Computers in Biology and Medicine 178 (2024) 108761

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Insights into traditional Large Deformation Diffeomorphic Metric Mapping and unsupervised deep-learning for diffeomorphic registration and their evaluation

Monica Hernandez *, Ubaldo Ramon Julvez

Computer Science Department, University of Zaragoza, Spain
Aragon Institute on Engineering Research, Spain

ARTICLE INFO

Keywords:
Diffeomorphic registration
Traditional vs. deep-learning
Evaluation

ABSTRACT

This paper explores the connections between traditional Large Deformation Diffeomorphic Metric Mapping methods and unsupervised deep-learning approaches for non-rigid registration, particularly emphasizing diffeomorphic registration. The study provides useful insights and establishes connections between the methods, thereby facilitating a profound understanding of the methodological landscape. The methods considered in our study are extensively evaluated in T1w MRI images using traditional NIREP and Learn2Reg OASIS evaluation protocols with a focus on fairness, to establish equitable benchmarks and facilitate informed comparisons. Through a comprehensive analysis of the results, we address key questions, including the intricate relationship between accuracy and transformation quality in performance, the disentanglement of the influence of registration ingredients on performance, and the determination of benchmark methods and baselines. We offer valuable insights into the strengths and limitations of both traditional and deep-learning methods, shedding light on their comparative performance and guiding future advancements in the field.

1. Introduction

The non-rigid registration of images is the process of determining the transformation that best warps the source image into the target image according to convenient non-rigid transformation models and image similarity metrics. Non-rigid image registration is a fundamental stage in many different medical applications involving spatial or temporal changes of anatomical or functional features [1–4]. Among the most relevant applications, inter-subject registration is used for the spatial normalization of inputs in template building, atlas-based segmentation, or deep-learning based disease classification [5–9]. Intra-subject registration is used for the fusion of multi-modal information, establishing comparisons from different imaging modalities, the capture of correlations between structure and function, the guidance of computerized interventions, or the analysis of the temporal evolution of diseases [10–12].

1.1. Traditional vs. unsupervised deep-learning methods

The variational formulation of the non-rigid registration problem from the minimization of an energy functional was inspired by Horn and Schunck's approach to solving the optical flow problem [13]. The solutions to both problems have evolved through decades, retaining the energy minimization approach as a backbone [14]. Large Deformation Diffeomorphic Metric Mapping (LDDMM) stands out for being a mathematically well-established approximation to the non-rigid registration problem through diffeomorphisms [15]. Diffeomorphisms enable shape analysis from transformations and thus, they constitute the inception point of Computational Anatomy applications [16,17]. The registration quality, the high accuracy, and the convenience of smooth and invertible transformations for medical applications have made diffeomorphic registration the target to reach by many research on non-rigid registration (e.g., the diffeomorphic versions of Demons [18] are preferred over Demons [19]).

Since the deep-learning explosion taking place in the second decade of the XXI century, deep-learning solutions have been proposed to solve a variety of computer vision and medical imaging problems. FlowNet [20] provided the first deep-learning solution to the optical flow problem and the working ideas were quickly adapted and extended to the problem of non-rigid registration and diffeomorphic registration in medical imaging [21–24]. Supervised deep-learning approaches led to unsupervised approaches that circumvented the costly

* Corresponding author at: Computer Science Department, University of Zaragoza, Spain.
E-mail address: mhg@unizar.es (M. Hernandez).

https://doi.org/10.1016/j.compbiomed.2024.108761
Received 29 November 2023; Received in revised form 4 June 2024; Accepted 13 June 2024
Available online 21 June 2024

need to compute ground truth transformations for training [4]. The interest in unsupervised deep-learning based image registration methods has exponentially increased, and the balance between traditional and deep-learning proposals has broken toward the latter.

Although both traditional and unsupervised deep-learning methods depart from the same minimization problem, the approaches used for finding the solutions are qualitatively different. Traditional methods for non-rigid or diffeomorphic registration seek transformations that minimize the energy functional through traditional optimization methods such as gradient-descent, Levenberg–Marquardt, Gauss–Newton, or quasi-Newton [25,26]. Given the source–target image pair, the optimization process is specific to that image pair. When enough regularization is used, the optimization follows a path of deformations belonging to the transformation model. Deep-learning methods seek a function represented with a deep neural network that allows for the computation of the transformation that minimizes the energy functional for the source–target image pair. The energy functional is turned into a loss function that is used during training to adjust the network parameters through stochastic optimization. In this case, the optimization path is not specific to the image pair but to the data used during training (network initialization, image pairs, and random sequentiation). Indeed, the models estimated during the optimization process are not guaranteed to provide solutions belonging to the transformation model.

With traditional methods, the good performance of the obtained solution for a given variational problem depends on reaching an acceptable local minimum. This mostly depends on the selected optimization strategy. With unsupervised deep-learning methods, the performance of the obtained solution depends on reaching a model with good generalization capabilities. This mostly depends on the amount of training data.

Despite the big difference in the approach, traditional and deep-learning methods still share the definition of the image similarity terms, the parametrization of the space of admissible transformations, and the need for regularization. The high modularity of the traditional non-rigid registration paradigm evidenced with the Insight Toolkit (ITK, www.itk.org) and FAIR [26] libraries makes it possible to adapt or extend the existing methods. For example, a traditional registration method can be adapted to deal with multimodality through a change in the image similarity term and the corresponding derivations of the expressions needed for optimization [26,27]. In addition, a small-deformation method can be turned into a diffeomorphic approach by including the stationary or the non-stationary parametrizations of diffeomorphisms for the representation of the transformation, selecting a convenient regularization, and computing the derivations needed for optimization (see, for example, the evolution of one of the variants in [28] to a diffeomorphic version in [29]).

Deep-learning approaches boost the benefit of this modularity since the derivations of the expressions needed for optimization that may be hard with traditional methods can be easily obtained through automatic differentiation. However, it is often difficult to find the connections, similarities, and differences between traditional and deep-learning methods due to the temporal distance between analogous proposals and the frequent use of different notations.

1.2. Evaluation of non-rigid registration

With the development of non-rigid image registration methods and applications, arose the need to provide evaluation metrics that allow establishing which methodological improvements outperform others for the applications of interest. The difficulty of the problem lies in the lack of ground truth deformations, leading us to rely on indirect methods such as the accuracy obtained when the methods are used in atlas-based segmentation or establishing point or surface correspondences, when these data come from manual expert delineations [30]. In addition, some evaluation proposals include the quantification of desirable properties of the transformations, such as the invertibility, smoothness, inverse consistency, transitivity, or enabling statistics [31,32].

It should be noted that there is often a compromise between these desirable properties and accuracy. Among methods with similar performance in atlas-based segmentation, the methods providing smooth, invertible, or enabling statistics transformations should be preferred. However, it is frequent to find recent evaluation studies where the highest accuracies are obtained at the expense of reducing smoothness or giving up invertibility, and the obtained accuracies are considered the only criterion to prevail over the state of the art [33]. It has been shown that smoothness, invertibility, or statistics enabling are obtained at the expense of reducing accuracy (compare, for example, the performance in [34] with its statistics enabling version [35]). Therefore, it is usual that methods with desirable properties are reported to be of inferior performance, and the unfair use of accuracy as the only criterion to establish superior performance is not discussed in the evaluations presented in the literature. This inertia arises from the lack of homogeneous evaluation protocols with a consistent use of datasets and evaluation metrics which hinders to assess whether a method is superior to another under fair and stable conditions.

There are recent interesting evaluation initiatives such as the Learn2Reg challenge [36], that aim at the homogenization and standardization of the evaluation protocols in non-rigid registration. However,

- Learn2Reg is limited to challenge participants. The segmentations on the test set are not available, and the owners of interesting methods do not participate in the challenge due to different reasons.
- Learn2Reg is more focused on results than on the underlying methodologies. Brief descriptions of the methods are provided, and parameters and models are missing. Reproducibility is hard to achieve for many methods.
- Challenge participants and organizers still give more importance to accuracy rather than desirable properties. The overall rank includes three metrics of accuracy and one metric of smoothness.
- Fast-inference deep-learning approaches take advantage to traditional methods. The overall rank includes one metric on time complexity at inference and no metrics on time complexity at training or memory complexity.

Learn2Reg is definitively a milestone toward improving evaluation protocols of non-rigid registration methods, but there is still a long way to go. Despite the valuable knowledge obtained from Learn2Reg, it is still hard to choose the best methods and baselines to beat when a novel non-rigid registration method is proposed. For example, in the case of OASIS challenge, improvements could go through facilitating the test set information and the process of submission, providing detailed information on the registration ingredients and parameters allowing the reproducibility of the results, providing constructive self-criticism on the fairness of the proposed metrics and ranks used for the comparisons, extending the evaluation from the performance in atlas-based segmentation to other applications and establishing realistic baselines for the different applications.

1.3. Contribution

The contribution of our work is to provide a homogeneous in notation, an extensive description, and a fair and consistent evaluation of traditional LDDMM and unsupervised deep-learning methods with a focus on diffeomorphic registration. On the one hand, we aim at establishing the inter- and intra-theoretical connections between traditional and deep-learning methods for a comprehensive understanding of their methodological insights. On the other hand, we aim at providing a complete, fair, and reproducible evaluation protocol to establish the methods and baselines to beat with future proposals.

In this work, we focus on the problem of T1w MRI non-rigid registration due to our interest in neurodegenerative diseases. We use our own NIREP-based evaluation protocol, which is consistent with our previous contributions to traditional LDDMM methods. In addition, we use Learn2Reg OASIS challenge evaluation protocol in the validation set to complete the results given in the challenge with interesting milestone methods that should not be forgotten. For the sake of reproducibility, we use a parameter configuration that allows a fair comparison among traditional methods. For the deep-learning approaches, we use the publicly available models generated by the proposing authors for the problem of T1w MRI non-rigid registration. Our objective is to increase knowledge in old and new methods for diffeomorphic registration with old and new evaluation protocols.

We provide interesting insights on both the methodological and quantitative and qualitative evaluation aspects of the considered methods with the hope that our work will serve as guidance for improving the understanding and the assessment of future proposals. The registration results are publicly available in https://ieee-dataport.org/documents/traditional-lddmm-vs-deep-learning-deformation-fields-nirep-and-oasis.

Our manuscript proceeds as follows. Section 2 provides an overview of the methodologies under traditional LDDMM methods. Section 3 provides an overview of the methodologies under the deep-learning methods considered in this study. Section 4 provides an overview of the most relevant evaluation protocols for non-rigid registration. Section 5 specifies the methods evaluated in this work. Section 6 describes the datasets used in the evaluation and the relevant implementation details for reproducibility. Section 7 shows the obtained evaluation results. Section 8 discusses the most relevant findings of this study. Finally, Section 9 provides the main conclusions of our work and directions worth to consider in future research.

2. Traditional LDDMM methods for diffeomorphic registration

In this section, we provide an overview of traditional LDDMM methods. We start from the original LDDMM formulation and go through the use of the stationary parametrization in LDDMM, the symmetric formulation of SyN, and PDE-constrained LDDMM approaches. Then, we review the most relevant EPDiff-constrained methods. We finish with a mention to the band-limited parametrization.

2.1. Initial setup

Let I_0 and I_1 be the moving (source) and fixed (target) images representing the input of the image registration problem. In the continuous domain, the images are represented by square-integrable functions $I_i : \Omega \rightarrow \mathbb{R}$, where Ω is a rectangular domain in $\mathbb{R}^d$. For volumetric images, $d = 3$. $Diff(\Omega)$ represents the Riemannian manifold of smooth diffeomorphisms on Ω. V is the tangent space of the Riemannian structure at the identity diffeomorphism, id. V is a space of smooth vector fields on Ω. $Diff(\Omega)$ has a Lie group structure, and V is the corresponding Lie algebra.

2.2. LDDMM

Large Deformation Diffeomorphic Metric Mapping (LDDMM) was proposed by Beg et al. [15] following the inspiration of Horn and Schunck method for the computation of the optical flow [13] and Christensen et al. method for greedy diffeomorphic registration [37]. The LDDMM problem is approached with a variational formulation from the minimization of the energy functional

$$E(v) = E_{\text{reg}}(v) + \lambda E_{\text{img}}(I_0 \circ \varphi^{-1}, I_1), \tag{1}$$

where v is the velocity field flow that parametrizes the problem, $\varphi^{-1} : \Omega \rightarrow \mathbb{R}^d$ is the diffeomorphic transformation that warps the moving I_0 into the fixed I_1 image, the total energy E is decomposed into the regularization E_{reg} and the image similarity metric E_{img}, and λ is a parameter that weights the contribution of E_{reg} and E_{img} to E.

LDDMM assumes that transformations live in an appropriate Riemannian manifold of diffeomorphisms, $Diff(\Omega)$. The Riemannian metric of $Diff(\Omega)$ is defined from the scalar product in V

$$\langle v, w \rangle_V = \langle Lv, Lw \rangle_{L^2} = \langle L^\dagger Lv, w \rangle_{L^2} = \int_\Omega \langle L^\dagger Lv(x), w(x) \rangle d\Omega, \tag{2}$$

where $L = (Id - \alpha\Delta)^s$, $\alpha > 0$, $s \in \mathbb{R}$ is the invertible self-adjoint differential operator associated with the differential structure of $Diff(\Omega)$. The metric is right-invariant with respect to the composition of diffeomorphisms. V is a Reproducing Kernel Hilbert Space (RKHS) of vector fields.

Instead of defining the energy on φ^{-1} directly, the variational problem is parametrized with $v_t \in L^2([0,1], V)$, which is a time-varying velocity field that represents the tangent vectors of the path of diffeomorphisms ϕ_t with beginning in the identity $\phi_0 = id$ and end in $\phi_1 = \varphi^{-1}$ and yield the minimum energy for the LDDMM problem. The transport equation

$$\frac{d\phi_t}{dt} = -v_t \circ \phi_t \tag{3}$$

with initial condition $\phi(0) = id$, corresponds with the Riemannian exponential map between the elements in the manifold of diffeomorphisms $Diff(\Omega)$ and the corresponding elements in the tangent space at the id, V.

In LDDMM, the regularization energy is defined from

$$E_{\text{reg}}(v) = \int_0^1 \|v_t\|_V^2 dt, \tag{4}$$

where $\|\cdot\|_V^2 = \langle \cdot, \cdot \rangle_V$. Thus, the length of the path of diffeomorphisms ϕ_t is given by the regularization energy. Under the exact matching assumption at convergence, $E_{\text{img}}(I_0 \circ \varphi^{-1}, I_1) = 0$ and the solution v_t yields a flow of diffeomorphisms ϕ_t which is a geodesic in $Diff(\Omega)$ with the Riemannian metric. This is the motivation below the word "metric" in LDDMM. In practice, the matching is not exact and the solutions depart slightly from belonging to geodesic paths.

The image similarity energy is defined from

$$E_{\text{img}}(I_0 \circ \varphi^{-1}, I_1) = \|I_0 \circ \varphi^{-1} - I_1\|_{L^2}^2, \tag{5}$$

although the energy minimization approach is amenable to the most commonly used image similarity metrics in medical image registration problems, such as normalized cross-correlation (NCC), its localized version (lNCC), mutual information (MI), and normalized gradient fields (NGF) [26,27,38]. Thus, the energy minimization problem of LDDMM is given by

$$E(v) = \int_0^1 \|v_t\|_V^2 dt + \frac{1}{\sigma^2} \|I_0 \circ \varphi^{-1} - I_1\|_{L^2}^2 \tag{6}$$

where σ is used as the weighting parameter between regularization and image similarity.

Gradient-descent is used in the optimization process. The derivation of the gradient $\nabla_v E(v)$ is obtained from the Hilbert space structure of V and the relationship between Gateaux derivatives and Frechet differentials. Starting from $v_t = 0_V, t \in [0,1]$, the gradient-descent leads the optimization toward a local minimum in the direction of the energy gradient with the gradient-descent update equation

$$v_t^{n+1} = v_t^n - \epsilon \nabla_v E(v_t^n) \tag{7}$$

where

$$\nabla_v E(v_t) = 2v_t - \frac{2}{\sigma^2} K \left(|D\phi_{t,1}| (I_0 \circ \phi_{t,0} - I_1) \nabla (I_0 \circ \phi_{t,0}) \right), \tag{8}$$

using the notation trick $\phi_{s,t} = \phi_t \circ \phi_s^{-1}$ and $K = (L^\dagger L)^{-1}$.

LDDMM was a revolutionary breakthrough in the medical image community due to its accuracy and its ability to provide high-quality smooth and invertible transformations. The solutions are given directly in the tangent space V and there is a direct way of computing the solutions in $Diff(\Omega)$ since the exponential map is simply solving an ODE (Eq. (3)). Solutions to the image registration problem given in V are crucial for their usability as inputs in Computational Anatomy applications such as the computation of population atlases (mean and median) and modes of variation (principal geodesic analysis) [39–42].

The most important limitation of LDDMM lies in the computational load as a result of the time-varying nature of v_t (non-stationary parametrization). Important advances in diffeomorphic registration, such as the stationary or the band-limited parametrizations, and second-order optimization aimed at reducing the computational complexity of the original formulation [32,43,44]. Other advances addressed the problem of improving baseline accuracies, extending E_{img} for multimodality, or imposing physical constraints (like incompressibility) [38,45–47]. All these advances have the original formulation of LDDMM as a backbone.

2.3. StLDDMM

Stationary LDDMM (StLDDMM) was proposed as an efficient alternative to non-stationary LDDMM by replacing the time-varying velocity fields by constant in time ones, known as stationary or steady velocity fields [43,44,48]. The method was originally proposed with gradient-descent [49] and subsequently improved with Gauss–Newton optimization [50]. The original idea of using the stationary parametrization in diffeomorphic registration was addressed independently by three different research groups, and StLDDMM shared the spotlight with Dartel [14] and diffeomorphic Demons [18].

The energy minimization of StLDDMM is given by

$$E(v) = \|v\|_V^2 + \frac{1}{\sigma^2}\|I_0 \circ \varphi^{-1} - I_1\|_{L^2}^2 \tag{9}$$

where the transport equation in the stationary case is simplified to

$$\frac{d\phi_t}{dt} = -v \circ \phi_t. \tag{10}$$

In this case, v does not depend on t. The transport ODE is usually solved by scaling and squaring for transformations, which adapts the scaling and squaring method for computing the group exponential in matrix groups to $Diff(\Omega)$ [48].

Gauss–Newton optimization involves the derivation of the gradient and the Hessian of the energy functional. They are obtained from the first and second-order relationships between Gateaux derivatives and Frechet differentials. The Gauss–Newton update equation is given by

$$v^{n+1} = v^n - \epsilon(H_v E(v^n))^{-1} \nabla_v E(v^n). \tag{11}$$

While for gradient-descent the optimization is sensitive to the initial selection and refinement strategy of parameter ϵ, Gauss–Newton is typically able to converge to acceptable local minima with $\epsilon = 1.0$. In addition, the method shows a super-linear convergence rate, which increases the efficiency of the optimization despite the extra burden in the computation of the Hessian.

It should be noticed that the solutions of the StLDDMM problem belong to one-parametric subgroups instead of geodesics. Therefore, the usability of stationary solutions is more limited than non-stationary ones. However, the computational efficiency made StLDDMM a competitive alternative to non-stationary LDDMM in applications strictly not requiring the use of geodesics.

2.4. SyN/ANTS

The Symmetric Normalization method proposed by Avants et al. [51] (SyN) is the best-performing method in the extensive evaluation framework conducted by Klein et al. [52]. The method was implemented with ITK as ANTS library, and it was the first Open Source version of a diffeomorphic registration method in the LDDMM family. SyN was progressively extended to work with several image similarity metrics and parametrizations. The exceptional accuracy achieved by this method in different applications made SyN a baseline to beat for new proposals of non-rigid registration methods. However, ANTS is implemented in the CPU, which difficult its use in applications requiring massive computations such as atlas building, spatial normalization in large datasets, or extensive evaluations [6,9,36].

The variational formulation of SyN departs from LDDMM in order to improve the symmetry between the forward and backward paths of diffeomorphisms. Thus, recalling with $v^t_{I_0 \to I_1}$ the velocity field flow from I_0 to I_1 and with $\phi^t_{I_0 \to I_1}$ the corresponding path of diffeomorphisms,

$$E(I_0, I_1, v_{I_0 \to I_1}, v_{I_1 \to I_0}) = \int_0^{0.5} \|v^t_{I_0 \to I_1}\|_V^2 + \|v^t_{I_1 \to I_0}\|_V^2 dt + \|I_0 \circ \phi^{0.5}_{I_0 \to I_1} - I_1 \circ \phi^{0.5}_{I_1 \to I_0}\|_{L^2}^2, \tag{12}$$

subject to $\phi^t_{I_0 \to I_1} \circ \phi^{1-t}_{I_1 \to I_0} = id$ and $\phi^{1-t}_{I_1 \to I_0} \circ \phi^t_{I_0 \to I_1} = id$. Therefore, the method imposes the minimization of the image similarity in the middle point of the forward and backward paths while constraining the $\|\cdot\|_V^2$ of the forward and backward velocity fields from the initial to the middle point of the forward and the backward paths, respectively. In addition, extra computations are included to guarantee the inverse consistency between the forward and the backward paths of diffeomorphisms. The LDDMM metric in V is replaced by Gaussian kernels of different sizes, enriching the RKHS structure of V. The optimization proceeds in a multiresolution strategy using gradient-descent. In their original paper, the authors used the gradient from LDDMM for the L^2 image similarity metric and provided the derivation of the gradient for lNCC. Later, they provided a reproducible evaluation setup including MI [53].

It is well-known that in non-symmetric LDDMM approaches the computed solutions favor minimizing the image similarity between the warped source and the target, while the image similarity between the target warped with the inverse diffeomorphism and the source tends to be greater. There are applications such as template building or shape analysis where it is important to work with symmetric and inverse-consistent solutions. In these applications, symmetric and inverse-consistent methods should be a more appropriate choice.

The secret of SyN performance is on the use of different Gaussian kernel regularizers within the multiresolution strategy and the use of lNCC as image similarity metric. The symmetry imposed by SyN is a desirable property, but other non-symmetric methods reached similar accuracies in standard datasets later on. However, the use of an image similarity in the middle of the diffeomorphism path decouples the problem into two simpler subproblems. Therefore, this may give an advantage in the registration of images with substantial differences in anatomy, such as normal-diseased pairs (e.g. Alzheimer's disease).

2.5. PDE-LDDMM

The family of PDE-constrained LDDMM methods proposed in [54] and extended in [46,55–57] is especially interesting. PDE-LDDMM consists in a formulation analytically but not numerically equivalent to Beg et al. LDDMM using an optimal control approach. The method extends the ideas of optical Stokes flow [58] to the diffeomorphic setting. PDE-LDDMM has been used for modeling compressible and incompressible diffeomorphisms, boundary-preserving nonlinear Stokes fluid diffeomorphisms, and mass and intensity preserving diffeomorphisms [46,47,59].

The constraints in PDE-LDDMM are derived from the inverse-consistency identity

$$\phi_t \circ \phi_t^{-1} = id \tag{13}$$

as follows. By differentiation with respect to time, we get

$$\frac{\partial}{\partial t}\phi_t \circ \phi_t^{-1}(x) + D\phi_t \circ \phi_t^{-1}\frac{d}{dt}\phi_t^{-1}(x) = 0 \tag{14}$$

$\forall x \in \Omega$. We now apply the transport equation (Eq. (3)) to get

$$\frac{\partial}{\partial t}\phi_t \circ \phi_t^{-1}(x) + D\phi_t \circ \phi_t^{-1} v_t \circ \phi_t^{-1}(x) = 0. \tag{15}$$

From the change of variables $\phi_t^{-1}(x) = y$ we get the deformation state equation

$$\frac{\partial \phi_t}{\partial t}(y) + D\phi_t v_t(y) = 0, \tag{16}$$

or, equivalently,

$$\partial_t \phi_t + D\phi_t \cdot v_t = 0. \tag{17}$$

The initial condition is $\phi_0 = id$.

2.5.1. PDE-LDDMM based on the image state equation

The original PDE-LDDMM method proposed by Hart et al. in [54] approached the LDDMM problem with a constrained variational formulation. The constraint was based on the restriction of the deformation state equation from maps to images. Thus, the problem is defined from the constrained minimization problem

$$E(v) = \int_0^1 \|v_t\|_V^2 dt + \frac{1}{\sigma^2}\|m(1) - I_1\|_{L^2}^2 \tag{18}$$

subject to

$$\partial_t m_t + \nabla m_t \cdot v_t = 0 \tag{19}$$

with initial condition $m(0) = I_0$ (see the analogy with Eq. (17)). The differentiation of the augmented Lagrangian with respect to the state variable m and its adjoint variable λ yield the optimality conditions

$$\begin{aligned} \partial_t m_t + \nabla m_t \cdot v_t &= 0 \\ -\partial_t \lambda_t - \nabla \cdot (\lambda_t \cdot v_t) &= 0 \\ m(0) &= I_0 \\ \lambda(1) &= \frac{2}{\sigma^2}(I_1 - m(1)). \end{aligned} \tag{20}$$

and the gradient (in L^2) needed for gradient-descent optimization

$$\nabla_{L^2} E(v) = 2L^\dagger L v + \lambda \nabla m. \tag{21}$$

From the optimal control point of view, v is the control, m is the state, and λ is the adjoint variable.

These equations were shown to be analytically equivalent to the ones derived in the original LDDMM [54]. The objective of the PDE-LDDMM approach was to avoid the expensive computations in the map space by the translation of the computations to the image space through the solution of the image state equation. Mang et al. proposed in [46] the extension of the problem with Gauss–Newton–Krylov optimization. In this case, the Hessian-vector product is given from

$$H_{L^2} E(v)\delta v = 2L^\dagger L \delta v + \delta\lambda \nabla m. \tag{22}$$

2.5.2. PDE-LDDMM based on the deformation state equation

Inspired by the previous PDE-LDDMM contributions, Hernandez et al. successfully explored the idea of obtaining more stability and accuracy by relying on the deformation state equation [34,57]. The authors proposed two different methods, one using the expressions of the state and adjoint variables that can be derived from the equivalence between Hart et al. PDE-LDDMM and original LDDMM. The second one directly imposed the deformation state equation as a constraint from inverse-consistency.

Thus, the first method uses the deformation state equation (Eq. (17)) for the computation of the forward and inverse paths, ϕ_t and ψ_t, and then it uses the expressions

$$\begin{aligned} J_t &= |D\psi_t| \\ m(t) &= I_0 \circ \phi_t \\ \lambda(t) &= J_t \lambda(1) \circ \psi_t \\ \delta m(t) &= \nabla I_0 \circ \phi_t \cdot \delta\phi_t \\ \delta\lambda(t) &= \delta J_t \lambda(1) \circ \psi_t + J_t \nabla\lambda(1) \circ \psi_t \cdot \delta\psi_t \end{aligned} \tag{23}$$

in the computation of the gradient $\nabla_{L^2} E(v)$ and the Hessian $H_{L^2} E(v)$ from Eqs. (21) and (22).

The second method solves Eq. (18) subject to

$$\partial_t \phi_t + D\phi_t \cdot v_t = 0 \tag{24}$$

with initial condition $\phi_0 = id$. The differentiation of the augmented Lagrangian with respect to the state variable ϕ and its adjoint variable ρ yield the optimality conditions

$$\begin{aligned} \partial_t \phi_t + D\phi_t \cdot v_t &= 0 \\ -\partial_t \psi_t - D\psi_t \cdot v_t &= 0 \\ -\partial_t \rho_t - \nabla \cdot (\rho_t \cdot v_t) &= 0 \\ \phi(0) &= id \\ \psi(1) &= id \\ \rho(1) &= \lambda(1) \cdot \nabla m(1) \end{aligned} \tag{25}$$

and the gradient needed for gradient-descent optimization is, in this case,

$$\nabla_{L^2} E(v) = 2L^\dagger L v + D\phi \cdot \rho. \tag{26}$$

The Hessian-vector product needed for Gauss–Newton–Krylov optimization is given from

$$H_{L^2} E(v)\delta v = 2L^\dagger L \delta v + D\delta\phi \cdot \rho + D\phi \cdot \delta\rho. \tag{27}$$

Originally, Runge–Kutta was used for the computation of the solutions of the ODEs. However, the equations needed for Semi-Lagrangian integration were derived in [34]. They improved the stability of the solvers while reducing the complexity with the number of time steps. Both methods outperformed the PDE-LDDMM proposal by Mang et al.

2.6. LDDMM and geodesic shooting

Given an initial velocity field $v_0 \in V$ at $t = 0$, the geodesic path under the right-invariant metric on $Diff(\Omega)$ is given by the solution of the Euler-Poincare (EPDiff) equation [60]

$$\partial_t v_t = -K\left((Dv)^T(L^\dagger L v) + D(L^\dagger L v)v + (L^\dagger L v)\nabla \cdot v\right). \tag{28}$$

This expression is obtained from a conservation of momentum law which expresses that the momentum of the diffeomorphic flow at any place along the geodesic can be generated from the momentum at the origin

$$\langle L^\dagger L v_t, w\rangle = \langle L^\dagger L v_0, D\phi_t^{-1} w \circ \phi_t\rangle, \forall w \in V. \tag{29}$$

Once the time-varying vector field flow v_t is computed from the EPDiff equation, the geodesic ϕ_t on $Diff(\Omega)$ can be obtained by the solution of the transport equation. This process is called geodesic shooting and it is one of the fundamentals in the study of Riemannian geometry [61].

Geodesic shooting on $Diff(\Omega)$ can be used to alleviate the computational complexity of LDDMM with a parametrization of the problem on the initial velocity field. Thus,

$$E(v_0) = \|v_0\|_V^2 + \frac{1}{\sigma^2}\|I_0 \circ \varphi^{-1} - I_1\|_{L^2}^2 \tag{30}$$

where φ^{-1} is computed from the solution at time $t = 1$ of the transport equation associated with the vector field v_t solution of Eq. (28) with initial condition v_0.

The dependence of φ^{-1} on v_0 is quite complex, therefore the derivation of the gradient $\nabla_{v_0} E(v_0)$ is cumbersome. Younes et al. approached the problem from the momentum conservation constraint in [62] yielding to quite complicated equations. Vialard et al. [63] proposed a PDE-constrained formulation with solutions on the space of scalar momentum, denoted with α. Thus,

$$E(\alpha_0) = \|\alpha_0 \nabla I_0\|_V^2 + \frac{1}{\sigma^2}\|m(1) - I_1\|_{L^2}^2 \tag{31}$$

where the initial momentum $L^\dagger L v_0 = \alpha_0 \nabla I_0$ and the problem is constrained to the state and EPDiff-equations. Singh et al. [40] proposed the PDE-constrained formulation with solutions on the vector momentum.

Later on, Zhang et al. proposed to compute the easier derivation of the gradient $\nabla_{v_1} E(v_0)$ and transport the vector backward using parallel transport through the adjoint Jacobi equations [32]. The idea was successfully applied in Hernandez et al. PDE-LDDMM methods with Gauss–Newton–Krylov optimization [35,56].

2.7. Mermaid

Mermaid (iMagE Registration via autoMAtIc Differentiation) is a suite for LDDMM diffeomorphic registration available in https://mermaid.readthedocs.io. The suite is focused on different PDE-LDDMM variants with origin in the geodesic shooting methods proposed in [40,63]. Automatic differentiation is used in the computation of the gradient of the different energy functionals and the optimizer is stochastic gradient descent. Although Mermaid's optimization is close to the one used with neural networks and the suite includes some learning-based energies (RDDMM), the optimization approach is strictly not based on data. Therefore, most of the methods in Mermaid can be considered as traditional.

The suite includes the stationary and EPDiff parametrizations. SSD, NCC, and lNCC are the metrics available for image similarity. The regularization energies include the ones used in LDDMM (named Helmholtz) and energies more typically used in non-rigid registration methods and optical flow, such as diffusion, curvature, or total variation. The suite also includes Gaussian and multi-Gaussian smoothing.

The registration models are distributed into:

- Advection for images and maps.
- EPDiff parametrization using vector-valued momentum for images and maps.
- EPDiff parametrization using scalar-valued momentum for images and maps.

According to our experience, not all combinations of the registration ingredients are possible: the input parameters are automatically changed by the code and the numerics are prone to become unstable for some combinations.

2.8. The band-limited parametrization

Last, but no least, a remarkable milestone in LDDMM is the proposal by Zhang and Fletcher [32] of expressing the variational formulation of geodesic shooting LDDMM in the space of band-limited initial vector fields. The band-limited parametrization supposes a significant reduction in memory and computation time. The reduction of the high-frequency components of the velocity fields augments the stability of the ODEs and smoothes the optimization curves as a consequence of the extra smoothness added with band-limiting to the variables. Although this parametrization was proposed for EPDiff methods, it has been successfully extended to other LDDMM approaches such as PDE-LDDMM methods [28].

3. Unsupervised deep-learning methods for diffeomorphic registration

In this section, we provide an overview of unsupervised deep-learning methods. We focus on those methods with own or available source code and models that have demonstrated exceptional accuracy at the time of their respective publications when assessing their performance in terms of the Dice similarity coefficient. We also include the methods considered as benchmarks in TransMorph paper [33].

3.1. VoxelMorph

VoxelMorph was proposed in 2018 [23,24] as an unsupervised deep-learning method for non-rigid image registration. VoxelMorph was among the first methods requiring no supervised information, such as ground truth transformations or anatomical landmark locations. The simplest version of VoxelMorph uses the small deformation parametrization for the solution of the problem ϕ. Thus,

$$\phi = id + u, \tag{32}$$

where $u : \Omega \to \mathbb{R}^d$ is the displacement field of the warp of the moving I_0 into the fixed I_1 image. The transformation ϕ is defined from the minimization of a loss function $\mathcal{L}$

$$\hat{\phi} = \arg\min_{\phi} \mathcal{L}(I_0, I_1, \phi). \tag{33}$$

The loss is borrowed from traditional methods and it is defined from the weighted contribution of image similarity and regularization losses

$$\mathcal{L}(I_0, I_1, \phi) = \mathcal{L}_{\text{sim}}(I_0 \circ \phi, I_1) + \lambda \mathcal{L}_{\text{reg}}(\phi). \tag{34}$$

The recommended image similarity loss is the mean squared error (MSE), based on the sum of squared differences (SSD), or the local normalized cross-correlation (lNCC), widely used in traditional formulations. The regularization loss is defined from

$$\mathcal{L}_{\text{reg}}(\phi) = \|\nabla u\|_{L^2}^2 = \int_\Omega \|\nabla u(x)\|_2^2 d\Omega \tag{35}$$

and penalizes the displacement field associated with the solution from showing large partial derivatives of first-order. This regularization loss was introduced in Horn and Schunck method for the estimation of the optical flow [13].

The authors proposed and compared two different architectures inspired by U-Net [64]. The U-Nets were combined with the spatial transformer [65] for the computation of the composition $I_0 \circ \phi$. They found that there was a compromise between the number of parameters and the accuracies obtained on the test datasets. VoxelMorph-II obtained the best registration results at the cost of using extra computational resources.

The codes for VoxelMorph were released in the GitHub repository https://github.com/voxelmorph/voxelmorph. The models in [23] were available in the repo. VoxelMorph library includes extensions to the use of the stationary parametrization and different image similarity metrics used in traditional methods. The library included useful tutorials on how to train and test VoxelMorph models on 2D and 3D data. These are probably the reasons why VoxelMorph has been adopted as the baseline to beat with subsequent proposals. The original VoxelMorph codes and models became legacy code in less than five years due to the fast evolution of the third-party dependencies. The authors are constantly updating and maintaining the library and more powerful models are released from time to time.

3.2. VoxelMorphDiff

The diffeomorphic version of VoxelMorph (VoxelMorphDiff) was proposed in 2018 [66,67]. Inspired by the supervised proposal by Krebs et al. [68], VoxelMorphDiff was approached using variational inference. Thus, the method is not only able to provide the diffeomorphic transformation that best warps the moving into the fixed image according to the loss function but also to provide uncertainty estimates.

Let z be a latent variable used for the parametrization of the transformation ϕ. Thus, we represent the solution of the problem by $\phi_z : \Omega \rightarrow \mathbb{R}^d$. The prior probability of z is modeled from a multivariate normal distribution of zero mean and covariance matrix Σ_z

$$p(z) = N(z; 0, \Sigma_z). \tag{36}$$

The authors assume z to be a stationary velocity field, and the solution to the problem is computed through the transport equation (Eq. (3)) by the scaling and squaring algorithm.

The smoothness of z is encouraged by the definition of $\Sigma_z^{-1} = \lambda L$, where λ controls the scaling of the velocity field z and L is the Laplacian of a neighborhood graph defined on the grid in which the image domain Ω is discretized.

Let I_1 be a noisy observation of the warped image $I_0 \circ \phi$. Then,

$$p(I_1|z; I_0) = \mathcal{N}(I_1; I_0 \circ \phi_z, \sigma^2 I), \tag{37}$$

where σ^2 represents the variance of the additive image noise.

The solution of the registration problem is obtained from the estimation of the posterior registration probability $p(z|I_1; I_0)$ and ϕ_z as the most likely transformation for the pair of images using maximum a posteriori (MAP) estimation. The approach comes with an estimate of the uncertainty of the registration.

Since the problem for $p(z|I_1; I_0)$ is intractable, the authors use a variational approach for the approximated posterior registration probability. Variational inference uses a convolutional neural network combined with the scaling and squaring method for transport equation integration and spatial transformer layers for the computation of warped images. The probabilistic formulation of the registration problem was first theoretically formulated in [14].

The posterior registration probability $p(z|I_1; I_0)$ is approximated with the probability $q_\psi(z|I_1; I_0)$ parametrized by ψ. The parameter ψ is obtained from the minimization of the Kullback—Leibler (KL) divergence between q_ψ and p. Thus, we seek for

$$\begin{aligned} &\min_\psi KL[q_\psi(z|I_1; I_0) \parallel p(z|I_1; I_0)] = \\ &\quad \min_\psi KL[q_\psi(z|I_1; I_0) \parallel p(z)] - E_q[\log p(I_0|z; I_1)]. \end{aligned} \tag{38}$$

At this point, the approximate posterior $q_\psi(z|I_1; I_0)$ is modeled with a multivariate normal with diagonal covariance matrix

$$q_\psi(z|I_1; I_0) = \mathcal{N}(z; \mu_{z|I_1,I_0}, \Sigma_{z|I_1,I_0}), \tag{39}$$

and the parameters $\mu_{z|I_1,I_0}$ and $\Sigma_{z|I_1,I_0}$ are estimated from a CNN parametrized by ψ with the minimizing function in Eq. (38) as the loss function. Thus,

$$\begin{aligned} &\mathcal{L}(\psi; I_1, I_0) = \\ &\quad KL[q_\psi(z|I_1; I_0) \parallel p(z)] - E_q[\log p(I_1|z; I_0)] = \\ &\quad KL[q_\psi(z|I_1; I_0) \parallel p(z)] - \frac{1}{2\sigma^2 K} \sum_k \| I_0 \circ \phi_{z_k} - I_1 \|^2, \end{aligned} \tag{40}$$

where K is the number of samples used. The first term in the final expression of the loss encourages the posterior q_ψ to be close to the prior $p(z)$. The second term favors those solutions where the warped image is similar to I_1, resembling VoxelMorph image similarity loss.

3.3. SynthMorph

SynthMorph departs from VoxelMorph with a training strategy which replaces image pairs with synthetic or true segmentation pairs [69]. The method is proposed from the observation that the use of image similarity metrics strictly from the images as loss functions results in models that are only able to predict acceptable solutions for image pairs and image differences similar to the data observed during training. SynthMorph exposes the models during training to a wide range of variability in segmentation pairs and differences that are expected to improve their generalization capability. In addition, the models are supposed to be agnostic to image modalities since images are replaced with segmentations.

The loss function is defined in terms of $\text{seg}(I_0)$ and $\text{seg}(I_1)$, the segmentations associated with the moving and the fixed images

$$\begin{aligned} &\mathcal{L}(\text{seg}(I_0), \text{seg}(I_1), \phi) = \\ &\quad \mathcal{L}_{\text{sim}}(\text{seg}(I_0) \circ \phi, \text{seg}(I_1)) + \lambda \quad \mathcal{L}_{\text{reg}}(\phi), \end{aligned} \tag{41}$$

where the image similarity metric is given by the Dice metric

$$\begin{aligned} &\mathcal{L}_{\text{sim}}(\text{seg}(I_0), \text{seg}(I_1), \phi) = \\ &\quad \frac{2}{L} \sum_{l=1}^{L} \frac{(seg(I_0) \circ \phi)_l \odot seg(I_1)_l}{(seg(I_0) \circ \phi)_l \oplus seg(I_1)_l}, \end{aligned} \tag{42}$$

where $(\cdot)_l$ represents the one-hot encoded segmentation of label l, L is the number of labels, and $\odot$ and $\oplus$ represent the point-wise multiplication and addition, respectively.

SynthMorph uses the stationary parametrization of diffeomorphisms. The regularization is defined from

$$\mathcal{L}_{\text{reg}}(\phi) = \frac{1}{2} \| \nabla u \|^2_{L^2}, \tag{43}$$

where u is the displacement field of ϕ, recall $\phi = id + u$.

The authors synthesized two different sets of training data, yielding two different models. The first set was obtained from random geometric shapes synthesized from noise distributions. The second set was obtained from brain image segmentations obtained from SynthSeg [70].

3.4. SyMNet

SyMNet was proposed in [71] with the idea of estimating both forward $\phi_t^{I_0 \rightarrow I_1}$ and backward $\phi_t^{I_1 \rightarrow I_0}$ diffeomorphic transformations for an image registration pair while enforcing the symmetry of the process. The method is heavily inspired by SyN and the suite of algorithms implemented in ANTS library.

SyMNet uses both the small displacement and the stationary parametrization of diffeomorphisms, although the former is preferred by the authors due to the improved consistency in the estimation of the inverse transformation [14]. The network is based on a U-Net architecture.

The loss function combines the image similarity loss that is decomposed into three image-based losses which measure the difference between the warped images at the beginning and end points of the forward and backward paths and in the middle and three regularization terms that control different aspects of the smoothness of the transformations. Thus,

$$\mathcal{L}(I_0, I_1) = \mathcal{L}_{\text{sim}} + \lambda_{\text{reg}} \mathcal{L}_{\text{reg}} + \lambda_{\text{mag}} \mathcal{L}_{\text{mag}} + \lambda_{\text{Jdet}} \mathcal{L}_{\text{Jdet}} \tag{44}$$

where

$$\begin{aligned} \mathcal{L}_{\text{sim}} = &-lNCC(I_0 \circ \phi_{0.5}^{I_0 \rightarrow I_1}, I_1 \circ \phi_{0.5}^{I_1 \rightarrow I_0}) \\ &-lNCC(I_0 \circ \phi_1^{I_0 \rightarrow I_1}, I_1) - lNCC(I_1 \circ \phi_1^{I_1 \rightarrow I_0}, I_0). \end{aligned} \tag{45}$$

The regularization loss combines the contribution of the regularization term for the forward and backward velocities with a constraint on the differences of the magnitude of the velocities to be similar, and the Jacobian determinant that is penalized against non-diffeomorphic solutions. Thus,

$$\mathcal{L}_{\text{reg}} = \|\nabla v^{I_0 \to I_1}\|^2_{L^2} + \|\nabla v^{I_1 \to I_0}\|^2_{L^2} \tag{46}$$

$$\mathcal{L}_{\text{mag}} = \|v^{I_0 \to I_1}\|^2_{L^2} - \|v^{I_1 \to I_0}\|^2_{L^2} \tag{47}$$

$$\mathcal{L}_{\text{Jdet}} = \int_{\Omega} \max(0, -|J_{\phi}(x)|) d\Omega. \tag{48}$$

The method differs from SyN in the image similarity loss, enriching the SyN image similarity with the errors at the beginning and end points of the paths of transformations. In addition, the regularization is substantially different from SyN, based on restricting the magnitude of the gradient of the velocity fields and complemented with a control on the magnitude of the velocity fields and a penalization on negative Jacobians.

3.5. LapIRN

LapIRN proposed the use of a Laplacian Pyramid Network architecture for diffeomorphic registration [72]. The idea of using a pyramidal image structure comes from coarse-to-fine or multi-resolution strategies that are commonly used in computer vision tasks such as scale-space, segmentation, object detection, optical flow estimation, and non-rigid registration. The use of a multi-resolution strategy enables the methods to capture large and small misalignments between the images. The computations at the coarsest resolution levels provide initial rough estimations of the transformations helping to avoid local minima far from the optimal solutions.

Multiresolution strategies are hard in neural network-based approaches, and cascade CNNs, where different networks for the different resolutions are combined in an end-to-end manner, seem the most reasonable approaches [73]. Thus, LapIRN follows a three-level cascade Laplacian pyramid framework, with an identical CNN-based registration network for each pyramid level. Each network uses as input the warped images upsampled from the immediately superior coarser scale and the transformations computed in that scale. LapIRN is trained in a coarse to fine manner, where the coarsest level network is trained first, and the finer-level networks are progressively added to the training of the whole cascade system. With respect to the alternative of separated network training, the training strategy followed in LapIRN helps achieve a balance of the loss weights overall the different resolution levels. The optimization loss for the method is defined following a similarity sub-pyramid for each level

$$\mathcal{L}^k_{\text{sim}}(I_0, I_1) = \sum_{i \le k} -\frac{1}{2^{k-i}} lNCC(I^i_0 \circ \phi^i, I^i_1), \tag{49}$$

where k denotes the pyramid level, (I^i_0, I^i_1) denote the image registration pair subsampled at level i, and ϕ^i denotes the transformation at level i. Typically, $k \in \{1, 2, \dots, L\}$ where $L = 3$.

LapIRN is implemented with small displacements and the stationary parametrization of diffeomorphisms. The regularization is respectively performed on the displacements or velocity fields obtained on the corresponding resolution level

$$\mathcal{L}^k_{\text{reg}} = \frac{1}{2^{L-k}} \|\nabla v\|^2_2. \tag{50}$$

According to the Learn2Reg challenge rank, LapIRN is one of the best-performing methods scoring in the top-three of the different registration tasks.

3.6. CycleMorph

CycleMorph stands for cycle-consistent unsupervised non-rigid image registration [74]. The method departs from the limitations observed in deep-learning approaches to preserve the topology of the objects in the image due to the lack of smoothness of the transformations, and the typically low accuracy of the models when inference is performed between the same images (surprisingly, the estimated transformation greatly differs from the identity in this trivial case). The authors propose a cycle consistency loss as a constraint that helps in the purpose of preserving the topology and an identity loss that helps with the same image registration problem. Thus, the loss function is defined as follows

$$\begin{aligned}&\mathcal{L}(I_0, I_1, \phi^{I_0 \to I_1}, \phi^{I_1 \to I_0}) = \\ &\quad \mathcal{L}_{\text{reg+sim}}(I_0, I_1, \phi^{I_0 \to I_1}) + \mathcal{L}_{\text{reg+sim}}(I_1, I_0, \phi^{I_1 \to I_0}) + \\ &\quad \mathcal{L}_{\text{cycle}}(I_0, I_1, \phi^{I_0 \to I_1}, \phi^{I_1 \to I_0}) + \\ &\quad \mathcal{L}_{\text{id}}(I_0, I_1, \phi^{I_0 \to I_1}, \phi^{I_1 \to I_0}).\end{aligned} \tag{51}$$

The loss $\mathcal{L}_{\text{reg+sim}}$ is selected as a typical total loss function with lNCC as image similarity and $\|\nabla \phi\|^2_{L^2}$ as regularization. The cycle consistency loss is defined from

$$\begin{aligned}&\mathcal{L}_{\text{cycle}}(I_0, I_1, \phi^{I_0 \to I_1}, \phi^{I_1 \to I_0}) = \\ &\quad \|I_1 \circ \phi^{I_1 \to I_0} - I_0\|_{L^1} + \|I_0 \circ \phi^{I_0 \to I_1} - I_1\|_{L^1}\end{aligned} \tag{52}$$

and robustly measures the L^1 differences between the warped images at the end-points of the path after the registration with the corresponding transformation. Finally, the identity loss is an additional regularization that guides the network to estimate the identity transform for identical pairs of images

$$\begin{aligned}&\mathcal{L}_{id}(I_0, I_1, \phi^{I_0 \to I_1}, \phi^{I_1 \to I_0}) = \\ &\quad -lNCC_{I_0 = I_1}(I_0 \circ \phi^{I_0 \to I_1} - I_1) - lNCC_{I_1 = I_0}(I_1 \circ \phi^{I_1 \to I_0} - I_0).\end{aligned} \tag{53}$$

Small deformation and a U-Net architecture complete the description of the proposed method.

CycleMorph and SymNet approach the problem of enhancing inverse consistency in the registration. Both methods overlap in the use of image similarity losses and regularization terms at the beginning and end points of the forward and backward paths. Then, they differ in the use of information in the middle of the path and the inclusion of different regularization terms. CycleMorph does not consider the option of using the stationary parametrization of diffeomorphisms although the method may be very easily extended. The model used in this work was generated and released by TransMorph's authors.

3.7. ViT-V-Net

ViT-V-Net builds on Vision Transformer architectures for image classification [75], which use a purely self-attention-based model that learns long-range spatial relations to focus on the relevant parts of the image for the task [76]. The ViT network cannot be used directly in the problem of image registration since the consecutive downsamplings have the effect of emphasizing low-resolution features without localization information. ViT-V-Net combines ViT and convolutional networks, providing an architecture suitable for image registration. The emphasis of the contribution is on the network, which departs from traditionally used U-Net. Small deformation, SSD, and diffusion regularizer are selected for the definition of the loss function.

3.8. TransMorph

So far, VoxelMorph library has served as base code for many of the subsequent proposals. TransMorph gives name to a suite of models closely related to VoxelMorph models [33]. The codes for TransMorph are available in the GitHub repository https://github.com/junyuchen24

5/TransMorph_Transformer_for_Medical_Image_Registration, where the parallelism with VoxelMorph can be appreciated.

TransMorph replaces the classical U-Net architecture of VoxelMorph with convolutional neural networks (CNNs) by a swin transformer, one of the adaptations of the original transformer architecture for natural language processing to computer vision applications. The suite includes the small deformation, stationary, and b-spline parametrizations; SSD and lNCC image similarity metrics; and the regularization energy used in VoxelMorph is complemented by the possibility of using the bending energy

$$\mathcal{L}_{bend} = \|\nabla^2 u\|_2^2, \tag{54}$$

closely related with LDDMM operator L.

In addition, the probabilistic version for a diffeomorphic TransMorph and a Bayesian version are available. The methods include the possibility of using the segmentations of the images during training and incorporate a loss based on the Dice similarity coefficient.

The proposal of TransMorph architecture came out after a thorough Neural Architecture Search (NAS) for the problem of image registration. The authors considered three models that use the hybrid Transformer-ConvNet architecture: ViTVNet, proposed for non-rigid image registration; Pyramid Vision Transformer (PVT), proposed for image classification, object detection, and semantic segmentation [77]; and Convolutional Neural Network and Transformer (CoTr), for 3D medical image segmentation [78]. In addition, they considered the pure transformer architecture of nnFormer, proposed for medical image segmentation [79]. The authors came out with a hybrid Transformer-ConvNet architecture with swin transformers to provide the best registration results according to the considered evaluation metrics.

TransMorph was trained with transformer architectures of different sizes ranging from tiny, to small, regular, and large versions. In addition, TransMorph comes with the code for conducting an interesting comparative with traditional methods such as SyN, NiftyReg, LDDMM, and deedsBVC. In addition, the codes include the possibility to compare the b-spline variant of TransMorph with MIDIR, a stationary variant of VoxelMorph where the stationary velocity fields are represented with b-splines [80].

3.9. NODEO

NODEO was proposed in [81] as a fresh learning-based approach to the problem using Neural Ordinary Differential Equations (NODEs). NODEs were first proposed in [82] as a learning-based approach to ODE solvers. The method is inspired by the analogies between the Euler method and ResNet [83] and replaces the residual network itself with a function leading the depth of the neural network from discrete to infinite dimension thus leveraging the accuracy of the solvers. Given an ODE in the shape of

$$\frac{dy}{dt} = f(y(t), t), \tag{55}$$

with initial condition $y(t_0) = y_0$, neural ODEs aim at learning the function f parametrized by θ in the shape of a neural network. Thus, the objective is to learn f_θ from

$$\frac{dz}{dt} = f_\theta(z(t), t) \tag{56}$$

using

$$\mathcal{L}(z(t_1)) = \mathcal{L}\left(z_0 + \int_{t_0}^{t_1} f_\theta(z(t), t)dt\right) \tag{57}$$

as loss function.

NODEO approaches the ODE solver of the original LDDMM problem with NODEs. Thus, the solution of the transport equation is estimated by NODES

$$\frac{d\phi_t}{dt} = -v_t^\theta(\phi_t), \tag{58}$$

where

$$\phi(s) = \phi(0) + \int_0^s -v_t^\theta(\phi_t)dt. \tag{59}$$

Notice the analogy between Eqs. (3) and (58).

The loss function is

$$\mathcal{L}(I_0, I_1, v_t^\theta) = \mathcal{L}_{\text{reg}}(v_t^\theta) + \mathcal{L}_{\text{sim}}\left(I_0, I_1, \phi(0) + \int_0^1 -v_t^\theta(\phi_t)dt\right). \tag{60}$$

The image similarity is lNCC while the regularization is borrowed from previous proposals

$$\mathcal{L}_{\text{reg}} = \lambda_{\text{reg}}\mathcal{L}_{\text{reg}} + \lambda_{\text{mag}}\mathcal{L}_{\text{mag}} + \lambda_{\text{Jdet}}\mathcal{L}_{\text{Jdet}}, \tag{61}$$

where

$$\mathcal{L}_{\text{reg}} = \|v_t^\theta\|_{L^2}^2 \tag{62}$$

$$\mathcal{L}_{\text{mag}} = \|\nabla u\|_{L^2}^2 \tag{63}$$

$$\mathcal{L}_{\text{Jdet}} = \int_\Omega \max(0, -|J_\phi(x)| + \epsilon)d\Omega. \tag{64}$$

The main difference between NODEO and other deep-learning approaches is that the training is conducted for each registration pair. Starting from a random initialization of v_t^θ, forward and backward propagation iteratively improve the estimation of the network parameters θ according to the minimization of the loss function given in Eq. (60).

Analyzing the codes available in the GIT repository https://github.com/yifannnwu/NODEO-DIR, two remarkable implementation details come to light. First of all, the authors used a stationary parametrization for the NODES. This means that v_t^θ does not depend on time and the authors are solving the problem for the stationary parametrization. For a number of time steps equal to two, the stationary parametrization is analogous to StLDDMM. For a number of time steps greater than two, the parametrization is analogous to Dartel. The computation of the inverse of the exponential mapping

$$\log : Diff(\Omega) \to V, \tag{65}$$

$\phi_t \to v_t$ is needed for the computation of the regularization loss. This is circumvented with the rough approximation

$$v_t \approx \phi_{t+1} - \phi_t \tag{66}$$

that introduces an intrinsic error in the solutions. Second, the authors claim the need to apply a Gaussian kernel in the last layer of the CNN architecture. This may be due to the method with the CNN layers is not able to learn sufficiently smooth models v_t^θ from the regularization losses. A more elegant explanation is that auto gradient is computing the gradient of the loss function in the space of L^2 functions, however, the gradient needs to be computed in V. This is a correct transformation of the derivatives considering the Gaussian RKHS structure of V.

4. Insights into the evaluation of traditional and deep-learning methods

Evaluation of non-rigid image registration nearly initiated with evaluation projects such as the Non-Rigid Image Registration Evaluation Project (NIREP) [31]. NIREP provided an evaluation framework to objectively compare the performance of non-rigid registration with a proposal of standard criteria. The authors provided a dataset of 16 brain MRI images together with manual segmentations over 32 gray matter regions. The framework was built on the idea that no metric alone is sufficient for performance evaluation. The authors proposed to use a set of diverse metrics for a good indication of the registration quality. These metrics included the overlap of the segmentations, the intensity variance that measures the sharpness of the atlas built from the registrations to a common template, the inverse consistency between the

forward and backward transformations, and the capacity of pairwise registrations to satisfy the transitivity property. From these metrics, the overlap of the segmentations prevailed through time.

The wide collaborative effort of Klein et al. [52] was the first extensive evaluation project given the brain MRI datasets (LPBA40, IBSR18, CUMC12, and MGH10), number of methods (14), and evaluation metrics (8). The evaluation metrics were in agreement with VALMET [84], an abandoned evaluation project preceding NIREP and all of them quantified the performance of non-rigid registration from the amount of overlap between segmentations and surfaces. Despite some of the authors of NIREP participated in this project, NIREP dataset was not included in the study, and, surprisingly, all the participants agreed on the inferior quality of IBSR18, CUMC12, and MGH10 images and segmentations. Indeed, the authors pointed out as caveats that the overlap metrics were insensitive to the presence of foldings in the transformations, and that the study did not inform on the intrinsic properties of the spatial deformations such as invertibility, inverse consistency, or transitivity. In addition, the study ranked the different methods as a whole, without any insight into the influence of the registration ingredients (transformation parametrization, regularization, image similarity, and optimization) in the outcome.

Despite the mentioned caveats, the overlap metrics suggested by Klein et al. (Target Overlap, Dice Similarity Coefficient, Jaccard index) and their datasets prevailed in subsequent evaluation studies leaving aside the evaluation of desirable properties related to the quality of the transformations. They widely constitute almost the only criteria to establish novel non-rigid registration methods in the state of the art. Fortunately, SyN resulted in one of the best DSC-performing methods and established a hard-to-beat and reproducible baseline with diffeomorphic solutions [53].

The following milestones in evaluation protocols for non-rigid image registration were organized around challenges: EMPIRE10 (lung CT; 2010) [85], CRC (lung CT and brain MRI; 2018) [86], CuRIOUS (intra-operative brain US and MRI; 2019) [87], ANHIR (histology; 2019) [88], and Learn2Reg (lung CT, brain MRI, thorax CT-MR, hidden; 2020, 2021, 2022) [36]. From them, EMPIRE10 and CRC recovered measurements of registration quality such as the ratio of singularities or inverse consistency.

Learn2Reg is probably the most comprehensive, extensive, and ambitious evaluation project to date. The challenge was proposed in 2020. This was a very opportune moment in which the community had witnessed a shift from traditional to supervised deep-learning methods with super efficient inference, and non-supervised proposals had just overcome the overhead of computing sample transformations for training. The challenge intended to serve as a unique benchmark to fairly evaluate the state of the art and upcoming proposals in different datasets and tasks. Learn2Reg proposed the use of the Dice Similarity Coefficient (DSC) and the 30th percentile as overlap metrics together with the 95th percentile Hausdorff distance (HD95), the standard deviation of the logarithm of shifted and clipped displacement Jacobians (SDlogJ)[1], a measurement of the quality of the transformations different from previous proposals (NIREP, EMPIRE10, CRC). In addition, the runtime at inference was included in the evaluation. The task regarding brain MRI registration is translated from Klein et al. datasets to OASIS (www.oasis-brains.org). FreeSurfer segmentations are used as a proxy for manual segmentations, yielding an evaluation-based on a bronze instead of on a gold standard.

As in Klein et al. study, the participants of Learn2Reg challenge evaluated the registration methods as a whole, therefore, we still do not have insights on the influence of the registration ingredients in the overall performance. In addition, the methods were limited to those research groups with the resources needed to participate in the challenge (human resources availability, time, software, and hardware), leaving aside milestones from the state of the art. The results were limited to the proposed evaluation metrics and methodological insights leading to the best performance were not provided.

In the problem of evaluation of non-rigid registration with OASIS, LapIRN, Convex Adam, and VoxelMorph ranked in the top three with average DSC values of 82, 81, and 80% in the test set. Surprisingly, Convex Adam combines a deep-learning based module for establishing large correspondences with traditional optimization (hybrid method) [89]. Even more, this hybrid approach ranked in the top positions of the different Learn2Reg applications recovering the interest in traditional approaches with the potential to collaborate with deep-learning solutions.

5. Methods evaluated in this work

Given the large families of traditional and deep-learning based non-rigid image registration methods, we need to focus our study. Thus, in this work, we selected traditional LDDMM and unsupervised deep-learning methods with available source code and models trained in the T1w MRI registration problem, preferably with diffeomorphic variants. Even though we conducted the evaluation in more than fifty variants of methods.

Since the number of different LDDMM variants is considerable and given the prevalence of the stationary parametrization of diffeomorphisms in unsupervised deep-learning approaches, we focused on the stationary versions of LDDMM leaving out of the scope of this work the non-stationary, EPDiff, or band-limited parametrizations. The methods with these interesting parametrizations will be studied in a subsequent work. Although traditional LDDMM methods with the stationary parametrization can be easily transformed into inverse consistent and symmetric methods, we decided to evaluate the methods as they were originally proposed and leave for future work the analysis of the advantages of these variants.

Table 1 gathers the traditional LDDMM and unsupervised deep-learning methods evaluated in our work. We provide a brief description of the method that allows the classification or identification of the model architecture, the parametrizations, the regularization terms, the image similarity metrics, and the optimization methods. For traditional methods, we selected the optimization that is feasible (e.g., MI and gradient descent in StLDDMM) or known to provide the best accuracy in previous evaluation studies (e.g., NGF and gradient descent in PDE-LDDMM).

Regarding the deep-learning methods, we decided to use the publicly available models generated by the original authors for the problem of T1w MRI registration. Our decision may carry problems of intra- and inter-fairness:

- The publicly available models may be suboptimal, better models may be achieved by extending the training phase.
- The publicly available models were not trained under the same conditions (image pairs and number of iterations) and a good method may underperform just because of an inferior training setup.

Nevertheless, some of these methods showed a competitive evaluation performance according to our criteria. In addition, our decision also carries interesting advantages:

- It is worth knowing the evaluation outcome of publicly available models and methods.
- The evaluation conducted with publicly available models is easily reproducible.

[1] The Python code for the computation of the SDlogJ is `log_jac_det = np.log(jacobian_determinant(disp_field[np.newaxis, :, :, :, :]) + 3) .clip(1.0e-9, 1.0e10)`.

Table 1
Summary of the registration ingredients of the methods evaluated in this work. We refer with v to the velocity field and u to the displacement field, $\phi = id + u$. The abbreviations used in the table are eq for equation, disp for displacement, and rep for representation.

Method	Model	Parametrization	Regularization	Image similarity	Optimization
SyN	LDDMM	Greedy SyN	Multi-kernel Gaussian	SSD, lNCC, MI (middle)	Gradient descent
StLDDMM	LDDMM	Stationary	$\|Lv\|_2^2$	SSD, NCC, lNCC, NGF	Gauss–Newton
StLDDMM	LDDMM	Stationary	$\|Lv\|_2^2$	MI	Gradient descent
PDE-LDDMM	State eq. constraint	Stationary, Galerkin rep.	$\|Lv\|_2^2$	SSD, NCC, lNCC	Gauss–Newton–Krylov
PDE-LDDMM	State eq. constraint	Stationary, Galerkin rep.	$\|Lv\|_2^2$	MI, NGF	Gradient descent
PDE-LDDMM	Map eq. constraint	Stationary, Galerkin rep.	$\|Lv\|_2^2$	SSD, NCC, lNCC	Gauss–Newton–Krylov
PDE-LDDMM	Map eq. constraint	Stationary, Galerkin rep.	$\|Lv\|_2^2$	MI, NGF	Gradient descent
Mermaid	Map	Stationary	$\|Lv\|_2^2$	SSD, NCC	Stochastic gradient descent
Mermaid	Scalar momentum map	Stationary	Multi-kernel Gaussian	SSD, NCC	Stochastic gradient descent
Mermaid	Vector momentum map	Stationary	Multi-kernel Gaussian	SSD, NCC	Stochastic gradient descent
NODEO	LDDMM through neural ODEs	Stationary	$\|\nabla u\|_2^2$, $J_{det} > 0$, $\|v\|_2^2$	SSD, lNCC	Adam
VoxelMorph	U-Net	Small disp./stationary	$\|\nabla u\|_2^2$	SSD, lNCC	Adam
VoxelMorph-Diff	Probabilistic, U-Net	Stationary	KL	SSD	Adam
SynthMorph	U-Net	Stationary	$\|\nabla u\|_2^2$	DSC	Adam
SymNet	U-Net	Small disp./stationary	$\|\nabla u\|_2^2$, $J_{det} > 0$, symm $\|v\|_2^2$	lNCC (beginning, middle, end)	Adam
LapIRN	Pyramid CNN	Small disp./stationary	$\|\nabla u\|_2^2/\|\nabla v\|_2^2$, $J_{det} > 0$	lNCC	Adam
TransMorph	Swin transformer	Small disp.	$\|\nabla u\|_2^2$	lNCC, DSC	Adam
TransMorph-Diff	Probabilistic Swim transformer	Stationary	KL	SSD	Adam
TransMorph-Bspl	Swin transformer	Stationary, b-spline rep.	$\|\nabla u\|_2^2$	lNCC	Adam
TransMorph-Bayes	Bayesian Swim transformer	Small disp.	$\|\nabla u\|_2^2$	lNCC	Adam
CycleMorph	U-Net	Small disp.	$\|\nabla u\|_2^2$	lNCC, $\mathcal{L}_{\text{cycle}}$, $\mathcal{L}_{\text{id}}$ (beginning, end)	Adam
MIDIR	U-Net	Stationary, b-spline rep.	$\|\nabla u\|_2^2$	MI	Adam
VIT	Vision transformer	Small disp.	$\|\nabla u\|_2^2$	SSD	Adam
PVT	Pyramid vision transformer	Small disp.	$\|\nabla u\|_2^2$	–	Adam
CoTr	CNN-Transformer	Small disp.	$\|\nabla u\|_2^2$	DSC	Adam
nnFormer	Swin transformer	Small disp.	$\|\nabla u\|_2^2$	–	Adam

Last, but not least, we considered including ConvexAdam from Learn2Reg22 in our study since the source code is available in https://github.com/multimodallearning/convexAdam. However, we found that the method apparently uses data extracted from the segmentation of the images and we did not find any information on how to obtain this data correctly.

6. Datasets and implementation details

6.1. NIREP

NIREP or NIREP16 was proposed in [31] for the evaluation of non-rigid registration. NIREP16 consists of 16 T1w Magnetic Resonance Imaging (MRI) images. These images were acquired at the Human Neuroanatomy and Neuroimaging Laboratory, University of Iowa. They were selected for the NIREP project from a database of 240 normal volunteers. Datasets correspond to 8 males and 8 females with a mean age of 32.5 ± 8.4 and 29.8 ± 5.8 years, respectively. The images are skull-stripped and aligned according to the anterior and posterior commissures. Image dimension is $256 \times 300 \times 256$ with a voxel size of $0.7 \times 0.7 \times 0.7$ mm. Images are distributed with the segmentation of 32 gray matter regions at the frontal, parietal, temporal, and occipital lobes. The most remarkable feature of this dataset is its excellent image quality. The geometry of the segmentations provides a specially challenging framework for deformable registration evaluation. The supplementary material gathers the details on the segmented regions and their visual appearance.

In our previous works, a subsampled version of this dataset has been extensively used for the evaluation of different LDDMM methods. The images of this dataset have been subsampled by reducing image dimension to $180 \times 210 \times 180$ with a voxel size of $1.0 \times 1.0 \times 1.0$ mm. Subsampling is needed to be able to run interesting but memory-demanding benchmark methods and to maintain the continuity of the evaluation results shown in previous works. In our experiments, the first image is selected as the source and warped to the remaining 15 images of the dataset.

6.2. OASIS Learn2Reg22

The open-access series of imaging studies, OASIS (https://www.oasis-brains.org/), is a project aimed at making neuroimaging data sets of the brain freely available to the scientific community. OASIS is divided into different projects with a focus on the study of the anatomical evolution of normal and diseased brains.

OASIS Learn2Reg22 dataset is a small sample made of 416 3D T1w MRI scans from different subjects. The dataset was first proposed in Learn2Reg21 challenge with the intention to assess the performance of non-rigid registration methods in the alignment of small structures of variable shape and size from monomodal MRI. Slight modifications in the Learn2Reg21 data were performed for Learn2Reg22 including the renaming of the files and the permutation and flip of some dimensions to provide the images in the space of the MNI 152 template. Thus, the image dimension in OASIS Learn2Reg22 dataset is $160 \times 224 \times 192$ while it is $160 \times 192 \times 224$ in OASIS Learn2Reg21. This is relevant for those models trained before 2022 with limitations on the input dimensions.

The original images were pre-processed for the HyperMorph paper [90]. Preprocessing included resampling and alignment to a common template and skull stripping. The segmentations were automatically obtained using FreeSurfer and SAMSEG from the neurite package.

A total of 35 brain structures are customarily used in the evaluation. The registration pairs are given by the challenge organizers. The supplementary material gathers the details on the segmented regions and their visual appearance.

The validation set, made up of 19 image pairs, can be used for the evaluation of non-rigid registration methods in the case that the validation set has not been involved in the model design. The segmentations are not available for the test set. Indeed, the evaluation of methods using the test set needs to go through an official submission of results to the challenge organizers. In our experiments, we used the validation set for evaluation.

6.3. Implementation details

The experiments were run on a machine equipped with one NVidia GeForce RTX 3090 Ti with 24 GB of video memory and an Intel Core i7 with 64 GB of DDR3 RAM. The C++ code of ANTS library was used for the SyN method. The LDDMM codes were developed in the GPU with MATLAB. The Python codes and models publicly available for Mermaid and the deep-learning methods were used for the remaining methods. We needed to install different Python environments due to the fast obsolescence of code dependencies.

For the traditional methods, we used the same implementation and parameters as in our previous works [27]. All methods were embedded into a multi-resolution scheme of three levels. Gradient-descent, Gauss–Newton, and Gauss–Newton–Krylov were implemented with an efficient method for the update of the step size based on offline backtracking line-search combined with a check on Armijo's condition. We used the stopping conditions in [46]. Otherwise, the optimization was stopped after 50 iterations in the case of gradient-descent and Gauss–Newton and after 5 inner × 10 outer iterations in the case of Gauss–Newton–Krylov.

Regularization parameters were selected from a search of the optimal parameters in NIREP16 and OASIS datasets. Thus, we used $\sigma^2 = 1.0$, $s = 2$, and a unit-domain discretization of the image domain Ω. We selected $\alpha = 0.0010$ for NIREP16 and $\alpha = 0.0025$ for OASIS. Details on the selection of α are given in the supplementary material.

ANTS was run with the following parameters
`synconvergence="[50x50x50,1e-6,10]",`
`synshrinkfactors="42x1",`
`synsmoothingsigmas="32x1vox".`
The selection of the number of iterations was in agreement with the number of iterations used in gradient-descent and the number of inner × outer iterations used in Gauss–Newton–Krylov optimization for PDE-LDDMM. The selection of the Gaussian smoothing parameters resulted in a minimal regularization with the objective of obtaining a maximal image matching.

Mermaid parameters were selected to be fairly compared with StLDDMM and PDE-LDDMM. Namely, we used a multiresolution strategy with 50 × 50 × 50 iterations, the same shrink factors than SyN, and an initial learning rate leading to a suitable convergence in all cases. The regularization was set to Helmholtz with the same parameters used in the traditional LDDMM methods. Mermaid codes used this regularization for the map model but it was automatically replaced for the scalar and vector momentum map models with a multi-kernel Gaussian. This means that Helmholtz regularizer is not implemented for these variants of Mermaid. Thus, we used the regularization parameters set with Mermaid codes
`multi_gaussian_stds= "[0.05, 0.1, 0.15, 0.2, 0.25]",`
`multi_gaussian_weights= "[0.06666666666666667,`
`0.13333333333333333, 0.19999999999999998,`
`0.26666666666666666, 0.3333333333333333]".`
As we will see in the results section, these parameters imposed a too strong regularization to provide competitive results. However, the large number of parameters and the time complexity of Mermaid made us leave the quest for a less restrictive parameter set out of the scope of this work.

NODEO was executed with the default parameters. Mean filter was used as smoothing kernel. The number of time steps was 2, yielding a stationary parametrization. The number of iterations was 300. The weighting parameters were $\lambda_{\text{reg}} = 0.0005$, $\lambda_{\text{mag}} = 0.05$, and $\lambda_{\text{Jdet}} = 2.5$.

For the deep-learning methods, different image size adjustments were performed in response to the sizes requested by the models. Images were zero-padded and cropped whenever possible. For SynthMorph the images needed to be cropped. For TransMorph, NIREP images and labels need to be resampled. OASIS L2R22 images were remapped into Learn2Reg21 space using flips and permutations.

7. Results

In this section, we show the most relevant results of the experiments conducted to evaluate the performance of the methods considered in this work. First, we provide and extensive evaluation in the NIREP16 database, where the traditional LDDMM methods considered in this work have been previously evaluated consistently throughout our previous works [27]. Next, we provide the evaluation in OASIS validation set in the framework of Learn2Reg22 challenge. As a complement of the evaluation conducted in our work, we provide the most relevant evaluation results in the atlas to IXI registration problem. The details are found in the supplementary material.

7.1. Evaluation metrics

In this study, we use the Dice Similarity Coefficient (DSC) for the evaluation of the accuracy of the methods. Given S and T two regions belonging to the warped source and the target labeling, the DSC is given by

$$DSC(S,T) = 2\frac{|S \cap T|}{|S| + |T|}, \tag{67}$$

where $|\cdot|$ denotes the cardinality of the regions. Thus, the DSC for a given structure is computed from the quotient of twice the cardinality of the intersection between the segmentations of the structures and the cardinality of their union, where segmentations are treated as sets. This metric ranges from 0 to 1. A DSC score of 1 means that the overlap is perfect while a DSC score of 0 means that the structures do not overlap. The DSC values of good-performing methods depend on the difficulty of the dataset. This means that a DSC of 0.5 in a difficult dataset may be as good as a DSC of 0.75 in an easier one. DSC is related to other metrics such as the Target Overlap or Jaccard coefficients, and it has been shown that the relative performance of the different methods is preserved through the used metric [52].

An informative way of comparing the performance of the methods is to plot the distribution of the DSC values in the shape of box-and-whiskers. The DSC values are stored in a matrix where the dimension is the number of structures times the number of experiments. Therefore, two different kinds of boxplots arise. Klein et al. proposed to plot the distribution of the DSC values averaged by structures [52]. This way, we can assess the distribution of the overall accuracy of the methods per experiment. On the other hand, Learn2Reg paper showed the plots of the distribution of the DSC values averaged by experiment [36]. This way, the boxplots represent the variability of the registration methods through the different structures.

According to our experience, the differences in performance among the registration methods can be better visualized with the proposal of Klein et al. Regarding Learn2Reg proposal, it is common that the differences in the difficulties for the registration of the structures are large. In consequence, the box plot and whisker sizes result bigger and the number of outliers rises considerably. Overall, the methods become less distinguishable in the comparison of these distributions.

A well-known issue with the use of box-and-whiskers of the DSC values is that the DSC may disproportionately penalize errors in small regions compared to large regions. Therefore, a method greatly performing in small regions may be penalized with the use of aggregated DSC scores. However, the use of box-and-whisker plots is reasonable to have a global assessment of a large set of methods. The alternative option would be to make a comparison of the accuracy of the methods on the different regions using stratified boxplots, as suggested in [91], but then one misses the big picture in studies like ours since we have observed that methods with similar performance are not usually consistently better than the others in all regions.

We measure the invertibility and smoothness of the transformations using different metrics based on the Jacobian determinant of the transformations (in the following, we use the word Jacobian to refer to its determinant). Namely, we use the Jacobian extrema, the percentage of negative Jacobians, and the standard deviation of the logarithm of the positive Jacobians (SDlogJ). For SDlogJ, we depart from Learn2Reg implementation in the sense that we directly compute the Jacobian of the transformations and we exclude from the standard deviation the points with negative Jacobian.

The Jacobian extrema allow measuring the greatest changes in volume, whether there are foldings in the transformations, and how aggressive they are. The percentage of negative Jacobians allows measuring whether there is a general tendency to fold, or foldings occur in a few isolated examples. The SDlogJ allows measuring the uniformity of the amount of deformation. An informative way of comparing these metrics is to plot the most illustrative ones in colored bubble charts as proposed in [36].

Apart from the quantitative evaluation, it is important to show some illustrative examples for a qualitative evaluation. In this case, the differences between the fixed and the warped images, an RGB coded map of the displacement fields, and the grids of the transformations are used in our work.

7.2. Evaluation in NIREP dataset

7.2.1. Quantitative assessment

Table 2 shows the mean and standard deviation of the DSC values after registration and the measurements obtained from the Jacobians. In addition, Fig. 1 shows, in the shape of box and whisker plots, the statistical distribution of the DSC values after averaging across the 32 segmented structures. In addition, Fig. 5 shows the results of pairwise right-tailed Wilcoxon rank-sum tests, that were conducted for the assessment of the statistical significance of the difference of medians for the distribution of the DSC values. The alternative hypothesis is that the median of the first distribution is higher than the median of the second one.

All the traditional methods with the exception of Mermaid svf-map obtained diffeomorphic or nearly diffeomorphic solutions, being diffeomorphisms the most frequent ones. For the deep-learning methods, the solutions were diffeomorphic or nearly diffeomorphic in NODEO, VM-Diff, VM-GIT, both versions of SynthMorph and SymNet, LapIRN-Diff, TransMorph-Diff IXI, TransMorph-Bspl IXI, VM-Diff IXI, and MIDIR. We consider a method with an average percentage close or inferior to 0.001 negative values nearly diffeomorphic, since negative Jacobians are obtained in around one hundred points. A method with an average percentage above 0.01 negative values is not considered nearly diffeomorphic, since negative Jacobians are obtained in a range from several hundred to thousands of points.

Compared with the best traditional DSC baseline value of 60.24 obtained by SyN-lNCC, StLDDMM outperformed this baseline with diffeomorphic and nearly diffeomorphic solutions (DSC of 61.52 obtained with NCC and DSC of 62.10 obtained with lNCC). PDE-LDDMM obtained an average DSC close to the baseline for the best-performing methods (60.11 for the method based on the image state equation and 60.92 for the method based on the deformation state equation). Mermaid svf-map also obtained an average DSC close to or outperforming the baseline but with non-diffeomorphic solutions (61.31 obtained with NCC). Compared with the best nearly diffeomorphic deep-learning DSC baseline value of 61.35, obtained by VM-GIT, only StLDDMM outperformed the value and Mermaid svf-map reached the value. With regard to the deep-learning methods, NODEO-lNCC, VM-GIT, SynthMorph-brains, both versions of SyMNet, LapIRN-Diff, and TransMorph-OASIS reached or overpassed the traditional baseline. The deep-learning baseline was reached or overpassed by NODEO-lNCC, SyMNet, and TransMorph-OASIS.

For the traditional methods, the best-performing metric was lNCC or NCC. For NODEO, lNCC was also the best-performing metric. However, for VoxelMorph the models trained with SSD seem to outperform lNCC. Finally, it takes our attention the low performance of the probabilistic diffeomorphic methods (VoxelMorph-Diff and TransMorph-Diff).

These observations are best complemented with the boxplots shown in Fig. 1. For some metrics the distributions of StLDDMM, PDE-LDDMM, and Mermaid svf-map outperform the distribution of ANTS-lNCC. In the case of StLDDMM the differences are significant ($p < 1.0e^{-4}$). From the figure, it is striking the superior performance of NODEO-lNCC.

The performance of VM-GIT, SyMNet-Disp, and LapIRN-Disp is superior to the traditional baseline. In this case, VM-GIT and SyMNet-Disp solutions are nearly diffeomorphic. In addition, the performance of SynthMorph-brains, SyMNet-Diff, and LapIRN-Diff are similar to the traditional baseline and the solutions are nearly diffeomorphic. TransMorph trained with OASIS also outperforms the traditional baseline but at the cost of very aggressive foldings in a large number of points (order of ten thousand points). Finally, it takes our attention that TransMorph trained with OASIS greatly outperforms the methods trained with IXI. One of the reasons may be that TransMorph-OASIS boosts its performance with a loss based on DSC. The pairwise right-tailed Wilcoxon rank-sum tests showed statistical significance coherent with all our observations of superiority ($p < 1.0e^{-4}$).

Fig. 4 shows a partial selection of the boxplots in Fig. 1 with the diffeomorphic and nearly diffeomorphic methods. The best-performing method is NODEO-lNCC. The second position is occupied by SyMNet-Disp. The third position is shared with VM-GIT and StLDDMM. Between the deep-learning and the traditional baselines, we find several StLDDMM and PDE-LDDMM methods, and SyMNet-Diff. Finally, some StLDDMM and PDE-LDDMM methods, SynthMorph-brains, and LapIRN-Diff reach the baseline distribution of SyN-lNCC. Below this baseline, we can find some PDE-LDDMM methods, Mermaid, NODEO-SSD, VM-Diff, TransMorph-Bspl, and MIDIR. The low performance of Mermaid svf scalar and vector momentum map is possibly due to the high regularization imposed by the software. The pairwise right-tailed Wilcoxon rank-sum tests showed statistical significance for the superiority of NODEO-lNCC with the exception of SymNet-Disp ($p = 0.0575$). Comparing SyMNet-Disp with VM-GIT and StLDDMM Wilcoxon test did not show statistical significance ($p = 0.98$ and $p = 0.79$).

We assessed whether the DSC distributions in Fig. 4 may be disproportionately penalizing errors in small regions compared to large regions. To this end, we used stratified box-plots over the 32 regions. We did not find evidence that the global DSC boxplots are penalizing any worthy registration method. These box-plots are shown in the supplementary material.

Fig. 6 shows a graphical representation of the relationship among the mean DSC values, SDlogJ, the percentage of negative Jacobian determinants, and $\max(J)$. Attending to the SDlogJ dimension, there is a clear threshold identifying methods with a large percentage of negative Jacobians (SDlogJ above 1). These methods include Mermaid svf-map, early versions of VoxelMorph, LapIRN-Disp, and the majority of models considered in TransMorph paper. From a close-up of the best DSC (above 59) and SDlogJ (below 0.5), we can see that all the methods are nearly diffeomorphic. The regularization of the baseline method SyN-lNCC is the highest, followed by LapIRN-Diff and PDE-LDDMM. There is an ascending trend between the DSC values and the SDlogJ values.

Table 2
Quantitative results on NIREP16. Mean and standard deviation of the Dice Similarity Coefficient (DSC), maximum and minimum of the Jacobian determinant, percentage of negative Jacobian determinants, and standard deviation of the logarithm of the Jacobian determinant for those points with positive values. The abbreviation sym is used for symmetric, svf is used for stationary velocity field, disp is used for displacement, and diff is used for diffeomorphic. In the column Model we detail either the model used for traditional methods or the name of the file with the trained model used for the deep-learning methods. Boldface indicate, for each family of methods, the one with the best DSC average. With ($\star n$) we indicate that the Jacobian computation failed in n experiments showing extremely large or even *nan* values for $max(J)$. The arrows indicate that high DSC values while not extreme Jacobian determinant values are preferable.

Method	Metric	Model	DSC (%) ↑	max (J) ↓	min (J) ↑	% of $\lvert J_\phi \rvert \leq 0$ ↓	SDlogJ ↓
Affine	–	–	43.56 ± 1.94	–	–	–	–
SyN	SSD	Sym-LDDMM	55.24 ± 2.04	4.73 ± 1.02	0.28 ± 0.09	0.000000	0.13 ± 0.01
SyN	lNCC	Sym-LDDMM	**60.24 ± 1.35**	4.11 ± 0.39	0.24 ± 0.03	0.000000	0.15 ± 0.00
SyN	MI	Sym-LDDMM	55.34 ± 1.33	4.55 ± 0.83	0.29 ± 0.04	0.000000	0.13 ± 0.01
StLDDMM	SSD	LDDMM	59.87 ± 1.77	9.76 ± 3.10	0.18 ± 0.03	0.000000	0.20 ± 0.01
StLDDMM	NCC	LDDMM	61.52 ± 1.44	12.37 ± 4.51	0.14 ± 0.03	0.000000	0.22 ± 0.01
StLDDMM	lNCC	LDDMM	**62.10 ± 1.54**	15.09 ± 4.46	0.02 ± 0.22	0.000046 ± 0.000123	0.27 ± 0.01
StLDDMM	MI	LDDMM	60.47 ± 1.46	14.55 ± 5.94	0.17 ± 0.02	0.000000	0.23 ± 0.01
StLDDMM	NGF	LDDMM	60.66 ± 1.71	9.19 ± 3.14	0.05 ± 0.24	0.000048 ± 0.000186	0.27 ± 0.01
PDE-LDDMM	SSD	State equation	57.48 ± 3.29	13.42 ± 18.96($\star$2)	0.00 ± 0.00	0.000000	0.22 ± 0.02
PDE-LDDMM	NCC	State equation	58.22 ± 4.37	8.22 ± 7.61($\star$1)	0.00 ± 0.00	0.000000	0.22 ± 0.02
PDE-LDDMM	lNCC	State equation	**60.11 ± 2.81**	7.41 ± 3.56	0.01 ± 0.01	0.000000	0.23 ± 0.02
PDE-LDDMM	MI	State equation	58.26 ± 1.78	4.71 ± 2.03	0.03 ± 0.03	0.000000	0.22 ± 0.01
PDE-LDDMM	NGF	State equation	58.73 ± 3.55	7.08 ± 4.29($\star$1)	0.07 ± 0.05	0.000000	0.25 ± 0.05
PDE-LDDMM	SSD	map equation	60.30 ± 2.41	24.67 ± 25.91	0.04 ± 0.02	0.000000	0.24 ± 0.02
PDE-LDDMM	NCC	map equation	**60.92 ± 1.88**	15.59 ± 5.89	0.05 ± 0.03	0.000000	0.24 ± 0.03
PDE-LDDMM	lNCC	map equation	58.81 ± 1.72	5.31 ± 1.13	0.10 ± 0.04	0.000000	0.19 ± 0.01
PDE-LDDMM	MI	Map equation	59.43 ± 1.47	7.93 ± 2.41	0.07 ± 0.02	0.000000	0.23 ± 0.01
PDE-LDDMM	NGF	Map equation	60.29 ± 1.74	18.36 ± 18.83	0.13 ± 0.03	0.000000	0.28 ± 0.01
Mermaid	SSD	svf map	60.23 ± 2.69	75.18 ± 27.38	−13.03 ± 6.65	0.24188 ± 0.03040	1.21 ± 0.07
Mermaid	SSD	svf scalar momentum map	51.21 ± 1.04	2.44 ± 0.66	0.56 ± 0.05	0.000000	0.11 ± 0.01
Mermaid	SSD	svf vector momentum map	48.84 ± 1.50	3.03 ± 0.54	0.35 ± 0.06	0.000000	0.26 ± 0.04
Mermaid	NCC	svf map	**61.31 ± 1.94**	100.69 ± 32.22	−28.69 ± 21.40	0.48353 ± 0.04501	1.66 ± 0.07
Mermaid	NCC	svf scalar momentum map	51.26 ± 1.05	2.40 ± 0.68	0.56 ± 0.05	0.000000	0.11 ± 0.01
Mermaid	NCC	svf vector momentum map	50.30 ± 0.93	1.56 ± 0.14	0.70 ± 0.03	0.000000	0.10 ± 0.01
NODEO	SSD	LDDMM through NODEs	58.75 ± 1.66	4.70 ± 0.72	0.37 ± 0.03	0.000000	0.15 ± 0.01
NODEO	lNCC	LDDMM through NODEs	**63.90 ± 1.62**[a]	11.57 ± 2.97	−0.36 ± 0.42	0.00189 ± 0.00236	0.30 ± 0.03
VM-I	SSD	`cvpr2018_vm1_12`	57.52 ± 2.93	86.12 ± 97.05	−13.04 ± 5.72	1.64137 ± 0.25119	2.95 ± 0.21
VM-I	lNCC	`cvpr2018_vm1_cc`	56.90 ± 1.93	59.70 ± 11.59	−14.89 ± 3.88	2.92957 ± 0.29345	3.90 ± 0.18
VM-II	SSD	`cvpr2018_vm2_12`	59.70 ± 2.52	53.13 ± 26.16	−9.26 ± 5.27	1.04256 ± 0.19816	2.37 ± 0.21
VM-II	lNCC	`cvpr2018_vm2_cc`	58.22 ± 2.25	46.91 ± 7.83	−7.65 ± 1.56	1.63442 ± 0.21334	2.95 ± 0.18
VM-Diff	SSD	`miccai2018_10_02_init1`	55.44 ± 2.53	19.91 ± 6.23	−0.19 ± 0.30	0.00006 ± 0.00006	0.37 ± 0.02
VM-GIT 2021	SSD	`vxm_dense_brain_T1_3D_mse`	**61.35 ± 2.24**	12.48 ± 2.17	0.04 ± 0.06	0.00001 ± 0.00003	0.28 ± 0.01
SynthMorph-shapes	DSC	`shapes-dice-vel-3-res-8-16-32-256f`	56.56 ± 1.35	6.45 ± 0.60	0.05 ± 0.02	0.000000	0.27 ± 0.01
SynthMorph-brains	DSC	`brains-dice-vel-0.5-res-16-256f`	**60.47 ± 1.59**	7.89 ± 0.84	0.01 ± 0.01	0.000003 ± 0.000013	0.28 ± 0.01
SymNet-Disp	lNCC	`SyMNet_fea8_140000`	**62.79 ± 1.83**	12.50 ± 1.48	−0.32 ± 0.35	0.00008 ± 0.00008	0.32 ± 0.01
SyMNet-Diff	lNCC	`SyMNet_smo30_update_80000`	61.13 ± 2.24	15.27 ± 2.25	−0.15 ± 0.27	0.00003 ± 0.00004	0.31 ± 0.01
LapIRN-Disp	lNCC	`LapIRN_disp_fea7`	**62.02 ± 1.31**	32.88 ± 5.84	−7.47 ± 1.63	1.33800 ± 0.27667	2.67 ± 0.26
LapIRN-Diff	lNCC	`LapIRN_diff_fea7`	60.44 ± 1.23	6.66 ± 0.78	0.16 ± 0.03	0.000000	0.22 ± 0.01
TransMorph OASIS	lNCC	`TransMorph_Validation_dsc0.857`	**61.70 ± 1.67**	22.00 ± 3.96	−3.78 ± 0.81	0.59623 ± 0.04706	1.81 ± 0.07
TransMorphLarge OASIS	lNCC	`TransMorphLarge_Validation_dsc0.8623`	61.60 ± 1.61	16.63 ± 2.57	−3.36 ± 0.57	0.33412 ± 0.04067	1.37 ± 0.08
TransMorph IXI	lNCC	`TransMorph_Validation_dsc0.744`	55.70 ± 1.41	23.75 ± 4.07	−4.75 ± 1.01	0.92994 ± 0.05306	2.24 ± 0.06
TransMorph-Diff IXI	SSD	`TransMorph_diff_Validation_dsc0.604`	51.20 ± 1.61	15.09 ± 4.55	−0.52 ± 0.46	0.00489 ± 0.00719	0.46 ± 0.04
TransMorph-BSpl IXI	lNCC	`TransMorph_bspl_Validation_dsc0.750`	**58.95 ± 1.69**	10.63 ± 1.90	0.06 ± 0.02	0.000000	0.28 ± 0.00
TransMorph-Bayes IXI	lNCC	`TransMorph_Bayes_Validation_dsc0.743`	58.88 ± 1.58	21.93 ± 6.46	−4.55 ± 1.49	0.67174 ± 0.06248	1.91 ± 0.09
VM-I IXI	–	`VoxelMorph_1_Validation_dsc0.720`	49.15 ± 1.49	35.46 ± 16.01	−8.15 ± 1.73	1.49607 ± 0.11320	2.83 ± 0.10
VM-II IXI	–	`VoxelMorph_2_Validation_dsc0.725`	52.35 ± 1.67	40.75 ± 15.04	−14.01 ± 4.94	1.12015 ± 0.11074	2.46 ± 0.12
VM-Diff IXI	SSD	`VoxelMorph_diff_Validation_dsc0.591`	52.31 ± 1.81	17.79 ± 4.32	−1.26 ± 0.84	0.00188 ± 0.00089	0.40 ± 0.01
CycleMorph IXI	lNCC	`CycleMorph_Validation_dsc0.729`	**55.23 ± 1.63**	37.62 ± 11.10	−8.82 ± 1.81	1.21097 ± 0.11454	2.55 ± 0.12
MIDIR IXI	MI	`MIDIR_Validation_dsc0.733`	50.91 ± 1.15	19.19 ± 3.12	0.02 ± 0.06	0.00000 ± 0.00001	0.36 ± 0.01
ViT IXI	SSD	`ViTVNet_Validation_dsc0.726`	**58.80 ± 1.60**	25.46 ± 5.35	−7.20 ± 1.69	1.08126 ± 0.06503	2.41 ± 0.07
PVT IXI	–	`PVT_Validation_dsc0.720`	53.71 ± 1.60	30.28 ± 4.28	−6.44 ± 2.39	1.60610 ± 0.08508	2.92 ± 0.07
CoTr IXI	DSC	`CoTr_Validation_dsc0.730`	31.95 ± 1.45	174.47 ± 7.73	−170.51 ± 4.20	46.05814 ± 0.02322	12.26 ± 0.00
nnFormer IXI	–	`nnFormer_VAlidation_dsc0.739`	49.59 ± 1.51	18.13 ± 7.34	−6.58 ± 2.10	0.80107 ± 0.06446	2.09 ± 0.08

[a] We indicate the method with the best DSC average.

7.2.2. Qualitative assessment

Fig. 7 shows sagittal views of the differences after registration of the best DSC-performing methods, bolded in Table 2. The methods reduce the differences after registration to different extents. Despite the excellent DSC values, NODEO shows errors in intensity matching in the boundary of the corpus callosum, caudate nucleus, and the parietal lobe. SynthMorph can be highlighted with a general poor visual image matching. Both versions of LapIRN show a general low image matching at the parietal and the external layer of the frontal lobe, and the boundary of the corpus callosum.

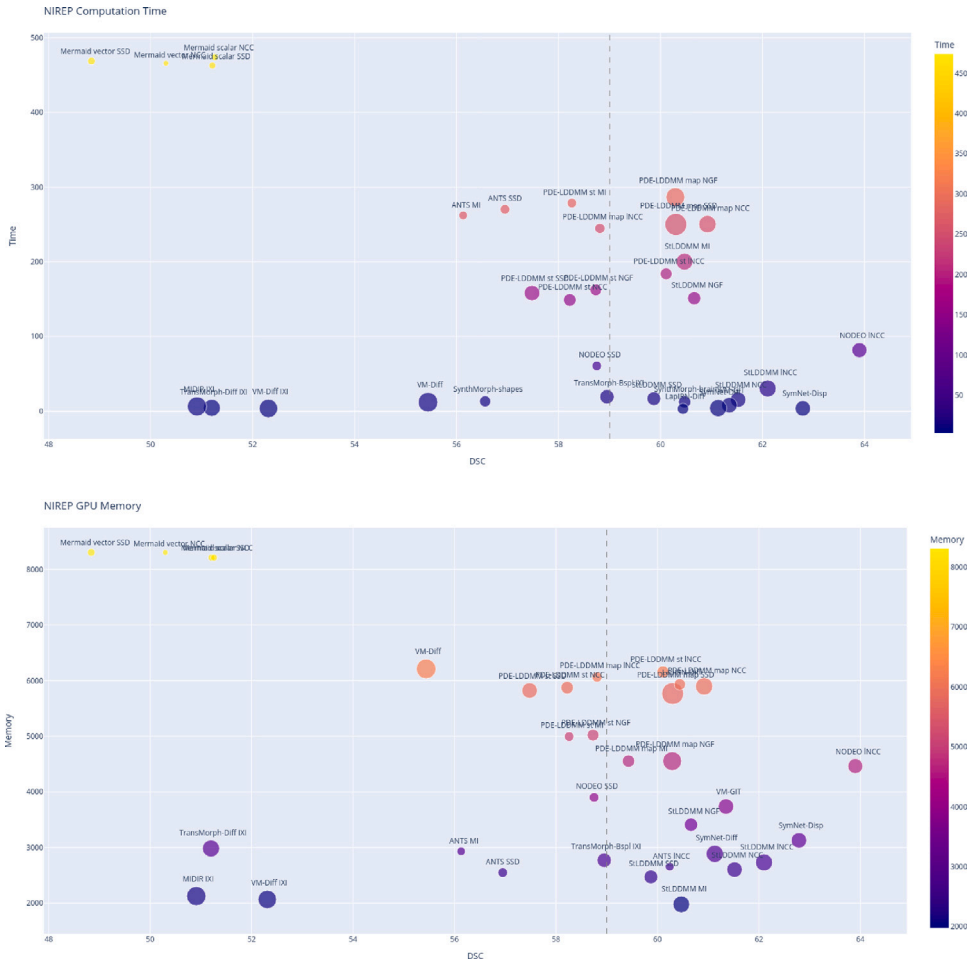

Fig. 1. NIREP16. Volume overlap obtained by the registration methods measured in terms of the DSC between the warped and the corresponding manual target segmentations. Box and whisker plots show the distribution of the DSC values averaged over the 32 NIREP manual segmentations. The boxes indicate the first, second, and third quartile of the DSC values. The whiskers indicate the minimum and maximum of the DSC values, leaving outside the outliers, which are marked with circles. The vertical purple line indicates the median of the baseline traditional method (ANTS lNCC) and the vertical blue line indicates the median of the baseline deep-learning method (VM-GIT), facilitating the comparisons.

Figs. 8–11 show sagittal views of the displacement fields and the transformation grids of the methods considered in this work. For each method, the variant with the best-performing metric is shown. The visual smoothness of traditional LDDMM methods can be also appreciated in VM-Diff, VM-GIT, LapIRN-Diff, TransMorph-Bspl, and MIDIR. It takes our attention the artifacts shown in NODEO. SynthMorph shapes seems to provide smoother displacements than SynthMorph brains. The artifacts shown in CoTr and nnFormer certainly explain the low performance of these methods.

The visual exploration of the transformation grids is even more informative than the visualization of the displacement fields. The foldings and the lack of smoothness of the non-diffeomorphic methods can be neatly appreciated in the figures. The patterns of deformation of Mermaid svf-map lack of smoothness all over the grid while the strong regularization is appreciated in the deformation patterns of Mermaid svf scalar and vector momentum map. NODEO specializes in obtaining foldings all over the cortex. Its patterns of deformation greatly differ from traditional methods. The patterns of deformation in VM-GIT look similar to StLDDMM and PDE-LDDMM. On the other hand, the deformation patterns of SynthMorph are qualitatively different between the shapes and the brains models. SyMNet also specializes in obtaining foldings all over the cortex, although the range of deformation is smaller than in NODEO. It drives our attention the flawed deformation obtained by LapIRN in the parietal lobe, which explains the intensity differences shown in Fig. 7 in this area. In addition, the deformation of LapIRN-Diff seems to follow a strong regularization with small and smooth deformation patterns located in the corpus callosum contrarily to what it is expected for a good-performing registration method for this image pair in this region. The grids in Fig. 11 show that the DSC accuracy with TransMorph OASIS models are obtained at the expense of foldings all over the brain. Regarding the IXI models, the transformations show unrealistic deformations in the great majority of experiments with the exception of TransMorph-Bspl and MIDIR, which show transformations visually similar to VM-GIT. It is striking the problems with the boundary shown by TransMorph-Diff and VM-Diff.

7.2.3. Computational complexity

Table 3 shows the total computation time and the VRAM peak memory reached through the computations in the NIREP16 database. ANTS was run on the CPU due to the lack of a GPU version of the library. Legacy VoxelMorph codes were also run on the CPU due to the incompatibility of the TensorFlow version needed with our GPU. Fig. 2 shows a graphical representation of the relationship among the mean DSC values (x-axis), the computational complexity (y-axis and color), and $\max(J)$ (circle size) of the nearly diffeomorphic methods.

The computation time in traditional methods that totally (SyN) or partially (LDDMM with MI and NGF) use the CPU was in the order of several minutes. It drives out attention the low efficiency of SyN-lNCC with respect to SyN-SSD or SyN-MI. The computation time of StLDDMM methods is in the order of less than half a minute, competing with some of the deep-learning models in efficiency (15 s for StLDDMM-NCC).

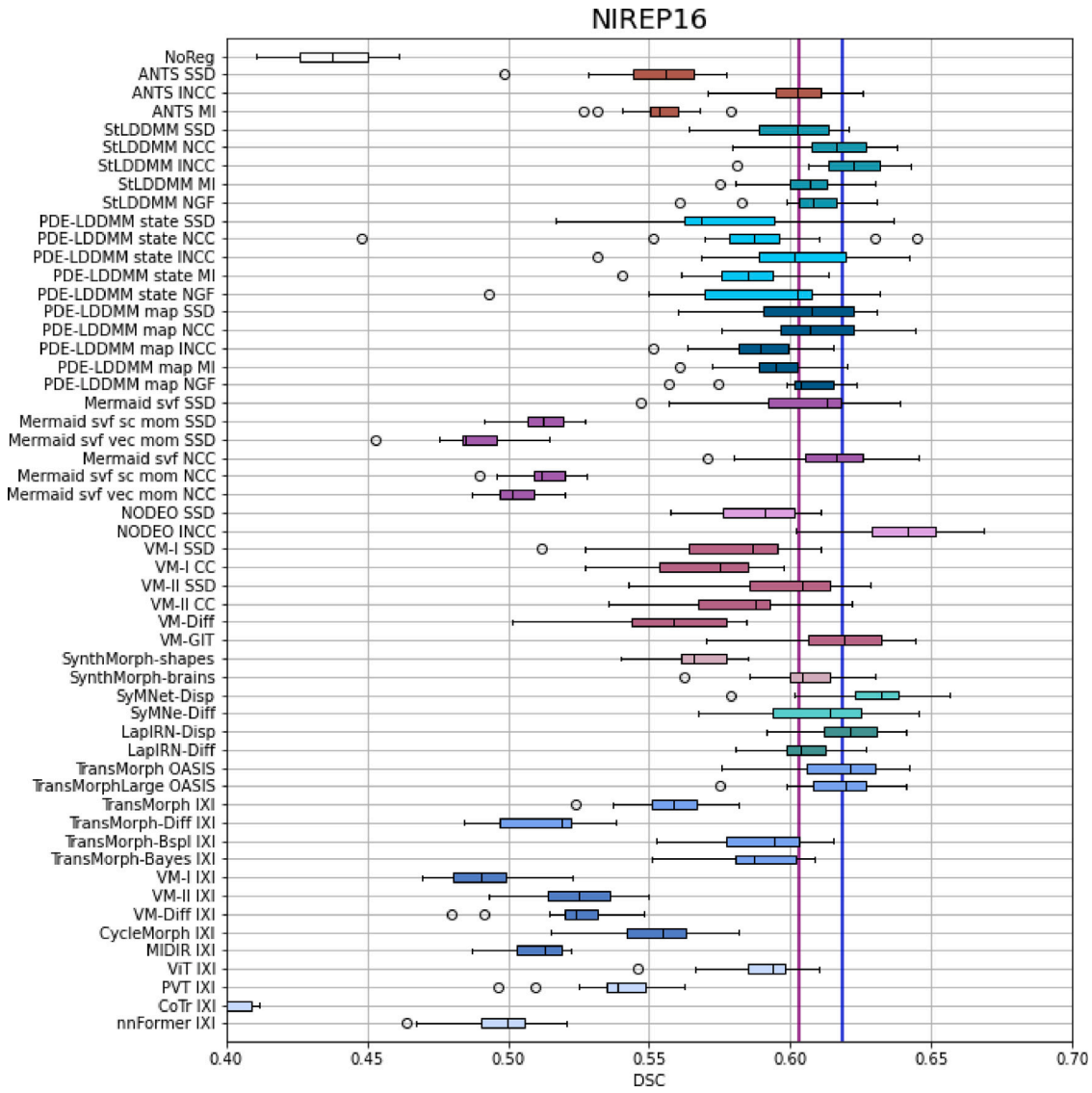

Fig. 2. NIREP 16. Relationship among the mean DSC values (x-axis), the computational complexity (y-axis and colorbar), and $\max(J)$ (bubble sizes), of the nearly diffeomorphic methods. Up figure shows the results for the computation time. Low figure shows the results for the VRAM peak memory. The vertical dashed line indicates the threshold for DSC-accurate methods. Best viewed with zooming.

The more complicated models and numerical implementation of PDE-LDDMM lead to computation times of two to five minutes. Mermaid suite of methods were the least efficient with a computation time close to 8 min in the GPU.

The most efficient deep-learning methods are SyMNet and LapIRN (around 4 s), followed by VoxelMorph models (8 s for VM-GIT). SynthMorph models computation time is close to 10 s. The computation time of TransMorph-related models ranges from 5 to 20 s. NODEO takes over a minute.

Both versions of SynthMorph stand out for their huge memory usage. However, it may be that the 18 GBs were allocated by TensorFlow and the actual memory usage of the algorithm would be much lower. Some of the TransMorph-related methods and Mermaid were among the most memory-consuming ones, reaching almost 10 GBs. Mermaid memory consumption was around 8 GBs.

From the best-performing methods with nearly diffeomorphic solutions, PDE-LDDMM with Gauss–Newton–Krylov optimization shared a cluster of memory-consumption between 5.5 and 6 GBs with LapIRN. Then, PDE-LDDMM with gradient descent optimization shared another cluster of memory-consumption between 4 and 5 GBs with NODEO-lNCC. ANTS, StLDDMM, SyMNet, and TransMorph-Bspl showed a memory-consumption between 2.5 and 3.5 GBs, which represents the cluster of the most memory-efficient methods.

7.3. Evaluation in OASIS dataset

7.3.1. Quantitative assessment

Table 4 shows the mean and standard deviation of the DSC values after registration and the measurements obtained from the Jacobians. In addition, Fig. 3 shows, in the shape of box and whisker plots, the statistical distribution of the DSC values after averaging across the 35 segmented structures, and after averaging across the 19 experiments, and vice-versa. Fig. 5 also shows the results of pairwise right-tailed Wilcoxon rank-sum tests.

For some LDDMM variants, the percentage of negative Jacobian increased with respect to NIREP results. However, all these methods remained nearly diffeomorphic. Mermaid svf-map was still non-diffeomorphic while diffeomorphic for scalar and vector momentum map models. For the deep-learning methods, the solutions were diffeomorphic or nearly diffeomorphic in NODEO-SSD, VM-Diff, VM-GIT, both versions of SynthMorph and SymNet, LapIRN-Diff, TransMorph-Diff IXI, TransMorph-Bspl IXI, VM-Diff IXI, and MIDIR, almost a total coincidence with NIREP with the exception of NODEO-lNCC.

Table 3
NIREP16. Computation time and maximum VRAM memory usage achieved by the registration methods considered in our study. The star symbol ⋆ in $time_{GPU}$ column indicates that some of the algorithm instructions were run on the CPU. In the peak VRAM column, it indicates that the memory usage is in the RAM.

Method	Metric	Model	$time_{GPU}$ (s)	peak VRAM (MBs)
SyN	SSD	SyM-LDDMM	270.31	2550 ⋆
SyN	lNCC	SyM-LDDMM	2065.24	2656 ⋆
SyN	MI	SyM-LDDMM	262.16	2932 ⋆
StLDDMM	SSD	LDDMM	16.88	2472
StLDDMM	NCC	LDDMM	15.39	2605
StLDDMM	lNCC	LDDMM	30.42	2733
StLDDMM	MI	LDDMM	200.26 ⋆	1981
StLDDMM	NGF	LDDMM	151.25 ⋆	3411
PDE-LDDMM	SSD	State equation	158.23	5823
PDE-LDDMM	NCC	State equation	148.93	5875
PDE-LDDMM	lNCC	State equation	183.89	6159
PDE-LDDMM	MI	State equation	278.74 ⋆	4997
PDE-LDDMM	NGF	State equation	162.49 ⋆	5023
PDE-LDDMM	SSD	Map equation	250.05	5769
PDE-LDDMM	NCC	Map equation	250.39	5899
PDE-LDDMM	lNCC	Map equation	244.78	6065
PDE-LDDMM	MI	Map equation	546.74 ⋆	4555
PDE-LDDMM	NGF	Map equation	286.60 ⋆	4555
Mermaid	SSD	svf map	463.56	8131
Mermaid	SSD	svf scalar momentum map	462.77	8213
Mermaid	SSD	svf vector momentum map	468.77	8305
Mermaid	NCC	svf map	483.03	8131
Mermaid	NCC	svf scalar momentum map	474.05	8213
Mermaid	NCC	svf vector momentum map	465.85	8305
NODEO	SSD	LDDMM through NODEs	60.71	3902
NODEO	lNCC	LDDMM through NODEs	81.79	4464
VM-I	SSD	`cvpr2018_vm1_l2`	6.01	5689 ⋆
VM-I	lNCC	`cvpr2018_vm1_cc`	6.19	5881 ⋆
VM-II	SSD	`cvpr2018_vm2_l2`	10.39	5424 ⋆
VM-II	lNCC	`cvpr2018_vm2_cc`	10.45	5586 ⋆
VM-Diff	SSD	`miccai2018_10_02_init1`	12.03	6212 ⋆
VM-GIT 2021	SSD	`vxm_dense_brain_T1w_3D_mse`	8.05	3739
SynthMorph shapes	DSC	`shapes-dice-vel-3-res-8-16-32-256f`	13.21	18 189
SynthMorph brains	DSC	`brains-dice-vel-0.5-res-16-256f`	12.41	18 189
SyMNet-Disp	lNCC	`SyMNet_fea8_140000`	3.67	3130
SyMNet-Diff	lNCC	`SyMNet_smo30_update_80000`	4.09	2888
LapIRN-Disp	lNCC	`LapIRN_disp_fea7`	4.54	5934
LapIRN-Diff	lNCC	`LapIRN_diff_fea7`	3.16	5934
TransMorph OASIS	lNCC	`TransMorph_Validation_dsc0.857`	20.69	2876
TransMorphLarge OASIS	lNCC	`TransMorphLarge_Validation_dsc0.8623`	26.92	4506
TransMorph IXI	lNCC	`TransMorph_Validation_dsc0.744`	9.70	6282
TransMorph-Diff IXI	SSD	`TransMorph_diff_Validation_dsc0.604`	4.45	2984
TransMorph-BSpl IXI	lNCC	`TransMorph_bspl_Validation_dsc0.750`	19.29	2772
TransMorph-Bayes IXI	lNCC	`TransMorph_Bayes_Validation_dsc0.743`	26.75	9796
VM-I IXI	–	`VoxelMorph_1_Validation_dsc0.720`	7.04	4104
VM-II IXI	–	`VoxelMorph_2_Validation_dsc0.725`	3.81	4154
VM-Diff IXI	SSD	`VoxelMorph_diff_Validation_dsc0.591`	3.67	2068
CycleMorph IXI	lNCC	`CycleMorph_Validation_dsc0.729`	9.66	3032
MIDIR IXI	MI	`MIDIR_Validation_dsc0.733`	6.38	2128
ViT IXI	SSD	`ViTVNet_Validation_dsc0.726`	9.52	4760
PVT IXI	–	`PVT_Validation_dsc0.720`	11.40	4724
CoTr IXI	DSC	`CoTr_Validation_dsc0.730`	4.57	7808
nnFormer IXI	–	`nnFormer_VAlidation_dsc0.739`	11.37	3874

Compared with the traditional DSC baseline value of 77.07 obtained by SyN-lNCC, StLDDMM performed similarly with nearly diffeomorphic solutions (DSC of 77.47, obtained with lNCC). PDE-LDDMM based on the deformation state equation obtained an average DSC close to the baseline with diffeomorphic solutions (76.12, obtained with NCC). In this case, PDE-LDDMM based on the image state equation underperformed the baseline for all the metrics. Compared with the best nearly diffeomorphic deep-learning baseline value of 77.78, obtained by SyMNet-Diff, among the traditional methods SyN-lNCC and StLDDMM performed similarly. Regarding the deep-learning methods, NODEO-lNCC, SynthMorph-shapes, LapIRN-Disp, TransMorph-OASIS, TransMorph-BSpl reached or overpassed the traditional and deep-learning baselines. From them, only SynthMorph and TransMorph-Bspl were nearly diffeomorphic.

Table 4
Quantitative results on OASIS L2R22. Same legend as in Table 2. With ($\star n$) we indicate that the Jacobian computation failed in n experiments showing extremely large or even *nan* values for $max(J)$. ($\star\star n$) indicates that Mermaid svf-map with NCC optimization exploded in n experiments, and the DSC could not be obtained. The arrows indicate that high DSC values while not extreme Jacobian determinant values are preferable.

Method	Metric	Model	DSC (%) ↑	max (J) ↓	min (J) ↑	% of $\lvert J_\phi\rvert \leq 0$ ↓	SDlogJ ↓
Affine	–	–	57.18 ± 5.17	–	–	–	–
SyN	SSD	Sym-LDDMM	68.23 ± 5.35	6.14 ± 2.05	0.08 ± 0.21	0.00004 ± 0.00013	0.16 ± 0.03
SyN	lNCC	Sym-LDDMM	**77.07 ± 3.49**	4.36 ± 0.67	0.23 ± 0.04	0.000000	0.18 ± 0.01
SyN	MI	Sym-LDDMM	74.58 ± 3.50	6.78 ± 3.01	0.17 ± 0.06	0.000000	0.17 ± 0.02
StLDDMM	SSD	LDDMM	73.32 ± 3.43	6.41 ± 2.75	0.26 ± 0.12	0.000000	0.16 ± 0.03
StLDDMM	NCC	LDDMM	76.06 ± 2.39	8.28 ± 3.44	0.16 ± 0.09	0.000004 ± 0.000016	0.18 ± 0.02
StLDDMM	lNCC	LDDMM	**77.47 ± 2.81**	32.16 ± 70.55	−0.52 ± 1.05	0.00273 ± 0.00594	0.30 ± 0.05
StLDDMM	MI	LDDMM	72.91 ± 3.83	32.41 ± 60.80	−0.16 ± 0.62	0.000867 ± 0.002033	0.28 ± 0.05
StLDDMM	NGF	LDDMM	73.21 ± 6.48	6.40 ± 3.98	−0.03 ± 0.61	0.000679 ± 0.002884	0.29 ± 0.07
PDE-LDDMM	SSD	State equation	70.44 ± 6.47	5.94 ± 3.34	0.01 ± 0.03	0.000005 ± 0.000013	0.19 ± 0.04
PDE-LDDMM	NCC	State equation	73.07 ± 6.16	14.10 ± 18.54	0.00 ± 0.01	0.000000	0.20 ± 0.03
PDE-LDDMM	lNCC	State equation	**73.30 ± 5.32**	11.51 ± 6.83($\star$1)	0.00 ± 0.01	0.000000	0.26 ± 0.04
PDE-LDDMM	MI	State equation	66.02 ± 13.78	65.53 ± 190.20($\star$1)	0.04 ± 0.04	0.002118 ± 0.006726	0.27 ± 0.13
PDE-LDDMM	NGF	State equation	69.40 ± 4.39	10.04 ± 2.75	0.06 ± 0.04	0.000000	0.32 ± 0.04
PDE-LDDMM	SSD	Map equation	72.80 ± 4.13	14.11 ± 11.87	0.13 ± 0.05	0.000000	0.20 ± 0.03
PDE-LDDMM	NCC	Map equation	**76.12 ± 2.61**	47.39 ± 80.02	0.08 ± 0.03	0.000000	0.22 ± 0.02
PDE-LDDMM	lNCC	Map equation	74.44 ± 3.67	20.35 ± 27.26	0.08 ± 0.04	0.000000	0.25 ± 0.02
PDE-LDDMM	MI	Map equation	71.82 ± 4.53	23.36 ± 27.97	0.09 ± 0.05	0.000000	0.28 ± 0.07
PDE-LDDMM	NGF	Map equation	71.78 ± 4.05	15.47 ± 10.86	0.17 ± 0.04	0.000000	0.29 ± 0.03
Mermaid	SSD	svf map	**69.26 ± 4.74**	50.72 ± 18.49	−2.69 ± 1.99	0.02159 ± 0.00791	0.50 ± 0.05
Mermaid	SSD	svf scalar momentum map	65.44 ± 4.13	1.78 ± 0.52	0.67 ± 0.09	0.000000	0.08 ± 0.02
Mermaid	SSD	svf vector momentum map	51.19 ± 8.87	1.72 ± 0.00	0.46 ± 0.00	0.000000	0.13 ± 0.00
Mermaid	NCC	svf map	73.78 ± 3.84 ($\star\star$ 8)	240.24 ± 89.83	−109.42 ± 69.73	0.97577 ± 0.12407	2.33 ± 0.14
Mermaid	NCC	svf scalar momentum map	64.18 ± 4.46	1.41 ± 0.25	0.77 ± 0.06	0.000000	0.05 ± 0.01
Mermaid	NCC	svf vector momentum map	50.40 ± 8.70	1.41 ± 0.00	0.64 ± 0.00	0.000000	0.11 ± 0.00
NODEO	SSD	LDDMM through NODEs	74.70 ± 3.46	6.16 ± 1.97	0.37 ± 0.04	0.000000	0.17 ± 0.02
NODEO	lNCC	LDDMM through NODEs	**79.13 ± 2.48**	20.63 ± 5.50	−0.56 ± 0.21	0.01450 ± 0.00806	0.48 ± 0.05
VM-I	SSD	`cvpr2018_vm1_l2`	71.57 ± 4.44	70.95 ± 24.72	−9.08 ± 3.28	1.25975 ± 0.26733	2.59 ± 0.26
VM-I	lNCC	`cvpr2018_vm1_cc`	74.11 ± 3.22	68.42 ± 39.51	−7.78 ± 1.63	1.27227 ± 0.20789	2.61 ± 0.20
VM-II	SSD	`cvpr2018_vm2_l2`	72.89 ± 4.54	102.32 ± 60.32	−9.36 ± 3.90	1.10000 ± 0.24781	2.43 ± 0.26
VM-II	lNCC	`cvpr2018_vm2_cc`	**75.26 ± 3.43**	45.52 ± 18.66	−4.78 ± 1.34	0.71020 ± 0.16217	1.97 ± 0.21
VM-Diff	SSD	`miccai2018_10_2_init1`	72.20 ± 4.50	26.34 ± 11.49	−0.08 ± 0.13	0.00007 ± 0.00016	0.37 ± 0.04
VM-GIT 2021	SSD	`vxm_dense_brain_T1_3D_mse`	75.25 ± 3.77	26.61 ± 11.66	0.03 ± 0.02	0.00000 ± 0.00002	0.32 ± 0.04
SynthMorph-shapes	DSC	`shapes-dice-vel-3-res-8-16-32-256f`	**77.45 ± 2.49**	7.56 ± 1.14	0.05 ± 0.01	0.000000	0.31 ± 0.01
SynthMorph-brains	DSC	`brains-dice-vel-0.5-res-16-256f`	75.08 ± 3.15	8.14 ± 1.20	0.03 ± 0.02	0.000000	0.29 ± 0.01
SymNet-Disp	lNCC	`SyMNet_fea8_140000`	76.85 ± 2.53	20.65 ± 5.87	−0.10 ± 0.20	0.00002 ± 0.00002	0.35 ± 0.02
SyMNet-Diff	lNCC	`SyMNet_smo30_update_80000`	**77.78 ± 2.66**	18.82 ± 3.78	−0.11 ± 0.24	0.00003 ± 0.00005	0.34 ± 0.01
LapIRN-Disp	lNCC	`LapIRN_disp_fea7`	**81.60 ± 2.44**	28.58 ± 12.40	−1.90 ± 0.34	0.28960 ± 0.09427	1.27 ± 0.19
LapIRN-Diff	lNCC	`LapIRN_diff_fea7`	77.03 ± 3.18	4.56 ± 1.25	0.24 ± 0.05	0.000000	0.15 ± 0.01
TransMorph OASIS	lNCC	`TransMorph_Validation_dsc0.857`	85.75 ± 1.42	36.01 ± 12.31	−6.13 ± 2.19	0.96647 ± 0.17053	2.28 ± 0.19
TransMorphLarge OASIS	lNCC	`TransMorphLarge_Validation_dsc0.8623`	**86.23 ± 1.40**[a]	35.69 ± 14.48	−5.91 ± 1.93	0.91601 ± 0.16652	2.22 ± 0.19
TransMorph IXI	lNCC	`TransMorph_Validation_dsc0.744`	76.30 ± 3.23	35.81 ± 12.28	−5.20 ± 0.85	1.14856 ± 0.23298	2.47 ± 0.24
TransMorph-Diff IXI	SSD	`TransMorph_diff_Validation_dsc0.604`	69.58 ± 3.05	22.06 ± 4.74	−1.89 ± 1.91	0.00285 ± 0.00235	0.42 ± 0.02
TransMorph-BSpl IXI	lNCC	`TransMorph_bspl_Validation_dsc0.750`	**77.04 ± 2.54**	14.65 ± 4.59	−0.09 ± 0.18	0.00010 ± 0.00012	0.32 ± 0.01
TransMorph-Bayes IXI	lNCC	`TransMorph_Bayes_Validation_dsc0.743`	75.00 ± 3.70	38.59 ± 15.44	−5.39 ± 1.53	1.20217 ± 0.25951	2.52 ± 0.26
VM-I IXI	–	`VoxelMorph_1_Validation_dsc0.720`	51.65 ± 6.30	25.90 ± 0.00	−9.20 ± 0.00	1.52077 ± 0.00000	2.85 ± 0.00
VM-II IXI	–	`VoxelMorph_2_Validation_dsc0.725`	52.13 ± 6.13	31.46 ± 0.00	−5.78 ± 0.00	1.31596 ± 0.00000	2.66 ± 0.00
VM-Diff IXI	SSD	`VoxelMorph_diff_Validation_dsc0.591`	49.43 ± 6.34	15.77 ± 0.00	−1.19 ± 0.00	0.00143 ± 0.00000	0.35 ± 0.00
CycleMorph IXI	lNCC	`CycleMorph_Validation_dsc0.729`	**71.70 ± 3.98**	42.68 ± 22.14	−4.79 ± 0.79	0.98096 ± 0.13966	2.30 ± 0.16
MIDIR IXI	MI	`MIDIR_Validation_dsc0.733`	51.62 ± 7.38	8.38 ± 0.00	0.10 ± 0.00	0.000000	0.29 ± 0.00
ViT IXI	SSD	`ViTVNet_Validation_dsc0.726`	**74.69 ± 4.11**	38.75 ± 19.39	−4.34 ± 1.01	0.78960 ± 0.12778	2.07 ± 0.16
PVT IXI	–	`PVT_Validation_dsc0.720`	62.63 ± 3.84	53.65 ± 22.44	−8.40 ± 1.73	1.94102 ± 0.14972	3.20 ± 0.12
CoTr IXI	DSC	`CoTr_Validation_dsc0.730`	32.45 ± 2.99	184.50 ± 0.00	−176.27 ± 0.00	46.65555 ± 0.00000	12.29 ± 0.00
nnFormer IXI	–	`nnFormer_Validation_dsc0.739`	67.00 ± 4.07	27.39 ± 5.54	−8.24 ± 2.25	1.20328 ± 0.16335	2.55 ± 0.17

[a] We indicate the method with the best DSC average.

For the traditional methods, the outperformance of lNCC or NCC metrics over the others seen in NIREP16 database prevails. However, lNCC tends to outperform SSD for the deep-learning methods. The performance of probabilistic diffeomorphic methods is low, as happened with NIREP16.

These observations are complemented with the boxplots shown in Fig. 3. Comparing the two alternatives of representing the DSC distributions (e.g. grouped by structures and grouped by experiments) the proposal of Klein et al. allowed a better comparative assessment due to the reasons already exposed in Section 7.1. In this case, the traditional baseline is among the best diffeomorphic methods and only the distribution of StLDDMM for the lNCC metric reaches the distribution of SyN-lNCC. NODEO-lNCC performance in OASIS is much more modest than in NIREP16, slightly outperforming the traditional baseline. It is striking the performance shown by both TransMorph-OASIS models. Since we are providing the results in OASIS validation set, it may be possible that these results come from a combination of using the DSC metric in the loss and data leakage.

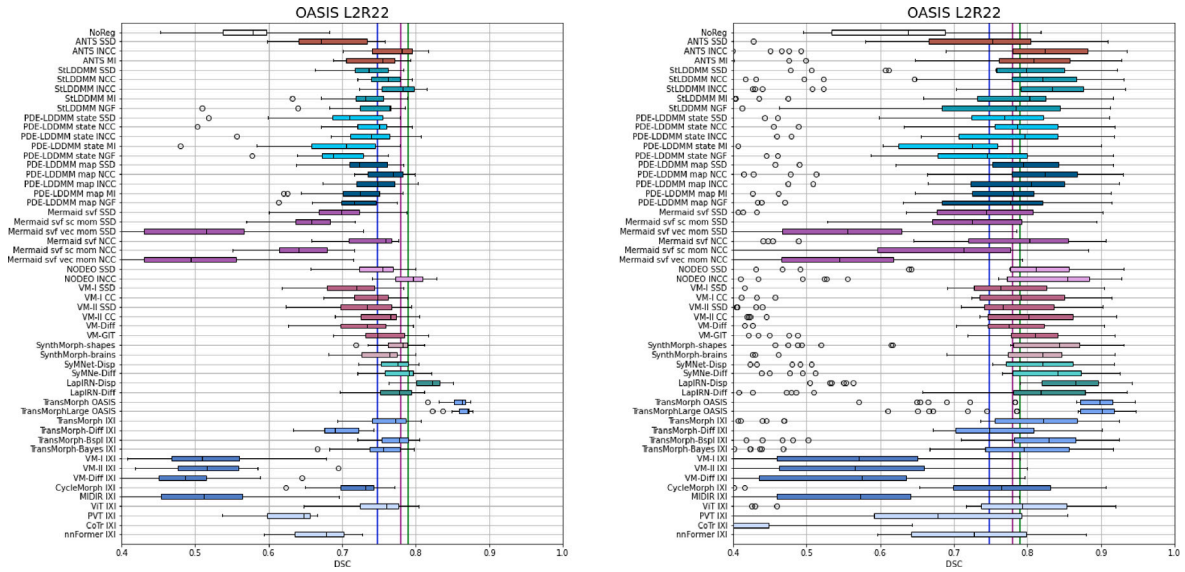

Fig. 3. OASIS L2R22. Volume overlap obtained by the registration methods measured in terms of the DSC between the warped and the corresponding manual target segmentations. Left, box and whisker plots show the distribution of the DSC values averaged over the 35 FreeSurfer segmentations. The vertical purple line indicates the median of the baseline traditional method (ANTS lNCC), the vertical blue line indicates the median of the deep-learning method (VM-GIT), and the vertical green line indicates the median of the baseline deep-learning method (SyMNet-Diff) facilitating the comparisons. Right, box and whisker plots show the distribution of the DSC values averaged over the number of experiments. The boxes indicate the first, second, and third quartile of the DSC values. The whiskers indicate the minimum and maximum of the DSC values, leaving outside the outliers, which are marked with circles. The vertical lines from the left plot are preserved for facilitating the comparisons.

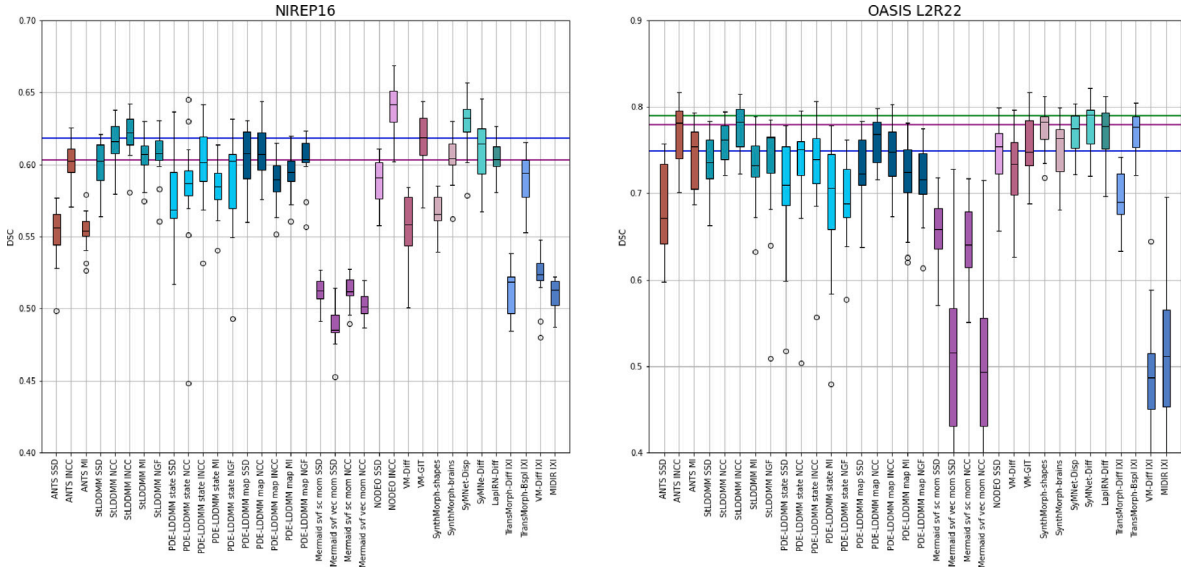

Fig. 4. NIREP16 and OASIS L2R22. Partial selection of the boxplots in Figs. 1 and 3 with the diffeomorphic and nearly diffeomorphic methods. The boxes indicate the first, second, and third quartile of the DSC values. The whiskers indicate the minimum and maximum of the DSC values, leaving outside the outliers, which are marked with circles. The vertical purple line indicates the median of the baseline traditional method (ANTS lNCC) and the vertical blue line indicates the median of the baseline deep-learning method (VM-GIT), facilitating the comparisons.

Overall, these quantitative results are qualitatively different from the results obtained in NIREP. The differences may be due to the different structures segmented in NIREP (mainly in the cerebral cortex) with respect to the structures segmented in OASIS (covering the whole brain and integrating different structures from the cerebral cortex into the same structure). NIREP seems a more challenging data, where the performance of the different image registration methods can be better appreciated.

In the partial selection of boxplots shown in Fig. 4 with the diffeomorphic and nearly diffeomorphic methods of Fig. 3, the best-performing methods show a similar distribution. They include ANTS-lNCC, StLDDMM with lNCC, SynthMorph-shapes, both versions of SyMNet, LapIRN-Diff, and TransMorph-Bspl. Comparing ANTS-lNCC with the others in the corresponding row of the Wilcoxon test did not show statistical significance ($p = 0.56$, $p = 0.50$, $p = 0.28$, $p = 0.73$, $p = 0.43$, and $p = 0.34$, respectively).

Regarding the graphical representation of the relationship among DSC, SDlogJ, the percentage of negative Jacobians, and $\max(J)$ given in Fig. 6, the observations obtained with NIREP still hold. From a close-up of the best DSC (above 75) and SDlogJ (below 0.5), we can see that the regularization of the baseline method SyN-lNCC is high, at the same level as StLDDMM NCC. However, the highest regularization is obtained by LapIRN-Diff. In this case, the ascending trend between the DSC and the SDlogJ values is appreciated in methods from 75.5 points of DSC. Although NODEO-lNCC was classified into non-diffeomorphic methods, the SDlogJ metric is below 0.5.

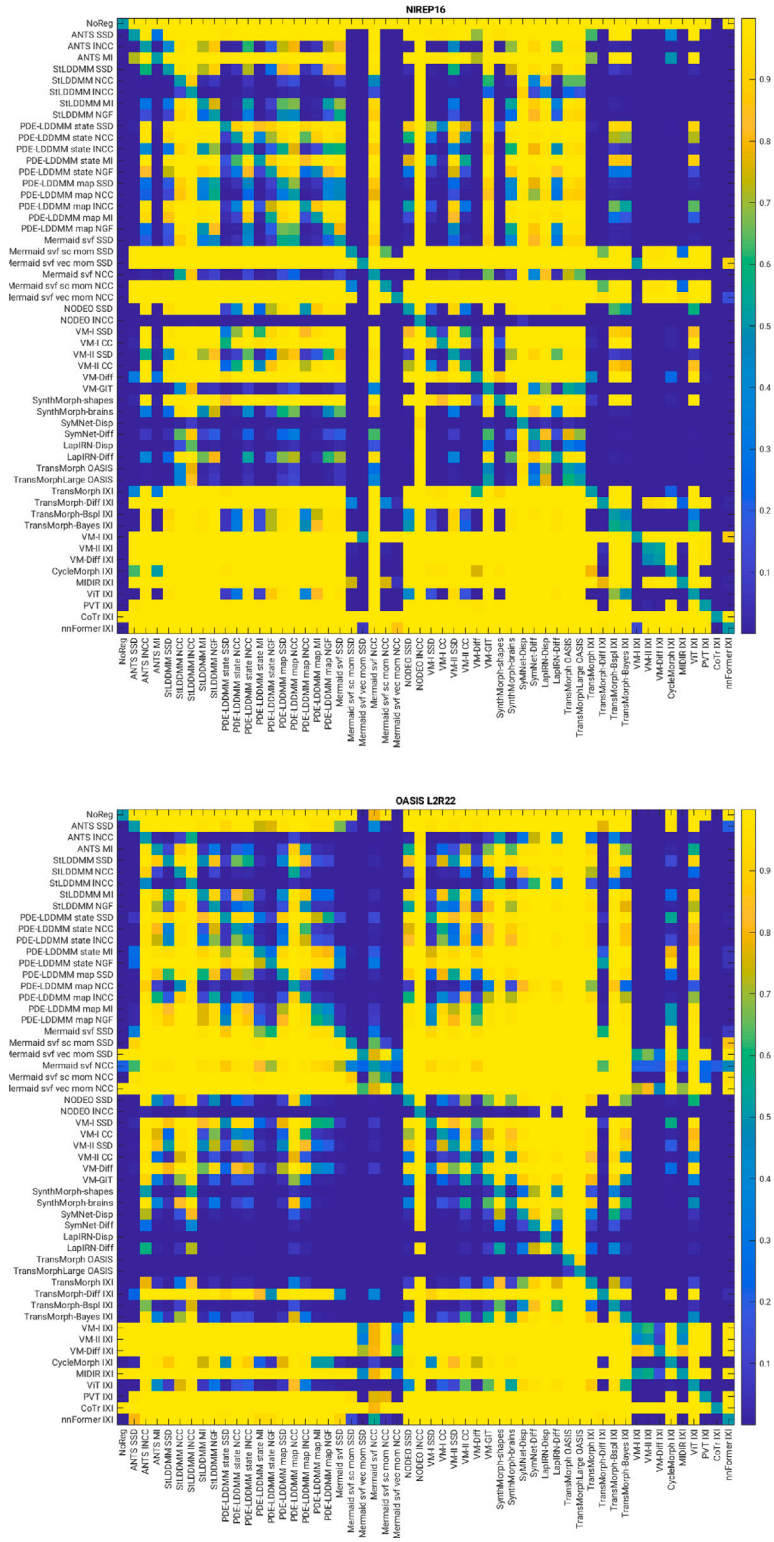

Fig. 5. Results of the pairwise right-tailed Wilcoxon rank-sum tests to assess the statistical significance of the differences among the methods. The matrix depicts the pairwise p-values between each two methods. An straightforward interpretation of the plots is that methods showing blue horizontal lines typically exhibit a statistically significant higher median DSC value distributions.

7.3.2. Qualitative assessment

Fig. 12 shows the sagittal view of the differences after registration of the best DSC-performing methods, bolded in Table 4. The methods show a better reduction of the differences after registration than in NIREP with similar appearances among similarly performing methods. It drives our attention the good visual reduction of differences shown by the traditional methods, NODEO lNCC, or TransMorph OASIS.

Figs. 13 and 14 show sagittal views of the transformation grids of the best-performing variants of the methods considered in this work. In this case, the lack of smoothness can also be appreciated in the figures. The specialization of NODEO in deforming the cortex

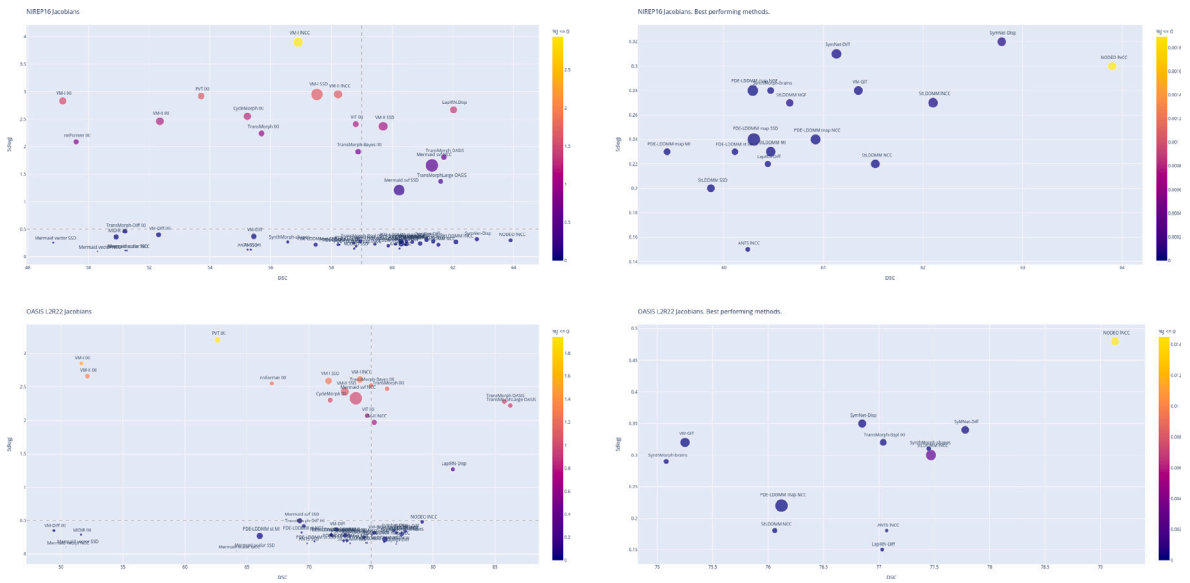

Fig. 6. NIREP16 and OASIS L2R22. Relationship among the mean DSC values (x-axis), the standard deviation of the Jacobian logarithms (SDlogJ, y-axes), the percentage of negative Jacobian determinants (colorbar), and $\max(J)$ (bubble sizes). The left figures show the results in all the considered methods. The right figures show a close-up of the methods with the best DSC and SDlogJ values.

is even more enhanced in this dataset. In this case, the patterns of deformation of traditional, VM-Diff, and VM-GIT do not look similar. The lack of resemblance between SynthMorph models persists. SyMNet and LapIRN-Disp specialize in obtaining small deformations all over the cortex. The patterns of deformation of LapIRN-Diff are small. The deformations of TransMorph OASIS models are too extreme and really unrealistic in locations such as the corpus callosum. The same happens with the deformations obtained with IXI models. The problems with the boundary shown by TransMorph-Diff and VM-Diff in NIREP persist.

7.3.3. Comparison with the state of the art results

According to Learn2Reg test results [36], LapIRN obtained a DSC of 82.0%, and the VoxelMorph methods produced by *3Idiots* and *Winter* groups obtained a DSC of 80.0 and 77.0%, respectively. Our results in the validation set were 81.0 and 77.0% for the LapIRN Disp and Diff models, and 75.25% for VM-GIT. There are profound methodological changes that may be responsible for the differences. LapIRN in Learn2Reg was a conditional version for hyperparameter tuning proposed in [92]. It seems that the VoxelMorph version proposed by *3Idiots* deeply modified different aspects of the original method such as the image similarity, the number of parameters, and training over patches instead of full images. On the other hand, Learn2Reg does not provide any information on the methodology underlying *Winter* submission. There is no information on whether the methods are nearly diffeomorphic, which may explain the lower performance with respect to our results with nearly diffeomorphic methods. Of course, the different training strategies and test datasets may also be responsible for these differences.

According to TransMorph validation results, LapIRN obtained a DSC of 86.1, VoxelMorph 84.7, TransMorph-OASIS 85.80, and TransMorph-Large 86.20%. We were able to reproduce the results with OASIS and TransMorph Vanilla and Large models. Therefore, the results in our study can be considered to complement the results shown in TransMorph paper for OASIS with extension to traditional LDDMM and the deep-learning methods considered in our work. From the gap in performance shown with LapIRN and VoxelMorph, it seems that there is room for improvement between the models used in our work and the models used in TransMorph paper. We suspect that this improvement may pass through the use of conditional networks for hyperparameter learning and the use of DSC losses during training.

8. Discussion

8.1. DSC accuracy as the only metric to evaluate the performance of non-rigid image registration

The problem of evaluation in non-rigid registration is quite complex. Ideally, the best evaluation protocol would go through having the ground truth transformation in a test dataset. As an alternative, the problem could be approached using a non-rigid registration method $\mathcal{M}$ for computing a ground truth transformation and evaluating the performance of the algorithms in the pair of images made of the moving and the warped images. However, this would be biased to the deformation model associated with $\mathcal{M}$. We also could use random transformations, but the test set would not represent a scenario close to real applications. In addition, the aperture problem in optical flow would give problems in low-textured areas and the filling-in effect may hide the real performance of the methods inside the brain. For these reasons, the non-rigid registration community has opted for an evaluation setup based on a bronze standard.

Recent state-of-the-art papers argue the superiority of deep-learning methods with respect to traditional methods in terms of accuracy exclusively measured in terms of the overlap between the warped and reference segmentations. Regularization in deep-learning methods is hard, therefore, the high DSC values are achieved through unrealistic deformations such as foldings or evident mismatches which question their usability in clinical applications.

There is a need to improve the evaluation protocols since the widely trusted DSC metric has well-known drawbacks leading to misleading evaluations. However, there seems to be an acquired inertia from early non-rigid registration evaluation protocols which seems difficult to break. This is not exclusive in the non-rigid image registration community [93]. Our study revealed some of the problems of solely using the DSC for evaluation, and our position is to start combining the DSC values with metrics of the smoothness of the transformation such as the percentage of negative Jacobians and the minimum Jacobian or the Jacobian extrema range.

From our results, the best-performing methods should be those exhibiting superior DSC while also maintaining a controlled percentage of negative Jacobians and a bounded minimum Jacobian. The value $\min(J) = -0.10$ may be an acceptable value for this bound.

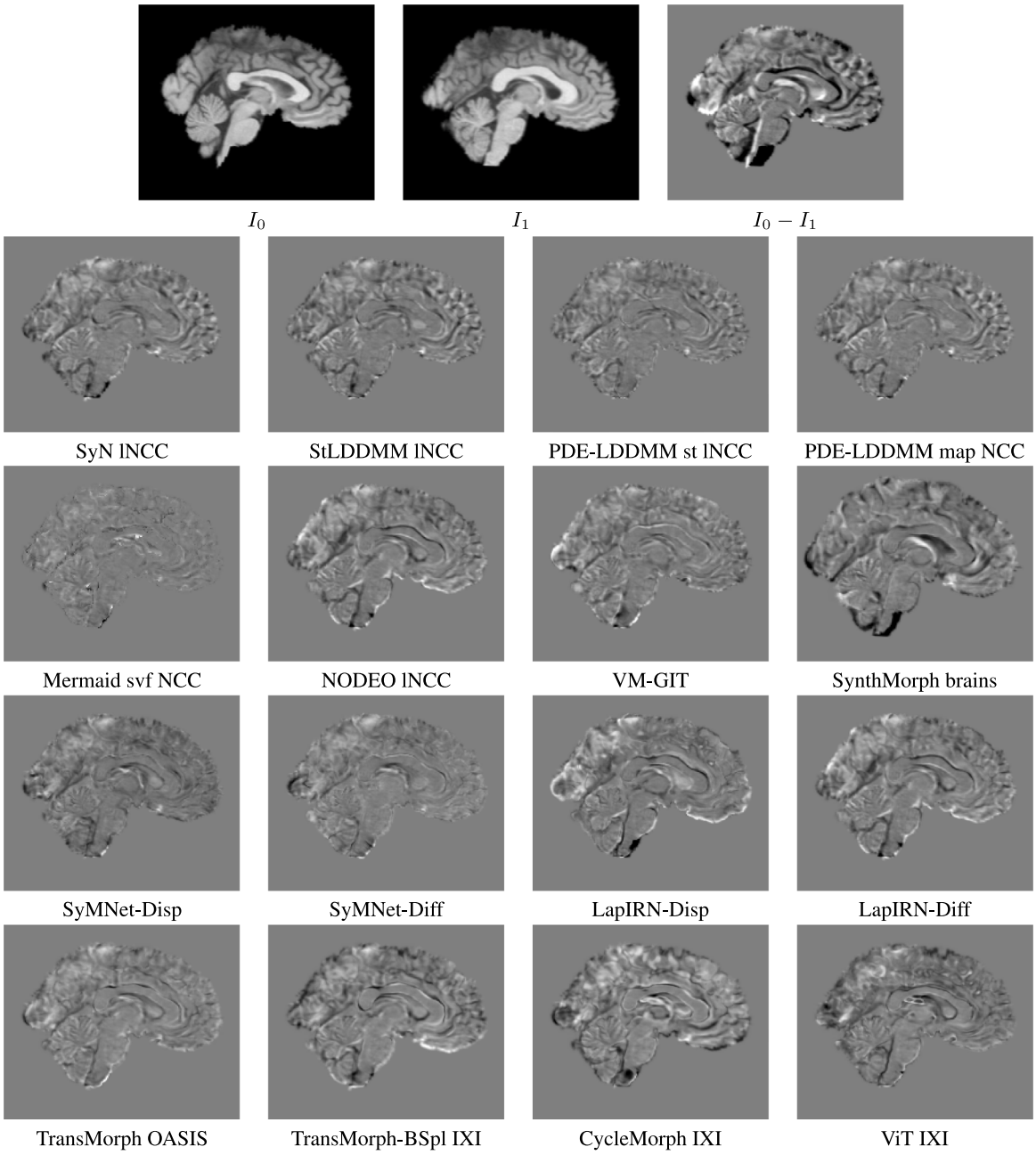

Fig. 7. NIREP16. Sagittal view of the differences after registration $I_0 \circ \phi - I_1$ in a representative experiment. The figure shows the methods with the best DSC mean in each of the considered families (bolded in Table 2). To enrich the comparison, the figure includes the diffeomorphic versions of SyMNet and LapIRN. Gray values indicate no or small differences after registration while white or black values indicate large differences.

8.2. Registration ingredients responsible for high DSC accuracies

For traditional methods, the image similarity metric plays an important role in the obtention of high accuracies. Cross-correlation metrics tend to outperform the others. For the deep-learning methods, both the SSD and cross-correlation metrics were used in the models with the highest accuracies. The replacement of images with segmentations in the training phase also provided high accuracies. For a suitable combination of transformation parametrization, architecture, and loss balance, the training phase with labeled images is able to converge to a model capable of obtaining high DSCs in the inference phase with the original gray-level images. However, the differences of the images after registration are much higher than the differences achieved with intensity-based metrics. We would like to remark that including the metric used for evaluation in the loss function can be a misleading practice, resulting in inflated baseline performance.

For traditional methods, the regularization parameter α greatly influenced the accuracy. A reasonable downgrade in regularization leads to a moderate lack of smoothness and nearly diffeomorphic transformations. We observed that too low regularization resulted in non-diffeomorphic solutions and ODE-stability problems. From previous works, we know that second-order optimization converges to a better local minimum.

For the deep learning methods, the small deformation parametrization leads to higher DSC values than the stationary parametrization. In the great majority of cases, the number of negative Jacobians considerably increased.

For the deep learning methods, the use of symmetric approaches and multiresolution strategies resulted into successful methods. In NIREP

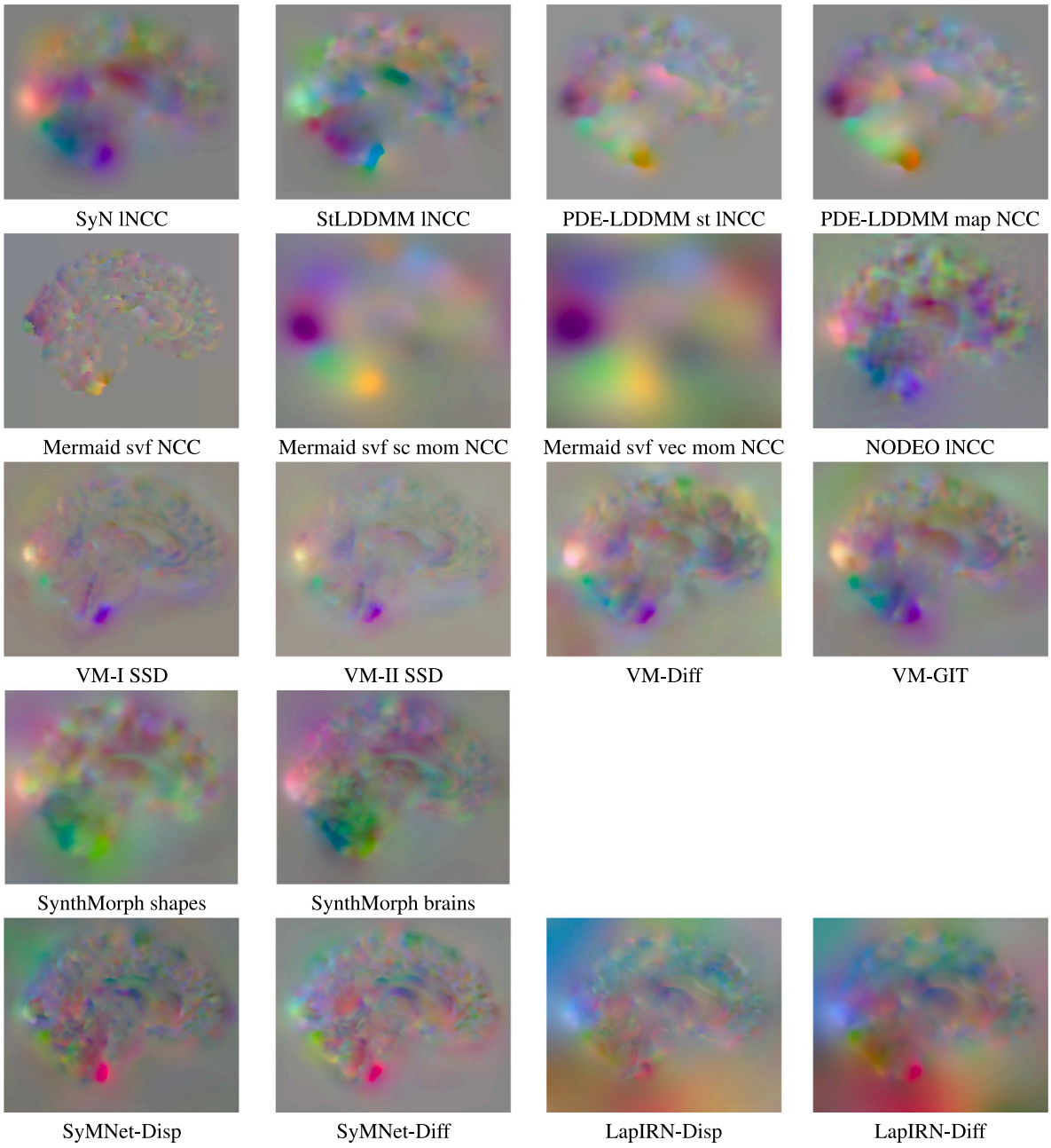

Fig. 8. NIREP16. Sagittal view of the displacement fields in a representative experiment. The figure shows the variant with the best-performing metric for each method. The RGB color map proposed in VoxelMorph paper is used for the color representation of the vector fields. To enrich the comparison, we show the displacement fields of the three variants of Mermaid. The displacements of TransMorph-related methods are shown in Fig. 9.

dataset, transformers did not outperform more conventional architectures. They did in OASIS dataset, but the improvement may be due to the use of DSC in the loss and the use of the validation set as the test set rather than the change in architecture. This observation is corroborated in studies such as Jia et al. [94]. Overall, it was striking the superior performance achieved with the inclusion of neural ODEs proposed in NODEO.

8.3. Registration ingredients responsible for diffeomorphic solutions

For traditional methods, the great majority of the considered variants of LDDMM provided diffeomorphic transformations, where smoothness can be appreciated in the displacement RGB representation and the transformation grids. The exception lies in Mermaid svf-map, where stochastic optimization converged to non-invertible solutions.

For deep learning methods, the use of the stationary parametrization is undoubtedly a necessary condition for obtaining diffeomorphic solutions. The b-splines based parametrization also allowed the obtention of nearly diffeomorphic solutions. Even though the models were not purely diffeomorphic in the great majority of cases but the smoothness can be visually appreciated as with the traditional methods.

The loss $J_{\text{det}} > 0$ acts as extra regularization preventing the model from developing non-diffeomorphic solutions. This observation was obtained from an in-house ablation study with SyMNet and LapIRN. However, it seems that this loss does not regularize enough in NODEO-lNCC. We suspect that VM-GIT and SynthMorph models may combine the stationary parametrization and the positive Jacobian restriction given the percentage of negative Jacobians. Finally, it is striking the high DSC accuracy and the low percentage of negative Jacobians obtained with SymNet-Disp model. Despite the obtained performance,

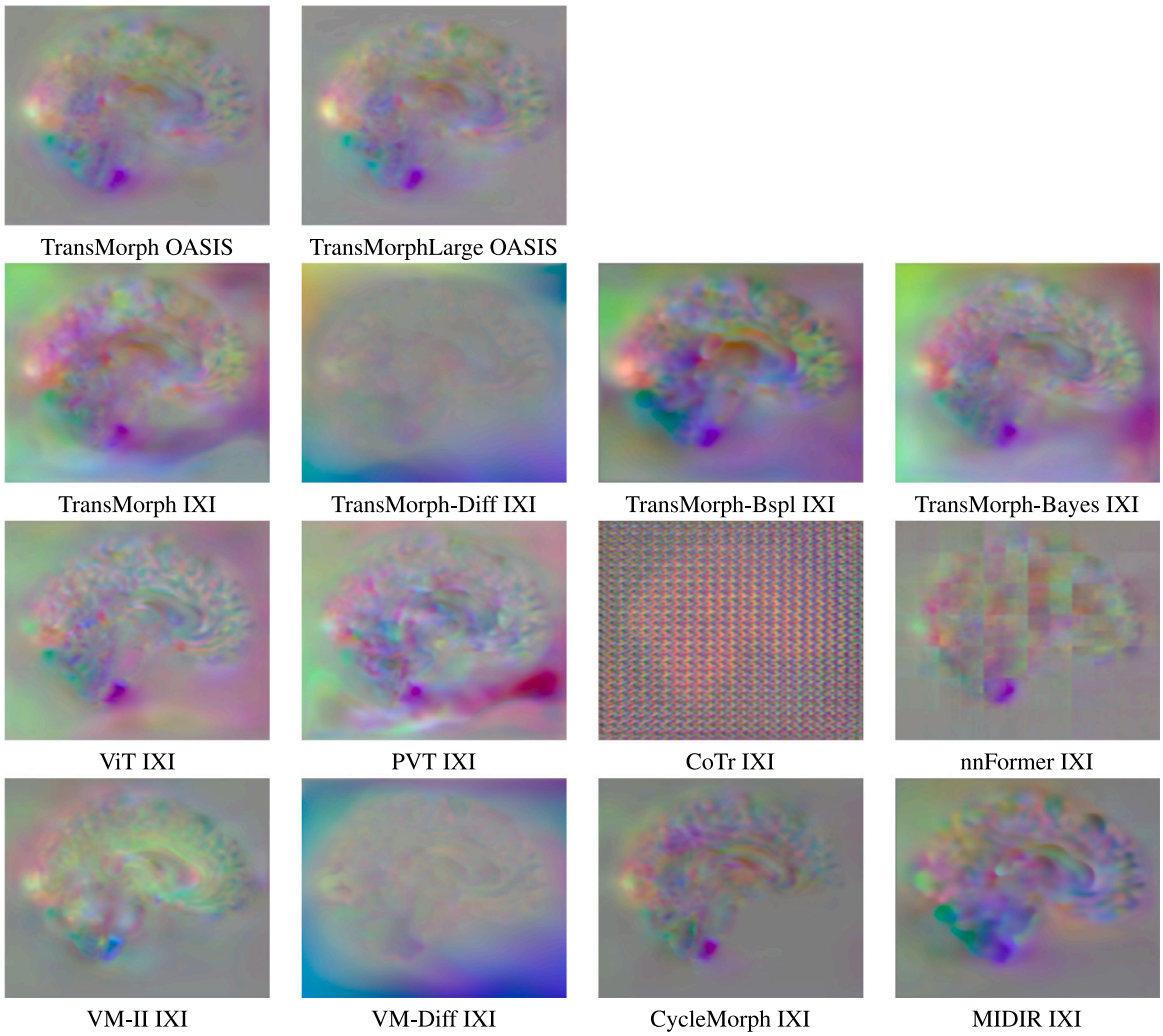

Fig. 9. NIREP16. Sagittal view of the displacement fields in a representative experiment. TransMorph related methods.

we believe that SyMNet-Diff could be preferable, since the small deformation parametrization has shown inconsistency problems between the forward and inverse transformations, $\phi = x+u$ and $\phi = x-u$ respectively (see [14] toy example in C-Shape experiment).

8.4. Diffeomorphism related properties, plausibility, and transformation quality

In early evaluation protocols, it was proposed to use the ability of the methods to provide solutions with inverse consistency and transitivity properties. Together with the diffeomorphic property, this was considered a guarantee of transformation quality provided the mathematically correct behavior of compositions and inversions. The inverse consistency is related to the closeness of $\phi \circ \phi^{-1}$ and $\phi^{-1} \circ \phi$ to the identity. Transitivity is intended to measure correspondence errors when two transformations are composed together. Therefore, given ϕ and ψ two invertible transformations, transitivity measures the closeness of $\phi \circ \psi \circ \psi^{-1} \circ \phi^{-1}$ to the identity. These metrics are intended to favor methods that are able to obtain pure diffeomorphic solutions with parametrizations that favor the inverse consistency and transitivity properties (e.g. stationary or non-stationary). In addition, methods including inverse consistency and transitivity restrictions are amenable to obtain a better score. Therefore, there is an evident bias in the selection of these metrics for evaluation and it would not be fair to use them in our study for establishing a general comparison with methods not imposing these restrictions.

In Learn2Reg, the quality of the transformations is assimilated to plausibility, which is identified with smoothness. This is an oversimplification of the concept plausible because the most probable transformation between two images should be smooth but many of the smooth transformations existing between the images may not be plausible. According to the Python codes provided by the challenge, the smoothness of the transformations is measured in terms of the standard deviation of the clipped logarithm of the shifted Jacobian of the displacement fields. This makes it hard to assess how the foldings in the transformations may affect the quantification of registration smoothness. Replacing this metric with the standard deviation of the log-Jacobian of the transformation, when this quantity is greater than zero, seems a more coherent quantification of registration smoothness, but it does not alleviate the problem of including a downgrade in the metric when the transformations show negative Jacobians. According to our results, our SDlogJ may still do a good job in the discrimination of non-diffeomorphic from nearly diffeomorphic solutions.

An intuitive definition of a plausible transformation for a non-rigid registration problem would be to assume that the images are printed on a super-elastic material or a super-viscous fluid and then impose deformation forces to obtain a global to local alignment of the cerebral structures without tearing the material. These forces would lead to a plausible transformation. We agree that plausibility would be a perfect concept for the quantification of registration quality. However, it is usually hard to know the physical model underlying the transformation between two images and, therefore, the subsequent quantification.

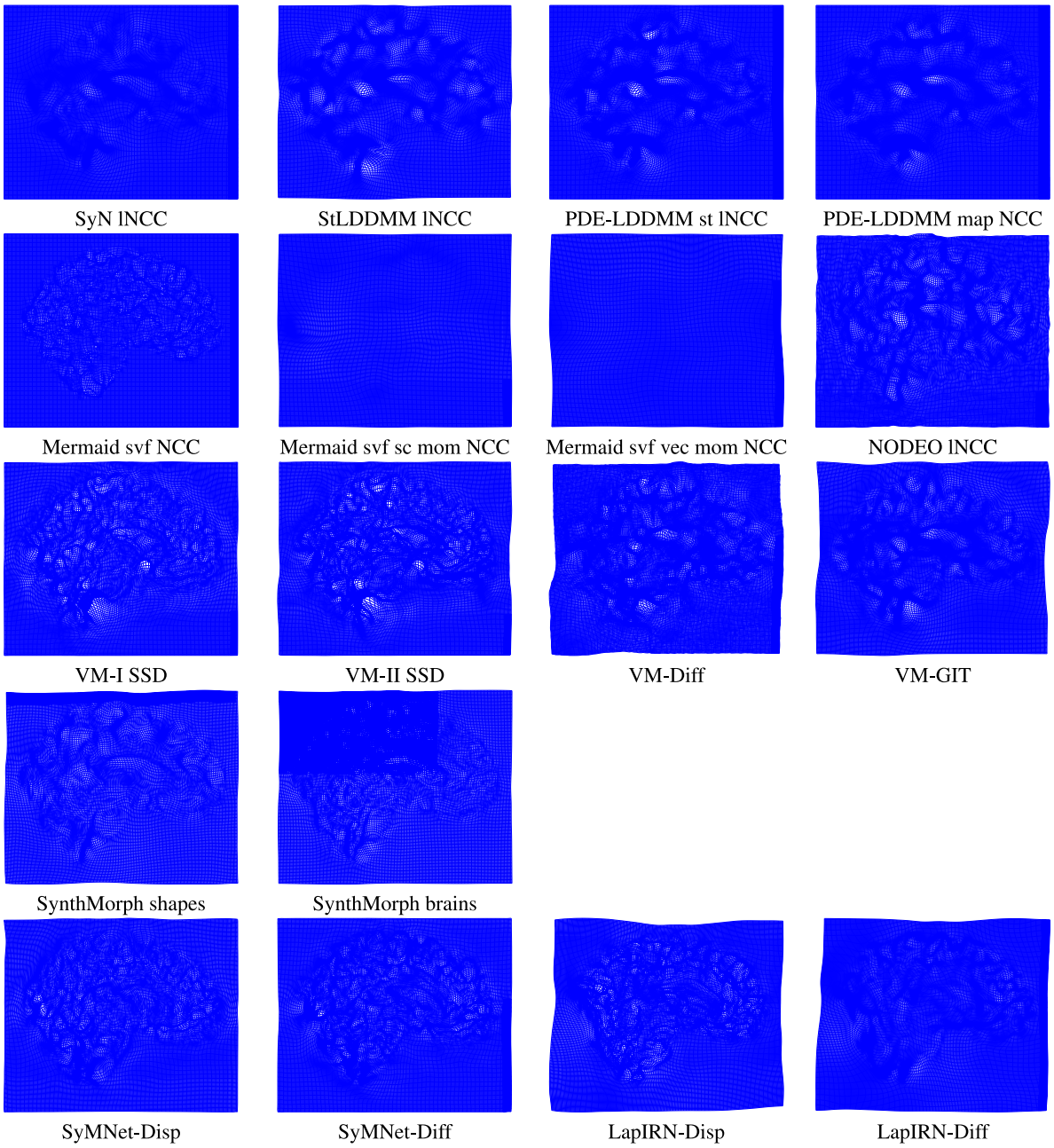

Fig. 10. NIREP16. Sagittal view of the transformation grids in a representative experiment. The figure shows the variant with the best-performing metric for each method. The grids of TransMorph related methods are shown in Fig. 11.

From the visual analysis of the transformation grids in Figs. 10, 11, 13, and 14, we are able to perceive that some methods provide more realistic transformations than others. Indeed, it is easier to identify non-realistic transformations than realistic ones. Our intuition is that smoothness is a necessary condition for plausibility in a great number of applications but it is far from being a sufficient one.

8.5. Which methods and models from our study may be established as benchmarks and baselines to beat with future proposals? which ones should not?

According to the results obtained in this study, we would suggest to consider the diffeomorphic or nearly diffeomorphic methods with the highest DSC values as the best candidates for establishing benchmark methods and baselines. This way, methods are selected among those with both high accuracy and desirable properties that make them potentially usable in clinical applications. As a benchmark for traditional methods, we would suggest StLDDMM with lNCC due to its high accuracy and efficiency. Together with the models VM-GIT, SynthMorph, SyMNet-Diff, LapIRN-Diff, and TransMorph-Bspl, they may constitute a competitive benchmark set. NODEO stands out for its accuracy, however, the artifacts shown in the displacement fields and the inconsistency in the obtention of nearly diffeomorphic solutions makes us recommend to use this method with caution. The DSC boxplots and Jacobian metrics obtained in this study for NIREP16 and OASIS datasets may serve as a baseline for future works. Furthermore, we recommend complementing the quantitative assessment with qualitative findings to provide a more comprehensive evaluation of future methods.

It drove our attention that, despite the similarity of the underlying methodologies below Mermaid and PDE-LDDMM, the latter greatly outperformed the former in terms of convergence stability, accuracy,

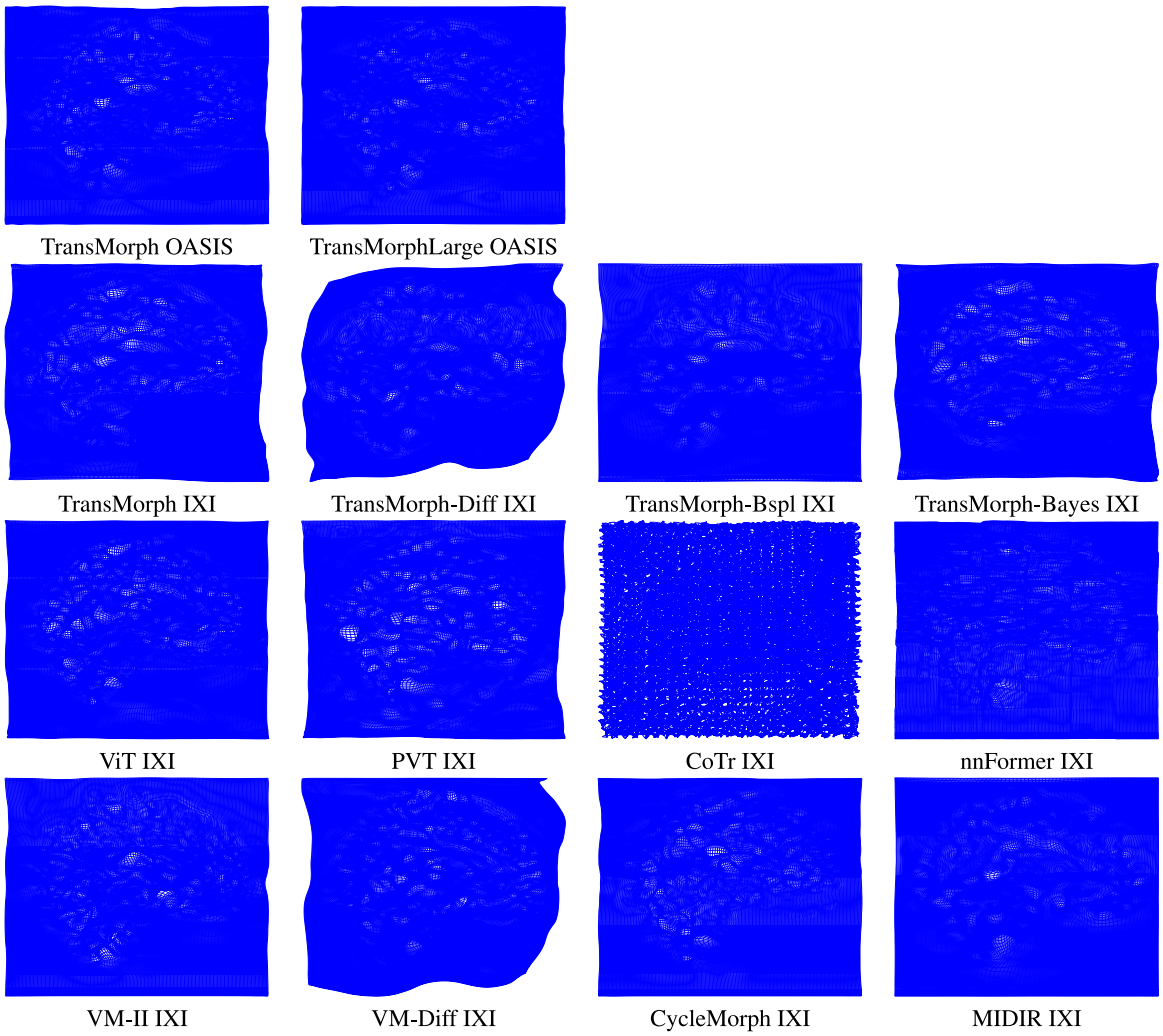

Fig. 11. NIREP16. Sagittal view of the transformation grids in a representative experiment. TransMorph related methods.

and smoothness. It remains to be studied whether the inferior performance of Mermaid svf-map may be due to the use of stochastic gradient descent or more specific design decisions such as the used ODE solvers (i.e. Euler vs. semi-Lagrangian). In addition, it should be studied how to relax the regularization effect of multi-kernel regularizers in Mermaid svf scalar and vector momentum maps.

From the different deep-learning approaches, probabilistic methods (VM-Diff and TransMorph-Diff) notably underperformed their non-probabilistic counterparts. Ashburner et al. first provided a probabilistic formulation of the image registration problem that was assimilated to an energy minimization problem due to the unfeasibility of computing the posterior with the techniques available in 2007 [14]. The problem was solved later on with variational inference, and probabilistic deep-learning methods were proposed for diffeomorphic registration [24,33,68]. The interest in the probabilistic approach lies in the capacity to provide the uncertainty in the obtained solutions, which may help to increase or decrease trust in the obtained transformations and trust may be crucial in clinical applications such as interventional ones. However, the usability of the uncertainty of a method with low accuracy is limited, so our study points out that further research is needed to improve the accuracy of probabilistic methods.

Finally, we lead our attention to the performance obtained with IXI models shown in the Supplementary Material. These models were generated to solve the atlas to image registration problem on the IXI dataset. Although the atlas-to-image registration problem is slightly different from the image-to-image problem, our results indicate that the resulting models are not able to adapt to this changing scenario. It remains to be studied whether the problem is with the special characteristics of an atlas image or with the IXI dataset itself.

8.6. Is it worthwhile to recover the balance between traditional and deep-learning proposals?

Our results show that traditional methods are able to compete with deep-learning methods when performance is measured from the combination of DSC accuracy and transformation quality measured in terms of invertibility and smoothness. However, only a few traditional methods show a computation time competitive with the few seconds that takes the inference phase of deep-learning methods. The use of more complicated models such as PDE-LDDMM inevitably increases the computational complexity to the order of a few minutes. The memory usage at inference is in a similar order of magnitude with interesting analogies between the families.

Despite the evident inferiority of traditional methods in computational complexity with respect to deep-learning methods at inference, there are consistent arguments that make it worth recovering the balance between traditional and deep-learning proposals:

- Traditional methods have complete control of the transformation model. Optimization smoothly minimizes the energy leading the solution in $Diff(\Omega)$ from the identity to a suitable local minimum. Deep-learning methods do not have control of the

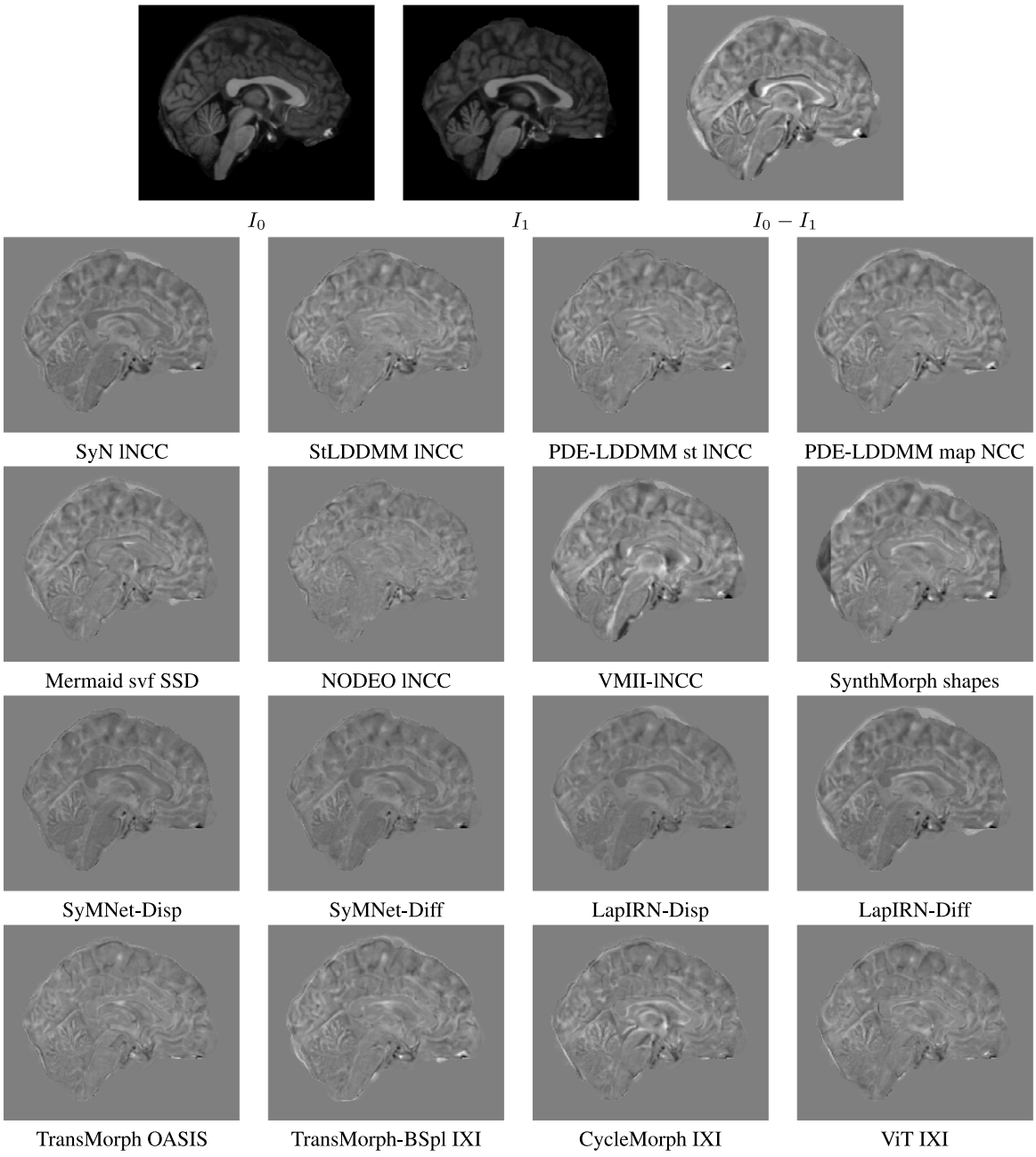

Fig. 12. OASIS. Sagittal view of the differences after registration $I_0 \circ \phi - I_1$ in a representative experiment. The figure shows the methods with the best DSC mean in each of the considered families (bolded in Table 4). To enrich the comparison, the figure includes the diffeomorphic versions of SyMNet and LapIRN. Gray values indicate no or small differences after registration while white or black values indicate large differences.

transformation model during training. Even after a considerable number of epochs with solid clues that the model can be considered to converge, we do not have any guarantee that the solution for an image pair belongs to the transformation model.

- The variability of strategies for training deep-learning models has a combinatorial order. It is not well understood the relationship between the image pair selection strategies used for training and the performance of the resulting model. The generalization obtained with a given training set is hard to reproduce with another dataset. The models are not able to deal correctly with unseen trivial cases (e.g. the registration of the same image in the image pair).
- The generalization capability of under-trained models is low (e.g. a reasonable number of epochs and a reasonable amount of data). Some of the best-performing models are trained in a wide range of datasets (e.g. VoxelMorph was trained in thousands of images from ADNI, OASIS, ABIDE, ADHD200, MCIC, PPMI, HABS, and Harvard GSP). The generalization capability of models trained over healthy populations to diseased ones is not well understood.
- There are clinical problems where there is not enough data for training good-performing models.
- Domain transfer, robustness, and usability remain underexplored problems.
- For deep-learning methods, the computational complexity during training is huge. Training for a batch sizes of one occupies the whole VRAM.

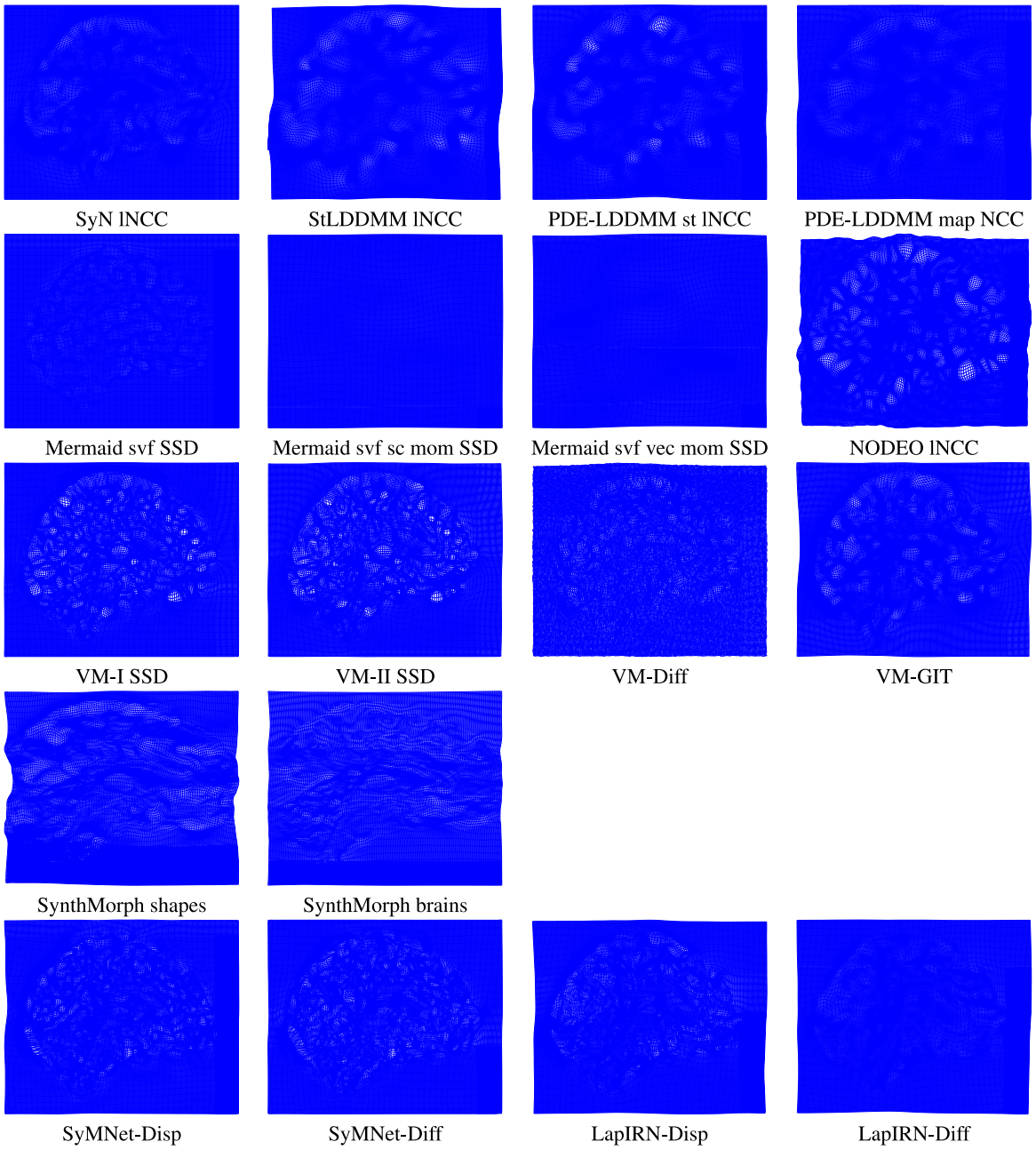

Fig. 13. OASIS. Sagittal view of the transformation grids in a representative example. The figure shows the variant with the best-performing metric for each method. The grids of TransMorph-related methods are shown in Fig. 14. Mermaid svf vec mom NCC experienced convergence problems in this experiment.

These are also opportunities for the improvement of deep-learning methods that may be solved from the cross-fertilization among traditional and deep-learning approaches. Traditional analogs may be used to elucidate whether training is regularizing the model to yield solutions close to the underlying model, to evaluate the generalization capability during training, or to assess whether the model is struggling with scarce data or shift domain problems.

Last but not least, one of the main conclusions of Learn2Reg challenge is that hybrid methods achieved the best DSC accuracies in most applications. So, combining traditional with deep-learning approaches may be a perfect tandem for many applications. As we showed, most of the deep learning-based methods are somehow inspired by traditional methods, as happened in other research fields. Thus, Learn2Reg pieces of evidence together with the theoretical and experimental insights shown in our study may serve to recover the interest in the research in traditional approaches, trying to bridge gaps between both worlds, better understanding, and close up.

9. Conclusions

In this work, we have provided an extensive methodological description and a fair and consistent evaluation of traditional LDDMM and unsupervised deep-learning methods with a focus on diffeomorphic registration. We have covered a wide spectrum of traditional methods belonging to what we like to call the LDDMM-verse. Regarding deep-learning methods, we have focused our study on methods with available source code and models trained in the T1w MRI registration problem, preferably with diffeomorphic variants. We have provided the most relevant theoretical insights of the considered methods, establishing their connections for a unifying view of the non-rigid registration problem and its solutions. We have followed old and new evaluation

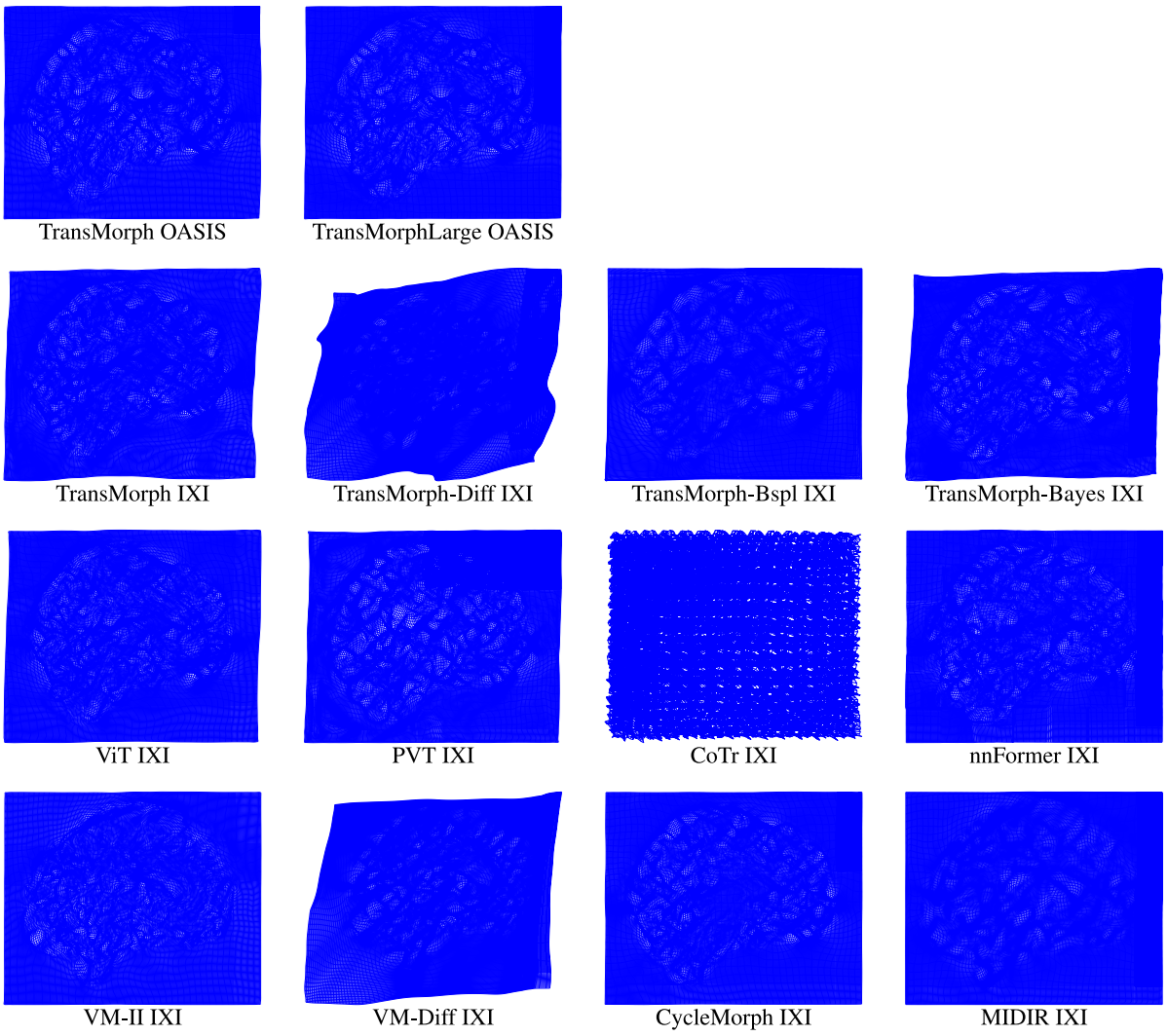

Fig. 14. OASIS. Sagittal view of the transformation grids in a representative experiment. TransMorph related methods.

protocols to give continuity to our previous contributions to the LDDMM family, while complementing Learn2Reg challenge with insightful ways of analyzing the results toward a fair and usable evaluation protocol.

The most important claim from our work is that the quantification of segmentation overlap should not be used alone to establish the performance of a method in the state of the art. Our target for performance quantification should be a combination of segmentation overlap and different metrics reflecting the smoothness and invertibility of the transformations. According to our results, the methods with high DSC scoring while showing diffeomorphic or nearly diffeomorphic solutions should be preferred, since non-diffeomorphic solutions can be visually perceived as non-realistic.

We found that most of the considered traditional methods are prone to obtain diffeomorphic solutions. Regarding deep-learning methods, our evaluation study pointed out the models that are able to obtain nearly diffeomorphic solutions in the datasets used for evaluation. If we restrict our analysis to methods with diffeomorphic or nearly diffeomorphic solutions, we found that traditional methods share a competitive DSC performance with the best-performing deep-learning models. Some traditional methods showed a computation time in the same order as some deep-learning methods and the increased computational complexity of other traditional methods was the result of using more complicated models with interesting and maybe worthy properties. We pointed out the methods that may be used as benchmarks in future evaluation studies and provided the baselines to beat in our evaluation protocol.

The logical subsequent steps from this work would be to extend the analysis to other methods in the LDDMM family, such as EPDiff-constrained LDDMM and band-limited methods with this new perspective on evaluation performance. In addition, deep-learning methods should be compared by the generation of models under the same training conditions to corroborate whether the registration ingredients influencing performance are the ones identified in this work.

There is still a difficult open question: How to lead training toward realistic transformations and how to establish quantitative metrics of plausibility as a proxy to assess transformation quality and realism. We believe that unsupervised methods should rely on stronger regularization strategies toward learning the underlying transformation models. In this sense, it may be even worth to recover supervised approaches. A possible way to approach the problem of plausibility quantification could be to use surrogates linked to clinical applications such as the quantification of the usability of the methods in Computational Anatomy applications or the measurement of the diagnostic capacity of artificial intelligence systems including non-rigid registration in their pipeline.

CRediT authorship contribution statement

Monica Hernandez: Writing – original draft, Visualization, Validation, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Ubaldo Ramon Julvez:** Writing – review & editing, Software, Methodology.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors would like to acknowledge the anonymous reviewers for their thorough and insightful comments. Their valuable feedback and constructive suggestions have significantly contributed to improving the quality of this manuscript. We also would like to acknowledge the authors of the deep-learning methods and the owners of NIREP and OASIS datasets for the publication of the models and data. This work was partially supported by the national research, Spain grants PID2019-104358RB-I00 (DL-Ageing project), PID2022-138703OB-I00 (Trust-B-EyE project), Government of Aragon Group, Spain Reference T64_20R (COS2MOS research group), Ubaldo Ramon-Julvez work is granted by the Government of Aragon, Spain. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.compbiomed.2024.108761.

References

[1] M.I. Miller, Computational anatomy: shape, growth, and atrophy comparison via diffeomorphisms, Neuroimage 23 (2004) 19–33.
[2] M.I. Miller, A. Qiu, The emerging discipline of computational functional anatomy, Neuroimage 45 (1) (2009) 16–39.
[3] A. Sotiras, C. Davatzikos, N. Paragios, Deformable medical image registration: A survey, IEEE Trans. Med. Imaging 32 (7) (2013) 1153–1190.
[4] H. Yang, J. Lyu, R. Tam, X. Tang, A survey on deep learning-based diffeomorphic mapping, in: Handbook of Mathematical Models and Algorithms in Computer Vision and Imaging, Springer, Cham, 2023.
[5] S. Joshi, B. Davis, M. Jomier, G. Gerig, Unbiased diffeomorphic atlas construction for compuational anatomy, Neuroimage 23 (2004) 151–160.
[6] X. Hua, A.D. Leow, N. Parikshak, S. Lee, M.C. Chiang, A.W. Toga, C.R. Jack, M.W. Weiner, P.M. Thompson, ADNI, Tensor-based morphometry as a neuroimaging biomarker for alzheimer's disease: an MRI study of 676 AD, MCI, and normal subjects, Neuroimage 43 (3) (2008) 458–469.
[7] M. Cabezas, A. Oliver, X. Llado, J. Freixenet, M. Cuadra, A review of atlas-based segmentation for magnetic resonance brain images, Comput. Methods Programs Biomed. 104 (3) (2011) e158–77.
[8] S.E. Spasov, L. Passamonti, A. Duggento, P. Lio, N. Toschi, ADNI, A parameter-efficient deep learning approach to predict conversion from mild cognitive impairment to Alzheimer's disease, Neuroimage 189 (2019) 276–287.
[9] A. Routier, N. Burgos, M. Diaz, M. Bacci, S. Bottani, et al., Clinica: an open source software platform for reproducible clinical neuroscience studies, Front. Neuroinform. 15 (2021) 689675.
[10] J.A. Schnabel, P.M. Heinrich, B.W. Papiez, J.M. Brady, Advances and challenges in deformable image registration: From image fusion to complex motion modelling, Med. Image Anal. 33 (2016) 145–148.
[11] A. Uneri, J. Goerres, T. de Silva, Deformable 3D-2D registration of known components for image guidance in spine surgery, in: Proc. of the 19th International Conference on Medical Image Computing and Computer Assisted Intervention, MICCAI'18, in: Lecture Notes in Computer Science (LNCS), Springer- Verlag, Berlin, Germany, 2016, pp. 124–132.
[12] J. Girija, G.N. Krishna, P. Chenna, 4D medical image registration: A survey, in: International Conference on Intelligent Sustainable Systems, ICISS'17, 2017.
[13] B.K. Horn, B.G. Schunck, Determining optical flow, Artificial Intelligence 17 (1981) 185–203.
[14] J. Ashburner, A fast diffeomorphic image registration algorithm, Neuroimage 38 (1) (2007) 95–113.
[15] M.F. Beg, M.I. Miller, A. Trouve, L. Younes, Computing large deformation metric mappings via geodesic flows of diffeomorphisms, Int. J. Comput. Vis. 61 (2) (2005) 139–157.
[16] M.I. Miller, S. Arguillere, D.J. Tward, L. Younes, Computational anatomy and diffeomorphometry: A dynamical systems model of neuroanatomy in the soft condensed matter continuum, WIREs Syst. Biol. Med. 10 (6) (2018).
[17] X. Pennec, S. Sommer, P.T. Fletcher, Riemannian Geometric Statistics in Medical Image Analysis, Academic Press, 2018.
[18] T. Vercauteren, X. Pennec, A. Perchant, N. Ayache, Diffeomorphic demons: Efficient non-parametric image registration, Neuroimage 45 (1) (2009) S61–S72.
[19] X. Pennec, P. Cachier, N. Ayache, Understanding the demon's algorithm: 3D non-rigid registration by gradient descent, in: Proc. of the 2nd International Conference on Medical Image Computing and Computer Assisted Intervention, MICCAI'99, in: Lecture Notes in Computer Science (LNCS), vol. 1679, Springer-Verlag, Berlin, Germany, 1999, pp. 597–605.
[20] A. Dosovitskiy, et al., FlowNet: Learning optical flow with convolutional networks, in: Proc. of the 14th IEEE International Conference on Computer Vision, ICCV'15, 2015.
[21] M.-M. Rohe, M. Datar, T. Heimann, M. Sermesant, X. Pennec, SVF-net, in: Proc. of the 20th International Conference on Medical Image Computing and Computer Assisted Intervention, MICCAI'17, in: Lecture Notes in Computer Science (LNCS), Learning deformable image registration using shape matching, Berlin, Germany, 2017, pp. 266–274.
[22] X. Yang, R. Kwitt, M. Styner, M. Niethammer, Quicksilver: Fast predictive image registration - a deep learning approach, Neuroimage 158 (2017) 378–396.
[23] G. Balakrishnan, A. Zhao, M. Sabuncu, A. Dalca, J. Guttag, An unsupervised learning model for deformable medical image registration, in: Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR'18, 2018, pp. 9252–9260.
[24] G. Balakrishnan, A. Zhao, M. Sabuncu, J. Guttag, A. Dalca, Voxelmorph: A learning framework for deformable medical image registration, IEEE Trans. Med. Imaging 38 (8) (2019) 1788–1800.
[25] J. Nocedal, S.J. Wright, Numerical Optimization, second ed., Springer, New York, 2006.
[26] J. Modersitzki, FAIR: Flexible Algorithms for Image Registration, SIAM, 2009.
[27] M. Hernandez, U. Ramon-Julvez, D.S. Tome, Partial differential equation-constrained diffeomorphic registration from sum of squared differences to normalized cross-correlation, normalized gradient fields, and mutual information: A unifying framework, Sensors 22 (2022) 3735.
[28] M. Hernandez, Primal–dual optimization strategies in huber-L1 optical flow with temporal subspace constraints for non-rigid sequence registration, Image Vis. Comput. 69 (2018) 44–67.
[29] M. Hernandez, Primal-dual convex optimization in large deformation diffeomorphic metric mapping: LDDMM meets robust regularizers, Phys. Med. Biol. 62 (23) (2017) 9067–9098.
[30] S. Klein, M. Staring, J. Pluim, Evaluation of optimization methods for nonrigid medical image registration using mutual information and B-splines, IEEE Trans. Image Process. 16 (12) (2007) 2879–2890.
[31] G.E. Christensen, X. Geng, J.G. Kuhl, J. Bruss, T.J. Grabowski, I.A. Pirwani, M.W. Vannier, J.S. Allen, H. Damasio, Introduction to the non-rigid image registration evaluation project (NIREP), in: Proc. of 3rd International Workshop on Biomedical Image Registration, WBIR'06, Vol. 4057, 2006, pp. 128–135.
[32] M. Zhang, T. Fletcher, Fast diffeomorphic image registration via Fourier-approximated Lie algebras, Int. J. Comput. Vis. (2018).
[33] J. Chen, E.C. Frey, Y. He, W.P. Segars, Y. Li, Y. Du, Transmorph: Transformer for unsupervised medical image registration, Med. Image Anal. 82 (2022) 102615.
[34] M. Hernandez, Combining the band-limited parameterization and semi-lagrangian Runge–Kutta integration for efficient PDE-constrained LDDMM, J. Math. Imaging Vision 63 (3) (2021a) 555–579.
[35] M. Hernandez, Efficient momentum conservation constrained PDE-LDDMM with Gauss–Newton–Krylov optimization, semi-lagrangian Runge–Kutta solvers, and the band-limited parameterization, J. Comput. Sci. 55 (2021b) 101470.
[36] A. Hering, et al., Learn2reg: Comprehensive multi-task medical image registration challenge, dataset and evaluation in the era of deep learning, IEEE Trans. Med. Imaging 42 (3) (2023) 697–712.
[37] G.E. Christensen, R.D. Rabbitt, M.I. Miller, Deformable templates using large deformation kinematics, IEEE Trans. Image Process. 5 (10) (1996) 1435–1447.
[38] T. Polzin, M. Niethammer, M.P. Heinrich, H. Handels, J. Modersitzki, Memory efficient LDDMM for lung CT, in: Proc. of the 19th International Conference on Medical Image Computing and Computer Assisted Intervention, MICCAI'18, in: Lecture Notes in Computer Science (LNCS), Springer- Verlag, Berlin, Germany, 2014, pp. 28–36.
[39] M.F. Beg, A. Khan, Computing an average anatomical atlas using LDDMM and geodesic shooting, in: Proc. of the 3rd IEEE International Symposium on Biomedical Imaging: From Nano to Macro, ISBI'06, 2006, pp. 1116–1119.
[40] N. Singh, J. Hinkle, S.C. Joshi, P.T. Fletcher, A vector momenta formulation of diffeomorphisms for improved geodesic regression and atlas construction, in: Proc. of the IEEE International Symposium on Biomedical Imaging: From Nano to Macro, ISBI'13, 2013.

[41] M. Zhang, P.T. Fletcher, Bayesian principal geodesic analysis for estimating intrinsic diffeomorphic image variability, Med. Image Anal. 25 (1) (2015) 37–44.
[42] N. Singh, J. Hinkle, S.C. Joshi, P.T. Fletcher, Hierarchical geodesic models in diffeomorphisms, Int. J. Comput. Vis. 117 (2016) 70–92.
[43] M. Hernandez, M.N. Bossa, S. Olmos, Registration of anatomical images using paths of diffeomorphisms parameterized with stationary vector field flows, Int. J. Comput. Vis. 85 (3) (2009) 291–306.
[44] M. Hernandez, Gauss-Newton inspired preconditioned optimization in large deformation diffeomorphic metric mapping, Phys. Med. Biol. 59 (20) (2014) 6085–6115.
[45] K.S. Kutten, N. Charon, M. Miller, et al., A diffeomorphic approach to multimodal registration with mutual information: Applications to clarity mouse brain images, 2016, ArXiv.
[46] A. Mang, G. Biros, An inexact Newton-Krylov algorithm for constrained diffeomorphic image registration, SIAM J. Imaging Sci. 8 (2) (2015) 1030–1069.
[47] A. Mang, G. Biros, Constrained H1 regularization schemes for diffeomorphic image registration, SIAM J. Imaging Sci. 9 (3) (2016) 1154–1194.
[48] V. Arsigny, O. Commonwick, X. Pennec, N. Ayache, A log-Euclidean framework for statistics on diffeomorphisms, in: Proc. of the 9th International Conference on Medical Image Computing and Computer Assisted Intervention, MICCAI'06, in: Lecture Notes in Computer Science (LNCS), vol. 4190, Springer- Verlag, Berlin, Germany, 2006, pp. 924–931.
[49] M. Hernandez, M.N. Bossa, S. Olmos, Registration of anatomical images using geodesic paths of diffeomorphisms parameterized with stationary vector fields, in: Proc. of the 11th IEEE International Conference on Computer Vision, ICCV'07, 2007.
[50] M. Hernandez, S. Olmos, Gauss-Newton optimization in diffeomorphic registration, in: Proc. of the 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro, ISBI'08, 2008.
[51] B.B. Avants, C.L. Epstein, M. Grossman, J.C. Gee, Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain, Med. Image Anal. 12 (2008) 26–41.
[52] A. Klein, et al., Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration, Neuroimage 46 (3) (2009) 786–802.
[53] B.B. Avants, N.J. Tustison, G. Song, P.A. Cook, A. Klein, J.C. Gee, A reproducible evaluation of ANTs similarity metric performance in brain image registration, Neuroimage 54 (3) (2011) 2033–2044.
[54] G.L. Hart, C. Zach, M. Niethammer, An optimal control approach for deformable registration, in: Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR'09, 2009.
[55] M. Hernandez, Band-limited stokes large deformation diffeomorphic metric mapping, IEEE J. Biom. Health Inf. 23 (1) (2019a) 362–373.
[56] M. Hernandez, PDE-constrained LDDMM via geodesic shooting and inexact Gauss-Newton-Krylov optimization using the incremental adjoint Jacobi equations, Phys. Med. Biol. 64 (2) (2019c) 025002.
[57] M. Hernandez, A comparative study of different variants of Newton-Krylov PDE-constrained Stokes-LDDMM parameterized in the space of band-limited vector fields, SIAM J. Imaging Sci. 12 (2) (2019b).
[58] P. Ruhnau, C. Schnorr, Optical stokes flow estimation: an imaging-based control approach, Exp. Fluids 42 (2007) 61–78.
[59] A. Mang, L. Ruthotto, A lagrangian Gauss Newton Krylov solver for mass- and intensity-preserving diffeomorphic image registration, SIAM J. Sci. Comput. 39 (5) (2017) B860–B885.
[60] M.I. Miller, A. Trouve, L. Younes, Geodesic shooting for computational anatomy, J. Math. Imaging Vision 24 (2006) 209–228.
[61] M. DoCarmo, Riemannian Geometry, Birkhauser, Boston, 1992.
[62] L. Younes, A. Qiu, R.L. Winslow, M.I. Miller, Transport of relational structures in groups of diffeomorphisms, J. Math. Imaging Vision 32 (2008) 41–56.
[63] F.X. Vialard, L. Risser, D. Rueckert, C.J. Cotter, Diffeomorphic 3D image registration via geodesic shooting using an efficient adjoint calculation, Int. J. Comput. Vis. 97 (2) (2011) 229–241.
[64] O. Ronneberger, F.P.B. T, U-Net: Convolutional Networks for Biomedical Image Segmentation, Vol. 9351, 2015.
[65] M. Jaderberg, K. Simonyan, A. Zisserman, K. Kavukcuoglu, Spatial transformer networks, in: Proc. of Conference on Neural Information Processing Systems, NEURIPS'15, Vol. 2, 2015, pp. 2017–2025.
[66] A. Dalca, G. Balakrishnan, J. Guttag, M. Sabuncu, Unsupervised learning for fast probabilistic diffeomorphic registration, in: Proc. of the 21st International Conference on Medical Image Computing and Computer Assisted Intervention, MICCAI'18, in: Lecture Notes in Computer Science (LNCS), vol. 11070, Springer-Verlag, Berlin, Germany, 2018, pp. 729–738.
[67] A. Dalca, G. Balakrishnan, J. Guttag, M. Sabuncu, Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces, Med. Image Anal. 57 (2019) 226–236.
[68] J. Krebs, H. Delingetter, B. Mailhe, N. Ayache, T. Mansi, Learning a probabilistic model for diffeomorphic registration, IEEE Trans. Med. Imaging 38 (9) (2019) 2165–2176.
[69] M. Hoffmann, B. Billot, D. Greve, J. Iglesias, B. Fischl, A. Dalca, Synthmorph: learning contrast-invariant registration without acquired images, IEEE Trans. Med. Imaging (2021).
[70] B. Billot, D. Greve, O. Puonti, A. Thielscher, K.V. Leemput, B. Fischl, A. Dalca, J. Iglesias, ADNI, Synthseg: Segmentation of brain MRI scans of any contrast and resolution without retraining, Med. Image Anal. 86 (2023) 102789.
[71] T. Mok, A.C. Chung, Fast symmetric diffeomorphic image registration with convolutional neural networks, in: Proc. of the IEEE Computer Society Conference OnComputer Vision and Pattern Recognition, CVPR'20, 2020, pp. 4644–4653.
[72] T. Mok, A. Chung, Large deformation diffeomorphic image registration with Laplacian pyramid networks, in: Proc. of the 23rd International Conference on Medical Image Computing and Computer Assisted Intervention, MICCAI'20, in: Lecture Notes in Computer Science (LNCS), Springer- Verlag, Berlin, Germany, 2020.
[73] S. Zhao, Y. Dong, E. Chang, Y. Xu, Recursive cascaded networks for unsupervised medical image registration, in: Proc. of the 16th IEEE International Conference on Computer Vision, ICCV'21, 2019, pp. 10599–10609.
[74] B. Kim, et al., CycleMorph: Cycle consistent unsupervised deformable image registration, Med. Image Anal. 71 (2021) 102036.
[75] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, in: Proc. of the International Conference on Learning Representations, ICLR'21, 2021.
[76] J. Chen, Y. He, E. Frey, Y. Li, Y. Du, ViT-V-Net: Vision transformer for unsupervised volumetric medical image registration, 2021, ArXiv.
[77] W. Wang, et al., Pyramid vision transformer: A versatile backbone for dense prediction without convolutions, in: Proc. of the 17th IEEE International Conference on Computer Vision, ICCV'21, Vol. 548–558, 2021, pp. 20804–20813.
[78] Y. Xie, J. Zhang, C. Shen, Y. Xia, Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation, in: Proc. of the 24th International Conference on Medical Image Computing and Computer Assisted Intervention, MICCAI'21, in: Lecture Notes in Computer Science (LNCS), vol. 12903, Springer- Verlag, Berlin, Germany, 2021.
[79] H. Zhou, J. Guo, Y. Zhang, L. Yu, L. Wang, Y. Yu, Nnformer: Interleaved transformer for volumetric segmentation, 2021, ArXiv.
[80] H. Qiu, C. Qin, A. Schuh, K. Hammernik, D. Rueckert, Learning diffeomorphic and modality-invariant registration using b-splines, Med. Imaging Deep Learn. (2021).
[81] Y. Wu, T.Z. Jiahao, J. Wang, P.A. Yushkevich, M.A. Hsieh, J.C. Gee, NODEO: A neural ordinary differential equation based optimization framework for deformable image registration, in: Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR'22, 2022, pp. 20804–20813.
[82] R.T.Q. Chen, Y. Rubanova, J. Bettencourt, D.K. Duvenaud, Neural ordinary differential equations, in: Proc. of Conference on Neural Information Processing Systems, NEURIPS'18, Vol. 31, 2018.
[83] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR'16, 2016, pp. 770–778.
[84] G. Gerig, J.M.C. M, Valmet: A new validation tool for assessing and improving 3d object segmentation, in: Proc. of the 4th International Conference on Medical Image Computing and Computer Assisted Intervention, MICCAI'01, in: Lecture Notes in Computer Science (LNCS), Springer- Verlag, Berlin, Germany, 2001.
[85] K. Murphy, B.V. Ginneken, J.M. Reinhardt, S. Kabus, K. Ding, X. Deng, K. Cao, K. Du, G.E. Christensen, V. Garcia, et al., Evaluation of registration methods on thoracic CT: the EMPIRE10 challenge, IEEE Trans. Med. Imaging 30 (11) (2011) 1901–1920.
[86] https://continuousregistration.grand-challenge.org.
[87] Y. Xiao, H. Rivaz, M. Chabanas, M. Fortin, I. Machado, Y. Ou, M.P. Heinrich, J.A. Schnabel, X. Zhong, A. Maier, et al., Evaluation of MRI to ultrasound registration methods for brain shift correction: the CuRIOUS2018 challenge, IEEE Trans. Med. Imaging 39 (3) (2019) 777–786.
[88] J. Borovec, et al., ANHIR: Automatic non-rigid histological image registration challenge, IEEE Trans. Med. Imaging 39 (10) (2020) 3042–3052.
[89] H. Siebert, L. Hansen, M. Heinrich, Fast 3D registration with accurate optimisation and little learning for Learn2Reg 2021, in: Biomedical Image Registration, Domain Generalisation and Out-of-Distribution Analysis, in: Lecture Notes in Computer Science, vol. 13166, 2022.
[90] A. Hoopes, M. Hoffmann, B. Fischl, J. Guttag, A. Dalca, Hypermorph: Amortized hyperparameter learning for image registration, in: Proc. of International Conference on Information Processing and Medical Imaging, IPMI'21, in: Lecture Notes in Computer Science (LNCS), vol. 12729, Springer-Verlag, Berlin, Germany, 2021, pp. 3–17.
[91] T. Rohlfing, Image similarity and tissue overlaps as surrogates for image registration accuracy: widely used but unreliable, IEEE Trans. Med. Imaging 31 (2) (2012) 153–163.
[92] T. Mok, A. Chung, Conditional deformable image registration with convolutional neural network, in: Proc. of the 24th International Conference on Medical Image Computing and Computer Assisted Intervention, MICCAI'21, in: Lecture Notes in Computer Science (LNCS), vol. 12904, Springer- Verlag, Berlin, Germany, 2021.
[93] A. Reinke, M.D. Tizabi, M. Baumgartner, et al., Understanding metric-related pitfalls in image analysis validation, Nat. Methods 21 (2024) 182–194.
[94] X. Jia, J. Bartlett, T. Zhang, W. Lu, Z. Qiu, J. Duan, U-Net vs transformer: Is U-Net outdated in medical image registration? in: Machine Learning in Medical Imaging, MLMI'22, in: Lecture Notes in Computer Science (LNCS), vol. 13583, Springer- Verlag, Berlin, Germany, 2022.