Stochastic Planner-Actor-Critic for Unsupervised Deformable Image Registration

Ziwei Luo[1*], Jing Hu[1*], Xin Wang[2†], Shu Hu[3],
Bin Kong[2], Youbing Yin[2], Qi Song[2], Xi Wu[1†], Siwei Lyu[3]
[1] Chengdu University of Information Technology, China
[2] Keya Medical, Seattle, USA
[3] University at Buffalo, SUNY, USA
algo_lzw@yahoo.com, jing_hu09@163.com, xi.wu@cuit.edu.cn, xinw@keyamedna.com

Abstract

Large deformations of organs, caused by diverse shapes and nonlinear shape changes, pose a significant challenge for medical image registration. Traditional registration methods need to iteratively optimize an objective function via a specific deformation model along with meticulous parameter tuning, but which have limited capabilities in registering images with large deformations. While deep learning-based methods can learn the complex mapping from input images to their respective deformation field, it is regression-based and is prone to be stuck at local minima, particularly when large deformations are involved. To this end, we present Stochastic Planner-Actor-Critic (SPAC), a novel reinforcement learning-based framework that performs step-wise registration. The key notion is warping a moving image successively by each time step to finally align to a fixed image. Considering that it is challenging to handle high dimensional continuous action and state spaces in the conventional reinforcement learning (RL) framework, we introduce a new concept 'Plan' to the standard Actor-Critic model, which is of low dimension and can facilitate the actor to generate a tractable high dimensional action. The entire framework is based on unsupervised training and operates in an end-to-end manner. We evaluate our method on several 2D and 3D medical image datasets, some of which contain large deformations. Our empirical results highlight that our work achieves consistent, significant gains and outperforms state-of-the-art methods.

Introduction

Deformable image registration (DIR) is an important task in medical imaging and has been actively studied for decades. DIR consists of establishing a spatial anatomical non-linear dense correspondence between a pair of fixed and moving images. The central task of DIR is the estimation of the ill-posed free-form transformation field that consists of a combination of global and local displacements (Eppenhof et al. 2019). Moreover, an accurate DIR on large deformation is needed due to soft organs (e.g., the brain, liver, and stomach) may undergo large deformations caused by patient re-positioning, surgical manipulation, or other physiological differences (Holden 2007). Large deformation diffeomorphic metric mapping (LDDMM) (Beg et al. 2005), derived from the group structure of the manifold of diffeomorphisms, is one of the most popular methods to tackle large deformations in DIR. However, achieving an optimal solution of the diffeomorphic image registration is computationally intensive and time-consuming, attempts at speeding up diffeomorphic image registration have thus been proposed to improve numerical approximation schemes (Wang and Zhang 2020).

Most existing DL-based solutions (Balakrishnan et al. 2019; Dalca et al. 2019; Mok and Chung 2020) are enforced to make a straightforward prediction, which is incapable to handle complicated deformations (Zhao et al. 2019). The step-wise image registration methods, such as R2N2 (Sandkühler et al. 2019) and RCN (Zhao et al. 2019), have shown the potential in DIR, in which the final deformation field (probably with large displacements) can be considered as a composition of the progressively predicted deformation field. But both of them are complex and computationally costly, and cannot deal with long step-wise registration. Inspired by the way that a human expert aligns two images by applying a sequence of local or global deformations, some RL-based image registration methods have been introduced in the past (Liao et al. 2017; Ma et al. 2017; Miao and Liao 2019; Sun et al. 2018; Hu et al. 2021). However, most of them merely focus on global rigid transformation since it only includes rotation and translation and can be easily represented by a low-dimensional discrete parametric model. Compared to rigid registration, DIR has huge, continuous state and action spaces, especially in 3D, which makes RL training extremely difficult. Krebs et al. (Krebs et al. 2017) proposed an RL-based approach for DIR, but their method uses supervised learning with ground truth deformation field, which is infeasible for most DIR tasks. And they incorporate the traditional statistical deformation model to reduce and discretize the action space, which leads to inferior performance in complex deformation registration.

In this paper, we propose a new RL architecture for unsupervised DIR problems, known as the *Stochastic Planner-Actor-Critic* (SPAC), to handle high dimensional continuous state and action spaces. As shown in Figure 1, the SPAC framework is formed with three core deep neural networks: the planner, the actor, and the critic. We introduce new a concept 'plan' which breaks the decision-making pro-

[*] Equal contribution.
[†] Corresponding authors.

cess into two steps, state → plan and plan → action. We call this process as meta policy, where the plan is a subspace of appropriate actions based on the current state, but it is not applied to the state directly, it is used to guide the actor to generate a tractable high-dimensional action that applies to the environment. The plan could be considered as an intermediate transition between state and action. As the input of the actor, the plan has a much lower dimension comparing with the state, which is easier for the actor to learn to predict actions. Meanwhile, the plan can be evaluated by the critic efficiently, since the Q function is easier to learn in the low-dimensional latent space. Furthermore, we employ an unsupervised registration learning strategy to learn the similarity of appearance between fixed and moving image pairs. The main contributions of our work can be summarized as follows:

- We describe a new RL framework, stochastic planner-actor-critic (SPAC), to handle large deformations by decomposing the monolithic learning process in DIR into small steps with high-dimensional continuous actions.
- To tackle the high-dimensional continuous action learning problem, we propose a stochastic meta policy that breaks the decision-making processing into two steps: state → low-dimensional plan and plan → deformation field action. The plan guides the actor to predict a tractable action, and the critic evaluates the plan, which makes the whole learning process feasible and computationally efficient.
- We design an registration environment which incorporates a K-means clustering module (Dice 1945) to obtain coarse segmentation maps to compute the Dice reward in an unsupervised manner, which obviates the need to collect real data with abundant and reliable ground-truth annotations. Besides, our method can be applied to entire 3D volumes.
- Experimental results on a variety of 2D/3D datasets show that the SPAC achieves state-of-the-art performance and consistently improves the results along with iterations.

Background

Deformable Image Registration

Deformable image registration (DIR) aims to learn a transformation between a fixed image and a moving image (Yang, Kwitt, and Niethammer 2016; Balakrishnan et al. 2018; de Vos et al. 2019). The DIR can be defined as an optimization problem. Given a pair of images (I_F, I_M), both of which on the image domain $\mathcal{X} \rightarrow \mathbb{R}^d$, where d is the dimension. I_F is the fixed image and I_M is the moving image. Denote Ω_w as a registration model parameterized by w. The output of it is a deformation filed, which can be warped on the moving image to align to the fixed image, denoted as $I_M \circ \Omega_w(I_F, I_M)$. We formulated the pairwise registration as a minimization problem based on energy function:

$$\min_w E(w) := G(I_F, I_M \circ \Omega_w(I_F, I_M)) + \lambda R(\Omega_w(I_F, I_M)) \tag{1}$$

where G represents a metric quantifying the similarity between the fixed image and the warped image, R represents a regularization constraining the deformation field, λ is a hyperparameter to balance these two terms. In the DL-based DIR task, the DL model tries to learn $\Omega_w(I_F, I_M)$ from a training dataset, which contains a large number of image pairs (I_F, I_M). The potential choice of G could be any similarity metric, such as the sum of squared differences (SSD), the normalized mutual information (NMI), or the negative normalized cross-correlation (NCC) (Haskins, Kruger, and Yan 2020; Balakrishnan et al. 2018).

Reinforcement Learning

RL is described by an infinite-horizon Markov decision process (MDP), defined by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{U}, r, \gamma)$. $\mathcal{S}$ is a set of states, $\mathcal{A}$ is action, and $\mathcal{U} : \mathcal{S} \times \mathcal{S} \times \mathcal{A} \rightarrow [0, \infty)$ represents the state transition probability density given state $s \in \mathcal{S}$ and action $\mathbf{a} \in \mathcal{A}$. $r : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ is the reward emitted from each transition, and $\gamma \in [0, 1]$ is the reward discount factor. Standard RL learns to maximize the expected sum of rewards from the episodic environments under the trajectory distribution ρ_π. It can be modified to incorporate an entropy term with the policy. Therefore, the resulting objective is defined as $\sum_{t=1}^{T} \mathbb{E}_{(\mathbf{s}_t, \mathbf{a}_t) \sim \rho_\pi} [r_t(\mathbf{s}_t, \mathbf{a}_t) + \alpha \mathcal{H}(\pi_\phi(\cdot|\mathbf{s}_t))]$, where α is a temperature parameter controlling the balance of the entropy $\mathcal{H}$ and the reward r_t.

Soft Actor-critic (SAC) (Haarnoja et al. 2018) is a promising framework for learning continuous actions, which is an off-policy actor-critic method that uses the above entropy-based framework to derive the soft policy iteration. Stochastic latent actor-critic (SLAC) improves the SAC by learning the representation spaces with a latent variable model which is more stable and efficient for complex continuous control tasks. It can improve both the exploration and robustness of the learned model. However, SLAC is far from enough for handling DIR, which has huge continuous action spaces such as voxel-wise estimation of a deformation field.

Stochastic Planner-Actor-Critic

Concretely, in our framework, the SPAC is formed with three core deep neural networks: the planner, the actor, and the critic with parameters ψ, ϕ, and θ, respectively (see Figure 1). The planner aims to generate a high-level plan in the low-dimensional latent space to guide the actor. In some sense, the plan can be considered as action clusters or action templates, which are high-level crude actions. Different from classic actor-critic models, the input of the actor is a stochastic plan instead of the state. That is, the generated plan is forwarded to the actor to further create the high dimensional action for DIR, and meanwhile, this plan is evaluated by the critic. We also add skip-connections from each down-sampling layer of the planner to the corresponding up-sampling layer of the actor. The information passed by skip-connections contains the details of the state $\mathbf{s}_t$ that are needed to reconstruct a natural-looking image. Using the proposed stochastic planner-actor-critic structure and supervision of similarity of appearance between fixed and moving image, SPAC could extend readily to complex DIR tasks.

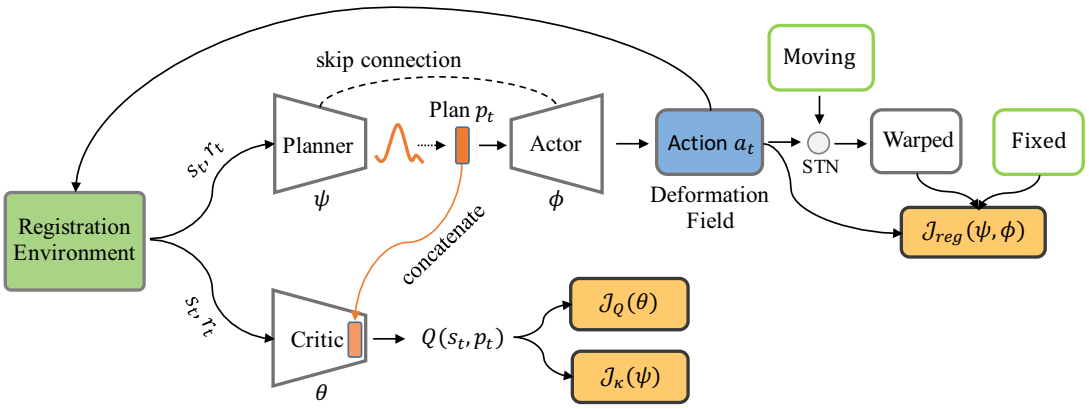

Figure 1: The network architecture of the proposed SPAC for DIR problem. At time step t, the registration environment receives action $\mathbf{a}_t$, and outputs state and reward (s_t, r_t). The plan $\mathbf{p}_t$ is sampled from the planner and evaluated by the critic. The action $\mathbf{a}_t$ is actually an immediate deformation field based on state s_t. The spatial transformer network (STN) is used as the warping function.

Problem Formulation

In forming SPAC, we apply the same notations as we defined for the conventional reinforcement learning in background section and introduce an additional component $\mathcal{P}$ for the continuous plan space to the infinite-horizon Markov decision process (MDP). Therefore, the MDP for SPAC can be defined by the tuple $(\mathcal{S}, \mathcal{P}, \mathcal{A}, \mathcal{U}, r, \gamma)$. $\mathcal{S}$ is a set of states, $\mathcal{P}$ is continuous plan, $\mathcal{A}$ is continuous action, and $\mathcal{U} : \mathcal{S} \times \mathcal{P} \times \mathcal{S} \times \mathcal{A} \to [0, \infty)$ represents the state transition probability density of the next state s_{t+1} given state $\mathbf{s}_t \in \mathcal{S}$, plan $p_t \in \mathcal{P}$ and action $\mathbf{a}_t \in \mathcal{A}$.

Step-wise Deformable Registration. Leveraging the sequential characteristic of reinforcement learning, we decompose the registration into T steps instead of predicting the deformation field in one-shot. At time step t, action $\mathbf{a}_t$ is the current deformation field generated by the Planner κ_ψ and Actor π_ϕ based on fixed image I_F and intermediate moving image I_{M_t}. Let $\Omega^t_{\psi,\phi}$ represents the accumulated deformation field composed by $\mathbf{a}_t$ and the previous deformation field $\Omega^{t-1}_{\psi,\phi}$. We can compute $\Omega^t_{\psi,\phi}$ with a recursive composition function:

$$\Omega^t_{\psi,\phi} = \begin{cases} 0 & \text{if } t = 0, \\ \mathcal{C}(\mathbf{a_t}, \Omega^{t-1}_{\psi,\phi}) & \text{otherwise,} \end{cases} \tag{2}$$

where

$$\mathcal{C}(\mathbf{a_t}, \Omega^{t-1}_{\psi,\phi}) = \Omega^{t-1}_{\psi,\phi} + (\mathbf{a}_t \circ \Omega^{t-1}_{\psi,\phi}). \tag{3}$$

To eliminate the warping bias in the multi-step recursive registration process (Zhao et al. 2019), we warp the initial moving image I_M using accumulated the deformation field $\Omega^t_{\psi,\phi}$. Then the warped image $I_{M_{t+1}}$ is used as the next moving image in time step $t+1$. Therefore, the registration result can be progressively improved by predicting deformation from coarse to the local refined. Using the notion of step-wise, the DIR optimization problem (Eq.(1)) in our SPAC framework can be rewritten as

$$\min_{\psi,\phi} E(\psi, \phi) := \frac{1}{T} \sum_{t=1}^{T} G(I_F, I_{M_t} \circ \Omega^t_{\psi,\phi}) + \lambda R(\Omega^t_{\psi,\phi}), \tag{4}$$

where we use a tuple (ψ, ϕ) instead of the parameter w in Eq.(1) since the deformation field will be learned from the SPAC framework. In the following, we provide the details of the environment and policy in our method.

DIR Environment

The overview of the step-wise deformable registration environment is shown in Figure 2. In the beginning, the environment only contains an image pair (I_F, I_M), then we performs K-means (MacQueen et al. 1967) with three clustering labels to obtain the corresponding segmentation maps (U_F, U_M) in an unsupervised manner. The generated segmentation map assigns each voxel to a virtual anatomical structure, which facilitates computing rewards. At time step t, the state $\mathbf{s}_t$ is the pair of fixed image I_F and moving image I_{M_t}, $\mathbf{s}_t = (I_F, I_{M_t})$. The next state s_{t+1} is obtained by warping I_M with composed deformable field $\Omega^t_{\psi,\phi}$: $s_{t+1} = (I_F, I_M \circ \Omega^t_{\psi,\phi})$. We incorporate the widely used spatial transformer network (STN) (Jaderberg et al. 2015) as the warping operator. The reward r_t is defined based on the Dice (Dice 1945) score:

$$r_t = \text{Dice}(U_F, U_M \circ \Omega^t_{\psi,\phi}) - \text{Dice}(U_F, U_M \circ \Omega^{t-1}_{\psi,\phi}), \tag{5}$$

where $\text{Dice}(U_1, U_2) = 2 \cdot \frac{|U_1 \cap U_2|}{|U_1| + |U_2|}$. This reward function explicitly assesses the improvement of the predicted deformation field $\Omega^t_{\psi,\phi}$.

Stochastic Meta Policy

In our formulation, we have a meta policy (κ, π), where the stochastic plan is modeled as a subspace of the deforma-

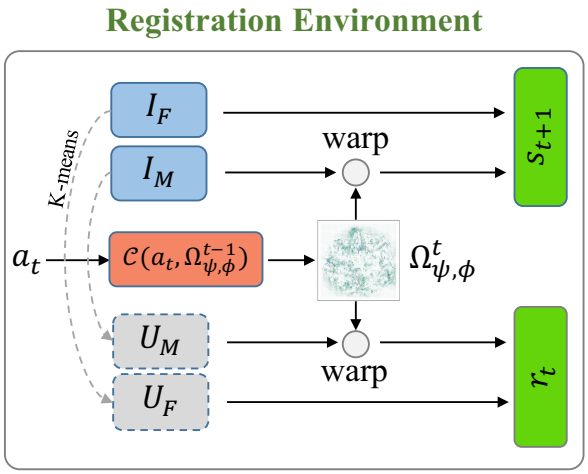

Figure 2: The architecture of the step-wise deformable registration environment. When the environment receives an action $\mathbf{a}_t$, it outputs the next state s_{t+1} and reward r_t. Specifically, the environment is composed by a pair of image (I_F, I_M), and generates the corresponding segmentation maps (U_F, U_M) with K-means clustering. $\mathcal{C}(a_t, \Omega^{t-1}_{\psi,\phi})$ is the compose function which composes action a_t to the accumulated deformation field $\Omega^t_{\psi,\phi}$. The next state s_{t+1} is obtained by concatenating I_F and the warped moving image $I_M \circ \Omega^t_{\psi,\phi}$. The reward r_t is obtained by Eq. (5) which represents the improvement of Dice score.

tion field that gives low-dimensional vector $\mathbf{p}_t$ given state $\mathbf{s}_t$. While the actor's action is actually a deterministic deformation field $\mathbf{a}_t$ determined by the plan $\mathbf{p}_t$. Consider a parameterized planner κ_ψ and actor π_ϕ, the stochastic plan is sampled as a representation: $\mathbf{p}_t \sim \kappa_\psi(\mathbf{p}_t|\mathbf{s}_t)$, and the action is generated by decoding the plan vector $\mathbf{p}_t$ to a high-dimensional deformation field: $\mathbf{a}_t = \pi_\phi(\mathbf{a}_t|\mathbf{p}_t)$ In practice, we reparameterize the the planner and the stochastic plan jointly using a neural network approximation $\mathbf{p}_t = f_\psi(\epsilon_t, \mathbf{s}_t)$, known as reparameterization trick (Kingma and Welling 2013), where ϵ_t is an input noise vector sampled from a fixed Gaussian distribution. Moreover, we maximize the entropy of plan to improve exploration and robustness. The augmented objective function is formulated as:

$$\max_{\psi,\phi} \sum_{t=1}^{T} \mathbb{E}_{(\mathbf{s}_t,\mathbf{p}_t,\mathbf{a}_t)\sim\rho_{(\kappa,\pi)}} \left[r_t(\mathbf{s}_t, \mathbf{p}_t, \mathbf{a}_t) + \alpha \mathcal{H}(\kappa_\psi(\cdot|\mathbf{s}_t)) \right].$$

where α is the temperature and $\rho_{(\kappa,\pi)}$ is a trajectory distribution under $\kappa_\psi(\mathbf{p}_t|\mathbf{s}_t)$ and $\pi_\phi(\mathbf{a}_t|\mathbf{p}_t)$.

Learning Planner and Critic

Different from conventional RL algorithms, the critic Q_θ evaluates plan $\mathbf{P}_t$ instead of action $\mathbf{a}_t$. since learning a low-dimensional plan in the DIR problem is easier and more effective. Specifically, the low-dimensional plan is concatenated to the downsampled vector of the critic and outputs soft Q function $Q_\theta(\mathbf{s}_t, \mathbf{p}_t)$ which is an estimation of the current state plan value, as shown in Figure 1.

When the critic is used to evaluate the planner, the rewards and the soft Q values are used to iteratively guide the stochastic policy improvement. In the evaluation step, following SAC (Haarnoja et al. 2018), SPAC learns a policy κ_ψ (planner) and fits the parametric Q-function $Q_\theta(\mathbf{s}_t, \mathbf{p}_t)$ (critic) using transitions sampled from the replay pool $\mathcal{D}$ by minimizing the soft Bellman residual:

$$J_Q(\theta) = \mathbb{E}_{(\mathbf{s}_t,\mathbf{p}_t)\sim\mathcal{D}} \left[\frac{1}{2} \Big(Q_\theta(\mathbf{s}_t, \mathbf{p}_t) - \big(r_t + \gamma \mathbb{E}_{\mathbf{s}_{t+1}} \left[V_{\bar{\theta}}(\mathbf{s}_{t+1}) \right] \big) \Big)^2 \right],$$

where $V_{\bar{\theta}}(\mathbf{s}_t) = \mathbb{E}_{\mathbf{p}_t\sim\kappa_\psi}[Q_{\bar{\theta}}(\mathbf{s}_t, \mathbf{p}_t) - \alpha \log \kappa_\psi(\mathbf{p}_t|\mathbf{s}_t)]$. We use a target network $Q_{\bar{\theta}}$ to stabilize training, whose parameters $\bar{\theta}$ are obtained by an exponentially moving average of parameters of the critic network (Lillicrap et al. 2015): $\bar{\theta} \to \tau\theta + (1-\tau)\bar{\theta}$. The hyper-parameter $\tau \in [0, 1]$. To optimize the $J_Q(\theta)$, we can do the stochastic gradient descent with respect to the parameters θ as follows,

$$\begin{aligned} \theta = \theta - \eta_Q \nabla_\theta Q_\theta(\mathbf{s}_t, \mathbf{p}_t) \Big(& Q_\theta(\mathbf{s}_t, \mathbf{p}_t) - r_t \\ & - \gamma \left[Q_{\bar{\theta}}(\mathbf{s}_{t+1}, \mathbf{p}_{t+1}) - \alpha \log \kappa_\psi(\mathbf{p}_{t+1}|\mathbf{s}_{t+1}) \right] \Big). \end{aligned} \tag{6}$$

Since the critic works on the planner, the optimization procedure will also influence the planner decisions. Following (Haarnoja et al. 2018), we can use the following objective to minimize the KL divergence between the policy and a Boltzmann distribution induced by the Q-function,

$$\begin{aligned} J_\kappa(\psi) =& \mathbb{E}_{\mathbf{s}_t\sim\mathcal{D}} \left[\mathbb{E}_{\mathbf{p}_t\sim\kappa_\psi} \left[\alpha \log(\kappa_\psi(\mathbf{p}_t|\mathbf{s}_t)) - Q_\theta(\mathbf{s}_t, \mathbf{p}_t) \right] \right] \\ =& \mathbb{E}_{\mathbf{s}_t\sim\mathcal{D}, \epsilon_t\sim\mathcal{N}(\mu,\sigma)} \big[\alpha \log(\kappa_\psi(f_\psi(\epsilon_t, \mathbf{s}_t)|\mathbf{s}_t)) \\ & - Q_\theta(\mathbf{s}_t, f_\psi(\epsilon_t, \mathbf{s}_t)) \big]. \end{aligned}$$

The last equation holds because $\mathbf{p}_t$ can be evaluated by $f_\psi(\epsilon_t, \mathbf{s}_t)$ as we discussed before. It should be mentioned that the hyperparameter α can be automatically adjusted by using one proposed method from (Haarnoja et al. 2018). Then we can apply the stochastic gradient method to optimize parameters as follows,

$$\begin{aligned} \psi =& \psi - \eta_\psi \Big(\nabla_\psi \alpha \log(\kappa_\psi(p_t|s_t)) + \\ & \big(\nabla_{p_t} \alpha \log(\kappa_\psi(p_t|s_t)) - \nabla_{p_t} Q_\theta(s_t, p_t) \big) \nabla_\psi f_\psi(\epsilon_t, s_t) \Big). \end{aligned} \tag{7}$$

Learning Planner and Actor with Unsupervised Registration

After getting action $\mathbf{a}_t$ and following meta policy (κ_ψ, π_ϕ), we can obtain $\Omega^t_{\psi,\phi}$ based on Eq.(2). We learn the similarity of appearance between the fixed image and the warped image using local normalized cross-correlation (NCC) (Balakrishnan et al. 2018): $G(I_F, I_{M_t}) = NCC(I_F, I_M \circ \Omega^t_{\psi,\phi})$.

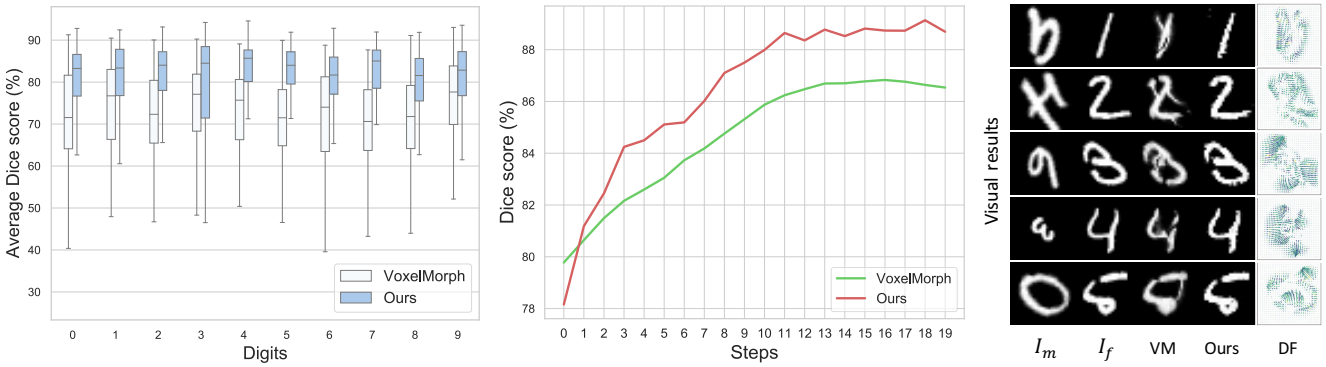

Figure 3: **Left:** The plot box of Dice scores over 9 fixed digits. **Center:** Step-wise comparison of our method and VoxelMorph (VM) (Balakrishnan et al. 2018). **Right:** Visual comparison of our method with VM. The scaled and rotated digits are transformed to other fixed digits. The Deformation Filed (DF) column shows the visualized deformable fields of our method.

Algorithm 1: Stochastic Planner-Actor-Critic

Input: I_F, I_M, U_F, U_M, replay pool $\mathcal{D}$
Init: ψ, ϕ, θ, $\bar{\theta}$, $\mathcal{D}$ and environment $\mathcal{E}$
for *each iteration* **do**
  for *each environment step* **do**
    $\mathbf{p}_t \sim \kappa_\psi(\mathbf{p}_t|\mathbf{s}_t)$, $\mathbf{a}_t \sim \pi_\phi(\mathbf{a}_t|\mathbf{p}_t)$
    $\mathbf{s}_{t+1}, r_t \sim \mathcal{U}(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{p}_t, \mathbf{a}_t)$
    $\mathcal{D} = \mathcal{D} \cup \{(\mathbf{s}_t, \mathbf{p}_t, \mathbf{a}_t, r_t, s_{t+1})\}$
  end
  for *each gradient step* **do**
    Sample from $\mathcal{D}$
    Update θ, ψ, ϕ with Eq.(6), Eq.(7), Eq. (8)
  end
end

A higher value of the NCC indicates a better alignment. In order to generate realistic warped images, we use a total variation regularizer (Rudin, Osher, and Fatemi 1992) to smooth the deformation field on its spatial gradients: $R(\Omega^t_{\psi,\phi}) = \|\nabla\Omega^t_{\psi,\phi}\|_2^2$. The final registration loss J_{reg} is defined as

$$J_{reg}(\psi,\phi) = \mathbb{E}_{s_t\sim\mathcal{D}}[-NCC(I_F, I_M \circ \Omega^t_{\psi,\phi}) + \lambda\|\nabla\Omega^t_{\psi,\phi}(\mathbf{s}_t)\|_2^2].$$

We can update ψ and ϕ from the planner and the actor by performing the following steps:

$$\psi = \psi - \eta\nabla_\psi J_{reg}(\psi,\phi), \quad \phi = \phi - \eta\nabla_\phi J_{reg}(\psi,\phi) \quad (8)$$

The pseudo-code of optimizing SPAC is described in Algorithm 1. All parameters of SPAC are optimized base on the samples from replay pool $\mathcal{D}$.

Experiments

Experimental Settings

Datasets. We evaluate our method on three types of images: MNIST digits, 2D brain MRI scans, and 3D liver CT scans. MNIST (LeCun et al. 1998) is regarded as a standard sanity check for the proposed registration method. The goal is to transform between two different 28×28 images of handwritten digits. In testing, we fixed ten digits from 0 to 9 as the atlases, and select 1000 randomly scaled and rotated digits as moving images to be aligned with the atlas.

The 2D brain MRI training dataset consists of 2302 pre-processed 2D scans from ADNI (Mueller et al. 2005), ABIDE (Di Martino et al. 2014) and ADHD (Bellec et al. 2017). The evaluation dataset uses 40 pre-processed slices from LONI Probabilistic Brain Atlas (LPBA) (Shattuck et al. 2008), each of which contains a segmentation ground truth of 56 manually delineated anatomical structures. All images are resampled to 128×128 pixels. The first slice of LPBA is served as the atlas, and all the remaining images are used as the moving image. For 3D registration, we use Liver Tumor Segmentation (LiTS) (Bilic et al. 2019) challenge data for training, which contains 131 CT scans with the segmentation ground truth manually annotated by experts. The SLIVER (Heimann et al. 2009) dataset has 20 scans with liver segmentation ground truth. We divide them into 10 pairs as the regular testing data. We also evaluated our method on the challenging Liver Segmentation of Pigs (LSPIG) (Zhao et al. 2019) dataset, which contains 17 paired CT scans from pigs, along with liver segmentation ground truth. All 3D volumes are resampled to $128 \times 128 \times 128$ pixels and pre-affined as standard pre-processing steps.

Baselines. This work focus on unsupervised deformable registration. We compare our method with several deep learning based image registration (DLIR) methods: VoxelMorph (VM) (Balakrishnan et al. 2019), VM-diff (Dalca et al. 2019), SYMNet (Mok and Chung 2020), R2N2 (Sandkühler et al. 2019) and RCN (Zhao et al. 2019). VM employs a U-Net structure with NCC loss to learn deformable registration, and VM-diff is a probabilistic diffeomorphic variant of VM. SYMNet is a state-of-the-art single-pass 3D registration method. R2N2 and RCN are sequence-based methods for 2D and 3D registration, respectively. To illustrate the effectiveness of the proposed method, we use the same network structure as VM. We also compare with two top-performing conventional registration algorithms, SyN (Avants et al. 2008) and Elastix (Klein et al.

Method	2D Registration LPBA	Time(s)	#Params	3D Registration SLIVER	LSPIG	Time(s)	#Params
SyN (Avants et al. 2008)	55.47±3.96	4.57	-	89.57±3.34	81.83±8.30	269	-
Elastix (Klein et al. 2009)	53.64±3.97	2.20	-	90.23±2.39	81.19±7.47	87.0	-
LDDMM (Beg et al. 2005)	52.18±3.48	3.27	-	83.94±3.44	82.33±7.14	41.4	-
VM (Balakrishnan et al. 2019)	55.36±3.94	0.02	105K	86.37±4.15	81.13±7.28	0.13	356K
VM-diff (Dalca et al. 2019)	55.88±3.78	0.02	118K	87.24±3.26	81.38±7.21	0.16	396K
R2N2 (Sandkühler et al. 2019)	51.84±3.30	0.46	3,591K	-	-	-	-
RCN (Zhao et al. 2019)	-	-	-	89.59±3.18	82.87±5.69	2.44	21,291K
SYMNet (Mok and Chung 2020)	-	-	-	86.97±3.82	82.78±7.20	0.18	1,124K
SPAC (t=20, SSIM reward)	56.43±3.76	0.16	107K	90.27±3.85	83.69±6.74	1.05	458K
SPAC (t=1, Dice reward)	55.21±3.55	0.02	107K	84.81±4.42	80.61±7.94	0.07	458K
SPAC (t=10, Dice reward)	56.12±3.68	0.08	107K	90.01±3.79	**84.67±6.05**	0.55	458K
SPAC (t=20, Dice reward)	**56.57±3.71**	0.16	107K	**90.28±3.66**	84.40±6.24	1.05	458K

Table 1: Dice score (%) results of our SPAC (t indicates the t-th step) with other methods over all datasets. The running times of 3D registration is test on SLIVER dataset. Note that R2N2 works only for 2D registration. The official RCN and SYMNet are implemented only for 3D registration.

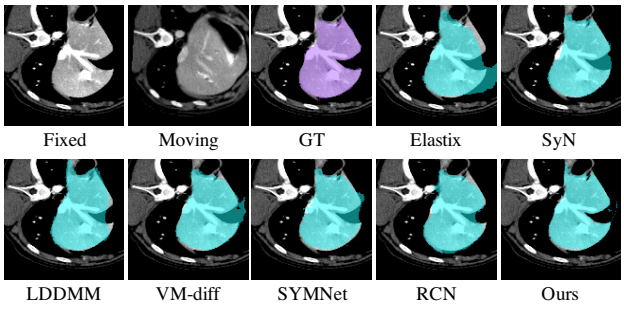

Figure 4: Visual results of our SPAC with other methods on 3D liver dataset. The warped moving image obtained by our SPAC is closer to the ground truth.

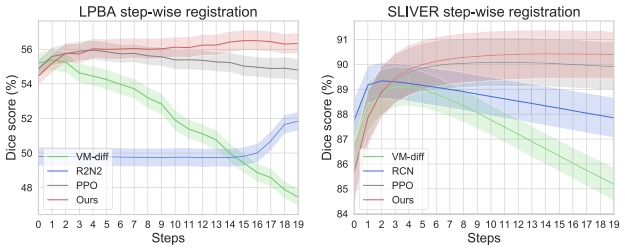

Figure 5: Step-wise registration results overall datasets. The RL-based methods (Our SPAC and PPO) perform more stable than other DL-based methods, and our SPAC achieved the best performance.

2009) with B-Spline (Rueckert et al. 1999). In addition, we provide the result of using SSIM as the reward function. We use Dice score as the evaluation metric.

Experimental Results

MNIST Digits Transform. Figure 3 shows the representative results and average Dice scores on MNIST digits transforms. In testing, the moving images are randomly scaled and rotated, which results in a larger and more challenging deformation field. Our method outperforms VM over all kinds of digits with significant improvements quantitatively and qualitatively, which also demonstrates that the proposed method has better generalizability and can work well on image pairs with large deformations.

Medical Image Registration. Table 1 summarizes the performance of our method with other state-of-the-art methods. The SPAC outperforms other methods over all 2D and 3D datasets. Moreover, the LSPIG dataset has large deformation fields and is quite different from the training dataset (LiTS) in structure and appearance. The quantitative results on LSPIG illustrate that our framework works well in large deformation dataset and has better generalizability than other conventional DL-based methods. Note that our method performs registration in a step-wise manner, which results in a slower speed than most one-step methods such as VM and SYMNet. But SPAC is still faster than other multi-step methods such as R2N2 and RCN. The SSIM reward based SPAC achieve good result on SLIVER but performs worse on LSPIG dataset compared with Dice reward based SPAC . We visualized an example of registration results on the LSPIG dataset by overlaying the warped moving segmentation map on the fixed image in Figure 4. The result shows that our model successfully learns registration even with a challenging, large deformation. The SPAC exceeds the performance of another step-wise method RCN, which demonstrates the effectiveness of our framework.

Analysis

Step-wise registration. The key idea of our method is to decompose the monolithic registration process into small steps by a lighter-weight CNN and progressively improves the transformed results. Figure 6 shows an example of step-wise registration process. The deformation fields visualized on the upper row illustrate that our method predicts transformation from coarse to the local refines step by step. We compared our method with PPO (Schulman et al. 2017) and other DL-based methods using step-wise registration in Figure 5. As the registration step increases, the performance of DL-based methods gets worse, while the RL-based methods

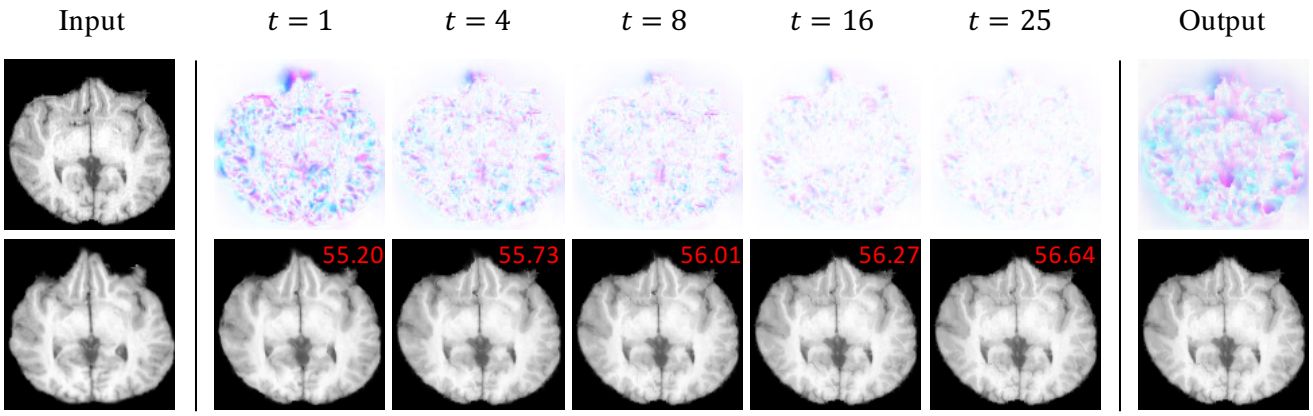

Figure 6: A step-wise registration example of our method on the LPBA dataset. Top row is the visualized displacement field, where deep color represents a large deformation. The Dice score (red) keeps increasing along the steps.

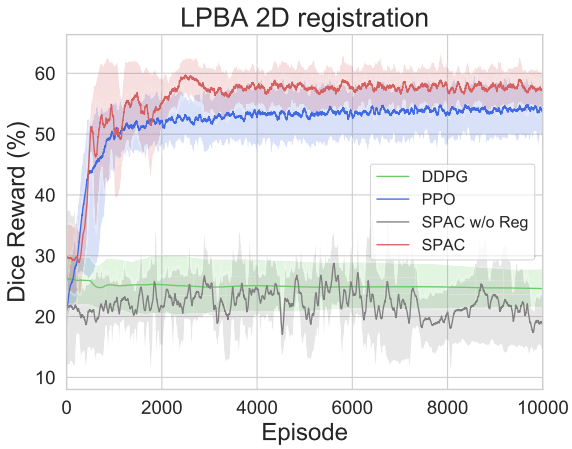

Figure 7: Learning curves of several RL-based methods on LPBA dataset.

are more stable. The Dice score of SPAC is increasing all the time on LPBA and SLIVER datasets.

Compare with other RL methods. To demonstrate our method in the reinforcement learning side. We modify our framework with other popular RL algorithms such as PPO (Schulman et al. 2017) and DDPG (Lillicrap et al. 2015) in Planner-Critic learning process. Moreover, we compare with the method discarding DL-based unsupervised registration loss (SPAC w/o Reg). The qualitative result of PPO-modified is shown in Table 2. The training curves of each method are shown in Figure 7. Our SPAC achieves batter performance than PPO. And the DDPG which uses deterministic policy is failed to convergence. The results also indicate that the RL agent can hardly deal with the DIR problem without unsupervised registration loss.

Ablation Experiments. We study the effect of some important settings in our framework, such as reinforcement learning, unsupervised registration learning, and evaluating

	LPBA	SLIVER	LSPIG
PPO-modified	55.82±3.49	89.30±3.63	83.55±6.24
SPAC-action	55.58±3.70	88.75±3.69	81.80±7.51
SPAC w/o RL	54.89±3.80	85.43±4.14	80.72±7.34
SPAC w/o Reg	44.67±3.74	79.34±4.02	72.45±6.25
SPAC	**56.57±3.71**	**90.28±3.66**	**84.40±6.24**

Table 2: Dice score (%) over several variants of our methods. 'SPAC-action' indicates that the critic evaluates the actor's action instead of planner in SPAC.

plan with the critic. Note that in the settings without using registration loss, we evaluate the deformation field as the only action, and both the planner and actor are trained with the RL objective. As summarized in Table 2, the result is unsatisfactory if we train SPAC without reinforcement learning, and it becomes worse if the training discards unsupervised registration loss. Critic evaluates actor's action (SPAC-action) results in an inferior performance compared with the SPAC (critic evaluates planner).

Conclusion

In this paper, a step-wise registration network based on reinforcement learning is proposed to handle large deformation problem which is especially observed in soft tissues. This method, SPAC, an off-policy actor-critic model, can efficiently learn good policies in spaces with high-dimensional continuous actions and states. Central to SPAC is the proposed component 'plan' which is defined in latent subspace and can guide the actor to generate high-dimensional actions. To the best of our knowledge, we are the first to propose a pure RL model to deformable medical image registration. Experiments based on diverse medical image datasets demonstrate that this architecture achieves significant gains over state-of-the-art methods, especially for the case with large deformations. With the superiority of good performance, we expect that the proposed architecture can potentially be extended to all deformable image registration tasks.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 61602065, Sichuan province Key Technology Research and Development project under Grant 2021YFG0038.

References

Avants, B. B.; Epstein, C. L.; Grossman, M.; and Gee, J. C. 2008. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Medical image analysis*, 12(1): 26–41.

Balakrishnan, G.; Zhao, A.; Sabuncu, M. R.; Guttag, J.; and Dalca, A. V. 2018. An unsupervised learning model for deformable medical image registration. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 9252–9260.

Balakrishnan, G.; Zhao, A.; Sabuncu, M. R.; Guttag, J.; and Dalca, A. V. 2019. VoxelMorph: a learning framework for deformable medical image registration. *IEEE transactions on medical imaging*, 38(8): 1788–1800.

Beg, M. F.; Miller, M. I.; Trouvé, A.; and Younes, L. 2005. Computing large deformation metric mappings via geodesic flows of diffeomorphisms. *International journal of computer vision*, 61(2): 139–157.

Bellec, P.; Chu, C.; Chouinard-Decorte, F.; Benhajali, Y.; Margulies, D. S.; and Craddock, R. C. 2017. The neuro bureau ADHD-200 preprocessed repository. *Neuroimage*, 144: 275–286.

Bilic, P.; Christ, P. F.; Vorontsov, E.; Chlebus, G.; Chen, H.; Dou, Q.; Fu, C.-W.; Han, X.; Heng, P.-A.; Hesser, J.; et al. 2019. The liver tumor segmentation benchmark (lits). *arXiv preprint arXiv:1901.04056*.

Dalca, A. V.; Balakrishnan, G.; Guttag, J.; and Sabuncu, M. R. 2019. Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces. *Medical image analysis*, 57: 226–236.

de Vos, B. D.; Berendsen, F. F.; Viergever, M. A.; Sokooti, H.; Staring, M.; and Išgum, I. 2019. A deep learning framework for unsupervised affine and deformable image registration. *Medical image analysis*, 52: 128–143.

Di Martino, A.; Yan, C.-G.; Li, Q.; Denio, E.; Castellanos, F. X.; Alaerts, K.; Anderson, J. S.; Assaf, M.; Bookheimer, S. Y.; Dapretto, M.; et al. 2014. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular psychiatry*, 19(6): 659–667.

Dice, L. R. 1945. Measures of the amount of ecologic association between species. *Ecology*, 26(3): 297–302.

Eppenhof, K. A.; Lafarge, M. W.; Veta, M.; and Pluim, J. P. 2019. Progressively trained convolutional neural networks for deformable image registration. *IEEE transactions on medical imaging*, 39(5): 1594–1604.

Haarnoja, T.; Zhou, A.; Hartikainen, K.; Tucker, G.; Ha, S.; Tan, J.; Kumar, V.; Zhu, H.; Gupta, A.; Abbeel, P.; et al. 2018. Soft actor-critic algorithms and applications. *arXiv preprint arXiv:1812.05905*.

Haskins, G.; Kruger, U.; and Yan, P. 2020. Deep learning in medical image registration: a survey. *Machine Vision and Applications*, 31(1): 1–18.

Heimann, T.; Van Ginneken, B.; Styner, M. A.; Arzhaeva, Y.; Aurich, V.; Bauer, C.; Beck, A.; Becker, C.; Beichel, R.; Bekes, G.; et al. 2009. Comparison and evaluation of methods for liver segmentation from CT datasets. *IEEE transactions on medical imaging*, 28(8): 1251–1265.

Holden, M. 2007. A review of geometric transformations for nonrigid body registration. *IEEE transactions on medical imaging*, 27(1): 111–128.

Hu, J.; Luo, Z.; Wang, X.; Sun, S.; Yin, Y.; Cao, K.; Song, Q.; Lyu, S.; and Wu, X. 2021. End-to-end multimodal image registration via reinforcement learning. *Medical Image Analysis*, 68: 101878.

Jaderberg, M.; Simonyan, K.; Zisserman, A.; and Kavukcuoglu, K. 2015. Spatial transformer networks. *arXiv preprint arXiv:1506.02025*.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Klein, S.; Staring, M.; Murphy, K.; Viergever, M. A.; and Pluim, J. P. 2009. Elastix: a toolbox for intensity-based medical image registration. *IEEE transactions on medical imaging*, 29(1): 196–205.

Krebs, J.; Mansi, T.; Delingette, H.; Zhang, L.; Ghesu, F. C.; Miao, S.; Maier, A. K.; Ayache, N.; Liao, R.; and Kamen, A. 2017. Robust non-rigid registration through agent-based action learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 344–352. Springer.

LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11): 2278–2324.

Liao, R.; Miao, S.; de Tournemire, P.; Grbic, S.; Kamen, A.; Mansi, T.; and Comaniciu, D. 2017. An artificial agent for robust image registration. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, 4168–4175.

Lillicrap, T. P.; Hunt, J. J.; Pritzel, A.; Heess, N.; Erez, T.; Tassa, Y.; Silver, D.; and Wierstra, D. 2015. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*.

Ma, K.; Wang, J.; Singh, V.; Tamersoy, B.; Chang, Y.-J.; Wimmer, A.; and Chen, T. 2017. Multimodal image registration with deep context reinforcement learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 240–248. Springer.

MacQueen, J.; et al. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, 281–297. Oakland, CA, USA.

Miao, S.; and Liao, R. 2019. Agent-based methods for medical image registration. In *Deep Learning and Convolutional Neural Networks for Medical Imaging and Clinical Informatics*, 323–345. Springer.

Mok, T. C.; and Chung, A. 2020. Fast symmetric diffeomorphic image registration with convolutional neural networks.

In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4644–4653.

Mueller, S. G.; Weiner, M. W.; Thal, L. J.; Petersen, R. C.; Jack, C. R.; Jagust, W.; Trojanowski, J. Q.; Toga, A. W.; and Beckett, L. 2005. Ways toward an early diagnosis in Alzheimer's disease: the Alzheimer's Disease Neuroimaging Initiative (ADNI). *Alzheimer's & Dementia*, 1(1): 55–66.

Rudin, L. I.; Osher, S.; and Fatemi, E. 1992. Nonlinear total variation based noise removal algorithms. *Physica D: nonlinear phenomena*, 60(1-4): 259–268.

Rueckert, D.; Sonoda, L. I.; Hayes, C.; Hill, D. L.; Leach, M. O.; and Hawkes, D. J. 1999. Nonrigid registration using free-form deformations: application to breast MR images. *IEEE transactions on medical imaging*, 18(8): 712–721.

Sandkühler, R.; Andermatt, S.; Bauman, G.; Nyilas, S.; Jud, C.; and Cattin, P. C. 2019. Recurrent Registration Neural Networks for Deformable Image Registration. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Shattuck, D. W.; Mirza, M.; Adisetiyo, V.; Hojatkashani, C.; Salamon, G.; Narr, K. L.; Poldrack, R. A.; Bilder, R. M.; and Toga, A. W. 2008. Construction of a 3D probabilistic atlas of human cortical structures. *Neuroimage*, 39(3): 1064–1080.

Sun, S.; Hu, J.; Yao, M.; Hu, J.; Yang, X.; Song, Q.; and Wu, X. 2018. Robust multimodal image registration using deep recurrent reinforcement learning. In *Asian Conference on Computer Vision*, 511–526. Springer.

Wang, J.; and Zhang, M. 2020. Deepflash: An efficient network for learning-based medical image registration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4444–4452.

Yang, X.; Kwitt, R.; and Niethammer, M. 2016. Fast predictive image registration. In *Deep Learning and Data Labeling for Medical Applications*, 48–57. Springer.

Zhao, S.; Dong, Y.; Chang, E. I.; Xu, Y.; et al. 2019. Recursive cascaded networks for unsupervised medical image registration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10600–10610.