This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

CARL: A Framework for Equivariant Image Registration

Hastings Greer[1] Lin Tian[1] François-Xavier Vialard[2,3] Roland Kwitt[4] Raúl San José Estépar[5]
Marc Niethammer[1]

[1]UNC Chapel Hill [2]LIGM, Université Gustave Eiffel [3]MOKAPLAN, INRIA Paris
[4]University of Salzburg [5]Harvard Medical School

Abstract

Image registration estimates spatial correspondences between image pairs. These estimates are typically obtained via numerical optimization or regression by a deep network. A desirable property is that a correspondence estimate (e.g., the true oracle correspondence) for an image pair is maintained under deformations of the input images. Formally, the estimator should be equivariant to a desired class of image transformations. In this work, we present careful analyses of equivariance properties in the context of multi-step deep registration networks. Based on these analyses we 1) introduce the notions of $[U, U]$ equivariance (network equivariance to the same deformations of the input images) and $[W, U]$ equivariance (where input images can undergo different deformations); we 2) show that in a suitable multi-step registration setup it is sufficient for overall $[W, U]$ equivariance if the first step has $[W, U]$ equivariance and all others have $[U, U]$ equivariance; we 3) show that common displacement-predicting networks only exhibit $[U, U]$ equivariance to translations instead of the more powerful $[W, U]$ equivariance; and we 4) show how to achieve multi-step $[W, U]$ equivariance via a coordinate-attention mechanism combined with displacement-predicting networks. Our approach obtains excellent practical performance for 3D abdomen, lung, and brain medical image registration. We match or outperform state-of-the-art (SOTA) registration approaches on all the datasets with a particularly strong performance for the challenging abdomen registration.

1. Introduction

Image registration is an important task in medical image computing [39]. There is extensive research into solving this problem with deep learning [2, 27, 36, 45]. In particular, unsupervised deep learning methods [12, 43] have attained strong performance on many benchmarks [19]. We present an equivariant registration layer that can be inserted into an unsupervised multi-step registration network, which makes the overall registration architecture equivariant to independently translated images.

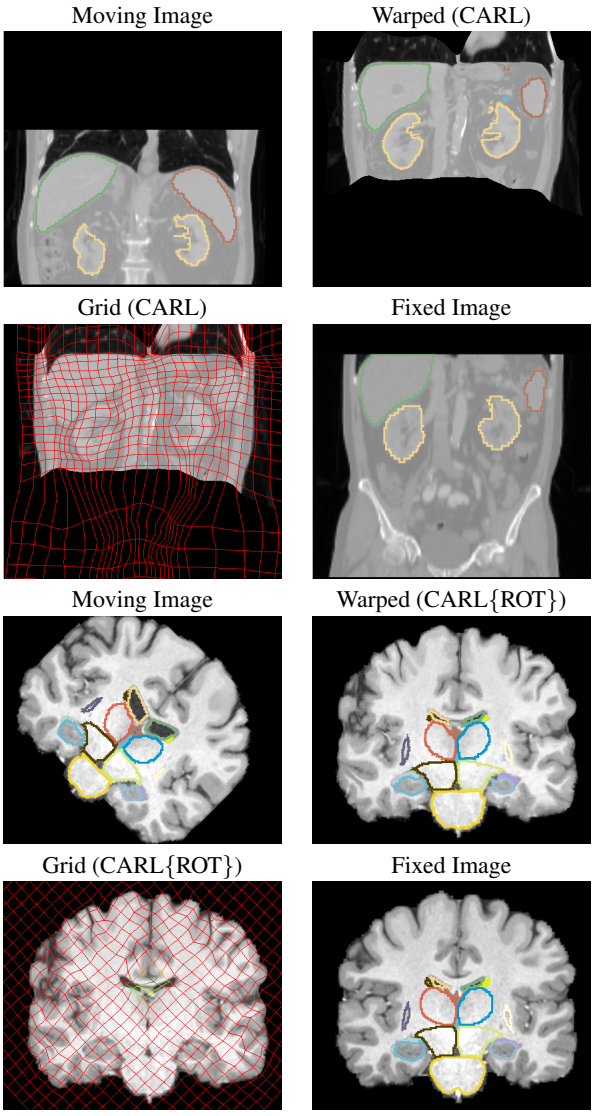


Figure 1. Results on a pair from the Abdomen1k dataset, and a synthetically rotated example from the HCP dataset. CARL tackles large displacements and arbitrary rotations via equivariance.

Notation. We denote the *moving* and *fixed* images as I^{M} and I^{F}, respectively. They are functions $\Omega \to \mathbb{R}^D$, where Ω is the image domain, and D is typically 1 for a single-channel imaging modality. We denote a free *transform* with a lowercase Greek letter such as φ. Further, we denote a *registration algorithm* as a capital Greek letter, such as Φ, Ξ, or Ψ. We note that neural networks that perform registration are a subset of registration algorithms. In our notation, a registration algorithm can be applied to a pair of images by square brackets, *i.e.*, $\Phi\left[I^{\mathrm{M}}, I^{\mathrm{F}}\right]$. Hence, when applied to a pair of images, a registration algorithm yields a map $\mathbb{R}^N \to \mathbb{R}^N$ (ideally a diffeomorphism) that can be post-composed with the moving image to yield a function that is close to the fixed image, *i.e.*,

$$I^{\mathrm{M}} \circ \Phi\left[I^{\mathrm{M}}, I^{\mathrm{F}}\right] \sim I^{\mathrm{F}}. \tag{1}$$

With this notation, we can define what it means for a registration algorithm to be *equivariant*. Conceptually, equivariance means that when the input images are warped, correspondence between the registered images is preserved; it is not required that the correspondence is *correct*, only that it is *the same* before and after applying a warp to the input images. We say a registration algorithm Φ is $[W, U]$ equivariant with respect to a class of transforms $\mathcal{T}$ if correspondence is preserved when the images are warped *independently*. Formally, this means

$$\begin{aligned} &\forall\, W, U \in \mathcal{T}: \qquad (2)\\ &\Phi\left[I^{\mathrm{M}} \circ W, I^{\mathrm{F}} \circ U\right] = W^{-1} \circ \Phi\left[I^{\mathrm{M}}, I^{\mathrm{F}}\right] \circ U\,. \end{aligned}$$

As a weaker form of equivariance, we say a registration algorithm is $[U, U]$ equivariant if correspondence is preserved when the images are warped *together*, *i.e.*,

$$\begin{aligned} &\forall\, U \in \mathcal{T}: \qquad (3)\\ &\Phi\left[I^{\mathrm{M}} \circ U, I^{\mathrm{F}} \circ U\right] = U^{-1} \circ \Phi\left[I^{\mathrm{M}}, I^{\mathrm{F}}\right] \circ U\,. \end{aligned}$$

From our analysis (see Sec. 3), it will be clear that both definitions are well justified.

2. Related work

We review related work at the intersection of equivariance (and how to achieve it) and image registration, as well as previous work on unsupervised learning of neural networks for image registration.

Equivariant encoders. The most impactful equivariant layers in the field of neural networks have been ordinary convolution [10] and attention [30], which are equivariant to translation and permutation, respectively. In this work, we focus on using attention and convolution to make a registration algorithm that is by construction $[W, U]$ equivariant to translation. However, there are also more recent approaches to create components with additional geometric equivariances such as to rotation. These include group equivariant convolution [5], StyleGAN's equivariant convolution [22], and spherical harmonic convolution [28]. These should be explored in future work to expand the deformations with respect to which our approach (based on what we call *coordinate attention*, see Sec. 4) is $[W, U]$ equivariant.

Keypoint-based equivariant registration architectures. Several registration approaches relying on keypoint matching have been proposed [3, 24]. Notably, any approach that aligns keypoints to keypoints using a least squares fit is $[W, U]$ equivariant to rigid motions if the keypoint selectors and descriptors are equivariant to rigid motions. The reason can be derived from the notion of equivariance as follows: *if two points correspond before warping the input images, they should correspond afterwards*. The keypoints align with each other before and after a warp is applied, and move with features on the input images when they are warped. SURF [3] and related methods fall under this category in the subfield of non-learning-based 2D image alignment. For medical image registration, learning-based approaches that leverage keypoints include EasyReg [21], KeyMorph [46], SAME++ [37] and E-CNN [4]. SAME++ [37] uses a pre-trained foundation model as a convolutional encoder, and extracts matched keypoints from it using methods traditionally used in 2D image feature alignment, such as checking for cycle consistency. This alignment is then refined using a trained deep network with overall strong performance on lung, abdomen, and head/neck computed tomography (CT) registration tasks. EasyReg, E-CNN and KeyMorph achieve equivariant registration by locating a fixed set of keypoints in each image and aligning matching keypoints. EasyReg's [21] affine pre-registration is by construction $[W, U]$ equivariant to affine deformations because it works by segmenting the brain images to be registered, and then bringing the centroids of the segmented regions into alignment. In effect, this uses each segmented region as a keypoint. In our experiments (see Sec. 6) we will see that this strategy works well even for out-of-distribution translations, and exhibits strong performance when detailed segmentations are available for training. In the KeyMorph work of [43], equivariance is achieved in an unsupervised setting by learning to predict keypoints in brain images. Unlike SURF and related approaches, KeyMorph avoids the need for a matching step by predicting keypoints in a fixed order. Finally, in the E-CNN work of [4], the authors leverage a rotationally equivariant encoder built using spherical harmonics to predict features, and the centers of mass of each feature channel are then used as keypoints and rigidly aligned in a differentiable manner. The network is trained by warping a single image with two different transforms and then penalizing the difference between the network output and the transform used to generate the pair.

Equivariance losses. Several works aim for equivariance via suitably designed loss terms. In the CoMIR work of [29], *e.g.*, the authors train an image encoder to be equivariant to rotations and deformations using an unsupervised contrastive loss with additional constraints. Furthermore, the concept of W-bipath consistency, introduced in [40], is effectively a penalty to encourage $[W, U]$ equivariance with respect to B-Spline transforms. In particular, [40] uses penalties to enforce that $\Phi[I^{\mathrm{M}}, I^{\mathrm{F}} \circ U] = \Phi[I^{\mathrm{M}}, I^{\mathrm{F}}] \circ U$ as well as inverse consistency, which together are sufficient conditions for $[W, U]$ equivariance.

Unsupervised image registration. In addition to prior work on equivariance, we also build on research into unsupervised image registration. The foundational concept [2] driving this subfield is to predict deformations using a neural network. Training minimizes a similarity and a regularity loss. It is feasible to learn multiple steps of registration by simply penalizing the final similarity and regularity [11, 33, 36], even when the steps have different transform models [12, 33]. Diffeomorphisms can be guaranteed by special vector field integration layers [6], inversion layers [20] or *encouraged* by a loss, such as diffusion [2, 33], or inverse consistency [11, 36]. Accuracy can be improved by learning features, computing cost volumes from the similarity between features in the fixed and moving images, and predicting a transform based on this cost volume [27]. From this subfield we adopt most of GradICON [8, 36, 38]: we use their loss and regularity hyperparameters to produce approximately diffeomorphic transforms, adopt their learning rate and multi-step, multi-resolution approach, and use their U-Net architecture for the *refinement layers* of CARL. While replacing *Step 1* of their approach with our custom architecture Ξ_θ might appear minor, it has profound consequences which we will work out in Secs. 3-5.2. In short, this replacement makes the entire network $[W, U]$ equivariant with respect to translation, whereas GradICON and most of the convolution-based deep registration networks only provide the weaker $[U, U]$ equivariance.

3. Registering vector-valued images that are restricted to be diffeomorphisms

In Sec. 1, we defined equivariance for registration as preserving the correspondence between points in the images if they are warped before being registered. To obtain a mathematical definition, we investigate a specific class of images I^{M} and I^{F} where the correspondence is well defined: we let them have $[0, 1]^D$ as their domain and range, and require them to be *diffeomorphisms*. While this restricted problem does not correspond to a real-world medical registration task, it allows a closed form solution, yielding insights that can be carried back to real registration problems where finding correspondences is not trivial. Because I^{M} is *invertible*, we can make Eq. (1) an *equality* and then solve for the registration algorithm, which we denote as Ξ:

$$\begin{aligned} & I^{\mathrm{M}} \circ \Xi\left[I^{\mathrm{M}}, I^{\mathrm{F}}\right] = I^{\mathrm{F}} \\ \Leftrightarrow \left(I^{\mathrm{M}}\right)^{-1} \circ I^{\mathrm{M}} \circ \Xi\left[I^{\mathrm{M}}, I^{\mathrm{F}}\right] &= \left(I^{\mathrm{M}}\right)^{-1} \circ I^{\mathrm{F}} \\ \Leftrightarrow \Xi\left[I^{\mathrm{M}}, I^{\mathrm{F}}\right] &= \left(I^{\mathrm{M}}\right)^{-1} \circ I^{\mathrm{F}}. \end{aligned} \tag{4}$$

Since I^{M} and I^{F} are images, and for these images Ξ is correct, a hypothetical registration algorithm that registers any pair of images correctly will output the same transforms as Ξ in situations where Ξ is valid, and hence will tend to share the equivariances of Ξ. In particular, Ξ is equivariant to the group $\mathrm{Diff}(\mathbb{R}^N) \times \mathrm{Diff}(\mathbb{R}^N)$ as follows. Let U and W be arbitrary members of $\mathrm{Diff}(\mathbb{R}^N)$ (diffeomorphisms), then we obtain the following commutative diagram:

$$\begin{array}{ccc} (I^{\mathrm{M}}, I^{\mathrm{F}}) & \xrightarrow{x, y \mapsto x \circ W, y \circ U} & (I^{\mathrm{M}} \circ W, I^{\mathrm{F}} \circ U) \\ \Xi \downarrow & & \downarrow \Xi \\ (I^M)^{-1} \circ I^F & \xrightarrow[\varphi \mapsto W^{-1} \circ \varphi \circ U]{} & W^{-1} \circ (I^M)^{-1} \circ I^F \circ U \end{array} \tag{5}$$

Now, if we have a new registration algorithm Φ, a class of input images and potential transforms, we can check if that particular network is equivariant by checking if for all W, U a diagram matching Eq. (5) holds (if so, the registration algorithm is $[W, U]$ equivariant w.r.t. diffeomorphisms):

$$\begin{array}{ccc} (I^{\mathrm{M}}, I^{\mathrm{F}}) & \xrightarrow{x, y \mapsto x \circ W, y \circ U} & (I^{\mathrm{M}} \circ W, I^{\mathrm{F}} \circ U) \\ \Phi \downarrow & & \downarrow \Phi \\ \Phi[I^M, I^F] & \xrightarrow[\varphi \mapsto W^{-1} \circ \varphi \circ U]{?} & W^{-1} \circ \Phi[I^M, I^F] \circ U \end{array} \tag{6}$$

In other words, if the registration algorithm results in φ for registering I^{M} to I^{F}, then it needs to be constructed such that it results in $W^{-1} \circ \varphi \circ U$ if images I^{M} and I^{F} are warped by W and U, respectively. In that case, W and U will not change the image correspondences.

In practice, we will often restrict the group from which W and U may be chosen. For example, if for a registration algorithm Φ, for any translations W and U, $\Phi[I^{\mathrm{M}} \circ W, I^{\mathrm{F}} \circ U] = W^{-1} \circ \Phi[I^M, I^F] \circ U$, we say that the algorithm Φ is $[W, U]$ equivariant with respect to translations.

Finally, we remark that some registration algorithms are only equivariant for applying the same transform to both images; that is, for any transform U from a group of transforms $\mathcal{T}$, $\Phi[I^{\mathrm{M}} \circ U, I^{\mathrm{F}} \circ U] = U^{-1} \circ \Phi[I^M, I^F] \circ U$. We say that such an algorithm is $[U, U]$ equivariant with respect to $\mathcal{T}$. This justifies the definitions we previewed in Sec. 1.

4. Coordinate attention

Our central contribution when seeking to realize $[W, U]$ equivariance is what we call *coordinate attention*, *i.e.*, a

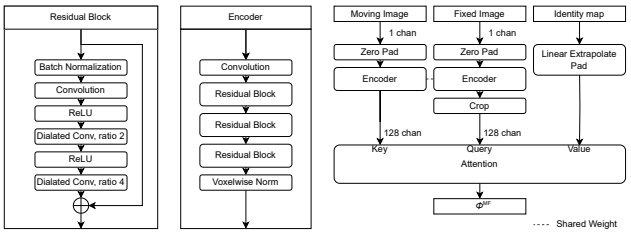


Figure 2. Architecture of Ξ_θ. The specific arrangement of pads and crops allows voxels to be mapped to points outside of Ω – this is necessary to represent translation.

standard attention block where the value vectors are the coordinates of each voxel. This has the effect of computing the center of mass of the attention mask associated with each query. As an illustrative example in Appendix **??**, we use coordinate attention to solve the diffeomorphism-to-diffeomorphism registration problem.

For registering medical images, we use an architecture we call Ξ_θ, a combination of coordinate attention with a feature encoder consisting of convolutional layers all operating at the same resolution, some of which are dilated. The encoder is exactly equivariant to integer pixel translations. Edge effects are reduced by padding with zeros before running images through the feature encoder, and cropping afterwards. The architecture is illustrated in Fig. 2. It is important that the coordinate attention formulation, which is "center of mass of softmax$(QV^\top)$", can be computed using a standard attention operation instead of a custom set of operations on all pairs of feature vectors. Attention has heavily optimized implementations such as flash attention [7] that only require $\mathcal{O}$(#voxels) memory and have excellent cache coherency. While flash attention is $\mathcal{O}$(image side length^6), in practice performance is adequate in our setting when Ξ_θ is applied to a volume of size $43 \times 43 \times 43$, which is sufficient for coarse registration. We will show (in Sec. 5.2) that it is then possible to add fine-grained registration while preserving $[W, U]$ equivariance.

5. Equivariance of Coordinate Attention with convolutional encoders (Ξ_θ)

We assume that the attention mask associated with each query vector has small spatial support. Finding a training procedure that reliably fulfilled this assumption across different datasets was nontrivial: we find that this assumption is satisfied after regularizing the network end-to-end with diffusion regularization for the first several epochs, and using GradICON regularization thereafter.

We assume that the feature encoders are translation equivariant like

$$\mathrm{Conv}_\theta(I \circ U) = \mathrm{Conv}_\theta(I) \circ U\,. \tag{7}$$

With these assumptions, we prove that Ξ_θ is $[W, U]$ equivariant to translations below.

Without positional embeddings or causal masking, (we do not use either) the attention mechanism is equivariant to permutations as follows: for P_1, P_2 permutations; and the output and K (Key), Q (Query), and V (Value) inputs represented each as a function from an index to a vector, and an attention block represented as $\mathbb{T}$,

$$\mathbb{T}[K \circ P_1, Q \circ P_2, V \circ P_1] = \mathbb{T}[K, Q, V] \circ P_2\,. \tag{8}$$

Additionally, because the attention weights in an attention block sum to 1, for an affine function f, we have

$$\mathbb{T}[K, Q, f \circ V] = f \circ \mathbb{T}[K, Q, V]\,. \tag{9}$$

A translation W by an integer number of voxels is both affine when seen as an operation on coordinates, $W_{x \mapsto x+r}$, and a permutation of the voxels when seen as an operation on voxel images $W_{\text{permutation}}$ (as long as we can neglect boundary effects). The map from indices to coordinates, coords, serves as the bridge between these two representations of a transform ($W_{x \mapsto x+r} \circ \text{coords} = \text{coords} \circ W_{\text{permutation}}$). As long as the attention masks have small spatial support, we can suppress boundary effects by padding with zeros before applying the operation. So, for translations W and U, we have

$$\Xi_\theta[I^M, I^F] := \mathbb{T}[\mathrm{Conv}_\theta(I^M), \mathrm{Conv}_\theta(I^F), \text{coords}]\ ,$$

from which we establish that Ξ_θ is $[W, U]$ equivariant with respect to translation as follows:

$$\begin{aligned}
&\Xi_\theta[I^M \circ W, I^F \circ U] \\
&= \mathbb{T}[\mathrm{Conv}_\theta(I^M \circ W), \mathrm{Conv}_\theta(I^F \circ U), \text{coords}] \\
&\overset{(7)}{=} \mathbb{T}[\mathrm{Conv}_\theta(I^M) \circ W, \mathrm{Conv}_\theta(I^F) \circ U, \text{coords}] \\
&\overset{(9)}{=} W^{-1} \circ \mathbb{T}[\mathrm{Conv}_\theta(I^M) \circ W, \mathrm{Conv}_\theta(I^F) \circ U, W \circ \text{coords}] \\
&= W^{-1} \circ \mathbb{T}[\mathrm{Conv}_\theta(I^M) \circ W, \mathrm{Conv}_\theta(I^F) \circ U, \text{coords} \circ W] \\
&\overset{(8)}{=} W^{-1} \circ \mathbb{T}[\mathrm{Conv}_\theta(I^M), \mathrm{Conv}_\theta(I^F) \circ U, \text{coords}] \\
&\overset{(8)}{=} W^{-1} \circ \mathbb{T}[\mathrm{Conv}_\theta(I^M), \mathrm{Conv}_\theta(I^F), \text{coords}] \circ U \\
&= W^{-1} \circ \Xi_\theta[I^M, I^F] \circ U\,.
\end{aligned} \tag{10}$$

The same argument can also be applied to $[W, U]$ equivariance to axis aligned π or $\frac{\pi}{2}$ rotations, provided that Conv_θ is replaced with an appropriate rotation equivariant encoder.

5.1. $[U, U]$ equivariance

So far, we have pointed out that Ξ_θ is $[W, U]$ equivariant to translation. Our next goal is to show that our full approach, *Coordinate Attention with Refinement Layers* (CARL), is

also $[W, U]$ equivariant to translation. To this end, we (1) show that VoxelMorph-like networks are $[U, U]$ equivariant to translation as they predict displacements, and then (2) show that a multi-step registration algorithm where the first step is $[W, U]$ equivariant and subsequent steps are $[U, U]$ equivariant is overall $[W, U]$ equivariant. Notably, any all-convolutional network is translationally equivariant[1]:

$$\begin{array}{ccc} X & \xrightarrow{f \mapsto f \circ U} & \tilde{X} \\ \mathrm{Conv}_\theta \downarrow & & \downarrow \mathrm{Conv}_\theta \\ Y & \xrightarrow[f \mapsto f \circ U]{} & \tilde{Y} \end{array} \tag{11}$$

To apply a convolutional network to registration, first it is made to have two inputs by channel-wise concatenation (cat), and it is chosen to have D output channels. Its output can then be interpreted (via interpolation) as a function from the domain of the input images to $\mathbb{R}^N$. If used to predict coordinates, this has a different equivariance than $[U, U]$ equivariance: to be $[U, U]$ equivariant with respect to translation, a registration algorithm needs to have a bottom arrow $f \mapsto U^{-1} \circ f \circ U$ where Eq. (11) has $f \mapsto f \circ U$. This is ameliorated in Quicksilver [45], VoxelMorph [2] and their successors by predicting displacements instead of coordinates. Overall, in VoxelMorph we have

$$\Phi_\theta[I^{\mathrm{M}}, I^{\mathrm{F}}](\vec{x}) := \mathrm{Conv}_\theta[\mathrm{cat}(I^{\mathrm{M}}, I^{\mathrm{F}})](\vec{x}) + \vec{x} \ , \tag{12}$$

and by parameterizing our translation transform U by a vector $\vec{r}$, we can compute

$$\begin{aligned} &\Phi_\theta[I^{\mathrm{M}} \circ U, I^{\mathrm{F}} \circ U](\vec{x}) \\ &\quad = \mathrm{Conv}_\theta[\mathrm{cat}(I^{\mathrm{M}} \circ U, I^{\mathrm{F}} \circ U)](\vec{x}) + \vec{x} \\ &\quad = \mathrm{Conv}_\theta[\mathrm{cat}(I^{\mathrm{M}}, I^{\mathrm{F}}) \circ U](\vec{x}) + \vec{x} \\ &\quad = \left(\mathrm{Conv}_\theta[\mathrm{cat}(I^{\mathrm{M}}, I^{\mathrm{F}})] \circ U\right)(\vec{x}) + \vec{x} \\ &\quad = \mathrm{Conv}_\theta[\mathrm{cat}(I^{\mathrm{M}}, I^{\mathrm{F}})](U(\vec{x})) + \vec{x} \\ &\quad = \mathrm{Conv}_\theta[\mathrm{cat}(I^{\mathrm{M}}, I^{\mathrm{F}})](U(\vec{x})) + (\vec{x} + \vec{r}) - \vec{r} \\ &\quad = U^{-1} \circ \left(\mathrm{Conv}_\theta[\mathrm{cat}(I^{\mathrm{M}}, I^{\mathrm{F}})](\vec{x}) + \vec{x}\right) \circ U \\ &\quad = U^{-1} \circ \Phi_\theta[I^{\mathrm{M}}, I^{\mathrm{F}}](\vec{x}) \circ U \ . \end{aligned} \tag{13}$$

5.2. Two-step registration

Multi-step registration is common practice, often using multiple algorithms. For example, ANTs [1] by default performs affine, then deformable registration, and returns the composition of the resulting transforms. Greer *et al.* [11] propose a general notation for this scheme: the TwoStep operator, which takes as input two registration algorithms, and yields a composite algorithm as follows:

$$\begin{aligned} &\mathrm{TwoStep}\{\Phi, \Psi\}[I^M, I^F] \\ &\qquad = \Phi[I^M, I^F] \circ \Psi[I^M \circ \Phi[I^M, I^F], I^F] \, . \end{aligned} \tag{14}$$

We show that if Φ is $[W, U]$ equivariant and Ψ is $[U, U]$ equivariant for U, W from a class of transforms $\mathcal{T}$, then $\mathrm{TwoStep}\{\Phi, \Psi\}$ is $[W, U]$ equivariant. In particular, application of the definition of Φ being $[W, U]$ equivariant, see Eq. (2), and Ψ being $[U, U]$ equivariant, see Eq. (3), yields

$$\begin{aligned} &\mathrm{TwoStep}\{\Phi, \Psi\}[I^M \circ W, I^F \circ U] \\ &= \Phi\left[I^M \circ W, I^F \circ U\right] \circ \\ &\quad \Psi\left[I^M \circ W \circ \Phi[I^M \circ W, I^F \circ U], I^F \circ U\right] \\ &\overset{(2)}{=} W^{-1} \circ \Phi[I^M, I^F] \circ U \circ \\ &\quad \Psi\left[I^M \circ \Phi[I^M, I^F] \circ U, I^F \circ U\right] \\ &\overset{(3)}{=} W^{-1} \circ \Phi[I^M, I^F] \circ \\ &\quad \Psi\left[I^M \circ \Phi[I^M, I^F], I^F\right] \circ U \\ &= W^{-1} \circ \mathrm{TwoStep}\{\Phi, \Psi\}[I^M, I^F] \circ U \ . \end{aligned} \tag{15}$$

This construction can be applied recursively for an N-step $[W, U]$ equivariant network, and can be combined with the downsampling operator (abbreviated as Down) from [36] (see Appendix **??**) to make the early stages operate at coarse resolution and subsequent ones at higher resolutions.

6. Experiments

Architecture. We propose the architecture *Coordinate Attention with Refinement Layers* (CARL),

$$\begin{aligned} &\mathrm{CARL} := \mathrm{TwoStep}\{\mathrm{TwoStep}\{\mathrm{Down}\{\mathrm{TwoStep}\{ \\ &\qquad \mathrm{Down}\{\Xi_\theta\}\}, \Psi_1\}\}, \Psi_2 \\ &\}, \Psi_3\} \end{aligned} \tag{16}$$

where Ξ_θ is a $[W, U]$ equivariant neural network composed of a translation-equivariant feature encoder and a coordinate attention layer as proposed, and Ψ_i are refinement layers: convolutional networks predicting displacements with architecture and feature counts from [11, 36], and so are $[U, U]$ equivariant. To facilitate a direct comparison of performance, the specific arrangement of TwoStep and Downsample layers is identical to GradICON: our only change is to replace the lowest resolution displacement predicting network with Ξ_θ. Following the suggestion of GradICON [36], the final refinement layer Ψ_3 is only added after the first 50,000 steps of training. *The overall network is* $[W, U]$ *equivariant with respect to translation.*

Losses. In all experiments, we train with LNCC image similarity (LNCC=1-local normalized cross correlation). For

[1] At least, for translations of integer multiples of the least common multiple of all convolution strides, and considering the input and output of the convolution to be functions.

regularization, we train initially with a diffusion loss for 1,500 steps, *i.e.*,

$$\begin{aligned}\mathcal{L} &= \text{LNCC}\left(I^{\text{M}} \circ \text{CARL}[I^{\text{M}}, I^{\text{F}}], I^{\text{F}}\right) \\ &+ \text{LNCC}\left(I^{\text{F}} \circ \text{CARL}[I^{\text{F}}, I^{\text{M}}], I^{\text{M}}\right) \\ &+ \lambda||\nabla(\text{CARL}[I^{\text{M}}, I^{\text{F}}]) - \mathbf{I}||_F^2\end{aligned} \tag{17}$$

and then continue training with the GradICON [36] regularizer for 100,000 steps, *i.e.*,

$$\begin{aligned}\mathcal{L} &= \text{LNCC}(I^{M} \circ \text{CARL}[I^{\text{M}}, I^{\text{F}}], I^{\text{F}}) \\ &+ \text{LNCC}(I^{F} \circ \text{CARL}[I^{\text{F}}, I^{\text{M}}], I^{M}) \\ &+ \lambda||\nabla(\text{CARL}[I^{\text{M}}, I^{\text{F}}] \circ \text{CARL}[I^{\text{F}}, I^{\text{M}}]) - \mathbf{I}||_F^2\,,\end{aligned} \tag{18}$$

where $\mathbf{I}$ is the identity matrix, and $\lambda > 0$. We use regularization parameter $\lambda = 1.5$ and an Adam optimizer with a learning rate of 0.0001. We remark that diffusion regularization is needed in early steps of training for stability, but GradICON delivers superior final regularity and correspondence by strongly forcing the network towards invertibility while still allowing large deformations.

Instance Optimization (IO). Following GradICON [36], we evaluate with and without 50 steps of Instance Optimization, *i.e.*, since our method does not use human annotations at train time, at deployment we can further optimize Eq. (18) on instances of test data to improve performance.

6.1. Deformed retina images

We first demonstrate the importance of several details of our implementation on a dataset consisting of synthetically warped retina segmentations from [36]. In particular, for this experiment, we train three network variants:

1) TwoStep$\{$Down$\{$TwoStep$\{$Down$\{\Psi_0\}, \Psi_1\}\}, \Psi_2\}$, a multi-step network with each step predicting a displacement field. This is the 1st stage architecture proposed in GradICON [36]. We expect this network to *not* be $[W, U]$ equivariant w.r.t. translation.
2) Down$\{$Down$\{\Xi_\theta\}\}$, a single-step coordinate attention network . We expect this network to be $[W, U]$ equivariant w.r.t translation, but less powerful than 1).
3) CARL, Eq. (16). We expect this network to be powerful and $[W, U]$ equivariant to translation.

Each is assessed *without* instance optimization. To assess differences between these network types, we create three variants of the retina dataset:

1) **Baseline**. The train set is constructed by warping 14 segmentations from the DRIVE dataset [35], and the test set is constructed by warping the remaining 6 segmentations, with elastic warp parameters as suggested in [36].
2) **Shift**. The same as Baseline, but with fixed images shifted by 200 pixels ($\sim \frac{1}{3}$ of retina diameter).
3) **Scale Shift**. The same as Shift, but with fixed images also scaled to 80% size (See Fig. 3).

Results for this experiment are shown in Fig. 3. We see that when training/testing on the **Baseline** dataset, as expected, all three networks have adequate performance, although the importance of the refinement layers is highlighted when Down$\{$Down$\{\Xi_\theta\}\}$ falls meaningfully behind. When trained and tested on the **Shift** variant, the equivariant networks succeed, while the displacement predicting network fails. Additionally, we see that the equivariant networks *generalize* to the **Shift** test dataset when trained only on **Baseline**. Finally, we see that the equivariant networks also learn on, *and generalize to*, the **Scale Shift** dataset. This result illustrates an important point that we did not prove but observe empirically: while Ξ_θ is formally only $[W, U]$ equivariant to integer translations, in practice it is $[W, U]$ equivariant to deformations that are locally approximately translations, such as scaling. We do see a large variance **between training runs** when training CARL on the **Shift** and **Scale Shift** datasets. That is, specifically when training on Shift or Scale Shift, most trained CARL networks have high performance across all test examples, but for a small fraction of training runs, performance is poor across the whole test set.

6.2. Performance comparison to other methods

We compare our method to existing approaches from the literature on three datasets. First, we evaluate on the 10 DirLab COPDGene test cases as this is a popular and difficult lung registration benchmark that allows us to compare our performance to diverse and highly optimized methods. Second, we evaluate on the HCP brain dataset, which allows us to directly compare to two existing equivariant registration methods, EasyReg and KeyMorph. Third, we evaluate on a novel registration task: registering images from the Abdomen1k dataset without registration-specific preprocessing. This dataset has large translational differences and different fields of view, which causes other unsupervised registration approaches to converge poorly. Our strong equivariance makes this task possible with the LNCC + regularization loss. We aim to present a picture of how our approach will work when applied as specified to new datasets. To this end, *we use the same hyperparameters for all three datasets*.

Lung registration. To evaluate CARL against the current state of the art for general registration, we use the DirLab lung registration task, which is well standardized and has gained considerable attention in the community. In particular, we train on 999 inhale/exhale CT pairs of images from the COPDGene dataset, and evaluate based on landmark distance in the 10 standard DirLab pairs. We follow the data preparation and augmentation suggested in GradICON [36]. Results are listed in Table 1, showing the commonly-used mean target registration error (mTRE) score as well as the percentage of negative values for the Jacobian determinant

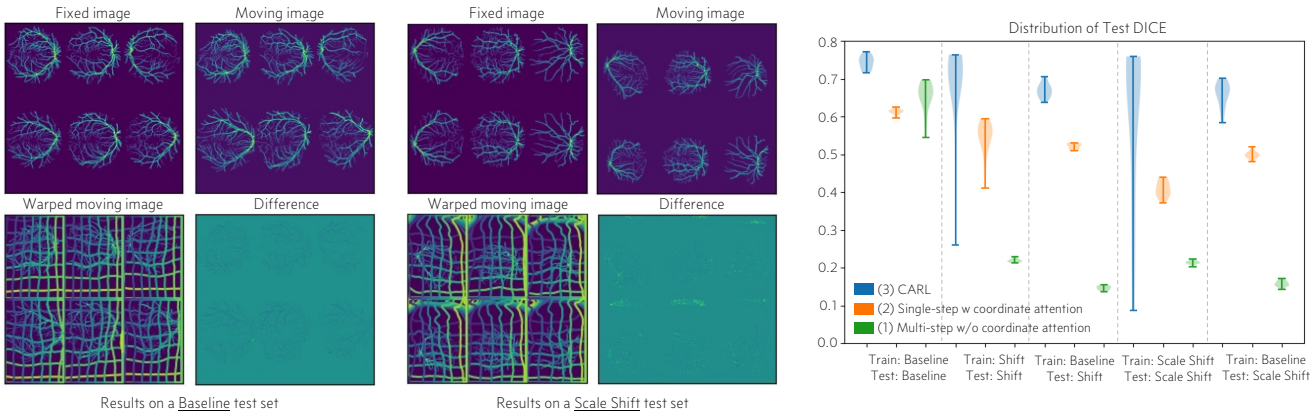


Figure 3. Equivariance allows CARL to generalize out-of-distribution. The *left four images* show the performance of CARL on a test set with the same distribution as the training set, where images are *aligned* in scale and translation. The *middle four images* show generalization to a test set **Scale Shift** where images are *misaligned* in scale and translation. The *right-most* figure shows the mean Dice distribution on the test set, over multiple training runs.

($\%|J|_{<0}$), indicative of the number of folds. Performance matches state of the art, exceeded significantly only by PTVReg and RRN which have hyper parameters selected specifically for the 10 challenge cases.

Brain registration. The existing equivariant registration methods that we compare to are presented as brain registration methods. We compare to them on T1w MRI images from the Human Connectome Project (HCP) dataset [31, 32, 41]. In detail, we perform evaluation and preprocessing as in [12], measuring DICE of midbrain structures **??** and find that CARL yields state-of-the-art performance. We also artificially translate or rotate one image by varying amounts and re-evaluate performance, verifying that GradICON sharply drops in performance under this shift, while our approach and EasyReg are unaffected (see Fig. 4).

Abdomen registration. Abdominal CT-CT datasets can capture the diversity of fields of view of abdominal CT scans in clinical practice, and the Abdomen1k [25] dataset is curated specifically to do so. Since our method is robust to these differences of field of view because it is $[W, U]$ equivariant to translation, we showcase its performance for inter-subject registration of the Abdomen1K [25] dataset. We train on the first 800 cases of the Abdomen1K dataset, validate on cases 900-1000, and test on cases 800-900. We opt for a minimalist preprocessing: we resample all images to [175, 175, 175] [2] and then divide by the maximum intensity. We test by randomly sampling 30 image pairs and then computing the mean Dice score over manual annotations of the liver, kidney, pancreas, and spleen. We report (Appendix, **??**) the exact pairs to allow future works to compare to our result. The most important comparison is to GradICON, as our final network, similarity and regularity loss are identical to GradICON's except that we replace the first U-Net based Ψ with our custom Ξ_θ. This change produces a dramatic improvement in non-instance-optimization registration accuracy and an improvement in instance-optimized registration accuracy. Results are listed in Table 1. We also check whether the model trained on the Abdomen1k dataset generalizes to the Learn2Reg Abdomen CT-CT challenge set [19]. We find excellent generalization: performance is better than other unsupervised approaches including the keypoint-based SAME++, but does not match the best supervised methods trained using the Learn2Reg CT-CT training set *and segmentations*.

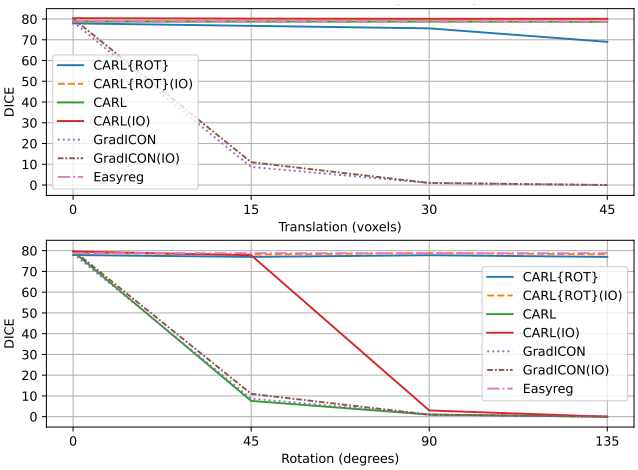


Figure 4. HCP evaluation while translating or rotating one image. CARL and EasyReg are unaffected by translation due to $[W, U]$ equivariance, and CARL{ROT} is additionally unaffected by rotation. GradICON DICE drops significantly when images are transformed differently due to its $[U, U]$ equivariance.

[2] This leads to nonisometric voxel spacing. We tried isometric voxel spacing, padding with background pixels if necessary, but non-isometric, non padded preprocessing had uniformly better performance

Table 1. Performance comparison on Abdomen1K, HCP, DirLab, and Learn2Reg Abdomen CT-CT dataset.

Method	%DICE↑	$\%\lvert J\rvert_{<0}\downarrow$
Abdomen1K		
ANTs [1]	45.4	0
VoxelMorph [2]	59.3	-
GradICON [36]	49.6	7e-1
GradICON(IO) [36]	71.0	4e-2
ConstrICON [12]	65.2	7e-4
ConstrICON(IO) [12]	66.8	2e-3
uniGradICON [38]	43.5	9e-1
uniGradICON(IO) [38]	54.8	2e-3
CARL	75.7	4e-1
CARL(IO)	77.2	2e-3
HCP		
ANTs [1]	77.2	0
GradICON [12]	78.6	1e-3
GradICON(IO) [12]	80.2	5e-4
ConstrICON [12]	79.3	4e-6
ConstrICON(IO) [12]	80.1	0
uniGradICON [38]	76.2	6e-5
uniGradICON(IO) [38]	78.9	2e-4
EasyReg [21]	78.8	-
KeyMorph [46]	67.2	0
CARL	79.6	6e-3
CARL(IO)	80.4	1e-3
CARL{ROT}	77.8	1e-3
CARL{ROT}(IO)	78.9	3e-3

Method	mTRE↓	$\%\lvert J\rvert_{<0}\downarrow$
DirLab		
ANTs [1]	1.79	0
Elastix [23]	1.32	-
VoxelMorph [2]	9.88	0
LapIRN [27]	2.92	0
RRN [15]	0.83	-
Hering el al [18]	2.00	6e-2
GraphRegNet [14]	1.34	-
PLOSL [44]	3.84	0
PLOSL(IO) [44]	1.53	0
PTVReg [42]	0.83	-
GradICON [36]	1.93	3e-4
GradICON(IO) [36]	1.31	2e-4
ConstrICON [12]	2.03	7e-6
ConstrICON(IO) [12]	1.62	3e-6
uniGradICON [38]	2.26	9e-5
uniGradICON(IO) [38]	1.40	9e-5
CARL	1.88	4e-6
CARL(IO)	1.34	6e-6

Method	%DICE↑	Unsupervised
L2R Abd. CTCT Test set		
ConvexAdam [34]	69	
LapIRN [27]	67	
Estienne [9]	69	
corrField [14, 17]	49	yes
PIMed [19]	49	
PDD-Net [16]	49	
Joutard [19]	40	
NiftyReg [26]	45	yes
Gunnarsson [13]	43	
L2R Abd. CTCT Val. set		
ConvexAdam [34]	71	
Estienne [9]	65	
Estienne [9]	52	yes
SAME++(IO) [37]	49	yes
uniGradICON [38]	48	yes
uniGradICON(IO) [38]	52	yes
CARL	50	yes
CARL(IO)	56	yes

6.3. Rotation equivariance

Our approach is formally $[W, U]$ equivariant to translations. In addition, we find empirically that it can be made nearly $[W, U]$ equivariant to rotations by data augmentation. The following changes to the training procedure are required: *First*, Ξ_θ is made formally equivariant to axis aligned π rotations by evaluating the encoder four times, once for each possible π rotation, and averaging. If the encoder is equivariant in the natural sense to transforms from a class $\mathcal{T}$, then Ξ_θ becomes $[W, U]$ equivariant to that class as mentioned in Sec. 4. *Second*, the receptive field of the encoder is broadened by adding an additional dilated convolution to the residual blocks with dilation 8. This empirically stabilizes training, but slightly damages translational equivariance via increased boundary effects. In addition, we change the extrapolation used when evaluating displacement field transforms at locations outside $[0, 1]$ (see **??** in the Appendix). This change allows the Jacobian of the final transform to be continuous across this boundary when the displacement field represents a large rotation. *Finally*, the dataset is augmented with random rotations from $\mathrm{SO}(3)$. Because this augmentation requires large rotations to align the images, pretraining with naive diffusion regularization (*i.e.* $\mathcal{L}_{\text{reg}}(\varphi) = ||\nabla\varphi - \boldsymbol{I}||_F^2$) is no longer appropriate as diffusion forbids transforms with Jacobians far from the identity map. During the pretraining step, it is therefore necessary to apply the diffusion regularizer to $R^{-1} \circ \varphi \circ Q$ instead of φ, where R and Q are the augmentations applied to the input images and φ is the output of the overall architecture. As the network becomes $[W, U]$ equivariant to arbitrary rotations, $\nabla(R^{-1} \circ \varphi \circ Q)$ becomes close enough to the identity matrix for diffusion regularization to be appropriate: see **??**. We refer to CARL trained using this procedure as CARL{ROT}. While baseline CARL can handle rotations of up to 45 degrees with instance optimization, CARL{ROT} performs correct registration when the input images are arbitrarily rotated (Fig. 4).

6.4. Limitations

Registering a pair of images with a forward pass takes 2.1 seconds. Instance optimization (IO) is slower: 209 seconds on an NVIDIA A100 GPU. However, we consider the runtime of IO a minor issue at the moment, and focus on accuracy, as we anticipate that hardware and software for computing the flash attention kernel will continue to improve.

7. Conclusion

We demonstrated that our proposed multi-step architecture has a novel combination of strong equivariance (by construction to translation and optionally for rotation by judicious data augmentation) and precise handling of complex and high-resolution deformations. We show that it has close to SOTA accuracy on lung registration when compared to a competitive leaderboard, SOTA performance on brain registration with empirically verified equivariance to translations and rotations, and significantly improves registration accuracy over previous unsupervised registration approaches for the challenging registration of raw abdomen scans with differing fields of view.

Acknowledgments This research was, in part, funded

by the National Institutes of Health (NIH) under awards 1R01AR082684, 1R21MH132982, 1R01HL149877, and RF1MH126732. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing official policies, either expressed or implied, of the NIH.

References

[1] Brian B Avants, Charles L Epstein, Murray Grossman, and James C Gee. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *MedIA*, 12(1):26–41, 2008. 5, 8

[2] Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca. VoxelMorph: a learning framework for deformable medical image registration. *TMI*, 38 (8):1788–1800, 2019. 1, 3, 5, 8

[3] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. SURF: Speeded up robust features. In *ECCV*, 2006. 2

[4] Benjamin Billot, Daniel Moyer, Neerav Karani, Malte Hoffmann, Esra Abaci Turk, Ellen Grant, and Polina Golland. Equivariant and denoising CNNs to decouple intensity and spatial features for motion tracking in fetal brain MRI. In *MIDL*, 2023. 2

[5] Taco Cohen and Max Welling. Group equivariant convolutional networks. In *ICML*, pages 2990–2999. PMLR, 2016. 2

[6] Adrian V Dalca, Guha Balakrishnan, John Guttag, and Mert R Sabuncu. Unsupervised learning for fast probabilistic diffeomorphic registration. In *MICCAI*, 2018. 3

[7] Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *NeurIPS*, 2022. 4

[8] Başar Demir, Lin Tian, Hastings Greer, Roland Kwitt, François-Xavier Vialard, Raúl San José Estépar, Sylvain Bouix, Richard Rushmore, Ebrahim Ebrahim, and Marc Niethammer. Multigradicon: A foundation model for multimodal medical image registration. In *International Workshop on Biomedical Image Registration*, pages 3–18. Springer, 2024. 3

[9] Théo Estienne, Maria Vakalopoulou, Enzo Battistella, Alexandre Carré, Théophraste Henry, Marvin Lerousseau, Charlotte Robert, Nikos Paragios, and Eric Deutsch. Deep learning based registration using spatial gradients and noisy segmentation labels. In *Segmentation, Classification, and Registration of Multi-modality Medical Imaging Data: MICCAI 2020 Challenges*, 2021. 8

[10] Kunihiko Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36: 193–202, 1980. 2

[11] Hastings Greer, Roland Kwitt, François-Xavier Vialard, and Marc Niethammer. ICON: Learning regular maps through inverse consistency. In *ICCV*, 2021. 3, 5

[12] Hastings Greer, Lin Tian, Francois-Xavier Vialard, Roland Kwitt, Sylvain Bouix, Raul San Jose Estepar, Richard Rushmore, and Marc Niethammer. Inverse consistency by construction for multistep deep registration. In *MICCAI*, 2023. 1, 3, 7, 8

[13] Niklas Gunnarsson, Jens Sjölund, and Thomas B Schön. Learning a deformable registration pyramid. In *MICCAI*, 2020. 8

[14] Lasse Hansen and Mattias P Heinrich. GraphRegNet: Deep graph regularisation networks on sparse keypoints for dense registration of 3D lung CTs. *TMI*, 40(9):2246–2257, 2021. 8

[15] Xinzi He, Jia Guo, Xuzhe Zhang, Hanwen Bi, Sarah Gerard, David Kaczka, Amin Motahari, Eric Hoffman, Joseph Reinhardt, R Graham Barr, Elsa Angelini, and Andrew Laine. Recursive refinement network for deformable lung registration between exhale and inhale CT scans. *arXiv preprint arXiv:2106.07608*, 2021. 8

[16] Mattias P Heinrich. Closing the gap between deep and conventional image registration using probabilistic dense displacement networks. In *MICCAI*, 2019. 8

[17] Mattias P Heinrich, Heinz Handels, and Ivor JA Simpson. Estimating large lung motion in copd patients by symmetric regularised correspondence fields. In *MICCAI*, 2015. 8

[18] Alessa Hering, Stephanie Häger, Jan Moltz, Nikolas Lessmann, Stefan Heldmann, and Bram van Ginneken. CNN-based lung CT registration with multiple anatomical constraints. *MedIA*, 72:102139, 2021. 8

[19] 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. *TMI*, 42(3):697–712, 2022. 1, 7, 8

[20] Joel Honkamaa and Pekka Marttinen. Asymreg: Robust symmetric image registration using anti-symmetric formulation and deformation inversion layers. *arXiv preprint arXiv:2303.10211*, 2023. 3

[21] Juan Eugenio Iglesias. A ready-to-use machine learning tool for symmetric multi-modality registration of brain mri. *Scientific Reports*, 13(1):6657, 2023. 2, 8

[22] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *NeurIPS*, 34:852–863, 2021. 2

[23] Stefan Klein, Marius Staring, Keelin Murphy, Max A Viergever, and Josien PW Pluim. Elastix: a toolbox for intensity-based medical image registration. *TMI*, 29(1):196–205, 2009. 8

[24] David G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60:91–110, 2004. 2

[25] Jun Ma, Yao Zhang, Song Gu, Cheng Zhu, Cheng Ge, Yichi Zhang, Xingle An, Congcong Wang, Qiyuan Wang, Xin Liu, Shucheng Cao, Qi Zhang, Shangqing Liu, Yunpeng Wang, Yuhui Li, Jian He, and Xiaoping Yang. Abdomenct-1k: Is abdominal organ segmentation a solved problem? *TPAMI*, 44(10):6695–6714, 2022. 7

[26] 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. *Comput. Methods Programs Biomed.*, 98(3):278–284, 2010. 8

[27] Tony CW Mok and Albert Chung. Large deformation diffeomorphic image registration with Laplacian pyramid networks. In *MICCAI*, 2020. 1, 3, 8

[28] Daniel Moyer, Esra Abaci Turk, P Ellen Grant, William M Wells, and Polina Golland. Equivariant filters for efficient tracking in 3d imaging. In *MICCAI*, 2021. 2

[29] Love Nordling, Johan Öfverstedt, Joakim Lindblad, and Natasa Sladoje. Contrastive learning of equivariant image representations for multimodal deformable registration. *ISBI*, pages 1–5, 2023. 3

[30] Ankur Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. A decomposable attention model for natural language inference. In *EMNLP*, pages 2249–2255, 2016. 2

[31] R. Jarrett Rushmore, Kyle Sutherland, Holly Carrington, Justine Chen, Michael Halle, Andras Lasso, George Papadimitriou, Nick Prunier, Elizabeth Rizzoni, Brynn Vessey, Peter Wilson-Braun, Yogesh Rathi, Marek Kubicki, Sylvain Bouix, Edward Yeterian, and Nikos Makris. Anatomically curated segmentation of human subcortical structures in high resolution magnetic resonance imaging: An open science approach. *Front. Neuroanat.*, 16, 2022. 7

[32] R. Jarrett Rushmore, Kyle Sutherland, Holly Carrington, Justine Chen, Michael Halle, Andras Lasso, George Papadimitriou, Nick Prunier, Elizabeth Rizzoni, Brynn Vessey, Peter Wilson-Braun, Yogesh Rathi, Marek Kubicki, Sylvain Bouix, Edward Yeterian, and Nikos Makris. HOA-2/SubcorticalParcellations: release-50-subjects-1.1.0. 2022. 7

[33] Zhengyang Shen, Xu Han, Zhenlin Xu, and Marc Niethammer. Networks for joint affine and non-parametric image registration. In *CVPR*, 2019. 3

[34] Hanna Siebert, Lasse Hansen, and Mattias P Heinrich. Fast 3d registration with accurate optimisation and little learning for learn2reg 2021. In *MICCAI*, 2021. 8

[35] Joes Staal, Michael D Abràmoff, Meindert Niemeijer, Max A Viergever, and Bram Van Ginneken. Ridge-based vessel segmentation in color images of the retina. *TMI*, 23(4):501–509, 2004. 6

[36] Lin Tian, Hastings Greer, François-Xavier Vialard, Roland Kwitt, Raúl San José Estépar, and Marc Niethammer. GradICON: Approximate diffeomorphisms via gradient inverse consistency. *CVPR*, 2022. 1, 3, 5, 6, 8

[37] 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. *arXiv preprint arXiv:2311.14986*, 2023. 2, 8

[38] Lin Tian, Hastings Greer, Roland Kwitt, Francois-Xavier Vialard, Raul San Jose Estepar, Sylvain Bouix, Richard Rushmore, and Marc Niethammer. uniGradICON: A foundation model for medical image registration. *arXiv preprint arXiv:2403.05780*, 2024. 3, 8

[39] 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. 1

[40] Prune Truong, Martin Danelljan, Fisher Yu, and Luc Van Gool. Warp consistency for unsupervised learning of dense correspondences. In *ICCV*, 2021. 3

[41] David C Van Essen, Kamil Ugurbil, Edward Auerbach, Deanna Barch, Timothy EJ Behrens, Richard Bucholz, Acer Chang, Liyong Chen, Maurizio Corbetta, Sandra W Curtiss, Stefania Della Penna, David Feinberg, Matthew F Glasser, Noam Harel, Andrew C Heath, Linda Larson-Prior, Daniel Marcus, Georgios Michalareas, Steen Moeller, Robert Oostenfeld, Steve E Peterson, Fred Prior, Brad L Schlaggar, Stephen M Smith, Abraham Z Snyder, Junqian Xu, Essa Yacoub, and WU-Minn HCP Consortium. The human connectome project: a data acquisition perspective. *NeuroImage*, 62 (4):2222–2231, 2012. 7

[42] Valery Vishnevskiy, Tobias Gass, Gabor Szekely, Christine Tanner, and Orcun Goksel. Isotropic total variation regularization of displacements in parametric image registration. *TMI*, 36(2):385–395, 2017. 8

[43] Alan Q. Wang, Evan M. Yu, Adrian V. Dalca, and Mert Rory Sabuncu. A robust and interpretable deep learning framework for multi-modal registration via keypoints. *MedIA*, 90: 102962, 2023. 1, 2

[44] Di Wang, Yue Pan, Oguz C Durumeric, Joseph M Reinhardt, Eric A Hoffman, Joyce D Schroeder, and Gary E Christensen. PLOSL: Population learning followed by one shot learning pulmonary image registration using tissue volume preserving and vesselness constraints. *MedIA*, 79:102434, 2022. 8

[45] Xiao Yang, Roland Kwitt, Martin Styner, and Marc Niethammer. Quicksilver: Fast predictive image registration–a deep learning approach. *NeuroImage*, 158:378–396, 2017. 1, 5

[46] Evan M. Yu, Alan Q. Wang, Adrian V. Dalca, and Mert R. Sabuncu. KeyMorph: Robust multi-modal affine registration via unsupervised keypoint detection. In *MIDL*, 2022. 2, 8