Deep Learning in Medical Image Registration: Magic or Mirage?

Rohit Jena[1,4] **Deeksha Sethi**[1] **Pratik Chaudhari**[1,2,*] **James C. Gee**[1,3,4,*]
[1]Computer and Information Science [2]Electrical and Systems Engineering
[3]Radiology [4] Penn Image Computing and Science Laboratory
{rjena, deesethi, pratikac}@seas.upenn.edu, gee@upenn.edu

Abstract

Classical optimization and learning-based methods are the two reigning paradigms in deformable image registration. While optimization-based methods boast generalizability across modalities and robust performance, learning-based methods promise peak performance, incorporating weak supervision and amortized optimization. However, the exact conditions for either paradigm to perform well over the other are shrouded and not explicitly outlined in the existing literature. In this paper, we make an explicit correspondence between the mutual information of the distribution of per-pixel intensity and labels, and the performance of classical registration methods. This strong correlation hints to the fact that architectural designs in learning-based methods is unlikely to affect this correlation, and therefore, the performance of learning-based methods. This hypothesis is thoroughly validated with state-of-the-art classical and learning-based methods. However, learning-based methods with weak supervision can perform high-fidelity intensity and label registration, which is not possible with classical methods. Next, we show that this high-fidelity feature learning does not translate to invariance to domain shift, and learning-based methods are sensitive to such changes in the data distribution. We reassess and recalibrate performance expectations from classical and DLIR methods under access to label supervision, training time, and its generalization capabilities under minor domain shifts.

1 Introduction

Deformable Image Registration (DIR) refers to the local, non-linear (hence deformable) alignment of images by estimating a dense displacement field. Many workflows in medical image analysis require images to be in a standard coordinate system for comparison, analysis, and visualization. In neuroimaging, communicating and comparing data between subjects requires the images to lie in a standard coordinate system [48, 96, 89, 32, 81, 85]. This assumption universally does not apply when brain image data are compared across individuals or for the same individual at different time points. Anatomical correspondences between diseased patients and normative brain templates help identify and localize abnormalities like tumors, lesions, or atrophy. Failed or anomalous correspondences impact diagnosis, treatment planning, and disease progression monitoring. DIR is also used to capture and quantify biomechanics and dynamics of different anatomical structures including myocardial motion tracking [74, 73, 7], improved monitoring of airflow and pulmonary function in lung imaging [66, 27, 97], and tracking of organ motion in radiation therapy [45, 14, 68, 78]. Latest breakthrough advances in imaging techniques like fluorescence and light-sheet microscopy [36, 69, 29, 98], *in-situ* hybridization, and multiplexing [65, 102] have led to image registration being imperative in advancing life sciences research. Relevant research includes a brain-wide mesoscale connectome of the mouse brain [67], uncovering behavior of individual neurons in *C.*

[*]Equal advising

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

 https://doi.org/10.52202/079017-3439

elegans [91], building cellular-level atlases of *C. elegans*, *Drosophila melanogaster*, and the mouse brain [105, 90, 96, 76, 71, 13].

Classical optimization-based and learning-based methods are the two reigning paradigms in DIR. Classical DIR methods are based on solving a variational optimization problem, where a similarity metric is optimized to find the best transformation that aligns the images. Most classical methods are formulated without any particular domain knowledge encoded in the optimization problem, and are therefore general and applicable to a wide range of problems. For instance, the popularly known registration toolkit ANTs [5] has been successfully applied to structural *and* functional neuroimaging data [48, 104, 43], CT lung imaging [66], cardiac motion modeling [53], developmental mouse brain atlases utilizing MRI and light sheet fluorescence microscopy [50] with virtually no change in the optimization algorithm. However, classical iterative methods have slow convergence, their performance is limited by the fidelity of image intensities, and they cannot incorporate learning to leverage a training set containing weak supervision such as anatomical landmarks, label maps or expert annotations. Deep Learning for Image Registration (DLIR) is an interesting paradigm to overcome these challenges. DLIR methods take a pair of images as input to a neural network and outputs a warp field that aligns the images, and their associated anatomical landmarks. The neural network parameters are trained to minimize the alignment loss over image pairs and landmarks in a training set. During inference, an image pair is provided and the network regresses a warp field. A primary benefit of this method is the ability to incorporate weak supervision like anatomical landmarks or expert annotations during training, which performs better landmark alignment without access to landmarks at inference time.

Motivation However, the benefits of using DLIR methods over classical DIR methods in terms of accuracy or robustness to domain shift are still topics with no clear consensus. Several DLIR methods claim that architectural choices and loss function design combined with amortized optimization of neural network parameters significantly outperform classical methods [63, 61, 17]. On the contrary, classical iterative methods that leverage implicit or explicit conventional priors have shown to outperform most deep learning methods on other challenging datasets [100, 79]. For example, in the context of lung registration, an implicit neural optimization method surpasses every deep learning baseline on the DIR-lab dataset [100]. In EMPIRE10 challenge without access to labeled data [66], classical methods are highly performant compared to deep learning methods. In the ANHIR histology registration challenge [12], the best performing algorithms were classical methods, and the deep learning method was fast and performed well, but did not have good generalization capabilities. Mok *et al.* [62] also mention that deep baselines typically fail 'spectacularly' on out of distribution data, and classical methods like Elastix and ANTs come out on top. However, these observations are relatively unstructured and not studied directly. The confounding variable of using labelmap supervision has urged the Learn2Reg 2024 LUMIR challenge [35] to be performed on fully unsupervised data. In our own empirical evaluations, we found that classical methods typically outperform deep methods under certain conditions and assumptions. Image registration is *NP-hard* being a non-convex optimization problem, and approximating the solution of NP-hard problems with deep learning methods is not guaranteed to be optimal, or even a minima of the registration loss at test-time. Deep learning methods also claim to provide amortized optimization since classical methods are extremely slow to run, however, modern GPU implementations [55, 59, 41] have patched this shortcoming of classical methods while providing state-of-the-art performance.

Contributions. The conditions needed for either paradigm to perform well over the other are clouded and not explicitly outlined in the existing literature. This has prolonged the tug-of-war between classical and deep learning methods. We perform a more structured problem setup and empirical evaluation to determine consensus on the benefits and limitations of each paradigm. First, we observe a strong correlation between the mutual information between per-pixel intensity and label maps, and the performance of classical registration methods. This strong correlation hints to the fact that the Jacobian projection in DLIR methods is unlikely to affect this correlation, and therefore, the performance of DLIR methods in the unsupervised setting. We empirically verify this hypothesis on a variety of state-of-the-art classical and DLIR methods, and address instrumentation bias in the existing literature. Secondly, since the label map is a deterministic function of the intensity image, DLIR methods can learn to perform better label matching when this constraint is enforced during training, by implicitly discovering the label map within the network features and predicting a warp field that minimizes the alignment error between label maps. This is a key strength of DLIR methods, that classical methods cannot leverage. Third, we show that even though learning methods implicit capture semantic information from the image which is not explicitly captured by classical methods,

this additional feature learning does not translate to invariance to domain shift, and DLIR methods are brittle to these changes. These empirical findings allow us to reassess and recalibrate performance expectations from classical and DLIR methods, using a systematic, unbiased and fair evaluation.

2 Related Work

2.1 Classical Optimization-Based Methods

Classical image registration algorithms employ iterative optimization on a variational objective to estimate the dense displacement field between two images. Some of the earliest approaches to deformable registration considered models for small deformations using elastic deformation assumptions [51, 23, 8, 31, 30, 20, 21], conceptualizing the moving image volume as an elastic continuum that undergoes deformation to align with the appearance of the fixed image. This was in conjunction with alternate formulations based on fluid-dynamical Navier-Stokes [22, 21] and Euler-Lagrange equations [2, 11, 4, 56, 58] and their subsequent optimization strategies. The seminal work of Beg.*et al.* [11] introduces an explicit Euler-Langrange formulation and a metric distance on the images as measured by the geodesic shortest paths in the space of diffeomorphisms used to transform the moving image to the fixed image. However, storing the explicit velocity fields is expensive in terms of compute and memory. This limitation motivated semi-Langrangian formulations [4, 3] to avoid storing velocity fields explicitly, and only storing the final diffeomorphism. ANTs [5, 1] is a widely used toolkit that employs the Euler-Langrange formulation with a symmetric objective function [2]. Yet another approach is to interpret deformable registration as an optical flow problem [70, 103], leading to the famous Demons algorithm and its diffeomorphic and symmetric variants [106, 93, 92, 95] implemented as part of the Insight Toolkit (ITK) [40, 25]. However, most of these methods are still computationally expensive to run owing to their CPU implementations. Recently, modern implementations leverage the massively parallelizable nature of the registration problem to run on GPUs, leading to orders of magnitude of speedups while retaining the robustness and accuracy of the classical methods [55, 59, 41]. However, as we show in Section 4, the registration performance of classical methods is limited by the fidelity of image intensities.

2.2 Deep Learning for Image Registration

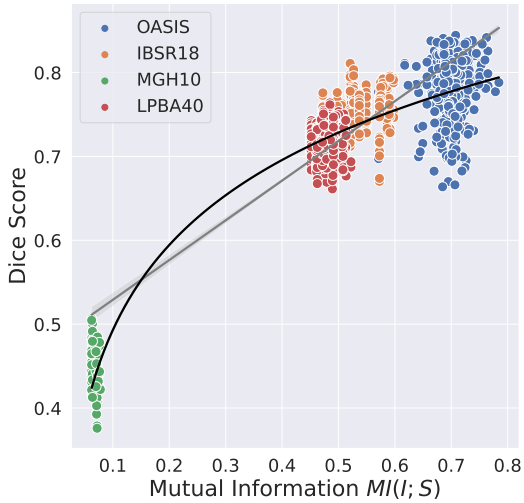

Figure 1: **Correlation between Dice Score and Mutual Information.** Classical registration methods like ANTs show a strong correlation between the Dice Score of registered pairs, and the mutual information between the corresponding image and label across 4 brain datasets.

In contrast to most classical methods, earliest Deep Learning for Image Registration (DLIR) methods employed supervised learning for registration tasks [15, 49, 77, 80] where the deformation field is obtained either manually or from a classical method. Voxelmorph [9] was one of the first approaches that introduced unsupervised learning for registration of in-vivo brain MRI images. Subsequent research expanded upon this paradigm, exploring diverse architectural designs [18, 52, 42, 62], loss functions [109, 108, 44, 24, 60, 107, 75, 16], and formulations based on incorporating inverse-consistency or symmetric transforms [61, 46, 47, 83, 109]. However, hyperparameter tuning became a challenge for DLIR methods since the methods had to be retrained for every new value of the regularization parameter. This motivated techniques such as conditional hyperparameter injection which addressed hyperparameter tuning [64, 38], while domain randomization and fine-tuning [37, 88, 72, 28] aimed to addressed generalizability of DLIR methods across domains. Recently, pretrained or foundation models are also proposed to address the generalizability of DLIR methods across different imaging and anatomy [54, 84]. However, these methods perform a monolithic prediction of the warp field from the input images, losing feedback from the intermediate stages of the registration process as done in classical methods. To refine the warp fields, recurrent or cascade-based archi-

https://doi.org/10.52202/079017-3439

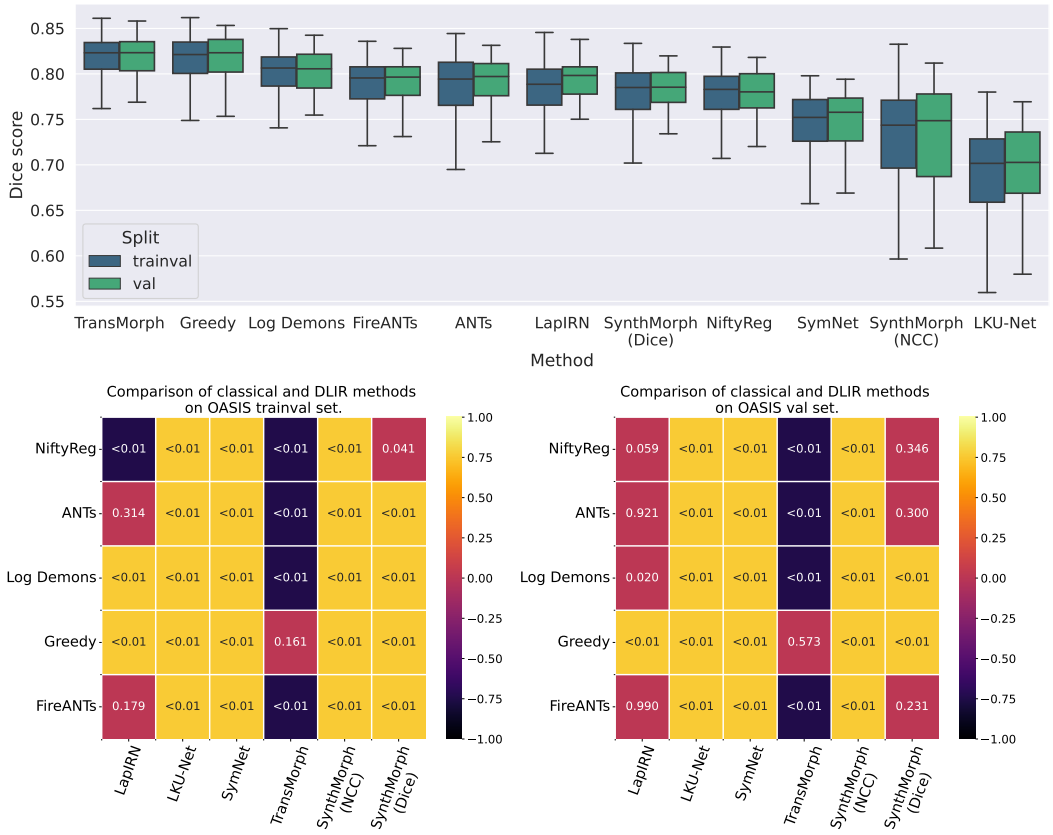

Figure 2: **Performance of classical and unsupervised DLIR methods on OASIS data.** Boxplots (**top**) show that classical methods on average are ranked higher than DLIR methods, both on the *trainval* and *val* splits. Interestingly, the performance of unsupervised DLIR methods does not improve on the *trainval* split compared to *val* split – showing that deep learning does not have an intrinsic advantage in label alignment. Tables (**bottom**) of p-values show the results of a pairwise two-sided t-test between the performance of classical and DLIR methods on the *trainval* and *val* splits. denotes a cell where the classical method is significantly better than the DLIR method ($p < 0.01$), a denotes the opposite, denotes no significant difference. Most of the cells are , indicating that classical methods are significantly better than DLIR methods.

tectures were proposed [108, 109, 107, 16]. However, cascade-based methods create a substantial memory overhead due to backpropagation through cascades and storage of intermediate volumes [6]. Another promising avenue is to leverage deep implicit priors [87] within optimization frameworks to improve the performance of optimization methods or incorporate implicit constraints of the optimized warp field [101, 99, 44, 39]. We refer the reader to [26, 33] for a comprehensive review of image registration techniques.

Despite the plethora of architectural formulations, loss functions, and output representations proposed in Deep Learning for Image Registration methods, we identify that these methods are highly sensitive to the domain gap between the distributions of training and test data, and in the unsupervised case, do not provide any benefit in terms of performance over classical methods. Their primary benefit is their ability to incorporate weak supervision like anatomical landmarks or expert annotations during training, which performs better landmark alignment on unseen image pairs (from the same distribution) without access to landmarks at inference time.

3 Preliminaries

We rehash the image registration problem statement to unify both classical and deep learning methods. Consider a dataset of image pairs $\mathcal{D} = \{(I_f^{(n)}, I_m^{(n)}) \mid n \in \mathbb{N}, 1 \leq n \leq N\}$, where $I_f^{(n)}$ and $I_m^{(n)}$ are the fixed and moving images defined over a spatial domain $\Omega \in \mathbb{R}^d$. We drop the superscript

n for simplicity. Also consider segmentation maps S_f and S_m for the fixed and moving images, respectively, defined over Ω. Given a family of transformations $T(\Omega)$, the goal of image registration is to estimate transformations $\varphi_\theta(f,m) \in T(\Omega)$ parameterized by θ that minimize the following objective:

$$\arg\min_\theta \sum_{f,m} \mathcal{L}(I_f, I_m \circ \varphi_\theta(f,m)) + \mathcal{R}(\varphi_\theta(f,m)) \tag{1}$$

where $\mathcal{L}$ is a dissimilarity function such as mean squared error, or negative local cross correlation, and $\mathcal{R}$ is a regularization term that encourages desirable properties of the transformation, such as smoothness or elasticity. We call Eq. (1) the *image matching* objective, since the transformations only need to align the intensity images. We can also call this the *unsupervised* objective, since it does not require any labeled data. If a suitably chosen label alignment loss $\mathcal{D}$ is added as well, the optimization problem becomes:

$$\arg\min_\theta \sum_{f,m} \mathcal{L}(I_f, I_m \circ \varphi_\theta(f,m)) + \mathcal{D}(S_f, S_m \circ \varphi_\theta(f,m)) + \mathcal{R}(\varphi_\theta(f,m)) \tag{2}$$

We call Eq. (2) the *label matching* objective, or a *weakly-supervised* objective. The image matching objective can subsume both DLIR and classical methods by choosing

$$\varphi_\theta(f,m) = \begin{cases} f_\theta(I_f, I_m), & \text{for deep networks,} \\ \varphi_{(f,m)}, & \text{for classical methods.} \end{cases} \tag{3}$$

where f_θ is a deep network parameterized by θ and $\varphi_{(f,m)}$ are optimizable free parameters that are indexed by the 2-tuple (f,m), i.e. $\theta = \bigcup_{f,m}\{\varphi_{(f,m)}\}$. In this paper, we consider methods that solve Eq. (1) using gradient-based methods. The gradient of Eq. (1) with respect to θ is given by (we remove the $\mathcal{R}$ term for simplicity):

$$\frac{\partial \mathcal{L}}{\partial \theta} = \sum_{f,m} \frac{\partial \mathcal{L}}{\partial \varphi_\theta(f,m)} \frac{\partial \varphi_\theta(f,m)}{\partial \theta} \tag{4}$$

The first term $\frac{\partial \mathcal{L}}{\partial \varphi_\theta(f,m)}$ is the training signal from the dissimilarity function which does not depend on the parameters θ for a given value of $\varphi_\theta(f,m)$ and choice of $\mathcal{L}$. The second term $\frac{\partial \varphi_\theta(f,m)}{\partial \theta}$ is the Jacobian of the transformation with respect to the parameters, which is a projection of the gradient from the space of warp fields to the space of arbitrary parameters. For classical methods, the Jacobian is the identity matrix, for deep networks it is determined by the functional relationship of the output with respect to network parameters. Therefore, the difference in training dynamics and overall performance gap between classical and deep learning methods is likely to be attributed to the choice of $\frac{\partial \varphi_\theta(f,m)}{\partial \theta}$.

4 Unsupervised DLIR does not improve label matching performance

A speculated claim of deep learning methods is that they can provide better label matching performance by simply training a network to minimize Eq. (1) in an unsupervised setting. Such improvements are claimed to come from architectural designs, which correspond to choice of Jacobian $\frac{\partial \varphi_\theta(f,m)}{\partial \theta}$. A variety of architectures and parameterizations [17, 63, 64, 62, 34, 82, 101] have been proposed to this effect. **However, we show that this is not the case.**

Image matching objectives ensure that intensities from the moving image are displaced to locations in the fixed image where they are most similar, without regard for alignment for any higher order structures. Intuitively, this will ensure label matching only to the extent that the intensity is predictive of the label. If an intensity value strongly corresponds to a particular label, then image matching will lead to label matching. Similarly, if a given intensity value corresponds to multiple possible labels, then image matching does not tell us which labels are matched via the image matching objective. More formally, considering the per-pixel intensity i and labels s as random variables, one can compute the mutual information between the intensity and label maps, denoted as $MI(i;s)$ to determine the predictability of one from the other. We now show that the label matching performance of classical methods is highly correlated with $MI(i;s)$. We consider a widely used classical method, ANTs [2, 5], to eliminate the effect of any Jacobian term. We consider four brain datasets - OASIS, LPBA40, MGH10, and IBSR18, which are acquired under different scanners, under different resolutions, and have different preprocessing, labelling and postprocessing protocols [57, 48]. For

 https://doi.org/10.52202/079017-3439

each dataset, we use ANTs for registering all pairs within the dataset and then evaluate the Dice score as an indicator of label matching performance. For each image I and its corresponding label map S, we compute the probability maps $p(i), p(s), p(i, s)$ using histogram binning, followed by the mutual information $MI(i; s) = H(s) - H(s|i)$. A Pearson's correlation coefficient between the Dice scores and the mutual information of the image and label (Fig. 1) reveals a strong linear ($\mathbf{r = 0.886}$) and logarithmic ($\mathbf{r = 0.933}$) relationship between the two quantities, shown by the gray and black lines respectively. Image matching improves label matching performance *only to the extent of the information about the label obtained from the image* (i.e. $MI(i; s)$). At a first glance, the Jacobian term $\frac{\partial \varphi_\theta(f,m)}{\partial \theta}$ seemingly does not have a role in improving this mutual information further.

Empirical Validation. We verify this claim empirically on the OASIS dataset, by minimizing Eq. (1) in both DLIR and classical methods. We split the OASIS dataset into a training set of 364 images and a validation set of 50 images. We choose 50 instead of 20 images as in the original split [35] to compute statistical significance. Dice score over 35 subcortical structures is used as the label matching metric. We choose SynthMorph [37], LapIRN [63], SymNet [61], LKU-Net [42] and TransMorph [19] as state-of-the-art DLIR baselines and ANTs [5], NiftyReg [59], Symmetric Log Demons [94], Greedy [106], FireANTs [41] as state-of-the-art classical baselines. For all DLIR methods, we use pretrained models if they are trained with Eq. (1), or train them with the architecture and hyperparameters provided in their original source code. The only exception is SynthMorph, which is trained on synthetically generated data and Dice loss of its corresponding synthetic labels (`shapes-sm` model). To compare SynthMorph's domain generalization capabilities with only the image matching objective, we add another model, dubbed '`shapes-sm-ncc`' that is trained on synthetically generated data as in the original pretrained model, but with the normalized cross-correlation of the aligned synthetic images. For all classical methods, we follow their recommended hyperparameters and run till convergence. All experiments are run on a cluster with 2 AMD EPYC 7713 CPUs and 8 NVIDIA A6000 GPUs.

Results. For all methods, we compute the Dice score of all 35 subcortical regions on images in the validation set (denoted as *val*), and all images (denoted as *trainval*). These Dice scores are sorted by median validation performance in Fig. 2(top). Moreover, we perform a two-sided t-test for each *(classical, DLIR)* pair, both on the trainval and validation sets, shown in Fig. 2(bottom). Fig. 2 shows the following conclusions: (a) the top performing classical method (Greedy) and the top performing DLIR method (TransMorph) achieve similar label matching performance on the val and the trainval set, i.e. the differences are *not* statistically significant ($p = 0.161$), (b) classical methods almost always perform better than DLIR methods, even on the training set showing that the Jacobian term does not improve label matching more than the mutual information between the image and label, and (c) for unsupervised DLIR methods, there is no improvement label matching performance in the training set compared to val set. The only role of the Jacobian term is to perform amortized learning, but without supervised objectives, this does not guarantee any additional boost in label matching.

The effect of instrumentation bias. The astute reader may observe that this result is in contrast to results shown in prior literature [61, 63, 101, 19, 10]. We note that this is due to instrumentation bias [86], where the baselines' performance may be misrepresented due to changes in hyperparameters, early stopping, or different preprocessing protocols. For instance, [10] mention that the default parameters of ANTs are not optimal, and choose a very different set of parameters (a Gaussian smoothing of 9 pixels, followed by an extremely small 0.4 pixels at the next scale). By stark contrast, we found the recommended parameters to work extremely well for all datasets considered in this paper. We speculate that these changes are done to tradeoff accuracy for speed, since classical methods converge slowly. However, this leads to misrepresentation of the performance of classical baselines. We found much better results (Fig. 2) for classical baselines simply by using their recommended scripts. We compare the discrepancy in performance between the baselines reported in the literature and the ones we obtained in Fig. 3. We follow the guidelines in [86] to evaluate all methods. To ensure our work does not introduce its own instrumentation bias for DLIR baselines, we compare the performance of our trained/pretrained models to the ones reported in the literature (Fig. 3). We make all evaluation scripts and trained models public[2] to encourage fairness and transparency in evaluations.

[2]https://github.com/rohitrango/Magic-or-Mirage/

Evaluation of classical methods reported by baselines					
Method	**Evaluated Baseline**	**Statistic**	**Reported value**	**Our eval**	**Difference**
SymNet	ANTs	Mean	0.680	0.787	0.107
PIRATE	ANTs	Mean	0.699	0.787	0.088
LapIRN	Demons	Mean	0.715	0.802	0.087
LapIRN	ANTs	Mean	0.723	0.787	0.064
NODEO	Demons	Mean	0.764	0.802	0.038
NODEO	ANTs	Mean	0.729	0.787	0.058
Voxelmorph	ANTs	Mean	0.749	0.787	0.038
Voxelmorph	NiftyReg	Mean	0.755	0.776	0.021
SynthMorph	ANTs	Median	0.770	0.797	0.027
Evaluation of DLIR baselines reported by us					
Method	**Dice supervision**	**Statistic**	**Reported value**	**Our eval**	**Difference**
SynthMorph	-	Median	0.780	0.785	0.005
TransMorph-Regular	✓	Mean	0.858	0.855	-0.003
LKU-Net	✓	Mean	0.886	0.904	0.018
LapIRN	✗	Mean	0.808	0.788	-0.020
SymNet	✗	Mean	0.743	0.748	0.005

Figure 3: **Instrumentation bias in evaluation of image registration algorithms.** We highlight a significant difference in evaluation metrics reported by baselines and our evaluation on the OASIS validation dataset. This difference can be attributed to deviation in hyperparameters from the recommended parameters or early stopping to save time. In either case, this misrepresentation leads to incorrect conclusions about the performance of the algorithm. The reported dice scores are anywhere from 2 to 10 Dice points lower than our evaluation, showing a non-trivial instrumentation bias. We report our own evaluation of DLIR algorithms and compare them with reported values to avoid introducing instrumentation bias in our evaluation.

5 Supervised DLIR methods demonstrate enhanced label matching

When label matching is introduced as an objective in Eq. (2), DLIR methods show superior performance than classical methods. Unlike the previous discussion, where only a pixelwise definition of $MI(i; s)$ was used to quantify the coaction of image intensities and label maps, we consider the entire image I and label volume S as high-dimensional random variables. Label maps are now a deterministic function of the image, i.e. $S = f(I)$, where f is the labelling protocol. In addition to image intensity, label maps are a function of morphological features, location, contrast, and the labelling protocol itself. When trained with the label maps as extra supervision, the network can infer these deterministic relationships to output a warp field that maximize both image similarity and label overlap. Classical intensity-based methods, on the other hand, do not have any mechanism to encode this additional relationship. Aligning intensities or intensity patches discards any functional relationship between high-level image features and labels. To show this, we repeat the same experiment setup as in Section 4 on the same splits, but with the label matching objective added as well.

Results. Fig. 4(top) shows the Dice scores for supervised classical and DLIR methods trained on the OASIS dataset, sorted by median validation performance. In this case, state-of-the-art DLIR methods outperform classical methods by a large margin, with notably higher Dice score on the *trainval* set than the *val* set, due to overfitting to the label matching for the training set. This is unlike unsupervised DLIR, where there was no improvement in label matching performance on the training set, emphasizing the fact that performing amortized training does not improve label matching performance by itself. These differences are statistically significant, with the exception of SymNet, which diverged under many training settings with the Dice loss, and only works marginally better than its unsupervised counterpart. SynthMorph is not trained on real data, and is added only as a reference for domain-agnostic performance.

This is an unsurprising result – the label matching objective provides additional training signal to the registration task, which is a highly ill-posed problem. Classical methods cannot incorporate this additional signal from a training dataset, and learning-based methods exploit this to achieve better registration on unseen data. Classical methods are, however, agnostic to modalties, intensity distributions, voxel resolutions, and anisotropy. The same registration algorithm (with possibly modified parameters) is applied to datasets with different characteristics, and they still retain their state-of-the-art performance. A related question arises for DLIR methods trained with label matching – does label matching performance transfer to other datasets?

 https://doi.org/10.52202/079017-3439

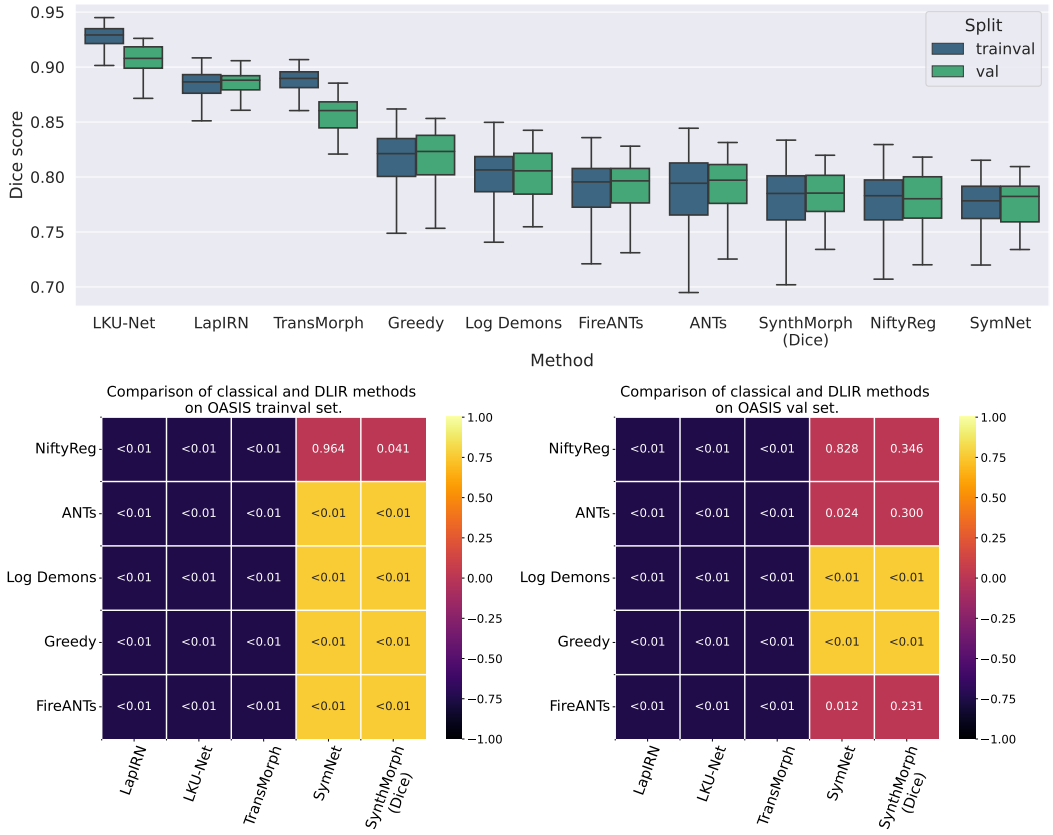

Figure 4: **Performance of classical and supervised DLIR methods on OASIS data.** Boxplots (**top**) show that DLIR methods show superior performance compared to classical methods. Unlike the unsupervised case, the effect of overfitting is clearly visible in the gap between the *trainval* and *val* splits. Tables (**bottom**) of p-values show the results of a pairwise two-sided t-test between the performance of classical and DLIR methods on the *trainval* and *val* splits. denotes a cell where the classical method is significantly better than the DLIR method ($p < 0.01$), a denotes the opposite, denotes no significant difference. State-of-the-art DLIR methods show significantly better performance than classical methods when label supervision is added.

6 DLIR methods do not generalize across datasets

A key strength of classical optimization registration algorithms is their agnostic nature to the image modality, physical resolution, voxel sizes, and preprocessing protocols. Most DLIR methods, on the contrary, have been evaluated extensively on the same distribution of validation datasets as the training data, it is unclear if the performance improvements transfer to other datasets of the same anatomy. To this end, we evaluate the performance of both the classical and DLIR methods on four brain datasets – CUMC12, LPBA40, MGH10, and IBSR18. These datasets represent community-standard brain mapping challenge data [48] for a comprehensive evaluation of 14 nonlinear classical registration methods, across various acquisition, preprocessing and labelling protocols. For all datasets, we follow the preprocessing steps followed by [48].

Each dataset contains a different set of labeled regions acquired manually using different labeling protocols. For each dataset, all previously considered registration algorithms are run on all image pairs, and the mean Dice score over all labeled regions is computed. The methods are then sorted by median validation performance in Fig. 5. For DLIR methods, we plot the performance with models trained with and without the label matching loss in the OASIS dataset, shown as blue and green boxplots respectively. Across all datasets, FireANTs, Greedy, ANTs and NiftyReg consistently perform better than DLIR methods. Among the DLIR methods, SynthMorph performs consistently better due to its domain-agnostic training paradigm. Remarkably, even though DLIR methods outperform classical methods on the OASIS dataset with label matching objective, the performance does not transfer to other datasets, even compared to its own unsupervised variant. This is a negative

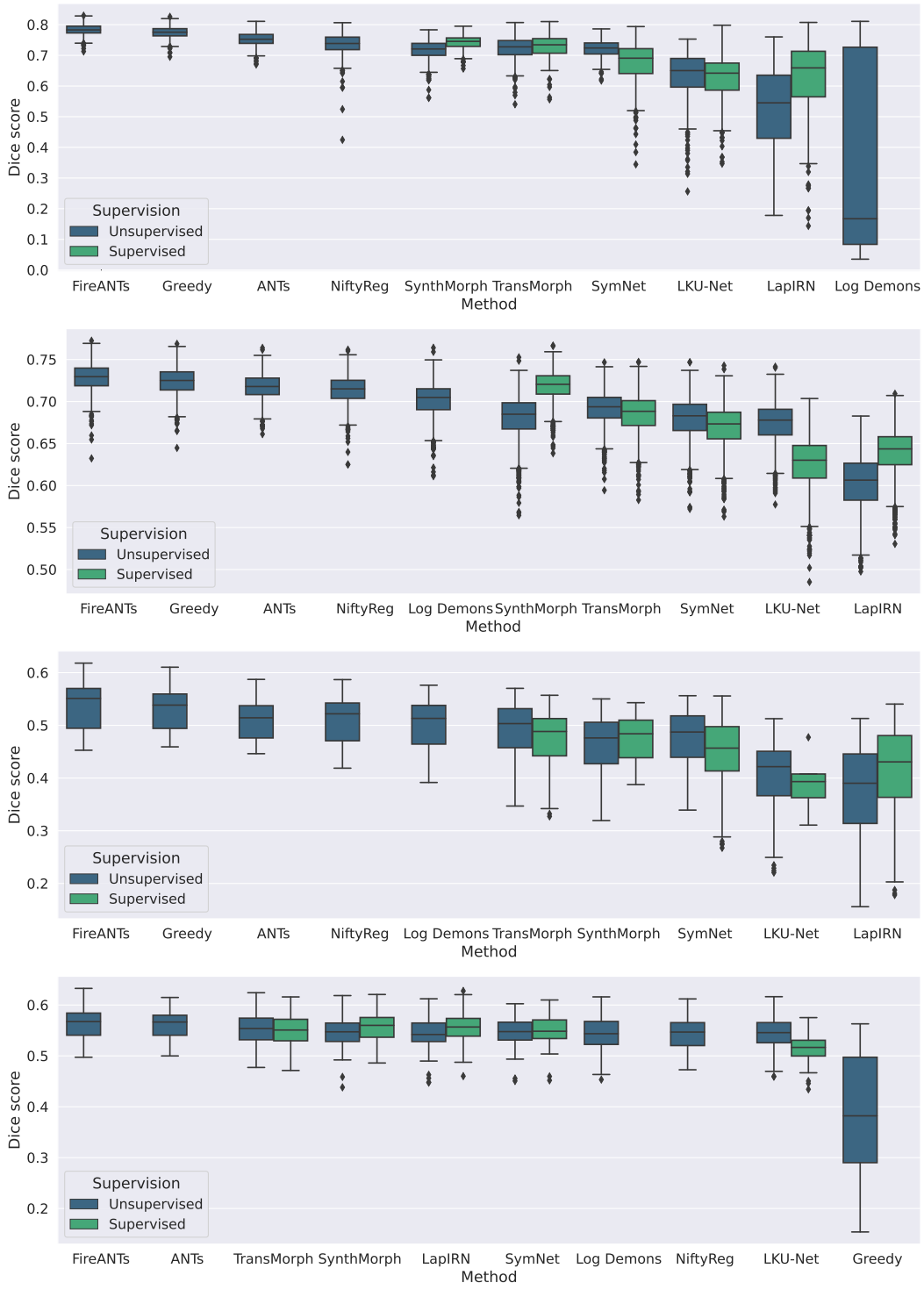

Figure 5: **Classical methods retain robustness across different datasets.** Boxplots show the performance of classical and DLIR methods trained on the OASIS dataset, on four T1-brain datasets. For DLIR methods, we plot the performance of the supervised and unsupervised models. Across all datasets, FireANTs and ANTs consistently outperform DLIR methods, showing robustness to domain shift. Among DLIR methods, SynthMorph and TransMorph show robust performance, and training with label matching objective does not lead to significant improvement.

https://doi.org/10.52202/079017-3439

result – implying that to improve performance on a new dataset, one must collect label maps from that dataset and retrain the model – existing collections of label maps are not sufficient to improve performance on new datasets. Unlike in tasks like segmentation, deep methods do not transfer their performance to out-of-distribution datasets, even with the same resolution, and the expected performance hierarchy does not hold

- **Expected**: Supervised DLIR ID > Supervised DLIR OOD > Classical
- **Observed**: Supervised DLIR ID > Classical > Supervised DLIR OOD

Practitioners should therefore be cautious when using prediction-based DLIR methods, especially when the training data is not representative of the test data, regardless of the presence of label maps.

7 Discussion

This study aims to provide a systematic and unbiased investigation of the performance of classical and DLIR methods under access to label supervision, and their generalization capabilities under small domain shifts. Preceding experiments show that classical methods provide an unprecedented level of robustness and generalizability across datasets, but are limited by the fidelity of the image matching objective. Supervised DLIR methods provide a promising step towards improving registration performance of anatomical regions by implicitly discovering these structures and predicting appropriate warp fields within the network architecture. However, this anatomical-awareness on the training dataset does not help in generalizing to other datasets, limiting the practical utility of these methods. The usability of anatomical landmarks and labelmaps to obtain domain-invariant registration performance still remains an open research problem. These results also have profound implications for annotated data collection and challenges the notion that large labeled datasets ensure robust generalization.

Although our study is performed on inter-subject registration with in-vivo neuroimaging datasets, none of our analysis, baselines, and evaluation make any domain or subject-specific modeling assumptions, and the datasets being community-standard benchmarks, the results are valuable and general, both within the neuroimaging and the biomedical communities at large. At the current state, a practitioner should choose predictive DLIR methods only if they have access to a large labeled dataset, *and* their application is limited to the same dataset distribution. In all other cases, classical optimization-based methods are the more accurate and reliable choice, even if labeled data exists but is not representative of the test data.

7.1 Limitations

Our work performs a comprehensive evaluation of state-of-the-art registration algorithms on a variety of neuroimaging datasets. However, our work does not consider hybrid methods, or representations that use optimal matching criteria based on correlation volumes or sparse correspondence features. Although our work considers large-scale community-standard neuroimaging datasets, the performance of these algorithms may differ on other anatomy or modalities. Our study also considers inter-subject registration only, although no method or evaluation incorporates any subject-specific assumptions. The effects on multimodal registration are not considered in this work. However, our work serves as a foundational step toward a more nuanced discussion on the longstanding technical challenges in image registration, and representations that are effective in mitigating these problems.

Acknowledgements

This work was supported by the National Institutes of Health (NIH) under grants RF1-MH124605, R01-HL133889, R01-EB031722, U24-NS135568.

References

[1] ANTsX. Antsx: Advanced normalization tools (ants). GitHub repository.

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

[3] Brian Avants and James C. Gee. Geodesic estimation for large deformation anatomical shape averaging and interpolation. *NeuroImage*, 23:S139–S150, January 2004.

[4] Brian B. Avants, P. Thomas Schoenemann, and James C. Gee. Lagrangian frame diffeomorphic image registration: Morphometric comparison of human and chimpanzee cortex. *Medical Image Analysis*, 10(3):397–412, June 2006.

[5] Brian B Avants, Nick Tustison, Gang Song, et al. Advanced normalization tools (ants). *Insight j*, 2(365):1–35, 2009.

[6] Shaojie Bai, Zhengyang Geng, Yash Savani, and J. Zico Kolter. Deep Equilibrium Optical Flow Estimation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 610–620, New Orleans, LA, USA, June 2022. IEEE.

[7] Wenjia Bai, Hideaki Suzuki, Jian Huang, Catherine Francis, Shuo Wang, Giacomo Tarroni, Florian Guitton, Nay Aung, Kenneth Fung, Steffen E Petersen, et al. A population-based phenome-wide association study of cardiac and aortic structure and function. *Nature medicine*, 26(10):1654–1662, 2020.

[8] Ruzena Bajcsy, Robert Lieberson, and Martin Reivich. A computerized system for the elastic matching of deformed radiographic images to idealized atlas images. *Journal of computer assisted tomography*, 7(4):618–625, 1983.

[9] Guha Balakrishnan, Amy Zhao, Mert R. Sabuncu, John Guttag, and Adrian V. Dalca. VoxelMorph: A Learning Framework for Deformable Medical Image Registration. *IEEE Transactions on Medical Imaging*, 38(8):1788–1800, August 2019. arXiv:1809.05231 [cs].

[10] Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca. Voxelmorph: a learning framework for deformable medical image registration. *IEEE transactions on medical imaging*, 38(8):1788–1800, 2019.

[11] M Faisal Beg, Michael I Miller, Alain Trouvé, and Laurent Younes. Computing large deformation metric mappings via geodesic flows of diffeomorphisms. *International journal of computer vision*, 61:139–157, 2005.

[12] Jiří Borovec, Jan Kybic, Ignacio Arganda-Carreras, Dmitry V Sorokin, Gloria Bueno, Alexander V Khvostikov, Spyridon Bakas, I Eric, Chao Chang, Stefan Heldmann, et al. Anhir: automatic non-rigid histological image registration challenge. *IEEE transactions on medical imaging*, 39(10):3042–3052, 2020.

[13] Bella E Brezovec, Andrew B Berger, Yukun A Hao, Feng Chen, Shaul Druckmann, and Thomas R Clandinin. Mapping the neural dynamics of locomotion across the drosophila brain. *Current Biology*, 34(4):710–726, 2024.

[14] Kristy K Brock, Sasa Mutic, Todd R McNutt, Hua Li, and Marc L Kessler. Use of image registration and fusion algorithms and techniques in radiotherapy: Report of the aapm radiation therapy committee task group no. 132. *Medical physics*, 44(7):e43–e76, 2017.

[15] Xiaohuan Cao, Jianhua Yang, Jun Zhang, Dong Nie, Minjeong Kim, Qian Wang, and Dinggang Shen. Deformable image registration based on similarity-steered cnn regression. In *Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part I 20*, pages 300–308. Springer, 2017.

[16] Junyu Chen, Eric C Frey, and Yong Du. Unsupervised learning of diffeomorphic image registration via transmorph. In *International Workshop on Biomedical Image Registration*, pages 96–102. Springer, 2022.

 https://doi.org/10.52202/079017-3439

[17] Junyu Chen, Eric C Frey, Yufan He, William P Segars, Ye Li, and Yong Du. Transmorph: Transformer for unsupervised medical image registration. *Medical image analysis*, 82:102615, 2022.

[18] Junyu Chen, Eric C. Frey, Yufan He, William P. Segars, Ye Li, and Yong Du. TransMorph: Transformer for unsupervised medical image registration. *Medical Image Analysis*, 82:102615, November 2022.

[19] Junyu Chen, Eric C. Frey, Yufan He, William P. Segars, Ye Li, and Yong Du. TransMorph: Transformer for unsupervised medical image registration. *Medical Image Analysis*, 82:102615, November 2022. arXiv:2111.10480 [cs, eess].

[20] Gary E Christensen and Hans J Johnson. Consistent image registration. *IEEE transactions on medical imaging*, 20(7):568–582, 2001.

[21] Gary E Christensen, Richard D Rabbitt, and Michael I Miller. Deformable templates using large deformation kinematics. *IEEE transactions on image processing*, 5(10):1435–1447, 1996.

[22] G.E. Christensen, S.C. Joshi, and M.I. Miller. Volumetric transformation of brain anatomy. *IEEE Transactions on Medical Imaging*, 16(6):864–877, December 1997. Conference Name: IEEE Transactions on Medical Imaging.

[23] Christos Davatzikos. Spatial transformation and registration of brain images using elastically deformable models. *Computer Vision and Image Understanding*, 66(2):207–222, 1997.

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

[25] Florence Dru, Pierre Fillard, and Tom Vercauteren. An ITK Implementation of the Symmetric Log-Domain Diffeomorphic Demons Algorithm. *The Insight Journal*, September 2010.

[26] Yabo Fu, Yang Lei, Tonghe Wang, Walter J Curran, Tian Liu, and Xiaofeng Yang. Deep learning in medical image registration: a review. *Physics in Medicine & Biology*, 65(20):20TR01, October 2020.

[27] Yabo Fu, Yang Lei, Tonghe Wang, Kristin Higgins, Jeffrey D. Bradley, Walter J. Curran, Tian Liu, and Xiaofeng Yang. LungRegNet: an unsupervised deformable image registration method for 4D-CT lung. *Medical physics*, 47(4):1763–1774, April 2020.

[28] Yabo Fu, Yang Lei, Jun Zhou, Tonghe Wang, S Yu David, Jonathan J Beitler, Walter J Curran, Tian Liu, and Xiaofeng Yang. Synthetic ct-aided mri-ct image registration for head and neck radiotherapy. In *Medical Imaging 2020: Biomedical Applications in Molecular, Structural, and Functional Imaging*, volume 11317, pages 572–578. SPIE, 2020.

[29] Davide Gambarotto, Fabian U Zwettler, Maeva Le Guennec, Marketa Schmidt-Cernohorska, Denis Fortun, Susanne Borgers, Jörn Heine, Jan-Gero Schloetel, Matthias Reuss, Michael Unser, et al. Imaging cellular ultrastructures using expansion microscopy (u-exm). *Nature methods*, 16(1):71–74, 2019.

[30] James C Gee and Ruzena K Bajcsy. Elastic matching: Continuum mechanical and probabilistic analysis. *Brain warping*, 2:183–197, 1998.

[31] James C Gee, Martin Reivich, and Ruzena Bajcsy. Elastically deforming a three-dimensional atlas to match anatomical brain images. 1993.

[32] Maged Goubran, Cathie Crukley, Sandrine De Ribaupierre, Terence M Peters, and Ali R Khan. Image registration of ex-vivo mri to sparsely sectioned histology of hippocampal and neocortical temporal lobe specimens. *Neuroimage*, 83:770–781, 2013.

[33] Grant Haskins, Uwe Kruger, and Pingkun Yan. Deep learning in medical image registration: a survey. *Machine Vision and Applications*, 31(1):8, January 2020.

[34] Mattias P. Heinrich, Heinz Handels, and Ivor J. A. Simpson. Estimating Large Lung Motion in COPD Patients by Symmetric Regularised Correspondence Fields. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Lecture Notes in Computer Science, pages 338–345, Cham, 2015. Springer International Publishing.

[35] Alessa Hering, Lasse Hansen, Tony CW Mok, Albert CS Chung, Hanna Siebert, Stephanie Häger, Annkristin Lange, Sven Kuckertz, Stefan Heldmann, Wei Shao, et al. Learn2reg: comprehensive multi-task medical image registration challenge, dataset and evaluation in the era of deep learning. *IEEE Transactions on Medical Imaging*, 42(3):697–712, 2022.

[36] Elizabeth MC Hillman, Venkatakaushik Voleti, Wenze Li, and Hang Yu. Light-sheet microscopy in neuroscience. *Annual review of neuroscience*, 42:295–313, 2019.

[37] Malte Hoffmann, Benjamin Billot, Douglas N Greve, Juan Eugenio Iglesias, Bruce Fischl, and Adrian V Dalca. Synthmorph: learning contrast-invariant registration without acquired images. *IEEE transactions on medical imaging*, 41(3):543–558, 2021.

[38] Andrew Hoopes, Malte Hoffmann, Bruce Fischl, John Guttag, and Adrian V Dalca. Hypermorph: Amortized hyperparameter learning for image registration. In *Information Processing in Medical Imaging: 27th International Conference, IPMI 2021, Virtual Event, June 28–June 30, 2021, Proceedings 27*, pages 3–17. Springer, 2021.

[39] Junhao Hu, Weijie Gan, Zhixin Sun, Hongyu An, and Ulugbek S. Kamilov. A Plug-and-Play Image Registration Network, March 2024. arXiv:2310.04297 [eess].

[40] Luis Ibanez, Will Schroeder, Lydia Ng, Josh Cates, et al. The itk software guide, 2003.

[41] Rohit Jena, Pratik Chaudhari, and James C Gee. Fireants: Adaptive riemannian optimization for multi-scale diffeomorphic registration. *arXiv preprint arXiv:2404.01249*, 2024.

[42] Xi Jia, Joseph Bartlett, Tianyang Zhang, Wenqi Lu, Zhaowen Qiu, and Jinming Duan. U-net vs transformer: Is u-net outdated in medical image registration? *arXiv preprint arXiv:2208.04939*, 2022.

[43] Di Jiang, Yuhui Du, Hewei Cheng, Tianzi Jiang, and Yong Fan. Groupwise spatial normalization of fmri data based on multi-range functional connectivity patterns. *Neuroimage*, 82:355–372, 2013.

[44] Ankita Joshi and Yi Hong. Diffeomorphic Image Registration using Lipschitz Continuous Residual Networks. page 13.

[45] Marc L Kessler. Image registration and data fusion in radiation therapy. *The British journal of radiology*, 79(special_issue_1):S99–S108, 2006.

[46] Boah Kim, Dong Hwan Kim, Seong Ho Park, Jieun Kim, June-Goo Lee, and Jong Chul Ye. Cyclemorph: cycle consistent unsupervised deformable image registration. *Medical image analysis*, 71:102036, 2021.

[47] Boah Kim, Jieun Kim, June-Goo Lee, Dong Hwan Kim, Seong Ho Park, and Jong Chul Ye. Unsupervised deformable image registration using cycle-consistent cnn. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22*, pages 166–174. Springer, 2019.

[48] Arno Klein, Jesper Andersson, Babak A. Ardekani, John Ashburner, Brian Avants, Ming-Chang Chiang, Gary E. Christensen, D. Louis Collins, James Gee, Pierre Hellier, Joo Hyun Song, Mark Jenkinson, Claude Lepage, Daniel Rueckert, Paul Thompson, Tom Vercauteren, Roger P. Woods, J. John Mann, and Ramin V. Parsey. Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. *NeuroImage*, 46(3):786–802, July 2009.

[49] Julian Krebs, Tommaso Mansi, Hervé Delingette, Li Zhang, Florin C Ghesu, Shun Miao, Andreas K Maier, Nicholas Ayache, Rui Liao, and Ali Kamen. Robust non-rigid registration through agent-based action learning. In *Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part I 20*, pages 344–352. Springer, 2017.

 https://doi.org/10.52202/079017-3439

[50] Fae A Kronman, Josephine K Liwang, Rebecca Betty, Daniel J Vanselow, Yuan-Ting Wu, Nicholas J Tustison, Ashwin Bhandiwad, Steffy B Manjila, Jennifer A Minteer, Donghui Shin, et al. Developmental mouse brain common coordinate framework. *bioRxiv*, 2023.

[51] Jan Kybic and Michael Unser. Fast parametric elastic image registration. *IEEE transactions on image processing*, 12(11):1427–1442, 2003.

[52] Leo Lebrat, Rodrigo Santa Cruz, Frederic de Gournay, Darren Fu, Pierrick Bourgeat, Jurgen Fripp, Clinton Fookes, and Olivier Salvado. CorticalFlow: A Diffeomorphic Mesh Transformer Network for Cortical Surface Reconstruction. In *Advances in Neural Information Processing Systems*, volume 34, pages 29491–29505. Curran Associates, Inc., 2021.

[53] Devavrat Likhite, Ganesh Adluru, and Edward DiBella. Deformable and rigid model-based image registration for quantitative cardiac perfusion. In *Statistical Atlases and Computational Models of the Heart-Imaging and Modelling Challenges: 5th International Workshop, STACOM 2014, Held in Conjunction with MICCAI 2014, Boston, MA, USA, September 18, 2014, Revised Selected Papers 5*, pages 41–50. Springer, 2015.

[54] Fengze Liu, Ke Yan, Adam P. Harrison, Dazhou Guo, Le Lu, Alan L. Yuille, Lingyun Huang, Guotong Xie, Jing Xiao, Xianghua Ye, and Dakai Jin. SAME: Deformable Image Registration Based on Self-supervised Anatomical Embeddings. In Marleen de Bruijne, Philippe C. Cattin, Stéphane Cotin, Nicolas Padoy, Stefanie Speidel, Yefeng Zheng, and Caroline Essert, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, Lecture Notes in Computer Science, pages 87–97, Cham, 2021. Springer International Publishing.

[55] Andreas Mang, Amir Gholami, Christos Davatzikos, and George Biros. CLAIRE: A distributed-memory solver for constrained large deformation diffeomorphic image registration. *SIAM Journal on Scientific Computing*, 41(5):C548–C584, January 2019. arXiv:1808.04487 [cs, math].

[56] Andreas Mang and Lars Ruthotto. A lagrangian gauss–newton–krylov solver for mass-and intensity-preserving diffeomorphic image registration. *SIAM Journal on Scientific Computing*, 39(5):B860–B885, 2017.

[57] Daniel S Marcus, Tracy H Wang, Jamie Parker, John G Csernansky, John C Morris, and Randy L Buckner. Open access series of imaging studies (oasis): cross-sectional mri data in young, middle aged, nondemented, and demented older adults. *Journal of cognitive neuroscience*, 19(9):1498–1507, 2007.

[58] Michael I. Miller, Alain Trouvé, and Laurent Younes. On the Metrics and Euler-Lagrange Equations of Computational Anatomy. *Annual Review of Biomedical Engineering*, 4(1):375–405, 2002. _eprint: https://doi.org/10.1146/annurev.bioeng.4.092101.125733.

[59] Marc Modat, Gerard R Ridgway, Zeike A Taylor, Manja Lehmann, Josephine Barnes, David J Hawkes, Nick C Fox, and Sébastien Ourselin. Fast free-form deformation using graphics processing units. *Computer methods and programs in biomedicine*, 98(3):278–284, 2010.

[60] Tony C. W. Mok and Albert C. S. Chung. Large Deformation Diffeomorphic Image Registration with Laplacian Pyramid Networks, June 2020. arXiv:2006.16148 [cs, eess].

[61] Tony CW Mok and Albert Chung. Fast symmetric diffeomorphic image registration with convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4644–4653, 2020.

[62] Tony CW Mok and Albert Chung. Affine medical image registration with coarse-to-fine vision transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20835–20844, 2022.

[63] Tony CW Mok and Albert CS Chung. Large deformation diffeomorphic image registration with laplacian pyramid networks. pages 211–221, 2020.

[64] Tony CW Mok and Albert CS Chung. Conditional deformable image registration with convolutional neural network. pages 35–45, 2021.

[65] Annette Moter and Ulf B Göbel. Fluorescence in situ hybridization (fish) for direct visualization of microorganisms. *Journal of microbiological methods*, 41(2):85–112, 2000.

[66] Keelin Murphy, Bram Van Ginneken, Joseph M Reinhardt, Sven Kabus, Kai Ding, Xiang Deng, Kunlin Cao, Kaifang Du, Gary E Christensen, Vincent Garcia, et al. Evaluation of registration methods on thoracic ct: the empire10 challenge. *IEEE transactions on medical imaging*, 30(11):1901–1920, 2011.

[67] Seung Wook Oh, Julie A Harris, Lydia Ng, Brent Winslow, Nicholas Cain, Stefan Mihalas, Quanxin Wang, Chris Lau, Leonard Kuan, Alex M Henry, et al. A mesoscale connectome of the mouse brain. *Nature*, 508(7495):207–214, 2014.

[68] Seungjong Oh and Siyong Kim. Deformable image registration in radiation therapy. *Radiation oncology journal*, 35(2):101, 2017.

[69] Omar E Olarte, Jordi Andilla, Emilio J Gualda, and Pablo Loza-Alvarez. Light-sheet microscopy: a tutorial. *Advances in Optics and Photonics*, 10(1):111–179, 2018.

[70] Karsten Østergaard Noe, Baudouin Denis De Senneville, Ulrik Vindelev Elstrøm, Kari Tanderup, and Thomas Sangild Sørensen. Acceleration and validation of optical flow based deformable registration for image-guided radiotherapy. *Acta Oncologica*, 47(7):1286–1293, 2008.

[71] Hanchuan Peng, Phuong Chung, Fuhui Long, Lei Qu, Arnim Jenett, Andrew M Seeds, Eugene W Myers, and Julie H Simpson. Brainaligner: 3d registration atlases of drosophila brains. *Nature methods*, 8(6):493–498, 2011.

[72] Javier Pérez de Frutos, André Pedersen, Egidijus Pelanis, David Bouget, Shanmugapriya Survarachakan, Thomas Langø, Ole-Jakob Elle, and Frank Lindseth. Learning deep abdominal ct registration through adaptive loss weighting and synthetic data generation. *Plos one*, 18(2):e0282110, 2023.

[73] Chen Qin, Shuo Wang, Chen Chen, Wenjia Bai, and Daniel Rueckert. Generative Myocardial Motion Tracking via Latent Space Exploration with Biomechanics-informed Prior, June 2022. arXiv:2206.03830 [cs, eess].

[74] Chen Qin, Shuo Wang, Chen Chen, Huaqi Qiu, Wenjia Bai, and Daniel Rueckert. Biomechanics-informed Neural Networks for Myocardial Motion Tracking in MRI, July 2020. arXiv:2006.04725 [cs, eess].

[75] Huaqi Qiu, Chen Qin, Andreas Schuh, Kerstin Hammernik, and Daniel Rueckert. Learning diffeomorphic and modality-invariant registration using b-splines. 2021.

[76] Lei Qu, Fuhui Long, and Hanchuan Peng. 3-d registration of biological images and models: registration of microscopic images and its uses in segmentation and annotation. *IEEE Signal Processing Magazine*, 32(1):70–77, 2014.

[77] Marc-Michel Rohé, Manasi Datar, Tobias Heimann, Maxime Sermesant, and Xavier Pennec. Svf-net: learning deformable image registration using shape matching. In *Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part I 20*, pages 266–274. Springer, 2017.

[78] Julian G Rosenman, Elizabeth P Miller, and Tim J Cullip. Image registration: an essential part of radiation therapy treatment planning. *International Journal of Radiation Oncology* Biology* Physics*, 40(1):197–205, 1998.

[79] Hanna Siebert, Lasse Hansen, and Mattias P Heinrich. Fast 3d registration with accurate optimisation and little learning for learn2reg 2021. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 174–179. Springer, 2021.

[80] Hessam Sokooti, Bob De Vos, Floris Berendsen, Boudewijn PF Lelieveldt, Ivana Išgum, and Marius Staring. Nonrigid image registration using multi-scale 3d convolutional neural networks. In *Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part I 20*, pages 232–239. Springer, 2017.

 https://doi.org/10.52202/079017-3439

[81] Joo Hyun Song, Gary E Christensen, Jeffrey A Hawley, Ying Wei, and Jon G Kuhl. Evaluating image registration using nirep. In *Biomedical Image Registration: 4th International Workshop, WBIR 2010, Lübeck, Germany, July 11-13, 2010. Proceedings 4*, pages 140–150. Springer, 2010.

[82] Takeshi Teshima, Isao Ishikawa, Koichi Tojo, Kenta Oono, Masahiro Ikeda, and Masashi Sugiyama. Coupling-based Invertible Neural Networks Are Universal Diffeomorphism Approximators. In *Advances in Neural Information Processing Systems*, volume 33, pages 3362–3373. Curran Associates, Inc., 2020.

[83] Lin Tian, Hastings Greer, François-Xavier Vialard, Roland Kwitt, Raúl San José Estépar, Richard Jarrett Rushmore, Nikolaos Makris, Sylvain Bouix, and Marc Niethammer. Gradicon: Approximate diffeomorphisms via gradient inverse consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18084–18094, 2023.

[84] Lin Tian, Zi Li, Fengze Liu, Xiaoyu Bai, Jia Ge, Le Lu, Marc Niethammer, Xianghua Ye, Ke Yan, and Daikai Jin. SAME++: A Self-supervised Anatomical eMbeddings Enhanced medical image registration framework using stable sampling and regularized transformation, November 2023. arXiv:2311.14986 [cs].

[85] Arthur W Toga and Paul M Thompson. The role of image registration in brain mapping. *Image and vision computing*, 19(1-2):3–24, 2001.

[86] Nicholas J Tustison, Hans J Johnson, Torsten Rohlfing, Arno Klein, Satrajit S Ghosh, Luis Ibanez, and Brian B Avants. Instrumentation bias in the use and evaluation of scientific software: recommendations for reproducible practices in the computational sciences, 2013.

[87] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep Image Prior. *International Journal of Computer Vision*, 128(7):1867–1888, July 2020. arXiv:1711.10925 [cs, stat].

[88] Hristina Uzunova, Matthias Wilms, Heinz Handels, and Jan Ehrhardt. Training cnns for image registration from few samples with model-based data augmentation. In *Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part I 20*, pages 223–231. Springer, 2017.

[89] David C Van Essen, Heather A Drury, Sarang Joshi, and Michael I Miller. Functional and structural mapping of human cerebral cortex: solutions are in the surfaces. *Proceedings of the National Academy of Sciences*, 95(3):788–795, 1998.

[90] Erdem Varol, Amin Nejatbakhsh, Ruoxi Sun, Gonzalo Mena, Eviatar Yemini, Oliver Hobert, and Liam Paninski. Statistical atlas of c. elegans neurons. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part V 23*, pages 119–129. Springer, 2020.

[91] Vivek Venkatachalam, Ni Ji, Xian Wang, Christopher Clark, James Kameron Mitchell, Mason Klein, Christopher J Tabone, Jeremy Florman, Hongfei Ji, Joel Greenwood, et al. Pan-neuronal imaging in roaming caenorhabditis elegans. *Proceedings of the National Academy of Sciences*, 113(8):E1082–E1088, 2016.

[92] Tom Vercauteren, Xavier Pennec, Aymeric Perchant, and Nicholas Ayache. Non-parametric diffeomorphic image registration with the demons algorithm. In *International conference on medical image computing and computer-assisted intervention*, pages 319–326. Springer, 2007.

[93] Tom Vercauteren, Xavier Pennec, Aymeric Perchant, and Nicholas Ayache. Symmetric Log-Domain Diffeomorphic Registration: A Demons-Based Approach. In Dimitris Metaxas, Leon Axel, Gabor Fichtinger, and Gábor Székely, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2008*, Lecture Notes in Computer Science, pages 754–761, Berlin, Heidelberg, 2008. Springer.

[94] Tom Vercauteren, Xavier Pennec, Aymeric Perchant, and Nicholas Ayache. Diffeomorphic demons: Efficient non-parametric image registration. *NeuroImage*, 45(1):S61–S72, March 2009.

[95] Tom Vercauteren, Xavier Pennec, Aymeric Perchant, Nicholas Ayache, et al. Diffeomorphic demons using itk's finite difference solver hierarchy. *The Insight Journal*, 1, 2007.

[96] Quanxin Wang, Song-Lin Ding, Yang Li, Josh Royall, David Feng, Phil Lesnar, Nile Graddis, Maitham Naeemi, Benjamin Facer, Anh Ho, Tim Dolbeare, Brandon Blanchard, Nick Dee, Wayne Wakeman, Karla E. Hirokawa, Aaron Szafer, Susan M. Sunkin, Seung Wook Oh, Amy Bernard, John W. Phillips, Michael Hawrylycz, Christof Koch, Hongkui Zeng, Julie A. Harris, and Lydia Ng. The Allen Mouse Brain Common Coordinate Framework: A 3D Reference Atlas. *Cell*, 181(4):936–953.e20, May 2020.

[97] Yan Wang, Xu Wei, Fengze Liu, Jieneng Chen, Yuyin Zhou, Wei Shen, Elliot K. Fishman, and Alan L. Yuille. Deep Distance Transform for Tubular Structure Segmentation in CT Scans. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3832–3841, Seattle, WA, USA, June 2020. IEEE.

[98] Asmamaw T Wassie, Yongxin Zhao, and Edward S Boyden. Expansion microscopy: principles and uses in biological research. *Nature methods*, 16(1):33–41, 2019.

[99] Jelmer M Wolterink, Jesse C Zwienenberg, and Christoph Brune. Implicit Neural Representations for Deformable Image Registration. page 11.

[100] Jelmer M Wolterink, Jesse C Zwienenberg, and Christoph Brune. Implicit neural representations for deformable image registration. In *International Conference on Medical Imaging with Deep Learning*, pages 1349–1359. PMLR, 2022.

[101] Yifan Wu, Tom Z. Jiahao, Jiancong Wang, Paul A. Yushkevich, M. Ani Hsieh, and James C. Gee. NODEO: A Neural Ordinary Differential Equation Based Optimization Framework for Deformable Image Registration. *arXiv:2108.03443 [cs]*, February 2022. arXiv: 2108.03443.

[102] Chenglong Xia, Jean Fan, George Emanuel, Junjie Hao, and Xiaowei Zhuang. Spatial transcriptome profiling by merfish reveals subcellular rna compartmentalization and cell cycle-dependent gene expression. *Proceedings of the National Academy of Sciences*, 116(39):19490–19499, 2019.

[103] Deshan Yang, Hua Li, Daniel A Low, Joseph O Deasy, and Issam El Naqa. A fast inverse consistent deformable image registration method based on symmetric optical flow computation. *Physics in Medicine & Biology*, 53(21):6143, 2008.

[104] Michael A Yassa, Shauna M Stark, Arnold Bakker, Marilyn S Albert, Michela Gallagher, and Craig EL Stark. High-resolution structural and functional mri of hippocampal ca3 and dentate gyrus in patients with amnestic mild cognitive impairment. *Neuroimage*, 51(3):1242–1252, 2010.

[105] Inwan Yoo, David GC Hildebrand, Willie F Tobin, Wei-Chung Allen Lee, and Won-Ki Jeong. ssemnet: Serial-section electron microscopy image registration using a spatial transformer network with learned features. pages 249–257, 2017.

[106] Paul A Yushkevich, John Pluta, Hongzhi Wang, Laura EM Wisse, Sandhitsu Das, and David Wolk. Ic-p-174: fast automatic segmentation of hippocampal subfields and medial temporal lobe subregions in 3 tesla and 7 tesla t2-weighted mri. *Alzheimer's & Dementia*, 12:P126–P127, 2016.

[107] Liutong Zhang, Lei Zhou, Ruiyang Li, Xianyu Wang, Boxuan Han, and Hongen Liao. Cascaded feature warping network for unsupervised medical image registration. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 913–916. IEEE, 2021.

[108] Shengyu Zhao, Yue Dong, Eric I-Chao Chang, and Yan Xu. Recursive cascaded networks for unsupervised medical image registration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[109] Shengyu Zhao, Tingfung Lau, Ji Luo, I Eric, Chao Chang, and Yan Xu. Unsupervised 3d end-to-end medical image registration with volume tweening network. *IEEE journal of biomedical and health informatics*, 24(5):1394–1404, 2019.

 https://doi.org/10.52202/079017-3439

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: All claims are shown empirically in the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The families of registration algorithms is limited to gradient-based methods to isolate dynamics of gradient-based methods. Our paper only uses brain datasets, since neuroimaging is one of the most (if not the most) popular modalities for studying registration algorithms. This is discussed in the limitations subsection.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Supplemental material contains scripts to reproduce all experiments of the paper. Code and pretrained models will be published to Github upon acceptance, with additional documentation, tutorials and instructions. Data is publicly available.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

https://doi.org/10.52202/079017-3439

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Code is provided in the supplemental material. Data is publicly available and instructions to reproduce the results are provided in the supplemental material.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: All recommended parameters of the existing methods are used (and mentioned in the paper).

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: p-values are reported for t-tests comparing classical and deep learning methods. Boxplots with interquartile ranges are reported for all experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Each baseline used in the paper has its own requirement specified in their work. We use a single machine for all experiments mentioned (specs are specified in the paper).

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: No research is performed involving new human subjects, animals, or environmental impact. Existing datasets comply with Code of Ethics. The proposed research is entirely computational. The proposed research has no immediate negative societal impact.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: Medical image registration has no immediate negative societal impact necessitating a dedicated discussion.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

https://doi.org/10.52202/079017-3439

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Appropriate citations are provided for existing code and data.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: Instructions to reproduce the results are provided in the paper and supplemental material. Only instructions to run the existing baselines are provided, no new method is proposed.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

https://doi.org/10.52202/079017-3439