Neural Tree Expansion for Multi-Robot Planning in Non-Cooperative Environments

Benjamin Rivière[1], Wolfgang Hönig[1], Matthew Anderson[1], and Soon-Jo Chung[1],

***Abstract*—We present a self-improving, Neural Tree Expansion (NTE) method for multi-robot online planning in non-cooperative environments, where each robot attempts to maximize its cumulative reward while interacting with other self-interested robots. Our algorithm adapts the centralized, perfect information, discrete-action space method from AlphaZero to a decentralized, partial information, continuous action space setting for multi-robot applications. Our method has three interacting components: (i) a centralized, perfect-information "expert" Monte Carlo Tree Search (MCTS) with large computation resources that provides expert demonstrations, (ii) a decentralized, partial-information "learner" MCTS with small computation resources that runs in real-time and provides self-play examples, and (iii) policy & value neural networks that are trained with the expert demonstrations and bias both the expert and the learner tree growth. Our numerical experiments demonstrate Neural Tree Expansion's computational advantage by finding better solutions than a MCTS with 20 times more resources. The resulting policies are dynamically sophisticated, demonstrate coordination between robots, and play the Reach-Target-Avoid differential game significantly better than the state-of-the-art control-theoretic baseline for multi-robot, double-integrator systems. Our hardware experiments on an aerial swarm demonstrate the computational advantage of Neural Tree Expansion, enabling online planning at 20 Hz with effective policies in complex scenarios.**

***Index Terms*—Distributed Robot Systems, Motion and Path Planning, Reinforcement Learning**

I. Introduction

MULTI-AGENT interactions in non-cooperative environments are ubiquitous in robotic applications such as self-driving, space exploration, urban air mobility, and human-robot collaboration. Planning, or sequential decision-making, in these settings requires a prediction model of the other agents, which can be generated through a game theoretic framework.

Recently, the success of AlphaZero [1] at the game of Go has popularized a self-improving machine learning algorithm: bias a Monte Carlo Tree Search with value and policy neural networks, use the tree statistics to train the networks with supervised learning and then iterate over these two steps to improve the policy and value networks over time. However, this algorithm is designed for classical artificial intelligence tasks (e.g. chess or Go), and applications in multi-robot domains require different assumptions: continuous state-action, decentralized evaluation, partial information, and limited computational resources. To the best of our knowledge, our work is the first to provide a complete multi-robot adaption from algorithm design to hardware experiment of the AlphaZero method.

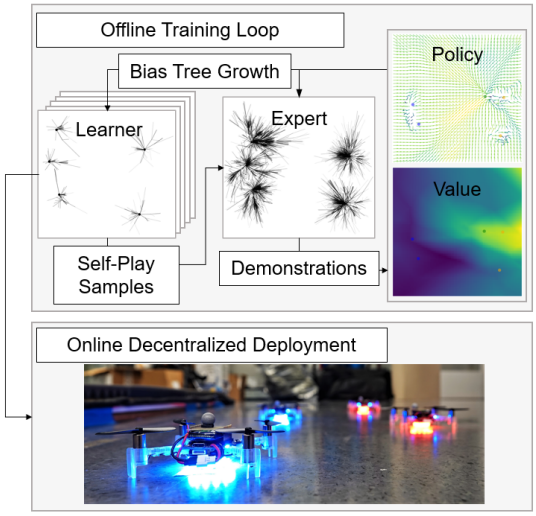


Fig. 1. We propose an AlphaZero-like method for multi-robot applications such as the Reach-Target-Avoid game. Offline, the learner finds relevant states through self-play, for which the expert generates demonstration data. The data is used to train policy and value neural networks, which are used to bias both tree growths. At runtime, the learner is deployed on each robot to generate an action with local information and little computation expense.

The overview of our algorithm is shown in Fig. 1. The key algorithmic innovation of our approach is to create two distinct MCTS policies to bridge the gap between high-performance simulation and real-world robotic application: the "expert" tree search is centralized and has access to perfect information and large computational resources, whereas the "learner" tree search is decentralized and has access to partial information and limited computational resources. During the offline phase, the neural networks are trained in an imitation learning style using the self-play states of the learner and the high-quality demonstrations of the expert. The expert's high-quality demonstrations enable policy improvement through iterations to incrementally improve the policy and value networks. The learner's self-play samples states that should appear more

Manuscript received: February, 21, 2021; Revised May 26, 2021; Accepted June 24, 2021.

This paper was recommended for publication by Editor Ani Hsieh upon evaluation of the Associate Editor and Reviewers' comments. This work was supported by Defense Advanced Research Projects Agency (DARPA). The views, opinions and/or findings expressed are those of the authors and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government. Preliminary work was in part funded by Raytheon. Video: https://youtu.be/mklbTfWl7DE. Code: https://github.com/bpriviere/decision_making.

[1] Graduate Aerospace Laboratories of the California Institute of Technology {briviere, whoenig, matta, sjchung}@caltech.edu.

Digital Object Identifier (DOI): see top of this page.

frequently at runtime. At deployment, each robot uses the learner to effectively plan online with partial information and limited computational budget. Our contribution is the Neural Tree Expansion algorithm that extends AlphaZero methods to (i) decentralized evaluation with local information, (ii) continuous state-action domain, and (iii) limited computational resources.

We validate our method in simulation and experiment. We demonstrate numerically that our approach generates compact trees of similar or better performance with 20 times fewer nodes, and the resulting policies play the Reach-Target-Avoid differential game with double-integrator dynamics significantly better than the current state of the art [2]. Our method is compatible with arbitrary game specifications; we demonstrate this by generating visual examples of canonical games in Fig. 2 and empirical evaluation of the Reach-Target-Avoid game for double-integrator and 3D Dubin's vehicle dynamics in Sec. IV. Our hardware experiments demonstrate that the solutions are robust to the gap between simulation and real world and neural expansion generates compact search trees that are effective real-time policies.

Related Work: Our work relates to multiple communities: planning, machine learning, and game theory. Planning, or sequential decision-making, problems can be solved in an online setting with Monte Carlo Tree Search (MCTS) [3]. MCTS searches through the large decision-making space by rolling out simulated trajectories and biasing the tree growth towards areas of high reward [4]. MCTS was first popularized by the Upper Confidence Bound for Trees algorithm [5] that uses a discrete-action, multi-armed bandit solution to balance exploration and exploitation in node selection. Recent work uses a non-stationary bandit analysis to propose a polynomial, rather than logarithmic, exploration term [6]. As an anytime algorithm, the space and time complexity of MCTS is user-determined by the desired number of simulations. Recent finite sample complexity results of MCTS [6, 7] show the error in root node value estimation converges at a rate of the order $n^{-1/2}$ where n is the number of simulations.

Application of MCTS to a dynamically-constrained robot planning setting requires extending the theoretical foundations to a continuous state and action space. In general, the recent advances in this area answer two principal questions: i) how to select an action, and ii) when a node is fully expanded. Regarding the former question, some solutions select an action using the extension of the multi-armed bandit in continuous domains [8], whereas our approach uses a policy network to generate actions. Regarding the latter question, a popular method to determine whether a node is expanded is to use progressive widening and variants; we adapt one such method, the Polynomial Upper Continuous Trees (PUCT) algorithm [9]. Despite the advance in theory for continuous action spaces, there have been relatively few studies of biasing continuous MCTS with deep neural networks [10].

The key idea of AlphaZero [1] is using MCTS as a policy improvement operator; i.e. given a policy neural network to guide MCTS, the resulting search produces an action closer to the optimal solution than that generated by the neural network. Then, the neural network is trained with supervised learning to imitate the superior MCTS policy, matching the quality of the network to that of MCTS in the training domain. By iterating over these two steps, the model improves over time. The first theoretical analysis of this powerful method is recently shown for single-agent discrete action space problems [6]. In comparison, our method is applied to a continuous state-action, multi-agent setting. Whereas AlphaZero methods use the policy network to bias the node selection process, i.e. given a list of actions, select the best one, our policy network is an action generator for the expansion process to create edges to children, i.e. given a state, generate an action. A neural expansion operator has previously been explored in motion planning [11], but not decision-making. In addition, our method's supervised learning step is closer to imitation learning, as used in DAgger [12], because the learner benefits from an adaptive dataset generation of using self-play to query from an expert.

Although the AlphaZero methods use a form of supervised learning to train the networks, they can be classified as a reinforcement learning method because the networks are trained without a pre-existing labelled dataset. Policy gradient [13] is a conventional reinforcement learning solution and there are many recent advances in this area [14]. Adding an underlying tree structure to deep reinforcement learning provides a higher degree of interpretability and a more stable learning process, enabled by MCTS's policy improvement property.

In contrast to data-driven methods, traditional analytical solutions can be studied and derived through differential game theory. The game we study, Reach-Target-Avoid, was first introduced and solved for simple-motion, 1 vs. 1 systems [15]. Later, multi-robot, single-integrator solutions have been proposed [16, 17]. Solutions considering multi-robots with non-trivial dynamics, such as the double-integrator [2], are an active area of research. Shepherding, herding, and perimeter defense are variants of the Reach-Target-Avoid game and are also active areas of research [18–21].

II. Problem Formulation

Notation: We denote the learning iteration with k and the physical timestep with a subscript t, which is suppressed for notation simplicity, unless necessary. Robot-specific quantities are denoted with i or j superscript, and, in context, the absence of superscript denotes a joint-space quantity, e.g. the joint state vector is the vertical stack of all individual robot vectors, $\mathbf{s}_t = [\mathbf{s}_t^1; \dots; \mathbf{s}_t^N]$ where N is the number of robots.

Definition 1: A partially observable stochastic game (POSG) is defined by a tuple: $\mathcal{G} = \langle \mathcal{I}, \mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \mathcal{Z}, \mathcal{O}, H \rangle$ where: $\mathcal{I} = \{1, ..., N\}$ is the set of robot indices, $\mathcal{S}$ is the set of joint robot states, $\mathcal{A}$ is the set of joint robot actions, $\mathcal{T}$ is the joint robot transition function where $\mathcal{T}(\mathbf{s}_t, \mathbf{a}_t, \mathbf{s}_{t+1}) = \mathbb{P}(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)$ is the probability of transitioning from joint state $\mathbf{s}_t$ to $\mathbf{s}_{t+1}$ under joint action $\mathbf{a}_t$, $\mathcal{R}$ is the set of joint robot rewards functions where $\mathcal{R}^i(\mathbf{s}, \mathbf{a}^i)$ is the immediate reward of robot i for taking local action $\mathbf{a}^i$ in joint state $\mathbf{s}$, $\mathcal{Z}$ is the set of joint robot observations, $\mathcal{O}$ is the set of joint robot observation probabilities where $\mathcal{O}(\mathbf{z}, \mathbf{s}, \mathbf{a}) = \mathbb{P}(\mathbf{z}|\mathbf{s}, \mathbf{a})$ is the probability of observing joint observation $\mathbf{z}$ conditioning on

the joint state and action, and H is the planning horizon. These joint-quantities can be constructed from the robot-specific quantity. The solution of a POSG is a sequence of actions that maximizes the expected reward over time and is often characterized by a policy or value function.

Assumption 1: A deterministic transition function is assumed, thereby permitting rewriting $\mathcal{T}$ with a dynamics function, f, as: $\mathcal{T}(\mathbf{s}_t, \mathbf{a}_t, \mathbf{s}_{t+1}) = \mathbb{I}(f(\mathbf{s}_t, \mathbf{a}_t) = \mathbf{s}_{t+1})$ where $\mathbb{I}$ is an indicator function. Similarly, rewriting $\mathcal{O}$ with a deterministic observation function results in $\mathbf{z}^i = h^i(\mathbf{s})$ for each robot i. Assumption 1 can be relaxed by considering specialized variants that are not the focus of this work. For example, realistic robotic scenarios with localization uncertainty from measurement noise can be handled with the observation widening variant [22].

Problem Statement: At time t, each robot i makes a local observation, $\mathbf{z}^i$, uses it to formulate an action, $\mathbf{a}^i$, and updates its state, $\mathbf{s}^i$, according to the dynamical model. Our goal is to find policies for each robot, $\pi^i : \mathcal{Z}^i \to \mathcal{A}^i$ that synthesizes actions from local observations through:

$$\mathbf{z}_t^i = h^i(\mathbf{s}_t), \quad \mathbf{a}_t^i = \pi^i(\mathbf{z}_t^i), \tag{1}$$

to approximate the solution to the general-sum, game theoretic optimization problem:

$$\begin{aligned} \mathbf{a}_t^{i^*} &= \underset{\{\mathbf{a}_\tau^i | \forall \tau\}}{\arg\max} \sum_{\tau=t}^{t+H} \mathcal{R}^i(\mathbf{s}_\tau, \mathbf{a}_\tau^i) \quad \text{s.t.} \\ \mathbf{s}_{\tau+1} &= f(\mathbf{s}_\tau, \mathbf{a}_\tau),\ \mathbf{s}_\tau^i \in \mathcal{X}^i,\ \mathbf{a}_\tau^i \in \mathcal{U}^i,\ \mathbf{s}_{\tau_0}^i = \mathbf{s}_0^i,\ \forall i, \tau \end{aligned} \tag{2}$$

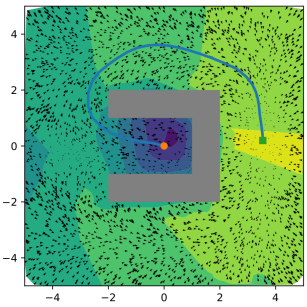
where $\mathcal{U}^i \subseteq \mathcal{A}^i$ is the set of available actions (e.g. bounded control authority constraints), and $\mathcal{X}^i \subseteq \mathcal{S}^i$ is the set of safe states (e.g. collision avoidance) and $\mathbf{s}_0^i$ is the initial state condition. The optimization problems for each robot i are simultaneously coupled through the evolution of the global state vector $\mathbf{s}$, where each robot attempts to maximize its own reward function $\mathcal{R}^i$. We evaluate our method on canonical cases and present visualizations in Fig. 2.

Reach-Target-Avoid Game: An instance of the above formulation is the Reach-Target-Avoid game [15] for two teams of robots, where team A gets points for robots that reach the goal region, and team B gets points for defending the goal by tagging the invading robots first. The teams are parameterized by index sets $\mathcal{I}_A$ and $\mathcal{I}_B$, respectively, where the union of the two teams represents all robots, $\mathcal{I}_A \cup \mathcal{I}_B = \mathcal{I}$. An example of the Reach-Target-Avoid game is shown in Fig. 2c, where the red robots try to tag the blue robots before the blue robots reach the green goal region. The x and o on the trajectory indicates tagged state and reached goal.

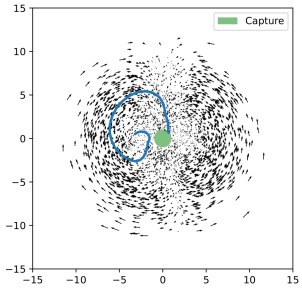
Dynamics: We consider the discrete-time double-integrator system for the i^{th} robot as a motivating example:

$$\mathbf{s}_{t+1}^i = \begin{bmatrix} \mathbf{p}_t^i \\ \mathbf{v}_t^i \end{bmatrix} + \begin{bmatrix} \mathbf{v}_t^i \\ \mathbf{a}_t^i \end{bmatrix} \Delta_t \tag{3}$$

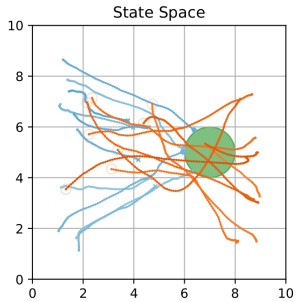
where $\mathbf{p}^i$ and $\mathbf{v}^i$ denote position and velocity and Δ_t denotes the simulation timestep. We use a simultaneous turn game formulation where at a given timestep, each team's action is chosen without knowledge of the other team's action.



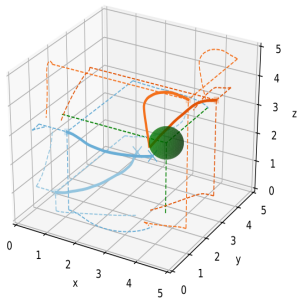
(a) NTE finds the intuitive value and policy function for the "bugtrap" motion planning problem. The robot starts at the orange dot, and terminates at the green square after reaching the goal.



(b) The state trajectories generated by NTE approximate the primary solution and barrier surface for the "homicidal chauffeur" game [15]. The plot is shown in Isaac's reduced space with an example trajectory in blue terminating in a capture condition.



(c) NTE scales to high dimensional team games for the 10 agent vs. 10 agent "Reach-Target-Avoid" with double-integrator dynamics.



(d) NTE is compatible with arbitrary dynamics, here is the "Reach-Target-Avoid" game with 3D Dubin's vehicle dynamics.

Fig. 2. Neural Tree Expansion (NTE) can be applied to decision-making problems such as single agent motion planning, canonical differential games, and high-dimensional team games.

Admissible State and Action Space: The admissible state space for each robot is defined by the following constraints: remain inside the position and velocity bounds, $\overline{p}$ and $\overline{v}$, and avoid collisions within the physical robot radius, r_p:

$$\|\mathbf{p}^i\|_\infty \leq \overline{p},\ \|\mathbf{v}^i\|_2 \leq \overline{v},\quad \|\mathbf{p}^j - \mathbf{p}^i\| > r_p,\ \forall i, j \in \mathcal{I} \tag{4}$$

For each robot i on team A ($\forall i \in \mathcal{I}_A$), the admissible state space has an additional constraint: avoid the robots on team B by at least the tag radius:

$$\|\mathbf{p}^i - \mathbf{p}^j\| > r_t, \quad \forall j \in \mathcal{I}_B \tag{5}$$

Then, the admissible state space for each team can be written compactly, e.g. $\mathcal{X}^i = \{\mathbf{s}^i \in \mathcal{S}^i \mid \text{s.t. (4), (5)}\}$, $\forall i \in \mathcal{I}_A$. The admissible action space for each robot is constrained by its acceleration limit: $\mathcal{U}^i = \{\mathbf{a}^i \in \mathcal{A}^i \mid \|\mathbf{a}^i\|_2 \leq \overline{a}\}$, $\forall i \in \mathcal{I}$, where $\overline{a}$ is the robot's acceleration limit. When a robot exits the admissible state or action space, or a robot on team A's position is within an r_g radius about the goal position $\mathbf{p}_g$, it becomes inactive.

Observation Model: Under Assumption 1, for each robot i, we define a measurement model that is similar to visual relative navigation $h^i : \mathcal{S} \to \mathcal{Z}^i$, which measures the relative state measurement between neighboring robots, as well as relative state to the goal. Specifically, an observation is defined as:

$$\mathbf{z}^i = \left[\mathbf{g} - \mathbf{s}^i, \{\mathbf{s}^j - \mathbf{s}^i\}_{j \in \mathcal{N}_A^i}, \{\mathbf{s}^j - \mathbf{s}^i\}_{j \in \mathcal{N}_B^i}\right], \tag{6}$$

where $\mathbf{g}$ is the goal position embedded in the state space, e.g. for 2D double-integrator $\mathbf{g} = [\mathbf{p}_g; 0; 0]$. Then, $\mathcal{N}_A^i$ and $\mathcal{N}_B^i$ denote the i^{th} robot's neighbors on team A and B, respectively. These sets are defined by each robot's sensing radius, r_{sense}:

$$\mathcal{N}_A^i = \{j \in \mathcal{I}_A \mid \|\mathbf{p}^j - \mathbf{p}^i\|_2 \leq r_{\text{sense}}\}. \tag{7}$$

Reward: The robot behavior is driven by the reward function; the inter-team cooperation behavior is incentivized by sharing the reward function and intra-team adversarial behavior is incentivized by assigning complementary reward functions only dependent on global state, $\mathcal{R}^i(\mathbf{s}) = -\mathcal{R}^j(\mathbf{s}),\ \forall i \in \mathcal{I}_A, j \in \mathcal{I}_B$. The reward can be defined by a single, robot-agnostic game reward, $\mathcal{R}(\mathbf{s})$ that team A tries to maximize and team B tries to minimize. The game reward is 0 until the terminal state where the game reward and traditional value function are identical and defined as:

$$V(\mathbf{s}_t) = \sum_{\forall i \in \mathcal{I}_A} \mathbb{I}(\|\mathbf{p}_t^i - \mathbf{p}_g\|_2 \leq r_g), \tag{8}$$

i.e. the value is the number of team A robots in the goal region. The indicator function payoff is known to be sparse and makes traditional search and reinforcement learning techniques ineffective [23]. The game termination occurs when all robots on team A are inactive; typically when they have reached the goal or been tagged by a robot on team B.

III. NTE Algorithm Description

We present the meta-algorithm, the expert and learner NTE, and the policy and value neural networks.

A. Meta Self-Improving Algorithm

The input of the meta-learning is the POSG game described in Sec. II, and the outputs are the policy and value neural networks, $\tilde{\pi}$ and $\tilde{V}$. The goal of the meta-learning is to improve the models across learning iterations, especially in relevant state domains, such that at runtime, the robots can evaluate the learner. The desired model improvement can be expressed by decreasing some general probability distribution distance between the policy network and the optimal policy function:

$$\text{dist}(\tilde{\pi}_k(\mathbf{z}), \pi^*(\mathbf{s})) \leq \text{dist}(\tilde{\pi}_m(\mathbf{z}), \pi^*(\mathbf{s})),\ \forall k > m,\ \forall \mathbf{s} \tag{9}$$

where π^* is the unknown optimal policy function that inputs a joint state and returns a joint action, $\tilde{\pi}_k$ is the policy network we train, and k, m are learning iteration indices. Specifically, we train robot-specific policies $\tilde{\pi}_k^i$ that map local observation to local action and compose them together to create the joint policy $\tilde{\pi}_k = [\tilde{\pi}_k^1; \ldots; \tilde{\pi}_k^{|\mathcal{I}|}]$ that maps joint observation, $\mathbf{z}$, to joint action, $\mathbf{a}$. Adapting the proof concept in [6] to our setting, the policy improvement can be shown by validating two properties and then iterating: (i) bootstrap

$$\text{dist}(\pi^*(\mathbf{s}), \pi_k^e(\mathbf{s}, \tilde{\pi}_k, \tilde{V}_k)) \leq \text{dist}(\pi^*(\mathbf{s}), \tilde{\pi}_k(\mathbf{z})),\ \forall \mathbf{s} \tag{10}$$

and, after generating an appropriate dataset, (ii) learning

$$\mathcal{D}_\pi = \{\mathbf{s}, \mathbf{a}\} \text{ with } \mathbf{a} = \pi_k^e(\mathbf{s}, \tilde{\pi}_k, \tilde{V}_k) \tag{11}$$

$$\text{dist}(\tilde{\pi}_{k+1}(\mathbf{z}), \pi^*(\mathbf{s})) \approx \text{dist}(\mathbf{a}, \pi^*(\mathbf{s})),\ \forall (\mathbf{s}, \mathbf{a}) \in \mathcal{D}_\pi \tag{12}$$

where $\tilde{V}_k$ is the value network and π^e is the expert that maps state to joint action while biased by the policy and value neural networks. Intuitively, the bootstrap property of MCTS (10) generates a dataset of policy samples superior to that of the policy network, and then the supervised learning property (12) matches the quality of the policy network to the quality of the new dataset.

Validating these two properties drives the design of our meta-learning algorithm in Algorithm 1. At each learning iteration k, each robot's policy network is trained in the following manner: a set of states is generated from self-play of the multi-agent learners, π_k^{le} and then the expert, π_k^e, searches on these states to create a dataset of learning targets for the supervised learning. The value network, $\tilde{V}_k$ is trained to predict the outcome of the game if each team were to play with the joint policy network $\tilde{\pi}_k$. Because the centralized expert has perfect information, coordination and large computational resources, the bootstrap property is more likely to hold. Furthermore, as the learner generates state space samples through self-play, the dataset is dense in frequently visited areas of the state space, and the models will be more accurate there, validating the learning property. Finally, we specify a simple POSG generator in Line 4 of Algorithm 1 to select opponent policies and game parameters for self-play.

Algorithm 1: Meta Self-Improving Learning

```
1  def Meta-Learning(G*):
2      π̃_0, Ṽ_0 = None, None
3      for k = 0, ..., K do
4          {G}_k = makePOSG(G*, k)
5          for i ∈ I do
6              {s} = selfPlay(π^le((.), π̃_k, Ṽ_k), {G}_k)
               /* Search from Algorithm 2 */
7              D^i_π = {s, π^e_k.Search(s, π̃_k, Ṽ_k))}
8              π̃^i_{k+1} = trainPolicy(D^i_π)
9          end
10         D_V = {s, selfPlay(π̃_k, {G}_k)}
11         Ṽ_{k+1} = trainValue(D_V)
12     end
```

B. Neural Tree Expansion

In order to specify the expert and learner policies, we first explain their common search tree algorithm shown in Algorithm 2 and adapted from [4] to our setting. For a complete treatment of MCTS, we refer the reader to [4].

The biased MCTS algorithm begins at some start state $\mathbf{s}$ and grows the tree until its computational budget is exhausted, typically measured by the number of nodes in the tree, L. Each node in the tree is a state, $\mathbf{s}$, each edge is an action $\mathbf{a}$, and each child is the new state after propagating the dynamics. Each node in the tree, n, is initialized with a state vector and an action edge to its parent node, n^p, i.e. $\text{n} = \textit{Node}(\mathbf{s}, (\text{n}^p, \mathbf{a}))$. Each node stores the state vector, $S(\text{n})$, the number of visits to the node, $N(\text{n})$, its children set, $C(\text{n})$, and its action set, $A(\text{n}, \text{n}'),\ \forall \text{n}' \in C(\text{n})$. The growth iteration

in the main function, *Search*, has four steps: (i) node selection, *Select*, selects a node to balance exploration of space and exploitation of rewards (ii) node expansion, *Expand*, creates a child node by forward propagating the selected node with an action either constructed by the neural network or by random sampling, (iii) *DefaultPolicy* collects terminal reward statistics by either sampling the value neural network or by rolling out a simulated state trajectory from the new node, and (iv) *Backpropagate* updates the number of visits and cumulative reward up the tree. The action returned by the search is the child of the root node with the most visits. The primary changes we make from standard MCTS are the integration of neural networks, highlighted in Algorithm 2.

The behavior of other agents is modelled in a turn-based fashion: each depth in the tree corresponds to the turn of an agent and their action is predicted by selecting the best node for their cost function. Intuitively, the MCTS search plans for all robots, assuming that they maximize their incentive. MCTS is known to converge to the minimax tree solution [4].

Algorithm 2: Neural Tree Expansion

```
/* π̃, Ṽ from Algorithm 1 */
def Search(s, π̃, Ṽ):
    n_0 ← Node(s, None)
    for l = 1, ..., L do
        n_l ← Expand(Select(n_0, π̃))
        v ← DefaultPolicy(S(n_l))
        Backpropagate(n_l, v)
    return A(n_0, arg max_{n'∈C(n_0)} N(n'))
def Expand(n, π̃):
    α ∼ U(0, 1)
    h if α < β_π then
        a ← [a^1, ..., a^{|I|}],  a^i ∼ π̃^i(h^i(s)),  ∀i ∈ I
    else
        a ← [a^1, ..., a^{|I|}],  a^i ∼ U^i,  ∀i ∈ I
    n' ← Node(f(s, a), (n, a))
    return n'
def DefaultPolicy(s, Ṽ):
    α ∼ U(0, 1)
    if α < β_V then
        v ∼ Ṽ(h_y(s))
    else
        while s is not terminal do
            a ← [a^1, ..., a^{|I|}],  a^i ∼ U^i,  ∀i ∈ I
            s ← f(s, a)
        v ← V(s)
    return v
```

C. Expert NTE

The expert, π^e, is a function from joint state $\mathbf{s}$ to joint action $\mathbf{a}$. The expert computes the action by calling *Search* in Algorithm 2 with a large number of nodes L_{expert}. The expert produces a coordinated team action by selecting the appropriate indices of the joint-space action, where the remaining, unused actions represent the predicted opponent team action. The expert's perfect information, centralized response, and large computational budget is necessary to guarantee the bootstrap property (10). We found that if the expert is given less computational resources, the learning process is not stable and the quality of the policy and value networks deteriorates over learning iterations. Many of the desirable properties of the expert for theoretical performance make it an infeasible solution for multi-robot applications, motivating the design of the learner.

D. Learner NTE

The learner for robot i, π^{le}, is a function from local observation $\mathbf{z}^i$ to local action $\mathbf{a}^i$. The learner computes the action by reconstructing the state from its local observation naively: $\tilde{\mathbf{s}}(\mathbf{z}) = \{\tilde{\mathbf{s}}^j\},\ \forall j \in \mathcal{N}_A^i \cup \mathcal{N}_B^i$, where we assume that the learner has prior knowledge of the absolute goal location. Then, the learner calls *Search* in Algorithm 2 with the estimated state and a small number of nodes L_{learner}, $L_{\text{learner}} < L_{\text{expert}}$. The final action $\mathbf{a}^i$ is selected from the appropriate index of the joint-space action returned by *Search*. Because the learner only selects a single action from the joint-space action, the learner is predicting the behavior of robots on both teams. This communication-less implicit coordination enables operation in bandwidth-limited or communication-denied environments.

E. Policy and Value Neural Networks

We introduce each neural network with its dataset generation and training in Algorithm 1, and its effect on tree growth via integration into Algorithm 2.

Policy Network: The policy network for robot i maps observations to the action distribution for a single robot and is used to create children nodes. The desired behavior of the policy network is to generate individual robot actions with a high probability of being near-optimal expansions given the current observation, i.e. generate edges to nodes with a high number of visits in the expert search.

The dataset for each robot i's policy network is composed of observation action pairs as computed in Line 7 of Algorithm 1. The action label is calculated by querying the expert at some state, extracting the root node's child distribution, and calculating the action label as the first moment of the action distribution, weighted by the relative number of visits:

$$\mathbf{a}_l^i = \sum_{\mathrm{n}' \in C(\mathrm{n}_0)} \frac{N(\mathrm{n}')}{N(\mathrm{n}_0)} A^i(\mathrm{n}_0, \mathrm{n}') \tag{13}$$

where $\mathbf{a}_l^i$ is the action label and n_0 is the root node. Recall that $C(\cdot)$ is a node's set of child nodes, $N(\cdot)$ is the number of visits to a node, and $A^i(\mathrm{n}_0, \mathrm{n}')$ is the ith robot's action from root node n_0 to child node n'. Next, we change the input from state to observation by applying robot i's observation model, $\mathbf{z}_l^i = h^i(\mathbf{s}_l)$. This is a global-to-local learning technique to automatically synthesize local policies from centralized

examples [24]. The collection of observation-action samples can be written in a dataset as $\mathcal{D}_\pi^i = \{(\mathbf{z}_l^i, \mathbf{a}_l^i) |\ l = 1, \ldots\}$.

The policy network training in Line 8 in Algorithm 1 is cast as a multivariate Gaussian learning problem, i.e. the output of the neural network is a mean, μ and variance $\mathbf{\Sigma}$. An action sample $\hat{\mathbf{a}}_l^i \sim \pi^{le}(\mathbf{z}_l^i)$ can then be computed by sampling ϵ and transforming it by the neural network output:

$$\hat{\mathbf{a}}_l^i = \mu(\mathbf{z}_l^i) + \mathbf{\Sigma}(\mathbf{z}_l^i)\epsilon, \quad \epsilon \sim \mathcal{N}(0, I) \tag{14}$$

The input, $\mathbf{z}_l^i$, is encoded with a DeepSet [25] feedforward architecture similar to [24] that is compatible with a variable number of neighboring robots. The maximum likelihood solution to the multivariate Gaussian problem is found by minimizing the following loss function:

$$\mathcal{L} = \mathbb{E} \sum_l (\mathbf{a}_l^i - \mu)^T \mathbf{\Sigma}^{-1} (\mathbf{a}_l^i - \mu) + \frac{1}{2} \ln |\mathbf{\Sigma}| \tag{15}$$

$$\tilde{\pi}^i = \underset{\tilde{\pi}^i \in \Pi^i}{\arg\min}\, \mathbb{E}\ \mathcal{L}(\mu(\mathbf{z}_l^i), \mathbf{\Sigma}(\mathbf{z}_l^i), \mathbf{a}_l^i) \tag{16}$$

where μ and $\mathbf{\Sigma}$ are generated by the neural network given $\mathbf{z}_l^i$, and $\mathbf{a}_l^i$ is the target.

The policy neural network, $\tilde{\pi}^i$, is integrated in Line 9–11 in Algorithm 2 in the expansion operation by constructing a joint-space action from decentralized evaluations of the policy network for all the agents, and then forward propagating that action. We found that using a neural expansion, rather than neural selection as in AlphaZero, is necessary for planning with a small number of nodes in environments with many robots. For example, in a 10 vs. 10 game such as that shown in Fig. 2c, the probability of sampling a control action from a uniform distribution that steers each robot towards the goal within 90 degrees is $(1/4)^{10}$. If the learner policy is evaluated with standard parameters (see Sec. IV-A), it will generate 5 children, which collectively are not likely to contain the desired joint action. Using the neural network expansion operator will overcome this limitation by immediately generating promising child nodes. Similar to [26], the expand operation switches between uniform random and neural network sampling at relative frequency $\beta_\pi \in [0, 1]$. The stochastic nature of both expansion modes enables the tree to maintain exploration.

Value Network: The value network is used to gather reward statistics in place of a policy rollout, and is called in the *DefaultPolicy* in Lines 17–19 of Algorithm 2. The value network uses an alternative state representation to be compatible with the estimated state for local computation:

$$\mathbf{y} = h_y(\mathbf{s}) = \left[\{\mathbf{s}^j - \mathbf{g}\}_{j \in \mathcal{N}_A^i}, \{\mathbf{s}^j - \mathbf{g}\}_{j \in \mathcal{N}_B^i}, n_{rg} \right] \tag{17}$$

where $h_y(\mathbf{s})$ is the alternative observation function and n_{rg} is the number of robots that have already reached the goal. The value network maps this alternative state representation to the parameters of a multi-variate Gaussian distribution. The desired behavior of the value network is to predict the outcome of games if they were rolled out with the current policy network. The value function implementation in *DefaultPolicy* is the same as AlphaZero methods.

The value network dataset in Line 10 of Algorithm 1 is generated for all robots at the same time and is composed of alternative state-value pairs. The $\mathbf{y}_l$ state can be generated from $\mathbf{s}$, and the value label, v_l is generated by self-play with the current policy network. The dataset, $\mathcal{D}_V$ can then be written as $\mathcal{D}_V = \{(\mathbf{y}_l, v_l) |\ \forall l = 1, \ldots\}$. Because AlphaZero methods use a policy selector rather than a generator, the dataset for the value network has to be made by rolling out entire games with MCTS. Instead, we generate the dataset by rolling out the policy network, which is much faster per sample, resulting in less total training time.

The value network is trained in Line 11 of Algorithm 1 with a similar loss function (15) as the policy network, using a learning target of the value labels, v_l, instead of the action $\mathbf{a}_l^i$. The value is also queried from the neural network in a similar fashion (14). The value network uses a similar model architecture as the policy network, permitting variable input size of $\mathbf{y}$. Integration of the value network in *DefaultPolicy* uses the same probabilistic scheme as the policy network with parameter β_V.

IV. Experimental Validation

A. MCTS and Learning Implementation

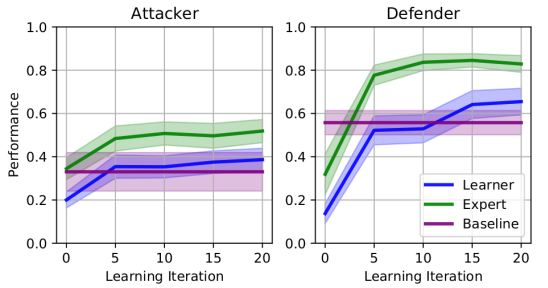
We implement Algorithm 1 in Python and Algorithm 2 in C++ with Python bindings. For the meta-algorithm, we only train the inner robot loop in Line 5 of Algorithm 1 once per team because we use homogeneous robots and policies. Our MCTS variant uses the following hyperparameters: $L_{\text{expert}} = 10\,000$, $L_{\text{learner}} = 500$, $C_p = 2.0$, $C_{pw} = 1.0$, $\alpha_{pw} = 0.25$ and $\alpha_d = (1 - 3/(100 - 10d))/20$ where d is the depth of the node. The neural frequency hyperparameters are $\beta_\pi = \beta_V = 0.5$. The double-integrator game parameters are chosen to match the hardware used in the physical experiments (see Sec. IV-E). We use position bounds $\overline{p} = 1, 2, 3$ m and constant velocity $\overline{v} = 1.0\,\mathrm{m/s}$ and acceleration $\overline{a} = 2.0\,\mathrm{m/s^2}$ bounds. The tag, collision, and sensing radii are: $r_t = 0.2\,\mathrm{m}$, $r_p = 0.1\,\mathrm{m}$, $r_{\text{sense}} = 2.0\,\mathrm{m}$. We train for up to 5 agents on each team. We use a simulation and planning timestep of $\Delta = 0.1\,\mathrm{s}$. Each team starts at opposite sides of the environment, and the goal is placed closer to the defenders' starting position.

We implement the machine learning components in PyTorch [27]. The datasets are of size $80\,000$ points per iteration for both value and policy datasets. The meta learning algorithm is trained until convergence. The policy and value network models both use DeepSet [25] neural network architecture, (e.g. [24]) where the inner and outer networks each have one hidden layer with 16 neurons and appropriate input and output dimensions. All networks are of feedforward structure with ReLU activation functions, batch size of 1028, and are trained over 300 epochs.

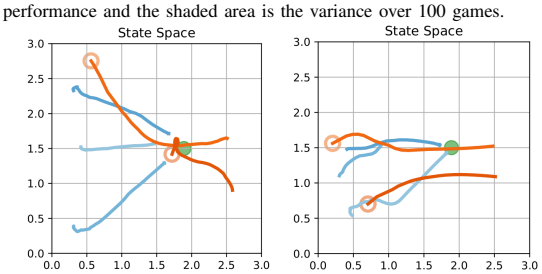
B. Variants and Baseline

In order to evaluate our method, we test the multiple learners and expert policies, each equipped with networks after k learning iterations. To isolate the effect of the neural expansion, we consider the $k = 0$ case for both learner and expert as an unbiased MCTS baseline solution.

As an additional baseline for the double-integrator game, we use the solution from [2]. Their work adapts the exact



(a) Double-integrator game evaluation: the thick lines indicate the average performance and the shaded area is the variance over 100 games.



(b) Learner defense (orange) finds emergent cooperative strategies to defend the goal (green).

(c) Baseline defense (orange) is vulnerable to learner's offensive (blue) dodge maneuver.

Fig. 3. Double-integrator performance and strategy examples.

differential game solution for simple-motion and single-robot teams proposed in [15] to a double-integrator, multi-robot team setting. However, their adapted solution is not exact because it assumes a constant acceleration magnitude input and relies on composition of pair-wise matching strategies.

C. Simulation Results

We evaluate our expert and learner by initializing 100 different initial conditions of a 3 attacker, 2 defender game in a 3 m space. Then, we rollout every combination of variants, learning iterations, and baseline for both team A and team B policies, for a total of 12 100 games. For a single game, the performance criteria for team A policies is the terminal reward and, in order to have consistency of plots (higher is better), the performance criteria for team B policies is one minus the terminal reward. An example game with a different number of agents and environment size is shown in Fig. 2c and its animation is provided in the supplemental video. The 10 vs. 10 game illustrates the natural scalability in number of agents of the decentralized approach and the generalizability of the neural networks, as they were only trained with data containing up to 5 robots per team.

The statistical results of the 3 vs. 2 experiment are shown in Fig. 3a where the thick lines denote the average performance value and the shade is the performance variance. We find the expected results; for both team A and team B, the learner with no bias has the worst performance, and learner with fully trained networks surpasses the centralized and expensive unbiased expert and approaches the biased expert. The baseline attacker is about the same strength as the unbiased expert, whereas the baseline defender is much stronger than the unbiased expert. In both cases, the fully-trained biased expert and learner are able to significantly outperform the baseline.

To investigate the qualitative advantages of our method, we looked at the games where our learner defense outperformed the baseline defense and found two principal advantages: first, the learner defense sometimes demonstrated emergent coordination that is more effective than a pairwise matching strategy, e.g. one defender goes quickly to the goal to protect against greedy attacks while the other defender slowly approaches the goal to maintain its maneuverability, see Fig. 3b. Second, the learner attacker is sometimes able to exploit the momentum of the baseline defender and perform a dodge maneuver, e.g. the bottom left interaction in Fig. 3c, whereas the learner defense is robust to this behavior. These examples show the learner networks can generate sophisticated, effective maneuvers.

As an additional experiment, we evaluate the learner (without retraining) in an environment with static and dynamic obstacles for 100 different initial conditions; an example is shown in the supplementary video. In this environment, the fully trained learner outperforms the unbiased learner 0.246 ± 0.022, this value is calculated by summing the performance criteria difference across attacking and defending policies. This result demonstrates the natural compatibility of tree-based planners with safety constraints and the robustness of the performance gain in out-of-training-domain scenarios.

D. Dynamics Extension

As shown in Fig. 2, NTE can be applied to arbitrary game settings and dynamics. We evaluate the same Reach-Target-Avoid game with 3D Dubin's vehicle dynamics as a relevant model for fixed-wing aircraft applications, shown in Fig. 2d. We consider the state, action, and dynamics: $\mathbf{s}_t = [x_t, y_t, z_t, \psi_t, \gamma_t, \phi_t, v_t]^T$, $\mathbf{a}_t = [\dot{\gamma}_t, \dot{\phi}_t, \dot{v}_t]^T$

$$\mathbf{s}_{t+1} = f(\mathbf{s}_t, \mathbf{a}_t) = \mathbf{s}_t + \begin{bmatrix} v_t \cos(\gamma_t) \sin(\psi_t) \\ v_t \cos(\gamma_t) \cos(\psi_t) \\ -v_t \sin(\gamma_t) \\ \frac{g}{v_t} \tan(\phi_t) \\ \mathbf{a}_t \end{bmatrix} \Delta_t \tag{18}$$

where x, y, z are inertial position, v is speed, ψ is the heading angle, γ is the flight path angle, and ϕ is the bank angle and g is the gravitational acceleration. The game is bounded to $\overline{p} = 5$ m with a maximum linear acceleration of $2.0\,\mathrm{m/s^2}$ and maximum angular rates of $36\,\mathrm{deg/s}$, and g is set to $0.98\,\mathrm{m/s^2}$ to scale to our game length scale.

We initialize 2 attacker, 2 defender games for 100 different initial conditions in a 5 m region and test the policy variants, without an external baseline, for a total of 81 000 games. The performance results are shown in Fig. 4, where we see the same trend that the learner and expert policies improve over learning iterations. In addition, the biased learner's performance quickly surpasses the unbiased expert ($k = 0$).

E. Hardware Validation

To test our algorithm in practice, we fly in a motion capture space, where each robot (CrazyFlie 2.x, see Fig. 1) is equipped with a single marker, and we use the Crazyswarm [28] for tracking and scripting. The centralized system simulates distributed operation by collecting the full state, computing local observations and local policies, and broadcasting only

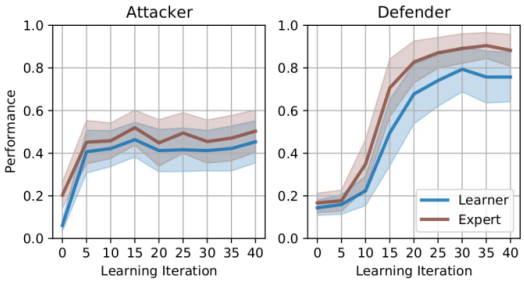


Fig. 4. 3D Dubin's vehicle game evaluation: the thick lines indicate the average performance and the shaded area is the variance over 100 games.

the output of each robot's learner policy. For a given double-integrator policy, we evaluate the learner to compute an action, forward-propagate double-integrator dynamics, and track the resulting position and velocity set-point using a nonlinear controller for full quadrotor dynamics. Planning in a lower-dimensional double-integrator state and then tracking the full system is enabled by the timescale separation of position and attitude dynamics of quadrotors.

We evaluate the double-integrator learner for up to 3 attacker, 2 defender games in an aerial swarm flight demonstration. We show the results of the experiments in our supplemental video. We use the same parameters as in simulation in Sec. IV-C. Our learner evaluation takes an average of $11\,\mathrm{ms}$ with a standard deviation of $6\,\mathrm{ms}$, with each robot policy process running in parallel on an Intel(R) Core(TM) i7-8665U. By comparison, the biased expert takes $329\pm144\,\mathrm{ms}$ to execute and the unbiased expert takes $277\pm260\,\mathrm{ms}$. Our computational tests show that the learner has a significant ($\approx$ 25 times) computational advantage over the baseline unbiased expert. Our physical demonstration shows that our learner is robust to the gap between simulation and real world and can run in real-time on off-the-shelf hardware.

V. Conclusion

We present a new approach for multi-robot planning in non-cooperative environments with an iterative search and learning method called Neural Tree Expansion. Our method bridges the gap between an AlphaZero-like method and real-world robotics applications by introducing a learner agent with decentralized evaluation, partial information, and limited computational resources. Our method outperforms the current state-of-the-art analytical baseline for the multi-robot double-integrator Reach-Target-Avoid game with dynamically sophisticated and coordinated strategies. We demonstrate our method's broad compatibility by further empirical evaluation of the Reach-Target-Avoid game for 3D Dubin's vehicle dynamics and visualization of canonical decision-making problems. We validate the effectiveness through hardware experimentation and show that our policies run in real-time on off-the-shelf computational resources. In future work, we will combine planning under uncertainty algorithms with deep learning to handle scenarios with model and measurement uncertainty.

References

[1] D. Silver *et al.*, "A general reinforcement learning algorithm that masters chess, shogi, and go through self-play", *Science*, vol. 362, no. 6419, pp. 1140–1144, 2018.
[2] M. Coon and D. Panagou, "Control strategies for multiplayer target-attacker-defender differential games with double integrator dynamics", in *IEEE 56th Annual Conf. on Decis. and Control*, 2017.
[3] M. J. Kochenderfer *et al.*, "Decision Making Under Uncertainty: Theory and Application", 1st Ed. The MIT Press, 2015.
[4] C. Browne *et al.*, "A survey of Monte Carlo tree search methods", *IEEE Trans. Comput. Intell. AI Games*, vol. 4, no. 1, pp. 1–43, 2012.
[5] L. Kocsis and C. Szepesvári, "Bandit based Monte-Carlo planning", in *Eur. Conf. Mach. Learn.*, vol. 4212, Springer, 2006.
[6] D. Shah, Q. Xie, and Z. Xu, "Non-asymptotic analysis of Monte Carlo tree search", in *SIGMETRICS (Abstracts)*, ACM, 2020, pp. 31–32.
[7] W. Mao, K. Zhang, Q. Xie, and T. Basar, "POLY-HOOT: Monte-Carlo planning in continuous space MDPs with non-asymptotic analysis", in *Neural Inf. Process. Syst.*, 2020.
[8] C. R. Mansley, A. Weinstein, and M. L. Littman, "Sample-based planning for continuous action Markov decision processes, and scheduling", in *Int. Conf. on Autom. Planning and Scheduling*, 2011.
[9] D. Auger, A. Couëtoux, and O. Teytaud, "Continuous upper confidence trees with polynomial exploration – Consistency", in *Eur. Conf. Mach. Learn.*, vol. 8188, Springer, 2013.
[10] T. M. Moerland, J. Broekens, A. Plaat, and C. M. Jonker, "A0C: Alpha zero in continuous action space", in *Eur. Workshop on Reinforcement Learn. 14*, 2018.
[11] B. Chen *et al.*, "Learning to plan in high dimensions via neural exploration-exploitation trees", in *Int. Conf. on Learn. Repres.*, 2020.
[12] S. Ross, G. J. Gordon, and D. Bagnell, "A reduction of imitation learning and structured prediction to no-regret online learning", in *Proc. Int. Conf. Artif. Intell. and Statist.*, 2011.
[13] R. S. Sutton, D. A. McAllester, S. P. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation", in *Neural Inf. Process. Syst.*, 1999.
[14] M. Prajapat, K. Azizzadenesheli, A. Liniger, Y. Yue, and A. Anandkumar, "Competitive policy optimization", CoRR abs/2006.10611 2020.
[15] R. Isaacs, "Differential games; a mathematical theory with applications to warfare and pursuit, control and optimization". Wiley, 1965.
[16] E. Garcia, D. W. Casbeer, and M. Pachter, "Optimal strategies for a class of multi-player reach-avoid differential games in 3d space", *IEEE Robot. Autom. Lett.*, vol. 5, no. 3, pp. 4257–4264, 2020.
[17] R. Yan, X. Duan, Z. Shi, Y. Zhong, and F. Bullo, "Matching-based capture strategies for 3D heterogeneous multiplayer reach-avoid differential games", CoRR 1909.11881 2019.
[18] A. A. Paranjape, S.-J. Chung, K. Kim, and D. H. Shim, "Robotic herding of a flock of birds using an unmanned aerial vehicle", *IEEE Trans. Robot.*, vol. 34, no. 4, pp. 901–915, 2018.
[19] J. Hu, A. E. Turgut, T. Krajník, B. Lennox, and F. Arvin, "Occlusion-based coordination protocol design for autonomous robotic shepherding tasks", *IEEE Trans. on Cogn. and Develop. Syst.*, pp. 1–1, 2020.
[20] S. Nardi, F. Mazzitelli, and L. Pallottino, "A game theoretic robotic team coordination protocol for intruder herding", *IEEE Robot. Autom. Lett.*, vol. 3, no. 4, pp. 4124–4131, 2018.
[21] D. Shishika, J. Paulos, and V. Kumar, "Cooperative team strategies for multi-player perimeter-defense games", *IEEE Robot. Autom. Lett.*, vol. 5, no. 2, pp. 2738–2745, 2020.
[22] Z. N. Sunberg and M. J. Kochenderfer, "Online algorithms for POMDPs with continuous state, action, and observation spaces", in *Int. Conf. on Autom. Planning and Scheduling*, 2018.
[23] C. Florensa, D. Held, M. Wulfmeier, M. Zhang, and P. Abbeel, "Reverse curriculum generation for reinforcement learning", in *Conf. on Robot Learn.*, vol. 78, 2017.
[24] B. Rivière, W. Hönig, Y. Yue, and S. Chung, "GLAS: Global-to-local safe autonomy synthesis for multi-robot motion planning with end-to-end learning", *IEEE Robot. Autom. Lett.*, vol. 5, no. 3, pp. 4249–4256, 2020.
[25] M. Zaheer *et al.*, "Deep sets", in *Neural Inf. Process. Syst.*, 2017.
[26] B. Ichter, J. Harrison, and M. Pavone, "Learning sampling distributions for robot motion planning", in *Proc. IEEE Int. Conf. on Robot. and Automat.*, 2018.
[27] A. Paszke *et al.*, "Pytorch: An imperative style, high-performance deep learning library", in *Neural Inf. Process. Syst.* 2019.
[28] J. A. Preiss, W. Hönig, G. S. Sukhatme, and N. Ayanian, "Crazyswarm: A large nano-quadcopter swarm", in *Proc. IEEE Int. Conf. Robot. Autom.*, 2017.