Learning rewards from exploratory demonstrations using probabilistic temporal ranking

Citation for published version:
Burke, M, Lu, K, Angelov, D, Straizys, A, Innes, C, Subr, K & Ramamoorthy, S 2023, 'Learning rewards from exploratory demonstrations using probabilistic temporal ranking', *Autonomous Robots*, vol. 47, no. 6, pp. 733-751. https://doi.org/10.1007/s10514-023-10120-w

Digital Object Identifier (DOI):
10.1007/s10514-023-10120-w

Published In:
Autonomous Robots

Autonomous Robots (2023) 47:733–751
https://doi.org/10.1007/s10514-023-10120-w

Learning rewards from exploratory demonstrations using probabilistic temporal ranking

Michael Burke[1] · Katie Lu[2] · Daniel Angelov[2] · Artūras Straižys[2] · Craig Innes[2] · Kartic Subr[2] · Subramanian Ramamoorthy[2]

Received: 22 February 2021 / Accepted: 16 June 2023 / Published online: 10 July 2023

Abstract
Informative path-planning is a well established approach to visual-servoing and active viewpoint selection in robotics, but typically assumes that a suitable cost function or goal state is known. This work considers the inverse problem, where the goal of the task is unknown, and a reward function needs to be inferred from exploratory example demonstrations provided by a demonstrator, for use in a downstream informative path-planning policy. Unfortunately, many existing reward inference strategies are unsuited to this class of problems, due to the exploratory nature of the demonstrations. In this paper, we propose an alternative approach to cope with the class of problems where these sub-optimal, exploratory demonstrations occur. We hypothesise that, in tasks which require discovery, successive states of any demonstration are progressively more likely to be associated with a higher reward, and use this hypothesis to generate time-based binary comparison outcomes and infer reward functions that support these ranks, under a probabilistic generative model. We formalise this *probabilistic temporal ranking* approach and show that it improves upon existing approaches to perform reward inference for autonomous ultrasound scanning, a novel application of learning from demonstration in medical imaging while also being of value across a broad range of goal-oriented learning from demonstration tasks.

Keywords Visual servoing · Reward inference · Probabilistic temporal ranking

1 Introduction

Informative path-planning for visual servo control and active viewpoint selection is a key ability for modern day autonomous robotics. However, these approaches typically assume that some notion of a goal or desired viewpoint is available, which may not always be the case. This work considers the case where the goal or cost function of an informative path-planning task is unknown, and needs to be inferred from expert demonstrations. The ability to teach robotic agents using expert demonstration of tasks promises exciting developments across several sectors of industry. This is particularly true of medical imaging, where task and anatomy variability can make it challenging to provide specifications and describe tasks directly, and it may be more natural to consider an apprenticeship learning (Abbeel & Ng, 2004) approach.

Indirect imitation learning (Bagnell, 2015) approaches formulate apprenticeship learning as a search problem within a solution space of plans, where some notional (unknown) *reward function* induces the demonstrated behaviour. A key learning problem is then to estimate this reward function. This *reward inference* approach is commonly known as inverse reinforcement learning (IRL) (Ng & Russell, 2000).

However, as illustrated in the ultrasound scanning task of Fig. 1, the demonstration process often followed by medical practitioners is a naturally exploratory one, involving an information gathering phase in addition to an optimisation phase, or a generally greedy motion towards a desired viewpoint. Human demonstrators regularly act so as to improve their states in complex tasks. For example, in a study on human approaches to combinatorial optimisation, Murawski and Bossaerts (2016) found that humans behave relatively

✉ Michael Burke
michael.g.burke@monash.edu

Subramanian Ramamoorthy
s.ramamoorthy@ed.ac.uk

1 Electrical and Computer Systems Engineering, Monash University, Clayton, Australia

2 School of Informatics, University of Edinburgh, Edinburgh, UK

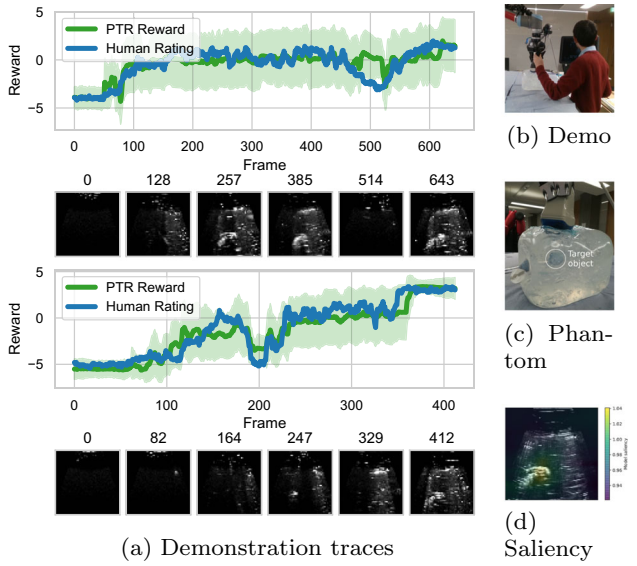


Fig. 1 This work introduces a temporal ranking strategy to learn reward functions **a** from human demonstrations **b** for autonomous ultrasound scanning. Probabilistic temporal ranking can learn to identify non-monotonically increasing rewards from demonstration image sequences containing exploratory actions, and successfully associates ultrasound features corresponding to a target object (**c**) with rewards (**d**)

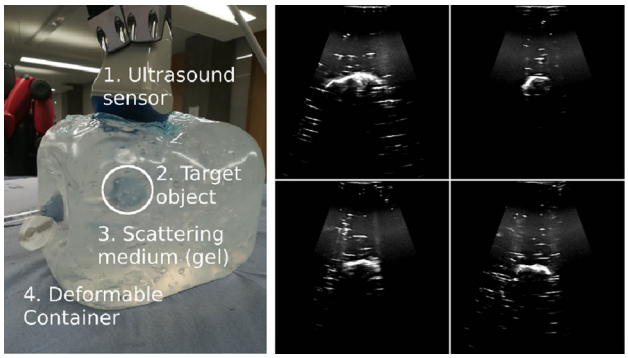


Fig. 2 This work considers the task of learning to search for and capture an image of a target object (2.) suspended in a scattering material (3.) housed within a deformable container (4.). Our goal is to learn a reward signal from demonstrations that allows us to move an ultrasound sensor (1.) to positions that produce clear images of the target object. High quality ultrasound images (right) captured by a human demonstrator show high intensity contour outlines, centre the target object of interest, and generally provide some indication of target object size

greedily, following something akin to a branch-and-bound algorithm for search. Similarly, in palpation experiments, Konstantinova et al. (2013) showed that humans searching for hard excursions in soft tissue searched extensively for nodules, but acted greedily when these had been found, refining the search using small focused, circular movements. Unfortunately, the exploratory nature of demonstrations such as these poses a challenge for many existing approaches to reward inference.

For example, the IRL approach to apprenticeship learning (Abbeel & Ng, 2004) aims to match the frequency (counts) of features encountered in the learner's behaviour with those observed in demonstrations. This technique provides necessary and sufficient conditions when the reward function is linear in the features encoding execution states, but results in ambiguities in associating optimal policies with reward functions or feature counts. An elegant reformulation of this using the principle of maximum entropy resolves ambiguities and results in a single optimal stochastic policy. Methods for maximum-entropy IRL (Ziebart et al., 2008; Wulfmeier et al., 2015; Levine et al., 2011) identify reward functions using maximum likelihood estimation, typically under the assumption that the probability of seeing a given trajectory is proportional to the exponential of the total reward along a given path. Unfortunately, these methods are fundamentally frequentist and thus struggle to cope with *repetitive sub-optimal demonstrations*, as they assume that frequent appearance implies relevance. i.e. If a feature is seen repeatedly across demonstration trajectories, it is deemed valuable, as are policies that result in observations of these features.

This makes these approaches unsuitable for a broad class of tasks that require *exploratory actions* or environment identification during demonstration. e.g. an expert using an ultrasound scan to locate a tumour (Fig. 1). Obtaining useful ultrasound images requires contact with a deformable body (see Fig. 2) at an appropriate position and contact force, with image quality affected by the amount of ultrasound gel between the body and the probe, and air pockets that obscure object detection. This means that human demonstrations are frequently and inherently sub-optimal, requiring that a demonstrator actively search for target objects, while attempting to locate a good viewpoint position and appropriate contact force. This class of demonstration violates many of the assumptions behind existing reward inference schemes.

In order to address this, this paper introduces probabilistic temporal ranking (PTR), a temporal ranking model of reward that addresses these limitations. PTR is a self-supervised approach and thus does not rely on the true reward or value function, using only a sequence of images or states to infer underlying reward functions.

Instead of assigning reward based on the maximum entropy model, PTR attributes reward using a ranking model. Here, we assume that, in general, an expert acts to improve their current state. This means that it is likely that observations at a later stage in a demonstrated trajectory are more important than those seen at an earlier stage. PTR uses this fact to generate time-based binary comparison outcomes, and then uses these to infer reward functions that support these ranks, under a probabilistic generative model that combines information about image or observation similarity (a Gaus-

sian process reward model over a learned latent space) with a noisy pairwise generative outcome model.

Importantly, PTR is able to handle cases where this temporal improvement is unsteady and non-monotonic, with intermediate performance dips. Experimental results show that probabilistic temporal ranking successfully recovers reward maps from demonstrations in tasks requiring significant levels of exploration alongside exploitation (where maximum entropy IRL fails), and obtains similar performance to maximum entropy inverse IRL when optimal demonstrations are available.

PTR is not only useful for active viewpoint selection problems, but the assumptions governing generally increasing rewards hold for a broad class of goal-oriented control problems, for example, swinging up a pendulum, or navigating to a specific location.

We highlight this through a number of simulated experiments, and illustrate the value of our approach in a challenging ultrasound scanning application, where demonstrations inherently contain a searching process, and show that we can train a model to find a tumour-like mass in an imaging phantom.[1] Ultrasound imaging is a safe and low cost sensing modality of significant promise for surgical robotics, and is already frequently used for autonomous needle steering and tracking (Liang et al., 2010; Chatelain et al., 2013). Chatelain et al. (2015) propose the use of ultrasound quality maps to improve image quality in robotic ultrasound scanning applications. Autonomous visual servoing systems have also been proposed in support of teleoperated ultrasound diagnosis (Abolmaesumi et al., 2002; Li et al., 2012), but these techniques tend to rely on hand designed anatomical target detectors or confidence maps. The scanner introduced in this work is fully autonomous, and relies entirely on a reward signal learned from demonstration, in what we believe is a first for medical imaging. Importantly, the probabilistic temporal ranking formulation provides more signal for learning, as a greater number of comparisons can be generated from each demonstration trajectory. This means that we can train a more effective prediction model from pixels than with maximum entropy IRL, which in turn opens up a number of avenues towards self-supervised learning for medical imaging and diagnosis.

In summary, the primary contributions of this paper are

- a temporal ranking reward model that allows for reward inference from sub-optimal, high dimensional exploratory demonstrations, and
- a method for autonomous ultrasound scanning using image sequence demonstrations.

[1] An imaging phantom is an object that mimics the physical responses of biological tissue, and is commonly used in medical imaging to evaluate and analyse imaging devices.

2 Related work

2.1 Reward or cost function inference

As mentioned previously, apprenticeship learning (Abbeel & Ng, 2004) is an alternative to direct methods of imitation learning (Bagnell, 2015) or behaviour cloning, and is currently dominated by indirect approaches making use of maximum entropy assumptions.

Maximum entropy or maximum likelihood inverse reinforcement learning models the probability of a user preference for a given trajectory ζ as proportional to the exponential of the total reward along the path (Ziebart et al., 2008),

$$p(\zeta|r) \propto \exp(\sum_{s,a\in\zeta} r_{s,a}). \tag{1}$$

Here, s denotes a state, a an action, and $r_{s,a}$ the reward obtained for taking an action in a given state. It is clear that this reward model can be maximised by any number of reward functions. Levine et al. (2011) use a Gaussian process prior to constrain the reward, while Wulfmeier et al. (2015) backpropagate directly through the reward function using a deep neural network prior. Maximum entropy inverse reinforcement learning approaches are typically framed as iterative policy search, where policies are identified to maximise the reward model. This allows for the incorporation of additional policy constraints and inductive biases towards desirable behaviours, as in relative entropy search (Boularias et al., 2011), which uses a relative entropy term to keep policies near a baseline, while maximising reward feature counts. Maximum entropy policies can also be obtained directly, bypassing reward inference stages, using adversarial imitation learning (Ho & Ermon, 2016; Finn et al., 2016; Fu et al., 2018; Ghasemipour et al., 2019), although reward prediction is itself useful for medical imaging applications.

Although maximum entropy IRL is ubiquitous, alternative reward models have been proposed. For example, Angelov et al. (2020) train a neural reward model using demonstration sequences, to schedule high level policies in long horizon tasks. Here, they capture overhead scene images, and train a network to predict a number between 0 and 1, assigned in increasing order to each image in a demonstration sequence. This ranking approach is similar to the pairwise ranking method we propose, but, as will be shown in later results, is limited by its rigid assumption of linearly increasing reward. Majumdar et al. (2017) propose flexible reward models that explicitly account for human risk sensitivity. Time contrastive networks (Sermanet et al., 2018) learn disentangled latent representations of video using time as a supervisory signal. Here, time synchronised images taken from multiple viewpoints are used to learn a latent embedding space where similar images (captured from different

viewpoints) are close to one another. This embedding space can then be used to find policies from demonstrations. Time contrastive networks use a triplet ranking loss, and are trained using positive and negative samples (based on frame timing margins).

Preference-based ranking of this form is widely used in inverse reinforcement learning to rate demonstrations (Wirth et al., 2017), and preference elicitation (Braziunas & Boutilier, 2006) is a well established area of research. For example, Brochu et al. (2010) use Bayesian optimisation with a pairwise ranking model to allow users to procedurally generate realistic animations. Lopes et al. (2009); Biyik et al. (2020) and Tucker et al. (2020) actively query demonstrators to learn reward functions. The latter make use of a Thurstonian model (Thurstone, 2017), computing the cumulative distribution function over the difference between rewards. This is intractable, and requires a Laplace approximation for inference. Sugiyama et al. (2012) use preference-based inverse reinforcement learning for dialog control. Here, dialog samples are annotated with ratings, which are used to train a preference-based reward model. These preference elicitation approaches are effective, but place a substantial labelling burden on users. In this work, we consider the non-interactive learning case where we are required to learn directly from unlabelled observation traces.

A number of extensions to maximum entropy IRL have been proposed to cope with sub-optimal demonstration sequences, typically through the inclusion of additional supervisory information about the quality of a demonstration sequence (Wu et al., 2019; Brown et al., 2019). In large part, these works define demonstration quality in terms of how noisy they are (Brown et al., 2019) or how far they deviate from some perfect demonstration. However, for discovery tasks such as ultrasound scanning, it is much harder to determine what constitutes an optimal or perfect demonstration, as all demonstrations require a degree of exploration before finding a good viewpoint.

Lee et al. (2016) use leveraged Gaussian processes to learn from both positive and negative demonstration examples. Similarly, Shiarlis et al. (2016) and Valko et al. (2013) consider inverse reinforcement learning in the looser case where only a subset of demonstrations are considered expert or successful, and the remaining 'failures' may contain elements key to success, or even be unlabelled successful demonstrations. These are semi-supervised learning approaches, as they rely on additional labelling information about the quality of demonstrations. In contrast, the PTR approach proposed in this paper is self-supervising, as it relies on time as a supervisory signal.

Brown et al. (2020) make use of a preference ranking approach to improve robot policies through artificial trajectory ranking using increasing levels of injected noise. Unlike (Brown et al., 2020), which uses preference ranking over trajectories, our work uses preference ranking within trajectories, under the assumption that a demonstrator generally acts to improve or maintain their current state. We modify a Bayesian image ranking model (Burke et al., 2017) that accounts for potential uncertainty in this assumption, and is less restrictive than the linearly increasing model of Angelov et al. (2020). Bayesian ranking models (Chu & Ghahramani, 2005) are common in other fields – for example, TrueSkill^{TM} (Herbrich et al., 2007) is widely used for player performance modelling in online gaming settings, but has also been applied to to train image-based style classifiers in fashion applications (Kiapour et al., 2014) and to predict the perceived safety of street scenes using binary answers to the question "Which place looks safer?" (Naik et al., 2014).

2.2 Active viewpoint selection

Given an appropriate reward model, autonomous ultrasound scanning requires a policy that balances both exploration and exploitation for active viewpoint selection or informative path planning. Research on active viewpoint selection (Sridharan et al., 2010) is concerned with agents that choose viewpoints which optimise the quality of the visual information they sense. Similarly, informative path planning involves an agent choosing actions that lead to observations which most decrease uncertainty in a model. Gaussian processes (GP) are frequently used for informative path planning because of their inclusion of uncertainty, data-efficiency, and flexibility as non-parametric models.

Binney and Sukhatme (2012) use GPs with a branch and bound algorithm, while (Cho et al., 2018) perform informative path planning using GP regression and a mutual information action selection criterion. More general applications of GPs to control include PILCO (Deisenroth & Rasmussen, 2011), where models are optimised to learn policies for reinforcement learning control tasks, and the work of Ling et al. (2016), which introduces a GP planning framework that uses GP predictions in H-stage Bellman equations.

These Bayesian optimisation schemes are well established methods for optimisation of an unknown function, and have been applied to many problems in robotics including policy search (Martinez-Cantin, 2017), object grasping (Yi et al., 2016), and bipedal locomotion (Calandra et al., 2014).

By generating policies dependent on predictions for both reward value and model uncertainty, Bayesian optimisation provides a mechanism for making control decisions that can both progress towards some task objective and acquire information to reduce uncertainty. GP's and Bayesian optimisation are often used together, with a GP acting as the surrogate model for a Bayesian optimisation planner, as in the mobile robot path planning approaches of Martinez-Cantin et al. (2009) and Marchant et al. (2014). Our work takes a similar approach, using GP-based Bayesian optimisation for

path planning in conjunction with the proposed observation ranking reward model.

The combination of preference-based learning with a policy trading-off exploration-exploitation is commonly studied within duelling bandit frameworks (Sui et al., 2017). Here, instead of learning from a reward signal, a policy is required to learn directly from preferential feedback. This differs from the the reward inference setting studied in this paper, where preference signals are generated a-priori by generating temporal comparisons from human demonstrations, and not online when the policy is deployed.

3 Probabilistic temporal ranking

This paper introduces probabilistic temporal ranking (PTR), a reward inference strategy for high dimensional exploratory image demonstrations. Below, we first describe a fully probabilistic temporal ranking model, which we then examine using a series of simulated experiments. We then introduce a deterministic neural approximation that can be efficiently trained in a fully end-to-end fashion, and is more suited to larger training sets, before moving on to our primary autonomous ultrasound scanning experiments, and the description of a Bayesian optimisation strategy for informative path-planning that facilitates this.

In general, we envisage PTR being used as in Fig. 3. First, a series of observations are collected while a human demonstrates an exploratory visual scanning task. PTR is then used to train a reward model by sampling pairwise temporal comparison outcomes from the demonstration sequences and performing model fitting. This reward model is then used by a suitable informative path-planning or active viewpoint selection policy (in this case Bayesian optimisation) to replicate the demonstration in new environments.

3.1 Fully probabilistic model

This paper incorporates additional assumptions around the structure of demonstration sequences, to allow for improved reward inference. We introduce a reward model that learns from pairwise comparisons sampled from demonstration trajectories. Here we assume that an observation or state seen later in a demonstration trajectory should typically generate greater reward than one seen at an earlier stage.

We build on the pairwise image ranking model of Burke et al. (2017), replacing pre-trained object recognition image features with a latent state, $\mathbf{x}_t \in \mathbf{R}^d$, learned using a convolutional variational autoencoder (CVAE),

$$\mathbf{x}_t \sim \mathcal{N}(\mu(\mathbf{Z}_t), \sigma(\mathbf{Z}_t)), \tag{2}$$

that predicts mean, $\mu(\mathbf{Z}_t) \in \mathbb{R}^d$, and diagonal covariance, $\sigma(\mathbf{Z}_t) \in \mathbb{R}^{d\times d}$, for input observation $\mathbf{Z}_t \in \mathbb{R}^{w\times h}$ captured at time t (assuming image inputs of dimension $w \times h$).

Rewards $r_t \in \mathbb{R}^1$ are modelled using a Gaussian process prior,

$$\begin{bmatrix} r' \\ r_t \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} K(\mathbf{X}', \mathbf{X}') + \Sigma_n & K(\mathbf{X}', \mathbf{x}_t) \\ K(\mathbf{x}_t, \mathbf{X}') & K(\mathbf{x}_t, \mathbf{x}_t) \end{bmatrix}\right). \tag{3}$$

Here, we use $\mathbf{X}'$ and r' to denote states and reward pairs corresponding to training observations. $\mathbf{X}' \in \mathbb{R}^{N\times d}$ is a matrix formed by vertically stacking N latent training states, and $K(\mathbf{X}', \mathbf{X}')$ a covariance matrix formed by evaluating a Matern32 kernel function

$$k(\mathbf{x}_t, \mathbf{y}_t) = \text{Matern32}(\mathbf{x}_t, \mathbf{y}_t, l), \quad \mathbf{x}_t, \mathbf{y}_t \in \mathbb{R}^d \tag{4}$$

for all possible combinations of latent state pairs $\mathbf{x}_t, \mathbf{y}_t$, sampled from the rows of $\mathbf{X}'$. $l \in \mathbb{R}^1$ is a length scale parameter with a Gamma distributed prior, $l \sim \Gamma(\alpha = 2.0, \beta = 0.5)$, and $\Sigma_n \in \mathbb{R}^{N\times N}$ is a diagonal heteroscedastic noise covariance matrix, with diagonal elements drawn from a Half Cauchy prior, $\Sigma_n \sim \text{HalfCauchy}(\beta = 1.0)$.

These priors are well calibrated to the inference task here, and should not need to be adjusted in other applications. A half Cauchy prior ($\beta = 1$) is a heavy tailed distribution that allows for diagonal covariance parameters (see Appendix, Fig. 14), favouring low noise rewards, but also allowing for higher noise if needed. Inference with this prior is thus capable of handling both noisy and more repeatable rewards. Decreasing beta would increase the prior probability of little variability in rewards for a given state.

Similarly, the Gamma distributed length scale prior places most probability mass over a length scale of about 1, but allows a range of values (see Appendix, Fig. 14). Given the standard normal prior used by the variational autoencoder, which compresses observations into a state space roughly constrained within the range $(-3, 3)$, this prior allows for both small local influence between latent states and rewards, or wider correspondences across the latent space if needed.

At prediction time, reward predictions r_t for image observations $\mathbf{Z}_t$ can be made by encoding the image to produce latent state $\mathbf{x}_t$, and conditioning the Gaussian in Equation (3) (Williams & Rasmussen, 2006).

Using this model, the generative process for a pairwise comparison outcome, $g \in \{0, 1\}$, between two input observation rewards r_{t_1} and r_{t_2} at time steps t_1 and t_2, is modelled using a Bernoulli trial over the sigmoid of the difference between the rewards,

$$g \sim \text{Ber}\left(\text{Sig}(r_{t_2} - r_{t_1})\right). \tag{5}$$

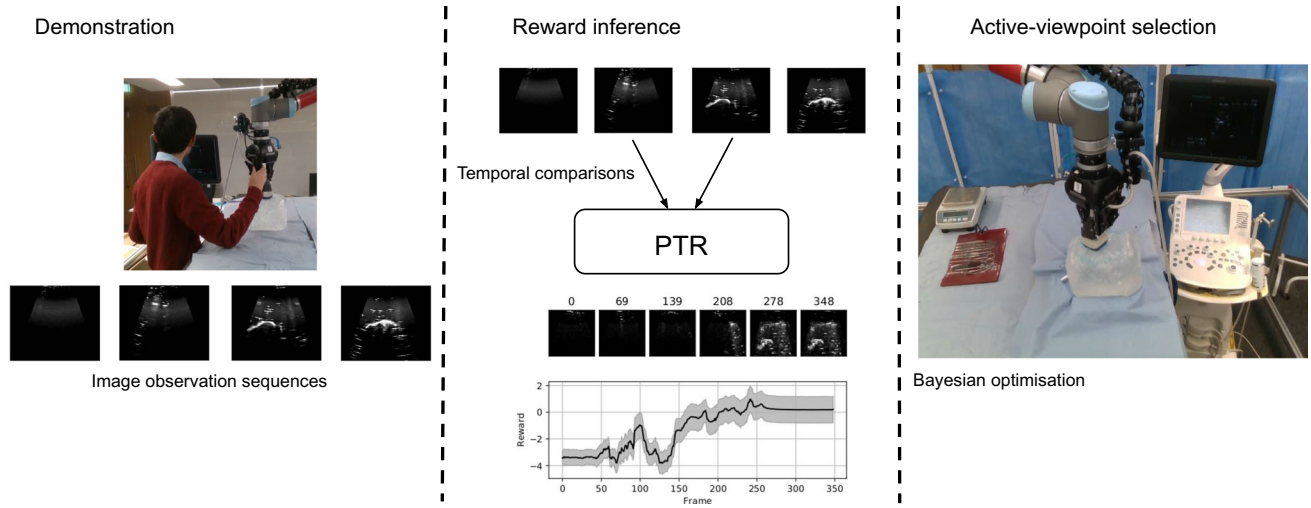


Fig. 3 Pipeline for autonomous ultrasound scanning. User demonstrations are used to collect image sequences, temporal comparisons are then sampled from these sequences and used by PTR to train a reward inference model, which is then used by a Bayesian optimisation policy for active-viewpoint selection in new environments

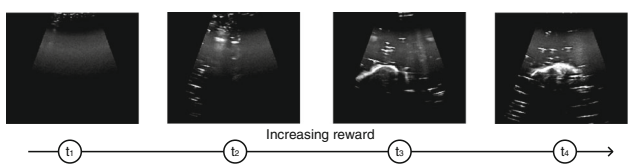


Fig. 4 Time is used as a supervisory signal, by sampling image pairs at times t_i, t_j, and setting $g = 1$ if $t_i > t_j$, $g = 0$ otherwise

This Bernoulli trial introduces slack in the model, allowing for tied or even decreasing rewards to be present in the demonstration sequence.

The Sigmoid used here produces a logit and allows for simple differentiation, which is helpful for approximate inference schemes and the neural approximation introduced below, avoiding the need for the Laplace approximation to the posterior used in Tucker et al. (2020); Biyik et al. (2020).

3.2 Reward inference using temporal observation ranking

The generative model above is fit to demonstration sequences using automatic differentiation variational inference (ADVI) (Kucukelbir et al., 2017) by sampling N observation pairs $\mathbf{Z}_{t_1}, \mathbf{Z}_{t_2}$ from each demonstration sequence, which produce a comparison outcome

$$g = \begin{cases} 1 & \text{if } t_2 \geq t_1 \\ 0 & \text{if } t_1 < t_2 \end{cases} . \tag{6}$$

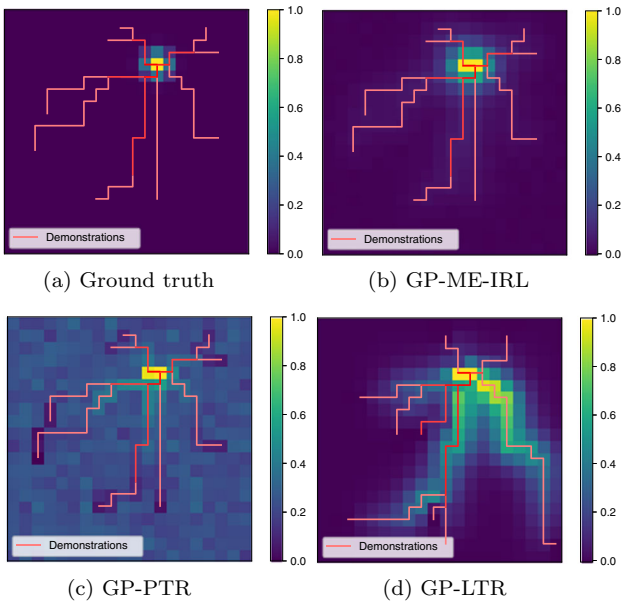
Intuitively, this temporal comparison test, which uses time as a supervisory signal (Fig. 4), operates as follows. Assume that an image captured at time step t_2 has greater reward than an image captured at t_1. This means that the sigmoid of the difference between the rewards is likely to be greater than 0.5, which leads to a higher probability of returning a comparison outcome $g = 1$. Importantly, this Bernoulli trial allows some slack in the model – when the difference between the rewards is closer to 0.5, there is a greater chance that a comparison outcome of $g = 1$ is generated by accident. This means that the proposed ranking model can deal with demonstration trajectories where the reward is non-monotonic. Additional slack in the model is obtained through the heteroscedastic noise model, Σ_n, which also allows for uncertainty in inferred rewards to be modelled.

Inference under this model amounts to using the sampled comparison outcomes from a demonstration trajectory to find rewards that generate similar comparison outcomes, subject to the Gaussian process constraint that images with similar appearance should exhibit similar rewards. After inference, we make reward predictions by encoding an input image, and evaluating the conditional Gaussian process at this latent state.

We briefly illustrate the value of this probabilistic temporal ranking approach in exploratory tasks using two simple grid world experiments.

3.3 Grid world–optimal demonstrations

The first experiment considers a simple grid world, where a Gaussian point attractor is positioned at some unknown location. Our goal is to learn a reward model that allows an agent (capable of moving up, down, left and right) to move



(a) Ground truth (b) GP-ME-IRL

(c) GP-PTR (d) GP-LTR

Fig. 5 Reward inference from optimal demonstrations. Demonstration trajectories are marked in red, and the colour map indicates the reward for each grid position. PTR and ME models have similar relative reward values, and policies trained using these rewards perform near identically. A linearly increasing reward model (LTR) attributes reward more evenly across a demonstration, resulting in sub-optimal policy performance here (Color figure online)

towards the target location. For these experiments, our state is the agent's 2D grid location.

We generate 5 demonstrations (grid positions) from random starting points, across 100 randomised environment configurations with different goal points. We then evaluate performance over 100 trials in each configuration, using a policy obtained through tabular value iteration (VI) using the reward model inferred from the 5 demonstrations. This policy is optimal, as the target location is known, so for all demonstrations the agent moves directly towards the goal, as illustrated for the sample environment configuration depicted in Fig. 5.

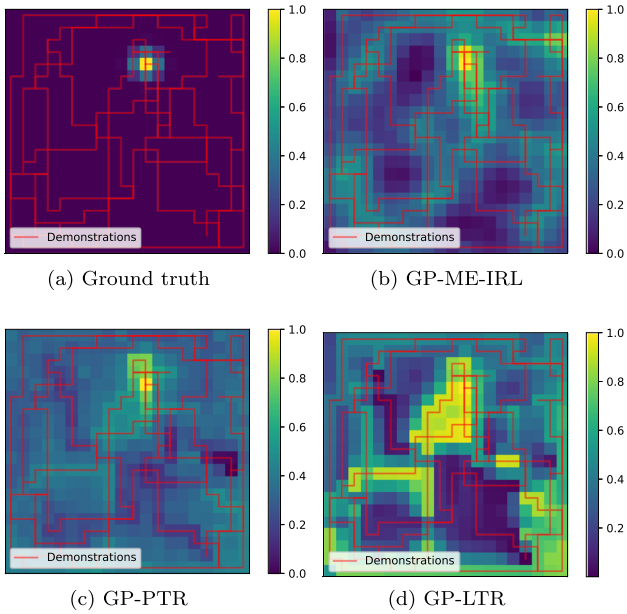
Table 1 shows the averaged total returns obtained for trials in environments when rewards are inferred from optimal demonstrations using the probabilistic temporal ranking[2] (GP-PTR), a Gaussian process maximum entropy approach (Levine et al., 2011) (GP-ME-IRL) and an increasing linear model assumption (Angelov et al., 2020) (GP-LTR). Value iteration is used to find a policy using the mean inferred rewards.

In the optimal demonstration case, policies obtained using both the maximum entropy and probabilistic temporal ranking approach perform equally well, although PTR assigns more neutral rewards to unseen states (Fig. 5). Importantly,

[2] We use PyMC3 (Salvatier et al., 2016) (GP-PTR) to build probabilistic reward models and ADVI (Kucukelbir et al., 2017) for model fitting.

Table 1 Averaged total returns using VI policy trained using inferred reward from optimal demonstrations

	Reward
GP-PTR	9.51 ± 4.92
GP-ME-IRL	9.58 ± 4.90
GP-LTR	7.39 ± 5.72



(a) Ground truth (b) GP-ME-IRL

(c) GP-PTR (d) GP-LTR

Fig. 6 Reward inference from exploratory demonstrations. Demonstration trajectories are marked in red, and the colour map indicates the reward for each grid position. Both the linearly increasing and maximum entropy reward models induce local maxima that result in sub-optimal policies (Color figure online)

as the proposed model is probabilistic, the uncertainty in predicted reward can be used to restrict a policy to regions of greater certainty by performing value iteration using an appropriate acquisition function instead of the mean reward. This implicitly allows for risk-based policies – by weighting uncertainty higher, we could negate the neutrality of the ranking model (risk-averse). Alternatively, we could tune the weighting to actively seek out uncertain regions with perceived high reward (risk-seeking).

3.4 Grid world–exploratory demonstrations

Our second experiment uses demonstrations that are provided by an agent that first needs to explore the environment, before exploiting it. Here, we use a Gaussian process model predictive control policy (see below) to generate demonstrations, and repeat the experiments above. As shown in Fig. 6, this policy may need to cover a substantial portion of the environment before locating the target.

Table 2 shows the averaged total returns obtained for trials in environments when rewards are inferred from exploratory demonstrations using probabilistic temporal ranking, the

Table 2 Averaged total returns using VI policy trained using inferred reward from exploratory demonstrations

	Reward
GP-PTR	7.42 ± 4.82
GP-ME-IRL	3.31 ± 4.24
GP-LTR	2.77 ± 4.30
T-REX	0.42 ± 1.59
D-REX	0.49 ± 2.10

Gaussian process maximum entropy approach and the linearly increasing reward assumption. Here, value iteration (VI) is used to find the optimal policy using the inferred rewards. A comparison with T-REX (Brown et al., 2019) and D-REX (Brown et al., 2020), trajectory ranking methods designed to learn reward functions from sub-optimal demonstrations are also included. It should be noted that T-REX is a supervised learning method, relying on additional labelling information about the quality of a demonstration, while PTR requires no information beyond the demonstration sequences. For these experiments, we generate labelling information for T-REX by using trajectory length as a rough heuristic for the quality of a demonstration. D-REX generates ranked trajectories by artificially injecting noise to demonstrations.

In this sub-optimal exploratory demonstration case, policies obtained using the maximum entropy approach regularly fail, while the probabilistic temporal ranking continues to perform relatively well. Figure 6 shows a sample environment used for testing. The sub-optimal behaviour of the exploring model predictive control policies used for demonstration can result in frequent visits to undesirable states, which leads to incorrect reward attribution under a maximum entropy model. Probabilistic temporal ranking avoids this by using the looser assumption that states generally improve over time. T-REX performs extremely poorly here, as it is unable to learn from the limited number of demonstrations provided. D-REX also fails here, as it is unable to separate the uninformative exploratory portions of the demonstrating from the final exploitative portion of the demonstration policy. While D-REX works well for demonstrations that are sub-optimal due to noise, in this case of exploratory goal oriented tasks, the assumptions made by PTR are more applicable.

Figure 7 shows the performance of PTR as trajectories become more exploratory. Here, 100 demonstrations were generated for a single environment and sorted by length. Reward models were then learned using subsets of 10 demonstrations of increasing length. Policies were trained to maximise these reward functions using value iteration. GP-ME-IRL rapidly degrades as trajectories become more exploratory. T-REX performs poorly here, as it needs both good and bad examples to learn a reward function. D-REX

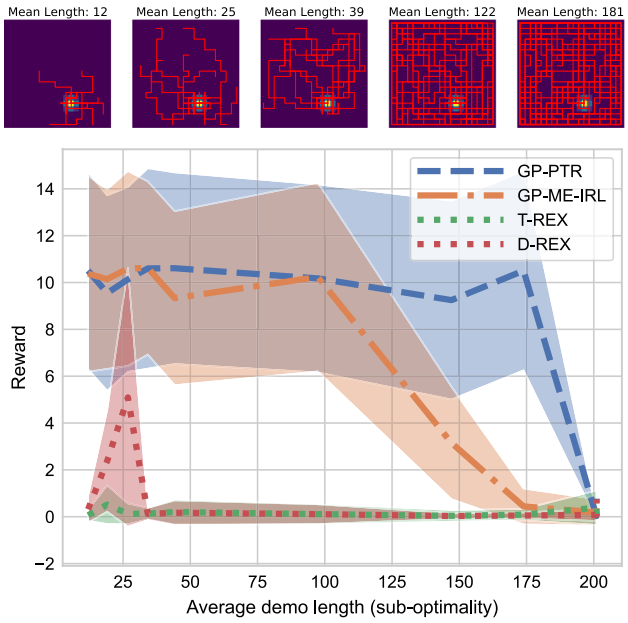


Fig. 7 Policy returns using rewards learned with trajectories of increasing length show the degradation of GP-ME-IRL as trajectories become more exploratory. T-REX performs poorly here, as it needs both good and bad examples to learn a reward function. PTR performs well for both optimal and exploratory demonstrations

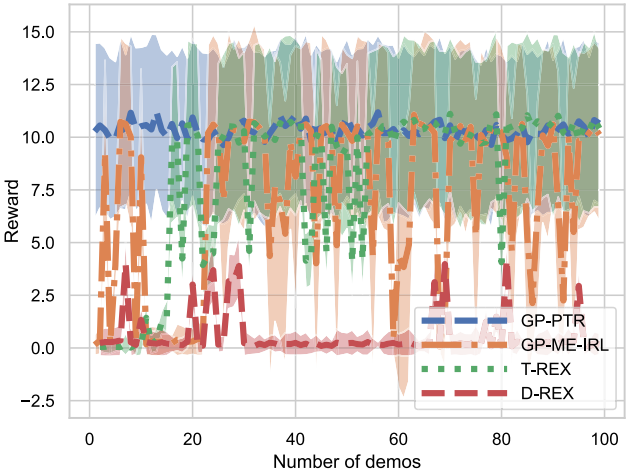


Fig. 8 Policy returns using rewards learned with increasing numbers of demonstration trajectories of increasing length. T-REX starts to perform better with more demonstration data, while GP-ME-IRL is highly dependent on the quality of demonstrations. PTR performs well even with a limited number of demonstration sequences

is unable to handle the exploratory trajectories. In contrast, PTR performs well for both optimal and exploratory demonstrations in these goal-oriented environments, failing only when demonstrations never reach the goal.

Figure 8 shows the rewards obtained by policies trained using rewards learned with increasing numbers of demonstrations. T-REX performs substantially better with more demonstration data and a good balance of optimal and

exploratory trajectories, but struggles to learn from limited data. GP-ME-IRL and D-REX, which make similar assumptions about the reward, are highly dependent on the quality of demonstrations and thus extremely unreliable in this setting. In contrast, PTR performs well even with a limited number of demonstration sequences.

3.5 A deterministic neural approximation to PTR

Given that learning from demonstration typically aims to require only a few trials, numerical inference under the fully Bayesian generative model described above is tractable, particularly if a sparse Gaussian process prior is used. However, in the case where a greater number of demonstrations or comparisons is available, we can approximate the fully probabilistic PTR model above with a deterministic model that can be trained in an end-to-end fashion, using the architecture in Fig. 9. Here, we replace the Gaussian process with a wide single layer fully connected network (FCN), $r_\psi(\mathbf{x})$, with parameters ψ, since single layer FCN's with i.i.d Gaussian weights are known to approximate a sample from a Gaussian processes (Neal, 1996) as model width tends to infinity. This approximation is trained by minimising a binary cross entropy loss over the expected comparison outcome alongside a variational autoencoder (VAE) objective,

$$\begin{aligned}\mathcal{LL} = &-\mathbb{E}_{\mathbf{x}_{t_1}\sim q_\theta}\left[\log p_\phi(\mathbf{Z}_{t_1}|\mathbf{x})\right] + \mathbb{KL}\left(q_\theta(\mathbf{x}|\mathbf{Z}_{t_1})||p(\mathbf{x})\right)\\ &-\mathbb{E}_{\mathbf{x}_{t_2}\sim q_\theta}\left[\log p_\phi(\mathbf{Z}_{t_2}|\mathbf{x})\right] + \mathbb{KL}\left(q_\theta(\mathbf{x}|\mathbf{Z}_{t_2})||p(\mathbf{x})\right)\\ &-\frac{1}{N}\sum_{i=1}^{N}\left[g_i\log\left(h(g_i)\right) + (1-g_i)\log\left(1-h(g_i)\right)\right]\end{aligned} \tag{7}$$

using stochastic gradient descent. Here, $\mathcal{LL}$ denotes the overall loss, $p(\mathbf{x})$ is a standard normal prior over the latent space, $q_\theta(\mathbf{x}|\mathbf{Z}_{t_i})$ denotes the variational encoder, with parameters θ, $p_\phi(\mathbf{Z}_{t_i}|\mathbf{x})$ represents the variational decoder, with parameters ϕ, and $h(g)$ is the comparison output logit (sigmoid). g_i is a temporal comparison outcome label, and $\mathbf{Z}_{t_i}$ denotes a training sample image, with $\mathbf{x}_t$ a sample from the latent space. Weight sharing is used for both the convolutional VAEs and FCNs.

Once trained, the reward model is provided by encoding the input observation, and then predicting the reward using the FCN. This allows for rapid end-to-end training using larger datasets and gives us the ability to backpropagate the comparison supervisory signal through the autoencoder, potentially allowing for improved feature extraction in support of reward modelling. However, this comes at the expense of uncertainty quantification, which is potentially useful for the design of risk-averse policies that need to avoid regions of uncertainty. We investigate these trade-offs and the efficacy of the approximate model in more general reinforcement learning environments below, and in the context of autonomous ultrasound scanning in Sect. 4.

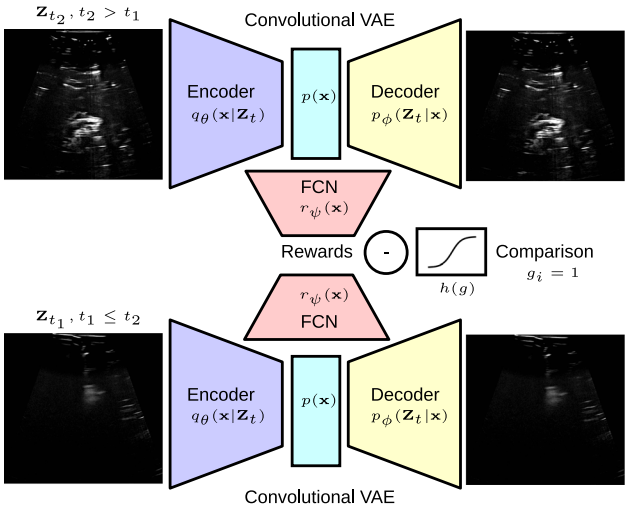


Fig. 9 Neural PTR approximation. Sampled images are auto-encoded, and a reward network predicts corresponding rewards, the sigmoid of the difference between these reward produces a comparison outcome probability. Weight sharing is indicated by colour. The network is trained jointly using a joint variational autoencoder and binary cross entropy loss (Color figure online)

3.6 General reinforcement learning environments

It should be noted that while reward inference using probabilistic temporal ranking is capable of handling sub-optimal exploratory demonstrations in goal-oriented environments, this assumption does not hold for more general tasks. To illustrate this, we applied the deterministic neural PTR approximation above (ML-PTR) to a range of low dimensional continuous control environments[3] (Brockman et al., 2016). Here, we collect 100 demonstrations from an agent trained using proximal policy optimisation (PPO) (Schulman et al., 2017), and infer rewards using ML-PTR. We then use the inferred reward function to train a PPO agent. We benchmark against AIRL (Fu et al., 2018) and GAIL (Ho & Ermon, 2016), popular imitation learning approaches for lower dimensional control tasks. We use the imitation toolbox (Gleave et al., 2022) and Stable Baselines3 (Raffin et al., 2021) for these experiments.

As shown in Table 3, the policy trained using the ML-PTR reward performs best on goal oriented tasks (learning to balance Pendulum-v1 and swing up Acrobot-v1), but fails on the continuous control tasks (Ant-v3, Hopper-v3, HalfCheetah-v3), where the assumption of generally increasing reward over a demonstration does not hold.

[3] No autoencoder is used, as the regularisation provided is unnecessary for these low dimensional control problems.

Table 3 Returns after training for 500000 environment steps. 100 demonstrations are used for reward learning or imitation learning, and the best result across 3 seeds reported (100 test episodes). As expected, PTR performs well on goal-oriented tasks, but fails elsewhere

Method	Pendulum-v1 Reward (Mean ± Std)	Hopper-v3 Reward (Mean ± Std)	HalfCheetah-v3 Reward (Mean ± Std)
Expert (PPO)	−150.4354 ± 88.1962	3437.0038 ± 8.1665	1193.6989 ± 62.7869
ML-PTR (PPO)	**−189.2376 ± 112.5798**	183.4452 ± 1.8466	130.8885 ± 46.1780
GAIL	−315.1454 ± 193.8659	**3415.7294 ± 1.9907**	**1220.1896 ± 92.7100**
AIRL	−948.4833 ± 107.4486	5.7232 ± 0.030	1000.2273 ± 63.5880

Method	Acrobot-v1 Reward (Mean ± Std)	Ant-v3 Reward (Mean ± Std)
Expert (PPO)	−73.8100 ± 10.0297	800.8259 ± 85.1205
ML-PTR (PPO)	−81.5600 ± 21.4487	−1558.8640 ± 2.1768
GAIL	−292.8000 ± 161.7387	**886.3507 ± 25.8765**
AIRL	**−77.8600 ± 19.3111**	−15.4500 ± 21.4068

Best performing approach is given in bold

4 Autonomous ultrasound scanning

For our primary experiment, we demonstrate the use of probabilistic temporal ranking in a challenging ultrasound scanning application. Here, we capture 10 kinesthetic demonstrations of a search for a target object using a compliant manipulator, and use only ultrasound image sequences (2D trapezoidal cross-sectional scans) to learn a reward model. Our goal is to use this reward model within a control policy that automatically searches for and captures the best image of a tumour-like mass[4] suspended within a deformable imaging phantom constructed using a soft plastic casing filled with ultrasound gel.

This task is difficult because it involves a highly uncertain and dynamic domain. Obtaining stable ultrasound images requires contact with a deformable imaging phantom at an appropriate position and contact force, with image quality affected by the thickness of the ultrasound gel between the phantom and the probe, while air pockets within the phantom object can obscure object detection. Moreover, since the phantom deforms, air pockets and gel can move in response to manipulator contact. This means that kinesthetic demonstrations are inherently sub-optimal and exploratory, as they require that a demonstrator actively search for target objects, while attempting to locate a good viewpoint position and appropriate contact force. As in real-world medical imaging scenarios, the demonstrator is unable to see through the phantom object from above, so demonstrations are based entirely on visual feedback from an ultrasound monitor.

[4] A roughly 30 mm x 20 mm blob of Blu tack original in a container of dimensions 200 mm x 150 mm x 150 mm.

4.1 Active viewpoint selection

Although the proposed reward model can be used with any policy, we demonstrate its use by means of a Bayesian optimisation policy, selecting action $\hat{\mathbf{a}}_t$ that drives an agent to a desired end-effector position $\hat{\mathbf{s}}_t$, drawn from a set of possible states (a volume of acceptable end-effector positions) using an upper confidence bound objective function that seeks to trade off expected reward returns against information gain or uncertainty reduction.

Here, we learn a mapping between reward and end-effector positions using a surrogate Gaussian process model with a radial basis function kernel,

$$r_t \approx \mathcal{GP}\left(\mathbf{0}, \mathrm{RBF}(\mathbf{s}_t, l_p)\right). \tag{8}$$

Length scale l_p is determined by maximum likelihood estimation, using a search within the length scale bounds $l_p \in [1\mathrm{e}{-5}, 0.01]\,m$. Fixed measurement noise, $\alpha = 0.5$, is used when fitting the Gaussian process to account for the variability in reward that may be obtained in a given end-effector position, resulting from the variability introduced by contact with the deformable phantom, potential tumour motion and ultrasound gel spreading effects.

The Gaussian process is iteratively trained using a buffer of visited states (in our experiments these are 3D Cartesian end-effector positions) and the corresponding rewards predicted using the image-based reward model. Actions are then chosen to move to a desired state selected using the objective function,

$$\hat{\mathbf{s}}_t = \mathrm{argmax}_{\mathbf{s}_t} \quad \mu(\mathbf{s}_t) + \beta\sigma(\mathbf{s}_t). \tag{9}$$

here $\mu(\mathbf{s}_t)$, $\sigma(\mathbf{s}_t)$ are the mean and standard deviation of the Gaussian process, and $\beta = 1$ is a hyperparameter controlling the exploration exploitation trade-off of the policy. This

objective function is chosen in order to balance the competing objectives of visiting states that are known to maximise reward with gaining information about values of states for which the model is more uncertain. In our ultrasound imaging application, actions are linear motions to a desired Cartesian state.

Since the GP starts with no prior information about reward, and is re-fit online after each position is visited, the Bayesian optimisation policy naturally transitions from exploration and becomes more exploitative as additional information is gained. Note that, for the ultrasound case, no policy is ever 'trained', instead we optimise the learned reward function online for each new environment using the fixed Bayesian optimisation strategy.

It should also be noted that any policy can be used to optimise rewards predicted using probabilistic temporal ranking. We selected a Bayesian optimisation strategy for online experiments due to its prevalence in active viewpoint selection literature, and because of its ability to deal with uncertainty in state rewards arising the dynamic structure of the deformable imaging phantom.

4.2 Reward inference evaluation

Figure 1 shows predicted reward sequences for sample expert demonstration traces held out from model training. It is clear that the ranking reward model captures the general improvements in image quality that occur as the demonstrator searches for a good scanning view, and that some searching is required before a good viewpoint is found. Importantly, the slack in the pairwise ranking model, combined with the model assumption that similar images result in similar rewards, allows for these peaks and dips in reward to be modelled, as probabilistic temporal ranking does not assume monotonically increasing rewards.

We qualitatively assessed the image regions and features identified using the reward model using saliency maps (Fig. 1d), which indicated that the proposed approach has learned to associate the target object with reward.

In order to quantitatively evaluate the performance of probabilistic temporal ranking for autonomous ultrasound imaging, approximately 5000 ultrasound images from a set of 10 demonstration sequences were ordered in terms of human preference by collecting 5000 human image comparison annotations and applying the ranking model of (Burke et al., 2017). We evaluate reward inference models in terms of how well they agree with this human labelling using Kendall's τ, a measure of the ordinal association between observation sets, and Spearman's ρ, a measure of rank correlation. Figure 10 shows these results.

We benchmark probabilistic temporal ranking (PTR) against a maximum entropy reward model with both a Gaussian process prior (GP-ME-VAE*) (Levine et al., 2011) and a neural network prior (Deep-ME-VAE*) (Wulfmeier et al., 2015), a monotonically increasing linear temporal ranking model (LTR) (Angelov et al., 2020), T-REX-VAE* (Brown et al., 2019) and a servoing reward model (Servo-VAE*) based on the cosine similarity of a latent image embedding to a final image captured. As in the grid world experiments, we provide T-REX with labelling information using trajectory length as a heuristic for demonstration quality. We also include a number of ablation results for probabilistic temporal ranking models. Model parameters are provided in the appendices.

Here, VAE* denotes the use of pre-trained image embedding learned independently using variational autoecoding, ML-PTR refers to a model trained without decoding the latent embedding (no autoencoder loss), ML-PTR-VAE refers to a maximum likelihood model trained jointly with both a variational autoencoding and pairwise ranking objective, and GP-PTR-VAE* denotes the use of the probabilistic temporal ranking with a pre-trained image embedding. It should be noted that all reward models were inferred without policy search, by directly optimising the reward objective.

It is clear that PTR outperforms baseline approaches. The maximum entropy reward models fail to learn adequate reward models. Servoing proved somewhat effective, but is unlikely to scale to more general problems and use cases. PTR improves upon LTR, illustrating the importance of allowing for non-monotonically increasing rewards. T-REX performs much better than maximum entropy approaches, but does not recover the underlying reward function. This is most likely due to the limited number of demonstrations available, but also potentially due to the fact that trajectory length is not always indicative of scan quality in this setting. In some demos, a set of very good quality images could be obtained after extensive exploration, while in others, a passable set of scans may have been obtained after relatively little searching. As a result, there is no clear or objective measure of the quality of a scanning sequence suitable for use with supervised learning approaches like T-REX.

The ablation results show that variational autoencoding produces better reward models. This is most likely due to its regularising effect, which helps to avoid over-fitting to insignificant image appearance differences. Directly optimising without this regularising effect (ML-PTR) essentially results in a monotonically increasing reward model, and produces similar results to LTR. Interestingly, learning an independent auto-encoding and using a single layer bottleneck reward network or Gaussian process, proved to be an extremely effective strategy.

We believe that maximum entropy reward inference fails for two primary reasons. First, probabilistic temporal ranking produces substantially more training data, as each pair of images sampled (50 000 pairs) from a demonstration provides a supervisory signal. In contrast, the maximum entropy approach treats an entire trajectory as a single data point (10

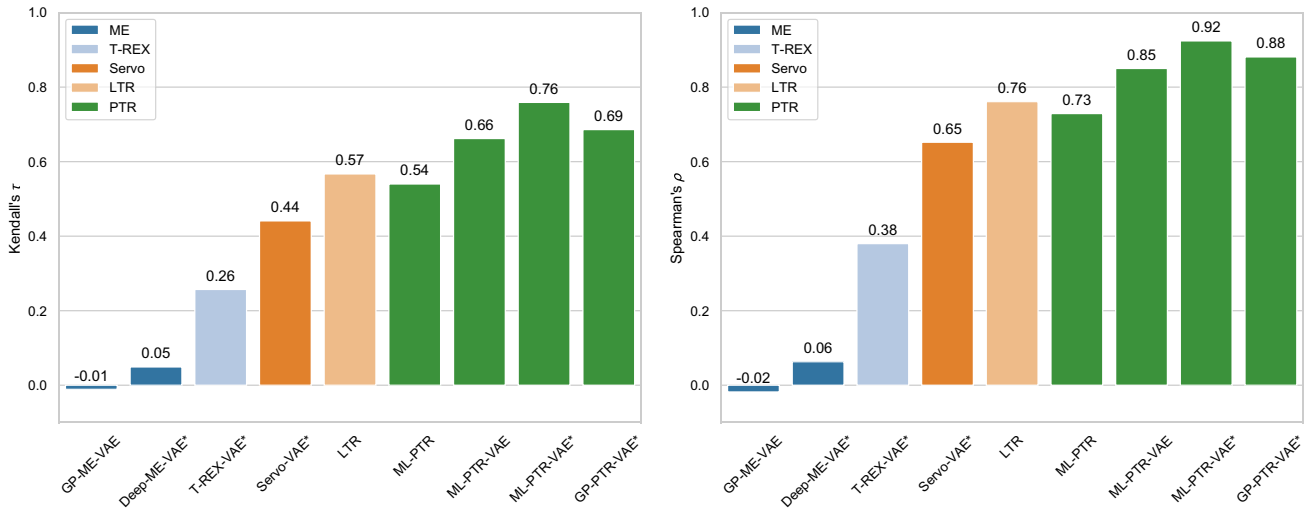


Fig. 10 Reward model association with human image ratings shows that temporal ranking (green) reward inference models strongly agree with human preferences. A Spearman rank correlation of $\rho = 1$ indicates an identical rank or ordering, while $\rho = -1$ indicates a completing opposing order. Similarly, Kendall's $\tau = 1$ indicates that the relative rank assigned to images is identical to that assigned by the human annotator, while $\tau = -1$ would indicate opposing ranks

trajectories), and thus needs to learn from far fewer samples, which is made even more challenging by the high dimensional image inputs. Secondly, the maximum entropy reward assumes that frequently occurring features are a sign of a good policy, which means that it can mistakenly associate undesirable frames seen during the scan's searching process for frames of high reward.

5 Policy evaluation

For policy evaluation, we compare probabilistic temporal ranking with a Gaussian process maximum entropy inverse reinforcement learning approach. For both models we use the same latent feature vector (extracted using a stand-alone variational autoencoder following the architecture in Fig. 9), and the same Bayesian optimisation policy to ensure a fair comparison.

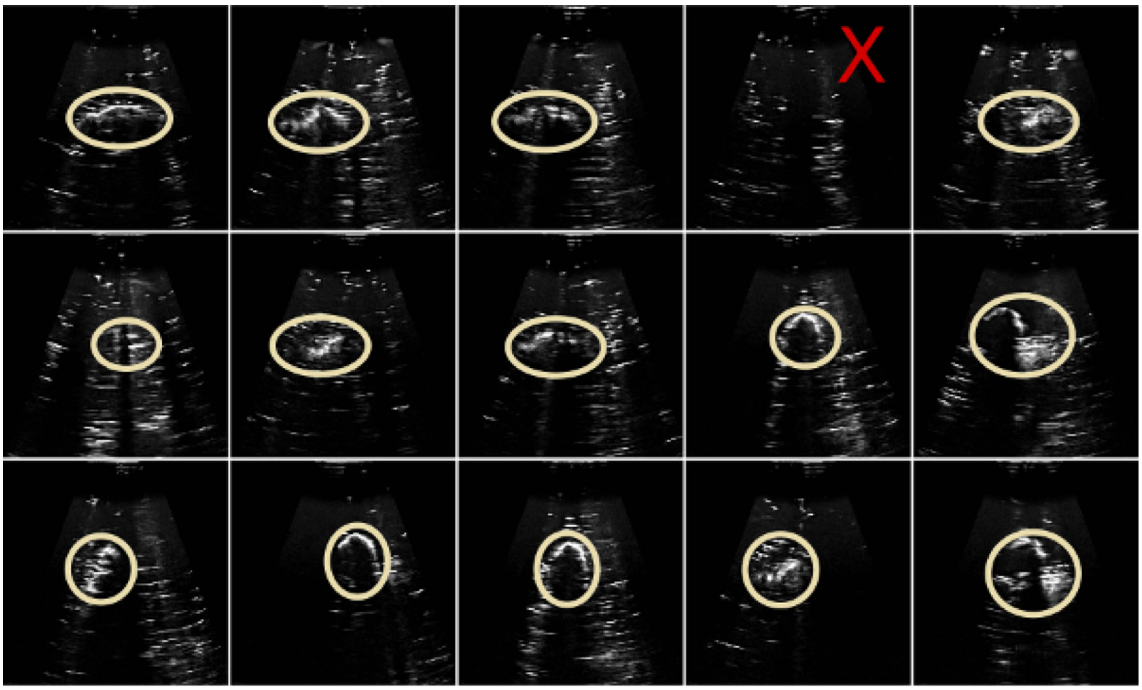
We compare the two approaches by evaluating the final image captured during scanning, and investigating the reward traces associated with each model.

Trials were repeated 15 times for each approach, alternating between each, and ultrasound gel was replaced after 10 trials. Each trial ran for approximately 5 min, and was stopped when the robot pose had converged to a stable point, or after 350 frames had been observed. A high quality ultrasound scan is one in which the contours of the target object stand out as high intensity, where the object is centrally located in a scan, and imaged clearly enough to give some idea of the target object size (see Fig. 2).

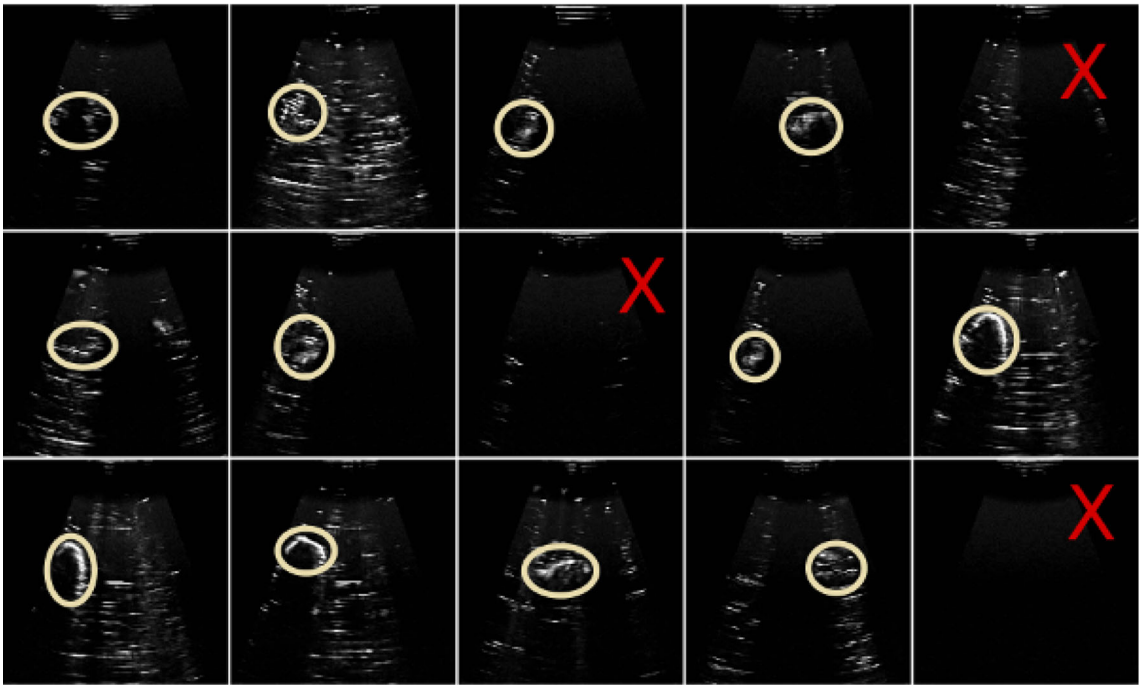
As shown in Fig. 11, the probabilistic temporal ranking model consistently finds the target object in the phantom, and also finds better rated images. Mean and standard deviations in image ratings were obtained using the rating model (see above) trained for reward evaluation using human image preference comparisons. The maximum entropy approach fails more frequently than the ranking approach, and when detection is successful, tends to find off-centre viewpoints, and only images small portions of the target object.

It is particularly interesting to compare the reward traces for the probabilistic temporal ranking model to those obtained using maximum entropy IRL when the Bayesian optimisation scanning policy is applied. Figure 12 overlays the reward traces obtained for each trial. The maximum entropy reward is extremely noisy throughout trials, indicating that it has failed to adequately associate image features with reward. Similar images fail to consistently return similar rewards, so the Bayesian optimisation policy struggles to converge to an imaging position with a stable reward score. In contrast, the reward trace associated with the probabilistic temporal ranking contains an exploration phase where the reward varies substantially as the robot explores potential viewpoints, followed by a clear exploitation phase where an optimal viewpoint is selected and a stable reward is returned.

Figure 13 shows the predicted reward over the search volume (a 50 mm × 50 mm × 30 mm region above the imagining phantom) for a PTR trial, determined as part of the Bayesian optimisation search for images with high reward, from reward and position samples (see Fig. 12). Here, we capture images at 3D end-effector locations according to the Bayesian optimisation policy, and predict the reward over the space of possible end-effector states using (8). Importantly, the Gaussian process proxy function is able to identify an ultrasound positioning region associated with high reward.



(a) Probabilistic temporal ranking reward (Average human image rating: 0.254 ± 0.079)



(b) Maximum entropy reward (Average human image rating: 0.119 ± 0.198)

Fig. 11 Final images obtained after policy convergence clearly show that images obtained using probabilistic temporal ranking are much clearer and capture the target object far more frequently than the maximum entropy reward. Target objects are circled, failures marked with a cross. (Images are best viewed electronically, with zooming. See anonymous companion site, https://sites.google.com/view/ultrasound-scanner, for higher resolution images.)

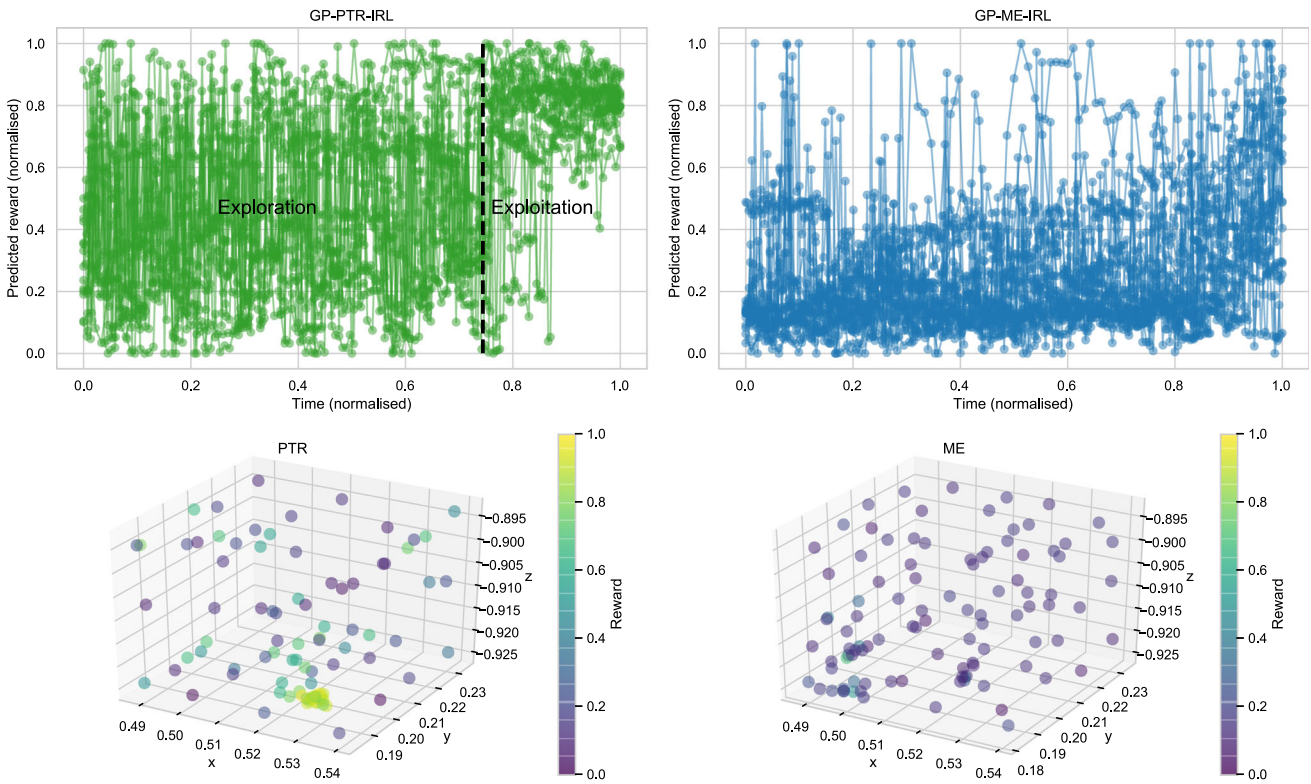


Fig. 12 Reward traces (top) show that the probabilistic temporal ranking reward is stable enough for the BO robot policy to explore the volume of interest (varying reward) before exploiting (stable reward). The maximum entropy reward is extremely noisy, indicating that it has failed to consistently associate high quality ultrasound image features with reward. This can also be observed when the 3D positions selected by the Bayesian optimisation policy are visualised (bottom), and coloured by the reward associated with the ultrasound images obtained when visiting these locations. The PTR policy first explores the allowable search region, before converging to an optimal viewing position (the cluster of high reward points). A policy using the maximum entropy reward model fails to locate the target object

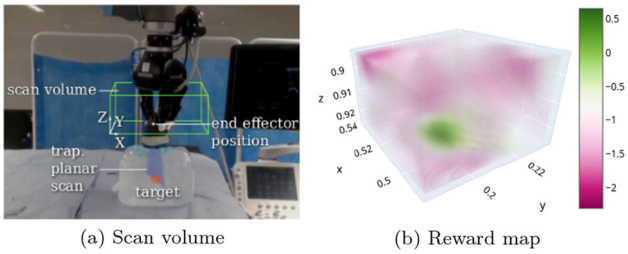


Fig. 13 A visualisation of the reward map (**b**) inferred by the Bayesian optimisation Gaussian process during scanning shows that it attributes high rewards (green) when the probe is pressed against the container directly above the target. The scan volume, or the support of the reward map, is illustrated using a green wireframe in the setup (**a**)

This corresponds to a position above the target object, where the contact force with the phantom is firm enough to press through air pockets, but light enough to maintain a thin, air-tight layer of gel between the probe and phantom.

6 Conclusion

This work introduced probabilistic temporal ranking, an approach to reward inference from exploratory demonstrations for visual servoing or active viewpoint selection tasks. Here, we take advantage of the fact that exploratory demonstrations, whether optimal or sub-optimal, often involve steps taken to improve upon an existing state. Results show that leveraging this to infer reward through a ranking model is more effective than common IRL methods in exploratory cases where demonstrations require a period of discovery in addition to reward exploitation and when observation traces are high dimensional.

This paper also shows how the proposed reward inference model can be used for a challenging ultrasound imaging application. Here, we learn to identify image features associated with target objects using kinesthetic scanning demonstrations that are exploratory, as they inevitably require a search for an object and position or contact force that returns a good image. Using this within a policy that automatically searches for positions[5] and contact forces that maximise a learned reward, allows us to automate ultrasound scanning.

[5] For videos and higher resolution scan images, along with post publication links to code see https://sites.google.com/view/ultrasound-scanner.

When comparing with human scanning, a primary challenge we have yet to overcome is that of spreading ultrasound gel smoothly over a surface. Human demonstrators implicitly spread ultrasound gel evenly over a target as part of the scanning process so as to obtain a high quality image. The Gaussian process policy used in this work is unable to accomplish this, which means scans are still noisier than those taken by human demonstrators. Moreover, human operators typically make use of scanning parameters like image contrast, beam width and scanning depth, which we kept fixed for these experiments. Nevertheless, the results presented here show extensive promise for the development of targeted automatic ultrasound imaging systems, and open up new avenues towards semi-supervised medical diagnosis.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10514-023-10120-w.

Acknowledgements We are particularly grateful to the Edinburgh RAD group and Dr Paul Brennan for valuable discussions and recommendations.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions. This work was supported by funding from the Turing Institute, as part of the Safe AI for surgical assistance project. S. Ramamoorthy was also supported by funding from the UKRI Trustworthy Autonomous Systems Node in Governance and Regulation (EP/V026607/1). K. Subr was supported by a Royal Society URF and K. Lu was partly supported by an EPSRC grant: Holey Sampling.

Declarations

Conflicts of interest S. Ramamoorthy is vice president at five.ai, an autonomous driving company whose focus lies outside the domain of this paper. The remaining authors confirm that no other conflict of interest exists.

Appendices

Hyper-parameter priors

See Fig. 14.

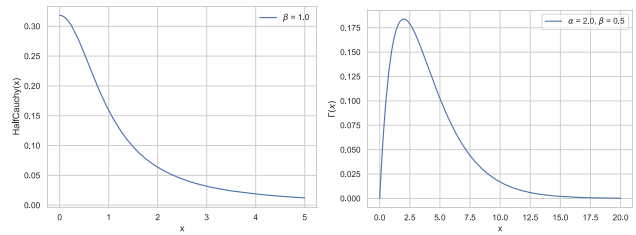


Fig. 14 Prior distributions used for PTR hyper-parameters

Table 4 Neural architecture parameters

Convolutional VAE	
Batch size	128
Training epochs	100
Adam optimiser	learning rate= $1e-4$
Input dims	$112 \times 112 \times 1 \in (0, 1)$
Encoder	
Conv 32	5×5 kernel, relu, strides 2
Conv 64	5×5 kernel, relu, strides 2
Conv 128	5×5 kernel, relu, strides 2
Conv 256	5×5 kernel, relu, strides 2
Dense FC	1024 neurons, relu
Dense FC	16×2 output (mean, variance)
Decoder	
Dense FC	1024 neurons, relu
Conv transpose 128	5×5 kernel, relu, strides 2
Conv transpose 64	5×5 kernel, relu, strides 2
Conv transpose 32	6×6 kernel, relu, strides 2
Conv transpose 1	6×6 kernel, relu, strides 2
Output dims	$112 \times 112 \times 1 \in (0, 1)$
Reward predictor	
Dense FC	relu, output dims 1

Network architectures

Table 4 shows the parameters and training settings used for the convolutional neural architecture experiments.

Table 5 shows the parameters and training settings used for the low-dimensional neural architecture experiments.

Learning to play atari breakout

Although this paper has primarily explored PTR in the context of exploratory demonstrations, PTR is also of use in a wide range of visual IRL tasks, particularly open ended 'survival' settings. In these more general cases, PTR rewards image features corresponding to time spent in an environment, and becomes more akin to an intrinsic motivation strategy (Barto, 2013).

Table 5 Neural architecture parameters (OpenAI Gym Environments)

RewardNet	
Batch size	64
Training epochs	1000
AdamW optimiser	learning rate$= 1e-4$
Encoder	
RunningNorm	
Dense FC	32×32, relu,
Dense FC	32×32, relu,
Reward predictor	
Dense FC	32×1, relu

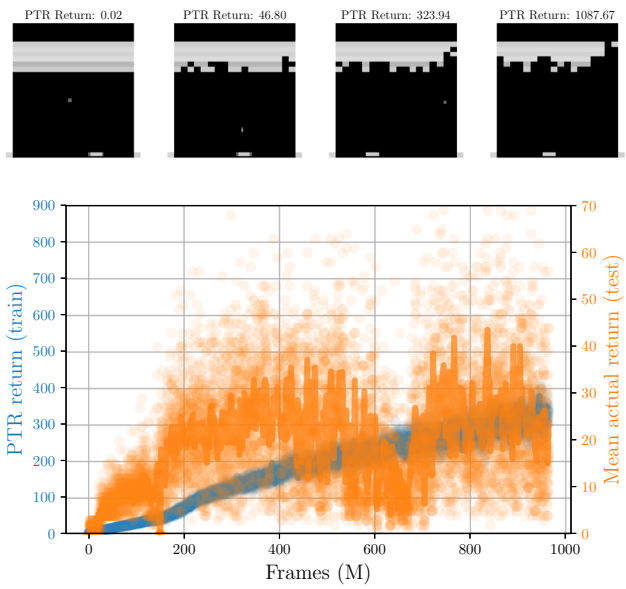


Fig. 15 Top: Game states associated with PTR rewards. Bottom: Returns (cumulative reward) obtained for an A3C policy trained using PTR. The blue axes and curve provides the increase in return as a function of frames seen during training, while the orange curve provides the corresponding true reward obtained by the trained policy. Axes are aligned by linearly scaling by the ratio of the maximum reward inferred using PTR and the maximum true reward of an A3C policy used to gather demonstrations

This is illustrated below using the Atari game Breakout, where a PTR reward function is learned from 20 demonstrations obtained by an agent trained using A3C[6] (Mnih et al., 2016). The PTR reward function was then used to train a second agent using A3C, which, as illustrated in Fig. 15, learns to play Breakout reasonably well.

[6] https://github.com/greydanus/baby-a3c

References

Abbeel, P.,& Ng, AY. (2004). Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning, ACM*, p 1 https://doi.org/10.1145/1015330.1015430

Abolmaesumi, P., Salcudean, S. E., Zhu, Wen-Hong., Sirouspour, M. R., & DiMaio, S. P. (2002). Image-guided control of a robot for medical ultrasound. *IEEE Transactions on Robotics and Automation, 18*(1), 11–23. https://doi.org/10.1109/70.988970

Angelov, D., Hristov, Y., Burke, M., Ramamoorthy, S. (2020). Composing diverse policies for temporally extended tasks. Robotics and automation letters (RA-L) arXiv:1907.08199.

Bagnell, JAD. (2015). An Invitation to Imitation. Tech. Rep. CMU-RI-TR-15-08, Carnegie Mellon University, Pittsburgh, PA https://www.ri.cmu.edu/pub_files/2015/3/InvitationToImitation_3_1415.pdf.

Barto, AG. (2013). Intrinsic motivation and reinforcement learning. In *Intrinsically Motivated Learning in natural and Artificial Systems*, Springer, pp. 17–47.

Binney, J., & Sukhatme, GS. (2012). Branch and bound for informative path planning. In *2012 IEEE International Conference on Robotics and Automation*, pp. 2147–2154 https://doi.org/10.1109/ICRA.2012.6224902.

Biyik, E., Huynh, N., Kochenderfer, M., & Sadigh, D. (2020). Active Preference-Based Gaussian Process Regression for Reward Learning. In *Proceedings of Robotics: Science and Systems*. https://doi.org/10.15607/RSS.2020.XVI.041.

Boularias, A., Kober. J., & Peters, J. (2011). Relative entropy inverse reinforcement learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 182–189 http://proceedings.mlr.press/v15/boularias11a/boularias11a.pdf.

Braziunas, D., & Boutilier, C. (2006). Preference elicitation and generalized additive utility. AAAI 21(2) https://www.aaai.org/Papers/AAAI/2006/AAAI06-253.pdf.

Brochu, E., Brochu, T., de Freitas, N. (2010). A Bayesian interactive optimization approach to procedural animation design. In *Proceedings of the 2010 ACM SIGGRAPH/Eurographics Symposium on Computer Animation, Eurographics Association*, pp. 103–112 https://doi.org/10.5555/1921427.1921443.

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., & Zaremba, W. (2016). Openai gym. arXiv:1606.01540.

Brown, D., Goo, W., Nagarajan, P., & Niekum, S. (2019). Extrapolating Beyond Suboptimal Demonstrations via Inverse Reinforcement Learning from Observations. In *International Conference on Machine Learning*, pp. 783–792 arXiv:1904.06387.

Brown, DS., Goo, W., & Niekum, S. (2020). Better-than-demonstrator imitation learning via automatically-ranked demonstrations. In *Conference on robot learning, PMLR*, pp. 330–359.

Burke, M., Mbonambi, S., Molala, P., & Sefala, R. (2017). Rapid Probabilistic Interest Learning from Domain-Specific Pairwise Image Comparisons. arXiv preprint arXiv:1706.05850

Calandra, R., Seyfarth, A., Peters, J., & Deisenroth, MP. (2014). An experimental comparison of Bayesian optimization for bipedal locomotion. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1951–1958 https://doi.org/10.1109/ICRA.2014.6907117.

Chatelain, P., Krupa, A., & Marchal, M. (2013). Real-time needle detection and tracking using a visually servoed 3D ultrasound probe. In *2013 IEEE International Conference on Robotics and Automation*, pp. 1676–1681 https://doi.org/10.1109/ICRA.2013.6630795.

Chatelain, P., Krupa, A., & Navab, N. (2015). Optimization of ultrasound image quality via visual servoing. In *2015 IEEE Interna-*

tional Conference on Robotics and Automation (ICRA), IEEE, pp. 5997–6002.

Cho, DH., Ha, JS., Lee, S., Moon, S., & Choi, HL. (2018). Informative path planning and mapping with multiple UAVs in wind fields. In *Distributed Autonomous Robotic Systems*, Springer, pp. 269–283. arXiv:1610.01303

Chu, W., & Ghahramani, Z. (2005). Preference learning with gaussian processes. In *Proceedings of the 22nd International Conference on Machine Learning*, pp. 137–144.

Deisenroth, M., & Rasmussen, CE. (2011). PILCO: A model-based and data-efficient approach to policy search. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pp.465–472 https://doi.org/10.5555/3104482.3104541.

Finn, C., Christiano, P., Abbeel, P., & Levine, S. (2016). A connection between generative adversarial networks, inverse reinforcement learning, and energy-based MODELS. arXiv preprint arXiv:1611.03852.

Fu, J., Luo, K., & Levine, S. (2018). Learning robust rewards with adversarial inverse reinforcement learning. In *International Conference on Learning Representations (ICLR)*. arXiv:1710.11248.

Ghasemipour, SKS., Zemel, R., & Gu, S. (2019). A Divergence Minimization Perspective on Imitation Learning Methods. In *Conference on Robot Learning (CoRL)*. arXiv:1911.02256.

Gleave, A., Taufeeque, M., Rocamonde, J., Jenner, E., Wang, SH., Toyer, S., Ernestus, M., Belrose, N., Emmons, S., & Russell, S. (2022). imitation: Clean imitation learning implementations. arXiv:2211.11972v1 [cs.LG], https://arxiv.org/abs/2211.11972, 2211.11972.

Herbrich, R., Minka, T., & Graepel, T. (2007). TrueSkill^{TM}: a Bayesian skill rating system. In *Advances in neural information processing systems* pp 569–576 https://papers.nips.cc/paper/3079-trueskilltm-a-bayesian-skill-rating-system.pdf

Ho, J., & Ermon, S. (2016). Generative adversarial imitation learning. In *Advances in neural information processing systems*, pp. 4565–4573. http://papers.nips.cc/paper/6391-generative-adversarial-imitation-learning.pdf.

Kiapour, MH., Yamaguchi, K., Berg. AC., & Berg, TL. (2014). Hipster wars: Discovering elements of fashion styles. In *European conference on computer vision*, Springer, pp. 472–488. https://doi.org/10.1007/978-3-319-10590-1_31.

Konstantinova, J., Li, M., Althoefer, K., Nanayakkara, T., Dasgupta, P. (2013). Palpation strategies for artificial soft tissue examination. In *3rd Joint Workshop on New technologies for Computer/Robot Assisted Surgery*

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., & Blei, D. M. (2017). Automatic differentiation variational inference. *Journal of Machine Learning Research, 18*(1), 430–474.

Lee, K., Choi, S., & Oh, S. (2016). Inverse reinforcement learning with leveraged gaussian processes. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE*, pp. 3907–3912.

Levine, S., Popovic, Z., & Koltun, V. (2011). Nonlinear inverse reinforcement learning with Gaussian processes. In *Advances in Neural Information Processing Systems*, pp. 19–27 https://papers.nips.cc/paper/4420-nonlinear-inverse-reinforcement-learning-with-gaussian-processes.pdf.

Li, T., Kermorgant, O., Krupa, A. (2012). Maintaining visibility constraints during tele-echography with ultrasound visual servoing. In *2012 IEEE International Conference on Robotics and Automation*, pp. 4856–4861 https://doi.org/10.1109/ICRA.2012.6224974.

Liang, K., Rogers, A. J., Light, E. D., von Allmen, D., & Smith, S. W. (2010). Three-dimensional ultrasound guidance of autonomous robotic breast biopsy: Feasibility study. *Ultrasound in Medicine & Biology, 36*(1), 173–177. https://doi.org/10.1016/j.ultrasmedbio.2009.08.014

Ling, CK., Low, KH., & Jaillet, P. (2016). Gaussian process planning with Lipschitz continuous reward functions: Towards unifying Bayesian optimization, active learning, and beyond. In *Thirtieth AAAI Conference on Artificial Intelligence*. arXiv:1511.06890.

Lopes, M., Melo, F., & Montesano, L. (2009). Active learning for reward estimation in inverse reinforcement learning. In W. Buntine, M. Grobelnik, D. Mladenić, & J. Shawe-Taylor (Eds.), *Machine Learning and Knowledge Discovery in Databases* pp. 31–46. Springer.

Majumdar, A., Singh, S., Mandlekar, A., & Pavone, M. (2017). Risk-sensitive Inverse Reinforcement Learning via Coherent Risk Models. In *Proceedings of Robotics: Science and Systems*, Cambridge, Massachusetts. http://www.roboticsproceedings.org/rss13/p69.pdf.

Marchant, R., & Ramos, F. (2014). Bayesian Optimisation for informative continuous path planning. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6136–6143. https://doi.org/10.1109/ICRA.2014.6907763.

Martinez-Cantin, R. (2017). Bayesian optimization with adaptive kernels for robot control. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3350–3356. https://doi.org/10.1109/ICRA.2017.7989380.

Martinez-Cantin, R., de Freitas, N., Brochu, E., Castellanos, J., & Doucet, A. (2009). A Bayesian exploration-exploitation approach for optimal online sensing and planning with a visually guided mobile robot. *Autonomous Robots, 27*(2), 93–103. https://doi.org/10.1007/s10514-009-9130-2

Mnih, V., Badia, AP., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., & Kavukcuoglu, K. (2016). Asynchronous methods for deep reinforcement learning. In *International Conference on Machine Learning*, pp. 1928–1937.

Murawski, C., & Bossaerts, P. (2016). How humans solve complex problems: The case of the knapsack problem. *Scientific Reports, 6*, 34851.

Naik, N., Philipoom, J., Raskar, R., & Hidalgo, C. (2014). Streetscore – predicting the perceived safety of one million streetscapes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 779–785. https://doi.org/10.1109/CVPRW.2014.121.

Neal, RM. (1996). Priors for infinite networks. In *Bayesian Learning for Neural Networks*, Springer. pp. 29–53. https://www.cs.toronto.edu/~radford/ftp/pin.pdf.

Ng, AY., & Russell, SJ. (2000). Algorithms for inverse reinforcement learning. In *ICML*, vol. 1, p. 2.

Raffin, A., Hill, A., Gleave, A., Kanervisto, A., Ernestus, M., & Dormann, N. (2021). Stable-baselines3: Reliable reinforcement learning implementations. *Journal of Machine Learning Research, 22*(268), 1–8.

Salvatier, J., Wiecki, T. V., & Fonnesbeck, C. (2016). Probabilistic programming in Python using PyMC3. *PeerJ Computer Science, 2*, e55. https://doi.org/10.7717/peerj-cs.55

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347.

Sermanet, P., Lynch, C., Chebotar, Y., Hsu, J., Jang, E., Schaal S., & Levine, S. (2018). Time-contrastive networks: Self-supervised learning from video. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1134–1141. arXiv:1704.06888.

Shiarlis, K., ao Messias, J., & Whiteson, S. (2016). Inverse reinforcement learning from failure. In *AAMAS 2016: Proceedings of the Fifteenth International Joint Conference on Autonomous Agents and Multi-Agent Systems*, pp. 1060–1068.

Sridharan, M., Wyatt, J., & Dearden, R. (2010). Planning to see: A hierarchical approach to planning visual actions on a robot using pomdps. *Artificial Intelligence, 174*(11), 704–725.

Sugiyama, H., Meguro, T., & Minami, Y. (2012). Preference-learning based inverse reinforcement learning for dialog control. In *Thirteenth Annual Conference of the International Speech Communication Association*. https://www.isca-speech.org/archive/archive_papers/interspeech_2012/i12_0222.pdf.

Sui, Y., Zhuang, V., Burdick, JW., & Yue, Y. (2017). Multi-dueling bandits with dependent arms. arXiv preprint arXiv:1705.00253.

Thurstone, LL. (2017). A law of comparative judgment. In *Scaling*, Routledge. pp. 81–92.

Tucker, M., Novoseller, E., Kann, C., Sui, Y., Yue, Y., Burdick, JW., & Ames, AD. (2020). Preference-based learning for exoskeleton gait optimization. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, pp. 2351–2357.

Valko, M., Ghavamzadeh, M., & Lazaric, A. (2013). Semi-supervised apprenticeship learning. In *European Workshop on Reinforcement Learning*, pp. 131–142.

Williams, CK., & Rasmussen, CE. (2006). Gaussian processes for machine learning, vol 2. MIT press Cambridge, MA http://www.gaussianprocess.org/gpml/chapters/RW.pdf.

Wirth, C., Akrour, R., Neumann, G., & Fürnkranz, J. (2017). A survey of preference-based reinforcement learning methods. *The Journal of Machine Learning Research, 18*(1), 4945–4990.

Wu, YH., Charoenphakdee, N., Bao, H., Tangkaratt, V., Sugiyama, M. (2019). Imitation learning from imperfect demonstration. In *International Conference on Machine Learning*, pp. 6818–6827.

Wulfmeier, M., Ondruska, P., Posner, I. (2015). Maximum entropy deep inverse reinforcement learning. arXiv preprint arXiv:1507.04888.

Yi, Z., Calandra, R., Veiga, F., van Hoof, H., Hermans, T., Zhang, Y., & Peters, J. (2016). Active tactile object exploration with Gaussian processes. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4925–4930. https://doi.org/10.1109/IROS.2016.7759723.

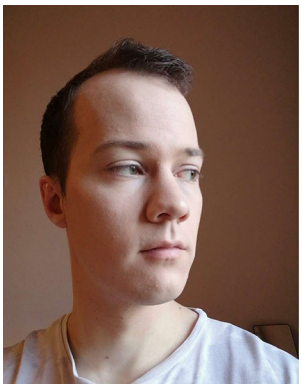
Ziebart, BD., Maas, A., Bagnell, JA., & Dey, AK. (2008). Maximum entropy inverse reinforcement learning. In *Proceedings of AAAI'08 Proceedings of the 23rd national conference on Artifical intelligence*, vol 3, pp. 1433–1438. https://www.aaai.org/Papers/AAAI/2008/AAAI08-227.pdf.

Michael Burke is a lecturer specialising in robot learning and control at Monash University, Australia. He was a research associate working on robot learning at the University of Edinburgh from 2018-2020 and led the Mobile Intelligent Autonomous Systems group at the Council for Scientific and Industrial Research (CSIR) in South Africa prior to this. He has a PhD in statistical signal processing from the University of Cambridge (2012-2016), a Masters of Science in electronic engineering from Stellenbosch University (2009-2011) and a Bachelors in electronic engineering from the University of Pretoria (2005-2008).

Katie Lu was an intern at the Robust Autonomy and Decisions group at the University of Edinburgh, following her completion of the university's MSc in Artificial Intelligence in 2019. She previously worked as a data scientist at a startup in San Diego. She received a BS in math and cognitive science from UCLA in 2016 and currently resides in her hometown of Los Angeles, where she is working as an intern in bioinformatics.

Daniel Angelov is a PhD student at the University of Edinburgh in Robotics and Autonomous Systems, doing research in the intersection of Learning from Demonstration, causality and robotics. Previously he obtained his MEng in Robotics from the University of Reading in 2015.

Artūras Straižys received the M.Eng. degree in Electrical and Electronics Engineering from Heriot-Watt University in 2013. He received the M.Res. degree in Robotics and Autonomous Systems from Edinburgh University in 2019. He is currently a Ph.D. student at University of Edinburgh. His research interests include medical robotics, robot control and skill transfer learning.



Craig Innes is a post-doctoral robotics researcher in the Institute for Action, Perception and Behaviour (IPAB) at the University of Edinburgh. His work focuses on the intersection between symbolic and statistical approaches to machine learning, and how notions of safety can be guaranteed in such systems.

Kartic Subr a Royal Society University Research Fellow and Senior Lecturer at the University of Edinburgh, researches models and algorithms for rapid approximation of solutions to computationally challenging problems. His group develops fast approximations of physically-based simulation to enable exciting applications in robotics, computer graphics and computer vision. Kartic has a PhD in Computer Science from University of California Irvine. His research has been shaped by his experiences in academic institutions including INRIA-Grenoble and University College London as well as companies such as Disney Research and NVIDIA.

Subramanian Ramamoorthy is a Professor and Chair of Robot Learning and Autonomy in the School of Informatics at the University of Edinburgh, where he is also Director of the Institute of Perception, Action and Behaviour. He is a Turing Fellow at the Alan Turing Insatitute and an Executive Committee Member for the Edinburgh Centre for Robotics. He received his PhD in Electrical and Computer Engineering from The University of Texas at Austin in 2007. He has been a Member of the Young Academy of Scotland at the Royal Society of Edinburgh, and has held Visiting Professor positions at the University of Rome “La Sapienza” and at Stanford University. His research focus is on robot learning and decision-making under uncertainty, with particular emphasis on achieving safe and robust autonomy in human-centered environments. This work has resulted in 100+ research articles, received best paper awards at ICRA, IROS, CoRL, TAROS, and EACL, and attracted funding from diverse sources including UKRI (EPSRC and MRC), European Commission (through FP7 and H2020), DARPA, DSTL and RAEng. Between 2017-2020, he served as Vice President - Prediction and Planning at FiveAI, a UK-based startup company focused on developing a technology platform for autonomous vehicles. He continues to be involved with the company as Scientific Advisor.