

PyroTrack: Belief-Based Deep Reinforcement Learning Path Planning for Aerial Wildfire Monitoring in Partially Observable Environments

Sahand Khoshdel¹, Qi Luo², Fatemeh Afghah¹

Abstract—Motivated by agility, 3D mobility, and low-risk operation compared to human-operated management systems of autonomous unmanned aerial vehicles (UAVs), this work studies UAV-based active wildfire monitoring where a UAV detects fire incidents in remote areas and tracks the fire frontline. A UAV path planning solution is proposed considering realistic wildfire management missions, where a single low-altitude drone with limited power and flight time is available. Noting the limited field of view of commercial low-altitude UAVs, the problem formulates as a partially observable Markov decision process (POMDP), in which wildfire progression outside the field of view causes inaccurate state representation that prevents the UAV from finding the optimal path to track the fire front in limited time. Common deep reinforcement learning (DRL)-based trajectory planning solutions require diverse drone-recorded wildfire data to generalize pre-trained models to real-time systems, which is not currently available at a diverse and standard scale. To narrow down the gap caused by partial observability in the space of possible policies, a belief-based state representation with broad, extensive simulated data is proposed where the beliefs (i.e., ignition probabilities of different grid areas) are updated using a Bayesian framework for the cells within the field of view. The performance of the proposed solution in terms of the ratio of detected fire cells and monitored ignited area (MIA) is evaluated in a complex fire scenario with multiple rapidly growing fire batches, indicating that the belief state representation outperforms the observation state representation both in fire coverage and the distance to fire frontline.

I. INTRODUCTION

The increase in wildfire frequency and severity in recent decades has significantly impacted human health and ecosystems. Economic costs of wildfire damage was approximately \$84.9 billion from 1980 to 2019 in the U.S. [1]. This highlights the critical need for effective wildfire management systems. As a result, the early management of wildfires, including early detection, monitoring, modeling, and suppression has gained increasing attention.

UAVs, equipped with advanced sensing technologies have offered many promising capabilities for high-resolution aerial imaging, and have shown significant potential in wildfire detection tasks, such as creating labeled data sets for wildfire detection with smoke occlusion [2]. The authors in [3] and [4] created a dual RGB/IR aerial dataset using UAVs. Compared to manned aerial systems, UAVs have transformed disaster management with shorter mission start-up time, improved robustness to smoke, chemicals, and heat, and ease of deployment. Despite the advantages

of utilizing UAVs for wildfire detection, tracking the fire frontline actively through the early stages, known as *wildfire monitoring*, remains a challenge. In wildfire monitoring, models aim to solve an optimization problem with the objective of fire coverage and with respect to UAV's trajectory along the fire frontiers with limited batteries while preventing damage caused by heat and smoke. The optimized UAV trajectory consists of coordinates in a 2D space over a planning horizon. In wildfire management operations, the altitude of the UAVs is usually pre-determined for safety to avoid collision with other UAVs and aircraft.

The complexity of wildfire environments, characterized by dynamic spatio-temporal patterns and influenced by factors like vegetation and wind, makes wildfire monitoring a computationally intensive task. Addressing this challenge, several studies have adopted a Partially Observable Markov Decision Process (POMDP) framework for trajectory optimization in wildfire monitoring, acknowledging the challenges posed by the vast scale of wildfires and the limited observability from low-altitude UAVs. Since model-based learning approaches require extensive pre-training data, developing an implicit representation of the environment's dynamics through a belief state could address this issue. Specifically, a belief-based approach can encapsulate information about the fire location and spreading behavior by performing Bayesian updates on belief states with previous observations, and thus substitute memory-mapping with perfect recalls. This belief-based approach can also keep track of the interactions among unobserved factors that directly affect the state transitions and rewards that are explicitly difficult to learn. This easy adaptation to complex dynamics makes belief-based methods superior to methods such as MPC and Kalman filters which have been extensively studied in the early literature.

The current belief-based models for UAV path planning encode the observation uncertainty because of low resolution or limited field of view (FOV) by a surrogate history-dependent distribution. In other words, belief-based models tend to estimate the environment dynamics implicitly through learning a representation of the underlying dynamics matrix. This approach may be more computationally burdensome as the model has no masked focus on the observed area and does not take into account the age of collected information. Previous POMDP-based fire monitoring models did not consider vegetation density, vegetation type, and realistic wind patterns influencing the fire spread, nor did they consider power limits, angle deviation, and the risk of overheating hardware. To the best of our knowledge, this is the first work to study UAV path planning in a realistic wildfire environment with various vegetation types

*This material is based upon work supported by the Air Force Office of Scientific Research under award number FA9550-20-1-0090 and the National Science Foundation under Grant Numbers CNS-2232048, and CNS-2204445.

¹Holcombe Department of Electrical and Computer Engineering, Clemson University, Clemson, SC, USA, {skhoshd, fafghah}@clemson.edu

²Department of Industrial Engineering, Clemson University, Clemson, SC, USA, qluo2@clemson.edu

and densities, considering the practical flight and power limitations of UAV and the age of observed information.

This paper proposes a belief-based DRL solution for UAV path planning to detect and monitor forest fires where low-altitude UAVs have a limited FoV of the environment. Our proposed method narrows the gap between observations and the full state by holding on to beliefs about cells outside the FOV. As a result, the agent will associate detection and monitoring rewards with the believed states of the environment and the UAV's current status. We demonstrate the effect of belief by comparing it to a purely memory-based observational representation in dynamic scenarios of the wildfire simulation. In summary, the contributions of the paper include the following:

- A multi-modal simulated wildfire framework (vegetation density and type, wind dynamics model, etc.)
- Belief state solution to a POMDP for wildfire monitoring (frontline tracking) regarding physical constraints.
- Uncertainty-aware state representation based on certainty map and age of information.

II. RELATED WORKS

Among the various methods for autonomous-UAV-based wildfire monitoring, this section highlights POMDP-based approaches. Some employ the FARSITE wildfire simulation model ([5], [6]), like [7], who used it in a distributed controller for monitoring dynamic wildfires with multiple UAVs, optimizing coverage relative to UAV altitude. The authors in [8] and [9] use FARSITE to develop a Kalman-based spread modeling framework as an estimate of where the fire front is propagating and optimize the trajectory of a UAV fleet based on the estimated state. [10] uses cellular automata to model the wildfire spread and Voronoi tessellation to generate waypoints for a single UAV to follow along the fire frontline.

All the aforementioned works use control-based or optimization approaches for the wildfire monitoring problem. However, rapidly growing fires need more controlling flexibility and the fire dynamics are often unknown a priori. More recently, UAV-based fire monitoring are modeled as MDP/POMDP and solved by learning the dynamics of the environment explicitly or implicitly. Specifically, the state for which the UAV decides to take the optimal estimated action is represented as a belief over the true hidden state of the environment. RL methods are a powerful tool in such scenarios where the environment model is not available to agents [11].

In this vein, [12] formulates the wildfire monitoring problem as a POMDP. The wildfire model considers the fuel and ignition states of the cells. The fire-spreading effects are modeled as the ignition probabilities of non-burning cells dependent on the proximity to ignited cells. This environment model does not fully capture the variability of contributing factors in a spread such as wind, vegetation, etc. Their reward function consists of front-line proximity, ignition coverage, banking tight circles over ignited areas, and redundant observations of multiple aircraft. This reward design is comprehensive but lacks considering the limited fuel/battery of the aircraft and direct penalties for overheated units onboard the aircraft. Furthermore, for their evaluation scenario, only a few fire patterns including circular, T-shaped, and arced fire shapes are considered,

TABLE I
PARAMETER AND SYMBOLS DESCRIPTION

Envir. Params.	Symbol	Agent Params.	Symbol
Total State	\mathcal{S}_{env}	Classification Error	\mathcal{E}
Ignition Status	F	State	\mathcal{S}
Remaining Fuel	f	Orientation	ϕ_U
Wind Magnitude	A	Belief	b
Wind Direction	ϕ	Certainty Factor	c
Parallel Wind Magnitude	$W_{ }$	Battery	P_U
Vegetation Density	ρ	Action	\mathcal{A}
Vegetation Type	V	Reward	\mathcal{R}
Vegetation Radius	R_v	Observation	\mathcal{Z}
Cell Neighborhood	$N_{i,j}, \epsilon_r$	Policy	π

overlooking general wildfire patterns and modeling forest fires.

[13] also models the wildfire tracking problem as a POMDP, while the environment and agent models include subsystems: the targets (fire fronts), the sensors (UAVs), and the tracker. The sensor state encapsulates the location, speed, and heading angles of each UAV. The target state models the 2D coordinates of the active flames. The tracker state, parameterized by a posterior mean vector and covariance matrix, aims to predict the locations of the fire fronts. Their policies will determine the forward acceleration and bank angle of the UAV, given observations about the aggregated state of the fire fronts with measurement errors. The evaluation is done for three main scenarios of one, two, and three UAVs, plus an evaluation of the robustness to wind. Despite the innovative approach of [13] and their extensive evaluation, wind disturbance is modeled with only adding a constant disturbance on the acceleration of the UAV and not on affecting the fire spreading.

III. SYSTEM MODEL

This section describes the forest fire and agent models, respectively. A summary of the environment and agent parameters and notation is shown in Table I.

A. Forest Wildfire Model

We consider a forest environment as a $N \times N$ grid being monitored by a single low-altitude drone. The trajectory of the UAV is defined as direct paths between cell centers. The frames captured by the UAV's camera along the path from one cell to the other are discarded for further processing to reduce the computational burden.

1) **Environment State Parameters:** The state of each of the cells within the grid at time t is represented by $\mathcal{S}_{env}^t(i, j) = \{F_{ij}^t, f_{ij}^t, W_{ij}^t, \rho_{ij}, V_{ij}\}$, listed as follows:

Ignition State The ignition indicator, $F_{ij}^t = \{0, 1, 2\}$, represents the 'not-ignited', 'ignited', and 'burned-out' states, respectively.

Remaining Fuel The remaining fuel within the cell, f_{ij}^t , controls the fire intensity within the cell. Ranging from an initial value $f_{ij}^0 = kf_0$, descending down to 0, after the cell finishes all the fuel within the cell is burnt out. The scale

factor k is defined as proportional to the vegetation density. The spread of the fire from a point within the cell to the corners is the factor controlling the probability of spread to the adjacent cells. However, for the sake of simplicity, we treat the remaining fuel as a co-variate of the progression of a spot fire within the cell.

Vegetation Density Level The density level, ρ_{ij} , corresponds to the amount of initial fuel available in a cell. In this work, we consider 5 different vegetation levels, ($\rho_{ij} = k\rho_0$; $k = \{1, \dots, 5\}$) modeling various density levels present in the environment.

Vegetation Type The vegetation type, V_{ij} controls how fast a cell finishes its fuel.

Wind Magnitude and Phase The magnitude and phase of the local wind around cell (i, j) , $W_{ij}^t = \{A_{ij}^t, \phi_{ij}^t\}$, directly affects the spread probability. The temporal and spatial pattern of the wind's magnitude and phase are described in the next section ($\vec{W}(x, y, t) = A(x, y, t) e^{j\phi(x, y, t)}$).

2) Environment State Initialization: To generate the simulated data, we first generate N_v random circular vegetation patches inside the grid environment, with a radius R_v in which: ($R_v^{min} \leq R_v \leq R_v^{max}$). The k^{th} patch has an assigned discrete vegetation density $\rho_v^k = \{1, \dots, 5\}$ and vegetation type $V_k = \{1, \dots, 5\}$, modeling the consumption rate of the fuel material. The vegetation density and type of every cell within this patch (V_{ij}) is set to V_k and considered constant across the progression simulation. The initial spot fires are chosen randomly within the vegetated regions. Finally, the wind magnitude and phase are initialized across the grid based on arbitrary patterns, some of which are formulated in Eq. 1. To avoid the complexity of the empirical fitted distributions and CFD simulator solutions, a simple spatial and temporal decomposition is considered such that for the wind phase, different cells follow the same temporal pattern but are different in a lag/lead phase, relative to each other. ($\phi(x, y, t) = \Delta\phi(t) + \phi_0(x, y)$).

$$\phi_0(x, y) = \left\{ \frac{x\pi}{m_\phi N}, \frac{y\pi}{m_\phi N}, t g^{-1}\left(\frac{y - \frac{N}{2}}{x - \frac{N}{2}}\right), \dots \right\}; m_\phi \geq \frac{1}{2} \quad (1)$$

where m_ϕ represents the spatial spread factor in the phase component. The lower bound for m_ϕ is to ensure a spatially-unique phase pattern induced by an initial phase limitation ($0 \leq \phi_0(x, y) \leq 2\pi$).

The wind magnitude follows the same spatial and temporal decomposition ($A(x, y, t) = \Delta A(t) + A_0(x, y)$). For the initial wind magnitude, due to the higher turbulence around fire centers, we consider a 2D Gaussian radial basis function (GRBF) around every ignited cell and choose the magnitude of every cell based on the RBF for which the cell is closer to its center. (Eq. 2)

$$A_0(x, y) = A_{max} \exp\left[\left(\frac{-1}{2\sigma_{rad}^2}\right) \min_{ij \in I} d_{xy,ij}^2\right] \quad (2)$$

$$\xRightarrow{(\epsilon_{rad} = 3\sigma_{rad})} = A_{max} \exp\left[-\left(\frac{9}{2\epsilon_{rad}^2}\right) \min_{ij \in I} d_{xy,ij}^2\right]$$

In Eq. 2, I is the set of initially ignited cells, N represents the grid size, $d_{xy,i'j'}$ is the distance to the source cell, ϵ_{rad}

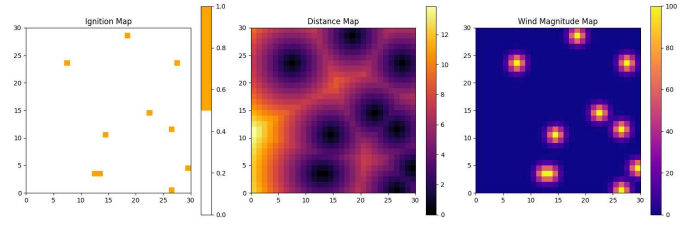


Fig. 1. **Sample Initialization of Wind Magnitude for $N = 30$, $N_{ign} = 10$, $A_{max} = 100$, $\epsilon_{rad} = 3$.** N_{ign} represents the number of initial ignitions. Maps: (Left: Ignition State, Middle: Distance from Fire, Right: Wind Magnitude)

is a cell radius for which the source cells RBF value is approximately zero. An initial configuration for the wind magnitude with noted parameters is displayed in Fig. 1.

3) Environment State Dynamics: The vegetation type and density are considered constant for every cell throughout the wildfire, ignoring vegetation departure and characteristic variation in a short monitoring period. The other variables including the remaining fuel, the wind's magnitude and direction vary over time, resulting in the ignition state transitions.

Fuel Consumption. A transition from ignition state '1' to '2' (burnout) happens when the fuel inside a cell finishes. Eq. 3 shows the vegetation density scales the initial amount of fuel for a cell, while the vegetation type controls the fuel consumption rate. It should be noted that the fuel consumption rate, in reality, depends on many factors other than the fuel, of which the most important are heat and oxygen density in the surrounding area. For the sake of simplicity, we consider the wind intensity as a rough measure of the oxygen density in the cell. Finally, as the cell fuel runs out, the ignition state of the cell changes to '2', indicating a burnt cell. (see (4))

$$f_{ij}^t = \rho_{ij} \exp(-V_{ij}(\frac{A_{ij}^t}{\max(A_{ij}^t)} t)) \quad (3)$$

$$(F_{ij}^{t+1} | F_{ij}^t = 1) = \begin{cases} 2 & ; (f_{ij}^{t+1} = 0) \\ 1 & ; O.W. \end{cases} \quad (4)$$

Wind Dynamics. Accurate simulation needs CFD modeling with multi-physics software, which is beyond the scope of this article. Therefore, we consider a simple sinusoidal temporal pattern to govern the phase and magnitude dynamics according to Eq. 5

$$\Delta\phi(t) = \left(\frac{\pi t}{T_p}\right); \Delta A(t) = A_b \sin\left(\frac{\pi t}{T_m}\right) \quad (5)$$

Here, A_b , T_m , T_p stand for the base wind magnitude level, magnitude variation period, and phase variation period.

Ignition State.

Fire spread is modeled by the transition of the ignition state from '0' to '1' for cells next to currently ignited cells. The probability of spread from a source cell (i, j) to an adjacent cell (i', j') , described as $S_{ij,i'j'}^t$, depends on factors such as inter-cell distance ($d_{ij,i'j'}$), source cell fuel (f_{ij}^t), and wind intensity (W_{ij}^t), calculated using wind magnitude (A_{ij}^t), phase (ϕ_{ij}^t), and wind alignment angle, as shown in Eq. 6. These factors are combined as F_{adj} , F_{fuel} , F_{wind} (Eq. 7), with the wind factor adjusted in no-wind scenarios. The

impact radius, σ_{spr} , gauges the fire's spread likelihood, and throughout the paper, $d_{ij,i'j'}$ denotes the distance between cells

$$p(F_{i'j'}^{t+1} = 1 | F_{i'j'}^t = 0) = \sum_{ij} p(F_{i'j'}^{t+1} = 1 | F_{i'j'}^t = 0, S_{ij,i'j'}^t = 1) p(S_{ij,i'j'}^t = 1) \quad (6)$$

$$p(F_{i'j'}^{t+1} = 1 | F_{i'j'}^t = 0, S_{ij,i'j'}^t = 1) = \underbrace{\left(\frac{1}{2\sigma_{spr}^2} e^{-\frac{d_{ij,i'j'}^2}{2\sigma_{spr}^2}} \right)}_{F_{adj}} \underbrace{\left(1 - \frac{f_{ij}^t}{f_{ij}^0} \right)}_{F_{fuel}} \left(\frac{1}{2} \left(1 + \frac{W_{||ij}^t}{\max A_{ij}^t} \right) \right)_{F_{wind}} \quad (7)$$

$$W_{||ij}^t = A_{ij}^t \cos(|\theta_{ij,i'j'}^t - \phi_{ij}^t|)$$

$$p(S_{ij,i'j'}^t = 1) = \frac{e^{-d_{ij,i'j'}^2}}{\sum_{ij} e^{-d_{ij,i'j'}^2}} \quad (8)$$

The Fig. 2 shows a simple scenario of the fire spread including ignition probability maps and fuel maps. Moreover, the effect of vegetation density on the initial amount of fuel within the cell and the vegetation type on the burnout rate of a cell are seen, respectively.

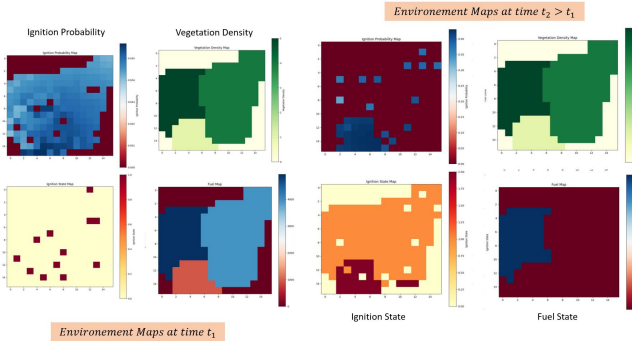


Fig. 2. The effect of fuel density and type on fuel consumption shown in a sample spread scenario. The denser vegetation patches have a higher initial fuel leading to a later burnout.

B. Agent Model

The UAV is modeled as a reinforcement learning agent within a POMDP framework, maintaining constant altitude and speed for a fixed FOV. The POMDP is represented by the tuple $(\mathbf{S}, \mathbf{A}, \mathbf{Z}, \mathbf{T}_{ss'}, \mathbf{R}_{ss'}, \mathbf{O}_{ss'}, \mathbf{p}_{ss'}, \gamma)$, covering state, action, and observation spaces, among others. Given the partial observability and complexity of environmental states, maximizing discounted rewards over observed sequences is challenging. The UAV's decision-making relies on current states and observations instead of inefficient history accumulation, following Markovity to form policy $\pi_{s,z}^a$.

1) **State Space:** The UAV's state is defined by its 2D coordinates, orientation, and battery level, expressed as $S_{UAV} = \{(x, y), \phi_U^t, P_U^t\}$. The orientation influences the available actions, specifically near grid edges or when adjusting the action's deviation angle. Notably, hovering actions remove the next step's deviation angle constraint.

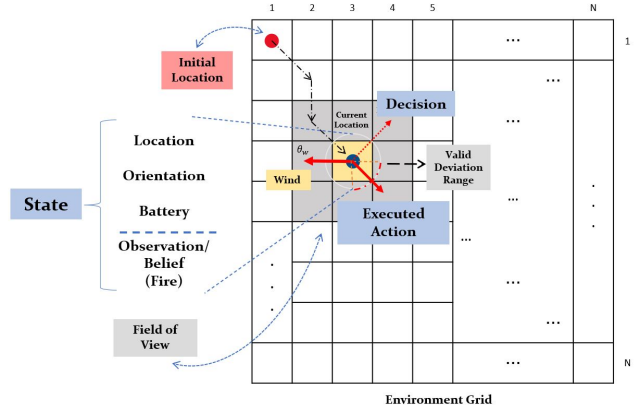


Fig. 3. UAV model for fire-frontline tracking. The valid deviation denotes the acceptable action space which may differ from the action with highest value, thus the most optimal action in the valid range is selected.

2) **Observation Space:** In a POMDP, the UAV's observations are incomplete, only capturing the ignition state and not reflecting the true environmental state, especially due to the hidden nature of wind and fuel levels. Additionally, the UAV's limited FOV leads to outdated information. Observed ignition states, affected by an onboard classification module, may inaccurately represent the actual states due to computational limitations. This is modeled as $\mathbf{O}_{ij}^t = \mathcal{E}(\mathbf{F}_{ij}^t)$, where the classification error matrix (\mathcal{E}) impacts the observed state.

3) **Action Space:** The UAV's action, θ^{t+1} , indicates its next movement direction within a state-dependent action space. This includes 4 main and 4 diagonal directions, plus hovering (\mathcal{A}_H), collectively forming $\mathcal{A} = \{\frac{k\pi}{4}; k \in \{0, \dots, 7\}\} \cup \mathcal{A}_H$. Given the dynamic, partially observable environment, we opt for discrete actions to enhance computational efficiency, aligning with practices in related works. The actions conform to grid constraints and deviation limits, as shown in Eq. 9.

$$\theta^{t+1} \in \mathcal{A}_S \subseteq \mathcal{A} \quad \text{if: } \|\theta^{t+1} - \theta^t\| < \Delta\theta_{dev} \quad (9)$$

C. Reward Function

For the reward function, we have to model both the main task and the constraints in the reward function. Hence, the reward function is an aggregation of a set of sub-functions modeling the objectives and constraints. (Eq. 10) $R_{obj}^t, R_{cstr}^t, R_{inf}^t$ represent the objective (main) reward, constraint penalty (power and battery), and information gain reward, respectively.

$$\mathbf{R}_{total}^t = \mathbf{R}_{obj}^t + \mathbf{R}_{cstr}^t + \mathbf{R}_{inf}^t \quad (10)$$

For this task, we consider detecting and tracking fires along the fire frontier as positive rewards representing the objective, while burnout (the UAV getting too close to the fire flame,) power consumption in movement/hovering, and the event of battery falling beyond a threshold (determining the recharge/return status) are modeled as negative rewards.

$$\mathbf{R}_{obj}^t = \alpha_{det} \frac{\|n_{det}\|}{\|n_{FoV}\|} + \alpha_{mon} e^{-d_{min}}; \quad (11)$$

$$S_{det} = \{(i, j) : F_{ij}^t = 1, (i, j) \in S_{FoV}\}$$

where $\alpha_{det}, \alpha_{mon}$ represent detection and monitoring reward coefficients, which aim to balance the two objectives of fire discovery and frontline tracking.

Conditions for which the negative rewards of battery depletion and burnout should be applied are shown with conditional identity functions.

$$\mathbf{R}_{\text{cstr}}^t = \alpha_{\text{mvm}} R_{\text{mvm}}^t + \alpha_P R_{\text{btr}} \mathbf{1}(P_U^t < P_U^{\text{thr}}) + \alpha_{\text{brn}} R_{\text{brn}} \mathbf{1}(F_{x,y}^t = 1) \quad (12)$$

To model the power consumption of physical movement accurately, we penalize the agent for moving γ_m times more than hovering and also consider a wind-dependent coefficient β_t which penalizes movement against the wind more than movement in its direction.

$$R_{\text{mvm}}^t = \begin{cases} R_H & ; A^t = A_H \\ \gamma_m \beta^t R_M & ; A^t = \frac{k\pi}{4}, k \in \mathbb{N} \end{cases} \quad (13)$$

$$\beta^t = 1 - \cos(A^t - \angle W^t)$$

Moreover, in the case of using belief maps, we add a similarity reward that shows the accuracy of a belief about an area after observing it within the FOV.

$$\mathbf{R}_{\text{inf}}^t = -\alpha_{\text{bel}} I(b_{\text{FOV}}; z_{\text{FOV}}) \quad (14)$$

$$I(X; Y) = D_{\text{KL}}(P_{XY}(x, y) || P_X(x) P_Y(y))$$

In Eq. 14, I represents mutual information between two random variables (vectors), which itself is the Kullback-Leiber (KL) divergence of their joint probability density and the product of their marginal probabilities. After aggregating the aforementioned partial rewards ($R_{\text{obj}}^t, R_{\text{cstr}}^t, R_{\text{inf}}^t$), the aggregation weights ($\alpha_{\text{det}}, \alpha_{\text{mon}}, \dots, \alpha_{\text{bel}}, \gamma_m$) should be hyper-tuned to obtain desirable results. Moreover, they can be adjusted dynamically through an episode to model the priority of the objectives or constraints in different phases of the mission. This dynamic focus helps the agent prevent the challenge of reward aggregation in multi-objective multi-constrained problems.

IV. PROPOSED METHOD

In this section, the two mission phases of the UAV are described. Next, the DQN used for value estimation is discussed. Finally, the solution to the POMDP approach using belief maps, in terms of state representation, is explained.

A. Mission Phases

Weight initialization is crucial in neural networks, including DQNs, due to the challenges posed by numerous local minimums in non-convex functions. Prior information plays a vital role, guiding the value network in the parameter space to be closer to true values. Our monitoring method adopts a 'Scan-and-Track' bi-phase approach, enabling the UAV to effectively initialize beliefs about the environment, countering the issue of initial ignorance of fire locations.

1) **Scanning Phase:** In the scanning phase, the agent calculates the shortest path along the whole environment based on its starting location, FOV size, and environment size, to create an initial fire grid map and update the wildfire model. This phase, conducted only in the first episode of each epoch, serves as DQN weight initialization. Here, the UAV follows a predetermined path, while enabling policy evaluation to enhance value estimates.

2) **Tracking Phase:** After one round of scanning the environment, we start the first episode of training by executing policies with an epsilon-greedy exploration-exploitation approach, where the epsilon is set to decay in a range of episodes in each epoch.

B. State Representation

As discussed previously, the observations in a POMDP are a function of the true state and several sources of error in between. In this section, we will discuss different approaches toward state representation as inputs to value/policy networks in the wildfire monitoring problem.

1) **Observation-Based Representation:** As the UAV moves from one area to another, the observations become outdated and the observed state of a cell within the old area at a specific time $Z_{i,j}^t$ is no longer a good estimate for it at a further time $Z_{i,j}^{t+k}$. By constructing an observation map Z^t which gets updated by replacing observations within the FOV at each time step, the current state of the environment is tracked with a rough estimate for slow progression scenarios or large FOVs.

Certainty Factor. To consider the uncertainty of observation of the past time t_{obs} , at the current time t , regarding the progress of the fires within the observed area, a certainty value for each cell is defined as follows:

$$c_{i,j}^t = 1 - \frac{t - t_{\text{obs}}}{t_{\text{max}}} \quad (15)$$

An element-wise multiplication of the certainty values and the observation matrix obtains an uncertainty-aware observation map. ($\tilde{Z}_{i,j}^t = Z_{i,j}^t c_{i,j}^t$). By feeding this compensated map to the value/policy network, the network focuses on the areas with higher certainty and adapts better to highly dynamic cases.

2) **Belief-Based Representation:** Here an alternative to representing the environment state in dynamic scenarios is proposed, in which the input to the value/policy network is the probability of the state being in the ignited states, and is defined as the belief state ($\mathbf{b}_{ij}^t = \text{Pr}\{F_{ij}^t = 1\}$). This probability is assigned by the agent based on a sequence of observations and its initial prior which comes from information about the vegetation type and density.

Belief Initialization:

In Bayesian models for i.i.d Bernoulli events, initial Beta distribution parameters α and β are determined by relative success count. However, with fire spread, adjacent cell ignition probabilities become codependent, breaking the Beta distribution's conjugacy. However, we use a Beta distribution initialized with known vegetation density and type as an improved prior to obtain smoother, faster and more stable convergence (quasi-convergence for highly dynamic cases).

$$\mathbf{b}_{ij}^0 = \text{Beta}(\alpha_{ij}, \beta_{ij}), \quad \alpha_{ij} = \frac{V_{ij}}{\rho_{ij}} \alpha_0, \quad \beta_{ij} = \frac{\rho_{ij}}{V_{ij}} \beta_0 \quad (16)$$

Belief Update: The belief update model represents the dynamics of the environment learned by the agent through samples collected at each time step. First, the limitations of Bayesian updates will be discussed and next, a beta-binomial model is proposed to approximate for belief updates.

1. **Bayesian Update:** Bayes' rule updates the belief at time t , using ignition probabilities and the likelihood of sustained ignition, as shown by the equation for belief updates. The sequential ignition process, allows belief updates based on

only ignition and burnout at time t .

$$\begin{aligned} \mathbf{b}_{ij}^{t+1} &= p(F_{ij}^{t+1} = 1) \\ &= p(F_{ij}^{t+1} = 1 | F_{ij}^t = 0)(1 - b_{ij}^t) \\ &\quad + p(F_{ij}^{t+1} = 1 | F_{ij}^t = 1)b_{ij}^t \\ &= (1 - \mathbf{b}_{ij}^t) \mathbf{p}_{ij}^{01t} + \mathbf{b}_{ij}^t (1 - \mathbf{p}_{ij}^{12t}) \end{aligned} \quad (17)$$

In Eq. 17, p_{ij}^{01t} and p_{ij}^{12t} denote the ignition and burnout probability of a cell at time t , respectively. The ignition probability at time t is a weighted sum of the fire spread probability from one cell to another one (Eq. 6) and the burnout process is a deterministic process that happens as soon as the fuel of a cell finishes.

$$\begin{aligned} \mathbf{p}_{ij}^{01t} &= \sum_{i',j'} p(F_{ij}^{t+1} = 1 | F_{ij}^t = 0, F_{i'j'}^t = 1) p(f_{i'j'}^t = 1) \quad (18) \\ &= \sum_{i',j'} b_{i'j'}^t p(f_{ij}^{t+1} = 1 | f_{ij}^t = 0, f_{i'j'}^t = 1) \\ &\stackrel{\text{Eq.2}}{\Rightarrow} \sum_{i',j'} b_{i'j'}^t \left(\frac{1}{2\sigma^2} e^{-\frac{d^2}{2\sigma^2}} \right) \left(\frac{W_{ij}^t}{\max(W_{ij}^t)} \right) \left(1 - \frac{f_{ij}^t}{f_{i'j'}^0} \right) \end{aligned}$$

Considering the perfect wind measurement in ignited cells, the remaining fuel $f_{i,j}^t$, yet remains unknown. Moreover, the burnout likelihood (p_{ij}^{12t}) of every cell is unknown despite the fuel consumption behavior known by the agent. This is due to the inconsistent monitoring of a cell and the inherent Gaussian noise implemented in the simulated data, accounting for other variables like heat and oxygen flow affecting consumption in reality. Thus, performing the Bayesian update is impossible without approximating wind and fuel measurements at given locations.

2. Heuristic Approach: An approximation to the Bayesian update for the Beta conjugate prior is to increase alpha and beta by the number of successes and failures of a Bernoulli trial in a group of i.i.d observations respectively. In the equation below, we consider N to be the number of cells observed within a FOV $N = (l_{FoV})^2$.

$$\alpha' = \alpha + \sum_{i=1}^K x_i \quad \beta' = \beta + N - \sum_{i=1}^K x_i \quad (19)$$

C. Deep Q-Learning

In large state and action spaces, tabular Q-learning is no longer a solution due to large state-action spaces. [14]. A Deep Q-network (DQN) is used in this case to approximate the true Q-values. This DQNs architecture is designed to take the observations/beliefs along with the UAV state parameters in two separate branches and fuse them later on. On one branch, the belief state or the observation (z_t or b_t), is fed through a CNN and compressed into an $8 \times 8 \times 256$ feature map after a few layers. Next, it is flattened and passed to a fully connected layer that outputs the latent representation of the environment in a vector of length 16. On the other branch, the UAV state ($S_{phy} = (x^t, y^t, P^t, \phi_U^t)$) are fed to three consecutive fully connected layers to reach the same dimension of the latent spatial feature vector.

V. EVALUATION

In this section, visual and numerical evaluations of the monitoring goal are presented. For visual evaluation, trajectories of the UAV in static and dynamic environment are shown

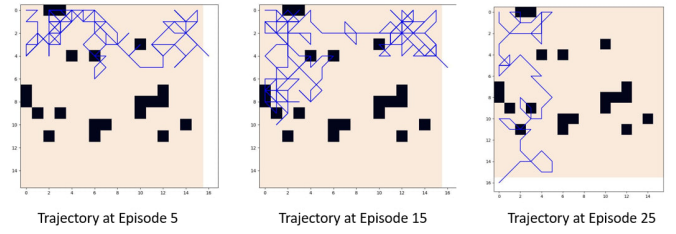


Fig. 4. Trajectories of the UAV in a static environment setting for episodes 5, 15, 25 from left to right. Burnt cells are shown in black. (Trajectory is plotted over the final burnt-out wildfire)

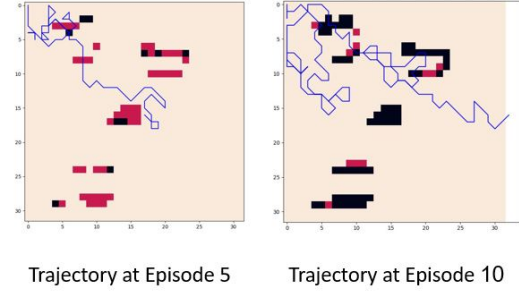


Fig. 5. Bridging over burnt cells - Trajectories of the UAV in a dynamic environment for episodes 5 and 10 from left to right. Burnt cells are shown in black, while ignited cells are shown in red.

with trajectories of the observation and belief representation. The experiment parameter values for the discussed results are shown in Table III.

A. Trajectory Analysis.

In static fire scenarios (Fig. 4), the UAV learns to identify paths with higher q-values, initially exploring more near the take-off area and later expanding its trajectory for broader exploration. In dynamic fire scenarios (Fig. 5), within 10 episodes, the UAV adapts to navigate using burnt areas as safe paths in a sparser fire (32×32 grid with 5×5 FOV). In the radial fire spread setting (Fig. 6), ($A_{Max} = 100, \sigma_{spr} = 1$), initial strategies using the observation map proved inefficient, but later, using belief states, the UAV learned to identify and use burnt areas as safe passages, optimizing its path in later episodes.

B. Coverage and Frontline Tracking Criteria.

To compare the UAVs performance in dynamic scenarios for belief v.s. observation state, two criteria are defined. First, the number of detected fire batches across the total grid ($\%Det = \frac{n_{det}}{n_{tot}}$). Second, for monitoring a new criterion called *MIA* (Monitored Ignited Area) is introduced which considers the percentage of ignited area under cover for each batch fire ($\frac{n_b}{l_{FoV}}$) and combines it with the normalized minimum distances (d_b) to their frontlines. MIA is calculated using Eq. 20. The results are summarized in Table II.

$$\begin{aligned} \mathbf{MIA} &= \mathbf{E}_{\forall b \in B} \left(\frac{n_b}{l_{FoV}^2} \frac{d_b^{max}}{d_b^{min}} \right) = \mathbf{E}_{\forall b \in B} \left(\frac{n_b}{l_{FoV}^2} \frac{l_{FoV}}{\sqrt{2} d_b^{min}} \right) \\ &= \frac{l_{FoV}}{\sqrt{2}} \mathbf{E}_{\forall b \in B} \left(\frac{n_b}{d_b^{min}} \right) \leq \frac{l_{FoV}^3}{\sqrt{2}} \end{aligned} \quad (20)$$

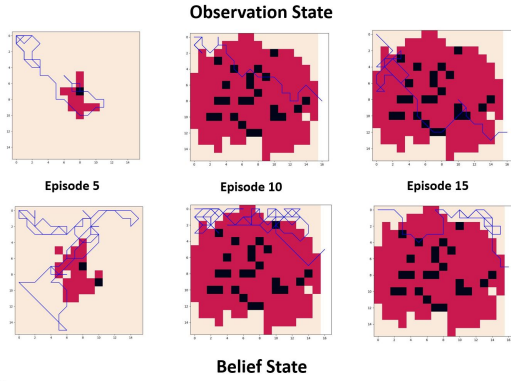


Fig. 6. Trajectory of the UAV, which learns in a dynamic radial spread setting with observation and belief maps for a 16×16 grid. (3×3 FOV and constant wind magnitude of $A = 0.8A_{max}$)

In the equation above, d_b^{max} is half the diagonal of the FOV, which equals the maximum distance of the UAV to an observed ignited cell. n_b and l_{FOV} respectively represent the number of covered cells and the size of the field of view

TABLE II
EVALUATION RESULTS FOR A 16×16 GRID WITH A 5×5 FOV

Evaluation Criteria	State Representation	Multi-Fire	
		Static (Batch)	Dynamic
Fire Coverage Ratio	Observation	86.2%	77.3%
	Belief	82.5%	79.4%
Time-Average MIA	Observation	23.25	11.91
	Belief	18.41	14.16

As seen in Table II, in static (slow) scenarios observation states yield a higher coverage ratio and MIA, compared to belief states. This approves expectations as the observations accurately represent the environment after the UAV covers the grid for the first time. In the dynamic case, however, as the observations become outdated, the belief helps the agent assign higher Q-values to actions for tracking the front line, instead of dangerously monitoring it from above.

VI. CONCLUSION

This paper develops a belief-based DRL solution for UAV path planning in dynamic forest wildfires considering various

TABLE III
EXPERIMENTAL SETTINGS FOR RESULTS IN TABLE II

Envir. Settings	Value	Agent Settings	Value
Grid Size	16	FOV	5
# Initial Ignitions	10	Steer limit	180°
# Veg. Patches	5	Detection Reward	10
Wind Mag. Var. Period	20	Monitoring Reward	10
Wind Phase Var. Period	80	Mvm/Hov Power Ratio	2
Wind Mag. Amp.	80	Belief Reward	40
Wind Mag. Var. Amp.	20	Burnout Penalty	-200
Wind Mag. Max	100	Burnout Limit	10
Num. Episodes	20	Max. Iterations	500

factors contributing to fire spread including the wind, and vegetation as well as the constraints of low-altitude drones (limited flight time and field of view). The belief-based state representation in such highly dynamic and partially observable environment where key factors of fire spread are hidden to the UAV with limited sensing and vision capabilities shows promise, by implicitly learning the wildfire spread model through estimating the ignition probability. The belief framework offers a memory-efficient statistic of the POMDP history suitable for low-altitude UAVs. Moreover, it considers the uncertainty of outdated regions through a certainty factor and offers tunable reward balance between objectives and constraints. Although this approach, may get limited by multi-modal highly co-variated data which results in complex spatial dependencies, but is capable of adapting to several monitoring tasks and scenarios. It is worth noting that the proposed method focuses on single-agent RL to exhibit state representation potential and future works are encouraged to extend this belief-based model to frameworks with multiple agents such as dec-POMDPs.

REFERENCES

- [1] A. Smith, "2010–2019: A landmark decade of us. billion-dollar weather and climate disasters," *National Oceanic and Atmospheric Administration*, 2020.
- [2] S. P. H. Boroujeni, A. Razi, S. Khoshdel, F. Afghah, J. L. Coen, L. O'Neill, P. Z. Fule, A. Watts, N.-M. T. Kokolakis, and K. G. Vamvoudakis, "A comprehensive survey of research towards ai-enabled unmanned aerial systems in pre-, active-, and post-wildfire management," 2024.
- [3] A. Shamsoshoara, F. Afghah, A. Razi, L. Zheng, P. Z. Fulé, and E. Blasch, "Aerial imagery pile burn detection using deep learning: The flame dataset," *Computer Networks*, vol. 193, p. 108001, 2021.
- [4] X. Chen, B. Hopkins, H. Wang, L. O'Neill, F. Afghah, A. Razi, P. Fulé, J. Coen, E. Rowell, and A. Watts, "Wildland fire detection and monitoring using a drone-collected rgb/ir image dataset," *IEEE Access*, vol. 10, pp. 121 301–121 317, 2022.
- [5] M. A. Finney, "Landscape fire simulation and fuel treatment optimization," *Methods for integrating modeling of landscape change: Interior Northwest Landscape Analysis System. Gen. Tech. Rep. PNW-GTR-610. Portland, OR: US Department of Agriculture, Forest Service, Pacific Northwest Research Station*, pp. 117–131, 2004.
- [6] —, *FARSITE, Fire Area Simulator—model development and evaluation*. US Department of Agriculture, Forest Service, Rocky Mountain Research Station, 1998, no. 4.
- [7] H. X. Pham, H. M. La, D. Feil-Seifer, and M. C. Deans, "A distributed control framework of multiple unmanned aerial vehicles for dynamic wildfire tracking," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 50, no. 4, pp. 1537–1548, 2018.
- [8] Z. Lin, H. H. Liu, and M. Wotton, "Kalman filter-based large-scale wildfire monitoring with a system of uavs," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 1, pp. 606–615, 2018.
- [9] E. Seraj, A. Silva, and M. Gombolay, "Multi-uav planning for cooperative wildfire coverage and tracking with quality-of-service guarantees," *Autonomous Agents and Multi-Agent Systems*, vol. 36, no. 2, p. 39, 2022.
- [10] A. Giuseppi, R. Germanà, F. Fiorini, F. Delli Priscoli, and A. Pietrabissa, "Uav patrolling for wildfire monitoring by a dynamic voronoi tessellation on satellite data," *Drones*, vol. 5, no. 4, p. 130, 2021.
- [11] A. T. Azar, A. Koubaa, N. Ali Mohamed, H. A. Ibrahim, Z. F. Ibrahim, M. Kazim, A. Ammar, B. Benjdira, A. M. Khamis, I. A. Hameed, and G. Casalino, "Drone deep reinforcement learning: A review," *Electronics*, vol. 10, no. 9, 2021. [Online]. Available: <https://www.mdpi.com/2079-9292/10/9/999>
- [12] K. D. Julian and M. J. Kochenderfer, "Distributed wildfire surveillance with autonomous aircraft using deep reinforcement learning," *Journal of Guidance, Control, and Dynamics*, vol. 42, no. 8, pp. 1768–1778, 2019.
- [13] P. Shobeiry, M. Xin, X. Hu, and H. Chao, *UAV Path Planning for Wildfire Tracking Using Partially Observable Markov Decision Process*. [Online]. Available: <https://arc.aiaa.org/doi/abs/10.2514/6.2021-1677>
- [14] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing atari with deep reinforcement learning," *arXiv preprint arXiv:1312.5602*, 2013.