IntNet: A Communication-Driven Multi-Agent Reinforcement Learning Framework for Cooperative Autonomous Driving

Leandro Parada[1], Kevin Yu[1], Panagiotis Angeloudis[1]

***Abstract*— Achieving safety in autonomous driving through Multi-Agent Reinforcement Learning (MARL) is a critical yet challenging task due to non-stationarity, partial observability, and the need for effective coordination among agents. Although earlier cooperative MARL methods have aimed to improve coordination by sharing encoded state observations, we find these strategies are insufficient in safety-critical scenarios. To address this gap, we present IntNet, a novel MARL framework that enhances safety by incorporating the transmission of vehicle intent and adaptive communication scheduling into a unified end-to-end learning paradigm, jointly optimising all components for safe coordination. Central to our approach is an observation prediction module that enables agents to forecast subsequent policy outputs, which they can then share across a vehicle-to-vehicle network. Our model-free architecture employs a Graph Attention Network to process incoming messages and a scheduler component to dynamically optimise the communication graph for bandwidth efficiency. We compare IntNet against state-of-the-art communicative MARL methods in complex urban environments with both autonomous and human-operated vehicles, achieving improved coordination, lower collision rates, and reducing communication efforts by up to 60%. Through extensive experiments, we assess the framework's learning efficiency, scalability, and the balance between information sharing and bandwidth usage for collision-free trajectories.**

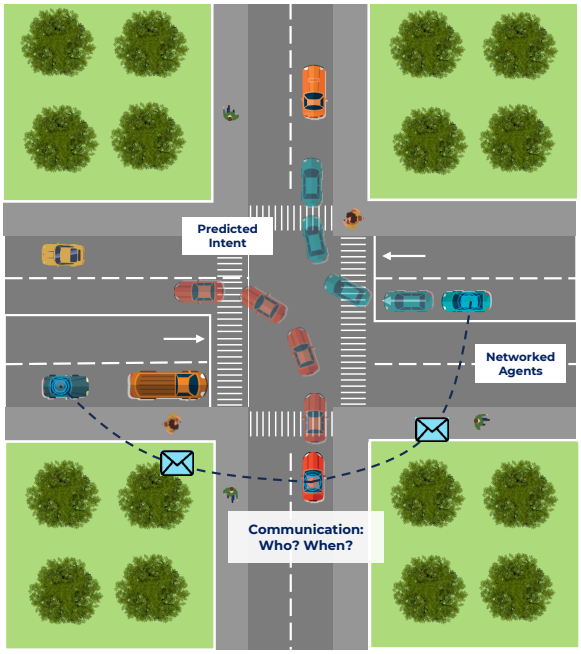


Fig. 1: Illustration of our MARL framework in an urban traffic context, showcasing the strategic exchange of state information and vehicle intent, coordinated by a scheduler to optimise bandwidth and enable real-time decision-making for autonomous vehicles.

I. INTRODUCTION

Deploying Connected and Autonomous Vehicles (CAVs) in dynamic urban environments requires effective coordination to ensure safety. While Multi-Agent Reinforcement Learning (MARL) has shown promise in coordinating autonomous tasks such as multi-robot path planning and navigation [17], applying it to autonomous driving presents distinct challenges. The complex and dynamic nature of urban settings and partial observability complicate the achievement of high levels of safety and coordination. Furthermore, the unpredictability of human driver behaviour adds a layer of complexity to this challenge.

Several applications of MARL in autonomous driving, spanning from intersection management to traffic optimisation, demonstrate its impact on the field. Notable research underscores MARL's efficacy in cooperative motion planning [1], [20], [23], as well as in controlling traffic signals to enhance urban mobility [8], [26]–[28]. Despite these considerable advancements, achieving effective coordination and safety remains the most critical challenge for MARL in autonomous driving [5]. While model-free MARL does not inherently provide safety guarantees, it offers a powerful methodology for addressing safety through trial-and-error learning, particularly in scenarios with diverse and complex configurations.

Agent coordination has been extensively studied within the MARL community by sharing state representations through a communication network [31]. However, we argue that for safety-critical applications such as autonomous driving—-characterised by highly dynamic scenarios-—simply sharing state features is not enough. Understanding the future intentions of others provides additional contextual information to the agents that may otherwise be unavailable, thus unlocking opportunities for safer, more efficient decision making Although some studies have explored intention-sharing in MARL [3], [9], their applications are limited to

Manuscript received: September 2, 2024; Revised: December 9, 2024; Accepted: January 6, 2025.

This paper was recommended for publication by Editor M. Ani Hsieh upon evaluation of the Associate Editor and Reviewers' comments.

[1]L. Parada, K. Yu, and P. Angeloudis are with the Centre for Transport Studies, Department of Civil and Environmental Engineering, Imperial College London, UKlap20@imperial.ac.uk

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

simplistic grid-like environments, lacking the intricacies of autonomous driving, such as high-dimensional observations (e.g., Camera, LiDAR) and autonomous traffic dynamics.

To address the existing gap, we introduce IntNet, an intention-sharing MARL framework tailored for CAVs navigating complex urban environments safely and efficiently. IntNet operates through two communication steps. In the first step, agents share encoded messages and their recent decision-making strategies based on actual, observed information. In the second step, they communicate their estimated intentions over a future prediction horizon H. Each agent uses an observation predictor to forecast environmental states, projecting future strategies over the horizon H to enhance coordination and proactive decision-making. To efficiently manage this process, a scheduler component is employed for bandwidth optimisation, while a graph attention network is used for processing incoming messages. The integration of each component within the end-to-end framework allows safety to emerge naturally from coordinated learning, rather than relying on hard constraints. The overall architecture employs an actor-critic loss alongside an independent supervised loss for training the observation prediction module. Training occurs directly from environmental interaction, eliminating the need for pre-training for the observation predictor. We benchmark IntNet against state-of-the-art Communicative MARL approaches across diverse urban scenarios and show how it can efficiently learn safer policies with low communication graph density.

The key contributions of this work are:

- We propose a novel intention-sharing framework for communicative MARL, integrating Graph Attention Networks (GATs) for processing shared messages and a scheduling mechanism to optimise communication, enhancing coordination and reducing uncertainty in safety-critical scenarios.
- We develop a prediction module for multi-agent settings that forecasts future observations without requiring pretraining, allowing agents to recursively plan future actions.
- We advance beyond prior work in grid-based environments by incorporating high-dimensional observations and realistic vehicle kinematics, providing a more representative evaluation of communicative MARL in autonomous driving scenarios.

The paper is organised as follows: Section II offers a brief literature review on MARL for autonomous driving and recent advancements in Multi-Agent Communicative Reinforcement Learning. Section III details the paper's methodology, encompassing the cooperative MARL framework, training function, and environment description. Section IV discusses the experiments and key findings. Finally, Section V concludes the paper.

II. RELATED WORK

A. MARL for Autonomous Driving

The use of MARL has significantly impacted the field of autonomous driving, with its applications spanning from intersection management to traffic flow optimisation. A few survey works have helped categorise the literature [5], [29]. Recent efforts have demonstrated MARL's efficacy in facilitating collaborative navigation and path planning [1], [20], [23] and traffic signal control [8], [26], [27]. These developments underline MARL's capacity to improve traffic efficiency through intelligent agent collaboration.

Further advancements in MARL for autonomous traffic management include addressing complex scenarios such as the double-merging problem [20], cooperative lane-changing in the presence of Human-Driven Vehicles (HDVs) [30], and highway ramp-merging strategies [2]. Key research areas focus on using Graph Convolutional Neural Networks (GCN) to enhance manoeuvring and safety [6] and applying MARL in challenging settings like near Emergency Vehicles [18] and in occluded situations [19]. Despite these promising efforts, safety and coordination in dynamic urban environments persist as largely unresolved issues in MARL for autonomous driving, underscoring the need for innovative solutions to address the unpredictability and complexity of these scenarios [5].

B. Multi-Agent Communicative Reinforcement Learning

Recent advances in Communicative Multi-Agent Reinforcement Learning (MACRL) highlight the importance of developing efficient communication methods among agents to enhance collaboration and save bandwidth. Early work by Foerster et al. [7] and Sukhbaatar et al. [22] introduced the concept of MARL agents learning to communicate for group decision-making, which was later refined by Jiang et al. [9] and Singh et al. [21], who introduced gating mechanisms for selective communication based on immediate surroundings. Further innovations include TarMAC by Das et al. [4], which uses attention mechanisms for focused communication, and work by Wang et al. [24] and Liu et al. [14] on communication graphs to manage bandwidth and improve precision. Niu et al.'s MAGIC system [16] represents a significant advancement, dynamically adjusting communication based on agents' conditions and environmental factors.

Intention-sharing in MARL has also been explored, with studies like those by Jiang et al. [9] and Chu et al. [3] demonstrating the benefits of agents sharing actions or policy outputs to reduce system non-stationarity. Kim et al. [12] further improved cooperation through intention-sharing using observation and action prediction modules with attention mechanisms, achieving superior performance in grid-based environments. Wang et al. [25] introduced a Bayesian method for decentralised multi-agent collaboration, enabling agents to infer hidden intentions through inverse planning and coordinate both high-level plans and low-level actions. Building on these approaches, our research introduces an intention-sharing MARL framework tailored for autonomous driving, employing Graph Attention Networks (GAT) and a message scheduling mechanism to handle the complexities of high-dimensional observations and communication bandwidth constraints effectively.

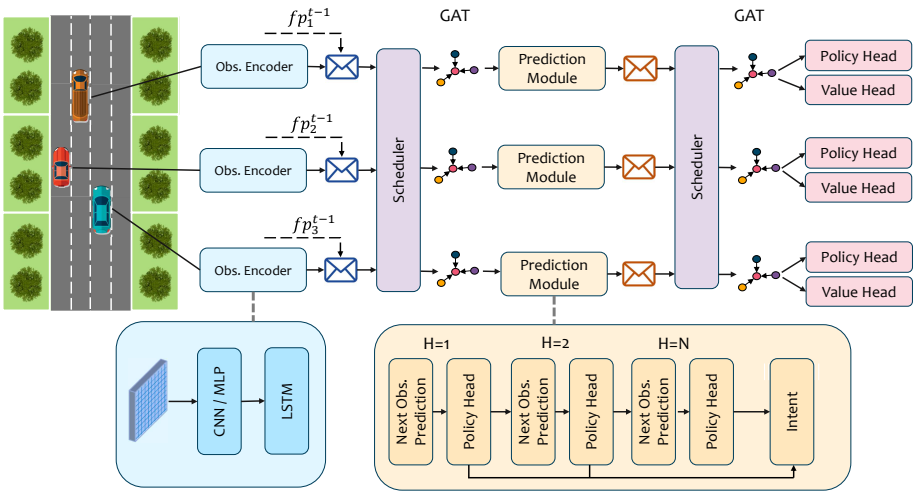


Fig. 2: This figure depicts IntNet, our MARL framework for safe, efficient CAV navigation, highlighting the sharing of encoded observations and vehicle intentions. The scheduler optimises the communication graph, while the GAT module processes incoming messages with attention to prioritise information for decision-making.

III. METHODOLOGY

A. Communication-based Multi-Agent Reinforcement Learning

We formulate the MACRL as a Decentralised Partially Observable Markov Decision Process (Dec-POMDP) [10], represented by the tuple $\langle N, S, A, O, \mathcal{T}, R, \gamma, \mathcal{M}, \mathcal{C}\rangle$. Here, $N = \{1, 2, \ldots, n\}$ is the number of agents. S denotes all possible states, where $s^t \in S$ represents the environment's state at time t. The set A comprises all possible actions, with each agent i selecting an action $a_i^t \in A$ at each time step.

Agents receive partial information via observations $o_i^t \in O$, and the transition between states, given agents' actions, is defined by the transition function $\mathcal{T} : S \times A \to S$. The reward function $R : S \times A \times S \to \mathbb{R}$ determines the immediate reward. Current and future rewards are balanced through the discount factor $\gamma \in [0, 1]$.

Communication occurs through messages $\mathcal{M}$ and is governed by the protocol $\mathcal{C} : \mathcal{M}^N \to \mathcal{M}$, which defines how messages are aggregated. The overall objective is to determine policies $\pi_i : O \times \mathcal{M} \to A$ for each agent that maximise the expected cumulative reward: $\max_{\pi_1,\pi_2,\ldots,\pi_N} \mathbb{E}\left[\sum_{t=0}^{T} \gamma^t r^t(s^t, a_1^t, a_2^t, \ldots, a_N^t, s^{t+1})\right]$, where $r^t(s^t, a^t, s^{t+1})$ is the immediate reward at time t, defined by $R(s^t, a^t, s^{t+1})$.

B. Intention-sharing MARL framework

Our proposed Intention-sharing MARL framework, as illustrated in Figure 2, is designed to facilitate communication among agents in a cooperative setting using a communicative actor-critic model. This framework operates through two communication steps: This framework operates through two communication steps: one for sharing current observations and policy fingerprints, and another for exchanging estimated intentions over a future prediction horizon.

1) Observation Encoding and Policy Fingerprint: In the first step, each agent i encodes its local observation o_i^t at time step t into a message m_i^t using an encoding function $f_{\text{enc}} : O \to \mathcal{M}$. The encoded message $m_i^t = f_{\text{enc}}(o_i^t)$ effectively summarises the observed state information that can be shared with other agents. Alongside this, each agent also broadcasts its policy fingerprint from the previous time step, denoted by $\mathbf{fp}_i^{t-1}$. This fingerprint summarises the agent's recent decision-making context and serves as a signature of its strategy. The policy fingerprint is defined as the probability distribution over the action space A:

$$\mathbf{fp}_i^{t-1} = \left\{\pi_\theta(a \mid o_i^{t-1}, m_{-i}^{t-1}) \mid a \in A\right\}, \tag{1}$$

Where o_i^{t-1} represents the local observation of agent i at time step $t-1$, and m_{-i}^{t-1} represents the messages received from all other agents at time step $t-1$. This approach enables agents to share both their current observations and their recent decision-making context, establishing a foundation for collaborative action planning.

2) Intention Prediction and Sharing: In the second step, each agent predicts and shares its future intentions over a specified time horizon H. These predictions are represented by a sequence of future policy fingerprints $\mathbf{I}_i^{t+H}$, which summarise the agent's projected decision-making strategies for the upcoming H steps. The intention set $\mathbf{I}_i^{t+H}$ is formulated as:

$$\mathbf{I}_i^{t+H} = \begin{bmatrix} \mathbf{fp}_i^{t+H} \\ \mathbf{fp}_i^{t+H-1} \\ \vdots \\ \mathbf{fp}_i^{t+1} \end{bmatrix} \tag{2}$$

where each future policy fingerprint $\mathbf{fp}_i^{t+h}$ for $h = 1, 2, \ldots, H$ is computed as:

$$\mathbf{fp}_i^{t+h} = \left\{ \pi_\theta(a \mid \hat{o}_i^{t+h}, m_{-i}^{t+h}) \mid a \in A \right\}. \tag{3}$$

In this equation, $\hat{o}_i^{t+h}$ represents the predicted observation of agent i at time step $t+h$, which is derived using an observation prediction module f_{obs}. The predicted observation $\hat{o}_i^{t+h}$ is calculated recursively starting from the agent's current observation o_i^t and the policy fingerprints from the other agents. The recursive prediction is defined as:

$$\hat{o}_i^{t+h} = \begin{cases} f_{\text{obs}}\left(o_i^t, \mathbf{fp}_{-i}^{t-1}\right) & \text{if} \quad h = 1 \\ f_{\text{obs}}\left(\hat{o}_i^{t+h-1}, \mathbf{fp}_{-i}^{t+h-1}\right) & \text{if} \quad h > 1 \end{cases} \tag{4}$$

where $\mathbf{fp}_{-i}^{t+h-1}$ denotes the collective policy fingerprints from all agents except i at time step $t+h-1$. At the first step ($h = 1$), the function f_{obs} uses the agent's current observation o_i^t and the policy fingerprints from the other agents at the previous time step $\mathbf{fp}_{-i}^{t-1}$. This initial prediction sets the stage for subsequent recursive steps.

For $h > 1$, f_{obs} recursively updates the predicted observation by taking as input the predicted observation from the previous step $\hat{o}_i^{t+h-1}$ along with the corresponding policy fingerprints $\mathbf{fp}_{-i}^{t+h-1}$. Through this recursive mechanism, f_{obs} effectively captures the evolving dynamics of the environment, informed by the continuous inte between the agent's own evolving state and the intentions of other agents. This allows each agent to form a coherent and forward-looking plan of its anticipated observations over the horizon H, thus enabling more informed and coordinated decision-making.

3) Communication Scheduling and Graph Attention: Given limited bandwidth and the high-dimensional state space in autonomous driving, our approach includes a scheduler for each communication step. This scheduler determines the communication links among agents at each time step t by constructing an adjacency matrix G^t. The element G_{ij}^t of this matrix is 1 if agent i communicates with agent j at time t, and 0 otherwise. This setup allows for optimising communication by reducing unnecessary data exchange, thereby enhancing bandwidth efficiency.

Agents use a Graph Attention Network (GAT) following each communication step to merge and process the received messages. The GAT employs an attention mechanism to dynamically weigh the importance of each message, allowing agents to prioritise the most relevant information. This prioritization is crucial for effective decision-making, particularly in high-dimensional, bandwidth-constrained environments where not all information can be equally considered. This process is mathematically represented as follows. Given a node i with feature representation h_i and its neighbors $\mathcal{N}(i)$, the attention coefficient α_{ij} between node i and its neighbor j is computed as:

$$\alpha_{ij} = \frac{\exp\left(\text{LeakyReLU}\left(\mathbf{a}^T[\mathbf{W}h_i \| \mathbf{W}h_j]\right)\right)}{\sum_{k \in N(i)\cup\{i\}} \exp\left(\text{LeakyReLU}\left(\mathbf{a}^T[\mathbf{W}h_i \| \mathbf{W}h_k]\right)\right)} \tag{5}$$

where $\|$ denotes concatenation, $\mathbf{W}$ is a weight matrix applied to each node, $\mathbf{a}$ is a weight vector for the attention mechanism, and LeakyReLU is a nonlinear activation function. The attention coefficients α_{ij} are normalised across all neighbours k of i using the softmax function, ensuring comparability.

The updated feature representation h_i' of node i, incorporating information from its neighbours, is then calculated as a weighted sum: $h_i' = \sigma\left(\sum_{j \in \mathcal{N}(i)\cup\{i\}} \alpha_{ij} \mathbf{W} h_j\right)$, where σ denotes an optional activation function. This formulation allows the GAT to adaptively prioritise information from the most relevant neighbours, efficiently integrating transmitted messages into meaningful features for enhanced decision-making. Finally, once the agents have processed the incoming messages, they determine the policy and value outputs through independent linear layers as depicted in figure 2.

C. Training

The observation prediction module f_{obs} and the policy are trained concurrently using data collected from episode rollouts. In contrast to previous approaches such as Nagabandi et al. [15], which pre-train the predictor, we do not pre-train f_{obs}. Instead, transitions stored during the rollouts are used to update both the prediction module and the policy simultaneously.

To train the observation prediction module, each episode step generates transitions $\tau = (o^t, a^t, o^{t+1})$, from which we compile a supervised dataset with inputs o^t, a^t and targets o^{t+1}. Observations are normalised, and variables are categorised as continuous or categorical. f_{obs} is trained using the Mean Squared Error (MSE) loss:

$$\mathcal{L}_{obs} = \frac{1}{M} \sum_{i=1}^{M} (o_i^{t+1} - \hat{o}_i^{t+1})^2, \tag{6}$$

where o_i^{t+1} represents the actual next observation, $\hat{o}_i^{t+1}$ denotes the predicted next observation, and M is the total number of samples.

In parallel, the policy is trained using an actor-critic loss, which incorporates the losses of both the actor (policy) and the critic (value):

$$\begin{aligned} \mathcal{L}_{AC} =& \mathbb{E}_{(\tau)} \left[\sum_{t=0}^{T} A^t \cdot \nabla_\theta \log \pi_i(a_i^t | o_i^t, m_i^t) \right] \\ &+ \lambda \sum_{t=0}^{T} \nabla_\rho (R_i^t - V_\rho(o_i^t))^2, \end{aligned} \tag{7}$$

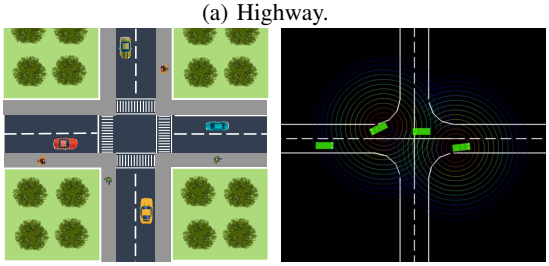
where $A^t = R_i^t - V_\rho(o_i^t)$ is the advantage function, and $V_\rho(o_i^t)$ is the value function.

This concurrent training setup allows the observation prediction module to improve the policy's understanding of environment dynamics in real time, without requiring a separate pre-training phase.

D. Environment Description

We use a modified version of Highway-env [13], adapted for multi-agent urban navigation where agents avoid collisions and reach their destinations to maximise rewards. While Highway-Env serves as the foundational simulator,

(a) Highway.



(b) Intersection.

(c) Lane merging.

Fig. 3: Scenarios used for the experiments.

the addition of real-time communication capabilities and dynamic graph learning significantly increases the complexity of the scenarios. Agents must process communication in real-time, construct dynamic graphs for decision-making, and coordinate in a multi-agent setting, which presents substantial computational and learning challenges. To enhance adaptability, we introduce randomness in vehicles' speeds, locations, and destinations, and limit agent communication to 60 meters to align with DSRC protocol constraints.

The elements of the gym environment can be summarised as follows:

- Agents: CAVs with limited visibility, capable of exchanging messages through a vehicle-to-vehicle communication network.
- Observation: LiDAR-like method using raycasting to segment the 360° field around the agent, detecting other vehicles within an 80-meter range. We use a total of 64 rays and update the observation every timestep.
- Action: Discrete meta-actions including acceleration, deceleration, speed maintenance, and lane changes (left and right). The speed values for each scenario are included in the Appendix. Reward: The reward function is designed for a cooperative game and incorporates components for speed (r_{speed}), collision avoidance (r_{col}), arrival (r_{arr}), and bonuses for cooperative lane merging (r_{merge}). To balance these components, we performed a preliminary grid search over weight configurations, ensuring safety was prioritised above all else. Within configurations that met safety requirements, we tuned other components, such as r_{speed}, to encourage efficient driving. Additionally, we tailored the reward weights to specific scenarios, as certain components, such as r_{arr}, are less relevant in environments like highway driving. This approach ensures that the reward function aligns with the unique objectives of each scenario while promoting coordination and safety. The specific values for each are included in the Appendix.

IV. Experiments

We benchmark our method against baselines across three scenarios: Highway, 4-way Intersection, and Lane Merging, depicted in Figure 3, which includes both schematic and actual environment representations to illustrate the observations. The Highway and Lane Merging environment contain both CAVs and human-driven vehicles (HDV), while the Intersection environment consists of only CAVs. To model the HDVs, we use the IDM/MOBIL car following model [11], which has been extensively used in the literature to represent human-driven behaviour with a simple analytical model.

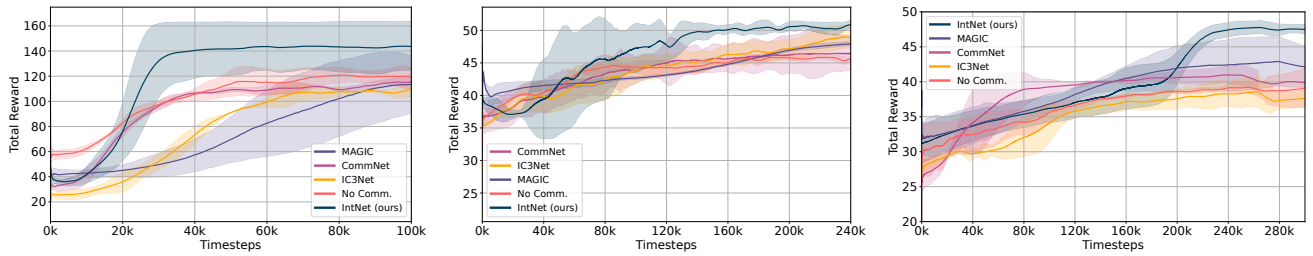
The primary evaluation metrics in our study are the average arrival and collision rates, expressed in percentages. The arrival rate quantifies the percentage of episodes in which all CAVs successfully reached their destinations, while the collision rate measures the percentage of episodes involving at least one CAV collision. Additionally, the average team return is evaluated to assess overall performance. To ascertain the robustness of our training outcomes, we utilise three independent seeds to generate learning curves and validate our results over 100 test episodes. The prediction horizon for all experiments was set to $H = 2$ based on a preliminary experiment, the results of which are provided in the Appendix.

A. Baseline comparison

We evaluated our method's performance against established communication-based MARL approaches, which primarily focus on sharing state or state features. The comparison included three baseline methods: No Communication, CommNet [22], IC3Net [21] and MAGIC [16].

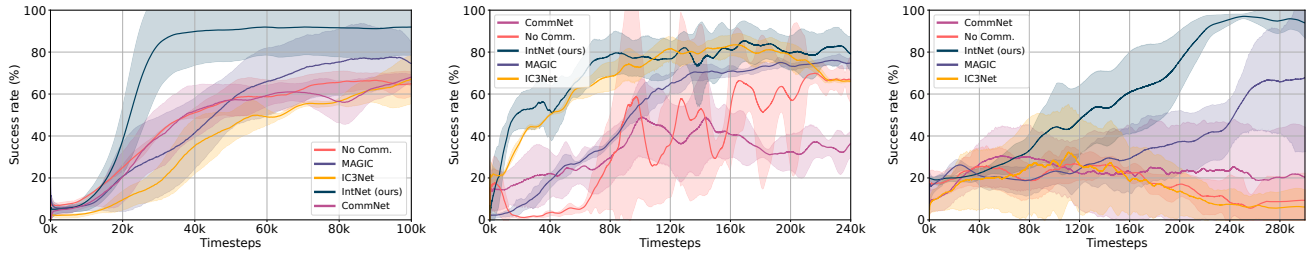
Figures 4 and 5 illustrate the team reward and arrival statistics, respectively, across three scenarios for all methods, with IntNet consistently achieving higher rewards and success rates. Variations in arrival rates indicate that intention-sharing is particularly beneficial in specific scenarios, such as lane merging where it's crucial for safe navigation, while its impact in the intersection scenario is smaller. Table I further details the arrival rate (%) test performance over 100 independent episodes in these scenarios, supporting these observations.

In the highway scenario, our proposed method enables agents to reach high success levels with just 100k training episodes, compared to baselines, which require around 200k episodes for similar success and safety outcomes. In the intersection scenario, although IC3Net shows success comparable to our IntNet method, ours provides a marginally quicker and more consistent learning curve. We see that



(a) Highway. (b) Intersection (c) Lane Merging

Fig. 4: Team reward of the different methods during training.



(a) Highway. (b) Intersection (c) Lane Merging

Fig. 5: Success rate of the different methods during training.

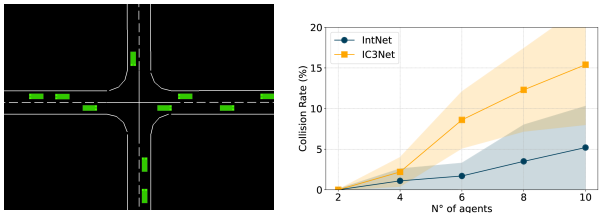
TABLE I: Test arrival rate (%) of different Communicative MARL methods.

Method	Highway	Intersection	Lane Merging
No Comm	62.6 ± 3.6	65.3 ± 8.4	10.3 ± 6.4
IC3Net	59.5 ± 6.1	70.7 ± 6.3	10.5 ± 8.3
CommNet	64.1 ± 8.4	41.0 ± 8.2	26.3 ± 7.6
MAGIC	77.6 ± 8.1	75.4 ± 6.7	66.1 ± 10.1
IntNet (Ours)	97.1 ± 2.6	89.7 ± 5.9	93.2 ± 9.9

all baseline algorithms underperform in the lane-merging scenario. This scenario demands the most cooperation out of the three, causing baseline algorithms to learn inconsistently and often favour solutions that increase the speed at the expense of safety, leading to more collisions.

B. Scalability

To assess the scalability of our methodology, we performed experiments in an intersection scenario, progressively increasing the number of agents within the network. The results, depicted in Figure 6 and table II, show the test collision rates (%) as the agent count varies. As expected, the environment's complexity escalates with each additional agent, making it increasingly difficult to maintain a zero collision rate. Nevertheless, our IntNet approach produces significantly lower collision rates compared to the IC3Net model when the number of agents is increased. The IC3Net model exhibits a growing number of collisions as more agents are added, with rates climbing over 15% when the environment includes ten agents, highlighting the challenges of scaling in more congested settings, even with communication.



(a) (b)

Fig. 6: (a) Intersection scenario with ten agents. (b) Effect of increasing the number of agents on the collision rate.

TABLE II: Test collision rates (%) for IntNet and IC3Net across various agent counts.

Agent Count	IntNet	IC3Net
2	0.0 ± 0.12	0.0 ± 0.05
4	1.1 ± 1.5	2.2 ± 1.8
6	1.7 ± 1.6	8.6 ± 3.5
8	3.5 ± 4.5	12.3 ± 5.12
10	5.2 ± 5.1	15.4 ± 7.4

C. Communication Efficiency

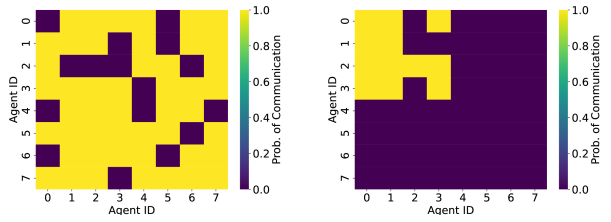
Our architecture includes a scheduler component that optimises message transmission, decreasing the communication network's load. We evaluated its impact through an experiment in an intersection scenario solely with CAVs, contrasting network performance with and without the scheduler. Without the scheduler, the system defaults to a fully connected network, facilitating communication among all agents within range. Table III compares network performance in the intersection environment with the scheduler, revealing that each agent communicates with fewer than half of the agents, on average, while achieving the same success level.

Agents communicate only at crucial times for safe navi-

gation, illustrated by Figure 7, showing the communication graph in the intersection environment at two key moments. Figure 7a reveals the graph as agents approach the intersection, and Figure 7b displays it as most agents exit, with links in 7b (vehicle IDs 0 to 3) indicating vehicles that are still in the intersection.

TABLE III: Effect of the communication scheduler on graph density and success rate.

	Graph Density	Success rate (%)
No Scheduler	1.0 ± 0.0	86.3 ± 5.1
Scheduler	0.41 ± 0.4	88.2 ± 5.8



(a) Entering the intersection. (b) Leaving the intersection

Fig. 7: Heatmaps of the communication graphs when entering and leaving the intersection.

V. CONCLUSIONS

This study introduces IntNet, a communication-driven MARL framework for cooperative autonomous driving, aimed at enhancing coordination and safety by sharing vehicle intent. We achieve this by incorporating an observation prediction module into a communicative actor-critic architecture. Our model-free training approach uses transitions recorded during episode rollouts to simultaneously train the observation predictor and the policy. The results demonstrate the efficacy of our intent-sharing strategy in fostering safer and more efficient policies in complex urban settings with multiple agents while minimising the communication graph density. Our findings also reveal that sharing intent not only enhances safety and efficiency but also accelerates learning convergence, streamlining the path to optimal policies in complex multi-agent environments. Future work will focus on extending the method's applicability to diverse urban scenarios and varying horizons H, with consideration given to communication delays and packet loss.

APPENDIX

A. Implementation details

1) Environment: The table below outlines the distinct sets of base parameters used for each of the three scenarios. The reward weights are listed in the order: $[r_{\text{speed}}, r_{\text{col}}, r_{\text{arr}}, r_{\text{merge}}]$. Values represent the weights assigned to each component in the respective scenarios.

2) IntNet details: The parameters for the proposed method are detailed in the table below:

TABLE IV: Simulation Parameters for the different scenarios.

Parameter	Highway	Intersection	Lane Merging
Lanes Count	4	4	3
N° Agents	4	4	3
N° HDVs	4	2	2
Max Steps	40	15	15
Speed Range	[20, 30]	[0, 9]	[20, 30]
Simulation Frequency	8	10	15
Reward Weights	[0.4, -5.0, 0.0, 0.0]	[1.0, -5.0, 1.0, 0.0]	[0.4, -1.0, 1.0, -0.5]

TABLE V: IntNet set of hyperparameters.

Hyper-parameter	Value
batch size	200
enc_hid_size	128
gat_num_heads	4
gat_num_heads_out	1
gat_hid_size	32
gat_encoder_out_size	32
gamma	0.9
learning_rate	0.0001
value_coeff	0.01

3) Prediction Horizon H: To determine the optimal prediction horizon H, we conducted a preliminary experiment on the Lane Merging Scenario. The architecture was trained with varying H values from 1 to 6, and the average and standard deviation of the test success rate were calculated. The results, shown in Figure 8, indicate that performance decreased for $H > 2$, with larger variability and diminishing returns. Based on these findings, we selected $H = 2$ for our experiments to balance predictive accuracy and practical utility.

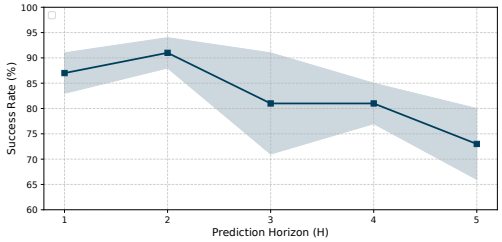


Fig. 8: Average test success rate with varying prediction horizons H in the Lane Merging Scenario.

REFERENCES

[1] E. Candela, L. Parada, L. Marques, T.-A. Georgescu, Y. Demiris, and P. Angeloudis. Transferring Multi-Agent Reinforcement Learning Policies for Autonomous Driving using Sim-to-Real. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2022.

[2] D. Chen, M. R. Hajidavalloo, Z. Li, K. Chen, Y. Wang, L. Jiang, and Y. Wang. Deep Multi-Agent Reinforcement Learning for Highway On-Ramp Merging in Mixed Traffic. *IEEE Transactions on Intelligent Transportation Systems*, 24(11):11623–11638, Nov. 2023. Conference Name: IEEE Transactions on Intelligent Transportation Systems.

[3] T. Chu, S. Chinchali, and S. Katti. Multi-agent Reinforcement Learning for Networked System Control. In *8th International Conference on Learning Representations, ICLR 2020*, Apr. 2020.

[4] A. Das, T. Gervet, J. Romoff, D. Batra, D. Parikh, M. Rabbat, and J. Pineau. TarMAC: Targeted Multi-Agent Communication, Feb. 2020. arXiv:1810.11187 [cs, stat].

[5] J. Dinneweth, A. Boubezoul, R. Mandiau, and S. Espié. Multi-agent reinforcement learning for autonomous vehicles: a survey. *Autonomous Intelligent Systems*, 2(1):27, Nov. 2022.

[6] J. Dong, S. Chen, P. Y. J. Ha, Y. Li, and S. Labi. A DRL-based Multiagent Cooperative Control Framework for CAV Networks: a Graphic Convolution Q Network. *arXiv:2010.05437 [cs, eess]*, Oct. 2020. arXiv: 2010.05437.

[7] J. N. Foerster, Y. M. Assael, N. de Freitas, and S. Whiteson. Learning to Communicate with Deep Multi-Agent Reinforcement Learning, May 2016. arXiv:1605.06676 [cs].

[8] T. Hu, Z. Hu, Z. Lu, and X. Wen. Dynamic traffic signal control using mean field multi-agent reinforcement learning in large scale road-networks. *IET Intelligent Transport Systems*, 2023.

[9] J. Jiang and Z. Lu. Learning Attentional Communication for Multi-Agent Cooperation, Nov. 2018. arXiv:1805.07733 [cs].

[10] L. P. Kaelbling, M. L. Littman, and A. R. Cassandra. Planning and acting in partially observable stochastic domains. *Artif. Intell.*, 101:99–134, 1998.

[11] A. Kesting, M. Treiber, and D. Helbing. General lane-changing model mobil for car-following models. *Transportation Research Record*, 1999(1):86–94, 2007.

[12] W. Kim, J. Park, and Y. Sung. Communication in Multi-Agent Reinforcement Learning: Intention Sharing. Oct. 2020.

[13] E. Leurent. An environment for autonomous driving decision-making, 2018.

[14] Y. Liu, W. Wang, Y. Hu, J. Hao, X. Chen, and Y. Gao. Multi-Agent Game Abstraction via Graph Attention Neural Network, Nov. 2019. arXiv:1911.10715 [cs].

[15] A. Nagabandi, G. Kahn, R. S. Fearing, and S. Levine. Neural Network Dynamics for Model-Based Deep Reinforcement Learning with Model-Free Fine-Tuning, Dec. 2017. arXiv:1708.02596 [cs].

[16] Y. Niu, R. Paleja, and M. Gombolay. Multi-Agent Graph-Attention Communication and Teaming. 2021.

[17] J. Orr and A. Dutta. Multi-Agent Deep Reinforcement Learning for Multi-Robot Applications: A Survey. *Sensors*, 23(7):3625, Jan. 2023. Number: 7 Publisher: Multidisciplinary Digital Publishing Institute.

[18] L. Parada, E. Candela, L. Marques, and P. Angeloudis. Safe and efficient manoeuvring for emergency vehicles in autonomous traffic using multi-agent proximal policy optimisation. *Transportmetrica A: Transport Science*, 0(0):1–29, 2023.

[19] L. Parada, H. Tian, J. Escribano, and P. Angeloudis. An End-to-End Collaborative Learning Approach for Connected Autonomous Vehicles in Occluded Scenarios. In *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*, 2023.

[20] S. Shalev-Shwartz, S. Shammah, and A. Shashua. Safe, Multi-Agent, Reinforcement Learning for Autonomous Driving. *arXiv:1610.03295 [cs, stat]*, Oct. 2016. arXiv: 1610.03295.

[21] A. Singh, T. Jain, and S. Sukhbaatar. Learning when to Communicate at Scale in Multiagent Cooperative and Competitive Tasks, Dec. 2018. arXiv:1812.09755 [cs, stat].

[22] S. Sukhbaatar, A. Szlam, and R. Fergus. Learning Multiagent Communication with Backpropagation, Oct. 2016. arXiv:1605.07736 [cs].

[23] D. Troullinos, G. Chalkiadakis, I. Papamichail, and M. Papageorgiou. Collaborative Multiagent Decision Making for Lane-Free Autonomous Driving. 2021.

[24] R. Wang, X. He, R. Yu, W. Qiu, B. An, and Z. Rabinovich. Learning Efficient Multi-agent Communication: An Information Bottleneck Approach, June 2020. arXiv:1911.06992 [cs].

[25] R. E. Wang, S. A. Wu, J. A. Evans, J. B. Tenenbaum, D. C. Parkes, and M. Kleiman-Weiner. Too many cooks: Coordinating multi-agent collaboration through inverse planning. *CoRR*, abs/2003.11778, 2020.

[26] T. Wang, J. Cao, and A. Hussain. Adaptive traffic signal control for large-scale scenario with cooperative group-based multi-agent reinforcement learning. *Transportation research part C: emerging technologies*, 125:103046, 2021.

[27] Z. Wang, K. Yang, L. Li, Y. Lu, and Y. Tao. Traffic signal priority control based on shared experience multi-agent deep reinforcement learning. *IET Intelligent Transport Systems*, 17(7):1363–1379, 2023.

[28] T. Wu, P. Zhou, K. Liu, Y. Yuan, X. Wang, H. Huang, and D. O. Wu. Multi-agent deep reinforcement learning for urban traffic light control in vehicular networks. *IEEE Transactions on Vehicular Technology*, 69(8):8243–8256, 2020.

[29] P. Yadav, A. Mishra, and S. Kim. A Comprehensive Survey on Multi-Agent Reinforcement Learning for Connected and Automated Vehicles. *Sensors*, 23(10):4710, Jan. 2023. Number: 10 Publisher: Multidisciplinary Digital Publishing Institute.

[30] W. Zhou, D. Chen, J. Yan, Z. Li, H. Yin, and W. Ge. Multi-agent Reinforcement Learning for Cooperative Lane Changing of Connected and Autonomous Vehicles in Mixed Traffic. *Autonomous Intelligent Systems*, 2(1):5, Dec. 2022. arXiv: 2111.06318.

[31] C. Zhu, M. Dastani, and S. Wang. A Survey of Multi-Agent Reinforcement Learning with Communication, Mar. 2022. arXiv:2203.08975 [cs].