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, 1

[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, 4

[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

CARL: A Framework for Equivariant Image Registration

Supplementary Material

8. Per Structure DICE Box Plot

To provide a more comprehensive picture of how different registration algorithms perform for different brain structures or organs (instead of purely reporting averages) Figs. 5 and 6 show anatomy-specific boxplots. We observe especially strong performance for CARL on the Abdomen1k dataset (Fig. 6) with excellent registration results for liver, kidney, and spleen.

9. Resolution, Downsampling & Coordinates

We use an internal convention that regardless of resolution, images have coordinates ranging from (0, 0, 0) to (1, 1, 1). Thus, a transform, a function from $[0,1]^D \to \mathcal{R}^N$, can be applied to an image of any resolution. This allows us to construct a multiresolution, multi-step registration algorithm using TwoStep and the operator Downsample defined in [11] as

$$\begin{aligned}&\text{Downsample}\{\Phi\}[I^{\mathrm{M}}, I^{\mathrm{F}}]\\ &= \Phi[\text{averagePool}(I^{\mathrm{M}}, 2), \text{averagePool}(I^{\mathrm{F}}, 2)]\,. \qquad (19)\\ &\text{TwoStep}\{\Phi, \Psi\}[I^M, I^F]\\ &= \Phi[I^M, I^F] \circ \Psi[I^M \circ \Phi[I^M, I^F], I^F]\,. \qquad (20)\end{aligned}$$

10. Implementing the diffeomorphism-to-diffeomorphism case

We can use coordinate attention to solve the diffeomorphism-to-diffeomorphism registration problem with a neural network $\Xi_{\mathcal{F}}$ (shown in the left part of Fig. 7).

The input functions I^{M}, I^{F}, and the output transform are approximated as arrays of voxels. The functional Ξ such that $\Xi[I^M, I^F] := (I^M)^{-1} \circ I^F$ (which only operates on images that are diffeomorphic) can be directly implemented, without training, using standard neural network components. We refer to this implementation as $\Xi_{\mathcal{F}}$. The intention is to map each voxel in the moving image into a high dimensional vector that will have a large dot product with the corresponding voxel in the fixed image with the same value, and then compute the attention matrix with the embedded fixed image voxels as the queries and the embedded moving image voxels as the keys. Subsequently, we can compute the center of mass of the attention masks (i.e., where each fixed image voxel matches on the moving image) by setting the *values* to be the raw coordinates of the moving image voxels. We choose for the embedding a 1×1 convolution with large weights followed by a sine-nonlinearity, which has the desired property of two vectors having a large dot product only when their input intensities are similar. Because our images are diffeomorphisms, we know a-priori that the input intensity of our moving image will only be close to intensities of the fixed image in a small region. We verify that this network, without any training, reproduces Ξ when applied to input images that are diffeomorphisms, see Fig. 7 (right).

10.1. Limitation on equivariance feasibility

In Sec. 10, we turned images into features using 1×1 convolution followed by a sine nonlinearity which, since it is a function applied pointwise, is perfectly equivariant. This worked since the images to be registered were diffeomorphisms, and hence each intensity vector was unique. However, since, as we are about to prove, we cannot achieve equivariance to arbitrary diffeomorphisms for registering real images, we have to sacrifice some equivariance in order to expand the set of valid inputs. This drives our choice to target translation and rotation equivariance out of the set of possible diffeomorphisms.

Claim. It is impossible to have an algorithm that is $[W, U]$ equivariant to any non-identity class of transforms and can be applied to arbitrary images.

Counterexample. Assume that Φ is a $[W, U]$ equivariant algorithm for all $W, U \in$ diffeomorphisms, and that is valid for all input images. We ask it to register the images $I^{\mathrm{M}}, I^{\mathrm{F}} := 0$. Then, for a non identity W and U picked to be identity,

$$\begin{aligned}\Phi[I^{\mathrm{M}}, I^{\mathrm{F}}] &= W^-1 \circ \Phi[I^{\mathrm{M}} \circ W, I^{\mathrm{F}} \circ U] \circ U \qquad (21)\\ \Phi[I^{\mathrm{M}}, I^{\mathrm{F}}] &= W^-1 \circ \Phi[I^{\mathrm{M}} \circ W, I^{\mathrm{F}}] \qquad (22)\\ \Phi[0, 0] &= W^{-1} \circ \Phi[0, 0] \qquad (23)\\ id &= W^{-1}\,, \qquad (24)\end{aligned}$$

where 0 indicates an image that is zero everywhere. This yields a contradiction.

Claim. It is impossible to have an algorithm that is $[W, U]$ equivariant to rotations and can be applied to rotationally symmetric images.

Counterexample. Assume that Φ is a $[W, U]$ equivariant algorithm for all $W, U \in$ rotations, and that Φ is valid for input images including at least one rotationally symmetric image I^{M} (such that for a non identity W, $I^{\mathrm{M}} \circ W = I^{\mathrm{M}}$.) We ask it to register the images $I^{\mathrm{M}}, I^{\mathrm{F}}$. Then, for a non identity W with respect to which I^{M} is symmetric and U

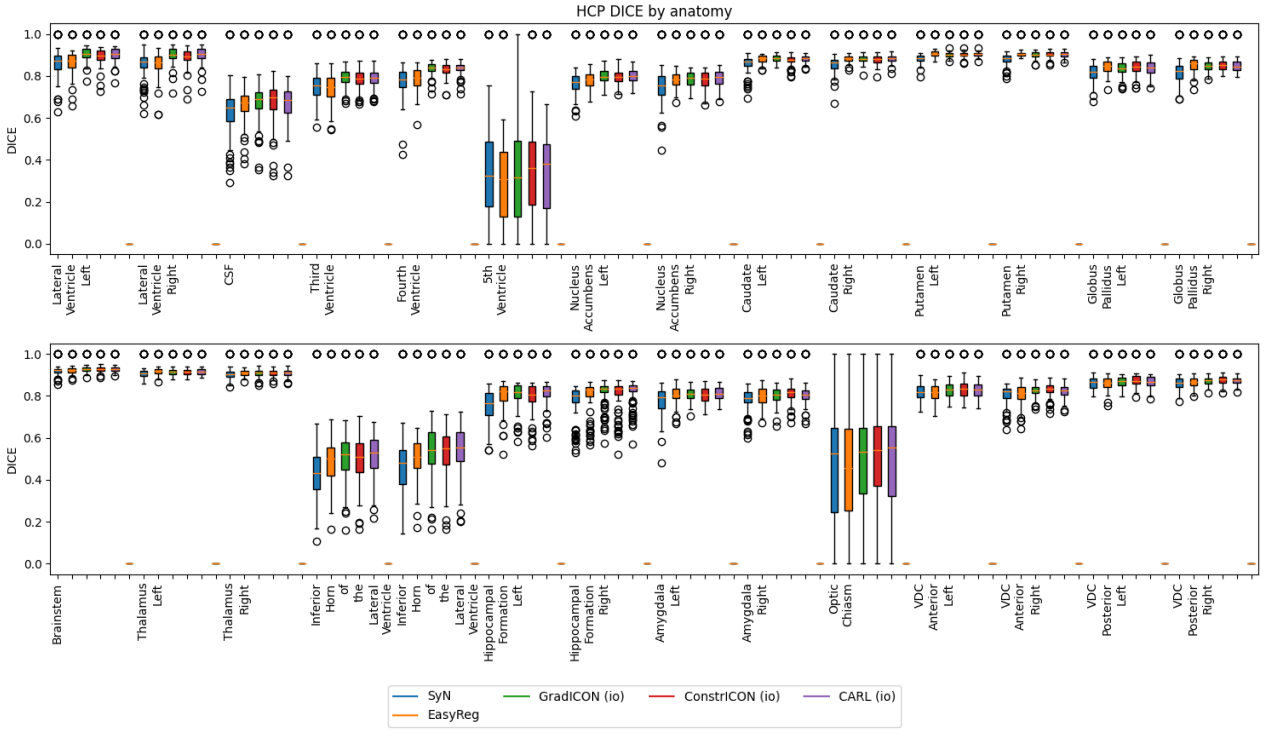

Figure 5. Per structure DICE scores on the HCP dataset. CARL ranks well on most structures.

picked to be identity,

$$\Phi[I^{\mathrm{M}}, I^{\mathrm{F}}] = W^{-1} \circ \Phi[I^{\mathrm{M}} \circ W, I^{\mathrm{F}} \circ U] \circ U \tag{25}$$
$$\Phi[I^{\mathrm{M}}, I^{\mathrm{F}}] = W^{-1} \circ \Phi[I^{\mathrm{M}} \circ W, I^{\mathrm{F}}] \tag{26}$$
$$\Phi[I^{\mathrm{M}}, I^{\mathrm{F}}] = W^{-1} \circ \Phi[I^{\mathrm{M}}, I^{\mathrm{F}}] \tag{27}$$
$$id = W^{-1}\,, \tag{28}$$

This yields a contradiction, as we assumed W was not identity.

We conclude that there is a tradeoff. If there is a valid input image I and a nonzero transform T such that $I \circ T = I$, then T cannot be in the class of transforms with respect to which Φ is $[W, U]$ equivariant. For a simple example, an algorithm that registers images of perfect circles cannot be $[W, U]$ equivariant to rotations. For a practical example, since brain-extracted brain images have large areas outside the brain that are exactly zero, algorithms that register such preprocessed brain images cannot be $[W, U]$ equivariant to transforms that are identity everywhere in the brain but have deformations outside the brain. To modify $\Xi_{\mathcal{F}}$ so that it can apply to a broader class of images other than "images that happen to be diffeomorphisms", we thus have to restrict the transforms with respect to which it is $[W, U]$ equivariant.

10.2. Guarantee given equivariance

While it is unfortunate that we cannot achieve [W, U] equivariance to arbitrary diffeomorphisms for arbitrary input images, there is great advantage to expanding the group of transforms $\mathcal{T}$ with respect to which our algorithm is [W, U] equivariant. The advantage is as follows: for any input image pair $I^{\mathrm{M}}, I^{\mathrm{F}}$ where the images can be made to match exactly by warping I^{M} by a transform U in $\mathcal{T}$, an algorithm that outputs the identity map when fed a pair of identical images and is $[W, U]$ equivariant with respect to $\mathcal{T}$ will output U for $I^{\mathrm{M}}, I^{\mathrm{F}}$. We see this as follows.

Assume Φ outputs the identity transform when fed identical fixed and moving images and $[W, U]$ equivariant with respect to $\mathcal{T}$, and $I^{\mathrm{M}} \circ U = I^{\mathrm{F}}$ for a $U \in \mathcal{T}$. Then

$$\Phi[I^{\mathrm{M}}, I^{\mathrm{F}}] \tag{29}$$
$$= \Phi[I^{\mathrm{M}}, I^{\mathrm{M}} \circ U] \tag{30}$$
$$= \Phi[I^{\mathrm{M}}, I^{\mathrm{M}}] \circ U \tag{31}$$
$$= U\,, \tag{32}$$

where $\Phi[I^{\mathrm{M}}, I^{\mathrm{M}} \circ U] = \Phi[I^{\mathrm{M}}, I^{\mathrm{M}}] \circ U$ holds because of the $[W, U]$ equivariance with W being the identity transform.

We note that before training, the CARL architecture emphatically does not have the property of outputting the iden-

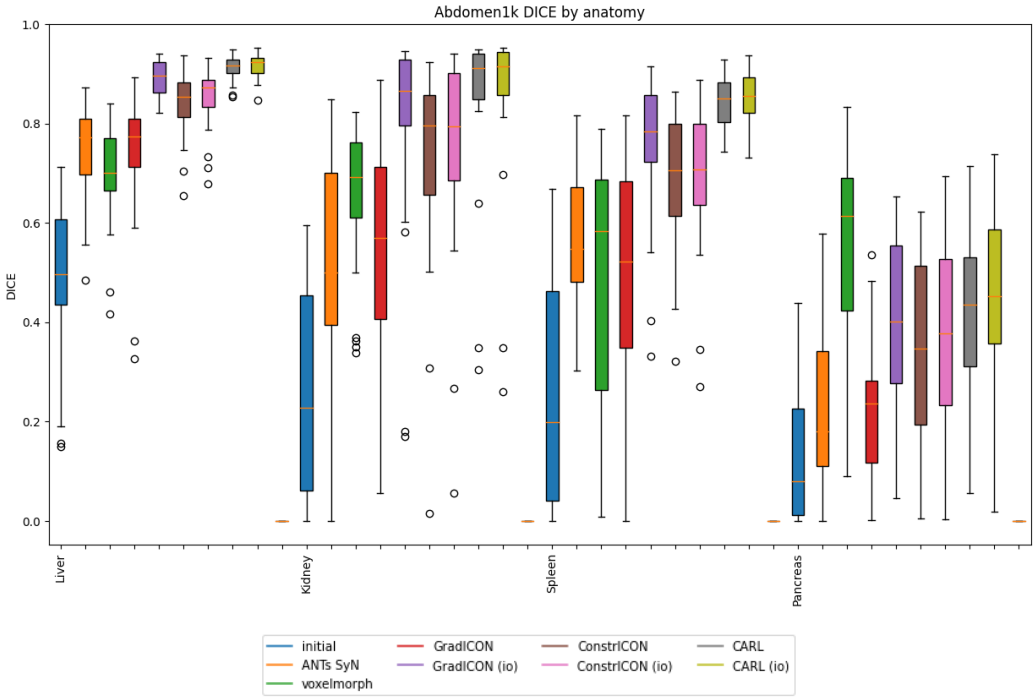

Figure 6. Per structure DICE scores on the Abdomen1k dataset. We observe that CARL is dramatically ahead of competing methods on liver, kidney, and spleen registration, but performs meaninfully worse than Voxelmorph on the pancreas.

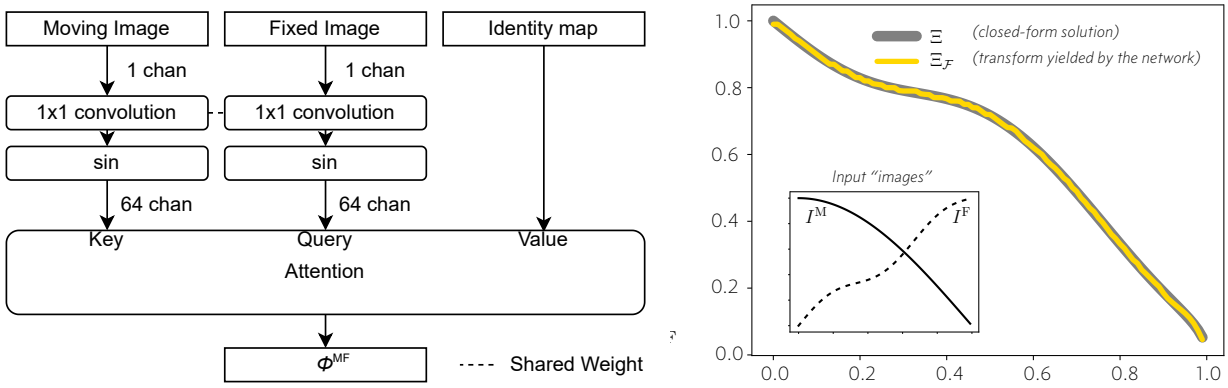

Figure 7. *Left*: Neural network $\Xi_{\mathcal{F}}$ implementing Ξ. *Right*: Result of registering the 1-dimensional "images" $I^{\mathrm{M}} : [0,1] \to [0,1], x \mapsto \cos(\frac{\pi}{2}x)$ and $I^{\mathrm{F}} : [0,1] \to [0,1], x \mapsto x + 0.07\sin(3\pi x)$ via Ξ and $\Xi_{\mathcal{F}}$, illustrating that the resulting maps are equivalent. Ξ is computable here as these images are invertible and smooth. The neural network output (gold) closely matches the analytical solution (i.e., $\Xi[I^{\mathrm{M}}, I^{\mathrm{F}}] = {I^{\mathrm{M}}}^{-1} \circ I^{\mathrm{F}} = \frac{2}{\pi}\cos^{-1}(x + 0.07\sin(3\pi x))$, gray). *Best-viewed in color.*

tity map when fed identical images: instead, it learns this property from the regularizer during training.

11. Performance implications of two step registration

We observed in Sec. 6.1 that Ξ_θ trains significantly better as the beginning of a multistep algorithm than on its own. Here, we examine why that may be, while removing as much complexity as possible. Our finding suggests

that two step registration assists training by functioning as a similarity measure with better capture radius.

First, we briefly train a single step network Φ on the **Baseline** task from Sec. 6.1, i.e. we stop training before convergence. Then, we examine the loss landscape of a trivial "fixedTranslation" neural network τ to register **Baseline**. This network has a single parameter, t, and it ignores its input images and always shifts images to the right by t: that is

$$\tau[I^{\mathrm{M}}, I^{\mathrm{F}}](\vec{r}) = \vec{r} + \begin{bmatrix} t \\ 0 \end{bmatrix} . \tag{33}$$

The optimal value of t is zero since there is no bias towards left or right shift of images in this dataset- but if we were to train τ on LNCC similarity, how well would the gradients drive t to zero?

We plot $LNCC(I^{\mathrm{M}} \circ \tau[I^{\mathrm{M}}, I^{\mathrm{F}}], I^{\mathrm{F}})$ against t compared to $LNCC(I^{\mathrm{M}} \circ TwoStep\{\tau, \Phi\}[I^{\mathrm{M}}, I^{\mathrm{F}}], I^{\mathrm{F}})$. We also plot $\frac{\partial}{\partial t} LNCC(I^{\mathrm{M}} \circ \tau[I^{\mathrm{M}}, I^{\mathrm{F}}], I^{\mathrm{F}})$ and $\frac{\partial}{\partial t} LNCC(I^{\mathrm{M}} \circ TwoStep\{\tau, \Phi\}[I^{\mathrm{M}}, I^{\mathrm{F}}], I^{\mathrm{F}})$ using PyTorch's back-propagation. Fig. 8 shows the result indicating that multi-step registration results in better capture radius.

We conjecture that when two step network TwoStep$\{\tau, \Phi\}$ is trained with an LNCC loss, the loss function seen by τ is not simply LNCC, but instead the loss function seen by τ is actually the performance of Φ (which is measured by LNCC), which is an implicit loss function with a better capture radius than the original LNCC loss function.

12. Computational Budget

Each 100,000 step training run of CARL takes 14 days on 4 RTX A6000 GPUs or 6 days on four A100 GPUS. In total, 336 GPU days were spent developing the CARL architecture and training the final models.

An additional 45 GPU days were spent training comparison methods. 14 server days were spent training KeyMorph variants, although the published KeyMorph code is io-bound and did not significantly load the server's GPU.

13. Comparison Methods Details

13.1. Abdomen1k

For Abdomen1k, we trained all methods using their published code and default hyperparameters.

13.2. DirLab Lung

On the DirLab challenge set, all results of comparison methods are taken from the literature. Results of ANTs, Elastix, Voxelmorph, and LapIRN are from [36]. The remainder are from their original publications.

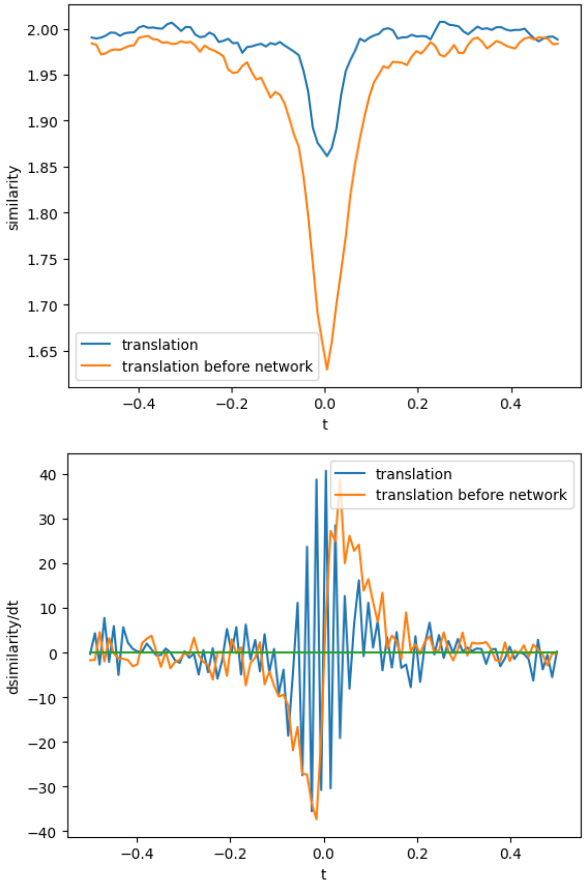

Figure 8. The loss of $TwoStep\{\tau, \Phi\}$ (i.e., translation before network) as a function of t is much better behaved than the loss of τ (i.e., translation) as a function of t. The capture radius of the former is larger and the loss is overall smoother close to the correct solution as shown on the bottom.

13.3. HCP

On HCP, we evaluated ANTs using code from [36]. Results of GradICON, and ConstrICON are from the ConstrICON publication. We evaluated KeyMorph by training a model on the HCP Dataset using KeyMorph's published code and hyperparameters for the IXI dataset. We evaluated Easyreg using its published weights, which are advertised to be appropriate for the HCP dataset. We measured the equivariance of the GradICON method using GradICON's published code and weights.

14. Abdomen 1k Test Pairs

We used the following 30 image pairs to evaluate our abdomen registration experiments.

00817 00872, 00808 00832, 00815 00863, 00857 00860, 00883 00848, 00826 00812, 00862 00803, 00849 00855, 00877 00800, 00857 00834, 00829 00875, 00813 00840,

00803 00802, 00803 00883, 00869 00801, 00848 00887, 00827 00854, 00803 00867, 00828 00856, 00863 00870, 00829 00844, 00829 00886, 00828 00858, 00837 00802, 00853 00871, 00882 00812, 00823 00880, 00837 00815, 00842 00864, 00854 00864

15. Extension to Rotational Equivariance

The first solution to obtain a registration network that exhibts rotation equivariance that comes to mind is to simply augment the training dataset with random rotations, and see if the network can still register it. This conceptually works fine for our main training with GradICON regularization, but breaks when directly applied to our diffusion regularized pretraining (which is empirically required for training a coordinate attention layer). That is , the training loss would look like

$$\begin{aligned} R, Q &\sim \text{Uniform(Rotations)} \\ I^M, I^F &\sim \text{Dataset} \\ \hat{I}^M, \hat{I}^F &:= (I^M \circ R), (I^F \circ Q) \\ \text{minimize} &: \mathcal{L}_{\text{sim}}(\hat{I}^M \circ \Phi[\hat{I}^M, \hat{I}^F], \hat{I}^F) + \mathcal{L}_{\text{reg}}(\Phi[\hat{I}^M, \hat{I}^F]) \end{aligned} \tag{34}$$

This cannot be reliably trained with a diffusion regularizer, because to align $\hat{I}^M$ to $\hat{I}^F$ will require transforms with Jacobians of the transformation map that are very far from the identity map as they will need to express large-scale rotations.

Our proposed solution is to move the augmentation "inside" the losses, in the following sense:

First, expand our augmented images $\hat{I}^M, \hat{I}^F$

$$\begin{aligned} R, Q &\sim \text{Uniform(Rotations)} \\ I^M, I^F &\sim \text{Dataset} \\ &\text{minimize} : \\ &\mathcal{L}_{\text{sim}}((I^M \circ R) \circ \Phi[I^M \circ R, I^F \circ Q], I^F \circ Q) \\ &+ \mathcal{L}_{\text{reg}}(\Phi[I^M \circ R, I^F \circ Q]) \end{aligned} \tag{35}$$

In this expanded form, change to the following:

$$\begin{aligned} R, Q &\sim \text{Uniform(Rotations)} \\ I^M, I^F &\sim \text{Dataset} \\ &\text{minimize} : \\ &\mathcal{L}_{\text{sim}}(I^M \circ (R \circ \Phi[I^M \circ R, I^F \circ Q] \circ Q^{-1}, I^F) \\ &+ \mathcal{L}_{\text{reg}}(R \circ \Phi[I^M \circ R, I^F \circ Q] \circ Q^{-1}) \end{aligned} \tag{36}$$

Then, collect like terms. It becomes clear that the augmentation is now *inside* the loss and connected to the network.

$$\begin{aligned} R, Q &\sim \text{Uniform(Rotations)} \\ I^M, I^F &\sim \text{Dataset} \\ \hat{\Phi}[I^M, I^F] &:= R \circ \Phi[I^M \circ R, I^F \circ Q] \circ Q^{-1} \\ \text{minimize} &: \mathcal{L}_{\text{sim}}(I^M, \hat{\Phi}[I^M, I^F], I^F) + \mathcal{L}_{\text{reg}}(\hat{\Phi}[I^M, I^F]) \end{aligned} \tag{37}$$

Now, while Φ outputs large rotations, on a rotationally aligned dataset $\hat{\Phi}$ outputs transforms with Jacobians near the identity, and so can be trained with diffusion regularization.

16. Improved Extrapolation of Displacement Fields

Displacement fields (disp) are stored as grids of vectors associated with coordinates in $[0,1]^D$. In ICON [11], Greer et al. noted that the method of extrapolating when evaluating a transform outside of this region is important when composing transforms, since transforms such as translations and rotations move some coordinates from inside $[0,1]^D$ to outside it. They propose coordinate by coordinate clipping before interpolating into *the displacement field*

$$\varphi_{\text{disp}}(x) = x + \text{interpolate}(\text{disp}, \text{clip}(x, 0, 1)) \,. \tag{38}$$

This formulation has a discontinuous Jacobian on the boundary of $[0,1]^D$, and in particular results in non-invertible transforms on the boundaries for large rotations.

We instead propose

$$\text{clip}(x) = x - \begin{cases} x & \text{if } x < 0 \\ 0, & \text{otherwise} \end{cases} - \begin{cases} (x-1) & \text{if } x > 1 \\ 0, & \text{otherwise} \end{cases} \tag{39}$$

$$\text{reflect}(x) = x - \begin{cases} 2x & \text{if } x < 0 \\ 0, & \text{otherwise} \end{cases} - \begin{cases} 2(x-1) & \text{if } x > 1 \\ 0, & \text{otherwise} \end{cases} \tag{40}$$

$$\varphi_{\text{disp}}(x) = x + 2\,\text{interpolate}(\text{disp}, \text{clip}(x)) \tag{41}$$

$$- \text{interpolate}(\text{disp}, \text{reflect}(x)) \tag{42}$$

which is identical inside $[0,1]^D$ but has continuous Jacobian over the boundary.

17. Investigation of internal features of Ξ_θ

We use interpretability techniques to investigate the features learned by the convolutional encoder Conv of Ξ_θ. Two images are selected from the Abdomen1k test set, and independently encoded using the convolutional network. We convert these images into features using Conv. From these

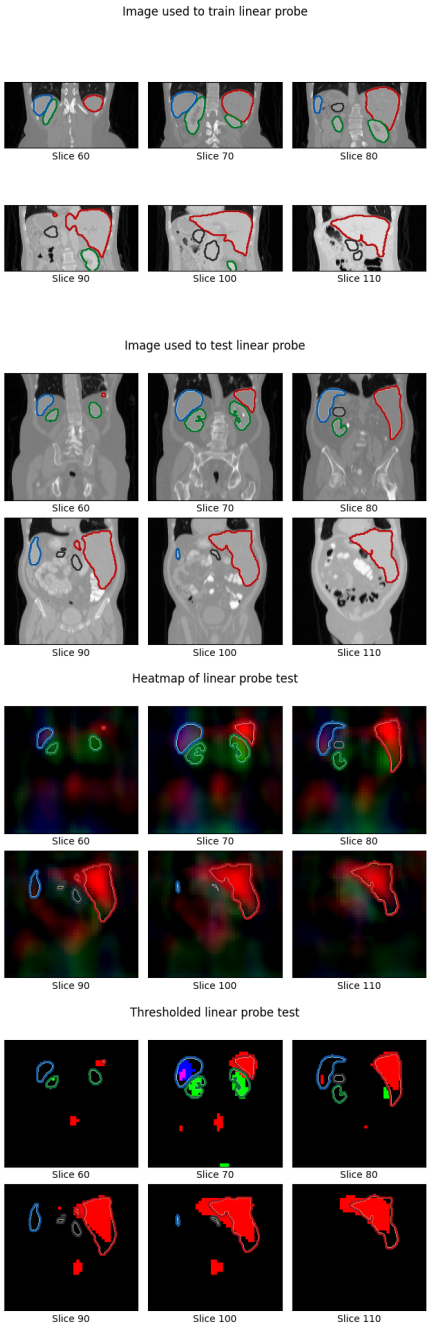

Figure 9. A linear probe is used to aid interpretability of the features learned by the convolutional encoder. The linear probe is trained on one image, and then its output heat maps are visualized on another image. The red, green, and blue channels are used to indicate the liver, kidney, and spleen respectively. The grey channel is used to indicate the pancreas, although no direction is found in the features that segments it.

voxelwise features, we train linear probes to segment the kidneys, liver, pancreas, and spleen of a single train image

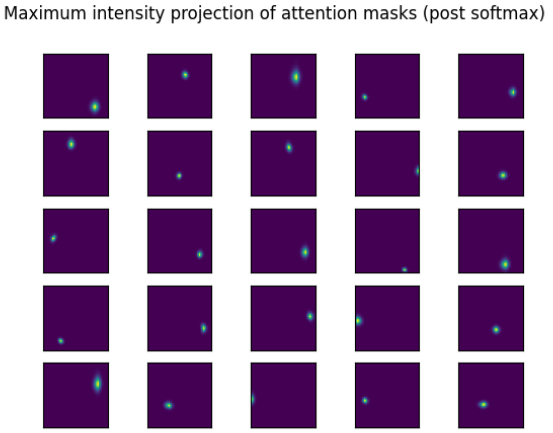

Figure 10. Sample attention masks from inside the coordinate attention block of CARL trained on Abdomen1k. The masks are compact, justifying the claim on which we build Sec. 5.

by minimizing least squares error between ground truth and predicted label, and then visualize the probe's output on a second test image. This linear probe suggests (see Fig. 9) that the features, which were learned without any segmentations, include directions that measure liver-ness, spleen-ness, and kidney-ness, but there is no pancreas-direction. This may explain why our model is less accurate at registering the pancreas than the other organs.

17.1. Verification that Ξ_θ's attention masks are compact

As an assumption in Sec. 5 was that post training, attention masks are spatially compact. We verify this by computing the attention masks associated with the query vectors of 25 random voxels when registering the pair in Fig. 9, maximum intensity projecting them to get 2-D heatmaps, and plotting. As expected, we see in Fig. 10 that each pixel in the moving image attends to a small region in the fixed image. As a result, these attention masks will not immediately interact with the boundary of the padded feature volume when an image is translated. This property is required for [W, U] equivariance.

18. Fully elaborated proof that Coordinate attention is [W, U] equivariant

Previously, we elided the difference between two definitions of an image: a function from $[0,1]^D \to \mathbb{R}$ suitable for composition with transforms, and a function from voxel indices to intensities suitable for discrete convolution and attention. Here, we fully make this distinction explicit. We will continue to consider *images* to be continuous, and discretize them as necessary by composing them with or interpolating them at the function coords which maps voxel indices to coordinates in $[0,1]^D$. This explicit style makes clear that the proof is formally correct, and also more directly maps to the implementation. We use the linear interpolation function interpolate(points, values, x) where x is the spatial location where we evaluate, and points and values are the locations where we know the value of the function (typically a grid).

Assumptions: We assume that the feature encoders are translation equivariant like

$$\mathrm{Conv}_\theta(I \circ U) = \mathrm{Conv}_\theta(I) \circ U\,. \tag{43}$$

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 $\mathbb{T}$,

$$\mathbb{T}[K \circ P_1, Q \circ P_2, V \circ P_1] = \mathbb{T}[K, Q, V] \circ P_2\,. \tag{44}$$

In plain language, changing the order of the queries causes the order of the output of the attention operation to have its order changed in the same way, and changing the order of the keys and values has no effect as long as they are changed together.

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

Lemma:

$$\mathbb{T}[K, Q, f \circ V] = f \circ \mathbb{T}[K, Q, V]\,. \tag{45}$$

Proof of lemma: Once attention weights w_i are computed, for each output token we produce a weighting function W

$$W(\mathbf{x}_i \dots) = \sum_j w_j \mathbf{x}_j \tag{46}$$

where w_i sum to 1.

We also have an affine function f, that is

$$f(\mathbf{x}) = A\mathbf{x} + \mathbf{b} \tag{47}$$

We then observe that b is preserved and hence $f \circ W = W \circ f$ as long as w_i sum to 1.

Finally, 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. This is a crucial empirical result that we find evidence for in Fig. 10

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

Proof: 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 (and hence do not interact with the boundary), we can suppress boundary effects by padding with zeros before applying the operation. So, for translations W and U, we have

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

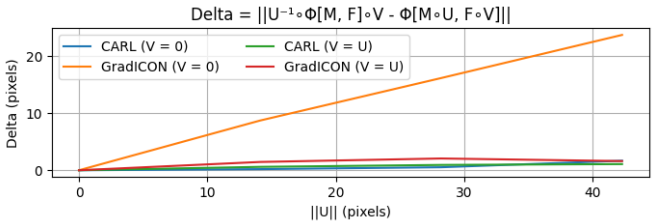

Figure 11. We directly confirm CARL is $[W, U]$ and GradICON is $[U, U]$ equivariant to translation on the deformed retina dataset.

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](x) = \text{interpolate}(\text{coords}, \mathbb{T}[\text{Conv}_\theta(I^M \circ W_{x\mapsto x+r} \circ \text{coords}), \text{Conv}_\theta(I^F \circ U_{x\mapsto x+r} \circ \text{coords}), \text{coords}], x) \\
&= \text{interpolate}(\text{coords}, \mathbb{T}[\text{Conv}_\theta(I^M \circ \text{coords} \circ W_{\text{permutation}}), \text{Conv}_\theta(I^F \circ \text{coords} \circ U_{\text{permutation}}), \text{coords}], x) \\
&\overset{(43)}{=} \text{interpolate}(\text{coords}, \mathbb{T}[\text{Conv}_\theta(I^M \circ \text{coords}) \circ W_{\text{permutation}}, \text{Conv}_\theta(I^F \circ \text{coords}) \circ U_{\text{permutation}}, \text{coords}], x) \\
&\overset{(45)}{=} \text{interpolate}(\text{coords}, W^{-1}_{x\mapsto x+r} \circ \mathbb{T}[\text{Conv}_\theta(I^M \circ \text{coords}) \circ W_{\text{permutation}}, \text{Conv}_\theta(I^F \circ \text{coords}) \circ U_{\text{permutation}}, W_{x\mapsto x+r} \circ \text{coords}], x) \\
&= W^{-1}_{x\mapsto x+r} \circ \text{interpolate}(\text{coords}, \mathbb{T}[\text{Conv}_\theta(I^M \circ \text{coords}) \circ W_{\text{permutation}}, \text{Conv}_\theta(I^F \circ \text{coords}) \circ U_{\text{permutation}}, W_{x\mapsto x+r} \circ \text{coords}], x) \\
&= W^{-1}_{x\mapsto x+r} \circ \text{interpolate}(\text{coords}, \mathbb{T}[\text{Conv}_\theta(I^M \circ \text{coords}) \circ W_{\text{permutation}}, \text{Conv}_\theta(I^F \circ \text{coords}) \circ U_{\text{permutation}}, \text{coords} \circ W_{\text{permutation}}], x) \\
&\overset{(44)}{=} W^{-1}_{x\mapsto x+r} \circ \text{interpolate}(\text{coords}, \mathbb{T}[\text{Conv}_\theta(I^M \circ \text{coords}), \text{Conv}_\theta(I^F \circ \text{coords}) \circ U_{\text{permutation}}, \text{coords}], x) \\
&\overset{(44)}{=} W^{-1}_{x\mapsto x+r} \circ \text{interpolate}(\text{coords}, \mathbb{T}[\text{Conv}_\theta(I^M \circ \text{coords}), \text{Conv}_\theta(I^F \circ \text{coords}), \text{coords}] \circ U_{\text{permutation}}, x) \\
&= W^{-1}_{x\mapsto x+r} \circ \text{interpolate}(\text{coords} \circ U^{-1}_{\text{permutation}}, \mathbb{T}[\text{Conv}_\theta(I^M \circ \text{coords}), \text{Conv}_\theta(I^F \circ \text{coords}), \text{coords}], x) \\
&= W^{-1}_{x\mapsto x+r} \circ \text{interpolate}((U_{x\mapsto x+r})^{-1} \circ \text{coords}, \mathbb{T}[\text{Conv}_\theta(I^M \circ \text{coords}), \text{Conv}_\theta(I^F \circ \text{coords}), \text{coords}], x) \\
&= W^{-1}_{x\mapsto x+r} \circ \text{interpolate}(\text{coords}, \mathbb{T}[\text{Conv}_\theta(I^M \circ \text{coords}), \text{Conv}_\theta(I^F \circ \text{coords}), \text{coords}], U_{x\mapsto x+r}(x)) \\
&= W^{-1} \circ \Xi_\theta[I^M, I^F] \circ U\,.
\end{aligned}
\tag{48}
$$

Again, 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.

Table 2. CARL: ablation of final refinement layer (no IO)

	Abdomen1k DICE	HCP DICE	DirLab mTRE	L2R DICE
w/o final refinement	74.1	78.8	2.58	49
with final refinement	75.7	79.6	1.88	50

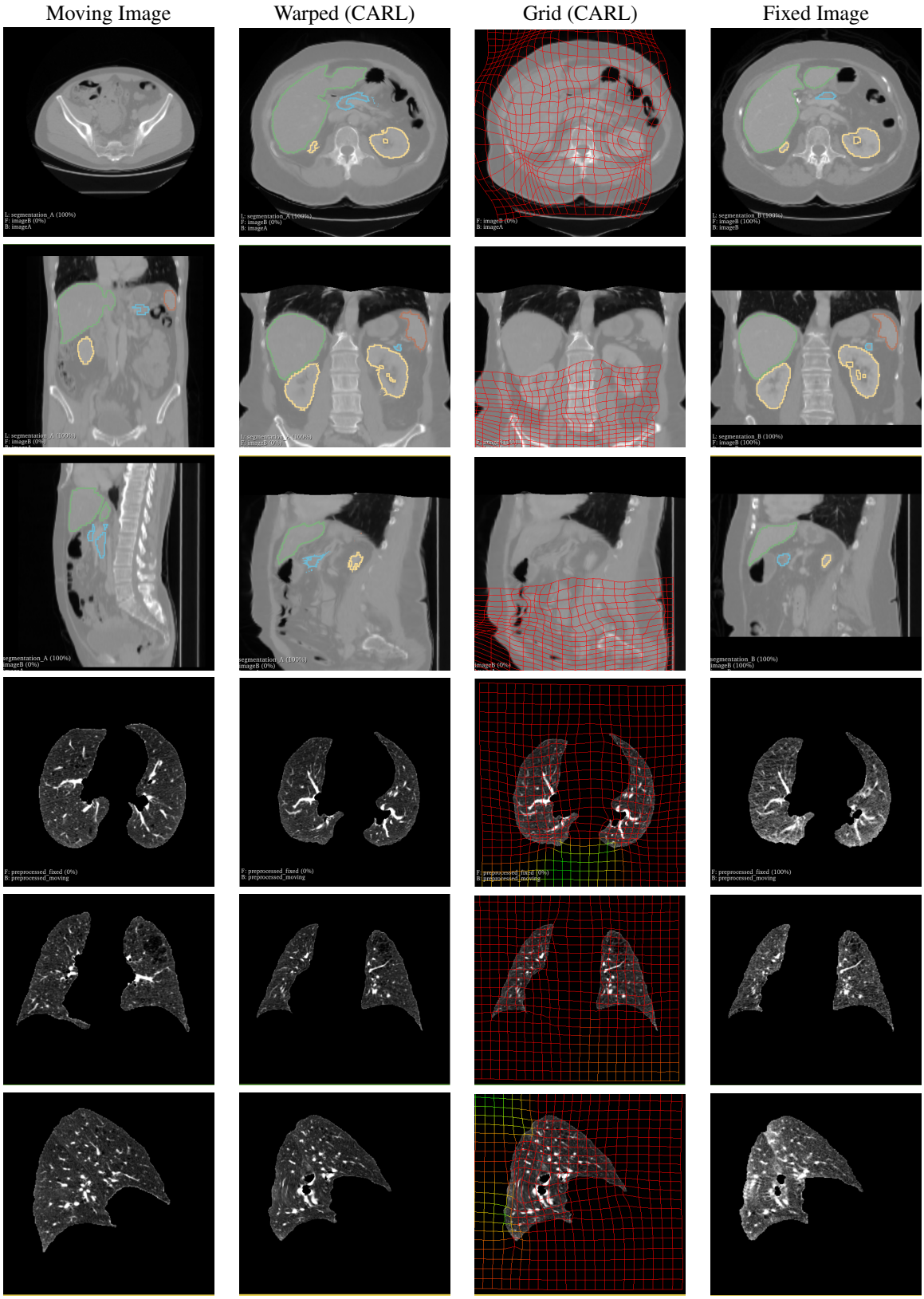

Figure 12. Detailed figures of our results on Abdomen1k (cases 00817 00872) and DirLAB case 1

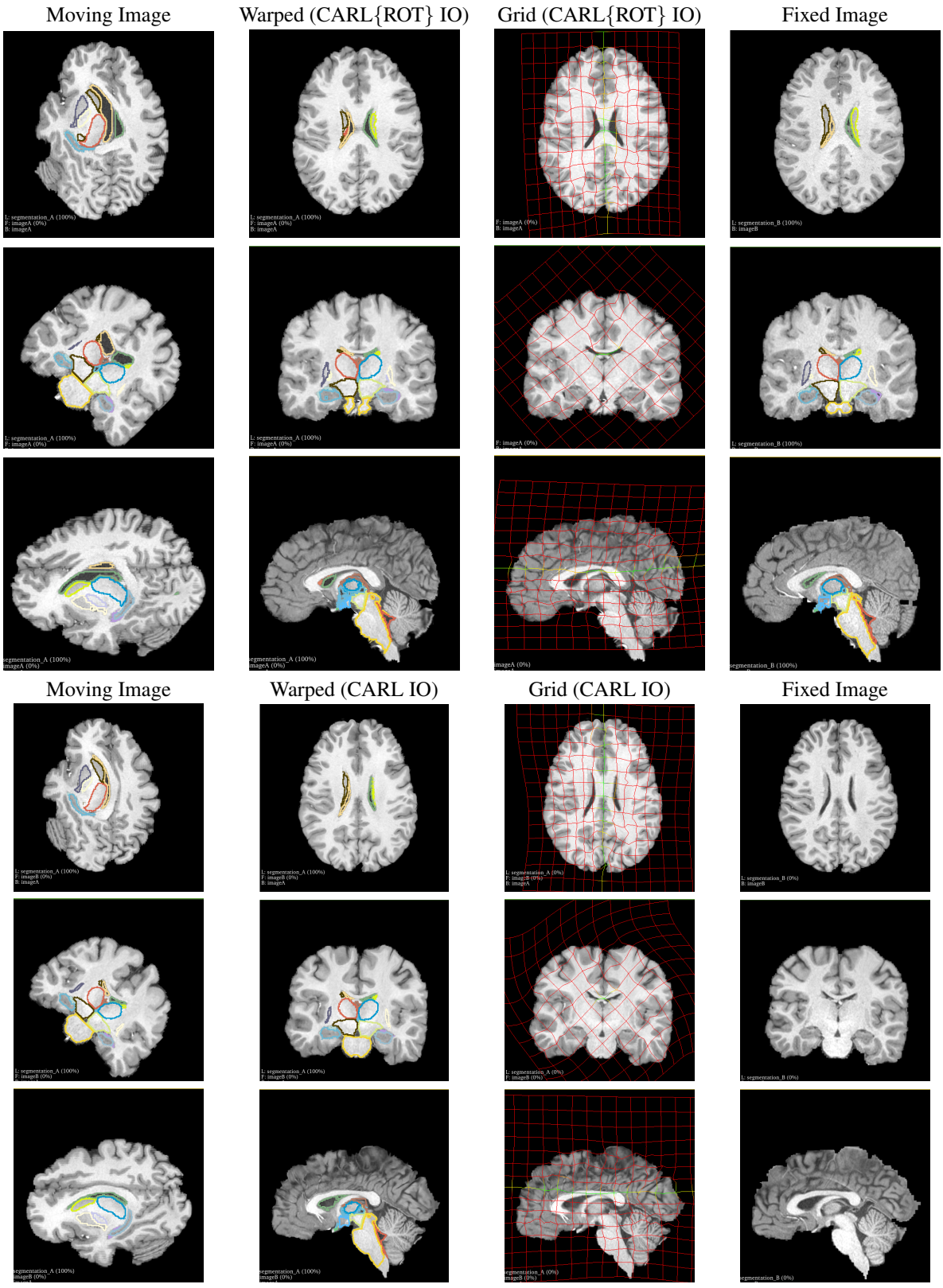

Figure 13. Detailed figures of our results on HCP with the moving image synthetically rotated by 45 degrees. Both CARL(IO) and CARL{ROT}(IO) handle a 45 degree rotation well. This is especially remarkable for CARL(IO), which is far out of its training distribution. 45 degrees is empirically the limit of CARL's tolerance for rotations, while CARL{ROT}'s DICE is unaffected by arbitrary rotations, as seen in Fig. 4. Also, observe that CARL{ROT}'s formal equivariance to rotation causes its deformation grid to move rigidly with the brain in the negative space surrounding it.

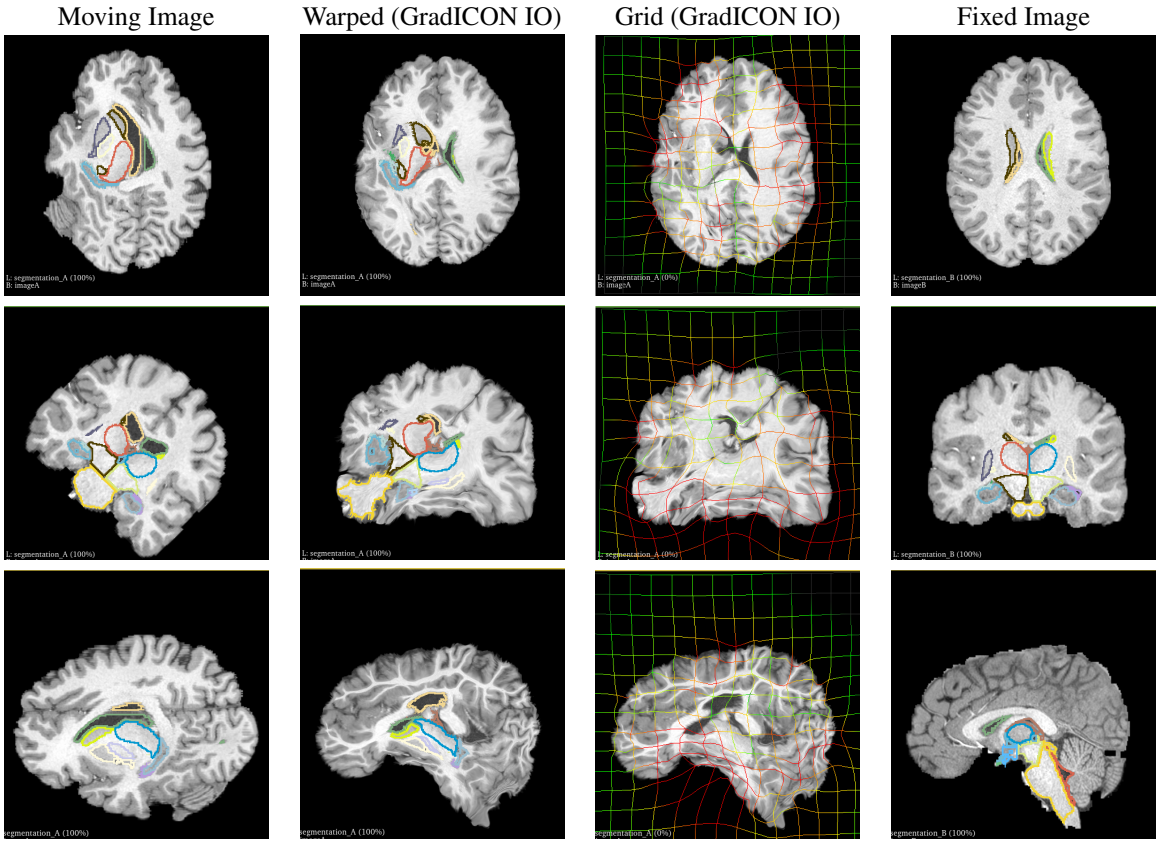

Figure 14. In contrast, GradICON cannot adapt to a 45 degree rotation which was outside its training distribution.