







# Collaborative Multi-agent Reinforcement Learning for Landmark Localization Using Continuous Action Space

Klemens Kasseroller<sup>1</sup>, Franz Thaler<sup>2</sup>, Christian Payer<sup>1</sup>,  
and Darko Štern<sup>2</sup>

<sup>1</sup> Institute of Computer Graphics and Vision, Graz University of Technology, Graz, Austria

<sup>2</sup> Gottfried Schatz Research Center: Biophysics, Medical University of Graz, Graz, Austria

`darko.stern@medunigraz.at`

**Abstract.** We propose a reinforcement learning (RL) based approach for anatomical landmark localization in medical images, where the agent can move in arbitrary directions with a variable step size. Using a continuous action space reduces the average number of steps required to locate a landmark by more than 30 times compared to localization using discrete actions. Our approach outperforms a state-of-the-art RL method based on a discrete action space and is inline with state-of-the-art supervised regression based methods. Furthermore, we extend our approach to a multi-agent setting, where we allow collaboration between agents to enable learning of the landmarks' spatial configuration. The results of the multi-agent RL based approach show that the position of occluded landmarks can be successfully estimated based on the relative position predicted for the visible landmarks.

**Keywords:** Reinforcement learning · Landmark localization · Collaborative multi-agent · Continuous action space

## 1 Introduction

Automatic localization of anatomical landmarks is an important step for a wide range of applications in medical image analysis, e.g. for registration or to initialize segmentation algorithms. Nevertheless, accurate anatomical landmark localization is also a challenging task due to anatomical and image intensity variations. Current state-of-the-art methods for anatomical landmark localization are based on supervised learning of Convolutional Neural Networks (CNNs) to either directly regress landmark coordinates [7] or their heatmap representation [11]. However, CNN based methods suffer from two major limitations. Either they require large amounts of memory to store the intermediate network outputs of the whole image on the GPU or, using patch-based approaches, depend on

an additional model for global guidance. Differently to supervised learning, reinforcement learning (RL) based approaches have the advantage, that the RL agent is able to internally keep a representation of the environment, i.e. the content of the medical images. Applied to a medical task such as anatomical landmark localization, this internal representation of the anatomy allows the RL agent to navigate through the image from any arbitrary starting position, without the need of an additional model for global guidance. Furthermore, by learning from patches, RL eliminates the need of storing the large intermediate network outputs of the whole image in GPU memory, which is a challenge, especially when working with large 3D volumes. Finally, the navigation through images based on the perception of local image information is similar to the way a physician localizes anatomical structures in a medical image. Indeed, the physician, based on their prior knowledge in human anatomy, can estimate the position of an anatomical structure relative to other structures in the image.

Anatomical landmark localization was first formulated as a RL task by Ghesu et al. [4]. In this approach, they utilized a Deep Q-Network (DQN) agent [10] to observe a sub-image and move with a fixed one-pixel step size on the four principal directions through the 2D image or on the six principal directions through the volumetric image. Since the agent is restricted to a discrete action space with a fixed step size, during inference the DQN approach needs a large number of steps before localizing the target landmark. Ghesu et al. [5] tackle this problem with a multi-scale framework to cover a larger field of view (FOV) and accordingly take large action steps. Their implementation, however, uses a separate neural network for each scale. Alansary et al. [1], similarly, use a multi-scale approach, where a single neural network is used for all scales to reduce training time. To localize multiple landmarks simultaneously, the same group extended their work by sharing the weights between convolution layers of multiple DQN agents [12] and by additionally combining the extracted information in the fully-connected layers before generating the actions [6]. A mutual challenge of all DQN-based approaches for landmark localization is the identification of the optimal stopping criterion of the agent. For that purpose, Maicas et al. [9] proposes an additional trigger action, which increases the action space of the DQN agent and consequently the complexity of the approach. A better-accepted approach is proposed in [4] where oscillation within a local neighborhood is used as an indicator for termination, however, this limits the accuracy of the method and prolongs inference time.

To overcome the above-mentioned limitations of DQN-based approaches, in this work, we propose a continuous action space for localizing anatomical landmarks. By allowing the agent to move in an arbitrary direction with variable step size, we reduce the number of steps the agent needs to localize the target landmark, which effectively speeds up inference time while improving the accuracy. To implement a continuous action space, we utilize an actor-critic approach proposed in [8]. In our setup, the problem of the stopping criterion is intrinsically addressed, since the agent stops moving when the movement displacement falls below the pixel size. Furthermore, inspired by multi-agent RL [2], we extend

our single-landmark/single-agent approach by introducing multiple agents to localize multiple landmarks simultaneously. Finally, we introduce communication between agents by providing every agent with its relative position to the other agents. This allows learning of a spatial configuration between landmarks, which serves as regularization and provides an estimate of landmarks' position even in the case where landmarks are missing or occluded.

## 2 Method

Anatomical landmark localization can be formulated as a Markov Decision Process (MDP) by defining an environment, states, actions and a reward function. The environment is the medical image  $I$  in which the agent navigates to localize the target landmark. The position in the environment is the state  $\mathbf{s} \in \mathbb{R}^D$  of the agent, where  $D$  is the number of image dimensions. The environment observed by the agent at the state  $\mathbf{s}$  is the observation  $o(\mathbf{s}) \subset I$ , which is restricted to a local image patch around  $\mathbf{s}$ . To allow the agent to move in arbitrary directions with a varying step size, we defined a continuous action space similarly as in [8]. We represent the action  $\mathbf{a}$  as a vector with  $\mathbf{a} = [a_1, \dots, a_D]^T \in \mathbb{R}^D$ . Which action the agent takes after observing the state  $\mathbf{s}$  is defined by the policy  $\pi$ . The agent's state after taking an action is obtained by adding the action vector  $\mathbf{a}$  to the current state  $\mathbf{s}$ , i.e.  $\mathbf{s}' = \mathbf{s} + \mathbf{a}$ . The reward function  $r$  for taking action  $\mathbf{a}$  is defined as:

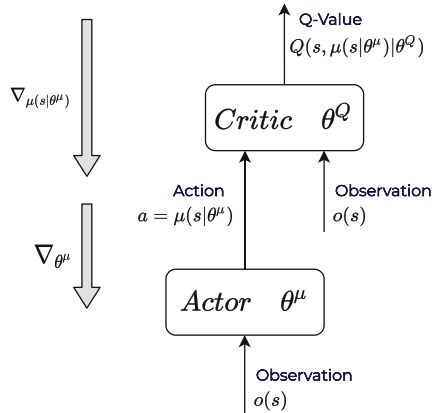
$$r = \|\mathbf{s} - \mathbf{g}\| - \|\mathbf{s}' - \mathbf{g}\|, \quad (1)$$

where  $\mathbf{g}$  is the position of the target landmark and  $\|\cdot\|$  is the Euclidean distance.

The MDP can be solved by sampling experience tuples of the form  $\langle \mathbf{s}, \mathbf{a}, r, \mathbf{s}' \rangle$  to determine the Q-function which can be written recursively as a Bellman equation:

$$Q_{t+1}(\mathbf{s}, \mathbf{a}) = \mathbb{E}[r + \max_{\mathbf{a}'} Q_t(\mathbf{s}', \mathbf{a}')] \quad (2)$$

with recursive step  $t$ . Differently from the DQN approach [10], where a single network can be used to model the Q-function due to the discrete action space, we use an actor-critic architecture as in [8] to allow continuous actions, see Fig. 1.



**Fig. 1.** Basic principle of the actor-critic architecture which allows continuous actions [8]. The actor network predicts an action and the critic network predicts a Q-value to evaluate the action.

Thus, a critic network parameterized with  $\theta^Q$  is modeling the Q-function, while a second, actor network  $\mu$  parameterized with  $\theta^\mu$  is used to learn the policy  $\pi$ . During inference only the actor network is used to generate the new position of the agent.

To improve the stability of the training we use *soft updates* for both, critic  $Q$  and actor  $\mu$  network as in [8]. Thus, we update the parameters of the target network  $\theta_T^{\{Q,\mu\}}$  with parameters of the current network  $\theta_C^{\{Q,\mu\}}$ . An alternating procedure is used to optimize the parameters of both critic  $\theta^Q$  and actor network  $\theta^\mu$ . By keeping the parameters of the actor network fixed, we optimize the critic parameters  $\theta^Q$  using the Bellman loss:

$$\arg \max_{\theta_C^Q} \frac{1}{N} \sum_i^N \left[ r_i + \gamma Q(\mathbf{s}'_i, \mu(\mathbf{s}'_i | \theta_T^\mu) | \theta_T^Q) - Q(\mathbf{s}_i, \mu(\mathbf{s}_i | \theta_T^\mu) | \theta_C^Q) \right]^2, \quad (3)$$

where  $N$  is the size of the mini-batch and  $\gamma \in [0, 1]$  is the discount factor used to weigh future rewards; the parameters  $\theta_T^Q$  and  $\theta_C^Q$  refer to the target and current network parameters of the critic respectively.

To optimize the parameters  $\theta^\mu$  of the actor network, we keep the parameters of critic  $\theta^Q$  fixed and maximize the expected Q-value by using the chain rule to compute the gradient:

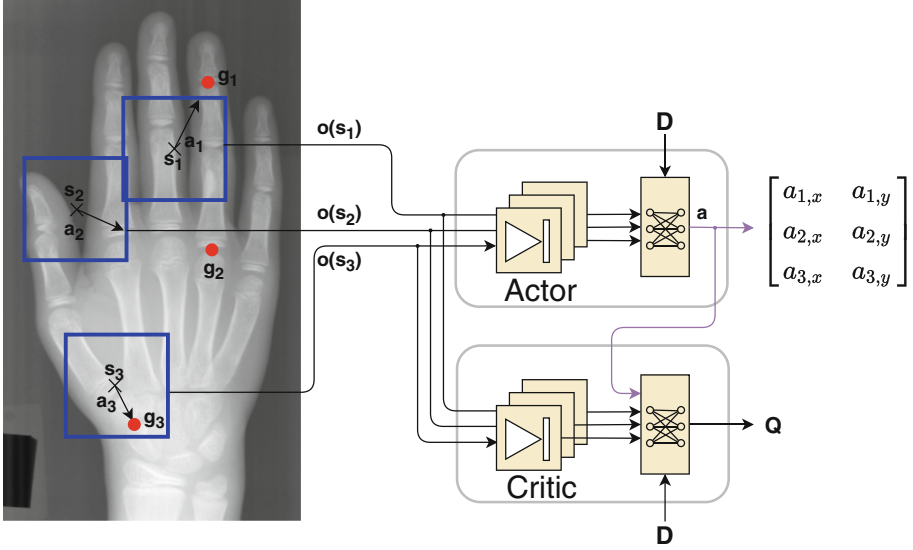
$$\nabla_{\theta_C^\mu} Q(\theta_C^Q, \theta_C^\mu) \approx \frac{1}{N} \sum_i^N \left[ \nabla_{\mu(\mathbf{s}_i | \theta^\mu)} Q(\mathbf{s}_i, \mu(\mathbf{s}_i) | \theta_C^Q) \nabla_{\theta_C^\mu} \mu(\mathbf{s}_i | \theta_C^\mu) \right]. \quad (4)$$

Differently from DQN [10], where the stopping criterion is usually determined by state oscillation, in the proposed approach with continuous action space we can define the stopping criterion as an action with an absolute length below one pixel.

So far we introduced an agent that performs a single action and therefore can only detect one landmark at a time. To simultaneously localize multiple landmarks in an image during inference, multiple agents have to be trained individually, each one trained for a different landmark. However, in such an approach no communication among agents exists. In this work we propose an approach for multi landmark localization inspired by a collaborative multi-agent system. Thus, the input to the actor network are now multiple observations, each corresponding to different agents, see Fig. 2. The actor network independently predicts the next action for each agent. Together with the observation of each agent, these actions are used as input to the critic network to predict a single Q-value optimized using Eq. 3. To allow collaboration between multiple agents and thus, to learn the spatial configuration of the landmarks, the position of each agent relative to each other is additionally provided to the actor and critic network. To this end, we define a list of pairwise offsets  $\mathbf{D}$  in the following form:

$$\mathbf{D} = \{\mathbf{s}^i - \mathbf{s}^j | i, j \in K, i \neq j\}, \quad (5)$$

where  $K$  is the number of agents.



**Fig. 2.** Schematic representation of our proposed method for landmark localization. One observation per agent is used as input to the actor network which predicts the next actions. To allow collaboration, we additionally provide the list of pairwise offsets  $\mathbf{D}$  (Eq. 5) to the network to enable learning of the spatial configuration of the anatomical landmarks. During training, an additional critic network is used which approximates the Q-function.

### 3 Experimental Setup

**Dataset.** We used a publicly available dataset of hand radiographs [3] acquired from different X-ray scanners to compare our method to both state-of-the-art RL and supervised learning based approaches for anatomical landmark localization. The dataset consists of 895 images with an average size of  $156 \times 2169$  pixels. Since the images do not contain information about the physical resolution, we follow [11] and assume a wrist width of 50 mm from which we calculate a physical resolution for each image. We downsample all images to a common long-axis size of 512 pixels and split the dataset with the ratio 80:20 into 716 images for training and 179 images for testing. Due to the long training time needed for RL based approaches, we used five representative landmarks from the 37 landmarks provided by the authors of [11].

**Implementation Details.** In our single-agent approach, the actor network consists of three consecutive convolution-pooling-convolution blocks followed by three fully-connected layers, after which a final fully-connected layer yields the network output. All convolution layers use an isotropic kernel size of 3 and ReLU activation, the number of filters of the first convolution layer is 32 and is doubled after every pooling layer. We employ average pooling with an isotropic kernel

size of 2 and zero padding. In the first three fully-connected layers we utilize 256 output neurons and ReLU activation, while the final fully-connected layer uses no activation function and directly outputs the predicted action. The actor and critic network are identical with following exceptions: the actor’s prediction is provided as an additional input at the first fully-connected layer of the critic network, and the critic network predicts a single Q-value. This approach we named Single-Agent Landmark Localization (SALL).

In the multi-agent approach, the observation of the agent is processed individually in a unique convolution path, resulting in one parallel convolution path per agent. These paths are concatenated before the first fully-connected layer to generate the action of all agents simultaneously. The list of pairwise offsets  $\mathbf{D}$  is provided as input to the first fully-connected layer as shown in Fig. 2. Same as with single-agent approach, the critic network of the multi-agent approach outputs a single Q-value. We named our multi-agent approach MALL. Additionally, we evaluated our multi-agent approach without collaboration between agents by omitting the pairwise offsets  $\mathbf{D}$  from the input to the actor and critic network. We named this approach in our experiments MALL<sub>noSC</sub> where SC stands for spatial configuration.

During training, the agent is initialized at a random position within the image. The agent progresses to the next image, if the distance between target and the agent’s current position is below one pixel or if the maximum number of steps is reached, i.e. 300 for DQN and 100 otherwise, leaving enough space for the agent to explore the entire state space. Furthermore, we limit the maximum step size per action to a distance of 50 pixels in each direction and we round the agent’s position to the position of the closest pixel. If the agent overshoots the image bounds, it’s position is moved to the closest position at the image border, out of bounds pixels of the observation are set to zero. Similarly to Lillicrap et al. [8], our actor network receives exploration noise from a Gaussian distribution during training. As hyperparameters we used  $\gamma = 0.85$ , a replay memory size of  $10^5$ , an exploration noise with a variance of 0.15, a soft update ratio of 0.125 and Adam optimizer with a learning rate of  $10^{-5}$  and  $10^{-3}$  for the actor and critic respectively. We trained for 30k episodes, the training time for the single-agent approaches was around three days and for the multi-agent approaches around 10 days on a workstation with Nvidia<sup>®</sup> Titan V GPU.

**Evaluation.** We divide our experiments into single-landmark localization experiments, where we train five SALL networks independently each predicting a different landmark, and multi-landmark localization experiments, where we train MALL as well as MALL<sub>noSC</sub> network to predict five landmarks simultaneously. For comparison, we use our implementation of DQN [4] and the original code of a state-of-the-art supervised learning based approach using heatmap regression, Spatial Configuration-Net (SCN) [11]. To ensure deterministic results and a fair comparison for all experiments, we use the center of the image as the agent’s initial position during inference. We compute the point error (PE) as the Euclidean distance between the agent’s final position and target landmark

in mm to evaluate the prediction accuracy. We present the average PE for all validation samples per landmark as well as it's overall average. Furthermore, we also determine the average number of steps the agent needs to localize the target landmark.

## 4 Results

The accuracy of the evaluated RL and supervised learning based approaches for landmark localization are presented in Table 1, separately for each landmark as well as all landmarks combined. In the same table we also show the average number of steps needed to terminate the RL based methods. The cumulative error distribution for all evaluated methods is shown in Fig. 3 again for each landmark separately as well as all landmarks combined. In Fig. 4 we show the results as error vectors drawn relative to the groundtruth landmark position of the respective image for all evaluated methods. In the same figure, we also show the results of the evaluated methods when the image is partially occluded starting from the landmark positioned between the metacarpal and phalanges bones simulated by uniform noise.

## 5 Discussion

In this work, we proposed a RL method for anatomical landmark localization using a continuous action space that allows the agent to move in an arbitrary direction with variable step size. This is different to existing state-of-the-art RL methods that use a discrete action space and a fixed step size, leading to a large number of steps and consequently long inference time to localize a landmark. As shown in Table 1, DQN [4], a state-of-the-art RL method for landmark localization, needs in average 193 forward passes of observation patches extracted from the inference image to localize a landmark. In contrast to that, our SALL method requires in average only 6.2 passes to reach the landmark. Hence, our experiments have shown that utilization of a continuous action space decreases the inference time by more than 30 times, which can be of high importance in e.g. time critical or energy efficient applications.

Our method has also shown to be more accurate compared to DQN, see Table 1 and Fig. 3, 4. While the average PE of the DQN method is  $1.19 \pm 0.9$  mm our method is able to localize landmarks with an average PE of  $0.86 \pm 0.74$  mm. This trend can be seen for all landmarks individually, while the largest difference can be observed for landmark 0, where our SALL method is 0.5 mm more accurate than DQN. One of the reasons why DQN is limited in accuracy is the stopping criterion, which is usually defined by oscillation of the agent and leads to ambiguity between the oscillating locations. This ambiguity is intrinsically resolved by our method, since the stopping criterion is predefined by a minimal displacement of the agent's position.

In comparison to the state-of-the-art SCN [11] method based on supervised learning and heatmap regression, our SALL method has shown inline results, see

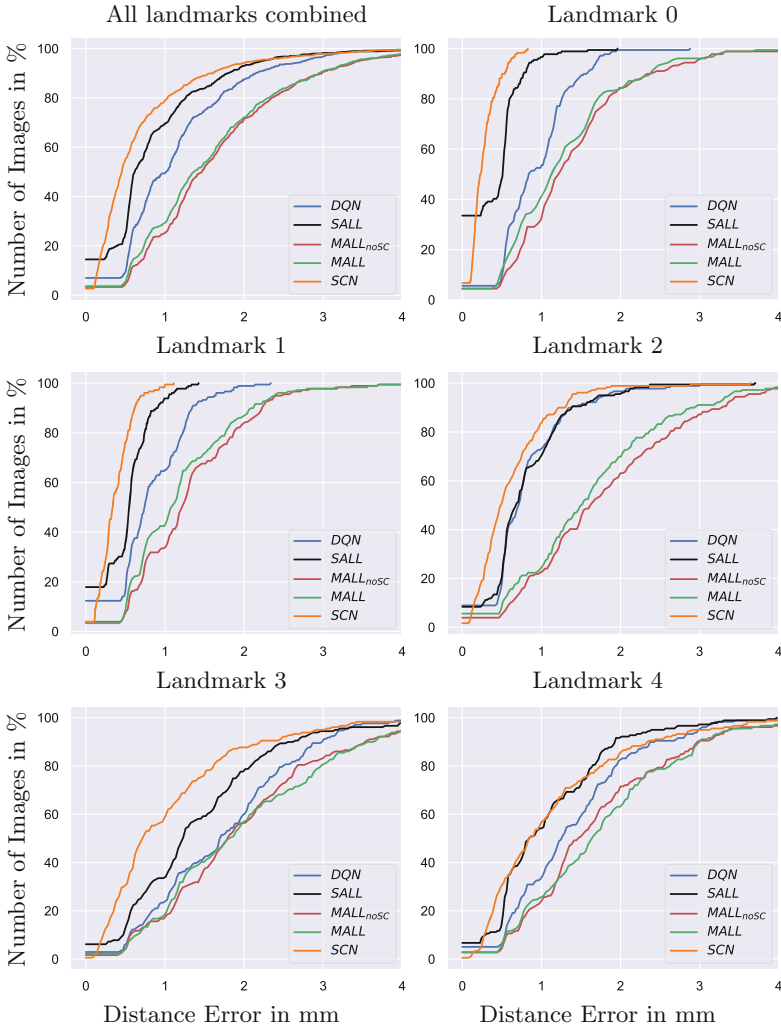
**Table 1.** Average point error (PE) of all evaluated approaches for landmark localization in mm, separately for each landmark (LM) and all landmarks combined (All), as well as the average number of steps required to terminate the RL based approaches during inference.

Algorithm		LM 0	LM 1	LM 2	LM 3	LM 4	All
SCN [11]	PE	<b>0.27 ± 0.17</b>	<b>0.37 ± 0.21</b>	<b>0.64 ± 0.49</b>	<b>1.09 ± 0.93</b>	1.15 ± 0.9	<b>0.7 ± 0.73</b>
	#steps	-	-	-	-	-	-
DQN [4]	PE	0.94 ± 0.46	0.87 ± 0.81	0.83 ± 0.54	1.9 ± 1.16	1.42 ± 0.84	1.19 ± 0.9
	#steps	253.6	85.8	177.5	219.1	229.3	193.0
SALL	PE	0.4 ± 0.35	0.52 ± 0.32	0.82 ± 0.53	1.44 ± 0.97	<b>1.09 ± 0.75</b>	0.86 ± 0.74
	#steps	<b>5.8</b>	<b>4.9</b>	<b>5.9</b>	<b>7.0</b>	<b>7.3</b>	<b>6.2</b>
MALL <sub>noSC</sub>	PE	1.41 ± 0.88	1.32 ± 0.71	1.82 ± 1.03	2.02 ± 1.14	1.71 ± 0.97	1.66 ± 0.99
	#steps	31.8	31.8	31.8	31.8	31.8	31.8
MALL	PE	1.35 ± 1.03	1.22 ± 0.72	1.63 ± 0.94	2.04 ± 1.2	1.78 ± 1.03	1.6 ± 1.04
	#steps	27.7	27.7	27.7	27.7	27.7	27.7

Table 1 and Fig. 3, 4. Although SALL achieves a better performance on landmark 4, a possible reason why SALL did not outperform SCN might be due to defining the agent’s position on a pixel level, which can be improved by allowing subpixel predictions. Furthermore, to achieve a high accuracy, a heatmap-based CNN method like SCN has to store the intermediate network outputs of the whole image in GPU memory which is a challenge, especially when working with large 3D volumes. Since our RL method processes local image patches extracted around the agent’s position, we are expecting similar performance for both 2D and 3D landmark localization tasks.

In this work, we additionally extend our single-agent approach (SALL) to a multi-agent approach (MALL) that is capable to simultaneously localize multiple landmarks. Differently to the recent work [6, 12], where the weights are shared between convolution layers of multiple DQN agents, our method has a separate convolutional path per agent before generating the action of every agent using fully-connected layers. Furthermore, in our method, we establish direct communication between agents by providing each agent with it’s relative position to all other agents. This communication between agents allows learning of the spatial configuration of anatomical landmarks, which is common in medical applications. Our experiments with partly occluded images (Fig. 4, right) show that both, heatmap based SCN [11] as well as the RL based DQN [4] are not able to localize the occluded landmark. The same behaviour is also shown by our single-agent RL method (SALL). It is interesting to see, that our multi-agent RL approach without list of pairwise offsets  $\mathbf{D}$  (MALL<sub>noSC</sub>) is failing to localize not only occluded but also visible landmarks when a large part of the image is replaced by random noise. A reason could be that the information extracted from each agent’s observation is combined in the actor before an action of each agent is generated. Thus, corrupted observations of individual agents strongly affect the performance of all other agents due to common fully-connected layers in the actor. Nevertheless, we would also expect a similar behaviour from other approaches that utilize a single network to generate the actions of multiple agents

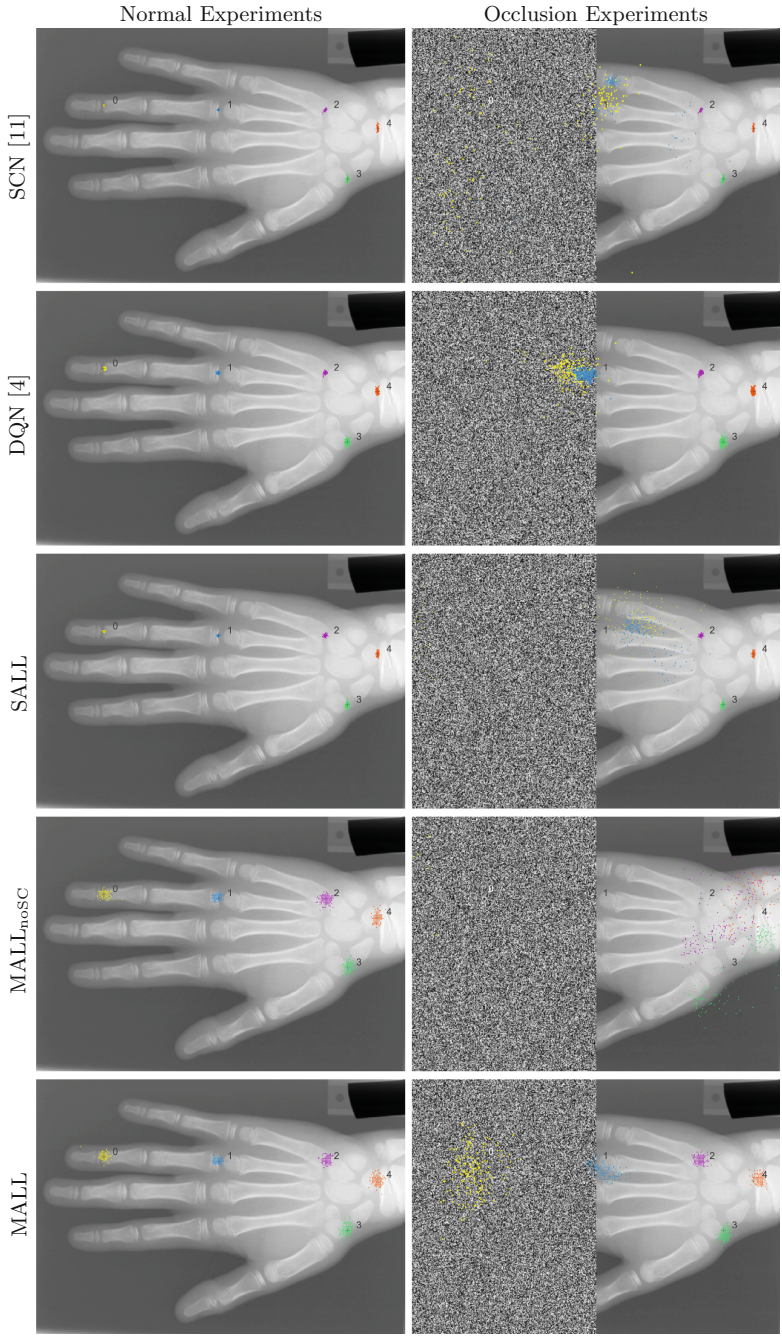




**Fig. 3.** Cumulative error distribution for all evaluated methods shown for all landmarks combined as well as for each landmark separately.

like [6]. Differently, when a list of pairwise offsets is provided to our multi-agent RL approach (MALL), it is able to not only localize the visible landmarks but also to use the information on their position to estimate the relative position of the occluded landmarks, see Fig. 4. Thus, our RL approach is able to successfully integrate the spatial configuration of the anatomical landmarks without the need of an additional model for global regularization, like statistical shape models or graphical models.

Additionally to evaluation of the proposed approach on volumetric images, in our future work we are planning to further investigate our multi-agent network



**Fig. 4.** Prediction results of all test images drawn on a single representative image as error vectors relative to the groundtruth landmark position of the respective image for the normal experiments (left) and the occlusion experiments (right). Each row corresponds to the method noted on the left.

architecture to improve the accuracy of the method. Namely, the average PE of MALL ( $1.6 \pm 1.04$  mm) is larger compared to SALL ( $0.86 \pm 0.74$  mm), which can be explained by an increased complexity of the MALL network compared to the SALL network. To improve the accuracy of MALL, the number of episodes can be increased, however, we trained both methods for the same number of episodes due to the long training time of MALL. A possible improvement to the MALL architecture is to use shared weights in each agent's convolutional path similarly to [6, 12], which would reduce the number of parameters and consequently also the training time.

## 6 Conclusion

In conclusion, our proposed RL based approach allows the agent to move in arbitrary directions with a variable step size to localize an anatomical landmark in medical images. Our results show, that the proposed continuous action space reduces the number of steps necessary to localize the landmark by more than 30 times in average compared to a state-of-the-art RL approach based on discrete actions. This consequently decreases the number of forward passes needed to localize the landmark, which can be of high importance in time critical or energy efficient applications. Moreover, compared to methods using a fixed step size, where the stopping criterion is often defined by oscillation of the agent's position and thus, limiting the accuracy of the method and prolonging inference time, the stopping criterion in the continuous action space is intrinsically defined by a minimal displacement of the agent's position. Furthermore, the movement with a variable step size is also more similar to how a physician advances through a medical image, since the proposed agent is able to adapt the step size depending on the distance from the anatomy of interest.

Our single-agent RL based method has shown a higher accuracy than DQN, a state-of-the-art RL approach for landmark localization. Compared to the state-of-the-art supervised learning based SCN approach, our single-agent RL approach achieved inline results. However, in contrast to SCN, our RL approach only requires patches and not the whole image as input, which can be beneficial when working with large volumetric images. In our extension to our multi-agent RL based approach, we also introduced communication among agents by providing each agent with it's relative position to the other agents which allowed learning of the spatial configuration of the landmarks. Thus, the results of our multi-agent RL based approach show that the position of the occluded landmarks can be successfully estimated based on the relative position predicted for the visible landmarks.

## References

1. Alansary, A., et al.: Evaluating reinforcement learning agents for anatomical landmark detection. *Med. Image Anal.* **53**, 156–164 (2019)

2. Foerster, J., Assael, I.A., de Freitas, N., Whiteson, S.: Learning to communicate with deep multi-agent reinforcement learning. *Adv. Neural Inf. Process. Syst.* **29**, 1–9 (2016)
3. Gertych, A., Zhang, A., Sayre, J., Pospiech-Kurkowska, S., Huang, H.: Bone age assessment of children using a digital hand atlas. *Comput. Med. Imaging Grap.* **31**(4–5), 322–331 (2007)
4. Ghesu, F.C., Georgescu, B., Mansi, T., Neumann, D., Hornegger, J., Comaniciu, D.: An artificial agent for anatomical landmark detection in medical images. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) *MICCAI 2016*. LNCS, vol. 9902, pp. 229–237. Springer, Cham (2016). [https://doi.org/10.1007/978-3-319-46726-9\\_27](https://doi.org/10.1007/978-3-319-46726-9_27)
5. Ghesu, F.C., et al.: Multi-scale deep reinforcement learning for real-time 3D-landmark detection in CT scans. *IEEE Trans. Pattern Anal. Mach. Intell.* **41**(1), 176–189 (2017)
6. Leroy, G., Rueckert, D., Alansary, A.: Communicative reinforcement learning agents for landmark detection in brain images. In: Kia, S.M., et al. (eds.) *MLCN/RNO-AI -2020*. LNCS, vol. 12449, pp. 177–186. Springer, Cham (2020). [https://doi.org/10.1007/978-3-030-66843-3\\_18](https://doi.org/10.1007/978-3-030-66843-3_18)
7. Li, J., Wang, Y., Mao, J., Li, G., Ma, R.: End-to-end coordinate regression model with attention-guided mechanism for landmark localization in 3D medical images. In: Liu, M., Yan, P., Lian, C., Cao, X. (eds.) *MLMI 2020*. LNCS, vol. 12436, pp. 624–633. Springer, Cham (2020). [https://doi.org/10.1007/978-3-030-59861-7\\_63](https://doi.org/10.1007/978-3-030-59861-7_63)
8. Lillicrap, T.P., et al.: Continuous control with deep reinforcement learning. In: *International Conference on Learning Representations* (2016)
9. Maicas, G., Carneiro, G., Bradley, A.P., Nascimento, J.C., Reid, I.: Deep reinforcement learning for active breast lesion detection from DCE-MRI. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) *MICCAI 2017*. LNCS, vol. 10435, pp. 665–673. Springer, Cham (2017). [https://doi.org/10.1007/978-3-319-66179-7\\_76](https://doi.org/10.1007/978-3-319-66179-7_76)
10. Mnih, V., et al.: Human-level control through deep reinforcement learning. *Nature* **518**(7540), 529–533 (2015)
11. Payer, C., Štern, D., Bischof, H., Urschler, M.: Regressing heatmaps for multiple landmark localization using CNNs. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 230–238 (2016)
12. Vlontzos, A., Alansary, A., Kamnitsas, K., Rueckert, D., Kainz, B.: Multiple landmark detection using multi-agent reinforcement learning. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 262–270 (2019)