FAST PREDICTIVE MULTIMODAL IMAGE REGISTRATION

Xiao Yang[1], *Roland Kwitt*[4], *Martin Styner*[1,3], *Marc Niethammer*[1,2]

[1]Department of Computer Science, UNC Chapel Hill, USA
[2]Biomedical Research Imaging Center, UNC Chapel Hill, USA
[3]Department of Psychiatry, UNC Chapel Hill, USA
[4]Department of Computer Science, University of Salzburg, Austria

ABSTRACT

We introduce a deep encoder-decoder architecture for image deformation prediction from *multimodal images*. Specifically, we design an image-patch-based deep network that jointly (i) learns an image similarity measure and (ii) the relationship between image patches and deformation parameters. While our method can be applied to general image registration formulations, we focus on the Large Deformation Diffeomorphic Metric Mapping (LDDMM) registration model. By predicting the *initial momentum* of the shooting formulation of LDDMM, we preserve its mathematical properties and drastically reduce the computation time, compared to optimization-based approaches. Furthermore, we create a Bayesian probabilistic version of the network that allows evaluation of registration uncertainty via sampling of the network at test time. We evaluate our method on a 3D brain MRI dataset using both T1- and T2-weighted images. Our experiments show that our method generates accurate predictions and that *learning* the similarity measure leads to more consistent registrations than relying on *generic* multimodal image similarity measures, such as mutual information. Our approach is an order of magnitude faster than optimization-based LDDMM.

Index Terms— deep learning, deformation prediction, multimodal image similarity

1. INTRODUCTION

Multimodal image registration seeks to estimate spatial correspondences between image pairs from different imaging modalities (or protocols). In general image registration, these correspondences are estimated by finding the spatial transformation which makes a transformed source image most *similar* to a fixed target image. For unimodal image registration, image similarity should be high if images are close to identical, which allows using simple image similarity measures such as the sum of squared intensity differences (SSD) between image pairs. Assessing image similarity *across* modalities is substantially more difficult as image appearance can vary widely, e.g., due to different underlying physical imaging principles. In fact, these differences are *desired* as they can, for example, highlight different tissue properties in brain imaging. Hence, more sophisticated multimodal similarity measures are required. Furthermore, image registration results are driven by both the chosen similarity measure *and* the chosen deformation model. Hence, especially for multimodal image registration, where assessing image similarity becomes challenging, considering the similarity measure *jointly* with the deformation model is important.

The most popular multimodal similarity measure is mutual information (MI) [1], but other *hand-crafted* multimodal similarity measures have also been proposed [2, 3, 4]. These approaches *assume* properties characterizing good image alignment, but do not *learn* them from data. Hence, more recently, learning-based approaches to measure image similarity have been proposed. These techniques include measuring image similarity by comparing observed and learned intensity distributions via KL-divergence [5]; or learning the similarity of image pixels/voxels or image features (e.g., Fourier/Gabor features) via max-margin structured output learning [6], boosting [7], or deep learning [8, 9]. Some methods avoid a complex similarity measure by applying image synthesis for the source/target image to change the task to unimodal registration [10, 11, 12]. However, the registration performance then heavily depends on the synthesis accuracy.

Once a similarity measure (learned or hand-crafted) is chosen, registration still requires finding the optimal registration parameters by numerical optimization. This can be particularly costly for popular nonparametric elastic or fluid registration approaches which are based on models from continuum mechanics [13, 14] and therefore require the optimization over functions, which results in millions of unknowns after numerical discretization. Approaches to avoid numerical optimization by prediction have been proposed. But, unlike our proposed method, they either focus on predicting displacements via optical flow [15, 16] or low-dimensional parametric models [17, 18, 19], or cannot address multimodal image registration [20], or do not consider jointly learning a model for deformation prediction and image similarity [21].

Specifically, Chou et al. [17] propose a multi-scale linear regressor for affine transformations or low-dimensional

978-1-5090-1172-8/17/$31.00 ©2017 IEEE

parameterized deformations via principal component analysis. Wang et al. [18] introduce a framework that involves key-point matching using sparse learning, followed by dense deformation field interpolation, which heavily depends on the accuracy of key-point selection. Cao et al. [21] propose a semi-coupled dictionary learning approach to jointly model unimodal image appearance and deformation parameters, but only a linear relationship between the two is assumed. The two closest methods to our proposed approach are [20, 19]. Yang et al. [20] use deep learning to model the nonlinear relationship between image appearance and LDDMM deformation parameters, but only consider unimodal atlas-to-image registration, instead of general image-to-image registration. Gutiérez-Becker et al. [19] learn a multimodal similarity measure using a regression forest with Haar-like features combined with a prediction model for a low-dimensional parametric B-spline model (5 nodes/dimension). In contrast, our approach (i) predicts the initial momentum of the shooting formulation of LDDMM [22], a nonparametric registration model[1]; and (ii) jointly learns a multimodal similarity measure from image-patches without requiring feature selection.

Contributions. We propose a deep learning architecture to *jointly* learn a multimodal similarity measure and the relationship between images and deformation parameters. Specifically, we design a deep encoder-decoder network to predict the initial momentum of LDDMM using multimodal image patches for *image-to-image registration* (opposed to atlas-to-image registrations as in [20]). We focus on LDDMM, but our method is applicable to other registration models. Our contributions are: (i) a patch-based deep network that generates accurate deformation parameter predictions using multimodal images; (ii) the simultaneous learning of a multimodal image similarity measure based only on multimodal image patches; (iii) an order of magnitude speedup, compared to optimization-based (GPU-accelerated) registration of 3D images; and (iv) a Bayesian extension of our model to provide uncertainty estimates for predicted deformations.

Organization. Sec. 2 reviews the initial momentum formulation of LDDMM and discusses our motivation for this parameterization. Sec. 3 introduces our deterministic and Bayesian network structure, as well as our method of speeding up the computation. Sec. 4 presents experimental results on a 3D autism brain dataset. Sec. 5 discusses potential future research directions, experiments, and possible extensions.

2. INITIAL MOMENTUM LDDMM SHOOTING

Given a moving image S and a target image T, LDDMM estimates a diffeomorphism φ such that $S \circ \varphi^{-1} \approx T$, where $\varphi \doteq \phi(1)$ is generated via a smooth flow $\phi(t), t \in [0,1]$. The parameter for the LDDMM shooting formulation is the *initial momentum vector field* $m(t)$, which is used to compute $\phi(t)$. The initial momentum is the dual of the initial velocity field $v(t)$ which is an element of a reproducing kernel Hilbert space V. The vector fields m and v are connected by a positive definite self-adjoint smoothing kernel K via $v = Km$ and $m = Lv$, where L denotes the inverse of K. The energy function for the shooting formulation of LDDMM is [25, 22]

$$E(m_0) = \langle m_0, Km_0 \rangle + \frac{1}{\sigma^2} ||S \circ \phi^{-1}(1) - T||^2, \text{where} \quad (1)$$

$$\begin{aligned} m_t + \text{ad}_v^* m = 0, m(0) = m_0, \\ \phi_t^{-1} + D\phi^{-1} v = 0, \quad \phi^{-1}(0) = \text{id}, \quad m - Lv = 0. \end{aligned} \quad (2)$$

Here, id is the identity map, ad* is the negated Jacobi-Lie bracket of vector fields, i.e., $\text{ad}_v m = Dmv - Dvm$, D denotes the Jacobian operator, and subscript t denotes the derivative w.r.t. time t. As in [20], we choose to *predict the initial momentum* m_0. This is motivated by the observation that the momentum parameterization (unlike parameterization via displacement or vector fields) allows patch-wise prediction, because the momentum is non-zero only on image edges (in theory, $m = \lambda \nabla I$ for images, where λ is a scalar field). Thus, from a theoretical point of view, no information is needed outside the patch for momentum prediction, and $m = 0$ in homogeneous regions. Furthermore, deformation smoothness is guaranteed via K *after* the prediction step. I.e., given a sufficiently strong regularizer, L, diffeomorphic deformations are obtained by integrating Eq. (2).

3. NETWORK STRUCTURE

Fig. 1 shows our network structure for 3D multimodal image deformation prediction. This network is an encoder-decoder network, where *two* encoders compute features from the moving/target image patches independently. The learned features from the two encoders are simply concatenated and sent to *three* decoders to generate one initial momentum patch for each dimension. The two encoder structure is different from the network in [20] since using two encoders instead of a single encoder with two initial input channels has the effect of reducing overfitting, as shown in Sec. 4. All convolutional layers use $3 \times 3 \times 3$ filters, and we choose PReLU [26] as the non-linear activation layer. Pooling and unpooling layers are usually used for multi-scale image processing, but unlike [20] where unpooling based on pooling index is straightforward, they are not suitable for our formulation since the two encoders perform pooling independently. Therefore, we follow the idea of [27] and use convolutional layers with a stride of 2 and deconvolution layers [28] as surrogates for pooling and unpooling layers. I.e., the network learns its pooling and unpooling operations during training. For training, we use the L1 norm as our similarity criterion. To predict the full image momentum during testing, we use a sliding window approach, predicting the momentum patch-by-patch, followed

[1] The relaxation formulation of LDDMM [23] is for example the basis of the successful ANTs [24] registration tools.

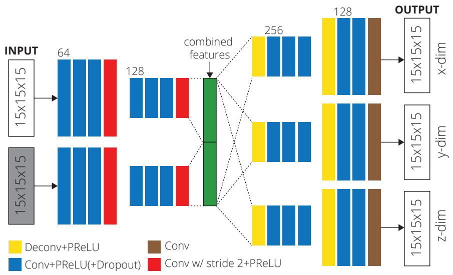

Fig. 1: 3D Bayesian network structure. The network takes 2 3D patches from moving/target images and generates 3 3D initial momentum patches (in 3 dim.). Numbers for the input and output denote input/output patch size, and numbers in the network indicate the number of features (output channels) for the conv. layers. To create a deterministic network, we remove all dropout layers.

by averaging the momenta in overlapping areas. We refer to this architecture as our *deterministic network.*

We extend the network to a *Bayesian network* by adding dropout layers [29] after all convolution+PReLU layers except for the "pooling/unpooling" (convolutional) layers [30]. Given the input(I)/output(O) of a convolutional layer and the convolutional layer's weight, W, adding dropout can be regarded as performing variational inference for the posterior distribution of the conv layer's weight $p(W|I, O)$ using a Bernoulli distribution $q(W) = W * (\text{Bernoulli}(p)_i)|_{i=1}^{N}$, where p is the dropout probability, N is the number of nodes in the conv layer, and $\text{Bernoulli}(p)_i$ randomly sets the ith node in the conv layer to 0. According to [30], training this dropout network is equivalent to minimizing the KL-divergence between the variational and true posterior $\text{KL}(q(W)|p(W|I, O))$. During testing, we keep the dropout layers and sample the network to obtain multiple momentum predictions for a single image. We choose the sample mean as the final momentum prediction and perform LDDMM shooting (by integrating Eq. (2)) for the momentum samples to generate multiple deformation fields. The variances of the deformation fields then estimate the predicted deformation *uncertainty*. We set the dropout probability to 0.3.

Patch pruning. We achieve a substantial speedup in computation time by using a *patch pruning* strategy. Specifically, we apply a large sliding window stride and ignore patches only containing the background of both the moving image and the target image. In our experiments, these technique reduced the number of patches to predict by approximately 99.85% for 229 × 193 × 193 3D brain images with 15 × 15 × 15 patch size and a sliding window stride of 14.

4. EXPERIMENTS

We assess our approach on the IBIS 3D Autism Brain image dataset [31], containing 375 T1w/T1w brain images (229 × 193 × 193) of 2 years subject with pediatric disease. We first register all images affinely to the ICBM 152 atlas [32] and select 359 images for training and the remaining 16 images for testing. For training, we randomly select 359 T1w-T1w image pairs and perform LDDMM registration using PyCA[2] on GPU with SSD as image similarity measure. We set the parameters for the LDDMM regularizer $L = a\Delta^2 + b\Delta + c$ to $[a, b, c] = [0.01, 0.01, 0.001]$, and σ in Eqn.1 to 0.2. *We then train the network to predict the momenta generated from the T1w-T1w registrations using their corresponding T1w and T2w images.* The network is implemented in `Torch` on a TITAN X GPU and is optimized (over 10 epochs) using `rmsprop` with a learning rate of $= 0.0001$, momentum decay $= 0.1$ and update decay $= 0.01$. For testing, we perform T1w-T2w pairwise registration predictions for all 16 test images, excluding self-registrations. This results in a total of 240 test cases. Each prediction result is compared to the T1w-T1w registrations obtained via LDDMM optimization (used as ground truth). The patch size for the 3D network is 15 × 15 × 15 and we use a sliding window size of 14 for both training and testing. For comparison to the ground truth deformation from LDDMM optimization, we trained another network using T1w-T1w data to perform prediction on the T1w-T1w registration cases. This network serves as the "upper limit" of our multimodal network's potential performance. We also implemented the architecture from [20] and train the network using the T1w-T2w data for comparison. The deformation errors are calculated as the 2-norm of the voxel-wise difference between the predicted deformations and the deformations obtained from LDDMM optimization.

Tab. 1(top) lists the evaluation results: our multimodal network (T1w-T2w) greatly reduces deformation error compared to affine registration, and only has a slight accuracy loss compared to the T1w-T1w network. This demonstrates registration consistency (to the T1w-T1w registration result) of our approach achieved by *learning* the similarity measure between the two modalities. Moreover, the deformation error data percentiles of Tab. 1 show that our network achieves a slight deformation error decrease of 2.1%∼7.3% for all data percentiles compared to [20]. This is likely due to less overfitting by using two small encoders.

We also test our network for registration tasks with limited training data. To do so, we randomly choose only 10 out of the 359 training images to perform pairwise registration, generating 90 T1w-T1w registration pairs. We then use these 90 registrations to train our T1w-T2w network model. Tab. 1 shows that although the network used only 10 images for training, performance only decreases slightly in comparison to our T1w-T2w network using 359 image pairs for training. Hence, by using patches, our network model can also be successfully trained with limited training images.

To further test our network's consistency in relation to the T1w-T1w prediction results, we calculate the deformation error of our T1w-T2w network w.r.t the T1w-T1w network. For comparison, we also run `NiftyReg` [33] B-spline registration on both T1w-T1w and T1w-T2w test cases using nor-

[2] `https://bitbucket.org/scicompanat/pyca`

	Deformation Error w.r.t LDDMM optimization on T1w-T1w data [voxel]						
Data Percentile	0.3%	5%	25%	50%	75%	95%	99.7%
Affine (Baseline)	0.1664	0.46	0.9376	1.4329	2.0952	3.5037	6.2576
Ours, T1w-T1w data	0.0353	0.0951	0.1881	0.2839	0.416	0.714	1.4409
[20], T1w-T2w data	0.0582	0.1568	0.3096	0.4651	0.6737	1.1106	2.0628
Ours, T1w-T2w data	0.0551	0.1484	0.2915	0.4345	0.6243	1.0302	2.0177
Ours, T1w-T2w data, 10 images	0.0663	0.1782	0.3489	0.5208	0.752	1.2421	2.3454
	Prediction/Optimization error between T1w-T2w and T1w-T1w [voxel]						
Data Percentile	0.3%	5%	25%	50%	75%	95%	99.7%
Ours	0.0424	0.1152	0.2292	0.3444	0.4978	0.8277	1.6959
`NiftyReg` (Baseline)	0.2497	0.7463	1.8234	3.1719	5.1124	8.9522	14.4666

Table 1: Evaluation results for the 3D dataset.

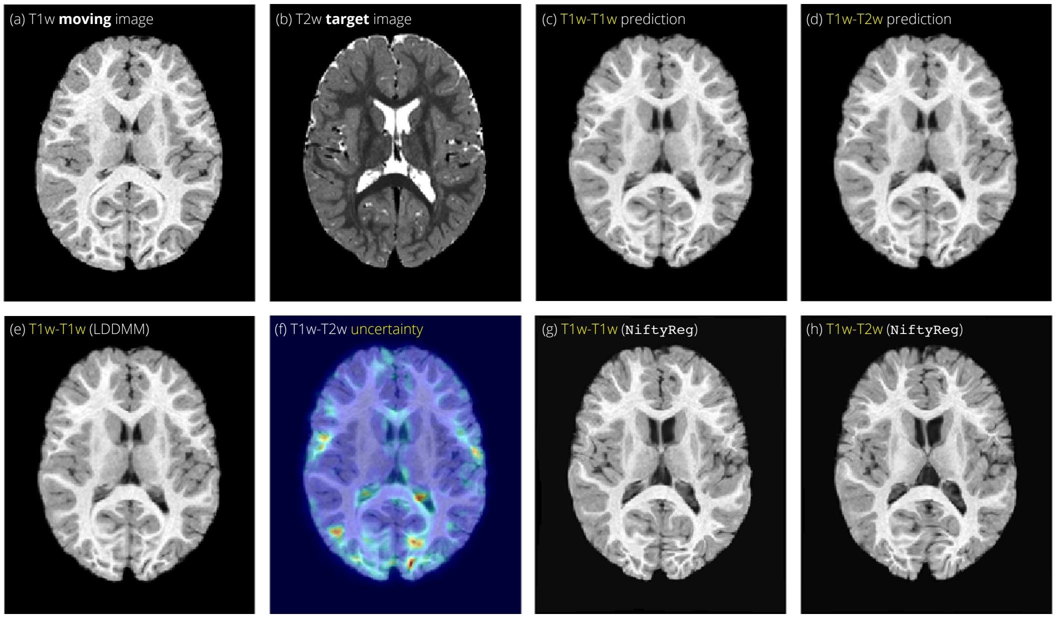

Fig. 2: Exemplary test case. (a) T1w moving image; (b) T2w target image; (c)-(d) deformation prediction result from T1w-T1w/T1w-T2w data; (e) deformation result by LDDMM optimization for T1w-T1w registration; (f) uncertainty of predicted T1w-T2w deformation as the 2-norm of the sum of variances of deformation fields in all spatial directions, mapped on the predicted T1w-T2w wrapped image. Yellow = more uncertainty, blue = less uncertainty; (g)+(h) `NiftyReg` registration result for T1w-T1w/T1w-T2w pair.

malized mutual information (NMI) with a grid size of 4 and a bending energy weight of 0.0001; we compare the deformation error between T1w-T2w and T1w-T1w registrations, see Tab. 1(bottom). Compared to `NiftyReg`, our method is more consistent for multimodal prediction. Fig. 2 shows one test case: using `NiftyReg` generates large differences in the ventricle area between the T1w-T1w and T1w-T2w cases, while our approach does not. We attribute this result to the shortcomings of NMI and not to `NiftyReg` as a registration method. We also computed the 2-norm of the sum of variances of deformation fields in all directions as the uncertainty of the deformation, shown in Fig. 2(f). We observe high uncertainty around the ventricle, due to the drastic appearance change in this area between the moving and the target image. **Computation time**. On average, our method requires 24.46s per case. Compared to (GPU) LDDMM optimization, we achieve a 36x speedup. Further speedups can be achieved by using multiple GPUs for independent patch predictions.

5. DISCUSSION AND SUPPORT

We proposed a fast method for multimodal image registration which simultaneously (i) learns the multimodal image similarity measure from image-patches and (ii) predicts registrations based on LDDMM, thereby guaranteeing diffeomorphic transformations. Different from [20], we use a different network structure for multimodal instead of unimodal image registration, and choose a new strategy to train the network. Our method shows good prediction performance and high consistency of the multimodal registration result in comparison to unimodal registration. Future work should test the registration performance via landmarks or volumetric overlap measures. Comparisons to other registration approaches and a direct LDDMM implementation with a standard multimodal similarity measure such as MI would also be desirable.

This work is supported by NSF EECS-1148870, NIH 1 R41 NS091792-01, R01-HD055741, R01-HD059854 and U54-HD079124. We thank NVIDIA for the GPU donation.

6. REFERENCES

[1] P. Viola and W. M. Wells III, "Alignment by maximization of mutual information," *IJCV*, vol. 24, no. 2, pp. 137–154, 1997.

[2] C. Meyer, J.L. Boes, B. Kim, and P. Bland, "Evaluation of control point selection in automatic, mutual information driven, 3D warping," in *MICCAI*, 1998.

[3] G. Hermosillo, C. Chefd'Hotel, and O. Faugeras, "Variational methods for multimodal image matching," *IJCV*, vol. 50, no. 3, pp. 329–343, 2002.

[4] P. Lorenzen, M. Prastawa, B. Davis, G. Gerig, E. Bullitt, and S. Joshi, "Multi-modal image set registration and atlas formation," *MedIA*, vol. 10, no. 3, pp. 440–451, 2006.

[5] C. Guetter, C. Xu, F. Sauer, and J. Hornegger, "Learning based non-rigid multi-modal image registration using Kullback-Leibler divergence," in *MICCAI*, 2005.

[6] D. Lee, M. Hofmann, F. Steinke, Y. Altun, ND. Cahill, and B. Schölkopf, "Learning similarity measure for multi-modal 3D image registration," in *CVPR*, 2009.

[7] F. Michel, M. Bronstein, A. M. Bronstein, and N. Paragios, "Boosted metric learning for 3D multi-modal deformable registration," in *ISBI*, 2011.

[8] X. Cheng and L. Zhang Y. Zheng, "Deep similarity learning for multimodal medical images," *Comput. Methods Biomech. Biomed. Eng. Imaging Vis.*, pp. 1–5, 2015.

[9] M. Simonovsky, B. Gutiérrez-Becker, D. Mateus, N. Navab, and N. Komodakis, "A deep metric for multimodal registration," *MICCAI*, 2016.

[10] Snehashis Roy, Aaron Carass, and J Prince, "Magnetic resonance image example based contrast synthesis," *TMI*, vol. 32, no. 12, pp. 2348–2363, 2013.

[11] Amod Jog, Snehashis Roy, Aaron Carass, and Jerry L Prince, "Magnetic resonance image synthesis through patch regression," in *ISBI*, 2013.

[12] Hien Van Nguyen, Kevin Zhou, and Raviteja Vemulapalli, "Cross-domain synthesis of medical images using efficient location-sensitive deep network," in *MICCAI*, 2015.

[13] M. Holden, "A review of geometric transformations for nonrigid body registration," *TMI*, vol. 27, no. 1, pp. 111–128, 2008.

[14] J. Modersitzki, *Numerical methods for image registration*, Oxford University Press, 2004.

[15] P. Weinzaepfel, J. Revaud, Z. Harchaoui, and C. Schmid, "DeepFlow: Large displacement optical flow with deep matching," in *ICCV*, 2013.

[16] A. Dosovitskiy, P. Fischery, E. Ilg, C. Hazirbas, V. Golkov, P. van der Smagt, D. Cremers, and T. Brox, "Flownet: Learning optical flow with convolutional networks," in *ICCV*, 2015.

[17] C-R. Chou, B. Frederick, G. Mageras, S. Chang, and S. Pizer, "2D/3D image registration using regression learning," *CVIU*, vol. 117, no. 9, pp. 1095–1106, 2013.

[18] Q. Wang, M. Kim, Y. Shi, G. Wu, and D. Shen, "Predict brain MR image registration via sparse learning of appearance & transformation," *MedIA*, vol. 20, no. 1, pp. 61–75, 2015.

[19] B. Gutiérez-Becker, D. Mateus, L. Peter, and N. Navab, "Learning optimization updates for multimodal registration," in *MICCAI*, 2016.

[20] X. Yang, R. Kwitt, and M. Niethammer, "Fast predictive image registration," in *DLMIA*, 2016.

[21] T. Cao, N. Singh, V. Jojic, and M. Niethammer, "Semi-coupled dictionary learning for deformation prediction," in *ISBI*, 2015.

[22] F.X. Vialard, L. Risser, D. Rueckert, and C.J. Cotter, "Diffeomorphic 3D image registration via geodesic shooting using an efficient adjoint calculation," *IJCV*, vol. 97, no. 2, pp. 229–241, 2012.

[23] M. F. Beg, M. I. Miller, A. Trouvé, and L. Younes, "Computing large deformation metric mappings via geodesic flows of diffeomorphisms," *IJCV*, vol. 61, no. 2, pp. 139–157, 2005.

[24] B. B. Avants, N. J. Tustison, G. Song, P. A. Cook, A. Klein, and J. C. Gee, "A reproducible evaluation of ants similarity metric performance in brain image registration," *Neuroimage*, vol. 54, no. 3, pp. 2033–2044, 2011.

[25] N. Singh, J. Hinkle, S. Joshi, and P.T. Fletcher, "A vector momenta formulation of diffeomorphisms for improved geodesic regression and atlas construction," in *ISBI*, 2013.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," *arXiv:1502.01852*, 2015.

[27] J.T. Springenberg, A.Dosovitskiy, T. Brox, and M.A. Riedmiller, "Striving for simplicity: The all convolutional net," *arXiv:1412.6806*, 2014.

[28] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *CVPR*, 2015.

[29] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *JMLR*, vol. 15, pp. 1929–1958, 2014.

[30] Y. Gal and Z. Ghahramani, "Bayesian convolutional neural networks with Bernoulli approximate variational inference," *arXiv:1506.02158*, 2015.

[31] J. J. Wolff, G. Gerig, J.D. Lewis, T. Soda, M.A. Styner, C. Vachet, K. Botteron, J.T. Elison, S.R. Dager, A.M. Estes, H.C. Hazlett, R.T. Schultz, L. Zwaigenbaum, and J. Piven, "Altered corpus callosum morphology associated with autism over the first 2 years of life," *Brain*, vol. 138, no. 7, pp. 2046–2058, 2015.

[32] V. Fonov, A.C. Evans, C.R. Almli K. Botteron, R.C. McKinstry, and D.L. Collins, "Unbiased average age-appropriate atlases for pediatric studies.," *NeuroImage*, vol. 54, no. 1, pp. 313–327, 2011.

[33] M. Modat, G.R. Ridgway, Z.A. Taylor, M. Lehmann, J. Barnes, D.J. Hawkes, N.C. Fox, and S. Ourselin, "Fast free-form deformation using graphics processing units," *Comput. Methods Prog. Biomed*, vol. 98, no. 3, pp. 278–284, 2010.