2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 25-29, 2020, Las Vegas, NV, USA (Virtual)

2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) | 978-1-7281-6212-6/20/$31.00 ©2020 IEEE | DOI: 10.1109/IROS45743.2020.9341668

# Graph Neural Networks for Decentralized Multi-Robot Path Planning

Qingbiao Li[1], Fernando Gama[2], Alejandro Ribeiro[2], Amanda Prorok[1]

***Abstract*— Effective communication is key to successful, decentralized, multi-robot path planning. Yet, it is far from obvious what information is crucial to the task at hand, and how and when it must be shared among robots. To side-step these issues and move beyond hand-crafted heuristics, we propose a combined model that automatically synthesizes local communication and decision-making policies for robots navigating in constrained workspaces. Our architecture is composed of a convolutional neural network (CNN) that extracts adequate features from local observations, and a graph neural network (GNN) that communicates these features among robots. We train the model to imitate an expert algorithm, and use the resulting model online in decentralized planning involving only local communication and local observations. We evaluate our method in simulations by navigating teams of robots to their destinations in 2D cluttered workspaces. We measure the success rates and sum of costs over the planned paths. The results show a performance close to that of our expert algorithm, demonstrating the validity of our approach. In particular, we show our model's capability to generalize to previously unseen cases (involving larger environments and larger robot teams).**

## I. Introduction

Efficient and collision-free navigation in multi-robot systems is fundamental to advancing mobility. The problem, generally referred to as Multi-Robot Path Planning (MRPP) or Multi-Agent Path Finding (MAPF), aims at generating collision-free paths leading robots from their origins to designated destinations. Current approaches can be classified as either *coupled* or *decoupled*, depending on the structure of the state space that is searched. While coupled approaches are able to ensure the optimality and completeness of the solution, they involve *centralized* components, and tend to scale poorly with the number of robots [1], [2]. Decoupled approaches, on the other hand, compute trajectories for each robot separately, and re-plan only in case of conflicts [3], [4], [5]. This can significantly reduce the computational complexity of the planning task, but generally produces sub-optimal and incomplete solutions. Balancing optimality and completeness with the complexity of computing a solution, however, is still an open research problem [6], [7].

This work focuses on multi-robot path planning for scenarios where the robots are restricted in observation and communication range, and possess no global reference frame for localization. This naturally arises when considering physical robots equipped with hardware constraints that limit their perception and communication capabilities [8]. These scenarios impose a decentralized structure, where at any given point in time, robots have only partial information of the system state. In this paper, we propose a combined architecture, where we train a convolutional neural network (CNN) [9] that extracts adequate features from local observations, and a graph neural network (GNN) to communicate these features among robots [10] with the ultimate goal of learning a decentralized sequential action policy that yields efficient path plans for all robots. The GNN implementation seamlessly adapts to the partial information structure of the problem, since it is computed in a decentralized manner. We train this architecture to imitate an optimal coupled planner with global information that is available offline at training time. Further, we develop a dataset aggregation method that leverages an online expert to resolve hard cases, thus expediting the learning process. The resulting trained model is used online in an efficient, decentralized manner, involving communication only with nearby robots. Furthermore, we show that the model can be deployed on much larger robot teams than the ones it was trained on.

[1]Qingbiao Li and Amanda Prorok are with the Department of Computer Science and Technology, University of Cambridge, Cambridge, United Kingdom {ql295,asp45}@cam.ac.uk

[2]Fernando Gama and Alejandro Ribeiro are with the Department of Electrical and Systems Engineering, University of Pennsylvania, Philadelphia, USA {fgama,aribeiro}@seas.upenn.edu

## II. Related work and Contribution

**Related work.** Classical approaches to multi-robot path planning can generally be described as either centralized or decentralized. *Centralized* approaches are facilitated by a planning unit that monitors all robots' positions and desired destinations, and returns a coordinated plan of trajectories (or way-points) for all the robots in the system. These plans are communicated to the respective robots, which use them for real-time on-board control of their navigation behavior. Coupled centralized approaches, which consider the joint configuration space of all involved robots, have the advantage of producing optimal and complete plans, yet tend to be computationally very expensive. Indeed, solving for optimality is NP-hard [11], and although significant progress has been made towards alleviating the computational load [12], [13], these approaches still scale poorly in environments with a high number of potential path conflicts.

*Decentralized* approaches provide an attractive alternative to centralized approaches, firstly, because they reduce the computational overhead, and secondly, because they relax the dependence on centralized units. This body of work considers the generation of collision-free paths for individual robots that cooperate only with immediate neighbors [14], [5], or with no other robots at all [3]. In the latter case, coordination is reduced to the problem of reciprocally avoiding other robots (and obstacles), and can generally be solved without the use of communication. Yet, by taking purely local objectives into account, global objectives (such as path efficiency) cannot be explicitly optimized. In the former case, it has been shown that monotonic cost reduction of global objectives can be achieved. This feat, however, relies on

strong assumptions (e.g., problem convexity and invariance of communication graph [15], [16]) that can generally not be guaranteed in real robot systems.

*Learning-based* methods have proven effective at designing robot control policies for an increasing number of tasks [17], [18]. The application of learning-based methods to multi-robot motion planning has attracted particular attention due to their capability of handling high-dimensional joint state-space representations, by offloading the online computational burden to an offline learning procedure. The work in [19] proposes a decentralized multi-agent collision avoidance algorithm based on deep reinforcement learning. Their results show that significant improvement in the quality of the path (i.e., time to reach the goal) can be achieved with respect to current benchmark algorithms (e.g., ORCA [3]). Also in recent work, Sartoretti et al. [7] propose a hybrid learning-based method called PRIMAL for multi-agent path-finding that uses both imitation learning (based on an expert algorithm) and multi-agent reinforcement learning. It is note-worthy that none of the aforementioned learning-based approaches consider inter-robot communication, and thus, do not exploit the scalability benefits of fully decentralized approaches. Learning what, how, and when to communicate is key to this aim.

Of particular interest to us is the capability of learning-based methods to handle high-dimensional joint state-space representations, useful when planning for large-scale collective robotic systems, by offloading the online computational burden to an offline learning procedure [20], [21], [19]. The fact that each robot must be able to accumulate information from other robots in its neighborhood is key to this learning procedure. From the point of view of an individual robot, its local decision-making system is incomplete, since other agents' unobservable states affect future values. *The manner in which information is shared is crucial to the system's performance, yet is not well addressed by current machine learning approaches.* Graph Neural Networks (GNNs) promise to overcome this deficiency [22], [10]. They capture the *relational* aspect of robot communication and coordination by modeling the collective robot system as a graph: each robot is a node, and edges represent communication links [23]. Although GNNs have been applied to a number of problem domains, including molecular biology [24], quantum chemistry [25], and simulation engines [26], they have only very recently been considered within the multi-robot domain, for applications of flocking and formation control [23], [20], [21].

**Contributions.** The application of GNNs to the problem of multi-robot path planning is novel. Our particular GNN implementation offers an efficient architecture that operates in a localized manner, whereby information is shared over a multi-hop communication network, through explicit communication with nearby neighbors only [10]. Our key contribution in this work is the development of a framework that can learn what information needs to be shared between robots, such that each robot can make a decision based on this local information; importantly, our robots only have local, relative information about their positions and goals (i.e., they possess no global reference frame). This framework is composed of a convolutional neural network (CNN) that extracts adequate features from local observations, and a graph neural network (GNN) that learns to explicitly communicate these features among robots. By jointly training these two components, the system is able to best determine what information is relevant for the team as a whole, and share this to facilitate efficient path planning. The proposed model is trained to imitate a centralized coupled planner, and makes use of a novel dataset aggregation method that leverages an online expert to resolve hard cases, thus expediting the learning process. We achieve performance that is close to that of optimal planners in terms of success rate and flowtime (sum of path costs), while also being able to generalize to previously unseen cases, such as larger robot teams and environments.

## III. Problem formulation

Let $\mathcal{V} = \{v_1, \dots, v_N\}$ be the set of $N$ robots. At time $t$, each robot perceives its surroundings within a given field of vision; although the robot knows where its own target destination is located, this information is clipped to the field of vision in a local reference frame (see Fig. 1). Furthermore, we assume no global positioning of the robots. This map perceived by robot $i$ is denoted by $\mathbf{Z}_t^i \in \mathbb{R}^{W_{\text{FOV}} \times H_{\text{FOV}}}$ where $W_{\text{FOV}}$ and $H_{\text{FOV}}$ are the width and height, respectively, and are determined by the field of vision radius $r_{\text{FOV}}$.

The robots can communicate with each other as determined by the communication network. We can describe this network at time $t$ by means of a graph $\mathcal{G}_t = (\mathcal{V}, \mathcal{E}_t, \mathcal{W}_t)$ where $\mathcal{V}$ is the set of robots, $\mathcal{E}_t \subseteq \mathcal{V} \times \mathcal{V}$ is the set of edges and $\mathcal{W}_t : \mathcal{E}_t \to \mathbb{R}$ is a function that assigns weights to the edges. Robots $v_i$ and $v_j$ can communicate with each other at time $t$ if $(v_i, v_j) \in \mathcal{E}_t$. The corresponding edge weight $\mathcal{W}_t(v_i, v_j) = w_t^{ij}$ can represent the strength of the communication (or be equal to 1 if we are only modeling whether there is a link or not). For instance, two robots $v_i$ and $v_j$, with positions $\mathbf{p}_i, \mathbf{p}_j \in \mathbb{R}^2$ respectively, can communicate with each other if $\|\mathbf{p}_i - \mathbf{p}_j\| \leq r_{\text{COMM}}$ for a given communication radius $r_{\text{COMM}} > 0$. This allows us to define an adjacency matrix $\mathbf{S}_t \in \mathbb{R}^{N \times N}$ representing the communication graph, where $[\mathbf{S}_t]_{ij} = s_t^{ij} = 0$ if $(v_j, v_i) \notin \mathcal{E}_t$.

In this work, we formulate the multi-agent path planning problem as a sequential decision-making problem that each robot solves at every time instant $t$, with the objective of reaching its destination. More formally, the goal of this work is to learn a mapping $\mathcal{F}$ that takes the maps $\{\mathbf{Z}_t^i\}_{v_i \in \mathcal{V}}$ and the communication network $\mathcal{G}_t$ at time $t$ and determines an appropriate action $\mathbf{u}_t$. We want the action $\mathbf{u}_t = \mathcal{F}(\{\mathbf{Z}_t^i\}, \mathcal{G}_t)$ to be such that it contributes to the global objective of moving the robots towards their destinations in the shortest possible time while avoiding collisions with other robots and with obstacles that might be present. The objective is to train the network to perform as well as a coupled centralized expert, while restricting robots to partial observations. The mapping $\mathcal{F}$ has to be restricted to involve communication only among nearby robots, as dictated by the network $\mathcal{G}_t$ at each time instant $t$.

## IV. Graph Neural Networks

In order to guarantee that the mapping $\mathcal{F}$ is restricted to communications only among nearby robots, we parametrize it by means of a GNN, which is a naturally decentralized solution (Sec. IV-B). We then train this GNN to learn appropriate actions that contribute to the global objective by means of supervised learning through an expert algorithm (i.e., imitation learning) (Sec. V-E).

### A. Graph Convolutions

Assume that each robot has access to $F$ observations $\tilde{\mathbf{x}}_t^i \in \mathbb{R}^F$ at time $t$. Let $\mathbf{X}_t \in \mathbb{R}^{N \times F}$ be the observation matrix where each row collects these $F$ observations at each robot $\tilde{\mathbf{x}}_t^i$, $i = 1, \ldots, N$,

$$\mathbf{X}_t = \begin{bmatrix} (\tilde{\mathbf{x}}_t^1)^\mathsf{T} \\ \vdots \\ (\tilde{\mathbf{x}}_t^N)^\mathsf{T} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_t^1 & \cdots & \mathbf{x}_t^F \end{bmatrix}. \tag{1}$$

Note that the columns $\mathbf{x}_t^f \in \mathbb{R}^N$ represent the collection of the observation $f$ across all nodes, for $f = 1, \ldots, F$. This vector $\mathbf{x}_t^f$ is a *graph signal* [27], since it assigns a scalar value to each node, $\mathbf{x}_t^f : \mathcal{V} \to \mathbb{R}$ so that $[\mathbf{x}_t^f]_i = x_t^{if} \in \mathbb{R}$.

To formally describe the communication between neighboring agents, we need a concise way of describing the graph $\mathcal{G}_t$ and relating it to the observations $\mathbf{X}_t$. Towards this end, we use the adjacency matrix $\mathbf{S}_t$. We note that other matrix descriptions of the graph, such as the Laplacian matrix or the Markov matrix are possible. We generically call $\mathbf{S}_t$ the *graph shift operator* (GSO) [27].

The operation $\mathbf{S}_t\mathbf{X}_t$ represents a linear combination of neighboring values of the signal due to the sparsity pattern of $\mathbf{S}_t$. More precisely, note that the value at node $i$ for observation $f$ after operation $\mathbf{S}_t\mathbf{X}_t \in \mathbb{R}^{N \times F}$ becomes

$$[\mathbf{S}_t\mathbf{X}_t]_{if} = \sum_{j=1}^{N} [\mathbf{S}_t]_{ij}[\mathbf{X}_t]_{jf} = \sum_{j:v_j \in \mathcal{N}_i} s_t^{ij} x_t^{jf} \tag{2}$$

where $\mathcal{N}_i = \{v_j \in \mathcal{V} : (v_j, v_i) \in \mathcal{E}_t\}$ is the set of nodes $v_j$ that are neighbors of $v_i$. Also, the second equality in (2) holds because $s_t^{ij} = 0$ for all $j \notin \mathcal{N}_i$.

The linear operation $\mathbf{S}_t\mathbf{X}_t$ is essentially *shifting* the values of $\mathbf{X}_t$ through the nodes, since the application of $\mathbf{S}_t$ updates the value at each node by a linear combination of values in the neighborhood. With the shifting operation in place, we can define a *graph convolution* [10] as linear combination of shifted versions of the signal

$$\mathcal{A}(\mathbf{X}_t; \mathbf{S}_t) = \sum_{k=0}^{K-1} \mathbf{S}_t^k \mathbf{X}_t \mathbf{A}_k \tag{3}$$

where $\{\mathbf{A}_k\}$ is a set of $F \times G$ matrices representing the filter coefficients combining different observations. Several noteworthy comments are in order with respect to (3). First, multiplications to the left of $\mathbf{X}_t$ need to respect the sparsity of the graph since these multiplications imply combinations across different nodes. Multiplications to the right, on the other hand, can be arbitrary, since they imply linear combination of observations within the same node in a weight sharing scheme. Second, $\mathbf{S}_t^k\mathbf{X}_t = \mathbf{S}_t(\mathbf{S}_t^{k-1}\mathbf{X}_t)$ is computed by means of $k$ communication exchanges with 1-hop neighbors, and is actually computing a summary of the information located at the $k$-hop neighborhood. Therefore, the graph convolution is an entirely local operation in the sense that its implementation is naturally distributed. Third, the graph convolution is actually computing the output of a bank of $FG$ filters where we take as input $F$ observations per node and combine them to output $G$ observations per node, $\mathcal{A}(\mathbf{X}_t; \mathbf{S}_t) \in \mathbb{R}^{N \times G}$. There are $FG$ graph filters involved in (3) each one consisting of $K$ filter taps, i.e., the $(f, g)$ filter can be described by filter taps $\mathbf{a}^{fg} = [a_0^{fg}, \ldots, a_{K-1}^{fg}] \in \mathbb{R}^K$ and these filter taps are collected in the matrix $\mathbf{A}_k$ as $[\mathbf{A}_k]_{fg} = a_k^{fg}$.

### B. Graph Neural Networks

A convolutional GNN [10] consists of a cascade of $L$ layers, each of which applies a graph convolution (3) followed by a pointwise nonlinearity $\sigma : \mathbb{R} \to \mathbb{R}$ (also known as activation function)

$$\mathbf{X}_\ell = \sigma\big[\mathcal{A}_\ell(\mathbf{X}_{\ell-1}; \mathbf{S})\big] \quad \text{for} \quad \ell = 1, \ldots, L \tag{4}$$

where, in a slight abuse of notation, $\sigma$ is applied to each element of the matrix $\mathcal{A}_\ell(\mathbf{X}_{\ell-1}; \mathbf{S})$. The input to each layer is a graph signal consisting of $F_{\ell-1}$ observations and the output has $F_\ell$ observations so that $\mathbf{X}_\ell \in \mathbb{R}^{N \times F_\ell}$. The input to the first layer is $\mathbf{X}_0 = \mathbf{X}_t$ so that $F_0 = F$ and the output of the last layer corresponds to the action to be taken at time $t$, $\mathbf{X}_L = \mathbf{U}_t$ which could be described by a vector of dimension $F_L = G$. The GSO $\mathbf{S}$ to be used in (4) is the one corresponding to the communication network at time $t$, $\mathbf{S} = \mathbf{S}_t$. At each layer $\ell$ we have a bank of $F_\ell F_{\ell-1}$ filters $\mathcal{A}_\ell$ described by a set of $K_\ell F_\ell F_{\ell-1}$ total filter taps $\{\mathbf{A}_{\ell k}\}_{k=0}^{K-1}$.

We note that, in the present framework, we are running one GNN (4) per time instant $t$, where each time step is determined by the moment the action is taken and the communication network changes. This implies that we need to carry out $\sum_{\ell=1}^{L}(K_\ell - 1)$ total communications before deciding on an action. Therefore, it is important to keep the GNN shallow (small $L$) and the filters short (small $K_\ell$).

In summary, we propose to parametrize the mapping $\mathcal{F}$ between maps $\mathbf{Z}_t$ and actions $\mathbf{U}_t$ by using a GNN (4) acting on observations $\mathbf{X}_t = \text{CNN}(\mathbf{Z}_t)$ obtained by applying a CNN to the input maps. We note that, by choosing this parametrization we are obtaining a mapping that is naturally distributed and that is adequately exploiting the network structure of the data.

## V. Architecture

The following sections describe the architecture, of which all components are illustrated in Fig. 1.

### A. Processing Observations

In an environment ($W \times H$) with static obstacles, each robot has a local field-of-view (FOV), the radius of which is defined by $r_{\text{FOV}}$, beyond which it cannot 'see' anything. The robot itself is located at the center of this local observation, and does not know its global position. The data available at robot $i$ is a map $\mathbf{Z}_t^i$ of size $W_{\text{FOV}} \times H_{\text{FOV}}$ (Fig. 1

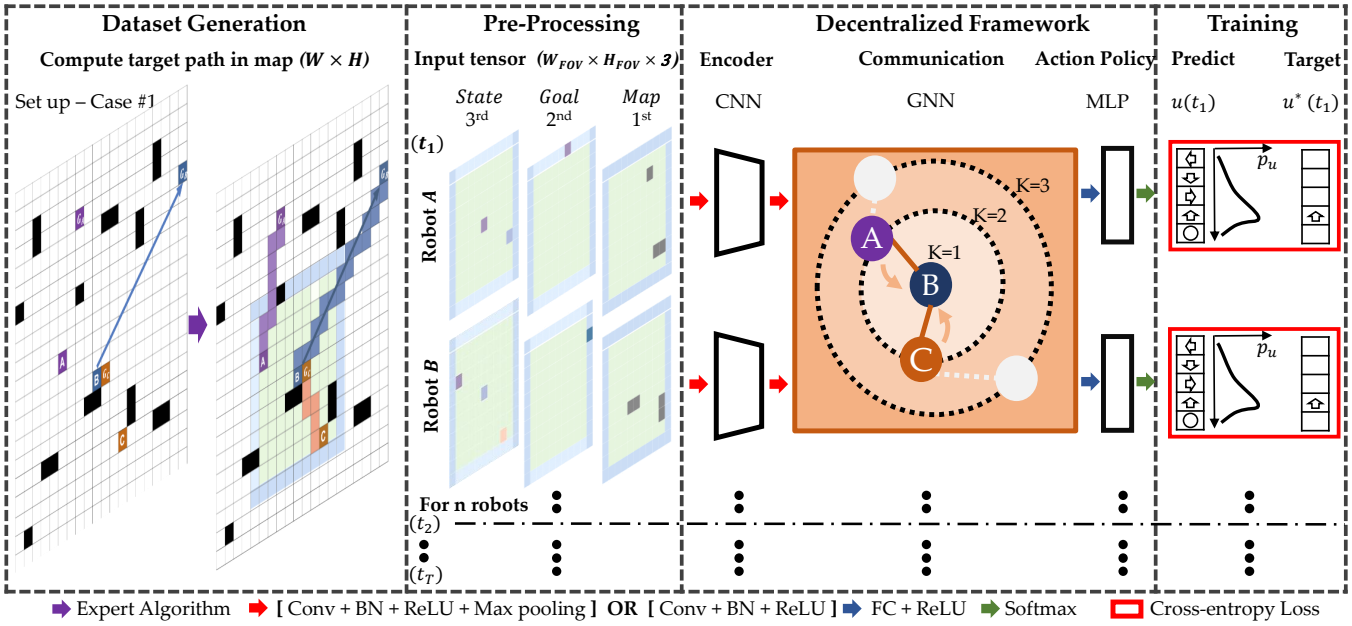



Fig. 1: Illustration of the proposed framework. *(i)* The input tensor is based on a binary map representation (1st channel: partial observation of the environment; 2nd channel: the position of goal ($\mathbf{p}^i_{goal}$), or its projection onto the boundary of the field-of-view; 3rd channel: self (agent) at center, with other agents within its field-of-view). *(ii)* The decentralized framework consists of a CNN to extract observations from the input tensor, a GNN to exchange information between the neighboring agents, and an MLP to predict the actions. *(iii)* Training is performed through cross-entropy loss over a discrete action space.

illustrates how we implement such partial observations). The input map $\mathbf{Z}^i_t$ is fed into a CNN that is run internally on each robot. This results in a vector $\tilde{\mathbf{x}}^i_t \in \mathbb{R}^F$ containing $F$ observations (1), $\tilde{\mathbf{x}}^i_t = \text{CNN}(\mathbf{Z}^i_t)$. These observations can then be communicated to nearby robots. The intuition behind using a CNN is to process the input map $\mathbf{Z}^i_t$ into a higher-level feature tensor $\tilde{\mathbf{x}}^i_t$ describing the observation, goal and states of other robots. This feature tensor is then transmitted via the communication network, as described in the following section,Sec. V-B.

### B. Communication

Each individual robot communicates its compressed observation vector $\tilde{\mathbf{x}}^i_t$ with neighboring robots within its communication radius $r_{\text{COMM}}$ over the multi-hop communication network, whereby the number of executed hops is limited by $K$. As described in Sec. IV-B, we apply our GNN to aggregate and fuse the states $(\tilde{\mathbf{x}}^j_t)$ within this $K$-hop neighborhood of robots $j \in \mathcal{N}_i$, for each robot $i$. The output of the communication GNN is a hyper-representation of the fused information of the robot itself and its $K$-hop neighbors, which is passed to the action policy, as described in Sec. V-C. We note that each robot carries a local copy of the GNN, hence resulting in a localized decision-making policy.

### C. Action Policy

We formulate the path-finding problem as a sequential classification problem, whereby an optimal action is chosen at each time step. We adopt a local multi-layer perceptron (MLP) to train our action policy network. More specifically, each node applies a MLP to the aggregated features resulting from the communication GNN. This MLP is the same across all nodes, resembling a weight-sharing scheme. The action $\tilde{\mathbf{u}}^i_t$ taken by robot $i$ is given by a stochastic action policy based on the probability distribution over motion primitives, which in our case consists of five discrete options (up, left, down, right, idle), and are represented by one-hot vectors. The final path is represented by the series of sequential actions.

### D. Network Architecture

We construct our CNN architecture by using `Conv2d-BatchNorm2d-ReLU-MaxPool2d` and `Conv2d-BatchNorm2d-ReLU` blocks sequentially three times. All kernels are of size 3 with a stride of 1 and zero-padding. In the GNN architecture, we deploy a single layer GNN (as described in Sec. IV-B) and set 128 as the number of input observations $F$ and output observations $G$. Note that we can tune the filter taps $K$ for non-communication ($K = 1$) and multi-hop communication ($K > 1$). In the action policy, we use a linear soft-max layer to decode the output observations $G$ from the GNN with 128 features into the five motion primitives.

### E. Learning from Expert Data

To train our models, we propose a supervised learning approach based on expert data (i.e., imitation learning). We assume that, at training time, we have access to an optimal trajectory of actions $\{\mathbf{U}^*_t\}$ for all the robots, and the corresponding maps obtained for this trajectory $\{\mathbf{Z}^i_t\}$, collected in a training set $\mathcal{T} = \{(\{\mathbf{U}_t\}, \{\mathbf{Z}^i_t\})\}$. Then, we train the mapping $\mathcal{F}$ so that the output is as close as possible to the corresponding optimal action $\mathbf{U}^*$ using a cross entropy loss $\mathcal{L}(\cdot,\cdot)$. If the mapping $\mathcal{F}$ is parametrized in terms of a GNN (4) then this optimization problem becomes

$$\min_{\text{CNN},\{\mathbf{A}_{\ell k}\},\text{MLP}} \sum_{(\{\mathbf{U}_t\},\{\mathbf{Z}^i_t\})\in\mathcal{T}} \sum_t \mathcal{L}(\mathbf{U}^*_t, \mathcal{F}(\{\mathbf{Z}^i_t\}, \mathcal{G}_t)). \quad (5)$$

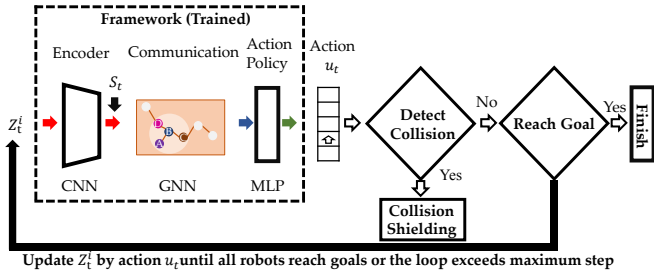



Fig. 2: Illustration of the inference stage: for each robot, the input map $\mathbf{Z}_t^i$ is fed to the trained framework to predict the action; collisions are detected and prevented by collision shielding. The input map $\mathbf{Z}_t^i$ is continuously updated until the robot reaches its goal or exceeds the timeout $T_{max}$.

We are optimizing over the filters in the CNN required to process the map as well as the set of matrices $\{\mathbf{A}_{\ell k}\}$ that contains the $\sum_{\ell=1}^{L} K_\ell F_{\ell-1} F_\ell$ learnable parameters of the communication GNN. Note that the number of parameters is independent of the size of the network $N$.

Imitation learning rests on the availability of an optimal solution (further elaborated in Sec. V-F, below). While this solution might be computationally expensive, or even intractable for large networks, we only need it at training time. Once trained, the GNN models can be deployed in different communication topologies [28], including those with a larger number of robots as is evidenced in the numerical experiments of Sec. VI-C. Given the decentralized nature of the parametrizations, the trained models are efficient in the sense that their computation is distributed among the agents, demanding only communication exchanges with one-hop neighbors.

### F. Expert Data Generation

As described in our problem statement in Sec. III, the robots operate in a grid world of size $W \times H$ with static obstacles randomly placed throughout the map. For each grid world, we generate *cases* randomly, i.e., problem instances, which consist of pairs of start and goal positions for all robots (we also refer to this as a *configuration*). We filter duplicates or invalid cases, and store the remaining cases in a setup pool, which is randomly shuffled at training time. For each case, we generate the optimal solution. Towards this end, we run an expert algorithm: Conflict-Based Search (CBS) [12] (which is a similar approach as taken in [7]). This expert algorithm computes our 'ground-truth paths' (the sequence of actions for individual robots), within a $300\,s$ timeout, for a given initial configuration. Our data set comprises $30{,}000$ cases for any given grid world and number of agents. This data is divided into a training set ($70\%$), a validation set ($15\%$), and a testing set ($15\%$).

### G. Policy Execution with Collision Shielding

At inference stage, we execute the action policy with a protective mechanism that we name *collision shielding*. Since it is not guaranteed that robots learn collision-free paths, we require this additional mechanism to guarantee that no collisions take place. Collision shielding is implemented as follows: *(i)* if the inferred action would result in a collision with another robot or obstacle, then that action is replaced by an idle action; *(ii)* if the inferred actions of two robots would result in an edge collision (having them swap positions), then those actions are replaced by idle actions. It is entirely possible that robots remain stuck in an idle state until the timeout is reached. When this happens, we count it as a failure case. The overall inference process is summarized in Alg. 1 and Fig. 2.

**Algorithm 1:** Generation of sequential actions.

**Input:** Input tensor, $\mathbf{x}_0^i, i \in [0, N]$, $N$ is the number of robots; timeout $T_{max} = 3T_{\text{MP}^*}$ as explained in Sec. VI-A; Policy $\pi$

**Output:** Predicted paths ($\hat{\chi}_i$) for each robot ($i$), consisting of sequential predicted actions $\hat{u}_t^i$, for all $t \in [0, T_{\text{MP}}]$ from initial position $\mathbf{p}_0^i$

```
for t in [0, T_max] do
    while not all robots at their goals do
        for robot i ∈ {1, ..., N} do
            obtain input tensor x_t^i and adjacency matrix S_t;
            û_t^i ← π(x_t^i, S_t) ;
            if robot i with action û_t^i collides with obstacle then
                û_t^i ← idle (collision shielding);
            end
        end
        if robot i, with action û_t^i, performs an edge collision with robot j then
            û_t^i ← idle (collision shielding);
        else
            record and update position p_{t+1}^i of robot i by û_t^i; update input tensor x_t^i and adjacency matrix S_t.
        end
    end
end
Evaluate χ̂ according to metrics (Sec. VI-A).
```

### H. Dataset Aggregation during Training

The use of collision shielding leads to failure cases due to potential deadlocks in the actions taken by the robots, where some of them remain stuck in an idle state. To overcome such deadlocks, we propose a dataset aggregation method that makes use of an *online expert* (OE) algorithm, during training. More specifically, every $C$ epochs, we select $n_{\text{OE}}$ random cases from the training set and identify which ones are stuck in a deadlock situation. Then, we run the expert starting from the deadlock configuration in order to unlock them into moving towards their goal. The resulting successful trajectory is added to the training set and this extended training set is then used in the following epochs. This process is detailed in Alg. 2. We note that no change is made to the validation or test sets. This dataset aggregation method is similar to the approach in DAgger [29], but instead of correcting every failed trajectory, we only correct trajectories from a randomly selected pool of $n_{\text{OE}}$ cases, as calls to our expert algorithm are time-consuming. Another key difference is that we need to resort to an explicit measure of failure (i.e., through the use of a timeout), since focusing on any deviations from the optimal path (as in the DAgger approach)

**Algorithm 2:** Training process with dataset aggregation.

**Input:** Input tensor, $\mathbf{x}_t^i, t \in [0, T_{\text{MP}^*}], i \in [0, N]$, $N$ is the number of robots; and adjacency matrix $\mathbf{S}_t$; target actions $u_t^{*,i}$ generated expert algorithm; cross-entropy loss $\mathcal{L}$; learning rate $\gamma$; $(\mathbf{x}_t^i, \mathbf{S}_t, u_t^{*,i}) \in$ offline dataset $D_{\text{offline}}$

**Output:** Proposed framework $\pi(\cdot : w)$

```
D ← D_offline ;
π(·: w) ← initialize parameters w ;
for epoch ∈ {1, ..., epoch_max} do
    for {s_t^i, S_t, u_t^{*,i}}_{i=1}^N ∈ D do
        for i ∈ {1, ..., N} do
            û_t^i = π(x_t^i, S_t : w) ;
            w ← w − γ · ∇_w L(û_t^i, u_t^{*,i})
        end
    end
    if mod (epoch, C) = 0 then
        for n_OE randomly selected cases from D_offline do
            Deploy π(·: w) based on Alg. 1;
            Upon timeout, deploy expert algorithm to
              solve failure case D_OE ;
            D ← D ∪ D_OE
        end
    end
end
```

may be misleading, because those paths may still lead to very competitive solutions in our problem setting.

## VI. Performance Evaluation

To evaluate the performance of our method, we perform two sets of experiments, *(i)* on networks trained and tested on the *same* number of robots, and *(ii)* on networks trained on a given number of robots, and tested on *previously unseen* team sizes (both larger and smaller).

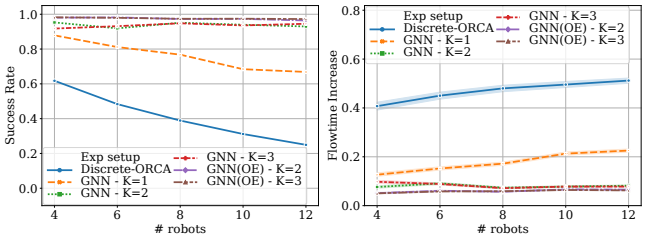
### A. Metrics

1) *Success Rate* ($\alpha$) $= n_{\text{success}}/n$, is the proportion of successful cases over the total number of tested cases $n$. A case is considered successful (*complete*) when *all* robots reach their goal prior to the timeout;

2) *Flowtime Increase* ($\delta_{\text{FT}}$) $= (\text{FT} - \text{FT}^*)/\text{FT}^*$, measures the difference between the sum of the executed path lengths (FT) and that of expert (target) path ($\text{FT}^*$). We set the length of the predicted path $T^i = T_{\max} = 3T_{\text{MP}^*}$ (Alg. 1 and Fig. 2), if the robot $i$ does not reach its goal. Here, $T_{\text{MP}^*}$ is the makespan of the solution generated by the expert algorithm.

### B. Experimental Setup

Our simulations were conducted using a 12-core, 3.2Ghz i7-8700 CPU and an Nvidia GTX 1080Ti GPU with 32 and 11GB of memory, respectively. The proposed network was implemented in PyTorch v1.1.0 [30], and was accelerated with Cuda v10.0 APIs. We used the Adam optimizer with momentum 0.9. The learning rate $\gamma$ scheduled to decay from $10^{-3}$ to $10^{-6}$ within 150 epochs, using cosine annealing. We set the batch size to 64, and L2 regularization to $10^{-5}$. The online expert on the GNN is deployed every $C = 4$ epochs on $n_{\text{OE}} = 500$ randomly selected cases from the training set.




a: Success rate ($\alpha$) b: Flowtime increase ($\delta_{\text{FT}}$)

Fig. 3: Results for success rate ($\alpha$) and flowtime increase ($\delta_{\text{FT}}$), as a function of the number of robots. For each panel, we vary the number of communication hops ($K \in [1, 2, 3]$), including results obtained through training with the online expert (OE). We also compare our framework with Discrete-ORCA [3][31].

### C. Results

We instantiate 600 different maps of size $20 \times 20$, of which 420 are used for training, 90 for validation, and 90 for testing. We generate 50 cases for each map. The obstacle density is set to 10%, corresponding to the proportion of occupied over free space in the environment. We consider a field of view of radius $r_{\text{FOV}} = 4$ and a communication radius of $r_{\text{COMM}} = 5$. At each time step, each robot runs a forwards pass of its local action policy (i.e., the trained network). At the end of each case (i.e., it is either solved or the timeout is reached), we record the length of each robot's path and the number of robots that reach their goals, to compute performance metrics according to Sec. VI-A.

*1) Effect of Communication on Flowtime and Success Rates:* Figures 3a and 3b show results for the success rate and flowtime increase, respectively, as a function of the number of robots. For each panel, we train a model for $N \in [4, 6, 8, 10, 12]$, and test it on instances of the same robot team size. In each experiment, we vary the number of communication hops ($K \in [1, 2, 3]$). Note that for $K = 1$ there is no communication involved. Similar to [3] and [31], we use a discrete version of a velocity-based collision-avoidance method (Discrete-ORCA) as an additional benchmark against which to test our method.

In both figures, we see a drop in performance for larger teams, but this drop is much more pronounced for the non-communicative GNN ($K = 1$). Our framework generally outperforms the Discrete-ORCA in terms of success rate and flowtime increase.

*2) Generalization:* Fig. 4a and 4b summarize the generalization capability of our model for success rate and flowtime increase, respectively. The experiment was carried out by testing networks across previously unseen cases. The tables specify the number of robots *trained on* in the rows, and the number of robots *tested on* in the columns. The results demonstrate strong generalization capabilities.

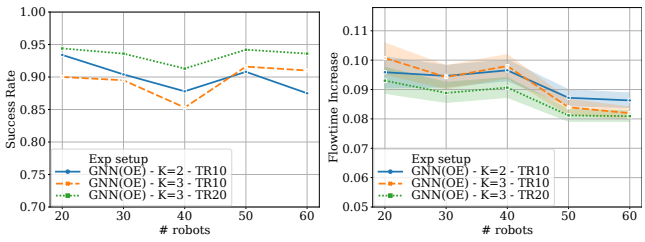
We perform subsequent experiments on larger robot teams to further test the generalization. Results in Fig. 5 show that our network, trained on only 10 robots scales to teams of sixfold size. We test the network in different grid map, where the map sizes are scaled to preserve the effective robot density. Notably, the results show *no* degradation of performance.

We train the GNN ($K \in [2, 3]$) with and without the online


a: Success rate $\alpha$ for GNN(OE) with K=3


b: Flowtime increase $\delta_{\mathrm{FT}}$ for GNN(OE) with K=3

Fig. 4: Success rate and flowtime increase. The rows represent the number of robots on which each model was trained, and columns represent the number of robots at test time. The heatmap maps performance to a color range where purple indicates the best performance and red indicates the worst performance.

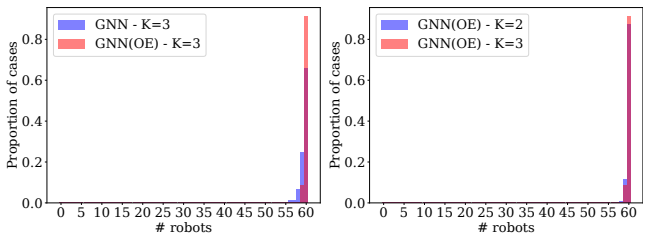



a: Success rate ($\alpha$) b: Flowtime increase ($\delta_{\mathrm{FT}}$)

Fig. 5: Results for success rate and flowtime increase, as a function of the number robots tested on. We vary the GNN implementation ($K \in [2, 3]$), trained ('TR10') on a $20 \times 20$ map with 10 robots, and GNN implementation ($K = 3$) trained ('TR20') on a $28 \times 28$ map with 20 robots. Testing was performed on maps that maintain constant effective robot density.

expert (OE) implementation on 10 robots, and test it on 60 robots in $50 \times 50$ environments, respectively. The grid maps are scaled to preserve the same effective density $\beta = \frac{n_{\text{robots}} + n_{\text{obs}}}{W \times H}$, where the number of obstacles $n_{obs} = \rho \times W \times H$, $\rho$ is the obstacle density in the map ($W \times H$) and $n_{obs}$ is the number of robots.

Different from our success rate metric, which only considers complete cases (all robots reach their goals), Fig. 6 presents the proportion of cases distributed over the number of robots reaching their goals. The distributions show that more than 75% of all robots *always* reach their goals across all implementations. In 97% of cases, more than 95% of robots (57 out of 60) reach their goals. For instance, there are 995 out of 1000 cases (99.5%), where at least 54 robots reach their goals with the GNN ($K = 3$) without OE implementation (worst implementation). We see from Fig. 6a how the GNN network with OE tends to generalize better than the GNN without OE, since the proportion of robots reaching the goal is larger. In Fig. 6b, we see how an increased communication hop count (from $K = 2$ to $K = 3$) contributes to a slightly larger proportion of robots reaching their goals.




a: Effect of online expert b: Effect of K

Fig. 6: Histogram of proportion of cases distributed over the number of robots reaching their goal; the network with hop count $K \in [2, 3]$, is trained on 10 robots and tested on 60 robots, with and without the OE.

## VII. Discussion and Future Work

Our results show that the decentralized framework generalizes to different numbers of robots, as seen in Sec. VI-C.1 and Sec. VI-C.2. We note that a single forward pass of our model (enabling a robot to predict its action) takes only $0.0019 \pm 2.15e^{-4}\ s$ on the workstation described in Sec. VI-B. In addition to the decentralized nature of our solution, this speed of computation is beneficial in real-world deployments, where each robot runs its own (localized) action policy. In contrast, the expert algorithm [12] is intractable for more than 14 agents in dense environments within the given timeout; this is corroborated by results in [6], [7].

The experiments in in Sec. VI-C.2 showed the capability of our decentralized policy to generalize to robot teams across different sizes. Fig. 4a and Fig. 4b showed that the framework trained in smaller robot teams ($n = 4,\ 6$) tends to perform worse than those trained in larger teams ($n = 8,\ 10,\ 12$), across any unseen instances (larger as well as smaller in size). The intuition for the cause of this phenomenon can be due to two main factors. Firstly, larger robot teams tend to cause more collisions, allowing the policy to learn how to plan more efficient paths more quickly. Secondly, policies trained on very small robot teams (e.g. 4 robots), tend to produce communication topologies that are idiosyncratic, and hence, may generalize more poorly. Results in Fig. 5 showed very strong generalization capabilities, with tests scaling to a factor of 6x of the instances trained on, without noticeable performance deterioration.

We also demonstrated that the use of our online expert leads to significant improvements (as seen in Fig. 3). Fig. 6a shows how the GNN with the online expert was able to increase the success rate of all 60 robots reaching goal given a framework trained on 10 robots, and contribute to a right-shift of the distribution.

There are some assumptions and corresponding limitations in the current implementation, which will be improved in future work. Firstly, we assumed that communication between robots was achieved instantly without delay. Time-delayed aggregation GNNs [20] can be introduced to extend our framework to handle time-delayed scenarios. Secondly, inter-robot live-locks and position swaps remain a challenge impeding 100% success. One potential solution to this is to deploy a policy gradient to add a penalty on the action causing such scenarios. Such a strategy (e.g., as implemented in [7]) is harder to train, and is left for future work.

## VIII. Conclusions

We considered the problem of collision-free navigation in multi-robot systems where the robots are restricted in observation and communication range, and possess no global reference frame for localization. We proposed a combined architecture, composed of a convolutional neural network that extracts adequate features from local observations, and a graph neural network that communicates these features among robots. The key idea behind our approach is that we jointly trained these two components, enabling the system to best determine what information is relevant for the team of robots as a whole. This approach was complemented by a data aggregation strategy that facilitated the learning process.

This work is the first to apply GNNs to the problem of multi-robot path planning. Our results show that we are very close to achieving the same performance as first-principles-based methods; in particular, we showed our model's capability to generalize to previously unseen cases involving much larger robot teams. Of particular importance is the fact that we can already scale our system to sizes that are intractable for coupled centralized solvers, while remaining computationally feasible through our decentralized approach.

## IX. Acknowledgments

We gratefully acknowledge the support of ARL grant DCIST CRA W911NF-17-2-0181. A. Prorok was supported by the Engineering and Physical Sciences Research Council (grant EP/S015493/1). We gratefully acknowledge their support. We also thank Binyu Wang and Zhe Liu for their help in providing benchmark algorithms. We thank Alexandre Raymond, Hehui Zheng, Nicholas Timmons, Rupert Mitchell and Weizhe Lin for the helpful discussion. Finally, we would like to thank the anonymous reviewers for the constructive feedback and help in improving this manuscript.

## References

[1] D. Silver, "Cooperative pathfinding," in *AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, 2005, pp. 117–122.

[2] T. Standley and R. Korf, "Complete algorithms for cooperative pathfinding problems," in *International Joint Conference on Artificial Intelligence*. Barcelona, Catalonia, Spain: AAAI Press, 2011, pp. 668–673.

[3] J. Van den Berg, M. Lin, and D. Manocha, "Reciprocal velocity obstacles for real-time multi-agent navigation," in *IEEE International Conference on Robotics and Automation*. IEEE, 2008, pp. 1928–1935.

[4] J. P. van den Berg and M. H. Overmars, "Prioritized motion planning for multiple robots," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*. Edmonton, Alta., Canada: IEEE, Aug 2005, pp. 430–435.

[5] W. Wu, S. Bhattacharya, and A. Prorok, "Multi-Robot Path Deconfliction through Prioritization by Path Prospects," *IEEE International Conference on Robotics and Automation*, 2020, arXiv: 1908.02361.

[6] M. Barer, G. Sharon, R. Stern, and A. Felner, "Suboptimal variants of the conflict-based search algorithm for the multi-agent pathfinding problem," in *European Conference on Artificial Intelligence*, 2014, pp. 961–962.

[7] G. Sartoretti, J. Kerr, Y. Shi, G. Wagner, T. Kumar, S. Koenig, and H. Choset, "Primal: Pathfinding via reinforcement and imitation multi-agent learning," *IEEE Robotics and Automation Letters*, vol. 4, pp. 2378–2385, 2019.

[8] L. Matignon, L. Jeanpierre, and A.-I. Mouaddib, "Coordinated multi-robot exploration under communication constraints using decentralized markov decision processes," in *AAAI Conference on Artificial Intelligence*, 2012.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.

[10] F. Gama, A. G. Marques, G. Leus, and A. Ribeiro, "Convolutional neural network architectures for signals supported on graphs," *IEEE Trans. Signal Process.*, vol. 67, no. 4, pp. 1034–1049, Feb. 2019.

[11] J. Yu and S. M. LaValle, "Structure and intractability of optimal multi-robot path planning on graphs," in *AAAI Conference on Artificial Intelligence*. AAAI Press, 2013, p. 1443–1449.

[12] G. Sharon, R. Stern, A. Felner, and N. R. Sturtevant, "Conflict-based search for optimal multi-agent pathfinding," *Artificial Intelligence*, vol. 219, pp. 40–66, 2015.

[13] C. Ferner, G. Wagner, and H. Choset, "ODrM*: Optimal multirobot path planning in low dimensional search spaces," in *IEEE International Conference on Robotics and Automation*, 2013, pp. 3854–3859.

[14] V. R. Desaraju and J. P. How, "Decentralized path planning for multi-agent teams with complex constraints," *Autonomous Robots*, vol. 32, no. 4, pp. 385–403, 2012.

[15] Y. Wang, N. Matni, and J. C. Doyle, "Separable and Localized System-Level Synthesis for Large-Scale Systems," *IEEE Transactions on Automatic Control*, vol. 63, no. 12, pp. 4234–4249, Dec. 2018.

[16] M. Rotkowitz and S. Lall, "A characterization of convex problems in decentralized control," *IEEE Transactions on Automatic Control*, vol. 50, no. 12, pp. 1984–1996, Dec. 2005.

[17] A. Rajeswaran, K. Lowrey, E. Todorov, and S. Kakade, "Towards generalization and simplicity in continuous control," in *Advances in Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 6553–6564.

[18] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, Sept. 2017, pp. 23–30.

[19] M. Everett, Y. F. Chen, and J. P. How, "Motion planning among dynamic, decision-making agents with deep reinforcement learning," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, Oct. 2018, pp. 3052–3059.

[20] E. Tolstaya, F. Gama, J. Paulos, G. Pappas, V. Kumar, and A. Ribeiro, "Learning decentralized controllers for robot swarms with graph neural networks," in *Conference Robot Learning 2019*. Osaka, Japan: Int. Found. Robotics Res., 30 Oct.-1 Nov. 2019.

[21] A. Khan, E. Tolstaya, A. Ribeiro, and V. Kumar, "Graph policy gradients for large scale robot control," in *Conference on Robot Learning*, 2020, pp. 823–834.

[22] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," *IEEE Transactions on Neural Networks*, vol. 20, no. 1, pp. 61–80, 2009.

[23] A. Prorok, "Graph Neural Networks for Learning Robot Team Coordination," *Federated AI for Robotics Workshop, IJCAI-ECAI/ICML/AAMAS 2018; arXiv:1805.03737 [cs]*, May 2018, arXiv: 1805.03737.

[24] D. K. Duvenaud, D. Maclaurin, J. Iparraguirre, R. Bombarell, T. Hirzel, A. Aspuru-Guzik, and R. P. Adams, "Convolutional networks on graphs for learning molecular fingerprints," in *Advances in Neural Information Processing Systems*, 2015, pp. 2224–2232.

[25] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, "Neural message passing for quantum chemistry," *International Conference on Machine Learning*, vol. 70, 2017.

[26] P. Battaglia, R. Pascanu, M. Lai, D. J. Rezende, and others, "Interaction networks for learning about objects, relations and physics," in *Advances in Neural Information Processing Systems*, 2016, pp. 4502–4510.

[27] A. Ortega, P. Frossard, J. Kovačević, J. M. F. Moura, and P. Vandergheynst, "Graph signal processing: Overview, challenges and applications," *Proc. IEEE*, vol. 106, no. 5, pp. 808–828, May 2018.

[28] F. Gama, J. Bruna, and A. Ribeiro, "Stability properties of graph neural networks," *arXiv:1905.04497v2 [cs.LG]*, 4 Sep. 2019.

[29] S. Ross, G. Gordon, and D. Bagnell, "A reduction of imitation learning and structured prediction to no-regret online learning," in *International Conference on Artificial Intelligence and Statistics*, Fort Lauderdale, FL, USA, 2011, pp. 627–635.

[30] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in PyTorch," in *NIPS Autodiff Workshop*, Long Beach, CA, USA, 2017, pp. 1–4.

[31] B. Wang, Z. Liu, Q. Li, and A. Prorok, "Mobile robot path planning in dynamic environments through globally guided reinforcement learning," *arXiv preprint arXiv:2005.05420*, 2020.