ICON: Learning Regular Maps Through Inverse Consistency

Hastings Greer[1] Roland Kwitt[2] François-Xavier Vialard[3] Marc Niethammer[1]

[1]UNC Chapel Hill [2]University of Salzburg [3]LIGM, Université Gustave Eiffel

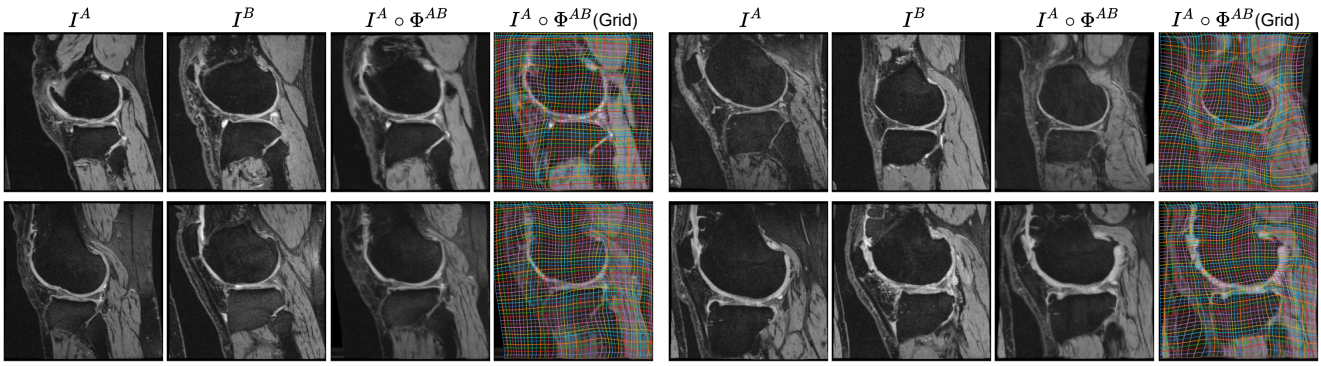

Figure 1: Example inverse consistent network (`ICON`) registration results for OAI knee images (see §4), obtained from a U-Net trained for *inverse consistency* (without any explicit loss to promote map regularity). All four panels show (*left to right*) the (1) moving image, (2) fixed image, (3) warped moving image and the (4) corresponding transformation grid (colored). Transformations are, as desired, smooth.

Abstract

Learning maps between data samples is fundamental. Applications range from representation learning, image translation and generative modeling, to the estimation of spatial deformations. Such maps relate feature vectors, or map between feature spaces. Well-behaved maps should be regular, which can be imposed explicitly or may emanate from the data itself. We explore what induces regularity for spatial transformations, e.g., when computing image registrations. Classical optimization-based models compute maps between pairs of samples and rely on an appropriate regularizer for well-posedness. Recent deep learning approaches have attempted to avoid using such regularizers altogether by relying on the sample population instead. We explore if it is possible to obtain spatial regularity using an inverse consistency loss only and elucidate what explains map regularity in such a context. We find that deep networks combined with an inverse consistency loss and randomized off-grid interpolation yield well behaved, approximately diffeomorphic, spatial transformations. Despite the simplicity of this approach, our experiments present compelling evidence, on both synthetic and real data, that regular maps can be obtained without carefully tuned explicit regularizers, while achieving competitive registration performance.

1. Motivation

Learning maps between feature vectors or spaces is an important task. Feature vector maps are used to improve representation learning [7], or to learn correspondences in natural language processing [4]. Maps between spaces are important for generative models when using normalizing flows [24] (to map between a simple and a complex probability distribution), or to determine spatial correspondences between images, *e.g.*, for optical flow [16] to determine motion from videos [12], depth estimation from stereo images [25], or medical image registration [39, 40].

Regular maps are typically desired; *e.g.*, diffeomorphic maps for normalizing flows to properly map densities, or for medical image registration to map to an atlas space [20]. Estimating such maps requires an appropriate choice of transformation model. This entails picking a parameterization, which can be simple and depend on few parameters (*e.g.*, an affine transformation), or which can have millions of parameters for 3D nonparametric approaches [14]. Regularity is achieved by 1) picking a simple transformation model with limited degrees of freedom, 2) regularization of the transformation parameters, 3) or implicitly through the data itself. Our goal is to demonstrate and understand how spatial regularity of a transformation can be achieved by encouraging *inverse consistency* of a map. Our motivating example is

image registration/optical flow, but our results are applicable to other tasks where spatial transformations are sought.

Registration problems have traditionally been solved by numerical optimization [28] of a loss function balancing an image similarity measure and a regularizer. Here, the predominant paradigm is *pair-wise* image registration[1] where many maps may yield good image similarities between a transformed moving and a fixed image; the regularizer is required for well-posedness to single out the most desirable map. Many different regularizers have been proposed [14, 28, 33] and many have multiple hyperparameters, making regularizer choice and tuning difficult in practice. Deep learning approaches to image registration and optical flow have moved to learning maps from *many image pairs*, which raises the question if explicit spatial regularization is still required, or if it will emanate as a consequence of learning over many image pairs. For optical flow, encouraging results have been obtained without using a spatial regularizer [10, 32], though more recent work has advocated for spatial regularization to avoid "vague flow boundaries and undesired artifacts" [18, 19]. Interestingly, for medical image registration, where map regularity is often very important, almost all the existing work uses regularizers as initially proposed for pairwise image registration [36, 42, 2] with the notable exception of [3] where the deformation space is guided by an autoencoder instead.

Limited work explores if regularization for deep registration networks can be avoided entirely, or if weaker forms of regularizations might be sufficient. To help investigate this question, we work with binary shapes (where regularization is particularly important due to the aperture effect [15]) and real images. We show that regularization is necessary, but that carefully encouraging *inverse consistency* of a map suffices to obtain approximate diffeomorphisms. The result is a simple, yet effective, nonparametric approach to obtain well-behaved maps, which only requires limited tuning. In particular, the in practice often highly challenging process of selecting a spatial regularizer is eliminated.

Our contributions are as follows: (1) We show that *approximate* inverse consistency, combined with off-grid interpolation, results in approximate diffeomorphisms, when using a deep registration model trained on large datasets. Foregoing regularization is insufficient; (2) Bottleneck layers are not required and many network architectures are suitable; (3) Affine preregistration is not required; (4) We propose randomly sampled evaluations to avoid transformation flips in texture-less areas and an inverse consistency loss with beneficial boundary effects; (5) We present good results of our approach on synthetic data, MNIST, and a 3D magnetic resonance (MR) knee dataset of the Osteoarthritis Initiative (OAI).

[1] A notable exception is congealing [45].

2. Background and Analysis

Image registration is typically based on solving optimization problems of the form

$$\theta^* = \operatorname*{argmin}_{\theta} \mathcal{L}_{\text{sim}}(I^A \circ \Phi_\theta^{-1}, I^B) + \lambda \mathcal{L}_{\text{reg}}(\theta) \ , \quad (1)$$

where I^A and I^B are moving and fixed images, $\mathcal{L}_{\text{sim}}(\cdot,\cdot)$ is the similarity measure, $\mathcal{L}_{\text{reg}}(\cdot)$ is a regularizer, θ are the transformation parameters, Φ_θ is the transformation map, and $\lambda \geq 0$. We consider images as functions from $\mathbb{R}^N$ to $\mathbb{R}$ and maps as functions from $\mathbb{R}^N$ to $\mathbb{R}^N$. We write $\|f\|_p$ for the L^p norm on a scalar or vector-valued function f.

Maps, Φ_θ, can be parameterized using few parameters (*e.g.*, affine, B-spline [14]) or nonparametrically with continuous vector fields [28]. In the nonparametric case, parameterizations are infinite-dimensional (as one deals with function spaces) and represent displacement, velocity, or momentum fields [2, 36, 42, 28]. Solutions to Eq. (1) are classically obtained via numerical optimization [28]. Recent deep registration networks are conceptually similar, but *predict* $\tilde{\theta}^*$, *i.e.*, an estimate of the true minimizer θ^*.

There are three interesting observations: *First*, for transformation models with few parameters (*e.g.*, affine), regularization is often not used (*i.e.*, $\lambda = 0$). *Second*, while deep learning (DL) models minimize losses similar to Eq. (1), the parameterization is different: it is over network *weights*, resulting in a predicted θ^* instead of optimizing over θ directly. *Third*, DL models are trained over *large collections of image pairs* instead of a single (I^A, I^B) pair. This raises the following questions: **Q1**) Is explicit spatial regularization necessary, or can we avoid it for nonparametric registration models? **Q2**) Is using a *single* neural network parameterization to predict *all* θ^* beneficial? For instance, will it result in simple solutions as witnessed for deep networks on other tasks [35] or capture meaningful deformation spaces as observed in [42]? **Q3**) Does a deep network parameterization itself result in regular solutions, even if only applied to a single image pair, as such effects have, *e.g.*, been observed for structural optimization [17]?

Regularization typically encourages spatial smoothness by penalizing derivatives (or smoothing in dual space). Commonly, one uses a Sobolev norm or total variation. Ideally, one would like a regularizer adapted to deformations one expects to see (as it encodes a prior on expected deformations *e.g.*, as in [29]). In consequence, picking and tuning a regularizer is cumbersome and often involves many hyperparameters. While avoiding explicit regularization has been explored for deep registration / optical flow networks [10, 32], there is evidence that regularization is beneficial [18].

Our key idea is to avoid complex spatial regularization and to instead obtain approximate diffeomorphisms by encouraging inverse consistent maps via regularization.

2.1. Weakly-regularized registration

Assume we eliminate regularization ($\lambda = 0$) and use the p-th power of the L^p norm of the difference between the warped image, $I^A \circ \Phi_\theta^{-1}$, and the fixed image, I^B, as similarity measure. Then, our optimization problem becomes

$$\theta^* = \arg\min_\theta \int (I^A(\Phi_\theta^{-1}(x)) - I^B(x))^p \, \mathrm{d}x \,, \; p \geq 1, \quad (2)$$

i.e., the image intensities of I^A should be close to the image intensities of I^B *after* deformation. Without regularization, we are entirely free to choose Φ_θ. Highly irregular minimizers of Eq. (2) may result as each intensity value I^A is simply matched to the closest intensity value of I^B regardless of location. For instance, for a constant $I^B(x) = c$ and a moving image $I^A(y)$ with a unique location y_c, where $I^A(y_c) = c$, the optimal map is $\Phi_\theta^{-1}(x) = y_c$, which is not invertible: only *one point* of I^A will be mapped to the *entire* domain of I^B. Clearly, more spatial regularity is desirable. Importantly, irregular deformations are common optimizers of Eq. (2).

Optimal mass transport (OMT) is widely used in machine learning and in imaging. Such models are of interest to us as they can be inverse consistent. An OMT variant of the discrete reformulation of Eq. (2) is

$$\theta^* = \arg\min_\theta \mathrm{d}x \sum_{i=1}^{S} (I^A(\Phi_\theta^{-1}(x_i)) - I^B(x_i))^p \quad (3)$$

for $p \geq 1$, where i indexes the S grid points x_i, $\Phi_\theta^{-1}(x_i)$ is restricted to map to the grid points y_i of I^A, and $\mathrm{d}x$ is the discrete area element. Instead of considering all possible maps, we attach a unit mass to each *intensity* value of I^A and I^B and ask for minimizers of Eq. (3) which transform the intensity distribution of I^A to the intensity distribution of I^B via *permutations* of the values only. As we only allow permutations, the optimal map will be *invertible* by construction. This problem is equivalent to optimal mass transport for one-dimensional empirical measures [31]. One obtains the optimal value by ordering all intensity values of I^A ($I_1^A \leq \cdots \leq I_S^A$) and I^B ($I_1^B \leq \cdots \leq I_S^B$). The minimum is the p-th power of the p-Wasserstein distance ($p \geq 1$) $\mathcal{W}_p^p = \sum_i |I_i^A - I_i^B|^p$. In consequence, minimizers for Eq. (2) are related to sorting, but do not consider spatial regularity. Note that solutions might not be unique when intensity values in I^A or I^B are repeated. Solutions via sorting were empirically explored for registration in [34] to illustrate that they, in general, do not result in spatially meaningful registrations. At this point, our idea of using inverse consistency (*i.e.*, invertible maps) as the only regularizer appears questionable, given that OMT often provides an inverse consistent model (when a matching, *i.e.*, a Monge solution, is optimal), while resulting in irregular maps (Fig. 2).

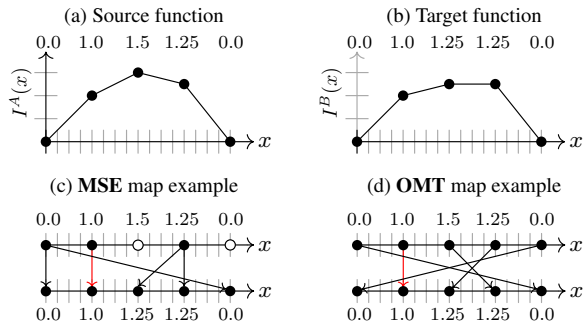

Figure 2: Source and target functions for a 1D registration example. Panels (c) and (d) show two possible solutions for mean square error (MSE) and OMT, respectively. In both cases, solutions may not be unique. However, for OMT, matching solutions will be one-to-one, *i.e.*, invertible. OMT imposes a stronger constraint than MSE on the obtainable maps, but irregular maps are still permissible.

Yet, we will show that a registration network, combined with an inverse consistency loss, encourages map regularity.

2.2. Avoiding undesirable solutions

Simplicity. The highly irregular maps in Fig. 2 occur for *pair-wise* image registration. Instead, we are concerned with training a network over an *entire image population*. Were one to find a global inverse consistent minimizer, a network would need to implicitly approximate the sorting-based OMT solution. As sorting is a continuous piece-wise linear function [5], it can, in principle, be approximated according to the universal approximation theorem [26]. However, this is a limit argument. Practical neural networks for sorting are either *approximate* [27, 11] or very large (*e.g.*, $O(S^2)$ neurons for S values [6]). Note that deep networks often tend to simple solutions [35] and that we do not even want to sort *all* values for registration. Instead, we are interested in more *local* permutations, rather than the global OMT permutations, which is what we will obtain for neural network solutions with inverse consistency.

Invertibility. Requiring map invertibility implies searching for a matching (a Monge formulation in OMT) which is an optimal permutation, but which may not be continuous[2]. Instead, our goal is a *continuous and invertible* map. We therefore want to penalize deviations from

$$\Phi_\theta^{AB} \circ \Phi_\theta^{BA} = \mathrm{Id} \,, \quad (4)$$

where Φ_θ^{AB} denotes a predicted map (by a network with weights θ) to register image I^A to I^B; Φ_θ^{BA} is the network output with reversed inputs and Id denotes the identity map.

Inverse consistency of maps has been explored to obtain symmetric maps for pair-wise registration [13, 8] and for

[2]It would be interesting to study how well a network approximates an OMT solution and if it naturally regularizes it.

registration networks [43, 36]. Related losses have been proposed on images (instead of maps) for registration [22, 21] and for image translation [44]. However, none of these approaches study inverse consistency for regularization. Likely, because it has so far been believed that additional spatial regularization is required for nonparametric registration.

2.3. Approximate inverse consistency

As we will show next, *approximate inverse consistency* by itself yields regularizing effects in the context of pairwise image registration.

Denote by $\Phi_\theta^{AB}(x)$ and $\Phi_\theta^{BA}(x)$ the output maps of a network for images (I^A, I^B) and (I^B, I^A), respectively. As inverse consistency by itself does not prevent discontinuous solutions, we propose to use *approximate* inverse consistency to favor C^0 solutions. We approximate the network's variance as two vector-valued independent spatial white noises $n_1(x), n_2(x) \in \mathbb{R}^N$ ($x \in [0,1]^N$ with N=2 or N=3 the image dim.) of variance 1 for each space location and dimension and define

$$\Phi_{\theta\varepsilon}^{AB}(x) = \Phi_\theta^{AB}(x) + \varepsilon n_1(\Phi_\theta^{AB}(x)) \ ,$$
$$\Phi_{\theta\varepsilon}^{BA}(x) = \Phi_\theta^{BA}(x) + \varepsilon n_2(\Phi_\theta^{BA}(x)) \ ,$$

with $\varepsilon > 0$. We then consider the loss $\mathcal{L} = \lambda \mathcal{L}_{\text{inv}} + \mathcal{L}_{\text{sim}}$, with inverse consistency component ($\mathcal{L}_{\text{inv}}$)

$$\mathcal{L}_{\text{inv}} = \left\|\Phi_{\theta\varepsilon}^{AB} \circ \Phi_{\theta\varepsilon}^{BA} - \text{Id}\right\|_2^2 + \left\|\Phi_{\theta\varepsilon}^{BA} \circ \Phi_{\theta\varepsilon}^{AB} - \text{Id}\right\|_2^2 \quad (5)$$

and similarity component ($\mathcal{L}_{\text{sim}}$)

$$\mathcal{L}_{\text{sim}} = \left\|I^A \circ \Phi_\theta^{AB} - I^B\right\|_2^2 + \left\|I^B \circ \Phi_\theta^{BA} - I^A\right\|_2^2 \ . \quad (6)$$

Importantly, note that there are *multiple* maps that can lead to the same $I^A \circ \Phi_\theta^{AB}$ and $I^B \circ \Phi_\theta^{BA}$. Therefore, among all these maps, minimizing the loss $\mathcal{L}$ drives the maps towards those that minimize the two terms in Eq. (5).

Assumption. *Both terms in Eq.* (5) *can be driven to a small value (of the order of the noise), by minimization.*

We first Taylor-expand one of the two terms in Eq. (5) (the other follows similarly), yielding

$$\left\|\Phi_{\theta\varepsilon}^{AB} \circ \Phi_{\theta\varepsilon}^{BA} - \text{Id}\right\|_2^2 \approx \left\|\Phi_\theta^{AB} \circ \Phi_\theta^{BA} + \varepsilon n_1(\Phi_\theta^{AB} \circ \Phi_\theta^{BA}) + \mathrm{d}\Phi_{\theta\varepsilon}^{AB}(\varepsilon n_2(\Phi_\theta^{BA})) - \text{Id}\right\|_2^2 \ . \quad (7)$$

Defining the right-hand side as A, developing the squares and taking expectation, we obtain

$$\mathbb{E}[A] = \left\|\Phi_\theta^{AB} \circ \Phi_\theta^{BA} - \text{Id}\right\|_2^2 + \varepsilon^2 \mathbb{E}\left[\left\|n_1 \circ (\Phi_{\theta\varepsilon}^{AB} \circ \Phi_{\theta\varepsilon}^{BA})\right\|_2^2\right] + \varepsilon^2 \mathbb{E}\left[\left\|\mathrm{d}\Phi_{\theta\varepsilon}^{AB}(n_2) \circ \Phi_\theta^{BA}\right\|_2^2\right] \ , \quad (8)$$

since, by independence, all the cross-terms vanish (the noise terms have 0 mean value). The second term is constant, *i.e.*,

$$\mathbb{E}\left[\left\|n_1 \circ (\Phi_{\theta\varepsilon}^{AB} \circ \Phi_{\theta\varepsilon}^{BA})\right\|_2^2\right] = \int \mathbb{E}\left[\|n_1\|_2^2(y)\right] \text{Jac}((\Phi_{\theta\varepsilon}^{BA})^{-1} \circ (\Phi_{\theta\varepsilon}^{AB})^{-1}) \, \mathrm{d}y = \text{const.} \ , \quad (9)$$

where we performed a change of variables and denoted the determinant of the Jacobian matrix as Jac. The last equality follows from the fact that the variance of the noise term is spatially constant and equal to 1. By similar arguments, the last expectation term in Eq. (8) can be rewritten as

$$\mathbb{E}\left[\left\|\mathrm{d}\Phi_{\theta\varepsilon}^{AB}(n_2) \circ \Phi_\theta^{BA}\right\|_2^2\right] = \int \text{Tr}(\mathrm{d}(\Phi_{\theta\varepsilon}^{AB})^\top \mathrm{d}\Phi_{\theta\varepsilon}^{AB}) \, \text{Jac}((\Phi_\theta^{BA})^{-1}) \, \mathrm{d}y \ , \quad (10)$$

where Tr denotes the trace operator. As detailed in the suppl. material, the identity of Eq. (10) relies on a change of variable and on the property of the white noise, n_2, which satisfies null correlation in space and dimension $\mathbb{E}[n_2(x)n_2(x')^\top] = \text{Id}_{\mathbb{R}^N}$ if $x = x'$ and 0 otherwise.

Approximation & H^1 regularization. We now want to connect the approximate inverse consistency loss of Eq. (5) with H^1 norm type regularization. Our assumption implies that $\Phi_\theta^{AB} \circ \Phi_\theta^{BA}, \Phi_\theta^{BA} \circ \Phi_\theta^{AB}$ are close to identity, therefore one has $\text{Jac}((\Phi_\theta^{BA})^{-1}) \approx \text{Jac}(\Phi_\theta^{AB})$. Assuming this approximation holds, we use it in Eq. (10), together with the fact that, $\Phi_{\theta\varepsilon}^{AB} \approx \Phi_\theta^{AB} + O(\varepsilon)$ to get at order ε^2 (see suppl. material for details) to approximate $\mathcal{L}_{\text{inv}}$, *i.e.*,

$$\mathcal{L}_{\text{inv}} \approx \left\|\Phi_\theta^{AB} \circ \Phi_\theta^{BA} - \text{Id}\right\|_2^2 + \left\|\Phi_\theta^{BA} \circ \Phi_\theta^{AB} - \text{Id}\right\|_2^2 + \varepsilon^2 \left\|\mathrm{d}\Phi_\theta^{AB} \sqrt{\text{Jac}(\Phi_\theta^{AB})}\right\|_2^2 + \varepsilon^2 \left\|\mathrm{d}\Phi_\theta^{BA} \sqrt{\text{Jac}(\Phi_\theta^{BA})}\right\|_2^2 \ . \quad (11)$$

We see that approximate inverse consistency leads to an L^2 penalty of the gradient, weighted by the Jacobian of the map. This is a type of Sobolev (H^1 more precisely) regularization sometimes used in image registration. In particular, the H^1 term is likely to control the compression and expansion magnitude of the maps, at least on average, on the domain. *Hence, approximate inverse consistency leads to an implicit H^1 regularization, formulated directly on the map.*

Inverse consistency with no noise and the implicit regularization of inverse consistency. Turning the noise level to zero also leads to regular displacement fields in our experiments when predicting maps with a neural network. In this case, we observe that inverse consistency is only approximately achieved. Therefore, one can postulate that the error made in computing the inverse entails the H^1 regularization as previously shown. The possible caveat of this hypothesis is that the inverse consistency error might not be independent

of the displacement fields, which was assumed in proving the emerging H^1 regularization. Last, even when the network should have the capacity to exactly satisfy inverse consistency for all data, we conjecture that the implicit bias due to the optimization will favor more regular outputs.

A fully rigorous theoretical understanding of the regularization effect due to the data population and its link with inverse consistency is important, but beyond our scope here.

3. Approximately diffeomorphic registration

We base our registration approach on training a neural network F_θ^{AB} which, given input images I^A and I^B, outputs a grid of *displacement* vectors, D_θ^{AB}, in the space of image I^B, assuming normalized image coordinates covering $[0,1]^N$. We obtain *continuous* maps by interpolation, *i.e.*,

$$\Phi_\theta^{AB} = D_\theta^{AB} + \mathrm{Id}, \quad D_\theta^{AB} = \mathrm{interp}(F_\theta^{AB}) \tag{12}$$

where $I^A \circ \Phi_\theta^{AB} \approx I^B$. Under the assumption of linear interpolation (bilinear in 2D and trilinear in 3D), Φ_θ^{AB} is continuous and differentiable except on a measure zero set. Building on the considerations of Sec. 2 we seek to minimize

$$\mathcal{L}(\theta) = \mathbb{E}_{p(I^A, I^B)} \left[\mathcal{L}_{\text{sim}}^{AB} + \lambda \mathcal{L}_{\text{inv}}^{AB} \right], \tag{13}$$

where $\lambda \geq 0$ and $p(I^A, I^B)$ denotes the distribution over all possible image pairs. The similarity and invertibility losses depend on the neural network parameters, θ, and are

$$\begin{aligned} \mathcal{L}_{\text{sim}}^{AB} &= \mathcal{L}_{\text{sim}}(I^A \circ \Phi_\theta^{AB}, I^B) + \mathcal{L}_{\text{sim}}(I^B \circ \Phi_\theta^{BA}, I^A) \\ \mathcal{L}_{\text{inv}}^{AB} &= \mathcal{L}_{\text{inv}}(\Phi_\theta^{AB}, \Phi_\theta^{BA}) + \mathcal{L}_{\text{inv}}(\Phi_\theta^{BA}, \Phi_\theta^{AB}) \end{aligned} \tag{14}$$

with

$$\mathcal{L}_{\text{sim}}(I, J) = \|I - J\|_2^2, \mathcal{L}_{\text{inv}}(\phi, \psi) = \|\phi \circ \psi - \mathrm{Id}\|_2^2. \tag{15}$$

For simplicity, we use the squared L^2 norm as similarity measure. Other measures, *e.g.*, normalized cross correlation (NCC) or mutual information (MI), can also be used. When $\mathcal{L}_{\text{inv}}^{AB}$ goes to zero, Φ_θ^{AB} will be approx. invertible and continuous due to Eq. (12). Hence, we obtain approximate C^0 diffeomorphisms without differential equation integration, hyperparameter tuning, or transform restrictions. Our loss in Eq. (13) is symmetric in the image pairs due to the symmetric similarity and invertibility losses in Eq. (14).

Displacement-based inverse consistency loss. A general map Φ_θ^{AB} may map points in $[0,1]^N$ to points outside $[0,1]^N$. Extrapolating maps across the boundary is cumbersome. Hence, we only interpolate displacement fields as in Eq. (12). We rewrite the inverse consistency loss as

$$\begin{aligned} \mathcal{L}_{\text{inv}}(\Phi_\theta^{AB}, \Phi_\theta^{BA}) &= \left\| (D_\theta^{AB} + \mathrm{Id}) \circ (D_\theta^{BA} + \mathrm{Id}) - \mathrm{Id} \right\|_2^2 \\ &= \left\| (D_\theta^{AB}) \circ \Phi_\theta^{BA} + D_\theta^{BA} \right\|_2^2 \end{aligned} \tag{16}$$

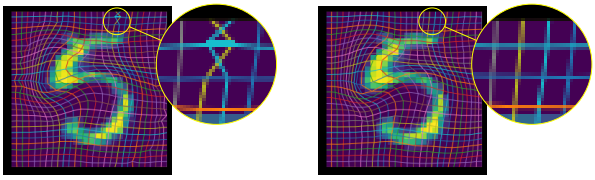

Figure 3: *Left*: Output generated by a network trained with inverse consistency, evaluated on a grid (not randomly); the loss cannot detect that these maps flip the pair of pixels in the upper right corner, as that error is not represented in the composed map. *Right*: Output from a network trained with random evaluation off of lattice points.

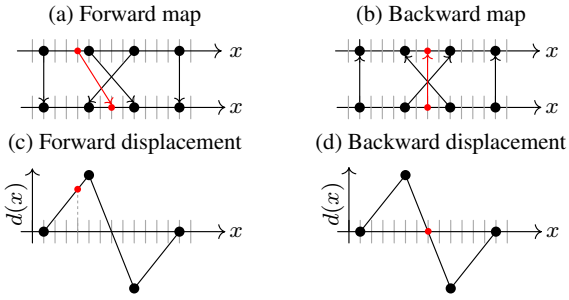

Figure 4: In this example, grid points (•) map to each other inverse consistently. The forward map (a) is inverted by the backward map (b) and folding of the space occurs as the middle two points swap positions. Off-grid points map under linear interpolation according to (c/d). The interpolated displacements for (•) do not result in an invertible map. Hence, this mismatch would be penalized by the inverse consistency loss, but only when evaluated off-grid.

and use it for implementation, as it is easier to evaluate.

Random evaluation of inverse consistency loss. $\mathcal{L}_{\text{inv}}^{AB}$ can be evaluated by approximating the L^2 norm, assuming constant values over the grid cells. In many cases, this is sufficient. However, as Fig. 3 illustrates, swapped locations may occur in uniform regions where a registration network only sees uniform background. This swap, composed with itself, is the identity as long as it is only evaluated at the center of pixels/voxels. Hence, the map appears invertible to the loss. However, outside the centers of pixels/voxels, the map is not inverse consistent when combined with linear interpolation. To avoid such pathological cases, we approximate the L^2 norm by random sampling. This forces interpolation and therefore results in non-zero loss values for swaps. Fig. 4 shows why off-grid sampling combined with inverse consistency is a stronger condition than only considering deformations at grid points. In practice, we evaluate the loss

$$\begin{aligned}
&\mathcal{L}_{\text{inv}}(\Phi_\theta^{AB}, \Phi_\theta^{BA}) = \left\|(D_\theta^{AB}) \circ \Phi_\theta^{BA} + D_\theta^{BA}\right\|_2^2 \\
&= \mathbb{E}_{x\sim\mathcal{U}(0,1)^N}\left[(D_\theta^{AB}) \circ \Phi_\theta^{BA} + D_\theta^{BA}\right]^2(x) \\
&\approx {}^1\!/_{N_p} \sum_i \left([(D_\theta^{AB}) \circ (D_\theta^{BA} + \mathrm{Id}) + D_\theta^{BA}](x_i + \epsilon_i)\right)^2 \quad (17) \\
&= {}^1\!/_{N_p} \sum_i \big([D_\theta^{AB} \circ (D_\theta^{BA} \circ (x_i + \epsilon_i) + x_i + \epsilon_i) \\
&\qquad + D_\theta^{BA} \circ (x_i + \epsilon_i)]\big)^2 \ ,
\end{aligned}$$

where N_p is the number of pixels/voxels, $\mathcal{U}(0,1)^N$ denotes the uniform distribution over $[0,1]^N$, x_i denotes the grid center coordinates and ϵ_i is a random sample drawn from a multivariate Gaussian with standard deviation set to the size of a pixel/voxel in the respective spatial directions.

4. Experiments

Our experiments address several aspects: First, we compare our approach to *directly* optimizing the maps Φ^{AB} and Φ^{BA} on a 2D toy dataset of 128×128 images. Second, on a 2D toy dataset of 28×28 images, we assess the impact of architectural and hyperparameter choices. Finally, we assess registration performance on real 3D MR images of the knee.

4.1. Datasets

MNIST. We use the standard MNIST dataset with images of size 28×28, restricted to the number "5" to make sure we have semantically matching images. For training/testing, we rely on the standard partitioning of the dataset.

Triangles & Circles. We created 2D triangles and circles (128×128) with radii and centers varying uniformly in $[.2, .4]$ and $[.4, .7]$, respectively. Pixels are set to 1 inside a shape and smoothly decay to -1 on the outside. We train using 6,000 images and test on 6,000 separate images[3].

OAI knee dataset. These are 3D MR images from the Osteoarthritis Initiative (OAI). Images are downsampled to size $192 \times 192 \times 80$, normalized such that the 1th percentile is set to 0, the 99th percentile is to 1, and all values are clamped to be in $[0, 1]$. As a preprocessing step, images of left knees are mirrored along the left-right axis. The dataset contains 2,532 training images and 301 test pairs.

4.2. Architectures

We experiment with four neural network architectures. All networks output displacement fields, D_θ^{AB}. We briefly outline the differences below, but refer to the suppl. material for details. The first network is an **MLP** with 2 hidden layers and ReLU activations. The output layer is reshaped into size $2 \times W \times H$. Second, we use a convolutional encoder-decoder network (**Enc-Dec**) with 5 layers each, reminiscent of a U-Net *without* skip connections. Our third network uses 6 convolutional layers without up- or down-sampling. The input to each layer is the concatenation of the outputs of all previous layers (**ConvOnly**). Finally, we use a **U-Net** with skip and residual connections. The latter is similar to **Enc-Dec**, but uses LeakyReLU activations and batch normalization. In all architectures, the final layer weights are initialized to 0, so that optimization starts at a network outputting a zero displacement field.

[3]Code to generate images and replicate these experiments is available at `https://github.com/uncbiag/ICON`

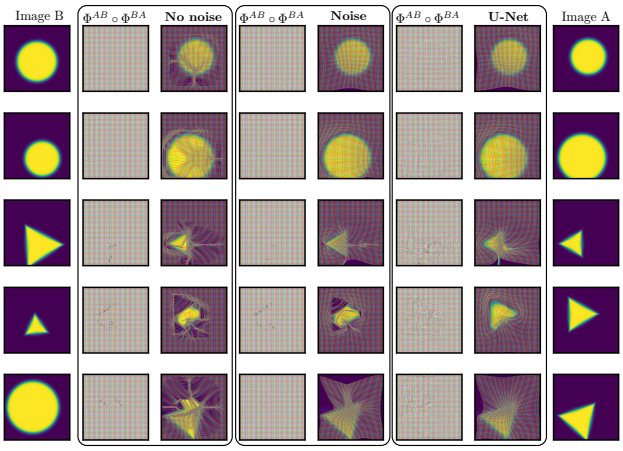

Figure 5: Comparison between **U-Net** results and **direct optimization** (no neural network; over Φ_θ^{AB} and Φ_θ^{BA}) w/ and w/o added noise, using the inverse consistency loss with $\lambda = 2,048$. Direct optimization w/o noise leads to irregular maps, while adding noise or using the **U-Net** improves map regularity (best viewed zoomed).

4.3. Regularization by approx. inverse consistency

Sec. 2.3 formalized that approximate inverse consistency results in regularizing effects. Specifically, when Φ_θ^{AB} is approximately the inverse of Φ_θ^{BA}, the inverse consistency loss $\mathcal{L}_{\text{inv}}^{AB}$ can be approximated based on Eq. (11), highlighting its implicit H^1 regularization. We investigate this behavior by three experiments: Pair-wise image registration (1) with artificially added noise (**noise**) and (2) without (**no noise**) artificially added noise, and (3) population-based registration via a **U-Net**. Fig. 5 shows some sample results, supporting our theoretical exposition of Sec. 2.3: Pair-wise image registration without noise results in highly irregular transformations even though the inverse consistency loss is used. Adding a small amount of Gaussian noise with standard deviation of 1/8th of a pixel (similar to the inverse consistency loss magnitudes we observe for a deep network) to the displacement fields before computing the inverse consistency loss, results in significantly more regular maps. Lastly, using a **U-Net** yields highly regular maps. Notably, all three approaches result in approximately inverse consistent maps. The behavior for pair-wise image registration elucidates why inverse consistency has not appeared in the classical (pair-wise) registration literature as a replacement for more complex spatial regularization. The proposed technique *only* results in regularity when inverse consistency errors are present.

MNIST								
Network →	**MLP**		**Enc-Dec**		**U-Net**		**ConvOnly**	
$\lambda \downarrow$	Dice	Folds	Dice	Folds	Dice	Folds	Dice	Folds
64	0.92	26.61	0.80	0.15	**0.93**	3.87	**0.93**	30.20
128	**0.92**	9.95	0.77	0.08	**0.92**	1.45	0.90	16.27
256	**0.91**	2.48	0.72	0.01	0.90	0.41	0.88	7.17
512	**0.90**	0.72	0.66	0.03	0.89	0.09	0.85	3.12
1,024	**0.88**	0.34	0.62	0.06	0.86	0.02	0.81	0.54
2,048	**0.87**	0.16	0.63	0.00	0.73	0.09	0.76	0.07
Triangles & Circles								
Network →	**MLP**		**Enc-Dec**		**U-Net**		**ConvOnly**	
$\lambda \downarrow$	Dice	Folds	Dice	Folds	Dice	Folds	Dice	Folds
64	**0.98**	1.24	0.94	3.50	**0.98**	2.74	0.97	12.57
128	**0.98**	0.73	0.90	2.71	**0.98**	1.59	0.96	10.15
256	**0.98**	0.27	0.88	1.11	0.97	1.14	0.96	8.49
512	**0.97**	0.10	0.87	0.65	0.96	0.70	0.94	6.61
1,024	**0.96**	0.03	0.86	0.22	0.95	0.25	0.92	3.91
2,048	**0.95**	0.03	0.85	0.15	0.94	0.09	0.89	2.18

Table 1: Network performance across architectures and regularization strength λ. **MLP / U-Net** perform best. All methods work.

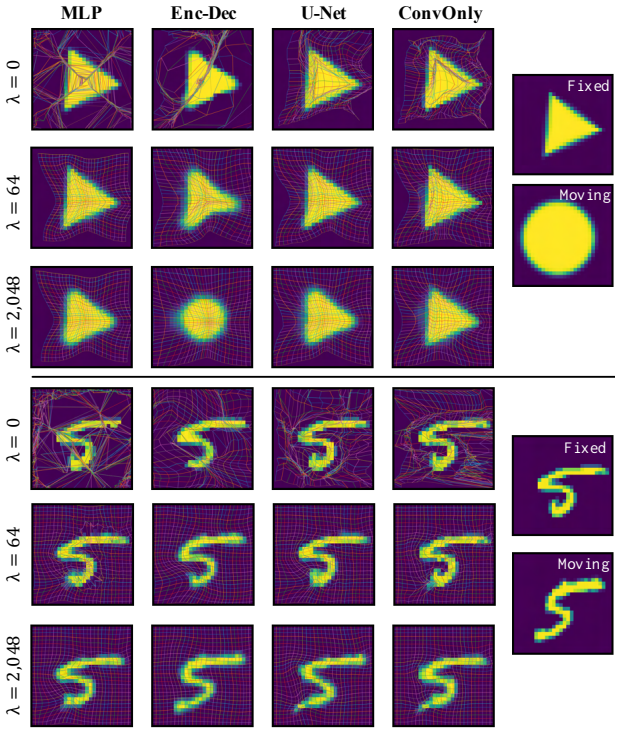

Figure 6: Comparison of networks as a function of λ. **U-Net** and **MLP** show the best performance due to their ability to capture long and short range dependencies. **Enc-Dec** and **ConvOnly**, which capture only long range and only short range dependencies, resp., also learn regular maps, but for a narrower range of λ. In all cases, maps become smooth for sufficiently large λ. Best viewed zoomed.

In summary, our theory is supported by our experimental results: approximate inverse consistency regularizes maps.

4.4. Regularization for different networks

Sec. 4.3 illustrated that approximate inverse consistency yields regularization effects which translate to regularity for network predictions, as networks will, in general, not achieve perfect inverse consistency. A natural next question to ask is "how much the results depend on a particular architecture"? To this end, we assess four different network types, focusing on MNIST and the triangles & circles data. We report two measures on held-out images: the *Dice score* of pixels with intensity greater than 0.5, and the mean number of *folds*, *i.e.*, pixels where the volume form $\mathrm{d}V$ of Φ is negative.

One hypothesis as to how network design could drive smoothness would be that smoothness is induced by convolutional layers (which can implement a smoothing kernel). If this were the case, we would expect the **MLP** to produce irregular maps with a high number of folds. Vice versa, since the **MLP** has no spatial prior, obtaining smooth transforms would indicate that smoothness is promoted by the loss itself. The latter is supported by Fig. 6, showing regular maps even for the **MLP** when λ is sufficiently large. Note that $\lambda = 0$ in Fig. 6 corresponds to an unregularized MSE solution, as discussed in Sec. 2.1; maps are, as expected, highly irregular and regularization via inverse consistency is clearly needed.

A second hypothesis is that regularity results from a *bottleneck* structure within a network, *e.g.*, a **U-Net**. In fact, Bhalodia *et al*. [3] show that autoencoders tend to yield smooth maps. To assess this hypothesis, we focus on the **Enc-Dec** and **ConvOnly** type networks; the former has a bottleneck structure, while the latter does not. Fig. 6 shows some support for the hypothesis that a bottleneck promotes smooth maps: for a specific λ, **Enc-Dec** appears to have more strongly regularized outputs compared to **U-Net**, with **ConvOnly** being the most irregular. Yet, higher values of λ (*e.g.*, 1,024 or 2,048) for **ConvOnly** yield equally smooth maps. Overall, a bottleneck structure does have a regularizing effect, but regularity can also be achieved by appropriately weighing the inverse consistency loss (see Tab. 1).

In summary, our experiments indicate that the regularizing effect of inverse consistency is a robust property of the loss, and should generalize well across architectures.

4.5. Performance for 3D image registration

For experiments on real data, we focus on the 3D OAI dataset. To demonstrate the versatility of the advocated inverse consistency loss in promoting map regularity, we refrain from affine pre-registration (as typically done in earlier works) and simply compose the maps of *multiple U-Nets* instead. In particular, we compose up to four U-Nets as follows: A composition of two U-Nets is initially trained on low-resolution image pairs. Weights are then frozen and this network is composed with a third U-Net, trained on high-resolution image pairs. This network is then optionally

Method	$\mathcal{L}_{\text{sim}}$	Dice	Folds	Time [s]
Demons	MSE	63.47	19.0	114
SyN	CC	65.71	0	1330
NiftyReg	NMI	59.65	0	143
NiftyReg	LNCC	67.92	203	270
vSVF-opt	LNCC	67.35	0	79
Voxelmorph (w/o affine)	MSE	46.06	83	0.12
Voxelmorph	MSE	66.08	39.0	0.31
AVSM (7-Step Affine, 3-Step Deformable)	LNCC	68.40	14.3	0.83
ICON (2 step ½ res., 2 step full res., w/o affine)	MSE	68.29	118.4	1.06
ICON (2 step ½ res., 1 step full res., w/o affine)	MSE	66.16	169.4	0.57
ICON (2 step ½ res., w/o affine)	MSE	59.36	49.35	0.09

Table 2: Comparison of ICON against the methods in [36], on cross-subject registration for OAI knee images.

frozen and composed with a fourth U-Net, again trained on high-resolution image pairs. During this multi-step training, the weighting of the inverse consistency loss is gradually increased. We train using ADAM [23] with a batch size of 128 in the low-res. stage, and a batch size of 16 in the high-res. stage. MSE is used as image similarity measure.

We compare our approach, InverseConsistentNet (ICON), against the methods of [36], in terms of (1) cartilage Dice scores between registered image pairs [1] (based on manual segmentations) and (2) the number of folds. The segmentations are not used during training and allow quantifying if the network yields semantically meaningful registrations. Tab. 2 lists the corresponding results, Fig. 1 shows several example registrations. Unlike the other methods in Tab. 2, except where explicitly noted, ICON does not require affine pre-registration. Since affine maps are inverse consistent, they are not penalized by our method. Notably, despite its simplicity, ICON yields performance (in terms of Dice score & folds) comparable to more complex, explicitly regularized methods. We emphasize that our objective is not to outperform existing techniques, but to present evidence that regular maps can be learned *without* carefully tuned regularizers.

In summary, using ICON yields (1) competitive Dice scores, (2) acceptable folds and (3) fast performance.

5. Limitations, future work, & open questions

Several questions remain and there is no shortage of theoretical/practical directions, some of which are listed next.

Network architecture & optimization. Instead of specifying a spatial regularizer, we now specify a network architecture. While our results suggest regularizing effects for a variety of architectures, we are still lacking a clear understanding of how network architecture and numerical optimization influence solution regularity.

Diffemorphisms at test time. We simply encourage inverse consistency via a quadratic penalty. Advanced numerical approaches (*e.g.*, augmented Lagrangian methods [30]) could more strictly enforce inverse consistency during *training*. Our current approach is only *approximately diffeomorphic* at test time. To guarantee diffeomorphisms, one could explore combining inverse consistency with fluid deformation models [14]. These have been used for deep registration networks [42, 41, 36, 37, 9] combined with explicit spatial regularization. We would simply predict a velocity field and obtain the map via integration. By using our loss, sufficiently smooth velocity fields would likely emerge. Alternatively, one could use diffeomorphic transformation parameterizations by enforcing positive Jacobian determinants [38].

Multi-step. Our results show that using a multi-step estimation approach is beneficial; successive networks can refine deformation estimates and thereby improve registration performance. What the limits of such a multi-step approach are (i.e., when performance starts to saturate) and how it interacts with deformation estimates at different resolution levels would be interesting to explore further.

Similarity measures. For simplicity, we only explored MSE. NCC, local NCC, and mutual information would be natural choices for multi-modal registration. In fact, there are many opportunities to improve registrations *e.g.* using more discriminative similarity measures based on network-based features, multi-scale information, or side-information during training, *e.g.*, segmentations or point correspondences.

Theoretical investigations. It would be interesting to establish how regularization by inverse consistency relates to network capacity, expressiveness, and generalization. Further, establishing a rigorous theoretical understanding of the regularization effect due to the data *population* and its link with inverse consistency would be important.

General inverse consistency. Our work focused on spatial correspondences for registration, but the benefits of inverse consistency regularization are likely much broader. For instance, its applicability to general mapping problems (*e.g.*, between feature vectors) should be explored.

6. Conclusion

We presented a deliberately simple deep registration model which generates approximately diffeomorphic maps by regularizing via an inverse consistency loss. We theoretically analyzed why inverse consistency leads to spatial smoothness and empirically showed the effectiveness of our approach, yielding competitive 3D registration performance.

Our results suggest that simple deep registration networks might be as effective as more complex approaches which require substantial hyperparameter tuning and involve choosing complex transformation models. As a wide range of inverse consistency loss penalties lead to good results, only the desired similarity measure needs to be chosen and extensive hyperparameter tuning can be avoided. This opens up the possibility to easily train extremely fast custom registration networks on given data. Due to its simplicity, ease of use, and computational speed, we expect our approach to have significant practical impact. We also expect that inverse consistency regularization will be useful for other

tasks, which should be explored in future work.

Acknowledgments. This research was supported by NIH awards R21-CA223304, 1R01-AR072013, P30 CA008748; by NSF award EECS-1711776; by the Austrian Science Fund: FWF P31799-N38; and by the Land Salzburg (WISS 2025): 20102-F1901166-KZP, 20204-WISS/225/197-2019. The work expresses the views of the authors and not of the funding agencies. The authors have no conflicts of interest.

References

[1] Felix Ambellan, Alexander Tack, Moritz Ehlke, and Stefan Zachow. Automated segmentation of knee bone and cartilage combining statistical shape knowledge and convolutional neural networks: Data from the Osteoarthritis Initiative. In *MIDL*, 2018.

[2] 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.

[3] Riddhish Bhalodia, Shireen Y Elhabian, Ladislav Kavan, and Ross T Whitaker. A cooperative autoencoder for population-based regularization of CNN image registration. In *MICCAI*, 2019.

[4] John Blitzer, Ryan McDonald, and Fernando Pereira. Domain adaptation with structural correspondence learning. In *EMNLP*, 2006.

[5] Mathieu Blondel, Olivier Teboul, Quentin Berthet, and Josip Djolonga. Fast differentiable sorting and ranking. In *ICML*, 2020.

[6] Wen-Tsuen Chen Wen-Tsuen Chen and Kuen-Rong Hsieh Kuen-Rong Hsieh. A neural sorting network with O(1) time complexity. In *IJCNN*, 1990.

[7] Xinlei Chen and Kaiming He. Exploring simple Siamese representation learning. In *CVPR*, 2021.

[8] Gary E Christensen and Hans J Johnson. Consistent image registration. *IEEE Transactions on Medical Imaging*, 20(7):568–582, 2001.

[9] Adrian V Dalca, Guha Balakrishnan, John Guttag, and Mert R Sabuncu. Unsupervised learning for fast probabilistic diffeomorphic registration. In *MICCAI*, 2018.

[10] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *ICCV*, 2015.

[11] Martin Engilberge, Louis Chevallier, Patrick Pérez, and Matthieu Cord. Sodeep: a sorting deep net to learn ranking loss surrogates. In *CVPR*, 2019.

[12] Denis Fortun, Patrick Bouthemy, and Charles Kervrann. Optical flow modeling and computation: A survey. *Computer Vision and Image Understanding*, 134:1–21, 2015.

[13] Gabriel L Hart, Christopher Zach, and Marc Niethammer. An optimal control approach for deformable registration. In *CVPR Workshops*, pages 9–16, 2009.

[14] Mark Holden. A review of geometric transformations for nonrigid body registration. *IEEE Transactions on Medical Imaging*, 27(1):111–128, 2007.

[15] Berthold Horn, Berthold Klaus, and Paul Horn. *Robot vision*. MIT press, 1986.

[16] Berthold KP Horn and Brian G Schunck. Determining optical flow. *Artificial Intelligence*, 17(1-3):185–203, 1981.

[17] Stephan Hoyer, Jascha Sohl-Dickstein, and Sam Greydanus. Neural reparameterization improves structural optimization. *CoRR*, abs/1909.04240, 2020.

[18] Tak-Wai Hui, Xiaoou Tang, and Chen Change Loy. Liteflownet: A lightweight convolutional neural network for optical flow estimation. In *CVPR*, 2018.

[19] Junhwa Hur and Stefan Roth. Iterative residual refinement for joint optical flow and occlusion estimation. In *CVPR*, 2019.

[20] Sarang Joshi, Brad Davis, Matthieu Jomier, and Guido Gerig. Unbiased diffeomorphic atlas construction for computational anatomy. *NeuroImage*, 23:S151–S160, 2004.

[21] Boah Kim, Dong Hwan Kim, Seong Ho Park, Jieun Kim, June-Goo Lee, and Jong Chul Ye. Cyclemorph: Cycle consistent unsupervised deformable image registration. *CoRR*, abs/2008.05772, 2020.

[22] 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 *MICCAI*, 2019.

[23] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[24] Ivan Kobyzev, Simon Prince, and Marcus Brubaker. Normalizing flows: An introduction and review of current methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[25] Hamid Laga, Laurent Valentin Jospin, Farid Boussaid, and Mohammed Bennamoun. A survey on deep learning techniques for stereo-based depth estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[26] Moshe Leshno, Vladimir Ya Lin, Allan Pinkus, and Shimon Schocken. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Networks*, 6(6):861–867, 1993.

[27] Tie-Yan Liu. *Learning to rank for information retrieval*. Now publishers, 2011.

[28] Jan Modersitzki. *Numerical methods for image registration*. Oxford University Press on Demand, 2004.

[29] Marc Niethammer, Roland Kwitt, and Francois-Xavier Vialard. Metric learning for image registration. In *CVPR*, 2019.

[30] Jorge Nocedal and Stephen Wright. *Numerical optimization*. Springer Science & Business Media, 2006.

[31] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

[32] Anurag Ranjan and Michael J Black. Optical flow estimation using a spatial pyramid network. In *CVPR*, 2017.

[33] Laurent Risser, François-Xavier Vialard, Robin Wolz, Maria Murgasova, Darryl D Holm, and Daniel Rueckert. Simultaneous multi-scale registration using large deformation diffeomorphic metric mapping. *IEEE Transactions on Medical Imaging*, 30(10):1746–1759, 2011.

[34] Torsten Rohlfing. Image similarity and tissue overlaps as surrogates for image registration accuracy: Widely used but unreliable. *IEEE Transactions on Medical Imaging*, 31(2):153–163, 2012.

[35] Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. The pitfalls of simplicity bias in neural networks. In *NeurIPS*, 2020.

[36] Zhengyang Shen, Xu Han, Zhenlin Xu, and Marc Niethammer. Networks for joint affine and non-parametric image registration. In *CVPR*, 2019.

[37] Zhengyang Shen, François-Xavier Vialard, and Marc Niethammer. Region-specific diffeomorphic metric mapping. In *NeurIPS*, 2019.

[38] Zhixin Shu, Mihir Sahasrabudhe, Riza Alp Guler, Dimitris Samaras, Nikos Paragios, and Iasonas Kokkinos. Deforming autoencoders: Unsupervised disentangling of shape and appearance. In *ECCV*, 2018.

[39] Aristeidis Sotiras, Christos Davatzikos, and Nikos Paragios. Deformable medical image registration: A survey. *IEEE Transactions on Medical Imaging*, 32(7):1153–1190, 2013.

[40] Nicholas J Tustison, Brian B Avants, and James C Gee. Learning image-based spatial transformations via convolutional neural networks: A review. *Magnetic Resonance Imaging*, 64:142–153, 2019.

[41] Xiao Yang, Roland Kwitt, and Marc Niethammer. Fast predictive image registration. In *Deep Learning and Data Labeling for Medical Applications*, 2016.

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

[43] Jun Zhang. Inverse-consistent deep networks for unsupervised deformable image registration. *CoRR*, abs/1809.03443, 2018.

[44] Pan Zhang, Bo Zhang, Dong Chen, Lu Yuan, and Fang Wen. Cross-domain correspondence learning for exemplar-based image translation. In *CVPR*, 2020.

[45] Lilla Zöllei, Erik Learned-Miller, Eric Grimson, and William Wells. Efficient population registration of 3D data. In *International Workshop on Computer Vision for Biomedical Image Applications*, 2005.