DeepAtlas: Joint Semi-supervised Learning of Image Registration and Segmentation

Zhenlin Xu$^{(\boxtimes)}$ and Marc Niethammer

University of North Carolina, Chapel Hill, NC, USA
`zhenlinx@cs.unc.edu`

Abstract. Deep convolutional neural networks (CNNs) are state-of-the-art for semantic image segmentation, but typically require many labeled training samples. Obtaining 3D segmentations of medical images for supervised training is difficult and labor intensive. Motivated by classical approaches for joint segmentation and registration we therefore propose a deep learning framework that jointly learns networks for image registration and image segmentation. In contrast to previous work on deep unsupervised image registration, which showed the benefit of weak supervision via image segmentations, our approach can use existing segmentations when available and computes them via the segmentation network otherwise, thereby providing the same registration benefit. Conversely, segmentation network training benefits from the registration, which essentially provides a realistic form of data augmentation. Experiments on knee and brain 3D magnetic resonance (MR) images show that our approach achieves large simultaneous improvements of segmentation and registration accuracy (over independently trained networks) and allows training high-quality models with very limited training data. Specifically, in a one-shot-scenario (with only one manually labeled image) our approach increases Dice scores (%) over an unsupervised registration network by 2.7 and 1.8 on the knee and brain images respectively.

1 Introduction

Image segmentation and registration are two crucial tasks in medical image analysis. They are also highly related and can help each other. *E.g.*, labeled atlas images are used via image registration for segmentation. Segmentations can also provide extra supervision (in addition to image intensities) for image registration and are used to evaluate registration results. Consequentially, joint image registration and segmentation approaches have been proposed. *E.g.*, approaches based on active-contours [11] and Bayesian [7] or Markov random field formulations [6]. While these

methods jointly estimate registration and segmentation results, they operate on *individual* image pairs (instead of a population of images) and require the computationally costly minimization of an energy function.

Deep learning (DL) has been widely and successfully applied to medical image analysis. For supervised image segmentation, CNN-based approaches are faster and better than classical methods when many labeled training samples are available [5]. DL-based registration achieves similar performance to optimization-based approaches but is much faster. As true transformations are not available, training either uses estimates from optimization-based methods [10] or is unsupervised [2]. Recent work [3] shows that weak supervision via an additional image segmentation loss between registered images can improve results over unsupervised training, which relies on the images alone. In practice, obtaining segmentations for 3D medical images is difficult and labor intensive. Hence, manual segmentations will often not be available for a large fraction of image data.

We propose **DeepAtlas**, to jointly learn deep networks for weakly supervised registration and semi-supervised segmentation. Our contributions are:

- *We propose the first approach to jointly learn two deep neural networks for image registration and segmentation.* Previous joint approaches require joint optimizations for each image pair. Instead, we jointly learn from a population of images during training, but can independently use the resulting segmentation and registration networks at test time.
- *Our joint approach only requires few manual segmentations.* Our two networks mutually guide each other's training on unlabeled images via an anatomy similarity loss. This loss penalizes the dissimilarity of the warped segmentation of the moving image and the segmentation of the target image. When registering image pairs consisting of a manually labeled image and the estimate of a labeled image (via its network-predicted segmentation), this loss provides anatomy consistency supervision for registration and forces the predicted segmentation to match the manual segmentation after registration.
- *We evaluate our approach on large 3D brain and knee MRI datasets.* Using few manual segmentations, our method outperforms separately learned registration and segmentation networks. In the extreme case, where only one manually segmented image is available, our approach facilitates one-shot segmentation and boosts registration performance at the same time.

2   Method

Our goal is to improve registration and segmentation accuracy when few manual segmentations are available for a large set of images by jointly learning a segmentation and a registration network. Figure 1 illustrates our approach consisting of two parts: weakly-supervised registration learning (solid blue lines) and semi-supervised segmentation learning (dashed yellow lines). Our loss is the weighted sum of the registration regularization loss ($\mathcal{L}_r$), the image similarity loss ($\mathcal{L}_i$), the anatomy loss ($\mathcal{L}_a$) penalizing segmentation dissimilarity, and the supervised segmentation loss ($\mathcal{L}_{sp}$). The losses $\{\mathcal{L}_r,\ \mathcal{L}_i,\ \mathcal{L}_a\}$ drive the weakly

supervised learning of registration (Sect. 2.1) and the losses $\{\mathcal{L}_a, \mathcal{L}_{sp}\}$ drive the semi-supervised learning of segmentation (Sect. 2.2). Section 2.3 details the implementation.

2.1   Weakly-Supervised Registration Learning

Given a pair of moving and target images I_m and I_t, a registration network $\mathcal{F}_R$ with parameters θ_r predicts a displacement field $\mathbf{u} = \mathcal{F}_R(I_m, I_t; \theta_r)$. This then allows warping the moving image to the target image space, $I_m^w = I_m \circ \Phi^{-1}$, where $\Phi^{-1} = \mathbf{u} + \mathrm{id}$ is the deformation map and id is the identity transform. A good map, Φ, maps related anatomical positions to each other. Unsupervised registration learning optimizes θ_r over an intensity similarity loss $\mathcal{L}_i$ (penalizing appearance differences between I_t and I_m^w) and a regularization loss $\mathcal{L}_r$ on $\mathbf{u}$ to encourage smooth transformations. Adding weak supervision by also matching segmentations between the target image (S_t) and the warped moving image ($S_m^w = S_m \circ \Phi^{-1}$) via an anatomy similarity loss $\mathcal{L}_a$ can improve registrations [3]. Weakly-supervised registration learning is then formulated as:

$$\theta_r^{\star} = \underset{\theta_r}{\operatorname{argmin}} \{\mathcal{L}_i(I_m \circ \Phi^{-1}, I_t) + \lambda_r \mathcal{L}_r(\Phi^{-1}) + \lambda_a \mathcal{L}_a(S_m \circ \Phi^{-1}, S_t)\}, \quad (1)$$

with weights $\lambda_r, \lambda_a \geq 0$. In practice, while a large set of images are often available, few of them have manual segmentations. In contrast to existing work, we estimate missing moving or target segmentations via our segmentation network (see Fig. 1). Hence, we provide weak supervision for *every* training image pair.
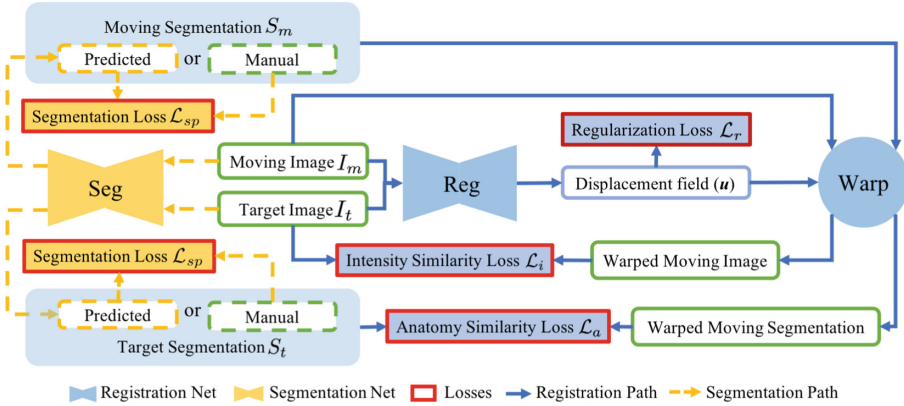

Fig. 1. *DeepAtlas* for joint learning of weakly supervised registration and semi-supervised segmentation. Unlabeled moving/target images are segmented by the segmentation network so that every training registration pair has weak supervision via the anatomy similarity loss which also guides segmentation learning on unlabeled images. (Color figure online)

2.2 Semi-supervised Segmentation Learning

The segmentation network $\mathcal{F}_S$ with parameters θ_s takes an image I as input and generates probabilistic segmentation maps for all semantic classes: $\hat{S} = \mathcal{F}_S(I; \theta_s)$. In addition to the typical supervised segmentation loss $\mathcal{L}_{sp}(\hat{S}, S)$ where S is a given manual segmentation, the anatomy similarity loss for registration $\mathcal{L}_a(S_m \circ \Phi^{-1}, S_t)$ also drives segmentation learning when S_m or S_t are predicted via $\mathcal{F}_S$ for unlabeled images. Specifically, we define these losses as:

$$\mathcal{L}_{seg} = \begin{cases} \lambda_a \mathcal{L}_a(S_m \circ \Phi^{-1}, \mathcal{F}_S(I_t)) + \lambda_{sp} \mathcal{L}_{sp}(\mathcal{F}_S(I_m), S_m), & \text{if } I_t \text{ is unlabeled;} \\ \lambda_a \mathcal{L}_a(\mathcal{F}_S(I_m) \circ \Phi^{-1}, S_t) + \lambda_{sp} \mathcal{L}_{sp}(\mathcal{F}_S(I_t), S_t), & \text{if } I_m \text{ is unlabeled;} \\ \lambda_a \mathcal{L}_a(S_m \circ \Phi^{-1}, S_t) + \lambda_{sp} \mathcal{L}_{sp}(\mathcal{F}_S(I_m), S_m), & \text{if } I_m \text{ and } I_t \text{ are labeled;} \\ 0, & \text{if both } I_t \text{ and } I_m \text{ are unlabeled.} \end{cases}$$

with weights $\lambda_a, \lambda_{sp} \geq 0$. $\mathcal{L}_a$ teaches $\mathcal{F}_S$ to segment an unlabeled image such that the predicted segmentation matches the manual segmentation of a labeled image via $\mathcal{F}_R$. In the case where the target image I_t is unlabeled, $\mathcal{L}_a$ is equivalent to a supervised segmentation loss on I_t, in which the single-atlas segmentation $S_m \circ \Phi^{-1}$ is the noisy true label. Note that we do not use two unlabeled images for training and $\mathcal{L}_a$ does not train the segmentation network when both images are labeled. We then train our segmentation network in a semi-supervised manner as follows:

$$\theta_s^{\star} = \underset{\theta_s}{\operatorname{argmin}} \ \mathcal{L}_{seg}. \tag{2}$$

2.3 Implementation Details

Losses: Various choices are possible for the intensity/anatomy similarity, the segmentation, and the regularization losses. Our choices are as follows.

Anatomy Similarity and Supervised Segmentation Loss: A cross-entropy loss requires manually tuned class weights for imbalanced multi-class segmentations [8]. We use a soft multi-class Dice loss which addresses imbalances inherently:

$$\mathcal{L}_{dice}(S, S^{\star}) = 1 - \frac{1}{K} \sum_{k=1}^{K} \frac{\sum_x S_k(x) S_k^{\star}(x)}{\sum_x S_k(x) + \sum_x S_k^{\star}(x)}, \tag{3}$$

where k indicates a segmentation label (out of K) and x is voxel location. S and $S^{\star}$ are two segmentations to be compared.

Intensity Similarity Loss: We use normalized cross correlation (NCC) as:

$$\mathcal{L}_i(I_m^w, I_t) = 1 - NCC(I_m^w, I_t), \tag{4}$$

which will be in $[0, 2]$ and hence will encourage maximal correlation.

Regularization Loss: We use the bending energy [9]:

$$\mathcal{L}_r(\mathbf{u}) = \frac{1}{N} \sum_{\mathbf{x}} \sum_{i=1}^{d} \|H(u_i(\mathbf{x}))\|_F^2 \tag{5}$$

where $\| \cdot \|_F$ denotes the Frobenius norm, $H(u_i(\mathbf{x}))$ is the Hessian of the i-th component of $\mathbf{u}(\mathbf{x})$, and d denotes the spatial dimension ($d = 3$ in our case). N denotes the number of voxels. Note that this is a second-order generalization of diffusion regularization, where one penalizes $\|\nabla u_i(\mathbf{x})\|_2^2$ instead of $\|H(u_i(\mathbf{x}))\|_F^2$.

Alternating Training: It is in principle straightforward to optimize two networks according to Eqs. 1 and 2. However, as we work with the whole 3D images, not cropped patches, GPU memory is insufficient to simultaneously optimize the two networks in one forward pass. Hence, we alternately train one of the two networks while keeping the other fixed. We use a 1:20 ratio between training steps for the segmentation and registration networks, as the segmentation network converges faster. Since it is difficult to jointly train from scratch with unlabeled images, we independently pretrain both networks. When only few manual segmentations are available, *e.g.*, only one, separately training the segmentation network is challenging. In this case, we train the segmentation network from scratch using a fixed registration network trained unsupervisedly. We start alternating training when the segmentation network achieves reasonable performance.

Networks: DeepAtlas can use any CNN architecture for registration and segmentation. We use the network design of [2] for registration; and a customized light 3D U-Net design for segmentation with LeakyReLU instead of ReLU, and smaller feature size due to GPU memory limitations.

3   Experiments and Results

We show on a 3D knee and a 3D brain MRI dataset that our framework improves both registration and segmentation when many images with few manual segmentations are available: i.e. N of M images are labeled ($N \ll M$).

Mono-networks: We train single segmentation/registration models as baselines. For segmentation, fully supervised networks are trained with N labeled images; the registration networks are trained via Eq. 1 using all M training images with N images labeled; the anatomy similarity loss, $\mathcal{L}_a$, is only used for training pairs where both images have manual segmentations. Models trained with $N = M$ manual segmentations (*i.e.*, with manual segmentations for all images) provide our upper performance bound. All mono-networks are trained for a sufficient number of epochs until they over-fit. The best models based on validation performance are evaluated.
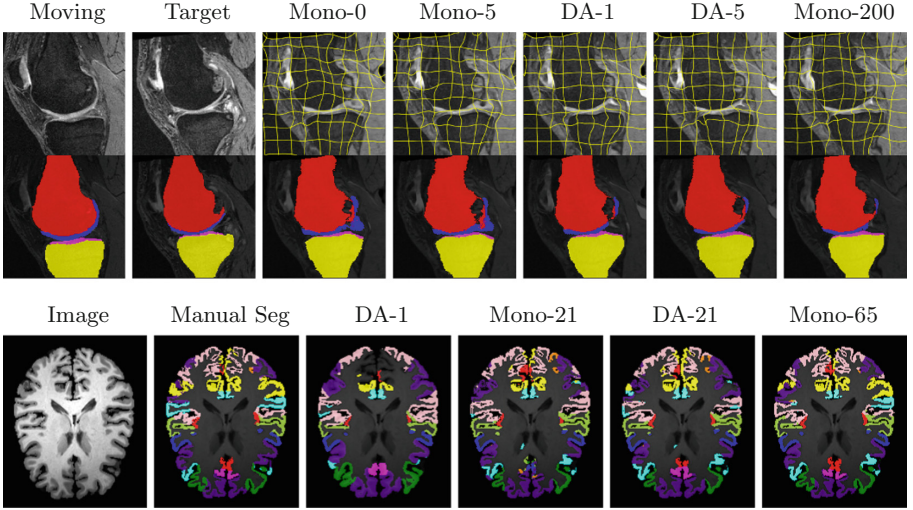
| Moving | Target | Mono-0 | Mono-5 | DA-1 | DA-5 | Mono-200 |

| Image | Manual Seg | DA-1 | Mono-21 | DA-21 | Mono-65 |

Fig. 2. Examples of knee MRI registration (top) and brain MRI segmentation (bottom) results. **Top:** The first two columns are the moving image/segmentation and the target image/segmentation followed by the warped moving images (with deformation grids)/segmentations by different models. **Bottom left to right:** original image, manual segmentation, and predictions of various models. Mono-i and DA-i represent the mono- and DA models with i manual segmentations respectively.

DeepAtlas (DA): We initialize the joint model with the trained mono-networks. In addition to the alternately trained DA models, we hold one network fixed all through training, termed **Semi-DeepAtlas (Semi-DA)**.

In one-shot learning (N = 1) experiments, training a supervised segmentation network based on a single labeled image is difficult; hence, we do not compute a segmentation mono-network in this case. For Semi-DA, we train a segmentation network from scratch with a fixed registration network that is trained unsupervised (N = 0). The DA model is initialized using the Semi-DA segmentation network and the unsupervised registration network.

Knee MRI Experiment: We test our method on 3D knee MRIs from the Osteoarthritis Initiative (OAI)[1] and corresponding segmentations of femur and tibia as well as femoral and tibial cartilage [1]. From a total of 507 labeled images, we use 200 for training, 53 for validation, and 254 for testing. To test registration performance we use 10,000 random image pairs from the test set. All images are affinely registered to an atlas built from the training images, resampled to isotropic spacing of 1 mm, cropped to $160 \times 160 \times 160$ and intensity normalized to [0, 1]. In addition, right knee images are flipped to be consistent with left knees. For training, the loss weights are $\lambda_r = 20{,}000$, $\lambda_a = 3$, and $\lambda_{sp} = 3$ based on

---

[1] https://nda.nih.gov/oai/.

approximate hyper-parameter tuning. Note that when computing $\mathcal{L}_r$ from the displacements, the image coordinates are scaled to $[-1, 1]$ for each dimension following the convention in the interpolation function of PyTorch.

Brain MRI Experiment: We also evaluate our method on the MindBooggle101 [4] brain MRIs with 32 cortical regions. We fuse corresponding segmentation labels of the left and right brain hemispheres. MindBoogle101 consists of images from multiple datasets, *e.g.*, OASIS-TRT-20, MMRR-21 and HLN-12. After removing images with incorrect labels, we obtain a total of 85 images. We use 5 images from OASIS-TRT-20 as validation set and 15 as test set. We use the remaining 65 images for training. Manual segmentations in the $N = 1$ and $N = 21$ experiments are only from the MMRR-21 subset; this simulates a common practical use case, where we only have few manual segmentations for one dataset and additional unlabeled images from other datasets, but desire to process a different, new dataset. All images are 1 mm isotropic, affinely-aligned, histogram-matched, and cropped to size $168 \times 200 \times 169$. We apply sagittal flipping for training data augmentation. We use the same loss weights as for the knee MRI experiment except for $\lambda_r = 5,000$, since cross-subject brain registrations require large deformations and hence less regularization.

Table 1. Segmentation and registration performance on 3D knee MRIs. Average (standard deviation) of Dice scores (%) for bones (femur and tibia) and cartilages (femoral and tibial). N of 200 training images are manually labeled.

N	Models	Segmentation Dice (%)			Registration Dice (%)		
		Bones	Cartilages	All	Bones	Cartilages	All
0	Mono	–	–	–	95.32(1.13)	65.71(5.86)	80.52(3.24)
1	Semi-DA	96.43(0.85)	76.67(3.24)	86.55(1.86)	–	–	–
	DA	**96.80(0.81)**	**77.63(3.22)**	**87.21(1.84)**	95.76(1.01)	70.77(5.68)	83.27(3.14)
5	Mono	96.51(1.69)	78.95(3.91)	87.73(2.37)	95.60(1.08)	68.13(5.98)	81.87(3.31)
	Semi-DA	96.97(1.26)	79.73(3.84)	88.35(2.22)	**96.38(0.81)**	73.48(5.26)	84.93(2.89)
	DA	**97.49(0.67)**	**80.35(3.64)**	**88.92(2.01)**	96.35(0.82)	**73.67(5.22)**	**85.01(2.86)**
10	Mono	97.29(1.03)	80.59(3.67)	88.94(2.07)	95.77(1.02)	69.45(5.93)	82.61(3.27)
	Semi-DA	97.60(0.76)	**81.21(3.58)**	89.40(1.99)	**96.66(0.72)**	74.67(5.01)	**85.66(2.73)**
	DA	**97.70(0.65)**	81.19(3.47)	**89.45(1.91)**	96.62(0.75)	**74.69(5.03)**	85.66(2.75)
200	Mono	98.24(0.34)	83.54(2.93)	90.89(1.56)	96.98(0.56)	77.33(4.34)	87.16(2.35)

Optimizer: We use `Adam`. The initial learning rates are 1e−3 for the mono-networks. Initial learning rates are 5e−4 for the registration network and 1e−4 for the segmentation network for Semi-DA and DA. Learning rates decay by 0.2 at various epochs across experiments. We use PyTorch and run on Nvidia V100 GPUs with 16 GB memory.

Results: All trained networks are evaluated using Dice overlap scores between predictions and the manual segmentations for the segmentation network,

or between the warped moving segmentations and the target segmentations for the registration network. Tables 1 and 2 show results for the knee and brain MRI experiments respectively in Dice scores (%). Figure 2 shows examples of knee MRI registrations and brain MRI segmentations.

General Results: For both datasets across different numbers of manual segmentations, Semi-DA, which uses a fixed pre-trained network to help the training of the other network, boosts performance compared to separately trained mono-networks. DA, where both networks are alternately trained, achieves even better Dice scores in most cases. Based on a Mann-Whitney U-test with a significance level of 0.05 and a correction for multiple comparisons with a false discovery rate of 0.05, our models (DA/Semi-DA) result in significantly larger Dice scores than the mono-networks for all experiments. This demonstrates that segmentation and registration networks can indeed help each other by providing estimated supervision on unlabeled data.

Knee Results: On knee MRIs, our method improves segmentation scores over separately learned networks by about 1.2 and 0.5, and registration scores increase by about 3.1 and 3.0, when training with 5 and 10 manual segmentation respectively. Especially for the challenging cartilage structures, our joint learning boosts segmentation by 1.4 and 0.7, and registration by 5.5 and 5.2 for $N=5$ and $N=10$ respectively.

Brain Results: Dice scores for segmentation and registration increase by about 2.6 and 3.5 respectively for the cortical structures of the brain MRIs.

One-Shot Learning: In the one-shot experiments on both datasets, reasonable segmentation performance is achieved; moreover, DA increases the Dice score over unsupervised registration by about 2.7 and 1.8 on the knee and brain data respectively. This demonstrates the effectiveness of our framework for one-shot learning.

Qualitative Results: DA achieves more anatomically consistent registrations than the mono-networks on the knee (Fig. 2) and Brain MRI samples (see supplementary material).

4   Conclusion

We presented our DeepAtlas framework for joint learning of segmentation and registration networks using only few images with manual segmentations. By introducing

Table 2. Segmentation and registration performance on 3D brain MRIs. Average (Standard deviation) of Dice scores (%) for 31 cortical regions. N of 65 training images are manually labeled.

N	Models	Seg Dice (%)	Reg Dice (%)
0	Mono	–	54.75(2.37)
1	Semi-DA	61.19(1.49)	–
	DA	**61.22(1.40)**	56.54(2.32)
21	Mono	73.48(2.58)	59.47(2.34)
	Semi-DA	75.63(1.66)	62.92(2.14)
	DA	**76.06(1.50)**	**62.92(2.13)**
65	Mono	81.31(1.21)	63.25(2.07)

an anatomical similarity loss, the learned registrations are more anatomically consistent. Furthermore, the segmentation network is guided by a form of data augmentation provided via the registration network on unlabeled images. For both bone/cartilage structures in knee MRIs and cortical structures in brain MRIs, our approach shows large improvements over separately learned networks. When only given one manual segmentation, our method provides one-shot segmentation learning and greatly improves registration. This demonstrates that one network can benefit from imperfect supervision on unlabeled data provided by the other network. Our approach provides a general solution to the lack of manual segmentations when training segmentation and registration networks. For future work, introducing uncertainty measures for the segmentation and registration networks may help alleviate the effect of poor predictions of one network on the other. It would also be of interest to investigate multitask learning via layer sharing for the segmentation and registration networks. This may further improve performance and decrease model size.

References

1. Ambellan, F., Tack, A., Ehlke, M., Zachow, S.: Automated segmentation of knee bone and cartilage combining statistical shape knowledge and convolutional neural networks: data from the osteoarthritis initiative. MedIA **52**(2), 109–118 (2019). https://doi.org/10.1016/j.media.2018.11.009
2. Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V.: An unsupervised learning model for deformable medical image registration. In: CVPR, pp. 9252–9260 (2018)
3. Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V.: Voxelmorph: a learning framework for deformable medical image registration. IEEE TMI (2019)
4. Klein, A., Tourville, J.: 101 labeled brain images and a consistent human cortical labeling protocol. Front. Neurosci. **6**, 171 (2012). https://doi.org/10.3389/fnins.2012.00171
5. Litjens, G., et al.: A survey on deep learning in medical image analysis. MedIA **42**, 60–88 (2017)
6. Mahapatra, D., Sun, Y.: Joint registration and segmentation of dynamic cardiac perfusion images using MRFs. In: Jiang, T., Navab, N., Pluim, J.P.W., Viergever, M.A. (eds.) MICCAI 2010. LNCS, vol. 6361, pp. 493–501. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15705-9_60
7. Pohl, K.M., Fisher, J., Grimson, W.E.L., Kikinis, R., Wells, W.M.: A Bayesian model for joint segmentation and registration. NeuroImage **31**(1), 228–239 (2006)
8. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28

9. Rueckert, D., Sonoda, L.I., Hayes, C., Hill, D.L.G., Leach, M.O., Hawkes, D.J.: Nonrigid registration using free-form deformations: application to breast MR images. IEEE TMI **18**(8), 712–721 (1999). https://doi.org/10.1109/42.796284
10. Yang, X., Kwitt, R., Styner, M., Niethammer, M.: Quicksilver: fast predictive image registration - a deep learning approach. NeuroImage **158**, 378–396 (2017)
11. Yezzi, A., Zollei, L., Kapur, T.: A variational framework for joint segmentation and registration. In: MMBIA, pp. 44–51, December 2001. https://doi.org/10.1109/MMBIA.2001.991698