Medical Image Analysis 89 (2023) 102875

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Multi-target landmark detection with incomplete images via reinforcement learning and shape prior embedding

Kaiwen Wan [a], Lei Li [b], Dengqiang Jia [c,d], Shangqi Gao [a], Wei Qian [e], Yingzhi Wu [f], Huandong Lin [g], Xiongzheng Mu [f], Xin Gao [g], Sijia Wang [e], Fuping Wu [h], Xiahai Zhuang [a,*]

[a] School of Data Science, Fudan University, Shanghai, 200433, China
[b] Institute of Biomedical Engineering, University of Oxford, Oxford, UK
[c] School of Naval Architecture, Ocean and Civil Engineering, Shanghai Jiao Tong University, Shanghai, China
[d] Hong Kong Centre for Cerebro-Cardiovascular Health Engineering (COCHE), Hong Kong, China
[e] Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai 200031, China
[f] Department of Plastic Surgery, Huashan Hospital, Fudan University, Shanghai 200040, China
[g] Department of Endocrinology and Metabolism, Zhong Shan Hospital, Fudan University, 200032 Shanghai, China
[h] Nuffield Department of Population Health, University of Oxford, Oxford, UK

ARTICLE INFO

Keywords:
Incomplete image
Shape prior
Landmark detection
Multi-agent reinforcement learning

ABSTRACT

Medical images are generally acquired with limited field-of-view (FOV), which could lead to *incomplete* regions of interest (ROI), and thus impose a great challenge on medical image analysis. This is particularly evident for the learning-based multi-target landmark detection, where algorithms could be misleading to learn primarily the variation of background due to the varying FOV, failing the detection of targets. Based on learning a navigation policy, instead of predicting targets directly, reinforcement learning (RL)-based methods have the potential to tackle this challenge in an efficient manner. Inspired by this, in this work we propose a multi-agent RL framework for simultaneous multi-target landmark detection. This framework is aimed to learn from incomplete or (and) complete images to form an implicit knowledge of global structure, which is consolidated during the training stage for the detection of targets from either complete or incomplete test images. To further explicitly exploit the global structural information from incomplete images, we propose to embed a shape model into the RL process. With this prior knowledge, the proposed RL model can not only localize dozens of targets simultaneously, but also work effectively and robustly in the presence of incomplete images. We validated the applicability and efficacy of the proposed method on various multi-target detection tasks with incomplete images from practical clinics, using body dual-energy X-ray absorptiometry (DXA), cardiac MRI and head CT datasets. Results showed that our method could predict whole set of landmarks with incomplete training images up to 80% missing proportion (average distance error 2.29 cm on body DXA), and could detect unseen landmarks in regions with missing image information outside FOV of target images (average distance error 6.84 mm on 3D half-head CT). Our code will be released via https://zmiclab.github.io/projects.html.

1. Introduction

Landmark detection is an essential step for biomedical image analysis (Zhan et al., 2016). Anatomical landmarks can describe morphological characteristics of anatomical structures, which is important in morphometric and medical analysis. Their locations can also be useful in many downstream computing tasks, such as image segmentation and registration (Seghers et al., 2007). However, in practical clinics, medical images usually have varying field-of-view (FOV) and cover incomplete region-of-interests (ROI). Such incomplete images lead to particular challenges for learning-based landmark detection.

We refer an image as *incomplete image* when it cannot cover the whole ROI containing all targets, or when it, as a training image, does not provide gold standard labels of all targets. Fig. 1(a) illustrates the saggital view of a complete CT image which covers all the anatomical landmarks of the head, and three incomplete images which only cover parts of the head. This incompleteness of medical images is common in clinical practice. For example, many of the CT images are acquired without a whole-head scan in clinic to reduce exposure to radiation, such as mandible modeling (Park et al., 2022), nasal cancer study (Li et al., 2019) and stroke diagnosis (Hakim et al., 2021).

* Corresponding author.
E-mail address: zhuangdata@aliyun.com (X. Zhuang).
URL: http://www.sdspeople.fudan.edu.cn/zhuangxiahai/ (X. Zhuang).

https://doi.org/10.1016/j.media.2023.102875
Received 10 January 2023; Received in revised form 5 May 2023; Accepted 13 June 2023
Available online 24 June 2023
1361-8415/© 2023 Elsevier B.V. All rights reserved.

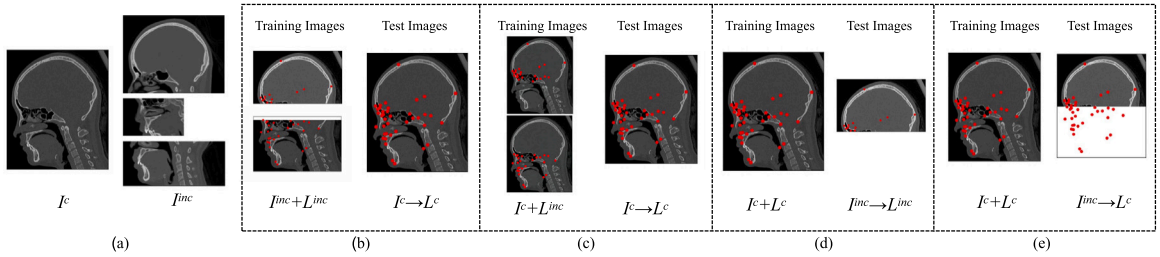

Fig. 1. Illustration of complex scenarios involving incomplete images for machine learning. Here, I^c (L^c) and I^{inc} (L^{inc}) respectively refer to complete images (labels) and incomplete images (labels); the dots (•) represent landmarks. (a) a complete image and three incomplete images of head CT in saggital view; (b) a scenario where the machine learning algorithm is required to learn from incomplete images (with incomplete gold standard labels) to infer all completes label inside of the complete test image ; (c) a scenario where the training images are complete but their gold standard labels are incomplete, and the target is to infer all complete label; (d) and (e) scenarios where training images are complete in both images and labels, but the test images are incomplete. Differently, in (d) the target is to infer only incomplete label, *i.e.,* landmarks, from the incomplete image given the available ROI, while in (e) the task is to predict all landmarks even for the ones not contained in the incomplete image. Note that images in this figure are all 2D saggital view of 3D images, and the landmarks on them are mapped from 3D space.

These images become incomplete when the studies involve the target landmarks of the whole head, namely a complete image in one study can become incomplete in another for wider ROI. This is even more common in dual-energy X-ray absorptiometry (DXA) scans, which are widely used to analyze *different* areas of the body. These DXA images from different studies are incomplete for the study whose targets involve the whole body (Briot, 2013). Finally, partial labeling is commonly seen in big medical image training data, since manually labeling is tedious and difficult even for experienced annotators. Moreover, due to the difference of targets in different studies, the collected label can be partial even for a complete image (Fries et al., 2019).

Fig. 1 summarizes and illustrates the common scenarios involving incomplete images in multi-target landmark detection. Particularly, Fig. 1(b) shows the most common scenario, where the training images contain only limited and incomplete ROI, consequently their labels are incomplete; but the target is to detect all landmarks given a complete image. This is common yet challenging, and the core is to learn the global structural information (also known as shape information) from limited FOV images, with an effective scheme to consolidate the global knowledge for inference. Fig. 1(c) represents a similar yet different application, where images are all complete, in both of the training and test stages; but the gold standard labels for training are incomplete. This can happen in multi-center studies, where different center has interests of different targets. The ultimate goal of machine learning here is to learn all the knowledge of local annotations, and assemble them into a global model for inference. Fig. 1(d) and (e) demonstrate the more practical situations, where our training images are perfectly complete. However, in clinic practice the acquired images can be of limited FOV and incomplete. It is therefore desirable that the artificial intelligent models can be robust in the presence of such incomplete test images. Furthermore, in certain extreme cases the missing part of the incomplete test images can be so important that we need to recover these missing landmarks, as Fig. 1(e) shows, the model is then expected to predict the targets even though they are not within the FOV of the test images.

The aforementioned scenarios can challenge the learning-based multi-target landmark detection algorithms. With incomplete images, the global structural information is not consistently presented, thus the algorithms may not be able to assemble the partial knowledge into global one. In addition, one algorithm developed for coping with incomplete training images may not be applicable to the scenario of incomplete test images, or vice versa. For deep learning-based methods, many models even require fixed sizes of input and output images, thus they cannot be directly extended for such incomplete images with random FOVs and sizes.

Reinforcement learning (RL), using an intelligent agent to study interacting with changing environments, has the potential of tackling the challenges (Kaelbling et al., 1996). Different from the conventional manner of learning and predicting targets directly, RL tries to learn a *navigation policy* with an artificial agent in landmark detection. Such agent can be trained to search for optimal trajectories to target landmark from any random initial position of the image. This patch-based strategy captures local image information more easily than the holistic image information. In the presence of incomplete images, RL-based algorithms based on such local information can robustly avoid disturbance caused by missing structures. Hence, RL has the potential of offering a solution for the learning and inference with incomplete images, thanks to the navigation strategy for identifying the global structure of incomplete information.

For incomplete images, this work further aims to develop a robust opportunity RL algorithm to detect multiple targets simultaneously. Potentially, RL with multiple agents can search for multiple targets and accomplish the mission simultaneously, with each agent targeting a single landmark. However, since each agent is navigating separately, this strategy may not fully utilize the global structural information, leading to additional challenges in scenarios of incomplete images. Embedding prior shape information into RL-based algorithms can be a solution. For example, Zhang et al. (2020) tried to add reward based on shape information to share them between different agents. Ghesu et al. (2018) used extra statistical shape modeling and robust estimation theory containing shape information to correct the initial prediction of RL. Nevertheless, existing studies mainly focus on dealing one scenario at one time, and a unified framework applicable to these general problems of incomplete images is yet to be researched and developed. Different scenarios may present different challenges for different methodologies, as Fig. 1 shows.

In this work, we propose a two-stage deep neural network (DNN) model for multi-target landmark detection with incomplete images. This model combines a multi-agent RL network and a statistical shape prior embedded network. The former network, extended from the single-agent RL framework, is aimed to search for multiple targets simultaneously from incomplete images; the latter, to provide prior knowledge of global structure for the former, is developed based on a statistical shape model. The two networks are trained alternately, and the output of RL network can provide more samples for the shape network during training, which enhances the robustness of the landmark detection model. Similar alternation is also applied in the inference (test) stage.

The rest of this work is organized as follows: Section 2 presents the related work. We elaborate on the methodologies in details in Section 3. Section 4 provides the validation work, with comprehensive experiments and results via five practical applications from clinic perspectives and four datasets. Finally, we conclude and discuss in Section 5.

2. Related work

2.1. Landmark detection algorithms

Most conventional landmark detection methods process only complete images in different ways to exploit the structural knowledge.

Several works tried to build a global structure containing the explicit relationship between landmarks. For example, Cootes et al. (1995) proposed a statistical shape model (SSM) to fit the object in target images, namely active shape model (ASM). ASM was later improved by adding global or local appearance information (Cootes et al., 2001; Cristinacce et al., 2006). Methods could also embed the shape information implicitly via regression or registration. For example, regression maps image appearance to landmark locations directly (Dantone et al., 2012; Yang and Patras, 2013). In contrast, registration builds a map between different images and it can propagate label information from atlas images to the target (Gass et al., 2014; Li et al., 2020). These traditional methods can obtain promising results, but are usually time expensive.

For landmark detection, there is a trend to shift from traditional methods to deep learning-based methods. Deep learning-based methods are much more efficient and can achieve results that may not be worse than traditional methods (Zhuang et al., 2019). Using powerful DNN to extract features, direct regression methods can predict the landmark locations in the target image in one single step (Sun et al., 2013; Zhou et al., 2013). A similar but different method technique is to introduce heatmap from convolution neural network to detect the landmarks. Instead of getting determined results, the methods output heatmap as the potential distribution of landmarks (Payer et al., 2016). Cascaded regression models were also proposed, which found coarse predictions at the beginning, and gradually updated localization in the following stages (Zheng et al., 2015; Lv et al., 2017). However, most deep learning-based methods struggle to handle high-dimensional image data (Kang et al., 2021) and may be sensitive to the bounding box of images that cover the ROI (Sagonas et al., 2016).

2.2. Usage of incomplete images

The use of incomplete images has been an ongoing research direction in recent years (Kerr et al., 2008). Traditional methods of processing incomplete images were usually based on image completion (Stasinski and Konrad, 2000; Vazquez et al., 2000, 2002), in which constructed images without prior knowledge of their contents and could lead to unreliable results. There are also traditional algorithms that can be applied directly to incomplete image data without using image reconstruction, such as using space variant operators to extract landmarks (Kerr et al., 2008). However, the incomplete images they studied were regularly or irregularly sampled from complete images, where missing pixels were distributed over the whole image. These images were different from the incomplete medical images discussed in this work, which were caused by limited ROI. Efforts have been devoted to deep learning-based methods for landmark detection of incomplete images. For instance, Wang et al. (2020) used a random mask-based data augmentation strategy to generate incomplete images for network training. Hanaoka et al. (2017) proposed a two-stage sampling algorithm that can be applied to incomplete CT images.

2.3. Statistical shape models for detecting landmarks and handling incomplete images

As mentioned in Section 2.1, SSM can be used for landmark detection. To get accurate shape priors, different methods have been proposed to construct shape representations, such as implicit level set representations (Cremers et al., 2007), Fourier surfaces (an extension of the classical Fourier transforms) (Staib and Duncan, 1996), and wavelet transform (Davatzikos et al., 2003). There were also efforts to combine the benefit to DNN with SSM. For example, Ma et al. (2018) leveraged the shape priors learned from SSM to improve the segmentation of DNN. Raju et al. (2022), Bhalodia et al. (2018) generated shape parameters of SSM with DNN directly.

Besides, SSM has also been employed to handle incomplete images. For example, Chintalapani et al. (2010) constructed SSM from complete CT images and inferred the missing information of incomplete images with statistical atlas. This method was improved by Grupp et al. (2015) with smoother extrapolation strategies. Several studies focus on constructing SSM directly from incomplete images. Meng et al. (2020) constructed SSM from partial and incomplete images with probabilistic principal component analysis. Jingting et al. (2016) proposed a framework for automatically building SSM from the incomplete shapes generated from a segmentation algorithm. However, the representation power of deep networks has not been explored yet when employing SSM to handle with incomplete images. To deal with this, we introduced DeepSSM for incomplete image problems.

2.4. Reinforcement learning for medical image analysis

As a well-studied topic of machine learning methods, RL has been applied to numerous tasks in medical image analysis (Zhou et al., 2021), including lesion detection (Maicas et al., 2017), registration (Liao et al., 2017), view plane localization (Alansary et al., 2018) and landmark detection (Alansary et al., 2019). Instead of locating the target directly, RL tried to learn the optimal path to the target from given position with an agent. Ghesu et al. (2016) adopted a deep RL agent to navigate in a 3D medical image for automatic landmark detection. This method was improved with a multi-scale strategy later (Ghesu et al., 2017). To detect multiple targets simultaneously, Vlontzos et al. (2019) extended the RL network to be able to train multiple agents at the same time. The structural relationship between landmarks has been considered in several methods for robust detection. For example, Leroy et al. (2020) proposed an RL scheme that used communicative agents to share information. Zhang et al. (2020) proposed an algorithm incorporating priors on physical structure, where graph communication layers and additional rewards were designed to exploit structural priors.

RL can detect absent anatomical landmarks and has also been used to handle incomplete images. Ghesu et al. (2016) concluded that RL can judge whether the anatomical landmark is absent in target image. Jain et al. (2020) proposed an RL-based multi-target landmark detection method, which could estimate primary target landmark locations even though the local images around targets are defaced. Ghesu et al. (2018) used a two-stage model, first training artificial agents to search for anatomical structures and then using elements of a statistical shape model to ensure the spatial coherence of the observed anatomical landmarks. However, most of these studies focus on prediction with incomplete images given complete images for training, which is solely one scenario of the incomplete image problems, as illustrated in Fig. 1(d). The complex nature and sophisticated scenarios involving incomplete images have not been fully explored yet. In contrast, this is the first study that aims to tackle this task, to the best of our knowledge.

3. Methodology

This work is aimed to develop an effective method for multi-target landmark detection in complex scenarios involving incomplete images. In practice, the images collected either for training or test could be presented in a nonstandard form. Specifically, they may not cover the whole ROI, so certain volumes containing target landmarks could be missing. These scenarios introduce the issue of unfixed number of landmarks, which challenges most of the deep learning-based approaches requiring fixed number of landmarks (see Table 1). Let $\{(I_i^c, X_i, \mathcal{J})|_{i=1,\dots,N}\}$ be a training set for a common task of landmark detection, where $\mathcal{J}$ is the index set of all target landmarks for a complete image such as I_i^c whose corresponding landmark set is X_i. By contrast, in the scenario of incomplete images a training set is denoted by $\{(I_i^{inc}, X_i, \mathcal{J}_i)|_{i=1,\dots,N}\}$, where I_i^{inc} represents an incomplete image. The set of available landmark index for I_i^{inc} is denoted as $\mathcal{J}_i$, which is a subset of $\mathcal{J}$.

With similar notation, let $\{I_i^{inc}|_{i=1,\dots,N'}\}$ be the test image set with incomplete images. A model, denoted as F, predicts the coordinates of all landmarks, *i.e.*, $\hat{X}_i = F(I_i^{inc}) \in \mathbb{R}^{|\mathcal{J}|\times N_{dim}}$, where N_{dim} is dimension

Table 1
Definition of symbols used in this paper.

Symbol	Definition	Symbol	Definition
I_i	An image with index i	$\mathcal{J}_i$	Available index set of I_i
I_i^c / I_i^{inc}	A complete/incomplete image with index i	S	Current state
I^c / I^{inc}	Scenarios involving complete/incomplete images	s_j	State of agent j
L^c / L^{inc}	Scenarios involving complete/incomplete label of images	R	Reward
L^{c-inc}	Scenarios involving complementary label of images	r_j	Reward got by agent j
N / N'	Number of training/test images	A	Current actions
N_{dim}	Dimension of the images	a_j	Action taken by agent j
$\mathbf{X} / \overline{X}$	Landmark set library/mean shape of $\mathbf{X}$	$M_{\text{direction}}$	Movement with given direction
$X_i / \hat{X}_i$	Landmark set of I_i/prediction of X_i	$\Gamma_j^{[t]}$	Patch centered on position of agent j at time t
$x_{ij} / \hat{x}_{ij}$	Landmark of X_i with index j/prediction of x_{ij}	ϵ	Greedy parameter
x_j^d	Ground-true location for agent j	D	Distance metric
$x_j^{[t]}$	The locations of agent j at time t	$\mathcal{J}$	Index set of all target landmarks

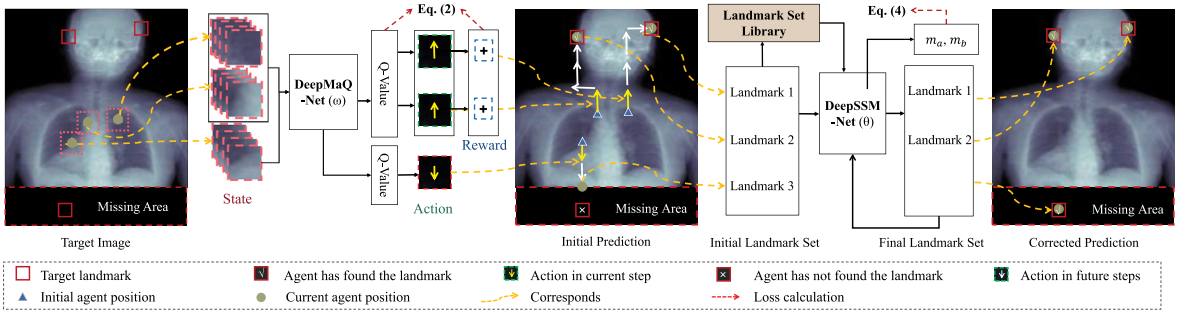

Fig. 2. Illustration of the shape-guided multi-agent reinforcement learning (SGMaRL) via DeepMaQ-Net and DeepSSM-Net. Here, we use three target landmarks for illustration, where one of them locates in the missing area. In the training stage, only agents whose targets exist in the image take actions and affect the calculation of loss function of DeepMaQ; while in the test stage, all agents navigate with shape guidance and regularization via DeepSSM-Net.

of the images. Intuitively, a workable model should have the following minimization problem:

$$\min \sum_{i=1}^{N'} \sum_{j \in \mathcal{J}_i} D(x_{ij}, \hat{x}_{ij}), \tag{1}$$

where $x_{ij} \in \mathbb{R}^{N_{dim}}$ and $\hat{x}_{ij} \in \mathbb{R}^{N_{dim}}$ are respectively the ground truth and prediction of coordinate for the landmark with index j in image I_i^{inc}; $D(\cdot, \cdot)$ is a distance metric, such as the Euclidean distance.

To fulfill the above objective, we propose a two-stage DNN based on shape-guided multi-agent reinforcement learning (SGMaRL). As illustrated in Fig. 2, this framework consists of two major components, *i.e.,* a deep multi-agent Q-learning (DeepMaQ) network for multi-target landmark detection, and a deep statistical shape model (DeepSSM) for shape guidance and regularization. We introduce these two algorithms respectively in Sections 3.1 and 3.2. Then we provide the implementation details in Section 3.3.

3.1. Deep multi-agent Q-learning

Due to static environments and discrete variables of action design (details are provided below), Deep Q-Network is capable for landmark detection on medical images (Yang and Xie, 2020), and Deep Q-Network with single agent (DeepSaQ) has been proved to be effective in incomplete images scenario (Ghesu et al., 2018). By learning a navigation policy with local patches as the input, DeepSaQ can detect one target and tackle the problem of missing structures elegantly. When multiple targets are presented, multiple DeepSaQs need to be trained and deployed, which can result in deterring training time and memory demand due to its inefficiency.

We therefore propose the DeepMaQ, *i.e.,* deep multi-agent Q-learning network, which can efficiently detect dozens of landmarks at same time in presence of incomplete images. For each target landmark, an AI learner, *i.e.,* an *agent*, is defined. The navigation policy can be optimized by defining a dynamics function of Markov decision process. For each frame, multi-series of local patches are constructed, as inputs, to guide the agents finding their paths to the corresponding target positions. All agents have different observations but navigate simultaneously. Additionally, each agent can be independent of others. This reduces the unnecessary dependency of the *effective agents*, when searching for their targets, on the *ineffective agents* whose corresponding targets are not within the image due to the incompleteness of the image.

In multi-target RL, we assign one agent for the detection of one landmark, thus the set of indices for agents is denoted as $\mathcal{J}$, the same as that of target landmarks. The learning capability and efficiency of a RL algorithm depend on the setting of state, action and reward of agents. In the following, we elaborate on their settings in details.

State: In our model, the current observations (states) of agents are represented by a series of cropped patches from an image. Current state, denoted as S, is then defined as the set of current states of all agents, $\{s_j|_{j \in \mathcal{J}}\}$, of which each s_j, associated with agent j, is a frame history buffer. We define this buffer using the concatenation of patches centered on the positions of agent j given the n time frames, *i.e.,* $s_j = \{\Gamma_j^{[t]}, \Gamma_j^{[t-1]}, \ldots, \Gamma_j^{[t-n+1]}\}$, where t and n represent current time and time lag term, respectively. The latest n cropped patches are used together to stabilize the search trajectories. Corresponding s_j is sliced from State S and then input to a DNN-based RL model (such as DeepMaQ), to decide which action should be taken for every agent j.

Action: An agent takes a series of actions to navigate to target position. In our model, current action is the concatenation of actions of all agents at current state, and is denoted as $A = \{a_j|_{j \in \mathcal{J}}\}$. For simplicity, here we consider four types of actions for 2D images, including move left, move right, move up and move down with a preset step length, thus the action of agent j from the action space is denoted as $a_j \in \{M_{\text{left}}, M_{\text{right}}, M_{\text{up}}, M_{\text{down}}\}$. Similarly for 3D images, we further allow the agent to move forward and backward, denoted respectively by $\{M_{\text{forward}}, M_{\text{back}}\}$. In the implementation, we adopt the multi-scale strategy coupling step lengths with scales, namely the agents move in larger step lengths in larger scale settings. The moving length will decrease when the agent terminates at certain scale. An agent receives feedback from the environment with an action, and the feedback is implemented via the reward mechanism.

Reward: At current state, an agent reaps a reward for taking an action, and different actions at different states of different agents receive different rewards. We denote our reward function of multi-agent RL as a concatenation of individual rewards from each agent, *i.e.,* $R = \{r_j|_{j\in\mathcal{J}}\}$, where $r_j = D(x_j^d, x_j^{[t-1]}) - D(x_j^d, x_j^{[t]})$, x_j^d is the ground-true location for agent j, $x_j^{[t-1]}$ and $x_j^{[t]}$ are the locations of agent j before and after taking an action, respectively, and D is Euclidean distance. Note that r_j is positive if the distance between location and ground truth becomes smaller and the unit is pixel. The reward function is the foundation of the loss function in RL.

The action-selection policy in the proposed DeepMaQ can be optimized by learning Q-value function. Here, we approximate a Q-value function with a Deep Q-Network, *i.e.,* the DeepMaQ-Net shown in Fig. 2. Hence, the outputs of DeepMaQ-Net are the estimated Q values and measure the benefits of taking different actions given current state for the agents. Based on the Bellman Equation (Alansary et al., 2019), we propose the following loss function for our multi-agent RL, *i.e.,* DeepMaQ,

$$\mathcal{L}_{MaQ}(\omega) = \sum_{i=1}^{N}\sum_{j\in\mathcal{J}_i}\mathbb{E}_{s_j,a_j,r_j,s'_j}\Big(r_j - Q_j(s_j,a_j;\omega) + \gamma\max_{a'_j}Q_j(s'_j,a'_j;\omega)\Big)^2, \tag{2}$$

where s_j, a_j, and r_j are current state, current action to take and immediate reward of taking this action for agent j, respectively; s'_j and a'_j are the new state and action to take after action a_j, respectively; $Q_j()$ represents the Q-value function; $\max_{a'_j}Q_j()$ denotes the maximal Q value for future actions to take, and γ is the discount factor.

3.1.1. DeepMaQ network

The DeepMaQ network is composed of four convolutional layers interleaved with three max-pool layers followed by four fully connected layers. The last layer was reshaped as a matrix according to the number of landmarks need to be detected. One can refer to Fig. 1 in supplementary material for details. The ith element in the jth column of output matrix represents the utility values of executing the ith action for agent j, which shall decide Q_j in Eq. (2). When agent j needs to take an action, the corresponding state column is extracted as an input to the DeepMaQ-Net, as shown in Fig. 2. One can see that when the number of target landmarks increases, we only need to increase the number of parameters in the last two fully connected layers. Hence, this architecture is suitable for various tasks of multi-target landmark detection.

3.2. Deep statistical shape model

Multi-agent RL can predict a set of landmarks which however may form an unrealistic shape of the object. This is attributed to the lack of global shape knowledge of the RL algorithm when no prior is provided. Hence, we introduce a statistical shape model (SSM) into the RL network, resulting in SGMaRL. To couple with the DeepMaQ, we further develop the DNN implemented SSM, referred to as the DeepSSM.

Given an image, the detected landmark set, denoted as X, is aligned to a pre-constructed SSM, as follows,

$$a^*,\ b^* = \arg\min_{a,\ b}\|X - T_a(\overline{X} + Pb)\|, \tag{3}$$

where $\overline{X}$ represents the mean shape of the landmark set library for constructing the SSM; T_a represents an affine transformation defining with vector a; P is the matrix formed from the largest k modes of the SSM, and vector $b \in \mathbb{R}^k$ refers to weights to be optimized. To find a^* and b^* which satisfy Eq. (3), conventional methods generally adopt an iterative algorithm, which nevertheless can be time-consuming. We therefore propose to train a DNN of SSM (DeepSSM), whose model is denoted as G, to infer a^* and b^* in an efficient manner. The loss function of DeepSSM is given by,

$$\mathcal{L}_{SSM}(\theta) = \sum_{i=1}^{N}\Big(X_i - T_{m_i^a}(\overline{X} + Pm_i^b)\Big)^2, \tag{4}$$

where $[m_i^a,\ m_i^b] = G(X_i;\theta)$ are output of the DeepSSM.

For the architecture, we adopt ResNet-18 (He et al., 2016) as the backbone of DeepSSM. The output vector of G is sliced into two sub-vectors, representing m^a and m^b, respectively. One can refer to Fig. 2 in supplementary material for details. For 2D images, m^a has 6 elements and m^b has k elements; while for 3D images, m^a has 12 elements and m^b also has k elements; k is the number of modes used in the SSM.

3.3. Implementation details

3.3.1. Training

There are two DNNs in SGMaRL, *i.e.,* DeepMaQ-Net and DeepSSM-Net, as Fig. 2 illustrates. The former is aimed to learn the navigation policy for every agent to find its target from a random starting position. The latter is trained to map and regularize a random shape to the closest one in the shape space defined by the SSM. The random shape here refers to the output of the DeepMaQ-Net and is defined by the set of detected landmarks, and the SSM is constructed using a pre-defined shape library. These two networks are trained alternately. Note that the outputs of half-way trained DeepMaQ-Net can provide useful samples for the training of DeepSSM-Net. Alg. 1 provides the pseudo code.

In the training of DeepMaQ-Net, which is a RL-based neural network, we assign random initial positions for agents at the beginning of the episode for a training image. For these agents whose target landmarks are within this image, we consider them as being *active* and allow them to take actions (navigate). By contrast, for those agents whose corresponding targets are in the missing areas, we consider them as being *passive* and keep them fixed in the initial positions without any navigation. Similarly, when an agent reaches its own terminal state, we then assign it as being passive and keep it in the terminal position with no more action during the rest navigation steps for other active agents in this episode. This lasts until the end of the episode.

In addition, we adopt three techniques to ensure a stable and efficient training of DeepMaQ-Net, including the ϵ-greedy (Alansary et al., 2019), replay-and-freeze, and multi-scale strategies (Zhuang and Shen, 2016). First, we propose to use ϵ-greedy strategy, where the agent selects an action uniformly at random with probability ϵ in each step. This is because when the model is insufficiently trained, the agent can be misled by taking the action corresponding to the maximal Q value. Here, ϵ will be set as a smaller value when DeepMaQ becomes convergence. Second, experience replay memory and freezing the target network can be used to minimize the effects of instability and divergence in deep Q-learning networks. The former (replay) avoids the problem of successive data sampling, and the latter (freezing) helps reducing rapid changes in Q values. Finally, the multi-scale strategy with multiple resolution of images is adopted to improve the capture range and convergence speed. Here, navigation step lengths are coupled with image resolution. Hence, navigation in course-scale images is with large physical step lengths and accelerates convergence towards the target plane; while navigation in fine-scale images steps in small length and fines tune the final estimation of plane parameter. The combination of multi-scale images contributes to better capture range and convergence speed.

Before training DeepSSM-Net, we first employ a set of subjects represented by complete landmark sets to construct a SSM (Cootes et al., 1995), with the results of mean shape $\overline{X}$ and mode matrix P in Eq. (3). We denote this landmark set as $\mathbf{X} = \{X_i|_{i=1,\dots,N}\}$. Then, we generate random samples using resulting SSM for a pre-training of DeepSSM-Net, where the samples are generated using $X^* = T_{m_*^a}(\overline{X} + Pm_*^b)$. m_*^a and m_*^b are the gold standard parameters of X^* for supervised learning.

Algorithm 1: Training stage of SGMaRL

Data: Training set $\{(I_i, X_i, \mathcal{J}_i)|_{i=1,\cdots,N}\}$, number of training images N, landmark set library $\mathbf{X}$, maximum number of episodes M, budget T, greedy parameter ϵ
Result: Optimal ω, θ

Initialize DeepMaQ-Net with random weights ω , the lag weight $\omega^- = \omega$, replay memory $\mathbf{D}$;
Construct SSM, pre-train DeepSSM-Net and expand $\mathbf{X}$;
for *episode = 1, M* **do**
  Select a random image I_i;
  Initialise sequences $\{x_j|_{j\in\mathcal{J}_i}\}$ with random landmarks, state $\{s_j^{[0]}|_{j\in\mathcal{J}_i}\}$ according to x_j ;
  for $t = 1, T$ **do**
    If agent $\{j|_{j\in\mathcal{J}_i}\}$ is not terminated, performs $a_j^{[t]}$ according to DeepMaQ-Net with probability $1-\epsilon$, otherwise selects a random action;
    Get $x_j^{[t+1]}$, $r_j^{[t]}$, $s_j^{[t+1]}$ and store transition $(s_j^{[t]}, a_j^{[t]}, r_j^{[t]}, s_j^{[t+1]})$ in $\mathbf{D}$;
    Sample random batch of transitions from $\mathbf{D}$, perform a gradient descent step on Eq. (2) with respect to DeepMaQ-Net parameters ω;
    Terminate agent $\{j|_{j\in\mathcal{J}_i}\}$ if it finds target, oscillates or reaches max steps, break when all agents are terminated;
  end
  Every c_1 steps, reset $\omega^- = \omega$ and update ϵ;
  Every c_2 steps, test DeepMaQ-Net with $\{I_i{}^c\}$ from training set, get predictions $\{\hat{X}_i\}$ and store them in $\mathbf{X}$;
  Sample random batch from $\mathbf{X}$, perform a gradient descent step on Eq. (4) with respect to DeepSSM-Net parameters θ;
end

Algorithm 2: Test stage of SGMaRL

Data: Test image set $\{I_i|_{i=1,\cdots,N'}\}$, number of test images N', index of all target landmarks $\mathcal{J}$, budget of DeepMaQ T, maximum number of subiterations with DeepSSM T', maximum number of iterations M'
Result: Landmarks prediction $\{X_i^{pred}|_{i=1,\cdots,N'}\}$

Select image I_i from $\{I_i|_{i=1,\cdots,N'}\}$;
for *iteration = 1, M'* **do**
  Initialise sequence $\{x_j|_{j\in\mathcal{J}}\}$ with random landmarks if iteration = 1, otherwise initialise $\{x_j|_{j\in\mathcal{J}}\}$ with X_i^{pred};
  for $t = 1,T$ **do**
    If agent $\{j|_{j\in\mathcal{J}}\}$ is not terminated, performs $a_j^{[t]}$ according to DeepMaQ-Net ,moves from $x_j^{[t]}$ to $x_j^{[t+1]}$;
    Terminate agent $\{j|_{j\in\mathcal{J}}\}$ if it oscillates or reaches max steps, update x_j with $x_j^{[t+1]}$, break when all agents are terminated;
  end
  for $t = 1,T'$ **do**
    Set $\Phi = \emptyset$;
    Add $\{j|_{j\in\mathcal{J}}\}$ to Φ if x_j is judged to be corrected according to DeepSSM-Net, break if $\Phi = \emptyset$;
    Calculate alternative landmarks $\{\tilde{x_j}|_{j\in\mathcal{J}}\}$ with DeepSSM-Net;
    Update $\{x_j|_{j\in\Phi}\}$ with $\{\tilde{x_j}|_{j\in\Phi}\}$ if $\tilde{x_j}$ is better than x_j according to DeepSSM-Net;
  end
  Update X_i^{pred} with $\{x_j|_{j\in\mathcal{J}}\}$, break if X_i^{pred} is convergence;
end

Finally, these generated samples can be added to $\mathbf{X}$ for the joint training of DeepMaQ-Net and DeepSSM-Net.

As Alg. 1 shows, in the joint training the outputs of DeepMaQ-Net, which are inputs of DeepSSM-Net, are also samples for training the latter network. Hence, they are further included into the library $\mathbf{X}$ for batch-based random sampling. Note that in the joint training, we use Eq. (4) as the unsupervised loss function for DeepSSM-Net.

3.3.2. Test

In the test stage, DeepMaQ-Net and DeepSSM-Net work in an interleaving and collaborative fashion, as the pseudo code in Alg. 2 illustrates. After initialization of the agents, DeepMaQ-Net navigates all the agents with the actions predicted from the network, accordingly to the similar rules in the training stage. After the navigation ends, DeepSSM-Net then regularizes the set of predicted landmarks iteratively until the optimal solution of landmark positions are achieved or maximal number of subiteration steps (T') is met. The resulting positions of landmarks are then fed into DeepMaQ-Net as the initial states of all the agents in DeepMaQ-Net for new round navigation. This interleaving application of DeepMaQ-Net and DeepSSM-Net last until the detection results converge or the iteration meets the maximal steps.

4. Experiment

4.1. Incomplete image scenarios

The proposed SGMaRL was evaluated in the following scenarios where incomplete images or labels could be encountered in either training or test stage, as illustrated in Fig. 1.

- **(Baseline) Train**: $I^c + L^c$, **Target**: $I^c \rightarrow L^c$. The training set consists of **c**omplete images with **c**omplete labels, and the task is to predict **c**omplete labels on **c**omplete images.
- **(Scenario I) Train**: $I^{inc} + L^{inc}$, **Target**: $I^c \rightarrow L^c$. The training set consists of **inc**omplete images with **inc**omplete labels, and the task is to predict **c**omplete labels on **c**omplete images.
- **(Scenario II) Train**: $I^c + L^{inc}$, **Target**: $I^c \rightarrow L^c$. The training set consists of **c**omplete images with **inc**omplete labels, and the task is to predict **c**omplete labels on **c**omplete images.
- **(Scenario III) Train**: $I^c + L^c$, **Target**: $I^{inc} \rightarrow L^{inc}$. The training set consists of **c**omplete images with **c**omplete labels, and the task is to predict **inc**omplete labels on **inc**omplete images.
- **(Scenario IV) Train**: $I^c + L^c$, **Target**: $I^{inc} \rightarrow L^c$. The training set consists of **c**omplete images with **c**omplete labels, and the task is to predict **c**omplete labels on **inc**omplete images.

For a given image, the missing landmarks are denoted as *complementary label* (L^{c-inc}). Noted that for complete images (I^c), it is possible that all landmarks are available and $L^{c-inc} = \emptyset$. But in some special cases, some landmarks may be absent, as shown in Fig. 7(a). For incomplete images (I^{inc}), $L^{c-inc} \neq \emptyset$. The landmarks near the edge of cropped images and those outsides belong to L^{c-inc} (see green points in Fig. 3(a)).

4.2. Dataset

We used four datasets to validate the proposed method in the scenarios mentioned above. The four datasets include (1) the whole-body dual energy X-ray absorptiometry (DXA) images, (2) the cardiac MRI images, (3) the whole-head CT images and (4) the half-head CT images. Appropriate dataset was chosen for each scenario according to clinical practice and the landmarks used in the experiment are of clinical significance.

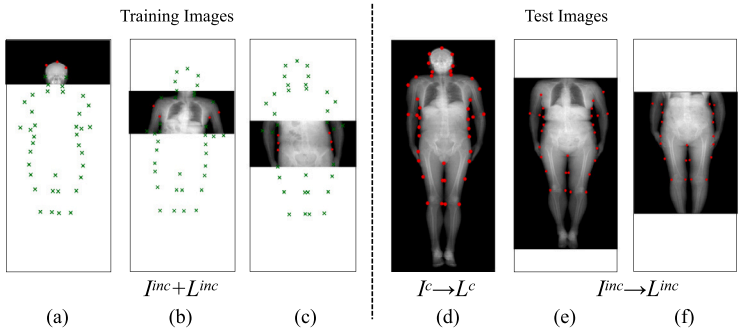

Fig. 3. Examples from the body dataset used in our experiments. (a), (b) and (c) represent training images while (d), (e) and (f) represent test images. The dots (•) represent landmarks known and the crosses (×) represent L^{c-inc}.

The **whole-body DXA dataset** consists of 99 DXA original samples, provided by Zhongshan Hospital affiliated to Fudan University, Shanghai, China (Gao et al., 2010). The participants from Shanghai Changfeng Study were scanned over the whole body with lunar iDXA produced by GE company. All these data were converted into images of JPG format with in-plane resolution of 3.0375 mm × 2.4001 mm. Each sample was labeled with 40 contour keypoints by experts, as shown in Fig. 3(d). These landmarks can be used to describe the geometric shape and build body shape features of each individual (Shepherd et al., 2017). The heights of the images in this dataset are large, so the dataset was used to verify whether SgMaRL can handle the complex changes in missing proportions in Scenario I.

The **cardiac MRI dataset** consists of 45 LGE CMR sequences, collected from Shanghai Renji hospital (Zhuang, 2016, 2019; Zhuang et al., 2022). Each sequence contains 10 to 18 slices with in-plane resolution of 1 mm × 1 mm, covering the main body of the ventricles. All 3D volumes were sliced and cropped to be 2D images with size of 256 × 256 in pixels. The segmentation ground truth for left ventricle (LV), right ventricle (RV) and myocardium (Myo) were provided by experts. Segmentation labels on each image were pre-processed to generate 33 landmarks for model training, and at the inference stage the predicted landmarks would be transformed to deliver the segmentation output. One can refer to Sec. 2.1 in supplementary material for the details of transformation between segmentation and landmarks. The landmarks on the contour of the target organ can guide follow-up segmentation work (Bhalodia et al., 2018). Images of this dataset have complex shape variability, so it was used to verify whether SgMaRL can preserve difficult shapes in Scenario II.

The **whole-head CT dataset** has 54 whole-head CT images, provided by Huashan Hospital affiliated to Fudan University, Shanghai, China (Qian et al., 2022). Each volume was manually labeled by experts with 25 anatomical skull landmarks, which located on the surface of upper-half head, as shown in Fig. 9(a). These landmarks can be used to investigate the genetic basis with morphology (Qian et al., 2022).

The **half-head CT dataset** consists of 41 upper-half head CT volumes from the Northern Han Chinese cohort, which was provided by Shanghai Institute of Nutrition and Health, Chinese Academy of Sciences. The number of landmarks in these volumes varies from 14 to 19 according to the position of intersecting surface. These volumes have much larger sample interval and lower resolution than samples from the whole-head dataset, as shown in Fig. 9(a) and (b). All of these volumes were reconstructed into an isotropic spacing of 1 mm along all three axes. Both whole-head dataset and half-head dataset are head CT scans. However, they have different FOVs and sample intervals. Thus we designed a comparison experiment based on these datasets in scenario III and IV.

4.3. Implementation details

Both DeepMaQ and DeepSSM were implemented on an NVIDIA GTX 1080Ti GPU in most experiments. However, both of them were implemented on an Intel Core i7-5700HQ CPU during test phrase to compare inference time with other methods. For DeepMaQ, we adopted oscillation property, which means if an agent passes one location for several times, it should be terminated. If the agent reaches the max steps, it would also be terminated. Note a terminated agent has reached its limit with given start position, but it is not guaranteed to find the target. The reward was set to be −100 to punish the agent when it moves across the border. The experience replay memory size, the discount factor γ and the time lag term were set as 10^5, 0.9 and 4, respectively. For 2D and 3D networks of DeepMaQ, the sizes of input images were 80×80 and $30 \times 30 \times 30$, respectively. The batch size was 48 for the former and 32 for the latter. Max steps for every agent was 200 for the former and 500 for the latter. DeepMaQ took 24–36 h and 48–72 h for training 2D and 3D models, respectively. For DeepSSM, we used complete labels for the construction of SSM and pre-training. The maximum number of iterations M' (see in **Alg. 2**) for test process was 5. The pre-training time of DeepSSM was around an hour for both 2D and 3D models, respectively.

4.4. Studies of *scenario I* using whole-body DXA

Datapreparation : The body dataset was randomly divided into 40 and 59 images for training and test, respectively. Each image in the training set was cropped horizontally with a predefined missing proportion (mp) $x\%$, which indicates that $x\%$ image area is randomly removed, and the height of cropped image is $1-x\%$ of the original one. To guarantee every agent learns the information of its corresponding landmark, each landmark index should be visited for at least once. Examples can be found in Fig. 3(a), (b) and (c). To evaluate our model in Scenario I, the experiments were conducted in two settings: (1) **Train:** $I^{inc}+L^{inc}$, **Target:** $I^{c}\rightarrow L^{c}$. All images in the test set were complete, as shown in Fig. 3(d). (2) **Train:** $I^{inc}+L^{inc}$, **Target:** $I^{inc}\rightarrow L^{inc}$. For training data, mp was set to 0%, 20%, 40%, 60% and 80%, respectively, while for test, it was set to 20% and 40% (see Fig. 3(e) and (f)). In the following, we notated the missing proportion of training images and test images as mp^{train} and mp^{test}, respectively. For comparisons, we also implemented (1) DeepMaQ, (2) SGMarL (T), a framework similar to SGMarL, replacing DeepSSM by the traditional SSM, and (3) ResNet, a ResNet-based landmark detection method (Terada et al., 2018). Euclidean distance was used as a distance metric and average distance error (ADE) was used to measure the evaluation accuracy of landmark detection.

Results : From Table 2, one can see that SGMaRL performed well with training or test data being either complete or incomplete. When testing on complete images, SGMaRL achieved 2.29 cm of ADE when $mp^{train}=80\%$, which was even better than the result of ResNet trained with complete images. When testing on incomplete images, ResNet almost failed in all settings. In contrast, when training with mp^{train} being 80%, SGMaRL could still achieved 2.95 cm and 5.06 cm with $mp^{test}=20\%$ and $mp^{test}=40\%$, respectively. Table 2 also shows that

Table 2
Comparative results between DeepMaQ and SGMaRL with different *mp*. SGMaRL(T), ResNet with mp^{train} = 0% was also included. ADE are in *cm*. Note that when mp^{test} = 40%, SGMaRL(T) failed with all mp^{train} settings, and thus corresponding results were not included.

mp^{train}	mp^{test}										
	0% (I^c)				20%				40%		
	DeepMaQ	SGMaRL	SGMaRL(T)	ResNet	DeepMaQ	SGMaRL	SGMaRL(T)	ResNet	DeepMaQ	SGMaRL	ResNet
0% (I^c)	1.12 ± 0.336	1.10 ± 0.281	1.09 ± 0.251	2.62 ± 1.24	1.24 ± 0.403	1.17 ± 0.311	1.17 ± 0.255	14.6 ± 9.40	1.36 ± 0.548	1.35 ± 0.532	13.9 ± 8.03
20%	1.10 ± 0.274	1.08 ± 0.244	1.07 ± 0.249	N/A	1.20 ± 0.350	1.18 ± 0.324	1.18 ± 0.319	N/A	1.27 ± 0.415	1.26 ± 0.407	N/A
40%	1.38 ± 0.593	1.30 ± 0.339	1.27 ± 0.341	N/A	1.25 ± 0.365	1.22 ± 0.316	1.22 ± 0.310	N/A	1.28 ± 0.353	1.27 ± 0.336	N/A
60%	7.06 ± 3.05	1.33 ± 0.328	Fail	N/A	5.00 ± 4.15	1.48 ± 0.469	Fail	N/A	2.17 ± 1.52	1.90 ± 1.23	N/A
80%	27.1 ± 3.79	2.29 ± 0.703	Fail	N/A	21.0 ± 3.62	2.95 ± 1.72	Fail	N/A	14.3 ± 4.08	5.06 ± 2.05	N/A

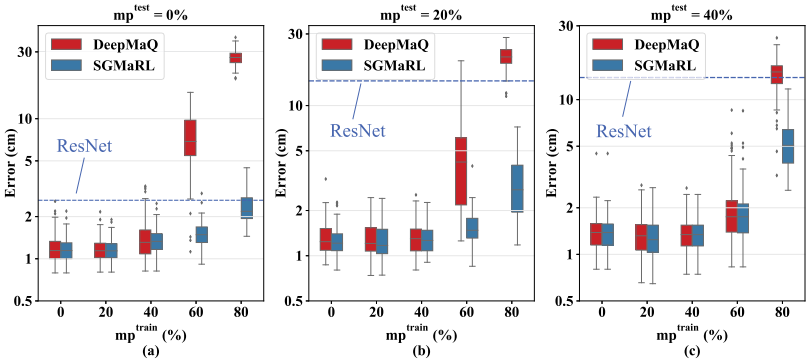

Fig. 4. Comparison between DeepMaQ and SGMaRL with different *mp*. Predictions of ResNet with mp^{train} = 0% were used as baselines.

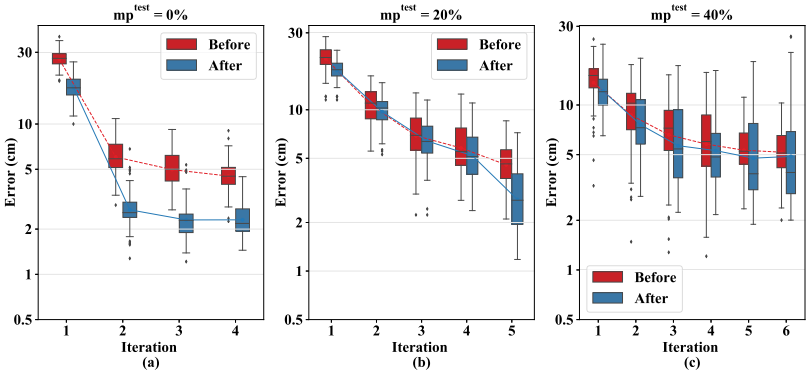

Fig. 5. Comparison before and after DeepSSM during iterations of SGMaRL with mp^{train} = 80%.

when $mp^{train} \geq 60\%$ or $mp^{test} = 40\%$, SGMaRL(T) failed because the traditional SSM cannot converge, while DeepSSM still worked. In the successful settings ($mp^{train} \leq 40\%$ and $mp^{test} \leq 40\%$), SGMaRL was slightly worse than SGMaRL(T) with the largest gap being only 0.03 cm.

For better illustration, we further visualized these results in Fig. 4. It shows that DeepMaQ and SGMaRL can both achieve acceptable results when mp^{train} is small (e.g. $mp^{train} \leq 40\%$). Moreover, when $mp^{train} = 80\%$, DeepMaQ failed while SGMaRL still worked. As shown in Fig. 4(a), when testing on complete images, there was no significant difference between predictions of DeepMaQ with $mp^{train} = 0\%$ and $mp^{train} = 20\%$ (p=0.733). When mp^{train} was less than 40%, DeepSSM could only deliver marginal assistance, and ADE of both DeepMaQ and SGMaRL varied in a small range from 1.08 cm to 1.38 cm. However, when mp^{train} rose from 60% to 80%, ADE of DeepMaQ increased largely from 7.06 cm to 27.1 cm. On the contrast, ADE of SGMaRL only increased slightly from 1.33 cm to 2.29 cm, thanks to the prior shape knowledge introduced by DeepSSM. Testing on incomplete images showed a similar tendency (see Fig. 4(b) and (c)).

To show how DeepSSM helps during the test of SGMaRL, we compared the predictions of SGMaRL before and after DeepSSM module in each iteration in Fig. 5. As shown in Fig. 5(b), before using DeepSSM, SGMaRL kept descending as the number of iterations increased. Besides, ADE of SGMaRL became smaller after adding DeepSSM in each iteration. We conclude that DeepSSM can improve predictions both in each iteration and between iterations of SGMaRL in general. Fig. 6 further provides visualization results. The predictions far from gold standards (GD) are denoted as *abnormal predictions*. Abnormal predictions occurred at the beginning and affected the detection accuracy. Moreover, as shown in Table 2 and Fig. 5, most abnormal predictions were improved, with the number of iterations increasing. This is probably because DeepSSM provided sufficient shape prior information for landmark detection. Besides, as shown in Table 2, when mp^{train} was 80%, ADE of SGMaRL was 5.06 cm with $mp^{test} = 40\%$, which was much worse than 2.29 cm and 2.95 cm with $mp^{test} = 0\%$ and $mp^{test} = 20\%$, respectively. Additionally, when $mp^{test} = 40\%$, Fig. 5(c) shows larger variance and slower descent of performance during iterations, comparing to $mp^{test} = 0\%$ and $mp^{test} = 20\%$. It is probably because larger mp^{test} with more L^{c-inc} on test images could suppress the advantages of DeepSSM.

4.5. Studies of *scenario II* using cardiac MRI

Data preparation : The aim was to predict all labels (LV, RV and Myo) for complete images in test set. The final predicted labels were

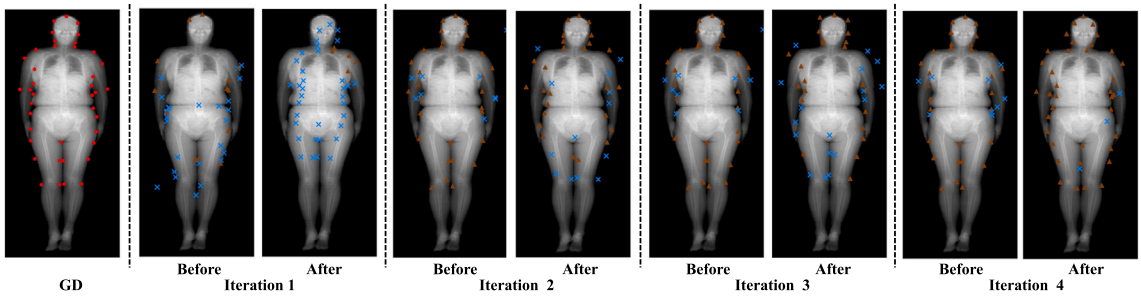

Fig. 6. The visualization of predictions before and after DeepSSM during iterations of SGMaRL with mp^{train} = 80%, mp^{test} = 0%. The dots (•) represent GD landmarks. The triangles (▲) represent predictions of landmarks which are close to GD while the crosses (×) represent abnormal predictions.

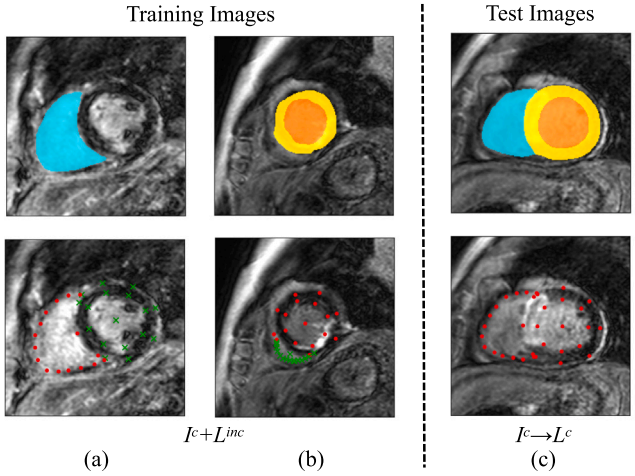

Fig. 7. Examples from the cardiac dataset used in our experiments. (a) and (b) represent training images while (c) represent test images. RV, LV and Myo are labeled in blue, orange and yellow, respectively. The dots (•) represent landmarks known and the crosses (×) represent the L^{c-inc}.

generated by DeepSSM based on the output of SGMaRL. For comparison, the experiments were conducted in two settings: (1) **Train:** $I^c + L^c$, **Target:** $I^c \rightarrow L^c$. The cardiac dataset was randomly divided into 20 and 25 samples for training and test, respectively. All training samples have labels of LV, Myo and RV. ResNet and UNet were selected as landmark detection method and segmentation method for comparison, respectively. ResNet delivered the landmarks to segmentation output via post-processing, while UNet outputted segmentation prediction directly. (2) **Train:** $I^c + L^{inc}$, **Target:** $I^c \rightarrow L^c$. All training images were complete. However, half of the training images were labeled with only LV and Myo (see Fig. 7(b)) while the others were labeled with only RV (see Fig. 7(a)). Because training with incomplete labels was more challenging, a little more samples were used in the training set. The cardiac dataset was randomly divided into 30 and 15 samples for training and test, respectively. Dice score, average symmetric surface distance (ASD), Hausdorff distance (HD) and ADE were used as metrics for the evaluation.

Results : Table 3 shows that SGMaRL can successfully predict landmarks and segmentation when training with complete or incomplete labels. For landmark detection, SGMaRL could reach 7.40 mm of ADE with complete labels, while ResNet reached 7.67 mm. With incomplete labels, SGMaRL could still achieve 8.19 mm of ADE. For segmentation, SGMaRL also had advantage over ResNet for most metrics, but there was no significant difference between HD of RV (p = 0.807). It also had great advantage over UNet in terms of HD. SGMaRL could reach 16.2 mm in terms of HD for RV, while UNet only could reach 54.3 mm. Note that UNet obtained better Dice scores and ASD compared to SGMaRL. This is because the segmentation obtained by SGMaRL was converted from predicted landmarks, which may introduce additional errors compared to the method that directly predicted pixel-wise segmentation.

To show the advantages of SGMaRL in preserving shape, we compared results of different methods and visualized them in Fig. 8. It shows that SGMaRL could predict smooth and accurate shape. In contrast, without structural information provided by DeepSSM, the prediction of DeepMaQ had irregular shape, due to the existence of abnormal landmarks prediction. The prediction of UNet might have an irregular boundary and many noises, which caused large HD (see Table 3). ResNet could predict a reasonable shape, but the prediction might be sometimes inaccuracy, which caused low Dice scores. Fig. 8 also shows SGMaRL can deal with variety of shapes effectively. In different slices, the shape of RV expressed complex variation. With the help of DeepSSM, SGMaRL can correct the abnormal landmarks and make prediction reasonable.

Table 3 further shows SGMaRL has additional advantages when training with incomplete labels. For example, the Dice score of RV rises from 0.780 to 0.825 after adding the DeepSSM. Besides, HD averagely decreased by 3.13 mm and 4.47 mm when training with complete and incomplete labels, respectively. It is possibly because training with incomplete labels could be quite challenging for DeepMaQ. With worse initial prediction, DeepSSM can improve HD more effectively. Moreover, SGMaRL has better segmentation performance in terms of contours than filled regions. For example, when training with complete labels, HD of all areas were improved with the assistance of DeepSSM. However, the Dice score of LV was 0.898 with DeepMaQ, while it dropped 0.023 with SGMaRL. We concluded that DeepSSM can improve HD and ADE effectively, but its influence on Dice score and ASD still needs further research.

4.6. Studies of *scenario III and IV* using head CT

Data preparation : We evaluated our methods in Scenario III and IV and conducted the experiments in two settings: (1) **Train:** $I^c + L^c$, **Target:** $I^{inc} \rightarrow L^{inc}$. We randomly selected 30 whole-head images from the whole-head dataset for training, while obtained 41 half-head images from the half-head dataset for test. The test set was denoted as **Real-Domain**, as shown in Fig. 9(b). The number of landmarks labeled by experts on image from Real-Domain does not exceed 19. (2) **Train:** $I^c + L^c$, **Target:** $I^{inc} \rightarrow L^c$. The training set is the same as setting (1), while for the test set, other 24 whole-head images were half cut as simulated upper-half head images. This test set was denoted as **Sim-Domain**, as shown in Fig. 9(c). To ensure the high similarity between Sim-Domain and Real-Domain, we varied the landmarks in training data of Sim-Domain from 14 to 19 via controlling the position of intersecting surface when half cutting the whole images. For setting (1), L^{inc} (landmarks labeled by expert) were considered for evaluation. For setting (2), we predicted all existed landmarks, which were divided into L^{inc} (landmarks inside of cropped images) and L^{c-inc} (landmarks near the edge of cropped images or those outsides). Landmark detection methods based on multi-atlas registration (MAS) and single-atlas

Table 3
Comparison between ResNet, UNet, DeepMaQ and SGMaRL with different experiment settings. The best average values have been shown in bold. The down arrow (↓) denotes smaller is better, while the up arrow (↑) denotes larger is better.

Method	LV			Myo			RV			ADE (mm) (↓)
	Dice (↑)	ASD (mm) (↓)	HD (mm) (↓)	Dice (↑)	ASD (mm) (↓)	HD (mm) (↓)	Dice (↑)	ASD (mm) (↓)	HD (mm) (↓)	
Train: $I^c + L^c$ Target: $I^c \rightarrow L^c$										
ResNet	0.835 ± 0.062	3.70 ± 1.37	13.5 ± 3.65	0.645 ± 0.092	3.77 ± 1.35	13.6 ± 4.23	0.774 ± 0.074	3.93 ± 1.23	**16.1 ± 3.02**	7.67± 3.46
UNet	**0.910 ± 0.045**	**2.49 ± 1.35**	29.5 ± 16.1	**0.816 ± 0.067**	**2.11 ± 1.03**	13.4 ± 10.0	**0.845 ± 0.197**	**3.69 ± 2.75**	54.3 ± 24.8	N/A
DeepMaQ	0.898 ± 0.045	2.59 ± 0.851	15.9 ± 8.49	0.754 ± 0.056	2.26 ± 0.783	14.9 ± 10.5	0.833 ± 0.085	2.91 ± 1.23	19.0 ± 7.51	7.78 ± 4.16
SGMaRL	0.875 ± 0.045	3.09 ± 0.732	**12.5 ± 5.84**	0.702 ± 0.064	2.92 ± 0.888	**11.7 ± 6.94**	0.778 ± 0.082	3.72 ± 1.23	16.2 ± 5.67	**7.40 ± 4.03**
Train: $I^c + L^{inc}$ Target: $I^c \rightarrow L^c$										
DeepMaQ	0.840 ± 0.074	3.82 ± 1.47	18.1 ± 7.57	0.650 ± 0.100	5.29 ± 1.51	16.2 ± 6.77	0.780 ± 0.073	**3.78 ± 1.20**	17.3 ± 7.60	8.66 ± 4.79
SGMaRL	**0.861 ± 0.065**	**3.04 ± 1.30**	**12.1 ± 5.43**	**0.693 ± 0.091**	**3.12 ± 1.32**	**12.0 ± 5.38**	**0.825 ± 0.086**	3.89 ± 1.35	**14.1 ± 4.53**	**8.19 ± 4.65**

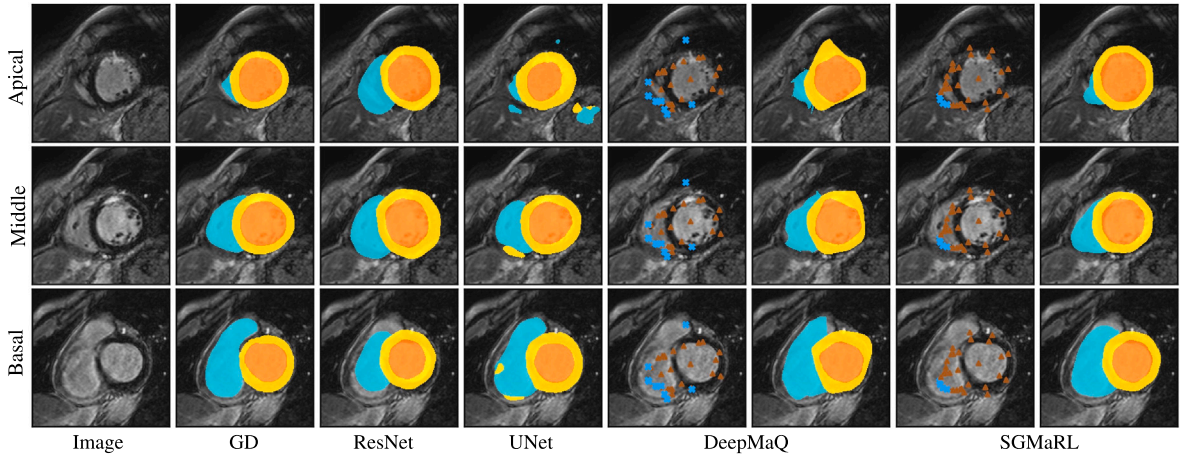

Fig. 8. Comparison of prediction results of ResNet, UNet, DeepMaQ and SGMaRL with GD in three typical slices. RV, LV and Myo are labeled in blue, orange and yellow, respectively. The triangles (▲) represent predictions of landmarks which are close to GD while the crosses (×) represent abnormal predictions.

Table 4
Comparative results MAS, SAS, DeepMaQ and SGMaRL with different types of landmarks in different domains. ADE are in mm.

Method	Real-Domain	Sim-Domain		Inference time
	L^{inc}	L^{inc}	L^{c-inc}	
MAS	6.10 ± 1.59	2.87 ± 0.782	6.80 ± 0.867	35 min 30 s
SAS	7.36 ± 1.50	5.58 ± 1.97	21.8 ± 10.1	1 min 11 s
DeepMaQ	8.99 ± 2.84	5.12 ± 1.85	11.4 ± 7.18	4 min
SGMaRL	7.25 ± 1.11	5.01 ± 1.83	6.84 ± 4.82	4 min 1 s

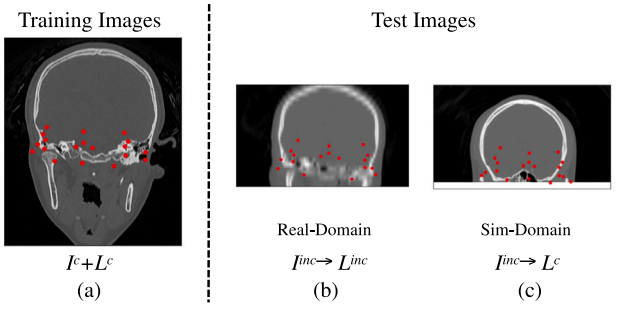

Fig. 9. Examples from the head datasets used in our experiments. (a) represents training images while (b) and (c) represent test images. The dots (•) represent landmarks known. Note that all images in this figure are 2D slices of 3D images from front view, and landmarks on them are mapped from 3D space. 3D head images are used in our experiment.

registration (SAS) from Zhuang et al. (2010) were used for comparison. All methods implemented on an Intel Core i7-5700HQ CPU. ADE was used for evaluation of landmarks detection accuracy.

Results : As shown in Table 4, our proposed method can successfully find different types of landmarks in both domains. The result of SGMaRL was better than SAS but worse than MAS. For example, ADE of L^{inc} with SGMaRL was 7.25 mm in Real-Domain, which was smaller than 7.36 mm by SAS, but larger than 6.10 mm by MAS. However, SGMaRL consumed far less time than MAS. MAS took more than 35 min while SGMaRL only took about 4 min on CPU with one case.

For most methods, the prediction of L^{inc} in Real-Domain was harder than L^{inc} in Sim-Domain, while L^{c-inc} in Sim-Domain was the most difficult. For example, DeepMaQ achieved ADE of 5.12 mm, 8.99 mm and 11.4 mm on L^{inc} in Sim-Domain, L^{inc} in Real-Domain and L^{c-inc} in Sim-Domain, respectively. Fig. 9 shows Real-domain test set has obvious appearance gap with the training images and poorer image quality, which increases the difficulty for detection task. For L^{c-inc}, it could be only inferred with structural information since their local appearance information was almost missing. With the help of DeepSSM, ADE of SGMaRL decreased by 0.11 mm, 1.74 mm, 4.56 mm in these landmarks, respectively. We concluded that DeepMaQ has worse prediction when target image has poor appearance information.

To investigate the robustness of the proposed method, we further divided L^{inc} and L^{c-inc} in Sim-domain into four categories, i.e., Easy, Normal, Hard and Extremely Hard, according to their detection difficulties and compared the performance on them. Fig. 10(a) illustrates the classification of the landmarks, and (b) shows the performance of the compared methods in the four categories. One can see that from Easy to Hard, SGMaRL had a stable performance, while ADE of other three methods increased. When tested on Extremely Hard samples, DeepMaQ and SAS performed poor, while the prediction error of SGMaRL only increased slightly, when compared to the results on Hard samples. The

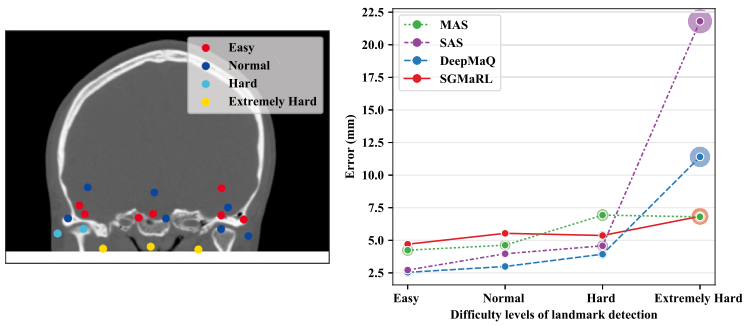

Fig. 10. The right graph shows comparative result of DeepMaQ, SGMaRL, MAS, and SAS on landmarks with different difficulty levels in Sim-Domain. The size of bubble represents the standard deviation. The left graph shows how different difficulty levels of landmarks located on a head image. Extremely hard represents landmarks which belong to L^{c-inc} on every image, while Hard represents the landmarks which belong to L^{c-inc} on this image but belong to L^{inc} on other images. The rest existent landmarks were sub-divided into Easy and Normal according to inter-observer errors.

results show that SGMaRL has great robustness on landmarks with different levels of detection difficulty.

5. Conclusion and discussion

In this paper, we proposed a shape-guided multi-agent RL algorithm for landmark detection with incomplete images. The algorithm contains two modules: DeepMaQ and DeepSSM. The former can detect multiple landmarks as targets synchronously with incomplete images for training; the latter can correct abnormal predictions and provide more accurate start positions for DeepMaQ. By combining them, the proposed method can exploit both local image information and shape information effectively, which is fast and robust for solving the general problem of multiple landmark detection with challenging incomplete images. Experiments conducted on different datasets show that our methods can perform landmark detection in different incomplete image problems and complex scenarios.

We have investigated the performance of our methods upon different mp in both training and test set. The results show that DeepSSM can achieve greater improvement when mp^{train} increases. We also have carried out comparison study of the prediction errors of our methods with different types of landmarks. The results show that (1) with prior shape information from DeepSSM, SGMaRL can even predict the landmarks outside the target image; (2) SGMaRL is robust on landmarks with different detection difficulties. Besides, we have extended our methods to segmentation task. The results show that DeepSSM can ensure the segmentation prediction of SGMaRL have a smooth shape based on landmarks as partial labels for training.

Though DeepSSM can provide great help, it can be explored for further improvement. As mentioned before, the assistance of DeepSSM became weaker as mp^{test} increasing. The details of DeepSSM are to be improved and refined to find abnormal landmarks more accurately, especially when mp^{test} is large. When performing segmentation, Dice score and ASD became worse after adding DeepSSM with complete labels for training. We will investigate the influence of DeepSSM on different metrics and hope modified DeepSSM can improve all metrics in different scenarios. Besides, learning the search strategy and localization in two stages may increase the possibility of abnormal predictions. Thus, it will be interesting to investigate a more elegant way to combine DeepMaQ with prior shape information. For this purpose, a new multi-agent RL sharing structural information with each agent will be studied in the future.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (61971142, 62111530195 and 62011540404) and the development fund for Shanghai talents (2020015).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.media.2023.102875.

References

Alansary, A., Folgoc, L.L., Vaillant, G., Oktay, O., Li, Y., Bai, W., Passerat-Palmbach, J., Guerrero, R., Kamnitsas, K., Hou, B., McDonagh, S.G., Glocker, B., Kainz, B., Rueckert, D., 2018. Automatic view planning with multi-scale deep reinforcement learning agents. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 277–285.

Alansary, A., Oktay, O., Li, Y., Le Folgoc, L., Hou, B., Vaillant, G., Kamnitsas, K., Vlontzos, A., Glocker, B., Kainz, B., et al., 2019. Evaluating reinforcement learning agents for anatomical landmark detection. Med. Image Anal. 53, 156–164.

Bhalodia, R., Elhabian, S.Y., Kavan, L., Whitaker, R.T., 2018. DeepSSM: A deep learning framework for statistical shape modeling from raw images. In: Reuter, M., Wachinger, C., Lombaert, H., Paniagua, B., Lüthi, M., Egger, B. (Eds.), Shape in Medical Imaging. Springer International Publishing, Cham, pp. 244–257.

Briot, K., 2013. DXA parameters: Beyond bone mineral density. Jt. Bone Spine 80 (3), 265–269.

Chintalapani, G., Murphy, R., Armiger, R.S., Lepisto, J., Otake, Y., Sugano, N., Taylor, R.H., Armand, M., 2010. Statistical atlas based extrapolation of CT data. In: Wong, K.H., Miga, M.I. (Eds.), Medical Imaging 2010: Visualization, Image-Guided Procedures, and Modeling, Vol. 7625. SPIE, International Society for Optics and Photonics, 762539.

Cootes, T., Edwards, G., Taylor, C., 2001. Active appearance models. IEEE Trans. Pattern Anal. Mach. Intell. 23 (6), 681–685.

Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J., 1995. Active shape models-their training and application. Comput. Vis. Image Underst. 61 (1), 38–59.

Cremers, D., Rousson, M., Deriche, R., 2007. A review of statistical approaches to level set segmentation: Integrating color, texture, motion and shape. Int. J. Comput. Vis. 72, 195–215.

Cristinacce, D., Cootes, T.F., et al., 2006. Feature detection and tracking with constrained local models. In: Bmvc, Vol. 1. No. 2. Citeseer, p. 3.

Dantone, M., Gall, J., Fanelli, G., Van Gool, L., 2012. Real-time facial feature detection using conditional regression forests. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. pp. 2578–2585.

Davatzikos, C., Tao, X., Shen, D., 2003. Hierarchical active shape models, using the wavelet transform. IEEE Trans. Med. Imaging 22 (3), 414–423.

Fries, J.A., Varma, P., Chen, V.S., Xiao, K., Tejeda, H., Saha, P., Dunnmon, J., Chubb, H., Maskatia, S., Fiterau, M., et al., 2019. Weakly supervised classification of aortic valve malformations using unlabeled cardiac MRI sequences. Nature Commun. 10 (1), 1–10.

Gao, X., Hofman, A., Hu, Y., Lin, H.-D., Zhu, C., Jeekel, J., juan Jin, X., yao Wang, J., Gao, J., Yin, Y., Zhao, N., 2010. The shanghai changfeng study: a community-based prospective cohort study of chronic diseases among middle-aged and elderly: objectives and design. Eur. J. Epidemiol. 25, 885–893.

Gass, T., Szekely, G., Goksel, O., 2014. Multi-atlas segmentation and landmark localization in images with large field of view. In: International MICCAI Workshop on Medical Computer Vision. Springer, pp. 171–180.

Ghesu, F.C., Georgescu, B., Grbic, S., Maier, A., Hornegger, J., Comaniciu, D., 2018. Towards intelligent robust detection of anatomical structures in incomplete volumetric data. Med. Image Anal. 48, 203–213.

Ghesu, F.C., Georgescu, B., Mansi, T., Neumann, D., Hornegger, J., Comaniciu, D., 2016. An artificial agent for anatomical landmark detection in medical images. In: Medical Image Computing and Computer-Assisted Intervention. MICCAI 2016, Springer International Publishing, Cham, pp. 229–237.

Ghesu, F.C., Georgescu, B., Zheng, Y., Grbic, S., Maier, A., Hornegger, J., Comaniciu, D., 2017. Multi-scale deep reinforcement learning for real-time 3D-landmark detection in CT scans. IEEE Trans. Pattern Anal. Mach. Intell. 41 (1), 176–189.

Grupp, R.B., Chiang, H., Otake, Y., Murphy, R.J., Gordon, C.R., Armand, M., Taylor, R.H., 2015. Smooth extrapolation of unknown anatomy via statistical shape models. In: Medical Imaging 2015: Image-Guided Procedures, Robotic Interventions, and Modeling, Vol. 9415. SPIE, International Society for Optics and Photonics, 941524.

Hakim, A., Christensen, S., Winzeck, S., Lansberg, M.G., Parsons, M.W., Lucas, C., Robben, D., Wiest, R., Reyes, M., Zaharchuk, G., 2021. Predicting Infarct Core From Computed Tomography Perfusion in Acute Ischemia With Machine Learning: Lessons From the ISLES Challenge. Stroke 52 (7), 2328–2337.

Hanaoka, S., Shimizu, A., Nemoto, M., Nomura, Y., Miki, S., Yoshikawa, T., Hayashi, N., Ohtomo, K., Masutani, Y., 2017. Automatic detection of over 100 anatomical landmarks in medical CT images: A framework with independent detectors and combinatorial optimization. Med. Image Anal. 35, 192–214.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition. CVPR, pp. 770–778.

Jain, A., Powers, A., Johnson, H.J., 2020. Robust automatic multiple landmark detection. In: 2020 IEEE 17th International Symposium on Biomedical Imaging. ISBI, IEEE, pp. 1178–1182.

Jingting, M., Lentzen, K., Honsdorf, J., Feng, L., Erdt, M., 2016. Statistical shape modeling from Gaussian distributed incomplete data for image segmentation. In: Clinical Image-Based Procedures. Translational Research in Medical Imaging. Springer International Publishing, Cham, pp. 113–121.

Kaelbling, L.P., Littman, M.L., Moore, A.W., 1996. Reinforcement learning: A survey. CoRR cs.AI/9605103.

Kang, S.H., Jeon, K., Kang, S.H., Lee, S.H., 2021. 3D cephalometric landmark detection by multiple stage deep reinforcement learning. Sci. Rep. 11 (1), 17509.

Kerr, D., Scotney, B., Coleman, S., 2008. Interest point detection on incomplete images. In: 2008 15th IEEE International Conference on Image Processing. pp. 817–820.

Leroy, G., Rueckert, D., Alansary, A., 2020. Communicative reinforcement learning agents for landmark detection in brain images. In: Machine Learning in Clinical Neuroimaging and Radiogenomics in Neuro-Oncology. Springer, pp. 177–186.

Li, L., Wu, F., Yang, G., Xu, L., Wong, T., Mohiaddin, R., Firmin, D., Keegan, J., Zhuang, X., 2020. Atrial scar quantification via multi-scale CNN in the graph-cuts framework. Med. Image Anal. 60, 101595.

Li, S., Xiao, J., He, L., Peng, X., Yuan, X., 2019. The tumor target segmentation of nasopharyngeal cancer in CT images based on deep learning methods. Technol. Cancer Res. Treat. 18, 1533033819884561.

Liao, R., Miao, S., de Tournemire, P., Grbic, S., Kamen, A., Mansi, T., Comaniciu, D., 2017. An artificial agent for robust image registration. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 31, No. 1.

Lv, J., Shao, X., Xing, J., Cheng, C., Zhou, X., 2017. A deep regression architecture with two-stage re-initialization for high performance facial landmark detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3317–3326.

Ma, J., Lin, F., Wesarg, S., Erdt, M., 2018. A novel Bayesian model incorporating deep neural network and statistical shape model for pancreas segmentation. In: Medical Image Computing and Computer Assisted Intervention. MICCAI 2018, Springer International Publishing, Cham, pp. 480–487.

Maicas, G., Carneiro, G., Bradley, A.P., Nascimento, J.C., Reid, I., 2017. Deep reinforcement learning for active breast lesion detection from DCE-mri. In: Medical Image Computing and Computer Assisted Intervention. MICCAI 2017, Springer International Publishing, Cham, pp. 665–673.

Meng, D., Keller, M., Boyer, E., Black, M., Pujades, S., 2020. Learning a statistical full spine model from partial observations. In: Shape in Medical Imaging. Springer International Publishing, Cham, pp. 122–133.

Park, S., Kim, H., Shim, E., Hwang, B.Y., Kim, Y., Lee, J.W., Seo, H., 2022. Deep learning-based automatic segmentation of mandible and maxilla in multi-center CT images. Appl. Sci. 12 (3).

Payer, C., Štern, D., Bischof, H., Urschler, M., 2016. Regressing heatmaps for multiple landmark localization using CNNs. In: Medical Image Computing and Computer-Assisted Intervention. MICCAI 2016, pp. 230–238.

Qian, W., Zhang, M., Wan, K., Xie, Y., Du, S., Li, J., Mu, X., Qiu, J., Xue, X., Zhuang, X., Wu, Y., Liu, F., Wang, S., 2022. Genetic evidence for facial variation being a composite phenotype of cranial variation and facial soft tissue thickness. J. Genet. Genom. 49 (10), 934–942.

Raju, A., Miao, S., Jin, D., Lu, L., Huang, J., Harrison, A.P., 2022. Deep implicit statistical shape models for 3D medical image delineation. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36, No. 2. pp. 2135–2143.

Sagonas, C., Antonakos, E., Tzimiropoulos, G., Zafeiriou, S., Pantic, M., 2016. 300 Faces in-the-wild challenge: database and results. Image Vis. Comput. 47, 3–18, 300-W, the First Automatic Facial Landmark Detection in-the-Wild Challenge.

Seghers, D., Slagmolen, P., Lambelin, Y., Hermans, J., Loeckx, D., Maes, F., Suetens, P., 2007. Landmark based liver segmentation using local shape and local intensity models. In: Proc. Workshop of the 10th Int. Conf. on MICCAI, Workshop on 3D Segmentation in the Clinic: A Grand Challenge. Citeseer, pp. 135–142.

Shepherd, J., Ng, B., Fan, B., Schwartz, A., Cawthon, P., Cummings, S., Kritchevsky, S., Nevitt, M., Santanasto, A., Cootes, T., 2017. Modeling the shape and composition of the human body using dual energy X-ray absorptiometry images. PLoS One 12, e0175857.

Staib, L.H., Duncan, J.S., 1996. Model-based deformable surface finding for medical images. IEEE Trans. Med. Imaging 15 5, 720–731.

Stasinski, R., Konrad, J., 2000. POCS-based image reconstruction from irregularly-spaced samples. In: Proceedings 2000 International Conference on Image Processing, Vol. 2. pp. 315–318.

Sun, Y., Wang, X., Tang, X., 2013. Deep convolutional network cascade for facial point detection. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition. pp. 3476–3483.

Terada, T., Chen, Y.W., Kimura, R., 2018. 3D facial landmark detection using deep convolutional neural networks. In: 2018 14th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery. ICNC-FSKD, pp. 390–393.

Vazquez, C., Dubois, E., Konrad, J., 2002. Reconstruction of irregularly-sampled images by regularization in spline spaces. In: Proceedings. International Conference on Image Processing, Vol. 3. p. III.

Vazquez, C., Konrad, J., Dubois, E., 2000. Wavelet-based reconstruction of irregularly-sampled images: application to stereo imaging. In: Proceedings 2000 International Conference on Image Processing, Vol. 2. pp. 319–322.

Vlontzos, A., Alansary, A., Kamnitsas, K., Rueckert, D., Kainz, B., 2019. Multiple landmark detection using multi-agent reinforcement learning. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 262–270.

Wang, P., Zhou, Q., Zhao, Y., Zhao, S., Ma, J., 2020. A robust facial landmark detection in uncontrolled natural condition. J. Phys. Conf. Ser. 1693 (1), 012202.

Yang, H., Patras, I., 2013. Privileged information-based conditional regression forest for facial feature detection. In: 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition. FG, pp. 1–6.

Yang, H., Xie, X., 2020. An actor-critic deep reinforcement learning approach for transmission scheduling in cognitive internet of things systems. IEEE Syst. J. 14 (1), 51–60.

Zhan, Y., Peng, Z., Hermosillo, G., Zhou, X., 2016. Chapter 3 - robust multi-landmark detection based on information theoretic scheduling. In: Medical Image Recognition, Segmentation and Parsing. In: The Elsevier and MICCAI Society Book Series, Academic Press, pp. 45–70.

Zhang, M., Xu, J., Abaci Turk, E., Grant, P.E., Golland, P., Adalsteinsson, E., 2020. Enhanced detection of fetal pose in 3D MRI by deep reinforcement learning with physical structure priors on anatomy. In: Medical Image Computing and Computer Assisted Intervention. MICCAI 2020, Springer International Publishing, Cham, pp. 396–405.

Zheng, Y., Liu, D., Georgescu, B., Nguyen, H., Comaniciu, D., 2015. 3D deep learning for efficient and robust landmark detection in volumetric data. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 565–572.

Zhou, E., Fan, H., Cao, Z., Jiang, Y., Yin, Q., 2013. Extensive facial landmark localization with coarse-to-fine convolutional network cascade. In: 2013 IEEE International Conference on Computer Vision Workshops. pp. 386–391.

Zhou, S.K., Le, H.N., Luu, K., V Nguyen, H., Ayache, N., 2021. Deep reinforcement learning in medical imaging: A literature review. Med. Image Anal. 73, 102193.

Zhuang, X., 2016. Multivariate mixture model for cardiac segmentation from multi-sequence MRI. In: Medical Image Computing and Computer-Assisted Intervention. MICCAI 2016, Springer International Publishing, Cham, pp. 581–588.

Zhuang, X., 2019. Multivariate mixture model for myocardial segmentation combining multi-source images. IEEE Trans. Pattern Anal. Mach. Intell. 41 (12), 2933–2946.

Zhuang, X., Li, L., Payer, C., Štern, D., Urschler, M., Heinrich, M.P., Oster, J., Wang, C., Smedby, Ö., Bian, C., et al., 2019. Evaluation of algorithms for multi-modality whole heart segmentation: an open-access grand challenge. Med. Image Anal. 58, 101537.

Zhuang, X., Rhode, K.S., Razavi, R.S., Hawkes, D.J., Ourselin, S., 2010. A registration-based propagation framework for automatic whole heart segmentation of cardiac MRI. IEEE Trans. Med. Imaging 29 (9), 1612–1625.

Zhuang, X., Shen, J., 2016. Multi-scale patch and multi-modality atlases for whole heart segmentation of MRI. Med. Image Anal. 31, 77–87.

Zhuang, X., Xu, J., Luo, X., Chen, C., Ouyang, C., Rueckert, D., Campello, V.M., Lekadir, K., Vesal, S., RaviKumar, N., et al., 2022. Cardiac segmentation on late gadolinium enhancement MRI: a benchmark study from multi-sequence cardiac MR segmentation challenge. Med. Image Anal. 81, 102528.