Partial Policy-based Reinforcement Learning for Anatomical Landmark Localization in 3D Medical Images

Walid Abdullah Al, Il Dong Yun* (*IEEE Member*)
Department of Computer and Electronic Systems Engineering,
Hankuk University of Foreign Studies, Yongin, South Korea

Utilizing the idea of long-term cumulative return, reinforcement learning (RL) has shown remarkable performance in various fields. We propose a formulation of the landmark localization in 3D medical images as a reinforcement learning problem. Whereas value-based methods have been widely used to solve RL-based localization problems, we adopt an actor-critic based direct policy search method framed in a temporal difference learning approach. In RL problems with large state and/or action spaces, learning the optimal behavior is challenging and requires many trials. To improve the learning, we introduce a partial policy-based reinforcement learning to enable solving the large problem of localization by learning the optimal policy on smaller partial domains. Independent actors efficiently learn the corresponding partial policies, each utilizing their own independent critic. The proposed policy reconstruction from the partial policies ensures a robust and efficient localization utilizing the sub-agents solving simple binary decision problems in their corresponding partial action spaces. Experiments with three different localization problems in 3D CT and MR images showed that the proposed reinforcement learning requires a significantly smaller number of trials to learn the optimal behavior compared to the original behavior learning scheme in RL. It also ensures a satisfactory performance when trained on a fewer images.

***Index Terms*—Actor-critic, landmark localization, medical image, partial policy, reinforcement learning**

I. Introduction

LANDMARK localization plays a vital role in medical image analysis, facilitating the automatic process for registration, classification, and segmentation [1], [2]. Besides speeding up the interpretation, it contributes to visualization and assessment-based applications. However, accurate landmark localization in 3D medical images is a challenging problem because of high inter-patient variations in terms of size, shape, and orientation, as well as the variations and artifacts caused by different parameter settings. Machine learning approaches are becoming more and more common to solve the localization problem under such variation. Standard approaches suggest classification or regression-based model in order to localize the landmarks. However, all of the previous learning approaches are mainly exploitative and may behave inconsistently for an exceptional test data. Long-term reward-oriented reinforcement learning (RL) algorithms offer ways to balance between exploration and exploitation, yielding a noteworthy performance in various fields of image processing [3], [4], [5]. With a few instances of implementation in object localization in terms of a bounding box, RL-enforced landmark localization is rarely found.

Value function approximation (e.g., deep Q-Network (DQN) [3]) is a widely used method to solve the RL problem for large state and/or action spaces, suggesting an indirect behavior learning. Compared to such value-based methods, explicit behavior learning by directly approximating the policy function [6] has the advantage of a better convergence. However, direct policy search method suffers from the high variance problem [7]. Actor-critic RL performs direct policy approximation while utilizing an additional value function approximator to reduce the variance, thus taking advantage of both the policy and value-based methods [8]. Nevertheless, a good exploration in order to obtain the optimal policy in a large space is challenging. Despite remarkable propositions and improvement to maintain the balance between exploration and exploitation, RL practically faces problem to successfully learn a task in a large state and/or action space and requires many trials [9].

In this paper, we formulate the landmark localization problem as a sequential decision-making problem in RL, where an agent initiated at a random position inside a 3D medical image (i.e., volume) observes the current state and takes subsequent actions to move towards the target landmark. We suggest learning the policy function directly using an actor-critic approach because of its advantage over pure policy or value-based approach. To ensure a successful behavior learning within a significantly smaller number of trials, we introduce a partial policy-based reinforcement learning model where multiple sub-agents learn assigned micro-tasks to successfully learn the original task. Partial policies with respect to the micro-tasks are obtained by projecting the original policy onto smaller sub-action spaces, enabling a disintegration of the complex decision problem into a set of simpler problems. We allow independent actors to update the corresponding partial policy functions each utilizing its own value function (i.e., critic). Fig. 1 shows a schematic illustration of the localization process using the partial policies for a 2D case.

II. Related Work

Prior research on landmark localization in both medical image and computer vision concentrated on the model-based methods where a major focus was on the classification-based approaches. For example, Shotton et al. [10] utilized a trained forest to perform per-pixel classification to recognize body

*Email: yun@hufs.ac.kr

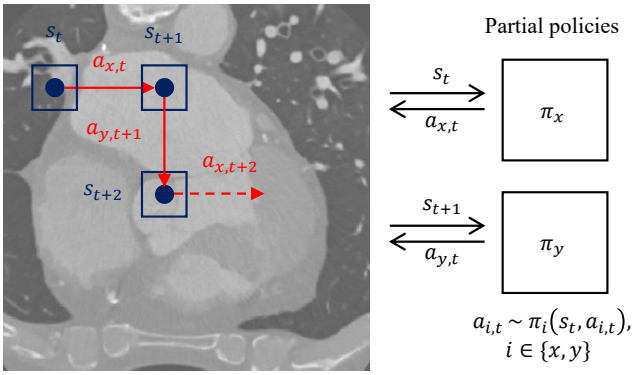

Fig. 1. **Landmark localization process using the partial policies.** A 2D case is used for illustration purposes. An agent initiated at any position at time t, observes the corresponding state s_t and decides an action to move towards the target landmark. Instead of a single RL agent with a single policy π, multiple sub-agents with simpler partial policies π_x and π_y repeatedly contribute to the successive state transitions. $a_{i,t}$ refers to the partial action at t, sampled from the partial policy distribution π_i on the i-th sub-action space, for the current state s_t.

parts as an intermediary step required for localizing body-joints in depth images. Zheng et al. [11] proposed a marginal space learning approach to localize aortic valve landmarks hierarchically, where a rough location is first obtained from a detected global object consisting of the landmarks, followed by a refinement session using local boosting tree based classifier. A generalized Procrustes analysis (GPA) was used to find the optimal global object. However, GPA does not guarantee the convergence of means [12]. Ionasec et al. [13] utilized a similar boosting tree classifier to localize the aortic valve landmarks. A forest classifier based method similar to [10] was also proposed for hand-joint localization in X-ray images. Nevertheless, classification approaches suffer from dataset imbalance problem because of the negligible positive samples, resulting in a biased classifier.

Regression-based approaches are becoming more and more popular instead of formulating the localization task as a classification problem. Regression models also showed significant improvement comparing to the classification models. These models suggest exploiting the predictions in different regions at a different distance from the target landmark. Criminisi et al. [14] proposed an efficient method for anatomy localization using a regression forest. Jung et al. [15] proposed the random tree walk (RTW) method for localizing human body joints in depth images, where a regression tree is trained to estimate the direction to the target joint for each position. A walker then starts walking using the learned direction and the expectation of the stepped positions is considered as the resultant joint position. An implication of RTW for localizing the aortic valve landmarks can be found in our previous work [16], where we performed a colonial walk initiating multiple random walks from different initial positions. The successful walker from the colony was elected by the minimum walk variance measure. Some joint models combining both the regression and classification approach also exist [17], [18]. Most recently, a stratified method is introduced by Oktay et al. [19], where image-patch driven local information as well as the global information in terms of organ-size, shape etc. are used for training.

There are a few approaches that do not fall under either of the above mentioned approaches. Simple connected component analysis [20]and coronary centerline tracking [21] is used for coronary ostia detection after aorta segmentation. However, robustness is challenged under image noise in case of connected component analysis, while centerline tracking algorithm has a high computational cost when operated for the whole surface of aorta. Elattar et al. [22] detected coronary ostia and aortic hinges on the aortic root surface, which is segmented using connected component analysis.

Most of the previous learning-based approaches are generally exploitative and may face generalization problem. RL provides an explorative learning scheme, which has shown remarkable improvements in various fields of image processing. To the best of our knowledge, only one instance of RL-based landmark localization exists in the literature, whereas a few instances of bounding box-based object localization in natural images can be found. Caicedo et al. [23] presented the object localization problem as a sequential decision-making problem in RL, where the process starts with a bounding-box covering the whole image, gradually applying transformation actions to the bounding box to finally localize the object. A better intersection-over-union (IoU) between the transformed box and the GT bounding box yields a positive reinforcement, and negative otherwise. The state is related to the image or sub-image inside the bounding box. Jie et al. [24] proposed a tree-structured RL with similar state and action representation. At each state, the agent applies two different actions i.e., scaling and translation, yielding two resultant states. Thus, the agent follows a recursive approach to finally find the object, representing a binary tree-like search. Ghesu et al. [25] present the only RL implementation in anatomical landmark localization in medical images, where the agent makes sequential position update actions to reach the landmark. The state is defined to be the region-of-interest (ROI) around the corresponding position. Relative distance change is used as the rewarding scheme. All these approaches mainly implemented Q-learning [26] to learn the action-value function while learning the optimal behavior indirectly. Direct behavior learning through optimizing the policy function shows better convergence comparatively, while having the problem of high variance. The actor-critic approach also approximates the policy function, however, uses an additional value function as the critic to reduce the variance [8].

Our work focuses on RL formulation for landmark localization in 3D medical images, where the agent action follows a definition similar to [25]. Unlike the previous approaches, we directly approximate the policy function following the actor-critic approach, where a state-value function is used as the policy evaluator or the critic. Moreover, we propose to learn multiple partial policies on different sub-action spaces instead of a single complex policy on the original action space, in order to improve the slow learning problem of RL and ensure a more robust localization.

The rest of the paper is organized as follows. Section III de-

scribes the formulation of landmark localization for the actor-critic RL. Section IV presents the partial policy-based RL for localization. Section V reports the experimental evaluation of the proposed approach. Finally, Section VI presents the concluding remarks.

III. Landmark Localization as RL

In our reinforcement learning-based localization scheme, an RL agent initiated at a random position interacts with the volume by taking consecutive discrete actions sampled from the learned policy distribution for the observed state at the current position, to finally reach the target landmark. During training, the agent tries to attain an optimal policy that maximizes the long-term return formulated as a discounted cumulative reward. The following is the description of the key elements of the Markov decision process (MDP) in the RL-wrapped localization scheme.

A. State

We represent the state as a function of the agent-position. For any position $\boldsymbol{q}$, the corresponding state $\boldsymbol{s} = \mathcal{S}(\boldsymbol{q})$ refers to a stack of the axial, coronal, and sagittal sub-images observed through a squared window centered at the corresponding position. Thus, we allow the agent at any position to observe an $m \times m \times 3$ block of surrounding voxels. Here, m is the window size. Such state is useful to provide a pseudo-3D view but requires less storage in the experience replay memory. This also helps traverse the state through a usual 2D CNN (of the policy and value networks), treating it as a 3-channeled image.

B. Action

Similar to Ghesu et al.'s approach [25], a discrete action space is considered where agent can take a unit step along either of the axes to update its position to a neighbouring voxel. Therefore, the agent holds three degrees of freedom to move, enabling six actions i.e., $right, left, up, down, slice\_forward, slice\_backward$. The first four moves are along the axial slice (X and Y axes), whereas, the last two actions allow the agent to jump across the slices moving along Z-axis. We represent our action space as follows:

$$\mathcal{A} = \{x^+, x^-, y^+, y^-, z^+, z^-\} \tag{1}$$

where x^+ , x^-, y^+, y^-, z^+ and z^- represent $right$, $left$, up, $down$, $slice\_forward$ and $slice\_backward$, respectively.

Using these simple actions yields a rather simple and deterministic transition. For a given position $\boldsymbol{q}$ and action a, the transitioned position $\boldsymbol{q'}$ can easily be obtained using the following transition function:

$$\begin{aligned} \boldsymbol{q'} &= \mathcal{T}(\boldsymbol{q}, a) = \big(q_x + U_x(a), q_y + U_y(a), q_z + U_z(a)\big) \\ U_i(a) &= \begin{cases} \eta, & \text{if } a = i^+ \\ -\eta, & \text{if } a = i^- \\ 0, & \text{otherwise} \end{cases} \\ i &\in \{x, y, z\} \end{aligned} \tag{2}$$

Here, η is the length of a unit-step. q_x, q_y, and q_z are the components of $\boldsymbol{q}$ along different axes. We denote such transition as $(\boldsymbol{q}, a, \boldsymbol{q'})$. There could be an additional action for no transition where agent remains at its current position without moving so that we can know that it has reached its destination landmark. However, adding such action would make the optimal policy finding harder because the state-action space (that should be explored) would become larger. Moreover, comparing to other actions, this action can render positive reward only for one state in the whole volume. Thus the model will suffered from sample selection bias and highly unlikely to trigger this action. Finally, we can ensure a satisfactory localization by using only just the aforementioned six actions. Because, eventually it would converge the target and move back and forth creating an oscillation of an amplitude of 1-2 voxels. The final localized position is the centroid of the oscillation, and may be approximated by taking the expectation of last few steps.

C. Reward

The agent at any position inside a 3D volume targets at choosing an action that maximizes the discounted cumulative reward. Therefore, we should encourage the agent to come closer to the target by giving an appropriate reward. We propose to use a simple binary reward function, where a positive reward is given if an action leads the agent closer to target landmark, and a negative reward is given otherwise. The reward is immediate after each action. The Euclidean distance measure is undertaken to assess the closeness. Hence, for a transition $(\boldsymbol{q}, a, \boldsymbol{q'})$, we can represent our reward function as follows:

$$\begin{aligned} \mathcal{R}(\boldsymbol{q}, a, \boldsymbol{q'}) &= \text{sign}(d_{\boldsymbol{pq}} - d_{\boldsymbol{pq'}}) \\ d_{\boldsymbol{ab}} &= ||\boldsymbol{a} - \boldsymbol{b}||_2 \end{aligned} \tag{3}$$

where $\boldsymbol{p}$ is the target landmark position. Such binary reward is widely used in reinforcement learning and useful for tracking the progress. Even in the case of a continuous real-valued reward definition, it is recommended to perform reward clipping, where all the positive and negative outcomes are labelled as +1 and -1, respectively.

D. Policy and value function

Policy function outputs the optimal action-probabilities for a given state, whereas value function outputs the expected cumulative return for a given state and/or a given action. We adopt a stochastic policy to map states to actions. Previous RL-based localization approaches focused on implicit policy learning through training a value function approximator. We exploit a direct and explicit policy learning that comes under the category of policy-based RL, performing direct parametrisation of the policy function. In the proposed approach, a non-linear policy function approximator represented by a multi-layer perceptron (MLP) on top of a deep-CNN is used, embedding the high level feature learning from the raw state inside the policy learning. The parametrised policy function can be represented as follows:

$$\pi_\theta(\boldsymbol{q}, a) = P(a|\mathcal{S}(\boldsymbol{q}), \theta) \tag{4}$$

where θ represents the weights of the deep policy network.

Direct policy search methods have a better convergence property while inducing a high variance problem. In actor-critic RL, an additional value function serves as a policy evaluator or critic to tackle the high variance problem. We use the state-value function approximator that tries to evaluate the policy, π_θ, for the current policy parameters, θ. We represent the value function approximator by another MLP stacked on top of the same CNN from the policy net, trying to approximate the state-value function to the true state-value (i.e., expected cumulative return for a state) for a given policy, as expressed as follows:

$$V_\omega(\boldsymbol{q}) \approx V^{\pi_\theta}(\mathcal{S}(\boldsymbol{q})) \tag{5}$$

where ω refers to the network parameters of the value approximator network. Therefore, both the policy and value function share the parameters of a common CNN while having their own exclusive MLP parameters. Fig. 3a presents the network architectures of the policy and value function in an actor-critic approach.

E. Learning

Using the aforementioned policy and value function approximator, actor-critic learning requires updating the parameters of both the policy (actor) and value (critic) networks. For a given state, actor-update aims at improving the policy to ensure a better cumulative return than the state-value inferred by the critic, whereas the critic aims at updating the value to approximate the cumulative return for the current policy. We perform the actor-critic RL in the widely used temporal differnce (TD) learning framework [27]. We used the simplest linear TD(0) approach, where TD-target and TD-error are respectively calculated for a transition $(\boldsymbol{q}, a, \boldsymbol{q'})$ using the following equations:

$$\begin{aligned} \tau(\boldsymbol{q}, a, \boldsymbol{q'}) &= \mathcal{R}(\boldsymbol{q}, a, \boldsymbol{q'}) + \gamma V_\omega(\boldsymbol{q'}) \\ \varepsilon(\boldsymbol{q}, a, \boldsymbol{q'}) &= \tau(\boldsymbol{q}, a, \boldsymbol{q'}) - V_\omega(\boldsymbol{q}) \end{aligned} \tag{6}$$

Thus, TD-target τ refers to the discounted cumulative return with a discount factor γ using the current policy, and TD-error ε refers to the advantage of the current policy over the critic-inferred state-value. Our goal is to approximate τ by the parametrised value function and update the policy towards the advantage. Hence, the cost functions for updating the value and policy parameters is stated as follows:

$$\begin{aligned} J_V(\omega) &= \mathbb{E}_{(\boldsymbol{q},a,\boldsymbol{q'})}\big[\big(\tau(\boldsymbol{q}, a, \boldsymbol{q'}) - V_\omega(\boldsymbol{q})\big)^2\big] \\ J_\pi(\theta) &= \mathbb{E}_{(\boldsymbol{q},a,\boldsymbol{q'})}\big[-\varepsilon(\boldsymbol{q}, a, \boldsymbol{q'}) \log \pi_\theta(\boldsymbol{q}, a)\big] \end{aligned} \tag{7}$$

In pure policy-based RL, expected return is used in the cost function of the policy network. The utilization of the advantage function formulated as TD-error in (6) serves as a better cost function, enabling a low variance in the policy approximation.

IV. Partial Policy-based RL

RL agent learns the optimal behavior from episodes of experience gathered by interacting with the environment. In problems with large state/action space, successful task learning is challenging as it requires a huge number of trials, becoming a major drawback of RL. Despite the various methods of exploration, the issue still practically remains. Moreover, availability of multiple actions triggering a positive feedback for a state makes the optimal policy learning harder. In the case of 3D localization, we have six actions along X,Y and Z axes. Except for a negligible portion in the state space, a number of alternative actions can be found triggering a positive feedback for most of the states of a volume. Consequently, learning the decision between a pair of actions along one axis alone, can ensure positive feedbacks for almost the entire state space. Therefore, actions along other axes does not get significance to influence the policy update. The situation is illustrated in Fig. 2, where localization in a 2D slice is presented for simplicity. The agent using such biased policy is most likely to generate only the actions along X-axis, failing to learn the task.

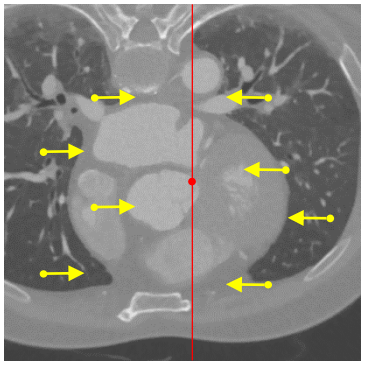

Fig. 2. **Optimal policy problem in presence of alternative actions triggering positive feedback.** The depicted policy ensure rewards in all states by deciding between the actions along X (horizontal) axis only, except for the states lying on the red vertical line passing through the red dotted target landmark.

For an efficient and effective learning of the optimal policy, we propose a partial policy-based learning approach. Instead of using one actor to update a large policy, we employ multiple micro-actors to learn the optimal behavior in partial sub-action spaces. Thus, the decomposition of the large problem into smaller problems enables an efficient, successful, and easier learning. Consequently, the proposed learning scheme can achieve the optimal policy within a fewer trials, compared to the conventional deep RLs. In this section, we first describe the dissection of the master policy to obtain partial policies, followed by the reconstruction of the original policy from the partial policies. Finally, we present the TD learning-framed actor-critic algorithm using the partial policies.

A. Partial policy

The objective of the partial policy is to obtain multiple simple policies on the projections of the actual action space, where the projected policies are able to reconstruct the policy on the original action space. We define smaller sub-action spaces i.e., partial action spaces, projecting the actual action space onto different Cartesian axes. Such dissection of the actual space also suggests multiple sub-agents corresponding

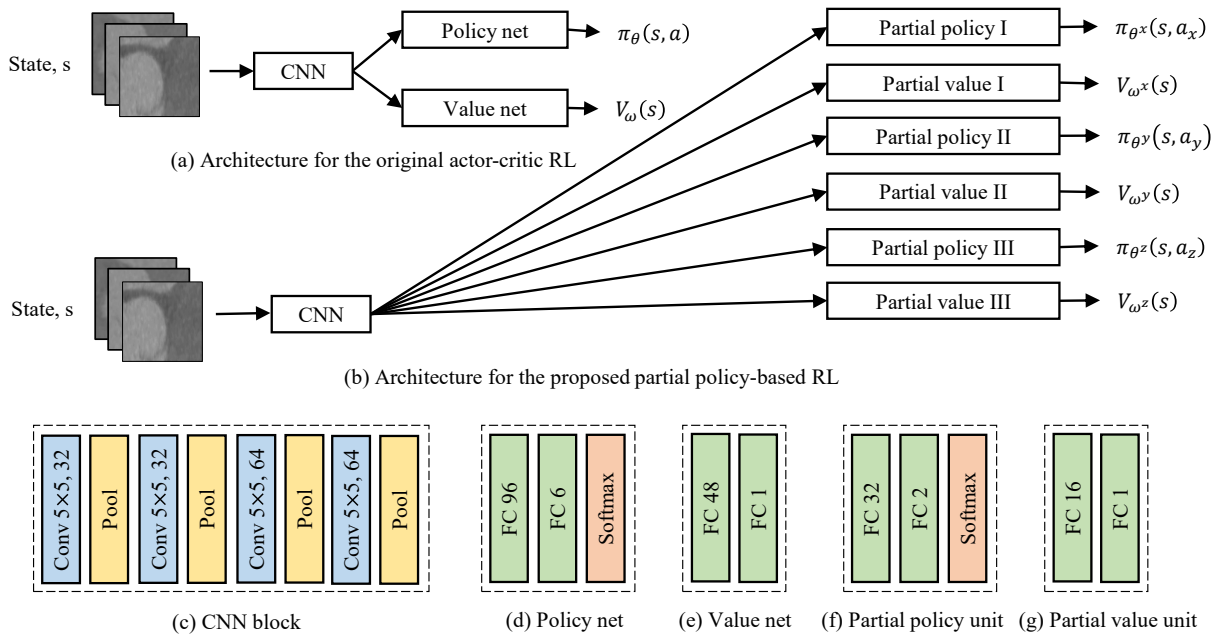

Fig. 3. **Architectures of the original actor-critic RL and the proposed partial policy-based RL.** For a state, the policy net outputs probabilities for all possible actions, and the value net gives a single scalar value as the long-term return. FC stands for fully-connected layer.

to the partial action spaces, each of them trying to maximize the expected cumulative reward by taking optimal actions sampled from the corresponding sub-action space. Thus, we use multiple sub-agents learning smaller sub-tasks, instead of one agent with a large task. The actual action space in our 3D localization problem is decomposed into the following partial action spaces:

$$\mathcal{A}_i = \{i^+, i^-\}, i \in \{x, y, z\}. \tag{8}$$

Partial policy refers to the policy undertaken by the sub-agents to map state to actions in the corresponding axial domain. Thus, partial policies are projections of the original policy, defining the stochastic behavior for the partial action spaces. Therefore, we can define three partial policies with respect to the partial action spaces. Three independent MLPs sharing a common preceding CNN are used to represent the partial policies. To evaluate the partial policies, we also define three value function approximator networks stacked on top of the same CNN. Therefore, we have 6 MLPs preceded by a common CNN. For each sub-action space, there is a sub-actor and sub-critic available to update the corresponding partial policy and value function. The partial policy and value function approximators for our problem can be expressed as follows:

$$\begin{aligned} &\pi_{\theta^i}(\boldsymbol{q}, a_i) = P(a_i|S(\boldsymbol{q}), \theta^i), a_i \in \mathcal{A}_i, i \in \{x, y, z\} \\ &V_{\omega^i}(\boldsymbol{q}) \approx V^{\pi_{\theta^i}}(S(\boldsymbol{q})), i \in \{x, y, z\} \end{aligned} \tag{9}$$

where, θ^i and ω^i are the network parameters for the i-th partial policy and value approximators. Fig. 3 shows the overall network architecture.

Introducing the partial policies, the learning problem become easier and simpler in the shrunk action spaces. The goal of learning a partial policy is to provide the sub-agent at a state with the probability of the actions in the corresponding partial action space. The sub-agent only needs to choose between two actions to maximize the cumulative return. Therefore, the partial policy learning shares the same idea with the simplest binary classification problem. Attaining the optimal partial policy is easier, enabling a better convergence. The original definitions of state and reward function are used without any alteration.

B. Reconstruction

Partial policy ensuring a simpler learning is not adequate without the definition of actual policy reconstruction from the partial policies in order to employ it appropriately. sub-agents can decide the optimal partial actions from the learned partial policy. Deriving the actual action $\boldsymbol{a}_m$ combining the partial actions $a_i, i \in x, y, z$ is equivalent to sampling the action from the actual reconstructed policy. The actual action can be reconstructed as follows:

$$\begin{aligned} \boldsymbol{a}_m &= \sum_{i \in \{x,y,z\}} a_i \zeta_i, \\ a_i &\sim \pi_{\theta^i}(\boldsymbol{q}, a_i) \end{aligned} \tag{10}$$

Here, ζ_i is the basis vector of the i-th axis. $\boldsymbol{a}_m$ is the actual action at position $\boldsymbol{q}$.

On the other hand, merging the partial policies to directly estimate the policy can be done by cascading the partial policies followed by normalization as expressed in the following:

$$\pi(\boldsymbol{q}, :) = \frac{1}{3} \bigcup_{i \in \{x,y,z\}} \{\pi_{\theta^i}(\boldsymbol{q}, i^+), \pi_{\theta^i}(\boldsymbol{q}, i^-)\} \tag{11}$$

Using either approach of (10) and (11) requires traversing all the three partial policy networks to estimate a single optimal

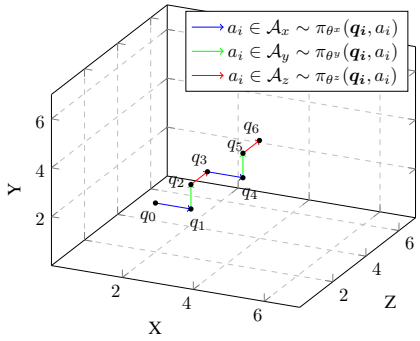

Fig. 4. **Sequential exploitation of the learned partial policies for localization.** Sub-agents are periodically exploited to make a number of step sequences, thus walking a Manhattan-like distance.

behavior for a single state. Moreover, the critic evaluates a policy by quantifying the next state arrived after an action sampled from that policy. Utilizing the above reconstruction suggests a common next state obtained by a collaborative decision on the partial actions. This no longer holds the assumption of independent partial policy learning, again making the problem difficult.

To ensure a greater efficiency, we propose a work-around to employ the partial policies to approximately represent the original policy. We suggest a periodic and sequential deployment of the partial policies, that can achieve the goal maintaining the efficiency. Thus, we perform a step-sequence instead of a single step. We define the k-th step-sequence s_k as follows:

$$
\begin{aligned}
s_k &= (a_{x,k}, a_{y,k}, a_{z,k}) \\
a_{x,k} &\sim \pi_{\theta^x}(\boldsymbol{q}_t, a_{x,k}) \\
a_{y,k} &\sim \pi_{\theta^y}(\boldsymbol{q}_{t+1}, a_{y,k}) \\
a_{z,k} &\sim \pi_{\theta^z}(\boldsymbol{q}_{t+2}, a_{z,k}) \\
\boldsymbol{q}_t &= \mathcal{T}(\boldsymbol{q}_{t-1}, a_{z,k-1}) \\
\boldsymbol{q}_{t+1} &= \mathcal{T}(\boldsymbol{q}_t, a_{x,k}) \\
\boldsymbol{q}_{t+2} &= \mathcal{T}_(\boldsymbol{q}_{t+1}, a_{y,k})
\end{aligned}
\tag{12}
$$

where $\boldsymbol{q}_t$ refers to the position at time step t, and $k = \lfloor \frac{t}{3} \rfloor$ is the order of step-sequence. We apply this sequence repeatedly, enabling a periodic selection of the partial policies. This also assures a balanced exploration in all sub-action spaces. The agent explicitly updates its position by taking a partial action from the defined sequence, contributing to the independent learning of the partial policies, because critic is able to give the feedback on the transitioned state solely reached by exploiting an individual partial policy. Only the responsible policy and value parameters are updated for a transition. The periodic application of the partial action sequence is depicted in Fig. 4. It is apparent that the actual agent is crossing a distance similar to Manhattan distance, periodically exploiting the sub-agents. The order of the partial actions in the unit sequence is not significant as long as they are repeated periodically in the overall action sequence.

C. Actor-critic RL for partial policy

The periodic application of the partial actions sampled from the partial policy functions establishes the foundation of partial policy-based actor-critic learning. The same TD-framework is used maintaining the originality of the reward and transition, because those are not affected by the partial policy. Three independent micro-actors are responsible for updating the partial policy networks, each having a corresponding critic.

At each step, one of the sub-agents are allowed to interact with the environment using the current parameters of the corresponding partial policy function, and utilize the critic to get directions to update the parameters. The critics also update the value based on the discounted return obtained by the current policy. Periodic deployment ensures the balance in learning all the partial policies and values. Algorithm 1 presents the comprehensive actor-critic method for partial policy-based reinforcement learning. While original actor-critic method operates on step, the partial policy-based actor-critic method operates on unit sequence consisting of three partial steps. For each partial steps, parameter updates occur in the corresponding partial policy function as well as in the value function. Though we used batch methods to update the networks from experience replay, the algorithm presented here uses the incremental method for easier interpretation. In batch-methods, we first gather episodes of experience and store them in an experience replay memory, then sample mini-batches from the memory to perform a stochastic gradient descent.

Algorithm 1 Actor-critic RL using partial policy

```
Initialize q, θ^x, θ^y, θ^z, ω
for each step sequence do
  for i ∈ {x, y, z} do
    Sample a_i ~ π_{θ^i}(q, a_i)
    Next position, q' = T(q, a_i)
    Reward, r = R(q, a_i, q')
    TD-target, τ(q, a_i, q') = r + γ V_{ω^i}(q')
    TD-error, ε(q, a_i, q') = τ(q, a_i, q') − V_{ω^i}(q)
    θ^i ← θ^i + α ∇_{θ^i} ε(q, a_i, q') log π_{θ^i}(q, a_i)
    ω^i ← ω^i − α ∇_{ω^i} (τ(q, a_i, q') − V_ω(q))^2
    q ← q'
  end for
end for
```

V. Experiment

We evaluate the proposed partial policy based RL method in three different problems in three datasets obtained from three different sites. Table I summarizes the evaluation datasets and the corresponding problems. The first dataset (Dataset-A) contains 71 contrast-enhanced coronary CT angiography (CCTA) volumes of 71 different patients. The corresponding problem is to localize the eight landmarks of the aortic valve (three hinge points, three commissure points, and two coronary ostia). Aortic valve (AV) landmark localization plays a vital role in preprocedural planning of transcatheter aortic valve implantation (TAVI) [11], which is an implant-based treatment method for severe aortic stenosis. Moreover, assessing the

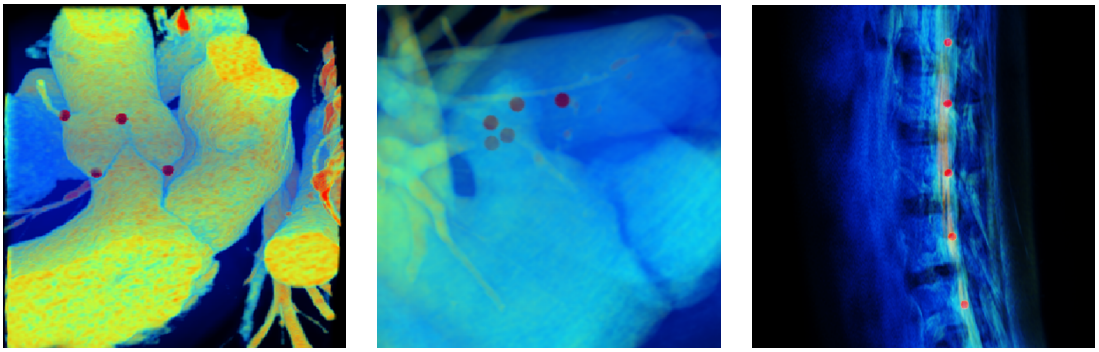

Fig. 5. **Proposed partial policy based RL-localized landmarks.** (*left*) Localized non-coronary and right coronary hinge points, the commisure point between them, and right ostium of the aortic valve in a CCTA volume are shown with other landmarks being occluded. (*middle*) Localized LAA seed-points for different initial positions in a CT volume. (*right*) Localized vertebra-centers in a spine MR volume.

TABLE I
EVALUATION DATASET AND PROBLEM DESCRIPTION.

	Data	**Voxel dimension**	**Problem**
Dataset-A	71 coronary CT	0.35 mm × 0.35 mm × 0.5 mm	Aortic valve landmarks
Dataset-B	150 cardiac CT	0.45 mm × 0.45 mm × 0.5 mm	LAA seed-point
Dataset-C	20 spine MR (Public)	0.58 mm × 0.58 mm × 1.5 mm	Vertebra centers

TABLE II
FOURFOLD CROSS VALIDATION TEST RESULTS FOR LOCALIZING THE AORTIC VALVE LANDMARKS IN CCTA VOLUMES (DATASET-A) AND LEFT ATRIAL APPENDAGE SEED-POINT IN CT VOLUMES (DATASET-B). FOR DATASET-A, LOCALIZATION ERROR IS PRESENTED AS THE EUCLIDEAN DISTANCE FROM THE GROUND TRUTH POSITION. THE RESULTS FOR THE TAVI AND NON-TAVI VOLUMES ARE PRESENTED SEPARATELY. FOR DATASET-B, THE PERCENTAGE OF THE LOCALIZATION FAILURE IS PRESENTED WITH RESPECT TO THE TOTAL NUMBER OF TRIALS FROM DIFFERENT RANDOM POSITION.

	Landmark/method	**Partial Policy RL**		**RL**	
		Mean ± SD	Median	Mean ± SD	Median
Non-TAVI error (mm)	Hinge points	1.96 ± 0.98	1.72	2.18 ± 1.21	1.92
	Commissure points	1.95 ± 0.93	1.69	2.16 ± 1.18	1.87
	Coronary ostia	1.91 ± 0.95	1.65	2.13 ± 1.20	1.81
TAVI error (mm)	Hinge points	2.08 ± 1.24	1.91	2.30 ± 1.27	2.11
	Commissure points	2.02 ± 1.15	1.85	2.24 ± 1.22	2.06
	Coronary ostia	1.94 ± 1.01	1.68	2.15 ± 1.32	1.94
LAAO failure (%)	LAA	7.21 ± 5.84	6.50	11.62 ± 7.71	9.22

valve is a clinical routine during any cardiac CT interpretation. However, it is a time consuming task because the valve anatomy is not easily perceived in the conventional CT views. Among the 71 volumes, 31 volumes are preprocedural CT obtained from actual TAVI-patients. Accurate localization in TAVI volumes is challenging because valvular calcification can significantly affect the anatomy in unpredictable ways.

Using the second dataset (Dataset-B) consisting of 150 cardiac CT volumes, we localized the left atrial appendage (LAA) seed-point, which can facilitate an automatic segmentation of the appendage. LAA segmentation is helpful for physicians because it is a major site of thrombosis potentially responsible for inducing stroke-risk in non-valvular atrial fibrillation [28]. Related prior works proposed different segmentation approaches, however, within a manually marked bounding box (i.e., volume of interest) [29]. The prior annotation of such bounding box enclosing LAA is a major obstacle of the approaches to become fully automatic. Therefore, localizing the aforementioned seed-point inside LAA can contribute to attaining an automatic segmentation method. Whereas the target points in the previous problem are specific, this problem suggests localizing any point inside the appendage. There is a large variation in appendage anatomy with an additional variation for different cardiac phase. The 150 volumes are obtained from 30 different patients in 5 different cardiac phases.

The third dataset (Dataset-C) is a public dataset consists of 20 MR images of spine targeted at vertebra recognition for spine structure analysis [30]. This dataset is available at the SpineWeb online repository. We implemented our proposed method to localize the centers of the vertebra. We localized 5 lumbar vertebra (L1 L5). Among the 20 volumes, one has problematic ground truth (GT) annotation and one volume captured the head to shoulder region where the intended vertebra

TABLE III
AORTIC VALVE LANDMARK LOCALIZATION ERROR COMPARISON WITH DIFFERENT APPROACHES.

Method	**AV Localization error (mm)**
	Mean ± SD
Partial Policy RL	1.98 ± 1.03
Actor-critic	2.19 ± 1.23
DQN	2.26 ± 1.35
RTW [16]	2.35 ± 1.48
Inter-observer difference [22]	2.38 ± 1.56

TABLE IV
FOURFOLD CROSS VALIDATION TEST RESULTS FOR LOCALIZING THE VERTEBRA CENTERS IN SPINE MR VOLUMES (DATASET-C) USING THE PARTIAL POLICY-BASED RL.

Vertebra	**Localization error (mm)**	
	Mean ± SD	Median
L1	2.79 ± 2.18	2.66
L2	2.61 ± 2.02	2.54
L3	2.58 ± 1.84	2.52
L4	2.86 ± 1.81	2.80
L5	3.10 ± 2.08	3.05
Overall	**2.79 ± 1.98**	**2.71**
[30]'s result*	2.87 ± 2.04	2.80

*Same dataset but different train/test split

are not present. Therefore, we proceeded our experiment using 18 volumes.

For all the datasets, necessary ground truth positions of the target landmarks were obtained from the corresponding site, which were used to process the reward signals for the RL agent. For all the experiments, a common set-ups for RL is used. We implemented both the original actor-critic RL and the proposed partial policy-based RL for Dataset-A and B to obtain a comparative evaluation. To compare the proposed method with the widely-used DQN, we also implemented DQN for Dataset-A. For a fair comparison, we used identical parameters and hyper-parameters for all the methods. A window size of $m = 50$ is used for the state, and the unit step size is set to 2 voxels. For each epoch, the agent was allowed to gather around 300 episodes of experience using its current policy, where each episode consists of 300 steps (or, 100 step-sequences in case of the partial policy). A replay memory of size 10^5 is used to store the transitions. For the partial policy-based approach, three replay memories are maintained to keep track of the corresponding partial transitions. Sampling mini-batches from the experienced transitions, we perform stochastic gradient descend to update the policy and value network. The learning rate for updating both the value and policy was $\alpha = 10^{-4}$, and the discount factor γ was set to 0.9. The CNN consists of 4 sets of convolutional, ReLU and max-pooling layer stacks (Fig. 3). The final layer was flattened to obtain a non-spatial representation. 6 different MLPs were connected to the final flat layer of the CNN, representing the partial policy and value functions. For the original RL, only 2 MLPs are connected to represent a single set of policy and value functions. All the policy nets have a final softmax-gating to generate action-probabilities. Though ϵ-greedy approach [3] is widely used for exploration in RL, Bayesian approach to allow the agent to define its own uncertainty has shown to perform better. Practically, the uncertainty is simulated by adding a dropout layer in the network [31]. We gradually annealed the dropout keep probability from 0.1 to 0.7 over the epochs. To localize the vertebra-centers in Dataset-C using the proposed method, we use the same architecture and hyper-parameter settings.

A four-fold cross validation is performed on patient-basis to evaluate the localization performance. For the first and third experiments, the localization error is calculated in terms of the Euclidean distance of the localized position from the corresponding ground truth. For the second experiment, such distance is not an appropriate measure because the assigned goal is to detect any point (seed) inside the left atrial appendage, and the annotated ground truth were also not specific. Therefore, the performance is measured using a binary comparison (i.e., whether the localized point was inside the appendage or not). During the test/validation session, the agent was not provided with any reward signal. For each test case, we conducted the localization process initiating the agent from different random positions inside the test volume, and presented the average localization result. Fig. 5 depicts the qualitative localization results of the proposed partial policy-based RL. In Table II, we present the localization error (for Dataset-A) and localization failure percentage (for Dataset-B) of the proposed partial policy approach against the original actor-critic RL. Table III presents the average AV landmarks localization error comparison for different methods. Table IV presents the average localization error of the vertebra centers in spine MR images. The average computation time for localization is 1.2 seconds, as tested with a 3.60GHz single-core CPU and a GeForce GTX TITAN Xp GPU. Computation time for all the cases is same because an equal number of steps is performed in all cases.

The proposed method showed an average error of 1.98 ± 1.03 mm localizing the AV landmarks in CCTA volume, whereas Elattar et al. [22]'s error for localizing the hinge points and ostia in CTA was 2.65 ± 1.57 mm, and Zheng et al. [11]'s error for localizing all the landmarks in C-arm CT was 2.11 ± 1.34 mm. The inter-observer difference in CTA was 2.38 ± 1.56 mm, as reported in [22]. The proposed method also showed improvement comparing to the random tree walk method in our previous work using the same dataset, where the average localization error was 2.35 ± 1.48 mm [16]. On the other hand, RTW takes only a few milliseconds to localize a landmark because of its simple feature computation. However, CNN-based feature computation in the currently proposed method is more useful and the current localization time of 1.2 seconds can still be considered efficient and allowable for such an improved accuracy. In our previous work, we also introduced a colonial walk method utilizing multiple walks from multiple initial positions and choosing the final walk by minimum walk variance. Such extension can also be performed with the current RL agent walk to improve the

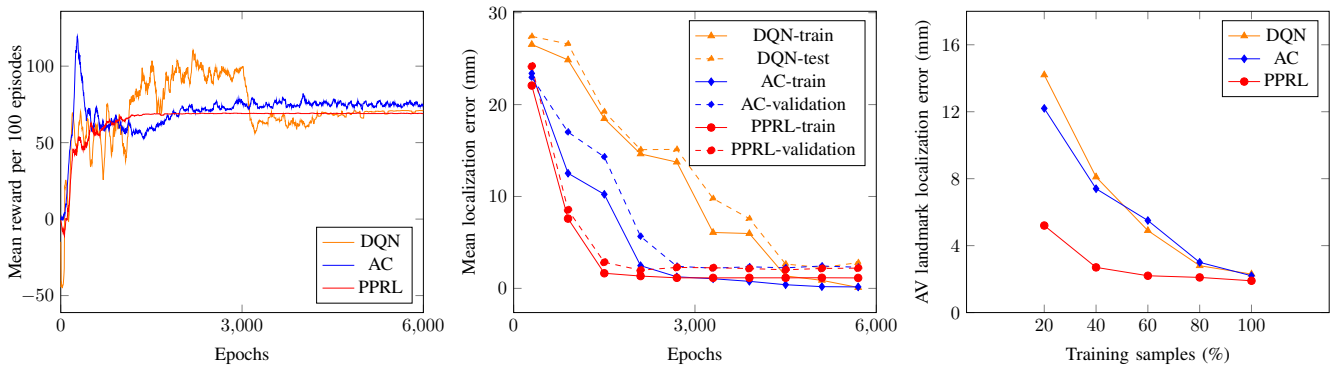

Fig. 6. **Learning curves of the proposed partial policy-based RL.** (*left* and *middle*) Average reward and localization error over different epochs, for AV landmark localization. Reward plot is smoothed out for better visualization. (*right*) Localization performance with respect to training data size. Training set size is gradually increased while validating with a common test set. The proposed method could achieve almost the maximum accuracy with a fewer training examples.

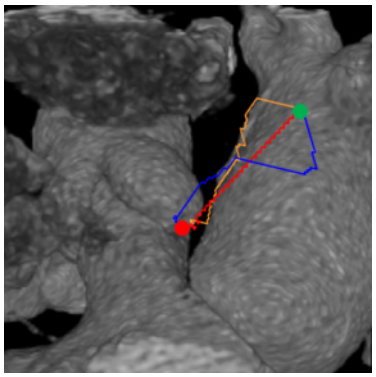

Fig. 7. **Search paths of different RL agents localizing a hinge point.** The transitions of the PPRL, acotr-critic, and DQN agents is indicated by red, blue, and orange lines, respectively. Green and red dots indicate the initial position and target landmark.

TABLE V
AORTIC VALVE LANDMARK LOCALIZATION RESULTS IN CALCIFIED TAVI VOLUMES BY AN AGENT TRAINED ON ONLY NON-TAVI VOLUMES.

Localization error (mm)	**Partial Policy RL**	**RL**
	Mean ± SD	Mean ± SD
Hinge points	2.52 ± 1.78	3.38 ± 2.19
Commissure points	2.35 ± 1.74	3.16 ± 2.17
Coronary ostia	2.26 ± 1.68	2.97 ± 2.02

accuracy even more, which we keep as our future work. As for the center locations of the vertebra in spine MR volumes, the proposed method showed an average error of 2.79 ± 1.98 mm, which is as good as [30]'s result.

The proposed partial-policy based method exhibited noteworthy improvement over the original actor-critic and the DQN approach. The average localization error is improved with a significant reduction of error variance. Learning the partial policies facilitated an improved localization with simpler decision process. At each step sequence, the sub-agents solve three binary decision problems. Consequently, the proposed method exhibited a noteworthy improvement comparing to directly learning the original policy. We also performed an additional experiment where the agent is trained only on the 40 non-TAVI volumes and attempts to localize the landmarks in the TAVI volumes. Thus, we could observe the agent behavior in a volume with valvular calcification, without providing any prior knowledge about calcified valves. Table V presents the average localization results for the proposed method and the conventional actor-critic RL. Partial policy could cope with the variation due to calcification significantly better than the conventional RL because of simpler decision space.

To compare the learning process and learned trajectories, we plot the average reward and localization error over the epoch from the learning process of the proposed method and the conventional ones, and illustrate optimal search paths in Fig. 6. A remarkable improvement is observed in the learnability of the proposed partial policy approach. It enabled a better and faster convergence. It converges within about half the epochs required by the conventional RLs (i.e., actor-critic and DQN), thus improving the slow learning problem in RL. Within a few-trials, the sub-agents could reach an optimal behavior, as depicted in the error plots. The search paths of the sub-agents also exhibit more confident transitions compared to the paths undertaken by the conventional agents (Fig. 7).

Preparing training data with ground truth acquisition is a difficult task in medical image processing. Therefore, it is advantageous to have a model that can give a satisfactory performance with knowledge of a fewer training data. Apart from the standard train/test splits, we randomly sampled about 20% of the dataset to be the test data (for Dataset-A). From the rest of the dataset, we gradually sampled 20%, 40%, 60%, 80%, and 100% to obtain five training subsets, where the last subset (100% training samples) refers to the whole training set. The test set is validated by the models trained on those subsets. For comparison, we used an identical split for the proposed PPRL, as well as the actor-critic and DQN approach. Fig. 6(*right*) presents our observation, where the proposed method could achieve very close to the maximum accuracy (achieved with 100% training samples) with a notably fewer training exmaples. Even with 20% of the training data, it could provide an average error of 3.8 mm, which is comparable to the 10 mm error of DQN and actor-critic. Thus, the partial-policy

based RL can potentially be useful in medical applications.

VI. Conclusion

Anatomical Landmark localization provides significant prior information for different applications in medical image processing. Our work presented a robust localization method formulated as reinforcement learning. For an efficient and successful learning with actor-critic method, we introduced a partial policy-based learning where multiple easier policies are learned on the sub-action spaces defined as projections of the original action space. Employing multiple sub-agents interacting with the environment, corresponding micro-actors and micro-critics independently update the deep partial policy and value networks, enabled a faster and better convergence. The experiment with aortic valve landmarks localization and left atrial appendage seed localization in 3D CT images, and vertebra localization in 3d spine MR images showed robust and improved performance, compared to the conventional actor-critic and widely used deep Q-learning approach. The proposed partial policy based approach required significantly fewer number of trials and fewer training data to achieve the optimal behavior, improving the slow learning problem of RL. The proposed method provides an efficient and potentially useful solution for localization, requiring an average localization time of 1.2 seconds.

Acknowledgment

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Education, Science, Technology (No. 2017R1A2B4004503).

References

[1] A. Sotiras, C. Davatzikos, and N. Paragios, "Deformable medical image registration: A survey," *IEEE Transactions on Medical Imaging*, vol. 32, no. 7, pp. 1153–1190, 2013.

[2] X. Han and Y. Zhou, "Systems and methods for segmenting medical images based on anatomical landmark-based features," Aug. 22 2017, US Patent 9,740,710.

[3] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[4] Y. Zhu, R. Mottaghi, E. Kolve, J. J. Lim, A. Gupta, L. Fei-Fei, and A. Farhadi, "Target-driven visual navigation in indoor scenes using deep reinforcement learning," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 3357–3364.

[5] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing atari with deep reinforcement learning," *arXiv preprint arXiv:1312.5602*, 2013.

[6] R. S. Sutton, D. A. McAllester, S. P. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," in *Advances in Neural Information Processing Systems*, 2000, pp. 1057–1063.

[7] V. R. Konda and J. N. Tsitsiklis, "Actor-critic algorithms," in *Advances in Neural Information Processing Systems*, 2000, pp. 1008–1014.

[8] I. Grondman, L. Busoniu, G. A. Lopes, and R. Babuska, "A survey of actor-critic reinforcement learning: Standard and natural policy gradients," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 6, pp. 1291–1307, 2012.

[9] Y. Duan, J. Schulman, X. Chen, P. L. Bartlett, I. Sutskever, and P. Abbeel, "RL^2: Fast reinforcement learning via slow reinforcement learning," *arXiv preprint arXiv:1611.02779*, 2016.

[10] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, and R. Moore, "Real-time human pose recognition in parts from single depth images," *Communications of the ACM*, vol. 56, no. 1, pp. 116–124, 2013.

[11] Y. Zheng, M. John, R. Liao, A. Nottling, J. Boese, J. Kempfert, T. Walther, G. Brockmann, and D. Comaniciu, "Automatic aorta segmentation and valve landmark detection in C-arm CT for transcatheter aortic valve implantation," *IEEE Transactions on Medical Imaging*, vol. 31, no. 12, pp. 2307–2321, 2012.

[12] A. Ross, "Procrustes analysis," *Course report, Department of Computer Science and Engineering, University of South Carolina*, 2004.

[13] R. I. Ionasec, I. Voigt, B. Georgescu, Y. Wang, H. Houle, F. Vega-Higuera, N. Navab, and D. Comaniciu, "Patient-specific modeling and quantification of the aortic and mitral valves from 4-D cardiac CT and TEE," *IEEE Transactions on Medical Imaging*, vol. 29, no. 9, pp. 1636–1651, 2010.

[14] A. Criminisi, D. Robertson, E. Konukoglu, J. Shotton, S. Pathak, S. White, and K. Siddiqui, "Regression forests for efficient anatomy detection and localization in computed tomography scans," *Medical Image Analysis*, vol. 17, no. 8, pp. 1293–1303, 2013.

[15] H. Y. Jung, S. Lee, Y. S. Heo, and I. D. Yun, "Forest walk methods for localizing body joints from single depth image," *PLoS ONE*, vol. 10, no. 9, p. e0138328, 2015.

[16] W. A. Al, H. Y. Jung, I. D. Yun, Y. Jang, H.-B. Park, and H.-J. Chang, "Automatic aortic valve landmark localization in coronary CT angiography using colonial walk," *PLoS ONE*, vol. 13, no. 7, p. e0200317, 2018.

[17] J. Gall, A. Yao, N. Razavi, L. Van Gool, and V. Lempitsky, "Hough forests for object detection, tracking, and action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 11, pp. 2188–2202, 2011.

[18] S. Schulter, C. Leistner, P. Wohlhart, P. M. Roth, and H. Bischof, "Accurate object detection with joint classification-regression random forests," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 923–930.

[19] O. Oktay, W. Bai, R. Guerrero, M. Rajchl, A. de Marvao, D. P. ORegan, S. A. Cook, M. P. Heinrich, B. Glocker, and D. Rueckert, "Stratified decision forests for accurate anatomical landmark localization in cardiac images," *IEEE Transactions on Medical Imaging*, vol. 36, no. 1, pp. 332–342, 2017.

[20] A. Hennemuth, T. Boskamp, D. Fritz, C. Kühnel, S. Bock, D. Rinck, M. Scheuering, and H.-O. Peitgen, "One-click coronary tree segmentation in CT angiographic images," in *International Congress Series*, vol. 1281. Elsevier, 2005, pp. 317–321.

[21] H. Tek, M. A. Gulsun, S. Laguitton, L. Grady, D. Lesage, and G. Funka-Lea, "Automatic coronary tree modeling," *The Insight Journal*, 2008.

[22] M. Elattar, E. Wiegerinck, F. van Kesteren, L. Dubois, N. Planken, E. Vanbavel, J. Baan, and H. Marquering, "Automatic aortic root landmark detection in CTA images for preprocedural planning of transcatheter aortic valve implantation," *The International Journal of Cardiovascular Imaging*, pp. 1–11, 2015.

[23] J. C. Caicedo and S. Lazebnik, "Active object localization with deep reinforcement learning," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2488–2496.

[24] Z. Jie, X. Liang, J. Feng, X. Jin, W. Lu, and S. Yan, "Tree-structured reinforcement learning for sequential object localization," in *Advances in Neural Information Processing Systems*, 2016, pp. 127–135.

[25] F. C. Ghesu, B. Georgescu, T. Mansi, D. Neumann, J. Hornegger, and D. Comaniciu, "An artificial agent for anatomical landmark detection in medical images," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2016, pp. 229–237.

[26] V. Mnih, N. Heess, A. Graves *et al.*, "Recurrent models of visual attention," in *Advances in Neural Information Processing Systems*, 2014, pp. 2204–2212.

[27] R. S. Sutton, "Learning to predict by the methods of temporal differences," *Machine Learning*, vol. 3, no. 1, pp. 9–44, 1988.

[28] M. Zoni-Berisso, F. Lercari, T. Carazza, and S. Domenicucci, "Epidemiology of atrial fibrillation: European perspective," *Clinical Epidemiology*, vol. 6, p. 213, 2014.

[29] C. Jin, J. Feng, L. Wang, J. Liu, H. Yu, J. Lu, and J. Zhou, "Left atrial appendage segmentation using fully convolutional neural networks and modified three-dimensional conditional random fields," *IEEE Journal of Biomedical and Health Informatics*, pp. 1–1, 2018.

[30] Y. Cai, S. Osman, M. Sharma, M. Landis, and S. Li, "Multi-modality vertebra recognition in arbitrary views using 3D deformable hierarchical

model," *IEEE Transactions on Medical Imaging*, vol. 34, no. 8, pp. 1676–1693, 2015.

[31] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *International Conference on Machine Learning*, 2016, pp. 1050–1059.