle f(\mathbf{A}), f(\mathbf{B}) \rangle}{\|f(\mathbf{A})\| \|f(\mathbf{B})\|} \quad , \tag{4}$$

with scalar product $\langle \cdot, \cdot \rangle$ and euclidean norm $\| \cdot \|$. Thus, an alternative interpretation of the NCC similarity measure is the cosine-similarity between two feature descriptors in a high-dimensional space. The descriptor is given by the intensity values of a centered image patch centered at a pixel $\mathbf{p}$. We will construct a similar metric, using semantic feature descriptors instead.

A semantic similarity metric for image registration   To align areas of similar semantic value, we propose a similarity metric based on the agreement of semantic feature representations of two images. Semantic feature maps are obtained by a *feature extractor*, which is pre-trained on a
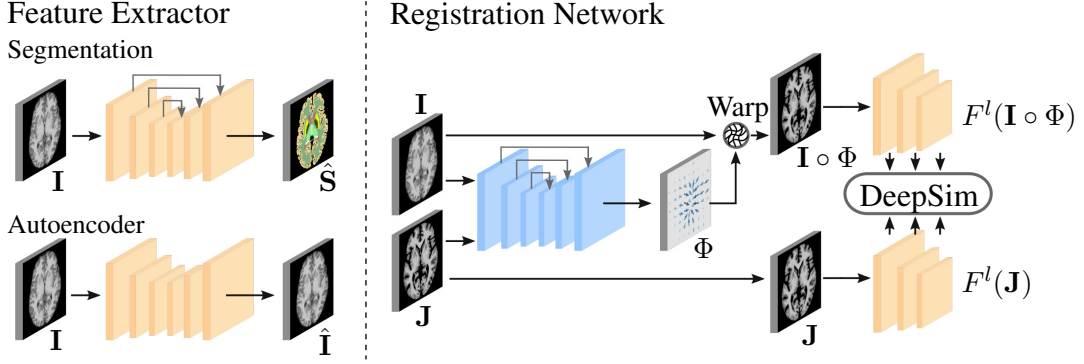
Figure 1: Two-step training: First, the feature extractor (yellow) is trained. We test both a segmentation model, and an auto-encoder. Next, the feature extractor weights are frozen and used in the loss computation of the registration network (blue).

surrogate task. To capture alignment of both localized, concrete features, and global, abstract ones, we calculate the similarity at multiple layers of abstraction. Given a set of feature-extracting functions $F^l \colon \mathbb{R}^{\Omega \times C} \to \mathbb{R}^{\Omega_l \times C_l}$ for L layers, we define

$$\mathrm{DeepSim}(\mathbf{I} \circ \Phi, \mathbf{J}) = \frac{1}{L} \sum_{l=1}^{L} \frac{1}{|\Omega_l|} \sum_{\mathbf{p} \in \Omega_l} \frac{\left\langle F_{\mathbf{p}}^l(\mathbf{I} \circ \Phi), F_{\mathbf{p}}^l(\mathbf{J}) \right\rangle}{\|F_{\mathbf{p}}^l(\mathbf{I} \circ \Phi)\| \|F_{\mathbf{p}}^l(\mathbf{J})\|} \ , \tag{5}$$

where $F_{\mathbf{p}}^l(\mathbf{J})$ denotes the lth layer feature extractor applied to image $\mathbf{J}$, at spatial coordinate $\mathbf{p}$. It is a vector of C_l output channels, and the spatial size of the lth feature map is denoted by $|\Omega_l|$.

Similarly to NCC, the neighborhood of the pixel is considered in the metric, as we compose F^l of convolutional filters with increasing receptive area of the composition. In contrast, it is not necessary to zero-mean the feature descriptors, as the semantic feature representations are trained to be robust to variances in image brightness present in the training data.

Feature extraction To aid registration, the functions $F^l(\cdot)$ should extract features of semantic relevance for the registration task, while ignoring noise and artifacts. We extract features from the encoding branch of networks trained on two surrogate tasks.

First, if segmentation masks are available, we can learn features on a supplementary segmentation task. Segmentation models excel at learning relevant kernels for the data while attaining invariance towards non-predictive features like noise, but require an annotated dataset for training. We denote the proposed similarity metric with feature extractors conditioned on this task as $\mathrm{DeepSim_{seg}}$.

Second, we can learn an abstract feature representation of the dataset in an unsupervised setting with auto-encoders. Auto-encoders learn efficient data encoding by training the network to ignore signal noise. A benefit of this approach is that no additional annotations are required. While variational methods for encoding tasks have several advantages, we choose a deterministic auto-encoder for its simplicity and lack of hyperparameters. We denote the similarity metric with feature extractors conditioned on this task as $\mathrm{DeepSim_{ae}}$.

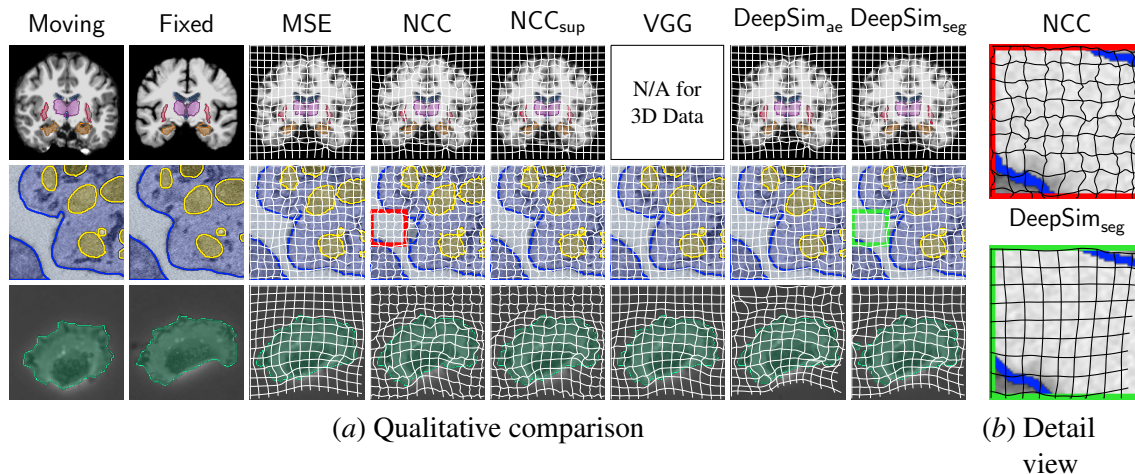(a) Qualitative comparison

(b) Detail view

Figure 2: Left: Qualitative comparison of registration models. Rows: Datasets Brain-MRI, Platelet-EM, PhC-U373. Columns 3-8: Registration models trained with various similarity metrics. Right: Detail view of highlighted noisy background areas on the Platelet-EM dataset. Select segmentation classes annotated in color. The transformation is visualized by morphed grid-lines.

4. Experiments

We empirically compare registration models trained with the unsupervised $DeepSim_{ae}$ and semi-supervised $DeepSim_{seg}$ to the baselines MSE, NCC, NCC_{sup} (NCC with supervised information), and VGG (a VGG-net based deep similarity metric from image generation). Our implementation of the baseline methods follows Avants et al. (2011), Balakrishnan et al. (2019), and Hou et al. (2017).

As our goal is not to advance the state of the art for any particular registration task, but instead to explore the value of our loss-function in a generic setting, we use well-established 2D

Figure 3: Hyperparameter tuning. For each model, we select the regularizer weighting factor λ with the highest validation mean dice overlap for further evaluation. Color scheme follows Figure 4.

and 3D U-Net (Ronneberger et al., 2015) architectures for both registration and segmentation models, sketched in Figure 1. For the auto-encoder task, we use the same architecture, but without the shortcut connections. Each network consists of three encoder and decoder stages. Each stage consists of one batch normalization (Ioffe and Szegedy, 2015), two convolutional, and one dropout layer (Gal, Yarin and Ghahramani, 2016). After the final decoder step, we smooth the model output with three more convolutional layers. We experimented with deeper architectures but found they do not increase performance. The activation function is LeakyReLu throughout the network, Softmax for the final layer of the segmentation network, Sigmoid for the final layer of the auto-encoder, and linear for the final layer of the registration network. The stages have $64, 128, 256$ channels for 2d datasets, and $32, 64, 128$ channels for 3d.
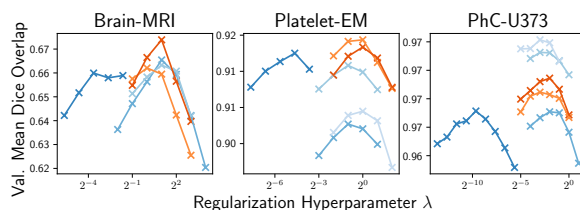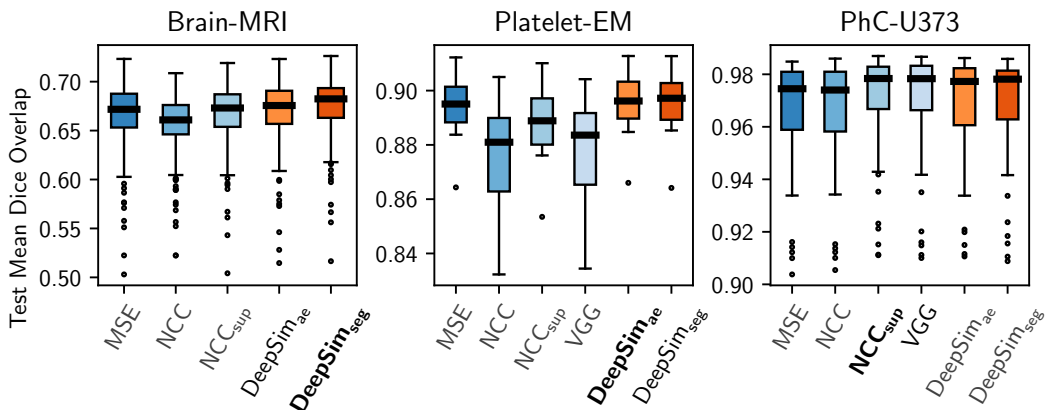
Figure 4: Mean dice overlap of registration models on the test set. Baselines in shades of blue, ours in red. Label of the best metric in bold, 2nd best black, others in grey. Boxplot with median, quartiles, deciles and outliers.

The segmentation model is trained with a cross-entropy loss function, the auto-encoder with the mean squared error, and the registration network with the loss given by Eq. 1. The optimization algorithm for all models is ADAM (Kingma and Ba, 2015), the initial learning rate is 10^{-4}, decreasing by a factor of 10 each time the validation loss plateaus. All models are trained until convergence. Due to the large 3D volumes involved, the choice of batch-size is often limited by available memory. We sum gradients over multiple passes to arrive at effective batch-sizes of 3-5 samples. The hyperparameter λ is tuned for each model independently on the validation set, we plot the validation score of tested values in Figure 3.

To show that our approach applies to a variety of registration tasks, we validate it on three 2D and 3D datasets of different modalities: 4000 T1-weighted *Brain-MRI* scans (Di Martino et al., 2014; LaMontagne et al., 2019), 74 slices of human blood cells of the *Platelet-EM* dataset (Quay et al., 2018), and 230 time-steps of the cell-tracking video *PhC-U373* (Maška et al., 2014; Ulman et al., 2017). All datasets are pre-aligned through affine transformations and split into train, validation, and test sections. We augment each pair of images with random affine transformations during training.

5. Results

Registration accuracy We measure the mean Sørensen Dice coefficient on the unseen test-set in Figure 4. Statistical significance testing of the results is performed with the Wilcoxon signed rank test for paired samples. A significance level of 5% gives Bonferroni-adjusted significance threshold $p = 0.002$. We further measure the effect size with Cohen's d and show the results in Table 1. Models trained with our proposed DeepSim_{ae} and DeepSim_{seg} outperform all baselines on the Brain-MRI and Platelet-EM datasets, with strong statistical significance and effect size. On the PhC-U373 dataset, all models achieve high dice-overlaps of > 0.97.

We monitor the mean dice overlap during training. The training accuracy is, with few exceptions, similar to the test accuracy, indicating that results generalize well. The empirical convergence speeds of the tested metrics differ. We observe that the DeepSim_{seg} converges faster than the baselines, especially in the first few epochs of training. See appendix A for a convergence plot.

Dataset	Method	Baseline							
		MSE		NCC		NCC_{sup}		VGG	
		p	d	p	d	p	d	p	d
Brain-MRI	$DeepSim_{ae}$	<0.001	0.14	<0.001	0.43	<0.001	0.10	–	–
	$DeepSim_{seg}$	<0.001	0.30	<0.001	0.60	<0.001	0.25	–	–
Platelet-EM	$DeepSim_{ae}$	0.016	0.12	<0.001	1.14	<0.001	0.51	<0.001	1.06
	$DeepSim_{seg}$	0.034	0.10	<0.001	1.12	<0.001	0.49	<0.001	1.04
PhC-U373	$DeepSim_{ae}$	<0.001	0.10	<0.001	0.11	<0.001	-0.06	<0.001	-0.05
	$DeepSim_{seg}$	<0.001	0.12	<0.001	0.13	0.002	-0.03	0.003	-0.02

Table 1: Results of the statistical significance test, performed with the Wilcoxon signed rank test for paired samples. Effect size measured with Cohen's d. Statistically insignificant results (significance level 0.05, Bonferroni-corrected to $p > 0.002$) and very small effect sizes ($|d| < 0.1$) in grey.

Qualitative examples & transformation grids We plot the fixed and moving images $\mathbf{I}$, $\mathbf{J}$ and the morphed image $\mathbf{I} \circ \Phi$ for each similarity metric model along with a more detailed view of a noisy patch of the Platelet-EM dataset in Figure 2, and perform a quantitative analysis of the transformation in Appendix B. On models trained with the baselines, we find strongly distorted transformation fields in noisy areas of the images. In particular, models trained with NCC and NCC_{sup} produce very irregular transformations, despite careful tuning of the regularization-hyper-parameter. The model trained by $DeepSim_{seg}$ is more invariant towards noise.

Anatomical regions The Brain-MRI dataset is annotated with the anatomical regions of the brain. We plot the dice overlap per region in a boxplot in Figure 5, and highlight regions where both of our metrics perform better than all baselines bold. Baseline methods (blue) perform very similar, despite NCC_{sup} as a supervised metric requiring more information over the unsupervised MSE and NCC.

6. Discussion & Conclusion

Registration models trained with DeepSim achieve high registration accuracy across datasets, allowing improved downstream analysis and diagnosis. Its consistency makes testing multiple traditional metrics unnecessary; instead of empirically determining whether MSE or NCC captures the characteristics of a data-set best, we can use DeepSim to learn the relevant features from the data.

Our experiments show that the availability of annotated anatomical regions can help in learning semantic features, but is not a necessity. Gains in performance in deep learning often require large, annotated datasets, which are expensive and time-consuming to obtain in biomedical settings. The feature extractor of $DeepSim_{ae}$ was trained on the unsupervised autoencoding task, requiring no additional data aside from the intensity images to be registered. It performed similarly to the semi-supervised $DeepSim_{seg}$ on two of the three datasets and outperformed the baselines.

The analysis of noisy patches in Figure 2 and Appendix B highlights a learned invariance to noise. The pixel-based similarity metrics are distracted by artifacts, leading to overly-detailed transformation fields. Models trained with DeepSim do not show this problem. While smoother transformation fields can be obtained for all metrics by strengthening the regularizer, this would
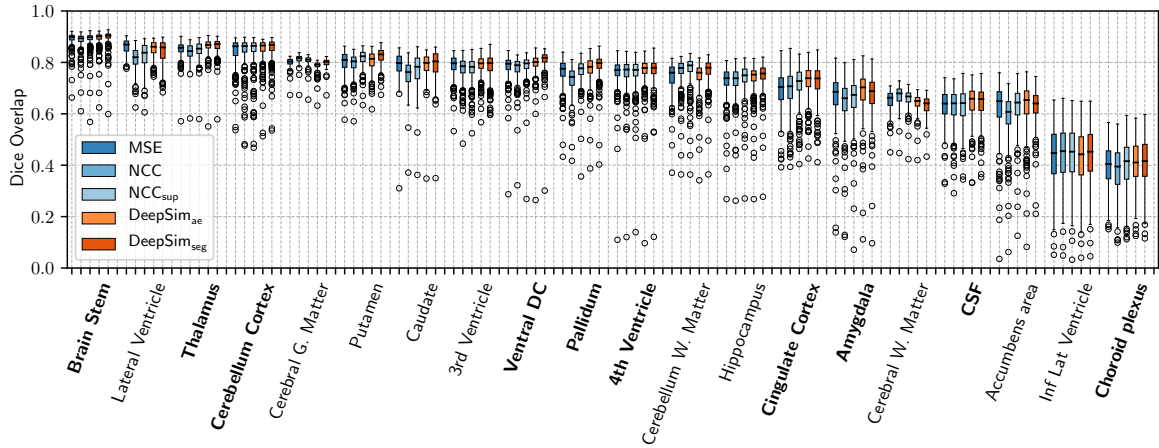
Figure 5: Dice overlaps of the anatomical regions of the Brain-MRI dataset. Baselines in shades of blue, our methods in red. Bold labels for regions where *both* of our methods score higher than all baselines. We combined labels of the left and right brain hemispheres into a single class. The boxplot shows median, quartiles, deciles and outliers.

negatively impact the registration accuracy of anatomically significant regions. Accurate registration of noisy, low-quality images allows for shorter acquisition time and reduced radiation dose in medical applications.

A weakness of DeepSim is the need to train a separate model for feature extraction. The design, training, and testing of a second model takes additional resources, and the presented approach necessitates a dataset to train the feature extractor with. In the context of deep learning-based registration, DeepSim_{ae} requires no additional data to what is required to train the registration model, while DeepSim_{seg} requires additional label maps. Compared to algorithmic registration methods, which optimize the registration map for each pair of input images separately, any deep-learning approach requires additional data.

DeepSim is a general metric, applicable to image registration tasks of all modalities and anatomies. Beyond the presented datasets, our good results in the presence of noise let us hope that DeepSim will improve registration accuracy in domains such as low dose CT, ultrasound, or microscopy, where details are often hard to identify, and image quality is poor. We further emphasize that the application of DeepSim is limited neither to deep learning nor to image registration. In Algorithmic image registration, a similarity-based loss is minimized via gradient descent-based methods. DeepSim can be applied to drive algorithmic methods, improving their performance by aligning deep, semantic feature embeddings. Similarly, DeepSim could be used for other image regression tasks, such as image synthesis, -translation, or -reconstruction.

Acknowledgments

References

Jennifer Alvén, Kerstin Heurling, Ruben Smith, Olof Strandberg, Michael Schöll, Oskar Hansson, and Fredrik Kahl. A Deep Learning Approach to MR-less Spatial Normalization for Tau PET Images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 355–363. 2019.

B. B. Avants, C. L. Epstein, M. Grossman, and J. C. Gee. Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. *Medical Image Analysis*, 12(1):26–41, 2008.

Brian B. Avants, Nicholas J. Tustison, Gang Song, Philip A. Cook, Arno Klein, and James C. Gee. A reproducible evaluation of ANTs similarity metric performance in brain image registration. *NeuroImage*, 54(3):2033–2044, 2011.

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.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*, 2016.

Steffen Czolbe, Oswin Krause, Ingemar Cox, and Christian Igel. A Loss Function for Generative Neural Networks Based on Watson's Perceptual Model. *Advances in Neural Information Processing Systems*, 2020.

Adrian V. Dalca, Guha Balakrishnan, John Guttag, and Mert R. Sabuncu. Unsupervised Learning for Fast Probabilistic Diffeomorphic Registration. *Medical Image Computing and Computer Assisted Intervention*, pages 729–738, 2018.

Adrian V. Dalca, Marianne Rakic, John Guttag, and Mert R. Sabuncu. Learning Conditional Deformable Templates with Convolutional Networks. *Advances in Neural Information Processing Systems*, pages 806—-818, 2019.

Bob D. de Vos, Floris F. Berendsen, Max A. Viergever, Hessam Sokooti, Marius Staring, and Ivana Išgum. A deep learning framework for unsupervised affine and deformable image registration. *Medical Image Analysis*, 52:128–143, 2019.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

Adriana Di Martino, Chao-Gan Yan, Qingyang Li, and Others. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular psychiatry*, 19(6):659–667, 2014.

Mirza Faisal Beg, Michael I Miller, Alain Trouvétrouv, and Laurent Younes. Computing Large Deformation Metric Mappings via Geodesic Flows of Diffeomorphisms. *International Journal of Computer Vision*, 61(2):139–157, 2005.

Zoubin Gal, Yarin and Ghahramani. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *International Conference on Machine Learning*, pages 1050—-1059, 2016.

James C Gee, Martin Reivich, and Ruzena Bajcsy. Elastically deforming a three-dimensional atlas to match anatomical brain images. *IRCS Technical Reports Series*, (192), 1993.

Grant Haskins, Jochen Kruecker, Uwe Kruger, Sheng Xu, Peter A. Pinto, Brad J. Wood, and Pingkun Yan. Learning deep similarity metric for 3D MR–TRUS image registration. *International Journal of Computer Assisted Radiology and Surgery*, 14(3):417–425, 2019.

Mattias P. Heinrich, Mark Jenkinson, Manav Bhushan, Tahreema Matin, Fergus V. Gleeson, Sir Michael Brady, and Julia A. Schnabel. MIND: Modality independent neighbourhood descriptor for multi-modal deformable registration. *Medical Image Analysis*, 16(7):1423–1435, 2012.

Xianxu Hou, Linlin Shen, Ke Sun, and Guoping Qiu. Deep feature consistent variational autoencoder. In *Winter Conference on Applications of Computer Vision*, pages 1133–1141. IEEE, 2017.

Xiaojun Hu, Miao Kang, Weilin Huang, Matthew R. Scott, Roland Wiest, and Mauricio Reyes. Dual-Stream Pyramid Registration Network. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 382–390. 2019a.

Yipeng Hu, Eli Gibson, Dean C. Barratt, Mark Emberton, J. Alison Noble, and Tom Vercauteren. Conditional Segmentation in Lieu of Image Registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 401–409. 2019b.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456. International Machine Learning Society (IMLS), 2015.

Diederik P. Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Julian Krebs, Tommaso Mansi, Hervé Delingette, Li Zhang, Florin C. Ghesu, Shun Miao, Andreas K. Maier, Nicholas Ayache, Rui Liao, and Ali Kamen. Robust non-rigid registration through agent-based action learning. In *Lecture Notes in Computer Science*, volume 10433, pages 344–352. Springer Verlag, 2017.

Pamela J LaMontagne, Tammie L S Benzinger, John C Morris, and Others. OASIS-3: Longitudinal Neuroimaging, Clinical, and Cognitive Dataset for Normal Aging and Alzheimer Disease. *medRxiv*, 2019.

Matthew C. H. Lee, Ozan Oktay, Andreas Schuh, Michiel Schaap, and Ben Glocker. Image-and-Spatial Transformer Networks for Structure-Guided Image Registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 337–345. 2019.

Hongming Li and Yong Fan. Non-rigid image registration using self-supervised fully convolutional networks without training data. In *Proceedings - International Symposium on Biomedical Imaging*, volume 2018-April, pages 1075–1078. IEEE Computer Society, 2018.

Lihao Liu, Xiaowei Hu, Lei Zhu, and Pheng-Ann Heng. Probabilistic Multilayer Regularization Network for Unsupervised 3D Brain Image Registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 346–354. 2019.

Martin Maška, Vladimír Ulman, David Svoboda, Pavel Matula, Petr Matula, Cristina Ederra, Ainhoa Urbiola, Tomás España, Subramanian Venkatesan, Deepak M W Balak, Pavel Karas, Tereza Bolcková, Markéta Štreitová, Craig Carthel, Stefano Coraluppi, Nathalie Harder, Karl Rohr, Klas E G Magnusson, Joakim Jaldén, Helen M Blau, Oleh Dzyubachyk, Pavel Krížek, Guy M Hagen, David Pastor-Escuredo, Daniel Jimenez-Carretero, Maria J Ledesma-Carbayo, Arrate Muñoz-Barrutia, Erik Meijering, Michal Kozubek, and Carlos Ortiz-de Solorzano. A benchmark for comparison of cell tracking algorithms. *Bioinformatics*, 30(11):1609–1617, 2014.

Nicolas Pielawski, Elisabeth Wetzer, Johan Öfverstedt, Jiahao Lu, Carolina Wählby, Joakim Lindblad, and Nataša Sladoje. CoMIR: Contrastive Multimodal Image Representation for Registration. *Advances in neural information processing systems*, 2020.

Matthew Quay, Zeyad Emam, Adam Anderson, and Richard Leapman. Designing deep neural networks to automate segmentation for serial block-face electron microscopy. In *International Symposium on Biomedical Imaging*, volume 2018-April, pages 405–408. IEEE Computer Society, 2018.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, volume 9351, pages 234–241. Springer Verlag, 2015.

Daniel Rueckert, Luke I Sonoda, Carmel Hayes, Derek LG Hill, Martin O Leach, and David J Hawkes. Nonrigid registration using free-form deformations: Application to breast mr images. *IEEE Transactions on Medical Imaging*, 18(8):712–721, 1999.

Martin Simonovsky, Benjamín Gutiérrez-Becker, Diana Mateus, Nassir Navab, and Nikos Komodakis. A deep metric for multimodal registration. In *Lecture Notes in Computer Science*, volume 9902, pages 10–18. Springer Verlag, 2016.

Jean-Philippe Thirion. Image matching as a diffusion process: an analogy with Maxwell's demons. Technical Report 3, 1998.

Vladimír Ulman, Martin Maška, Klas E.G. Magnusson, Olaf Ronneberger, Carsten Haubold, Nathalie Harder, Pavel Matula, Petr Matula, David Svoboda, Miroslav Radojevic, Ihor Smal, Karl Rohr, Joakim Jaldén, Helen M. Blau, Oleh Dzyubachyk, Boudewijn Lelieveldt, Pengdong Xiao, Yuexiang

Li, Siu Yeung Cho, Alexandre C. Dufour, Jean Christophe Olivo-Marin, Constantino C. Reyes-Aldasoro, Jose A. Solis-Lemus, Robert Bensch, Thomas Brox, Johannes Stegmaier, Ralf Mikut, Steffen Wolf, Fred A. Hamprecht, Tiago Esteves, Pedro Quelhas, Ömer Demirel, Lars Malmström, Florian Jug, Pavel Tomancak, Erik Meijering, Arrate Muñoz-Barrutia, Michal Kozubek, and Carlos Ortiz-De-Solorzano. An objective comparison of cell-tracking algorithms. *Nature Methods*, 14 (12):1141–1152, 2017.

Tom Vercauteren, Xavier Pennec, Aymeric Perchant, and Nicholas Ayache. Non-parametric diffeomorphic image registration with the demons algorithm. In *Lecture Notes in Computer Science*, volume 4792, pages 319–326, 2007.

Zhenlin Xu and Marc Niethammer. DeepAtlas: Joint Semi-supervised Learning of Image Registration and Segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 420–429. 2019.

Xiao Yang, Roland Kwitt, Martin Styner, and Marc Niethammer. Quicksilver: Fast predictive image registration - A deep learning approach. *NeuroImage*, 158:378–396, 2017.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. *Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018.

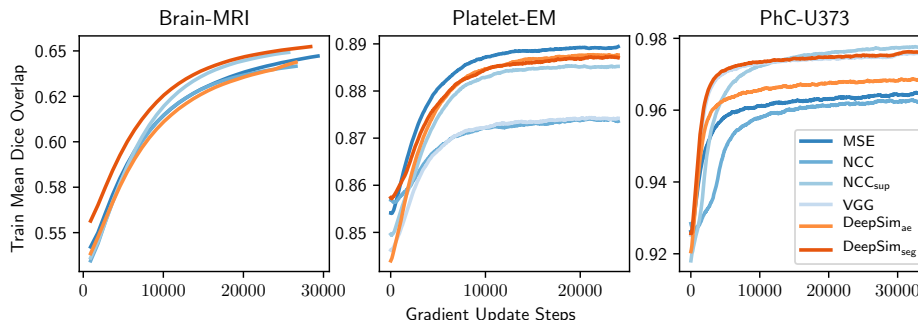## Appendix A. Optimization convergence

Figure 6: Mean dice overlap of the registration models during training. The training duration per model on a single RTX 2080 GPU is approximately seven days for the Brain-MRI dataset and one day for the Platelet, PhC-U373 datasets.

Appendix B. Regularity of the transformation

Method	Dataset					
	Brain-MRI		Platelet-EM		PhC-U373	
	$\sigma^2(J_\Phi)$	$|J_\Phi| < 0\,(\%)$	$\sigma^2(J_\Phi)$	$|J_\Phi| < 0\,(\%)$	$\sigma^2(J_\Phi)$	$|J_\Phi| < 0\,(\%)$
MSE	0.19	0.42	0.23	0.40	0.03	0.02
NCC	0.23	0.93	0.54	4.15	0.29	0.71
NCC$_{\text{sup}}$	0.11	0.28	0.54	4.03	0.28	0.57
VGG	–	–	0.09	0.00	0.16	0.13
DeepSim$_{\text{ae}}$	0.13	0.20	0.10	0.04	0.20	0.35
DeepSim$_{\text{seg}}$	0.09	0.12	0.13	0.14	0.18	0.32

Table 2: Regularity of the transformation. The determinant of the Jacobian of the transformation $|J_\Phi|$ is a measure of how the image volume is compressed or stretched by the transformation. We assess transformation smoothness by the variance of the voxel-wise determinant $\sigma^2(J_\Phi)$, a lower variance indicates a more volume-preserving transformation. We assess the regularity of the transformation by measuring the percentage of voxels for which the determinant is < 0, which indicates domain folding.
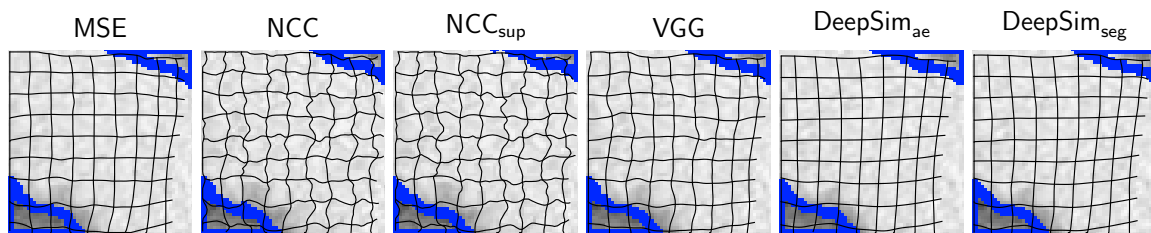
Figure 7: Addendum to Figure 2(b). Detail view of noisy background areas on the Platelet-EM dataset. The cell-boundary is annotated in blue. The transformation is visualized by morphed grid-lines.