
Peer reviewed version

Link to published version (if available):
10.1109/IROS.2018.8593751

Link to publication record in Explore Bristol Research
PDF-document

This is the author accepted manuscript (AAM). The final published version (version of record) is available online via IEEE at https://ieeexplore.ieee.org/document/8593751. Please refer to any applicable terms of use of the publisher.

University of Bristol - Explore Bristol Research
General rights
This document is made available in accordance with publisher policies. Please cite only the published version using the reference above. Full terms of use are available: http://www.bristol.ac.uk/red/research-policy/pure/user-guides/ebr-terms/
Deep Learning for Exploration and Recovery of Uncharted and Dynamic Targets from UAV-like Vision

William Andrew¹, Colin Greatwood² and Tilo Burghardt¹

Abstract—This paper discusses deep learning for solving static and dynamic search and recovery tasks – such as the retrieval of all instances of actively moving targets – based on partial-view Unmanned Aerial Vehicle (UAV)-like sensing. In particular, we demonstrate that abstracted tactic and strategic explorational agency can be implemented effectively via a single deep network that optimises in unity: the mapping of sensory inputs and positional history towards navigational actions. We propose a dual-stream classification paradigm that integrates one Convolutional Neural Network (CNN) for sensory processing with a second one for interpreting an evolving long-term map memory. In order to learn effective search behaviours given agent location and agent-centric sensory inputs, we train this design against 400k+ optimal navigational decision samples from each set of static and dynamic evolutions for different multi-target behaviour classes. We quantify recovery performance across an extensive range of scenarios; including probabilistic placement and dynamics, as well as fully random target walks and herd-inspired behaviours. Detailed results comparisons show that our design can outperform naive, independent stream and off-the-shelf DRQN solutions. We conclude that the proposed dual-stream architecture can provide a unified, rationally motivated and effective architecture for solving online search tasks in dynamic, multi-target environments. With this paper we publish³ key source code and associated models.

I. INTRODUCTION

In this work, we propose a map-based, unified deep learning framework (see Fig. 1) applicable to recovery tasks in structured environments where an agent with local sensing and spatial-temporal long-term memory is tasked with visiting within a confined space as quickly as possible each of a known-size set of static or dynamically moving targets.

In the special case where agents return and target locations are fully known over time and space, the task can be mapped to the static or dynamic Travelling Salesman Problem (TSP) [2], respectively. Practical solutions for this problem class have in the past been computed using Dynamic Programming [3] or Ant Colony System optimisation [4].

We consider a more realistic scenario, where target locations are initially unknown, consequently requiring exploratory agency. The search for targets with unknown locations typically arises in search and rescue (SAR) applications [5], [6], [7]. Here specifically, application of the proposed methodology is motivated by requirements to discover constituent positions of dynamic herds in confined-area outdoor farming, such that close-up UAV-based visual animal identification can take place [8]. The described recovery task may classically be interpreted as a partially-observable Markov decision process (POMDP), described in various surveys on visually-motivated robotic navigation [9], [10] prior to deep learning. However, in this paper, we propose to cast the task into a framework for optimising deep recurrent classification. That is, mapping positional history and current sensory inputs to new navigational actions via a single Deep Recurrent Neural Network (DNN). Similar to Zhang et al. [11] in

³This work was supported by the EPSRC Centre for Doctoral Training in Future Autonomous and Robotic Systems (FARSCOPE) at the Bristol Robotics Laboratory (BRL).

1Department of Computer Science, ²Department of Aerospace Engineering, Faculty of Engineering, University of Bristol, UK {will.andrew, colin.greatwood}@bristol.ac.uk, tilo@cs.bris.ac.uk

²Source code available at: https://data.bris.ac.uk/data and https://github.com/CWOA/GTRF

Fig. 1. Unified Dual-Stream Deep Architecture for Search and Recovery Tasks. The proposed design models the interpretation of sensory (tactical) and historic navigational (strategic) information within a single deep network (dotted yellow), which allows for unified back-propagation of navigation decision errors across both domains. Starting at the top right, the flow chart shows the agent’s sensory input 𝐼, that is either (a) an abstracted, or (b) rendered, nearby environment sample. The input is processed via (c) a sensory/visual CNN utilising a basic AlexNet [1] design. In a second, parallel stream an evolving positional history memory 𝑀 (holding either (d) spatial, or (e) spatial-temporal, long term information) is used as tensor input into (f) a network interpreting explorational histories. Both streams are concatenated into (g) a shallow integration network that culminates in a SoftMax map towards a score vector output 𝑉 over the possible navigational actions 𝑎 ∈ 𝐴. During training, the entire deep network (dotted yellow) is optimised based on triples (𝐼, 𝑀, 𝑉) using one-hot encoding of 𝑉 given 𝑎 and cross-entropy loss. During inference at time step 𝑡, the network receives a sensory input 𝐼 and (h) selects the top-ranking navigational action 𝑎 based on 𝑉, which is (i) performed and, in-turn, (j) the positional history 𝑀 is updated. In order to initiate a next iteration time-step 𝑡 + 1, a new sensory 𝐼 is sampled from the environment 𝐸 closing the operational loop.
their recent work on reinforcement learning for exploration, we integrate explicit long-term memory into the design, experimenting with both storage of spatial as well as spatial-temporal information (see Figure 1(d) and (e)).

In order to focus on the core of the proposed methodology and to be able to operate on tractable computational grounds, we abstract from real-world layouts to (1): model time discretely, (2): represent space as a simple two-dimensional grid world and (3): assume grid-cell agent localisation to be resolved perfectly. Proposed as early as 1990 [12], grid worlds remain popular for exploring the viability of AI solutions today [13], [11]. In our work, agent-centric visual sampling – approximating a low-flying UAV with a downward-facing camera – is generated either from (1): occupancy of local sub-grids of the world as content-independent abstractions, or (2): rendering from 3D Gazebo simulations [14], [11] for domain-specific, high-resolution images of particular target and environment content (e.g. grazing cattle herds).

Principal contributions of this work can be summarised as:

- A novel dual-stream CNN-based navigation paradigm and architecture for discovering the positions of static or dynamic targets in uncharted environments.
- Demonstrable out performance of implemented baselines when training on optimal navigation decisions.
- Proof of concept across conducted experiments in meaningful feature learning within a realistic simulation environment applicable to real-world UAV scenarios.

II. FURTHER RELATED WORK

A. Categorisation for Navigation

Robotic navigational research naturally relies heavily on the categorisation of sensory input. In visually-motivated navigation, approaches can largely be categorised as map-based and mapless navigation [9], [10]. Mapless vision approaches – with no global environment representation – can be found to: be based in optical-flow [15] and template appearance matching [16], [17], track landmark features [18], [19] and more recently relevant, directly classify visual input via CNNs [20], [21]. In this work, we formulate a 2D global approximation of the environment (storing visited positions) inspired by occupancy grid maps [22], [23], [24], as opposed to post-exploration map-building approaches [25] and topological map representations [26].

B. Deep Reinforcement Learning for Exploration

Seminal work in deep reinforcement learning (DRL) occurred in 2013 when researchers from DeepMind Technologies used DNNs as an approximator for learning control policies from high-dimensional sensory input such as images. Trained with a variant of Q-learning towards the goal of replicating human-level performance in playing Atari 2600 games, the term Deep Q-Network (DQN) was coined [27], [28]. Deep Recurrent Q-Networks (DRQN) [29] extend this work in a recurrent design for partial agent-environment observability with the utilisation of a Long-Short Term Memory (LSTM) layer [30] in place of the first post-convolutional fully connected layer. These works gave rise to a new genre of reinforcement learning research, and aligned with this work: learning agent navigation in complex [31] and similar [32] environments, map-less navigation [33] achieved visually [34] or via other sensor measurements. One particularly noteworthy paper employs DRL towards target-driven visual agent navigation in simulated indoor environments [35] – bearing resemblance with the problem at hand here.

One could even argue that our work falls well under the domain of classical reinforcement learning, yet we propose its application here to be unnecessarily complex and intensive: within generated episodes, target locations (containing reward) can be known (even though computationally expensive due to TSP’s complexity bounds) and a globally optimal path that visits all targets whilst minimising path length can be inferred (see Section IV). Put differently, an understanding of what constitutes a good solution is known and we can train against this. We will demonstrate this knowledge to be a strong basis for DNN training, accelerating the learning process and making it applicable to sample sets considerably smaller than those required to employ DRQN successfully.

III. PROBLEM FORMALISATION

A. Base Case – Static Target Recovery

For target recovery in a grid world $E$ under discrete time $t$, let $E$ be defined as a rectangular 2D matrix with dimensions $w \times h$. The agent $G$ and static targets $r^i \in R$ both have discrete coordinates $(x,y)$ within the map boundaries $0 \leq x < w$, $0 \leq y < h$ where $x,y \in \mathbb{Z}$. Exactly $|R|$ targets are placed according to some unknown (possibly random) distribution $P$ in this world. Note that $G$ may take position anywhere, whilst multiple targets cannot occupy the same map position:

\[ \forall i,j \in |R|: r^i_x \neq r^j_x \land r^i_y \neq r^j_y \hspace{1cm} (1) \]

where $i \neq j$ and $i,j \in \mathbb{Z}$.

Possible agent actions $a \in A$ are the four possible navigation directions for non-boundary coordinates: forward ($-y$), backward ($+y$), left ($-x$), right ($+x$) or $A = \{f,l,r\}$, respectively, for a top-left origin. Particular actions are inapplicable in case they would move the agent outside the map (e.g. if $G_x = G_w = 0$, corresponding possible actions are $A[G] = \{b,r\}$, $\forall w, h > 0$). Performing one particular action (e.g. ‘move right one unit’, or $x := x + 1$) is defined to take one agent step in discrete time $t$. The agent is deemed to have recovered a target $r^i$ iff $G_x = r^i_x \land G_y = r^i_y$ at some time-step $t$. A recovery solution $S$ to an episode (i.e. visiting all $r^i \in R$) is defined as an ordered action sequence given initial agent coordinates (e.g. $S = [f,f,l,b,r,r,b]$), $G_x^{init} = G_y^{init} = 1$). An episode terminates in $S$ once full target visitation is achieved, or in failure due to time expiry: $t > t_e$.

The quality of a solution $S$ is defined by its cardinality $|S|$, the quality of exploratory agency is defined as the average solution quality (e.g. measured as target recovery rate) computed by scenario sampling given $P$.

Per time step, let the agent $G$ be provided with two observable inputs: its own current position $(G_x^i,G_y^i)$ in coordinates of $E$, and an agent-centric $3 \times 3$ grid image $I$ resolving target occupancy in $E$ topologically adjacent to $G$’s
location – in essence, neighbouring targets and unoccupied cells can be sensed (see Figure 1(a)). For the base case, this abstracts from realistic visual inputs (such as Figure 1(b)) and provides a content-independent problem isolation of tactical sensing from more fine-grained visual recognition tasks. The current position is continuously recorded to keep a \( w \times h \) spatial occupancy map \( M \) up-to-date, encoding for each position of world \( E \): 1) if exploration has taken place, and 2) if the agent is currently located there (see Figure 1(d)). \( M \) in this form provides long-term positional history. Given this setup, the problem can be stated as: how well can we optimise the quality of exploratory action (see examples depicted in Figure 2) by learning of \( P \) using only samples of navigation decisions \((I,M,a)\)?

B. Dynamic Case – Actively Moving Target Recovery

Here the base case is extended for target motion over time. Targets now apply individual actions \( a_i \in A_{ext} \) at velocity \( \frac{1}{2} u/t \) (spatial grid units per time-step) – whereas the agent displaces at velocity \( 1u/t \). We select \( s = 3 \), given \( s > 1 \) makes agent-target visitation eventually always possible. The action set is extended here to include action ‘do nothing’ denoted by \( n \) yielding \( A_{ext} = \{ f,b,l,r,n \} \) to permit optimised agent-target path intersection. Furthermore, single cell target occupation conditions given in Equation 1 are mandated \( \forall t_i \in \{ t_0, t_1, ..., t_e \} \). Architectural and temporal modifications are enacted on the agent’s memory such that both agent and recovery locations are encoded spatiotemporally in the map (see Figure 1(e)), i.e. time annotations are memorised. These modifications are made to allow the system to learn about the uncertainty of target locations by relating agent paths over time and spatiotemporal recovery points along these paths – note that this information reveals properties of target motion beyond placement statistics according to \( P \).

IV. GROUND TRUTH SYNTHESIS

To yield ground-truth solutions to generated episodes for the purpose of training data synthesis of appropriate navigation decisions \((I,M,a)\), we employ the use of two strategies detailed as follows.

A. Fast Approximation: Closest Unvisited Target (CU)

Approximative solutions can be generated fast by determining the position of the closest unvisited target \( r^i \in R \), selecting an appropriate action \( a \in A \) towards visiting \( r^i \) and repeating until all targets have been visited. This forms an approximation in a nearest-neighbour fashion to any environment configuration irrespective of \( w,h \) or \(|R|\).

Action selection is determined by first finding angle \( \theta \) between the agent \( G \) and the closest unvisited target \( r^i \) using:

\[
\theta = \text{atan}2(r^i_y - G_y, r^i_x - G_x).
\]

Second, the action \( a \in A \) is selected via \( \frac{\pi}{2} \)-wide intervals:

\[
a = \begin{cases} 
  \text{rand}(f,r), & \text{if } \frac{\pi}{4} < \theta < \frac{3\pi}{4} \\
  \text{rand}(b,r), & \text{if } -\frac{3\pi}{4} < \theta < -\frac{\pi}{4} \\
  \text{rand}(b,l), & \text{if } -\frac{\pi}{4} < \theta < -\frac{3\pi}{4} \\
  \text{rand}(f,l), & \text{if } \frac{3\pi}{4} < \theta < \frac{5\pi}{4} \\
  f, & \text{otherwise.}
\end{cases}
\]

where \( \text{rand}(X) \) randomly selects an element \( x \in X \).

B. Optimal Solution: Permutation of Targets (PT)

Globally-optimal solution(s) are some ordering on \( R \) that are associated with minimal \(|S|\) for a given \( G_x^{\text{init}}, G_y^{\text{init}} \). A target ordering can be transformed into navigation decisions by using Equations 2 and 3. The number of possible orderings is factorially dependent on the cardinality of the target set, that is, within \( O(|R|!) \). Episode-specific, optimal target sequence orderings are generated via exhaustive search, whose runtime intrinsically degrades exponentially with growing \(|R|\), but is independent of dimensions \( w \times h \). Experimental implications are discussed in Section VI-A and related performance data is summarised in Figure 4.

V. IMPLEMENTATION

A. Architectural Details

As introduced, Figure 1 illustrates and explains the end-to-end deep architecture used here. The reader can observe there that network output consists of a class membership score vector \( V \) with \(|V| = |A| \) and \( \sum_{v \in V} = 1 \) as enforced by a final SoftMax layer. In general, the model-selected action \( a \in A \) is taken to be the maximal-likelihood value of \( V \). Note that, if enacting \( a \) results in the agent moving outside of the environment boundaries, a random valid action is performed instead and equally, loop-detection may also alter action
transfers such as back-propagation through time (BPTT) \[37\] based training can be employed here as opposed to resource-efﬁcient temporally independent/non-correlated, standard batch-trained designs. Furthermore, since all training instances are mod-elled temporally independent/non-correlated, standard batch-based training can be employed here as opposed to resource-intensive back-propagation methods for recurrent architectures such as back-propagation through time (BPTT) \[37\] and real-time recurrent learning (RTRL) \[38\].

### B. Infinite Loop Detection

An inherent property of model output (i.e. next-step action selection) is the potential for performing inﬁnite agent loops. In their simplest forms, inﬁnite loops are easily detectable via direct subtring search into past actions (e.g. \{ f, b, f, b, ...\}, \{r, b, l, f, ...\}). However, speciﬁc subtring rules for more complex loop formations – consisting of many actions – are difﬁcult to formulate and do not generalise well. Instead, loop detection here is primitively indicated in the event that the agent visits a particular grid coordinate \( \alpha \) times (we empirically set \( \alpha = 3 \) to minimise solution length).

Upon indication of a loop, action selection control (see Figure 1\((h)\)) is given to an alternate algorithm, which in turn navigates the agent towards the closest preferred unvisited location in the occupancy map \( M \) for current \( G_x, G_y \). This is achieved by examining occupancy map values surrounding the agent in a \( \beta_{visit} = 1 \) radius. Unvisited locations in that radius vote to determine the next action. If no unvisited cell exists within that radius value \( \beta \), it is incremented \( \beta := \beta + 1 \).

### C. Training Setup

Synthetic training data produced by the selected ground-truth synthesis method is utilised to train the dual-stream DNN model defined previously. For training end-to-end, a single instance \((I, M, a)\) consists of inputs: agent-centric visual input \( I \) and occupancy grid map \( M \), whilst the output is an one-hot action-class vector encoding ground-truth action \( a \) used for back-propagation. To verify pure ground-truth classiﬁcation performance, \( k = 10 \) fold cross validation is performed over the respective experimental dataset. At each fold \( k_i \), training is performed for 50 epochs over the partitioned training set with weights initialised randomly from a truncated normal distribution and using a batch size of 64. Costs are optimised via categorical cross-entropy loss using stochastic gradient descent (SGD) with momentum \[39\] and a ﬁxed learning rate \( \epsilon = 0.001 \). Mean and standard deviation of cross-validated classiﬁcation accuracies for each experiment (if applicable) are given in Table II \((h)\).

### D. Baseline Algorithms

We implement three baseline methodologies to compare our dual-stream, uniﬁed deep learning approach against:

1) Naive Solution (NS): As a simple, naïve strategy to provide a primitive baseline, we employ an algorithm whereby if \((1)\): an unvisited target is currently present within the visual ﬁeld (veriﬁed by examining \( M \)), select an angle-appropriate action towards it using Equation 3. Otherwise, \((2)\): navigate towards a next unvisited location using the voting strategy given in Section V-B (otherwise used to break loops), and repeat \((1)\) until the episode terminates.

2) Deep Recurrent Q-Network (DRQN): We also implement off-the-shelf DRQN \[29\] for comparison employing a 1000-time-step strong experience replay buffer. We establish this baseline to validate the hypothesis that explorational agency beneﬁts from known solution samples and the inclusion of structured, map-based long-term memory. Visualisation of the environment was modiﬁed with a white border surrounding the map used to create the partial visual input \( I \) given to an agent. This serves to provide agent knowledge of the environment boundaries – which is an implicit property of the occupancy grid map \( M \)’s architecture for our method. Additionally, target visitation is visually signiﬁed to the agent who can then no longer receive reward from that target.

3) Split Stream Network (SSN): To validate the motivation of our proposed dual-stream approach – that explorational agency beneﬁts from information exchange between sensor and positional history inputs trained under a single architecture – we split the proposed dual-stream architecture during training. That is, resulting split stream networks 1 and 2 optimise on \((I, a)\) and \((M, a)\) using network architectures \((c)\) and \((f)\) as given in Figure 1 respectively. Network outputs are element-wise summed and normalised to yield the ﬁnal action-class score vector used for action selection.

### VI. Experiments

All experiments were conducted using a 3.6 GHz AMD 8-core Bulldozer CPU with a Nvidia GTX 1080Ti GPU and 32 GB of DDR3 RAM. Throughout experiments, ‘random’
numbers are generated using the P-RNG Merseenne Twister algorithm [40]. We fix the number of targets to $|R| = 5$, set the environment $E$ dimensions $w = h = 10$ and unless specified otherwise, synthesise 20,000-episode strong training datasets per experiment containing approximately 400,000 instances of $(I, M, a)$ data tuples. Note also that the agent’s initial position is randomly generated at the beginning of every episode $C_{x_i y_i}^{\text{init}} \in [0, w - 1], C_{x_j y_j}^{\text{init}} \in [0, h - 1]$. Targets remain static throughout episodes (solving the base case given in Section VI-A where distribution $P$ is two-dimensional) in all experiments up until those given in Section VI-F. For testing, 20,000 episodes are presented to the trained models to solve; each model recovery performance is then measured and reported. Results for all experiments are given in Table I and evaluated over the following subsections.

A. Episode and Ground Truth Generation

In this experiment, we compare and contrast the aforementioned ground-truth strategy (see Section IV) – that is, Closest Unvisited Target (CU) and Permutation of Targets (PT) – for training data generation utilised for subsequent model training. Figure 3 illustrates comparative results on newly-generated episodes with fixed environment dimensions $w = h = 10$ and random target distribution. As theorised for PT, the time required to solve episodes increases factorially with the number of targets, i.e. $O(|R|!)$, whilst CU requires negligible linear time (see Figure 3 (middle)). The tradeoff being that as $|R| \rightarrow \infty$, the average difference in generated solution lengths diverge (see Figure 3 (left)) – since PT guarantees optimality. Yet, as illustrated in Figure 3 (right), employing the CU strategy yields optimality in the majority of instances ($\sim 60\%$). Comparison of both training synthesis strategies against validation episodes demonstrably proves that PT prevails (see both Table I and Figure 4) in learning the conditions for optimal decisions given random target distribution. Thus, we employ the PT approach for episode solution synthesis for all other experiments and establish this result (i.e. using PT) as the base case benchmark for static, uniformly randomly distributed targets.

B. Baseline Comparison

These experiments evaluate the three implemented baseline algorithms (see Section V-D) against our proposed dual-stream architecture. For each algorithm, targets are uniformly randomly spatially distributed within $E$. Resulting performance statistics illustrate under-performance in all aspects in comparison to our benchmark. The employed naïve approach (NS) yields highest baseline performance and is accordingly shown for comparison in Figure 4 against benchmark results achieved by our solution. Off-the-shelf DRQN [29] demonstrably performs poorly despite access to experiences containing decisions leading to reward. This occurs arguably as a result of the agent having no global notion of current environment localisation necessary to locate possible future reward in an overall context. Splitting inputs into separately-trained neural network streams (SSN) is also shown to yield poor performance since model exposure to any single input ($I$ or $M$) alone is insufficient in producing optimal reasoning (the problem becomes increasingly partially-observable). This being said, knowledge of occupancy grid $M$ (encoding $G_x, G_y$, visited cells, etc.) inherently encodes more information on $E$ than $I$, which
is reflected in our findings of significantly higher mAP classification accuracy in the network optimising on \((M, a)\). Additionally, performing end-to-end training on two separate networks to address the imbalance attributes significant additional computation over our proposed network architecture.

**C. Simulated Full Visual Input (+S)**

To this point, sensory environment observability has been encoded by small \(3 \times 3\) sensory abstractions. To simulate a more realistic target location recovery assignment – where visual target detection is no longer trivial – we employ the task of visual outdoor farming census related to [8]; discovering cattle positions within a field using a quadrotor UAV.

The task is simulated utilising randomly-oriented 3D cattle models within the Gazebo robot simulation environment [14] (see Figure 5 (top)). The simulation environment is used to directly replace the agent’s sensory field \(I\) with a \(50 \times 50\) pixel image (see Figure 5 (bottom)) for a 100\(^\circ\) FoV camera, whilst \(M\) remains identical in architecture and operation. The UAV is flown in discrete 2m increments within an \(x-y\)-plane at fixed height \(z = 3.5m\).

The experimental setup remains identical to the benchmark case with agent and cattle targets uniformly randomly spatially distributed. Whilst obtained results (see Table I and Figure 4) indicate an advantage for the simulator case, these results are not directly comparable within the setup, since environment observability improves in a three dimensional projection beyond 2D gridding given the camera FoV since environment observability improves in a three dimensional projection beyond 2D gridding given the camera FoV, visibility of object shadows, etc.

Whilst efforts could have been made to strictly enforce agent visibility to a 1 grid ground cell radius (as for the grid world), this would have failed to model real scenarios well since visual artefacts (e.g. shadows) are present in real-world sensing. Note that the employed visual processing CNN architecture (AlexNet [1]) clearly demonstrates that it can recover and utilise additional visual information found in such realistic robotic scenarios (see Table II Section VI-C).

**D. Learning Static Recovery under Spatial Distributions**

Experiments to this point have dealt with fully random uniform target distributions. Here we show that simpler and arguably more realistic spatial distributions can also be learnt (naturally, far more effectively) under the solution proposed here. We employ three additional distributions as follows:

1) **Fixed Grid**: As a baseline distribution case, targets are distributed in a fixed, static grid across episodes (spatial distribution \(P\) is known completely), whilst agent position initialisation varies as before. Since there are many orders of magnitude less possible environment initialisation configurations, \(wh - |R| = (10 \times 10) - 5 = 95\) – since only the agent position is varied per-episode – one would expect highly improved model performance, which indeed turns out to be the case both in pure ground-truth classification accuracy and online model performance (see Table II row ‘Fixed Grid’).

2) **Equidistant Grid**: Targets are distributed spatially in a grid that satisfies target-target nearest neighbour (NN) equidistance (Voronoi-like distribution) – a property that is observed in real cattle-herd distributions [41]. The position of the grid in \(E\) is varied randomly per-episode and inter-target NN spacing satisfies 2 grid-unit spacing. Since the grid is fixed in shape across episodes, positional knowledge of just two targets provides full information of remaining target positions. Ergo, the problem is then vastly less complex than the fully random case, which is reflected in model performance on this distribution (see Table II rows ‘Equidistant’).

3) **Gaussian**: Target positions are sampled from a two-dimensional Gaussian with constant parameters \(\mu_x = 3, \mu_y = 5\) and \(\sigma_x = \sigma_y = 1\) at the beginning of each episode. Values are sampled from the distribution as required. Results demonstrably show strong model performance in efficiently discovering target positions – relatable to an ability to effectively learn the distributional parameters of the underlying 2D Gaussian \(P\) (see Table II rows ‘Gaussian’).

**E. Memorising Recovery Locations (+M)**

In these experiments, the visitation map \(M\) is modified such that target locations are marked upon agent visitation.
Herd-inspired motion is implemented with an overall group grid (as for Section VI-D.2) with random position in $E$ applied. Targets are initially distributed in an equidistant manner not the case for fully random distributions – knowledge of randomly positioned target $r_i$ reveals no information about the position of $r_j$, and yields no increase (i.e. slight decrease) in performance (see Table I, compare rows 4, 5 and 10).

**F. Learning Dynamic Multi-Target Recovery**

Thus far, targets have remained static throughout episodes. Here we extend our implementation to include individual target motion in two forms. Formalisation of this, now dynamic recovery problem, is given in Section III-B. Note also that infinite loop detection – applicable to the static recovery task – is disabled here.

1) Random Walk: Targets implement random walk via random action selection $\text{rand}(A)$ with a fixed velocity of $\frac{1}{2}u/t$ (one grid unit every 3 time-steps). To determine globally-optimal solutions towards training data generation, $\varepsilon = t_e = 300$ random actions are pre-determined for each target such that Equation 1 – mandating single grid cell target occupancy – remains satisfied for every time-step $t_i \in [0, \varepsilon]$. The yielded matrix containing target coordinates over time is then exhaustively searched for each possible combination of future target visitation orderings towards finding the shortest solution using a full ‘lookahead’ extension of the PT solver.

2) Herd-like Motion: Looking towards more realistic scenarios closer to a robotic UAV application in ‘smart farming’, cattle herd-inspired distribution initialisation and motion is applied. Targets are initially distributed in an equidistant grid (as for Section VI-D.2) with random position in $E$. Herd-inspired motion is implemented with an overall group direction $d_i \in A$ randomly selected at the beginning of each episode. Direction $d_i$ is applied to each target at velocity $\frac{1}{3}u/t$ with a $10\%$ per-individual likelihood that they instead perform a random action $\text{rand}(A_{ext})$ excluding in-axis directions for $d_i$ (e.g. if $d_i = r$, $A_{ext} = \{f, b, n\}$).

This motion behaviour approximation is supported by literature observing collective dynamics for grazing cattle herds [42]. Finally, upon reaching the boundaries of $E$, a new group direction $d_{i+1}$ is randomly selected where $d_i \neq d_{i+1}$. Identically to random walk, $\varepsilon = t_e = 300$ individual motion actions are pre-determined and solved for optimally via the full ‘lookahead’ extension of the PT solver.

Following network training on the random walk experiment, the so far employed 20,000-episode dataset cardinality was found to be insufficient for optimal performance leading to significant model overfitting (see Figure 6). This observation illustrates the significant increase in problem complexity introduced by individual target motion. Increasingly larger datasets improving validation accuracy were thus synthesised at the cost of overall computation time. However, reasonable synthesis, training and evaluation times were quickly exceeded – we opted to empirically limit dynamic experiment datasets to 60,000 episodes as a reasonable accuracy versus time tradeoff. Findings suggest that there is more validation accuracy to be gained by providing even larger datasets of episode datasets (see Figure 6 right).

Quantitative ground-truth classification and online model performance statistics given in Table I demonstrably indicate the capability of the proposed unified network to generalise well to the case of individual target dynamics. In particular, herd-inspired motion generally improves search and recovery capabilities (see Table I, compare rows 13 vs 14 and 15).

Finally, experimental outcomes to this point culminate in our last, highlighted experiment: ‘Herd Walk (+S)’, combining target dynamics enacting herd-like motion and visual sensory input rendered via the cattle-census simulation environment (see Section VI-C and Figure 5). As for the static case, we observe increased environment observability and reduced partiality in view compared to the 2D case. Yet, the experiment clearly validates the employed dual-stream, single network architecture in yielding competitive explorational decisions whilst processing higher resolution, complex visual imagery in unity with spatial-temporal navigational memory.

**VII. Conclusion**

This work demonstrates that recovery tasks can be effectively modelled by combining visually-motivated sensing and map-based positional histories under a single deep classification architecture, comprising the combined optimisation of both information streams in unity. We have shown that this approach outperforms the various tested baselines including split stream optimisation. Our proposed architectural choices demonstrably generalise well to a wide range of scenarios, target distributions and dynamism given training data comprised of good or optimal decision strategies.

Future work will look towards deployment of models developed on-board, real-world UAV visual navigation complementing individual target identification tasks. Within this, the visitation map $M$ will be augmented probabilistically to account for outdoor UAV localisation variance introduced by using GPS. Additionally, environment-observability benefits introduced by varying UAV altitude will be investigated.