

Interpretable End-to-end Urban Autonomous Driving with Latent Deep Reinforcement Learning

Jianyu Chen¹, Shengbo Eben Li², Masayoshi Tomizuka¹

Abstract—Unlike popular modularized framework, end-to-end autonomous driving seeks to solve the perception, decision and control problems in an integrated way, which can be more adapting to new scenarios and easier to generalize at scale. However, existing end-to-end approaches are often lack of interpretability, and can only deal with simple driving tasks like lane keeping. In this paper, we propose an interpretable deep reinforcement learning method for end-to-end autonomous driving, which is able to handle complex urban scenarios. A sequential latent environment model is introduced and learned jointly with the reinforcement learning process. With this latent model, a semantic birdeye mask can be generated, which is enforced to connect with a certain intermediate property in today’s modularized framework for the purpose of explaining the behaviors of learned policy. The latent space also significantly reduces the sample complexity of reinforcement learning. Comparison tests with a simulated autonomous car in CARLA show that the performance of our method in urban scenarios with crowded surrounding vehicles dominates many baselines including DQN, DDPG, TD3 and SAC. Moreover, through masked outputs, the learned policy is able to provide a better explanation of how the car reasons about the driving environment. The codes and videos of this work are available at our github repo[†] and project website[‡].

Index Terms—Autonomous driving, Deep reinforcement learning, End-to-end driving policy, Probabilistic graphical model, Interpretability.

I. INTRODUCTION

Most of today’s autonomous driving systems are using a highly modularized hand-engineered approach, for example, perception, localization, behavior prediction, decision making and motion control, etc [1], [2]. Take the perception module as an example: even though some learning techniques are used, its design still needs tedious hand-engineered work like selecting representation features of each types of road users. Even though working well in a few driving tasks, this modularized framework starts to touch its performance limitation in urban driving scenarios because (1) too much human heuristics can lead to conservative driving policy; (2) it is hard to generalize as we might need to redesign the heuristics for each new scenario and task, and (3) these modules are strongly

¹Department of Mechanical Engineering, University of California, Berkeley, CA 94720, USA. Email: jianyuchen@berkeley.edu, tomizuka@berkeley.edu

²State Key Lab of Automotive Safety and Energy, School of Vehicle and Mobility, Tsinghua University, Beijing 100084, China. Email: lishbo@tsinghua.edu.cn

[†]<https://github.com/cjy1992/interp-e2e-driving>

[‡]<https://sites.google.com/berkeley.edu/interp-e2e/>

entangled with each other, and the whole system becomes expensive to scale and maintain.

Those limitations might be avoided with end-to-end autonomous driving approaches, in which a driving policy can be learned and generalized to new tasks without much hand-engineered involvement [3]–[5]. Moreover, the learned policy can be continuously optimized in driving, which is possible to achieve superhuman performance. Two main branches for end-to-end autonomous driving are imitation learning (IL) [3], [4], [6], [7], which learns a driving policy by imitating the collected expert driving data, and reinforcement learning (RL) [8]–[10], which learns a policy by self exploration and reinforcement. However, existing end-to-end methods are criticized by two main shortcomings: 1) The learned policies are quite lack of interpretability because neural network is like a black-box. When a deep neural network is learned directly from raw observations to control command, we can not explain how it works. 2) They can only deal with simple driving tasks such as lane keeping or car-following. However, urban autonomous driving is much more complex due to highly dynamic road traffic and strong road user interaction. The various urban scenarios and street views significantly increase the sample complexity, making it extremely challenging to learn a good end-to-end driving policy.

This paper introduces the maximum entropy RL with sequential latent variables to address the problems in end-to-end autonomous driving. The latent space is employed to encode the complex urban driving environment, including visual inputs, spatial features, road conditions and road users’ states. Historical high-dimensional raw observations are compressed into this low-dimensional latent space with a sequential latent environment model, which is learned jointly with maximum entropy reinforcement learning process.

The introduced latent space enables an interpretable explanation of how the policy reasons about the environment by decoding the latent state to a semantic birdeye mask. During training, this mask is enforced to connect with some intermediate properties in today’s modularized framework, for example, localization & mapping, object detection, and behavior prediction, thus providing an explanation on the learned policy. Meanwhile, the latent space provides a much more compact state representation, which significantly reduces sample complexity of learning the driving policy, resulting in a large performance improvement. We implemented our method to learn an end-to-end driving policy from raw camera and lidar inputs in CARLA simulator. Experimental evaluation demonstrates that our method significantly outperforms prior methods in crowded

urban scenarios. Examples of decoded semantic birdeye masks are presented to illustrate how our autonomous car understands the driving situations.

II. RELATED WORKS

Recent advances in machine learning enables the possibility of learning based end-to-end approaches for autonomous driving. There are two main approaches: imitation learning (IL) and reinforcement learning (RL). IL learns a driving policy from expert driving data [3], [4], [6], [7]. With expert samples as labelled data, a driving policy is often easy to train, and it generally works well in structured driving tasks if one can collect enough expert data. However, there are fundamental limitations for IL: (1) IL is data hungry, and moreover its performance is limited to the average of the demonstration data; (2) IL is unable to learn skills that are not provided or rare in the demonstration data. This makes it difficult to deal with some dangerous scenarios such as near collision cases because they might never be demonstrated by the expert.

Combined with deep learning techniques, RL shows its power on tackling complex decision making and planning problems, bringing a series of breakthroughs in recent years. Agents trained with deep RL techniques achieves super-human-level performance in game playing [11]–[13], go playing [14], [15], and robotics [16], [17]. Related deep RL algorithms range from value based methods such as DQN [11], [12] and double DQN [18], actor-critic based methods such as A3C [19], DDPG [9] and TD3 [20], policy optimization based methods such as TRPO [21] and PPO [22], and maximum entropy RL methods such as SAC [23], [24]. With RL, a policy can be learned automatically without any expert data. It can explore various kinds of possible cases including some dangerous ones, and then learn related skills. It also has the potential to achieve superhuman performance.

Researchers have been trying to apply deep RL techniques to the domain of autonomous driving. Wolf et al. [8] used DQN to learn to steer an autonomous car to keep in the track in simulation. Its action space is discrete and only allows coarse steering angles. Lillicrap et al. [9] proposed a continuous control deep RL algorithm which learns a deep neural network policy that is able to drive the autonomous car on a simulated racing track. Chen et al. [25] proposed a hierarchical deep RL framework to solve driving scenarios with complex decision making such as traffic light passing. Kendall et al. [10] demonstrated the first application of deep RL to real world autonomous driving. They learned a deep lane keeping policy using a single front-view camera image as input. There are a lot of other related works not mentioned here. However, existing works are either for simple scenarios without complex road conditions and multi-agent interactions, or use manually designed feature representations.

Another problem of learning-based approaches for autonomous driving is that they are lack of interpretability. The learned deep neural network policy is like a black box, which is not ideal since autonomous driving is a safety critical real world application. It is important for us to know whether and

how the autonomous car understand the environment. Some works have made efforts in this direction. Bojarski et al. [26] visualized NVIDIA’s deep neural network based driving system by extracting the convolutional layer feature maps and finding the salient objects. Kim et al. [27] used a visual attention model with a causal filter to visualize the attention heatmap. Sauer et al. [28] analyzed the decision making process of the deep neural network by using gradient-weighted class activation maps to obtain the attention of the CNN. However, the interpretable information they provide — mostly just tell which part of the observed image is within attention — is rather weak.

Probabilistic graphical model (PGM) is a generic and powerful tool to formulate many machine learning problems [29]. Sequential latent model [30]–[34] is one of its very relevant applications to this work, which uses PGM to formulate stochastic time sequence processes with latent variables. Close connections are also found between PGM and maximum entropy reinforcement learning [35]–[37]. Some recent works propose to integrate sequential latent model learning and reinforcement learning [33], [34], [38], [39]. Such methods show great potential in end-to-end learning of deep policies with high dimensional inputs. However, no prior works have used this branch of techniques to formulate and solve autonomous driving problems. Furthermore, they do not provide interpretability of the learned model, and do not take multiple sources of sensor inputs, which is essential for autonomous driving systems.

III. PGM FOR ENVIRONMENT MODELING AND REINFORCEMENT LEARNING

A. Probabilistic Graphical Model (PGM)

Probabilistic graphical model (PGM) is probabilistic, but uses a graph to represent conditional dependence between random variables [29]. They are widely used in Bayesian statistics and Bayesian learning. Fig.1 shows a simple example of PGM. There are in total 4 nodes A, B, C and D. These nodes can represent random variables meaning observable quantities, unobservable latents, or unknown parameters. The edges between nodes represents conditional dependencies. In Fig.1, C is conditioned on A and B, while D is conditioned on C. Each edge is associated with a conditional probability, such as $p(C|A, B)$, $p(D|C)$. With the ability to describe complex causal effects and probabilistic transitions, PGM can be used as a universal model in stochastic RL. We will now introduce how PGM is used to model a discrete-time dynamic environment.

B. PGM for Sequential Latent Environment Modeling

To obtain optimal policy, it is crucial to accurately model the environment. The environment in its nature has the following characteristics: (1) High dimensional observations: either for a human being or an autonomous car, the raw observations for them are usually high dimensional, such as RGB images; (2) Time-sequence probabilistic dynamics: the state of the environment will change with time, thus time sequence relations should be modeled; (3) Partially observable: the observation at the current time alone might not be enough to recover full

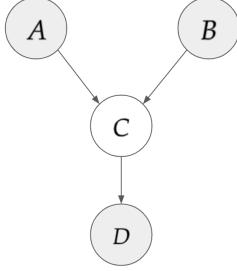


Fig. 1: A simple example of probabilistic graphical model

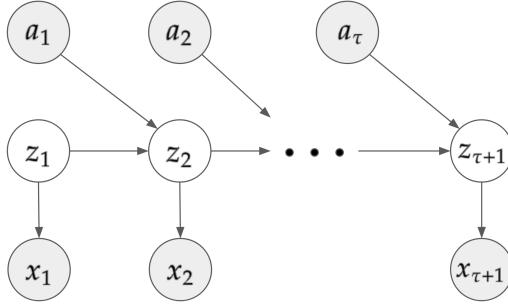


Fig. 2: A PGM sequential latent environment modeling

state of the environment, historical information needs to be summarized by historical observations.

Here we introduce a probabilistic sequential latent environment model, which satisfies the above stated characteristics. Similar structures of this model is adopted by multiple literatures [30], [33], [34]. As shown in Fig.2, x_t represents the observation at time step t , which can be high dimensional sensor inputs such as RGB images. a_t is the action chosen at t . z_t is the latent state variable at t , which is a description of the current situation summarizing historical information, e.g, the position, velocity, intention of other road participants, the drivable areas, and the road markings. The observation x_t is a decoding of the latent state z_t , defined by $p(x_t|z_t)$. The latent state z_t together with the action a_t , decide the latent state at the next time step by the state transition function $p(z_{t+1}|z_t, a_t)$.

This environment model is quite generic, as there is no restrictions of the format and physical meaning of observation, action, and latent state. Furthermore, the observation decoding function $p(x_t|z_t)$ and state transition function $p(z_{t+1}|z_t, a_t)$ can be arbitrarily complex, such as deep neural networks.

By introducing an additional filtering function $p(z_{t+1}|z_t, x_{t+1}, a_t)$, the latent state can be inferred in a recursive bayesian filter way. Given a new observation x_{t+1} , we have $p(z_{t+1}) = p(z_{t+1}|z_t, x_{t+1}, a_t) p(z_t)$, where a_t is the action executed at the last time step. The latent state for the first time step is obtained by $p(z_1) = p(z_1|x_1)$. We can also make probabilistic predictions by rolling out the future states based on the state transition function:

$$p(z_{\tau:\tau+H}|a_{\tau:\tau+H-1}) = p(z_\tau) \prod_{t=\tau}^{\tau+H-1} p(z_{t+1}|z_t, a_t) \quad (1)$$

Furthermore, with the decoding networks, we can not only

decode to the raw observations for unsupervised learning, but can also decode to any other representations, such as a semantic mask to provide interpretable explanations.

We can fit the parameters ψ of this PGM from dataset, which is composed of observation-action trajectory sequences $\mathcal{D} = \{(x_{1:\tau}^i, a_{1:\tau}^i)\}_{i=1}^N$, by maximizing the likelihood of the data:

$$\max_{\psi} \prod_{i=1}^N p(x_{1:\tau}^i | a_{1:\tau}^i) \quad (2)$$

C. PGM for Reinforcement Learning

Under the settings of reinforcement learning [40], at each time step, an agent observes the state z_t , executes action a_t generated by its policy $a_t \sim \pi(a_t|z_t)$, and then gets the reward $r(z_t, a_t)$. The state is then updated according to the state transition $z_{t+1} \sim p(z_{t+1}|z_t, a_t)$. Assume there are H time steps in an episode and the initial state is generated by $z_1 \sim p(z_1)$, then the objective of reinforcement learning is to find an policy that optimizes the expected accumulative rewards:

$$\pi^* = \operatorname{argmax}_{\pi} \mathbb{E}_{\substack{z_1 \sim p(z_1) \\ a_t \sim \pi(a_t|z_t) \\ z_{t+1} \sim p(z_{t+1}|z_t, a_t)}} \sum_{t=1}^H r(z_t, a_t) \quad (3)$$

Note here we do not explicitly write the discount factor γ in the accumulative rewards, instead we incorporate the discount factor by modifying the state transition model [35]. If the initial state transitions are given by $p(z_{t+1}|z_t, a_t)$, adding a discount factor is equivalent to undiscounted problem under the modified state transitions $\bar{p}(z_{t+1}|z_t, a_t) = \gamma p(z_{t+1}|z_t, a_t)$, where there is an additional transition with probability $1 - \gamma$, regardless of action, into an absorbing state with reward zero. The discount factor allows convergence of the value function in infinite-horizon settings. Without loss of generality, we will omit γ from the PGM related derivations in this paper, but it can be inserted trivially in all cases simply by modifying the state transition models as mentioned above. The discount factor is revisited as an explicit consideration in our reinforcement learning algorithm implementation in V-C.

Maximum entropy reinforcement learning (MaxEnt RL) [23], [35], [41] modifies the above standard RL by adding an entropy regularization term $\mathcal{H}(\pi(a_t|z_t)) = -\log \pi(a_t|z_t)$ to the reward. Now considering we are using a parametric function as the policy π_ϕ , for example a deep neural network with weights ϕ , then the objective of MaxEnt RL can be written as:

$$\phi^* = \operatorname{argmax}_{\phi} \mathbb{E}_{\substack{z_1 \sim p(z_1) \\ a_t \sim \pi_\phi(a_t|z_t) \\ z_{t+1} \sim p(z_{t+1}|z_t, a_t)}} \sum_{t=1}^H [r(z_t, a_t) - \log \pi_\phi(a_t|z_t)] \quad (4)$$

There are several reasons why we would like to use MaxEnt RL instead of standard RL [42]. First, it performs better exploration. Standard RL requires specific exploration strategies such as adding noise to the policy. However, MaxEnt RL has a stochastic policy by default, thus the policy itself includes the exploration strategy, which is optimized during RL training.

In practice, the performance of MaxEnt RL is often better and more robust than standard RL algorithms.

Second, MaxEnt RL can be interpreted as learning a PGM. As shown in Fig.3, z_t represents the state, a_t is the action, and O_t is a binary random variable. The use of O_t is to indicate whether the agent is acting optimally at time step t. Its conditional probability is defined by:

$$p(O_t = 1|z_t, a_t) = \exp(r(z_t, a_t)) \quad (5)$$

thus higher reward indicates higher optimality. Therefore, to make the agent act optimally, we want to maximize the probability of optimality in the whole trajectory $p(O_{1:H})$. Let's now look at its log probability:

$$\begin{aligned} \log p(O_{1:H}) &= \log \int \int p(O_{1:H}, z_{1:H}, a_{1:H}) dz_{1:H} da_{1:H} \\ &= \log \int \int p(O_{1:H}, z_{1:H}, a_{1:H}) \\ &\quad \frac{q(z_{1:H}, a_{1:H})}{q(z_{1:H}, a_{1:H})} dz_{1:H} da_{1:H} \quad (6) \\ &= \log \mathbb{E}_{q(z_{1:H}, a_{1:H})} \left[\frac{p(O_{1:H}, z_{1:H}, a_{1:H})}{q(z_{1:H}, a_{1:H})} \right] \\ &\geq \mathbb{E}_{q(z_{1:H}, a_{1:H})} [\log p(O_{1:H}, z_{1:H}, a_{1:H}) \\ &\quad - \log q(z_{1:H}, a_{1:H})] \end{aligned}$$

The above inequality is obtained by adding a variational distribution $q(z_{1:H}, a_{1:H})$ and then applying Jensen's inequality. The variational distribution should be the trajectory distribution generated by the current policy $\pi(a_t|z_t)$:

$$q(z_{1:H}, a_{1:H}) = p(z_1) \pi(a_H|z_H) \prod_{t=1}^{H-1} p(z_{t+1}|z_t, a_t) \pi(a_t|z_t) \quad (7)$$

The optimality distribution of the trajectory is:

$$\begin{aligned} p(O_{1:H}, z_{1:H}, a_{1:H}) &= p(O_{1:H}|z_{1:H}, a_{1:H}) p(z_{1:H}, a_{1:H}) \\ &= \exp \left(\sum_{t=1}^H r(z_t, a_t) \right) \quad (8) \\ &\quad p(z_1) \prod_{t=1}^{H-1} p(z_{t+1}|z_t, a_t) \end{aligned}$$

By cancellation of repeated terms, the inequality (6) becomes:

$$\log p(O_{1:H}) \geq \mathbb{E}_{q(z_{1:H}, a_{1:H})} \sum_{t=1}^H [r(z_t, a_t) - \log \pi(a_t|z_t)] \quad (9)$$

Note that we can maximize the left side by maximizing the right side, and the right side of the inequality is exactly the same objective of MaxEnt RL. This means, we can use MaxEnt RL to maximize the likelihood of optimality variables in the PGM in Fig.3. In this sense, the reinforcement learning problem is reformulated into a learning problem for the PGM shown in Fig.3.

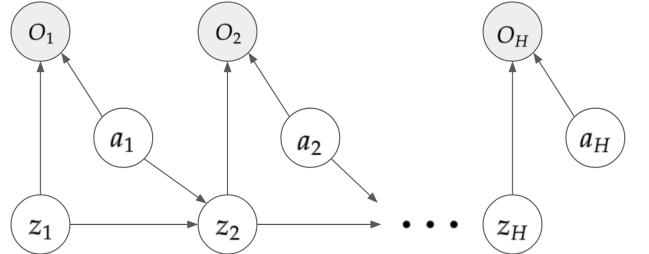


Fig. 3: A PGM for maximum entropy reinforcement learning.

IV. INTERPRETABLE END-TO-END URBAN AUTONOMOUS DRIVING

A. PGM for Interpretable Urban Autonomous Driving

There are two main building blocks for urban autonomous driving. The first is the perception and recognition module, which helps the autonomous car to understand the current driving situation, such as where is the ego vehicle, what is the road condition, and where are the surrounding road participants. Furthermore, it needs to be able to reason about what will happen in the future, such as where will the ego car and surrounding road participants go. These information should be obtained given the historical high dimensional raw sensor inputs. The second module is planning and control, which helps the autonomous car decide what action to take.

Using the methods mentioned in Section III, the above two building blocks can be formulated by two PGMs separately, and it's natural to combine the two PGMs into a single one. Inspired by recent works that combines latent representation learning and reinforcement learning [33], [34], [38], we present our PGM for urban autonomous driving, as shown in Fig.4. Same to the notations in Section III, z_t represents for the latent state, a_t represents for action, O_t represents for the optimality variable, and x_t represents for the sensor inputs. Note here we allow sensor inputs from multiple sources.

We have a newly introduced variable, m_t , which we call the mask. It contains semantic meanings of the environment in a human understandable way. Details about this mask is described in Section IV-B. The main purpose of the mask is to provide interpretability for the system. At training time we need to provide the ground truth labels of the mask, but at test time, the mask can be decoded from the latent state, showing how the system is understanding the environment semantically.

After learning this PGM in Fig.4, the following modules can be obtained:

- 1) **Policy** $p(a_t|z_t)$: Given the latent state, the policy model tells how to choose the action.
- 2) **Inference** $p(z_{t+1}|x_{1:t+1}, a_{1:t})$: With historical sensor inputs and actions, the inference model infers the current latent state.
- 3) **Latent dynamics** $p(z_{t+1}|z_t, a_t)$: This helps predict the future states.
- 4) **Generative models** $p(x_t|z_t), p(m_t|z_t)$: $p(x_t|z_t)$ decodes the latent state z_t to raw sensor inputs x_t , showing how much information the latent state captures. $p(m_t|z_t)$ generates the semantic mask m_t to provide interpretability.

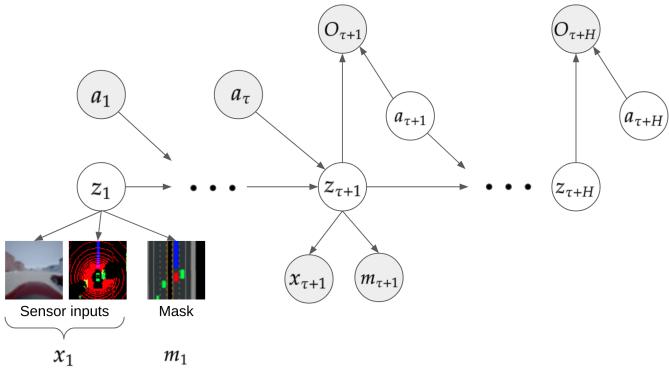


Fig. 4: A PGM for interpretable end-to-end urban autonomous driving

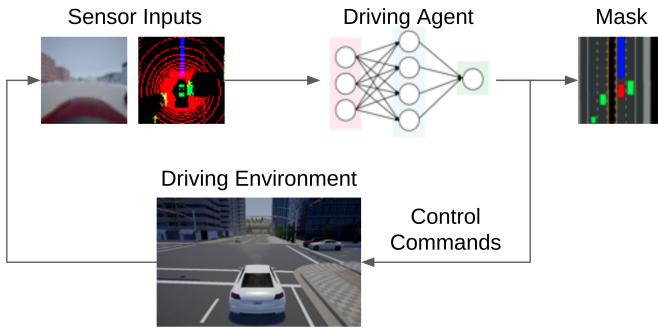


Fig. 5: The interpretable end-to-end urban autonomous driving agent

The whole model can be trained end-to-end. After training, an intelligent driving agent containing an interpretable environment model and a driving policy is obtained. As shown in Fig.5, the agent takes multi-modal sensor inputs from the driving environment, and then output control commands to drive the car in urban scenarios. In the meantime, the agent generates a semantic mask to interpret how it understand the current driving situation.

B. Sensor Inputs and Mask

We use two sensors to provide the observations, camera and lidar. For camera, the sensor input is a front-view RGB image, which can be represented by a tensor of $\mathbb{R}^{64 \times 64 \times 3}$. For lidar, we project the point clouds to the ground plane and render them into a 2D lidar image. The lidar image is represented by a tensor of $\mathbb{R}^{64 \times 64 \times 3}$, with each pixel rendered in red or green depending on whether there are lidar points at or above ground level existing in the corresponding pixel cell. Desired route constituted of waypoints are rendered in blue.

We use camera and lidar together because they are both important sensor sources and provide complementary information. Lidar point clouds provides accurate spatial information of other road participants and obstacles in 360 degrees of view. While the front-view camera is good at providing information of the road conditions.

The semantic mask provides bird-view semantics of the road conditions and objects, which is represented by a tensor of

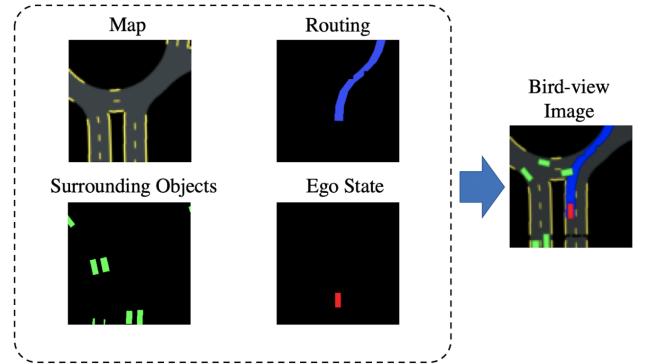


Fig. 6: The bird-view semantic mask for urban autonomous driving

$\mathbb{R}^{64 \times 64 \times 3}$. As shown in Fig.6, the mask is composed of the following four parts:

1) **Map:** Map contains information of road conditions. Drivable areas and lane markings are rendered in the map.

2) **Routing:** Routing contains information of waypoints, which is provided by a route planner. It is rendered as a thick blue polyline.

3) **Detected Objects:** Historical bounding boxes of detected surrounding road participants (e.g., vehicles, bicycles and pedestrians) are rendered as green boxes.

4) **Ego State:** The bounding box of the ego vehicle is rendered as a red box.

V. JOINT LEARNING OF ENVIRONMENT MODEL AND DRIVING POLICY

A. Variational Inference for Joint Model Learning and Policy Learning

The environment model and driving policy can be learned jointly by learning the PGM shown in Fig.4. For convenience, we first introduce some notations. Denote a trajectory to be composed of sensor inputs, masks, actions and rewards:

$$\vec{x} = x_{1:\tau+1}, \vec{m} = m_{1:\tau+1}, \vec{a} = a_{1:\tau}, \vec{r} = r_{1:\tau} \quad (10)$$

The dataset is then composed of this kind of trajectories collected during the exploration phase $\mathcal{D} = \{(\vec{x}^i, \vec{m}^i, \vec{a}^i, \vec{r}^i)\}_{i=1}^N$. We further denote:

$$\begin{aligned} \vec{z} &= z_{1:\tau+1}, \vec{z}^w &= z_{1:\tau+H}, \vec{z}^p &= z_{\tau+1:\tau+H}, \\ \vec{O}^p &= O_{\tau+1:\tau+H}, \vec{a}^p &= a_{\tau+1:\tau+H} \end{aligned} \quad (11)$$

where the superscript p stands for post, and w stands for whole. The learning objective is to maximize the log likelihood of the sensor inputs, mask and the optimality variables:

$$\log \prod_{(\vec{x}, \vec{m}, \vec{a}, \vec{r}) \in \mathcal{D}} p(\vec{x}, \vec{m}, \vec{O}^p | \vec{a}) = \sum_{(\vec{x}, \vec{m}, \vec{a}, \vec{r}) \in \mathcal{D}} \log p(\vec{x}, \vec{m}, \vec{O}^p | \vec{a}) \quad (12)$$

This can be maximized by stochastic gradient descent (SGD), which optimizes parametric functions by gradient descent, with the gradient estimated by sampling a batch of data points. To make SGD applicable to our problem, $p(\vec{x}, \vec{m}, \vec{O}^p | \vec{a})$ needs to

be represented by parametric functions, then auto-differentiation tools (e.g., TensorFlow) can be used to calculate its gradient. We can use variational inference [43] to compute this log likelihood. We first introduce the latent variables \vec{z}^w and \vec{a}^p :

$$\log p(\vec{x}, \vec{m}, \vec{O}^p | \vec{a}) = \log \int \int p(\vec{x}, \vec{m}, \vec{O}^p, \vec{z}^w, \vec{a}^p | \vec{a}) d\vec{z}^w d\vec{a}^p \quad (13)$$

Then introduce a variational distribution $q(\vec{z}^w, \vec{a}^p | \vec{x}, \vec{a})$ into (13):

$$\begin{aligned} & \log p(\vec{x}, \vec{m}, \vec{O}^p | \vec{a}) \\ &= \log \int \int p(\vec{x}, \vec{m}, \vec{O}^p, \vec{z}^w, \vec{a}^p | \vec{a}) \frac{q(\vec{z}^w, \vec{a}^p | \vec{x}, \vec{a})}{q(\vec{z}^w, \vec{a}^p | \vec{x}, \vec{a})} d\vec{z}^w d\vec{a}^p \end{aligned} \quad (14)$$

The variational distribution is defined as:

$$\begin{aligned} q(\vec{z}^w, \vec{a}^p | \vec{x}, \vec{a}) \\ = q(\vec{z} | \vec{x}, \vec{a}) \pi(a_{\tau+H} | z_{\tau+H}) \prod_{t=\tau+1}^{\tau+H-1} p(z_{t+1} | z_t, a_t) \pi(a_t | z_t) \end{aligned} \quad (15)$$

where $q(\vec{z} | \vec{x}, \vec{a})$ is the inference of latent states given historical sensor inputs and actions. The rest part of the right hand side represents the trajectory distribution by executing policy $\pi(a_t | z_t)$ with latent state transition $p(z_{t+1} | z_t, a_t)$.

Now eliminate the integration in (14) by introducing expectation, and apply Jensen's inequality we have:

$$\begin{aligned} \log p(\vec{x}, \vec{m}, \vec{O}^p | \vec{a}) &= \log \mathbb{E}_{q(\vec{z}^w, \vec{a}^p | \vec{x}, \vec{a})} \left[\frac{p(\vec{x}, \vec{m}, \vec{O}^p, \vec{z}^w, \vec{a}^p | \vec{a})}{q(\vec{z}^w, \vec{a}^p | \vec{x}, \vec{a})} \right] \\ &\geq \mathbb{E}_{q(\vec{z}^w, \vec{a}^p | \vec{x}, \vec{a})} \left[\log p(\vec{x}, \vec{m}, \vec{O}^p, \vec{z}^w, \vec{a}^p | \vec{a}) \right. \\ &\quad \left. - \log q(\vec{z}^w, \vec{a}^p | \vec{x}, \vec{a}) \right] \\ &= \text{ELBO} \end{aligned} \quad (16)$$

where ELBO stands for evidence lower bound. We can maximize the original log likelihood by maximizing the ELBO. Let's now derive $p(\vec{x}, \vec{m}, \vec{O}^p, \vec{z}^w, \vec{a}^p | \vec{a})$ by probability factorization according to the PGM in Fig.4:

$$\begin{aligned} & p(\vec{x}, \vec{m}, \vec{O}^p, \vec{z}^w, \vec{a}^p | \vec{a}) \\ &= p(\vec{x}, \vec{m}, \vec{O}^p, z_{\tau+2:\tau+H}, \vec{a}^p | \vec{z}, \vec{a}) p(\vec{z} | \vec{a}) \\ &= p(\vec{x} | \vec{z}) p(\vec{m} | \vec{z}) p(\vec{O}^p, z_{\tau+2:\tau+H}, \vec{a}^p | z_{\tau+1}) p(\vec{z} | \vec{a}) \\ &= p(\vec{x} | \vec{z}) p(\vec{m} | \vec{z}) \frac{p(\vec{O}^p, \vec{z}^p, \vec{a}^p)}{p(z_{\tau+1})} p(\vec{z} | \vec{a}) \end{aligned} \quad (17)$$

According to the soft optimality assumption:

$$\begin{aligned} & p(\vec{O}^p, \vec{z}^p, \vec{a}^p) = p(\vec{z}^p, \vec{a}^p) p(\vec{O}^p | \vec{z}^p, \vec{a}^p) \\ &= p(\vec{a}^p) p(z_{\tau+1}) \prod_{t=\tau+1}^{\tau+H-1} p(z_{t+1} | z_t, a_t) \exp \left(\sum_{t=\tau+1}^{\tau+H} r(z_t, a_t) \right) \end{aligned} \quad (18)$$

We thus have:

$$\begin{aligned} p(\vec{x}, \vec{m}, \vec{O}^p, \vec{z}^w, \vec{a}^p | \vec{a}) &= p(\vec{x} | \vec{z}) p(\vec{m} | \vec{z}) p(\vec{a}^p) \\ &\quad \prod_{t=\tau+1}^{\tau+H-1} p(z_{t+1} | z_t, a_t) \exp \left(\sum_{t=\tau+1}^{\tau+H} r(z_t, a_t) \right) p(\vec{z} | \vec{a}) \end{aligned} \quad (19)$$

Substituting the variational distribution (15) into (16), we have:

$$\begin{aligned} \text{ELBO} &= \mathbb{E}_{q(\vec{z}^w, \vec{a}^p | \vec{x}, \vec{a})} [\log p(\vec{x} | \vec{z}) + \log p(\vec{m} | \vec{z}) + \log p(\vec{z} | \vec{a}) \\ &\quad + \log \prod_{t=\tau+1}^{\tau+H-1} p(z_{t+1} | z_t, a_t) + \sum_{t=\tau+1}^{\tau+H} r(z_t, a_t) - \log q(\vec{z} | \vec{x}, \vec{a}) \\ &\quad - \log \prod_{t=\tau+1}^{\tau+H} \pi(a_t | z_t) - \log \prod_{t=\tau+1}^{\tau+H-1} p(z_{t+1} | z_t, a_t) + \log p(\vec{a}^p)] \end{aligned} \quad (20)$$

Notice the cancellations in (20), we have:

$$\begin{aligned} \text{ELBO} &= \mathbb{E}_{q(\vec{z}^w, \vec{a}^p | \vec{x}, \vec{a})} [\log p(\vec{x} | \vec{z}) + \log p(\vec{m} | \vec{z}) + \log p(\vec{z} | \vec{a}) \\ &\quad - \log q(\vec{z} | \vec{x}, \vec{a})] \\ &\quad + \mathbb{E}_{q(\vec{z}^w, \vec{a}^p | \vec{x}, \vec{a})} \left[\sum_{t=\tau+1}^{\tau+H} (r(z_t, a_t) - \log \pi(a_t | z_t) + \log p(a_t)) \right] \end{aligned} \quad (21)$$

The first part of the right hand side of (21) corresponds to learning the environment model, while the second part corresponds to learning the driving policy, we will derive the details of the two parts in V-B and V-C, respectively.

B. Environment Model Learning

The environment model can be learned via optimizing the first part of (21):

$$\mathbb{E}_{q(\vec{z} | \vec{x}, \vec{a})} [\log p(\vec{x} | \vec{z}) + \log p(\vec{m} | \vec{z}) + \log p(\vec{z} | \vec{a}) - \log q(\vec{z} | \vec{x}, \vec{a})] \quad (22)$$

where we replace $\mathbb{E}_{q(\vec{z}^w, \vec{a}^p | \vec{x}, \vec{a})}$ with $\mathbb{E}_{q(\vec{z} | \vec{x}, \vec{a})}$ because this part of ELBO is only related to $z_{1:\tau+1}$. Now let's further derive the components in (22) by unfolding them with time. Considering the conditional dependence of PGM in Fig.4. The generative models can be unfolded as:

$$\log p(\vec{x} | \vec{z}) = \log \prod_{t=1}^{\tau+1} p(x_t | z_t) = \sum_{t=1}^{\tau+1} \log p(x_t | z_t) \quad (23)$$

$$\log p(\vec{m} | \vec{z}) = \log \prod_{t=1}^{\tau+1} p(m_t | z_t) = \sum_{t=1}^{\tau+1} \log p(m_t | z_t)$$

The prior model can be unfolded using the latent state transition function:

$$\begin{aligned} \log p(\vec{z} | \vec{a}) &= \log \left[p(z_1) \prod_{t=1}^{\tau} p(z_{t+1} | z_t, a_t) \right] \\ &= \log p(z_1) + \sum_{t=1}^{\tau} \log p(z_{t+1} | z_t, a_t) \end{aligned} \quad (24)$$

The posterior inference model can be unfolded as:

$$\begin{aligned} \log q(\vec{z}|\vec{x}, \vec{a}) &= \log \left[q(z_1|\vec{x}, \vec{a}) \prod_{t=1}^{\tau} q(z_{t+1}|z_t, \vec{x}, \vec{a}) \right] \\ &\approx \log \left[q(z_1|x_1) \prod_{t=1}^{\tau} q(z_{t+1}|z_t, x_{t+1}, a_t) \right] \\ &= \log q(z_1|x_1) + \sum_{t=1}^{\tau} \log q(z_{t+1}|z_t, x_{t+1}, a_t) \end{aligned} \quad (25)$$

Note here we approximate $q(\vec{z}|\vec{x}, \vec{a})$ and $q(z_{t+1}|z_t, \vec{x}, \vec{a})$ with $q(z_1|x_1)$ and $q(z_{t+1}|z_t, x_{t+1}, a_t)$ for simplicity. If we want to obtain the exact accurate values, bi-directional recurrent neural networks should be used to obtain the posterior probabilities conditioned on the whole trajectory sequence (\vec{x}, \vec{a}) [30].

We can now unfold (22) with time:

$$\begin{aligned} \mathbb{E}_{q(\vec{z}|\vec{x}, \vec{a})} [\log p(\vec{x}|\vec{z}) + \log p(\vec{m}|\vec{z}) + \log p(\vec{z}|\vec{a}) - \log q(\vec{z}|\vec{x}, \vec{a})] \\ \approx \mathbb{E}_{q(\vec{z}|\vec{x}, \vec{a})} \left[\sum_{t=1}^{\tau+1} \log p(x_t|z_t) + \sum_{t=1}^{\tau+1} \log p(m_t|z_t) \right. \\ - D_{KL}(q(z_1|x_1) || p(z_1)) \\ \left. - \sum_{t=1}^{\tau+1} D_{KL}(q(z_{t+1}|z_t, x_{t+1}, a_t) || p(z_{t+1}|z_t, a_t)) \right] \end{aligned} \quad (26)$$

C. Driving Policy Learning

The driving policy can be learned via optimizing the second part of (21):

$$\begin{aligned} \max_{q(\vec{z}^p, \vec{a}^p|\vec{x}, \vec{a})} \sum_{t=\tau+1}^{\tau+H} [r(z_t, a_t) - \log \pi_\phi(a_t|z_t) + \log p(a_t)] \\ = \mathbb{E}_{\substack{z_{\tau+1} \sim p(z_{\tau+1}|\vec{x}, \vec{a}) \\ a_t \sim \pi_\phi(a_t|z_t) \\ z_{\tau+1} \sim p(z_{\tau+1}|z_t, a_t)}} \sum_{t=\tau+1}^{\tau+H} [r(z_t, a_t) - \log \pi_\phi(a_t|z_t)] \end{aligned} \quad (27)$$

where $\log p(a_t)$ is ignored since we assume uniform action prior. The optimization problem (27) then becomes a standard MaxEnt RL problem.

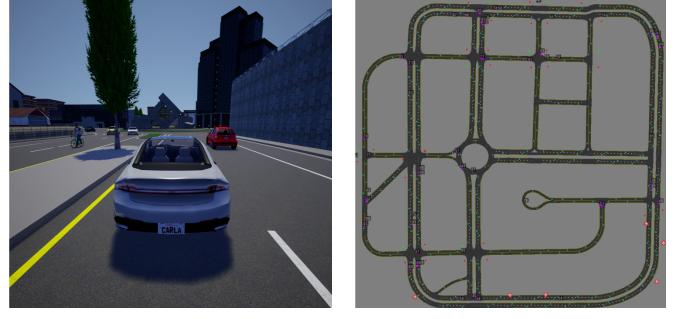
We use soft actor-critic (SAC) [23] to solve this MaxEnt RL problem. SAC is a function approximation version of the soft policy iteration (SPI). SPI is an extension of the standard policy iteration to the maximum entropy case, which is to iteratively apply the soft policy evaluation:

$$\begin{aligned} \mathcal{T}^\pi Q(z_t, a_t) &= r(z_t, a_t) \\ &+ \gamma \mathbb{E}_{z_{t+1} \sim p} \left[\mathbb{E}_{a_{t+1} \sim \pi} [Q(z_{t+1}, a_{t+1}) - \log \pi(a_{t+1}|z_{t+1})] \right] \end{aligned} \quad (28)$$

and the soft policy improvement:

$$\pi_{\text{new}} = \underset{\pi'}{\operatorname{argmin}} D_{KL} \left(\pi'(\cdot|z_t) \middle\| \frac{\exp(Q^{\pi_{\text{old}}}(z_t, \cdot))}{Z^{\pi_{\text{old}}}(z_t)} \right) \quad (29)$$

where $Z^{\pi_{\text{old}}}(z_t)$ is the normalization term.



(a) Sample view of CARLA simulator (b) Map layout of the simulated city

Fig. 7: Simulation environment

The function approximation implementation is to optimize the loss functions that address the soft policy evaluation and soft policy improvement. The loss functions are the Bellman residual in (28):

$$J_Q = \mathbb{E}_{z_\tau \sim q(\vec{z}|\vec{x}, \vec{a})} \left[\frac{1}{2} (Q(z_\tau, a_\tau) - \bar{Q}(z_\tau, a_\tau)) \right] \quad (30)$$

and the KL divergence in (29):

$$J_\pi = \mathbb{E}_{\substack{z_{\tau+1} \sim q(\vec{z}|\vec{x}, \vec{a}) \\ a_{\tau+1} \sim \pi(a_{\tau+1}|z_{\tau+1})}} [\log \pi(a_{\tau+1}|z_{\tau+1}) - Q(z_{\tau+1}, a_{\tau+1})] \quad (31)$$

Note

$$\begin{aligned} \hat{Q}(z_\tau, a_\tau) &= r_\tau \\ &+ \gamma \mathbb{E}_{\substack{z_{\tau+1} \sim q(\vec{z}|\vec{x}, \vec{a}) \\ a_{\tau+1} \sim \pi(a_{\tau+1}|z_{\tau+1})}} [\bar{Q}(z_{\tau+1}, a_{\tau+1}) - \log \pi(a_{\tau+1}|z_{\tau+1})] \end{aligned} \quad (32)$$

where \bar{Q} is a delayed Q network.

Thus, the joint learning algorithm becomes to use SGD to maximize the model learning part of ELBO in (26) and minimize J_Q in (30) and J_π in (31).

VI. EXPERIMENTS

A. Simulation Setup

We train and evaluate our proposed method on CARLA simulator [44]. CARLA is a high-definition open-source simulation platform for autonomous driving research. It simulates not only the driving environment and vehicle dynamics, but also the raw sensor data inputs such as camera RGB image and lidar point cloud. Fig.7 (a) shows a sample view of the driving simulation environment we use.

Fig.7 (b) shows the map layout of the virtual town in CARLA we use for training. It includes various urban scenarios such as intersections and roundabouts. The range of the map is $400m \times 400m$, with about $6km$ total length of roads. 100 vehicles are running autonomously in the virtual town to simulate a multi-agent environment. The vehicles will randomly choose a direction at intersections, then follow the route, while slowing down for front vehicles and stopping when the front traffic light becomes red.

B. Implementation Details

1) **Reward Function:** We use the following reward function in our experiments:

$$r = 200 r_{\text{collision}} + v_{\text{lon}} + 10 r_{\text{fast}} + r_{\text{out}} - 5 \alpha^2 + 0.2 r_{\text{lat}} - 0.1 \quad (33)$$

where $r_{\text{collision}}$ is the reward related to collision, which is set to -1 if the ego vehicle collides and 0 otherwise. v_{lon} is the longitudinal speed of the ego vehicle. r_{fast} is the reward related to running too fast, which is set to -1 if it exceeds the desired speed (8 m/s here) and 0 otherwise. r_{out} is set to -1 if the ego vehicle runs out of lane, and 0 otherwise. α is the steering angle of ego vehicle in rad. r_{lat} is the reward related to lateral acceleration, which is calculated by $r_{\text{lat}} = -|\alpha|v_{\text{lon}}^2$. The last constant term is added to prevent the ego vehicle from standing still.

2) **Network Architecture:** The parametrized neural networks in our method includes the generative models $p(x_t|z_t)$ and $p(m_t|z_t)$, the latent dynamics $p(z_{t+1}|z_t, a_t)$, the filtering model $q(z_{t+1}|z_t, x_{t+1}, a_t)$ and $q(z_1|x_1)$, the Q network $Q(z_t, a_t)$, and the policy network $\pi(a_t|z_t)$. Here we followed the two-layer hierarchical latent space structure as in [34], such that $z_t^1 \in \mathbb{R}^{32}$ and $z_t^2 \in \mathbb{R}^{256}$. Each sensor input size and mask size is $64 \times 64 \times 3$, such that $x_t, m_t \in [0, 255]^{64 \times 64 \times 3}$.

$p(x_t|z_t)$ and $p(m_t|z_t)$ both consist of 5 deconvolutional layers ((256, 4, 1), (128, 3, 2), (64, 3, 2), (32, 3, 2), and (3, 5, 2), with each tuple means (filters, kernel size, strides)). $p(z_{t+1}|z_t, a_t)$ consists of two fully connected layers with hidden units number 256, followed by a Gaussian output layer. $q(z_{t+1}|z_t, x_{t+1}, a_t)$ and $q(z_1|x_1)$ both consist of 5 convolutional layers ((32, 5, 2), (64, 3, 2), (128, 3, 2), (256, 3, 2), and (256, 4, 1), with each tuple means (filters, kernel size, strides)) to first encode the sensor inputs x_t into features of size 256. Then two fully connected layers with hidden units number 256 are followed, with a Gaussian output layer. $Q(z_t, a_t)$ consists of two fully connected layers with hidden units number 256, followed by a linear output layer. $\pi(a_t|z_t)$ consists of two fully connected layers with hidden units number 256, followed by a Gaussian layer, and a tanh bijector.

3) **Training Details:** At each new episode, the ego vehicle is placed in a random feasible start position in the virtual town. Other vehicles are also located to new random positions. The maximum episode length is 500, the time interval for adjacent frames is 0.1 second. We use a frame skip of 4 for temporal extension, which means the action is fixed for every 4 environment steps.

The hyperparameters are the same with [23]. One gradient step is applied per each skipped frame environment step (e.g., in our case it is one gradient step per every 4 environment steps). The Q network and policy are trained with batch size 256 and learning rate 0.0003. The sequential latent model is trained with batch size 32 and learning rate 0.0001. The length of sequential model used for training is $\tau = 10$. The discount factor $\gamma = 0.99$.

VII. EVALUATION RESULTS

During evaluation, we use the same stochastic policy that is used during training. 10 episodes are performed at each

evaluation step and the average return is calculated. Same with the training phase, all vehicles are randomly relocated in the whole map for each new episode. No frame skip is performed at the evaluation phase.

A. Variants of Proposed Method

Besides our proposed method, we also trained and evaluated other two variants of the method, and then compare the three methods:

1) **Sensor Inputs and Mask (Proposed):** This is our proposed, which takes the sensor inputs and generate the mask.

2) **Sensor Inputs Only:** Here we consider the case that no mask is provided. So only the camera and lidar sensor inputs are inputted and reconstructed. The model learning part is then trained in an unsupervised way without mask labels.

3) **Mask Input Only:** Assume we already have a good perception and localization system that can accurately detect vehicles, localize ego vehicle, and provide accurate road condition information, we can then directly generate the mask and use it as our input. In this case, only the mask is inputted and reconstructed. Note this method can be regarded as an extension of our paper [45], which uses offline data to train a non-sequential variational auto-encoder (VAE) to learn the latent state, and then apply SAC on the latent space.

B. Baseline RL Algorithms

We compare our proposed methods with the following state-of-the-art model-free RL algorithms:

1) **DQN [12]:** DQN is the first deep reinforcement learning method proposed by DeepMind. It uses deep neural network approximate the Q value and uses deep learning to approximate the bellman operation.

2) **DDPG [9]:** DDPG is an actor-critic algorithm on the deterministic policy gradient which is able to handle continuous action spaces. Besides a deep Q network that is approximated with bellman operation, there is a policy network which is optimized with policy gradient.

3) **TD3 [20]:** For value-based and actor-critic based RL methods such as DQN and DDPG, function approximation errors will lead to overestimated value estimates and sub-optimal policies. TD3 improves the function approximation errors by taking the minimum value between a pair of critics and delaying policy updates.

4) **SAC [23]:** SAC is a fundamentally different RL algorithm compared to the above methods, which is within the MaxEnt RL framework. We have briefly introduced the algorithm in Section V-C.

To make a fair comparison, we use the same encoding networks with our proposed method for those baseline algorithms, but now without decoders. We use recurrent neural networks (RNN), since our proposed method also considers time sequence. The type of RNN we use is long short term memory (LSTM), with LSTM size of 40 and output size of 100.

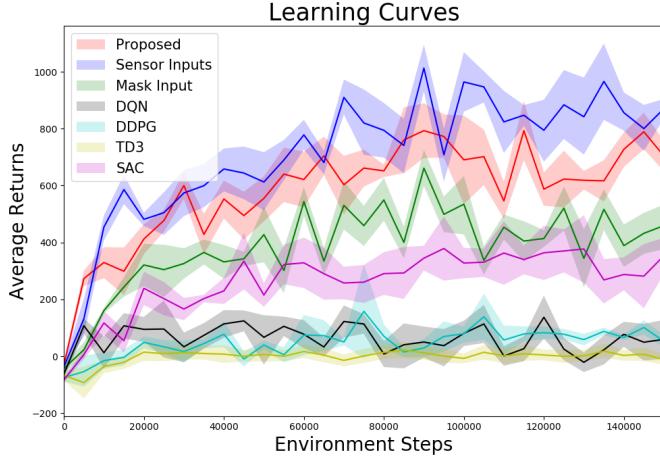


Fig. 8: Comparison of learning curves with baseline RL algorithms. Average returns calculated with 5 trials, each with 10 episodes. Shaded area indicates standard deviation.

C. Evaluation Results

The performance comparison is shown in Fig.8. We draw the learning curves composed of average returns (the average discounted cumulative rewards of multiple testing episodes) vs environment steps. We can see that all variants of our proposed method are significantly better than the baselines. Actually, the baselines almost do not work at all. Note that our baselines implemented here are already better than existing RL methods for autonomous driving, which mostly only take front-view camera image, do not consider time sequence, and do not use some state-of-the-art RL algorithms such as SAC and TD3.

VIII. INTERPRETABILITY

Besides the performance, our proposed method also has significant advantage in terms of interpretability by decoding a semantic mask from the latent state. However, since the baseline RL algorithms do not have a latent space, they are not able to provide an interpretable semantic mask. In this section, we will explain how our method is able to interpret how the autonomous car understand the environment.

A. Detection & Localization Functionality

It is essential to localize the autonomous car and understand the road conditions around the car. Traditionally, this is enabled by a separate localization & mapping system, which requires the collection of an HD map and designing of complex SLAM [46] algorithms. However, our method is able to obtain all those information within the end-to-end RL training process, without storing any HD maps or manually designing any localization algorithms.

On the other hand, Object detection is of fundamental importance, as failing to detect road participants and obstacles might lead to serious incidents. The environment model obtained in our method also has the ability to detect surrounding vehicles by fusing camera and lidar sensor inputs.

Fig.9 shows random sampled frames of the sensor inputs, ground truth masks, and reconstructions during running with the learned model and policy. For each sample, the first row contains the raw sensor inputs and ground truth mask (left to right: camera, lidar, bird-view mask). The second row contains the corresponding reconstructed images from the latent state. Note here only the raw sensor inputs are observed, the ground truth bird-view image is displayed only for comparison. From the reconstructed bird-view mask, we can see that it can accurately locate the ego car and decode the map information (e.g, drivable areas and road markings), even though there is no direct information from the raw sensor inputs indicating the ego car is in an intersection. We can also see that our model can accurately detect the surrounding vehicles (green boxes) given raw camera and lidar observations.

B. Quantified Evaluation

We also quantify the interpretability of our method by calculating the average pixel difference between the decoded masks and the ground truth masks during massive simulation tests in the virtual city. The metric is defined as:

$$e = \frac{1}{N} \sum_{i=1}^N \frac{\text{sum}(|\hat{m}_i - m_i|)}{W \times H \times C} \quad (34)$$

where \hat{m}_i is the predicted mask, m_i is the ground truth mask, N is the number of samples we evaluate. W , H and C are the size of the mask image. In our case, $W = H = 64$, $C = 3$. Values in m_i and \hat{m}_i are RGB values scaled to $[0, 1]$. After evaluating $N = 10^4$ frames in the simulation, we got the average pixel difference to be $e = 0.032$, which indicates high accuracy when decoding the birdeye semantic mask images.

C. Failure Cases Interpretation

Although we can learn significantly better driving policy than baseline RL methods as shown in Section VII-C, we can still observe some failure cases such as collisions with surrounding vehicles during testing. Our method can help interpret why the agent fails. Fig.10 shows examples of our failure cases interpretation. Same as before, the first row shows the sensor inputs and ground truth masks, while the second row shows the reconstructed sensor inputs and masks. The left example shows a case where the agent collides with another vehicle in an intersection. From the reconstructed mask we can see the agent does not recognize the surrounding vehicle. This might be caused by the low resolution of the sensor inputs, as we can hardly see the vehicle in the camera image. The right example shows a case where the agent collides with a vehicle occupying part of its lane. From the reconstructed mask we can see that although the agent recognize the vehicle, it mistakenly localize it in its own lane. This might because this is a very rare situation and almost all training data is composed of vehicles running in their own lanes.

IX. CONCLUSIONS

In this paper, we proposed an interpretable end-to-end reinforcement learning algorithm for autonomous driving in urban

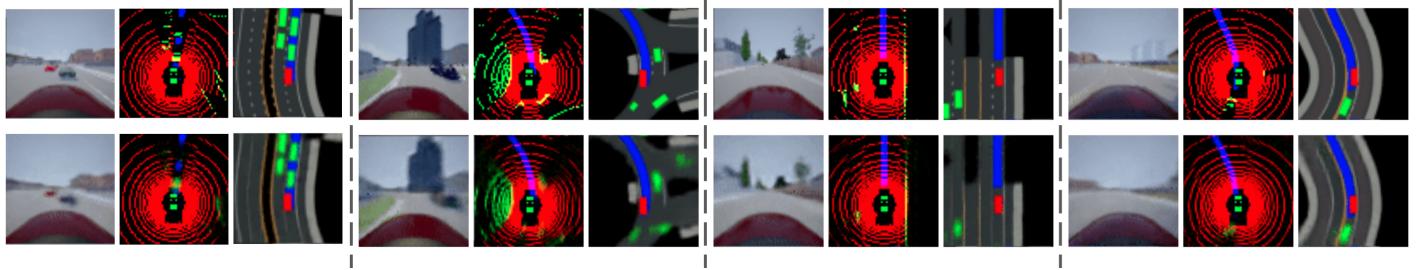


Fig. 9: Randomly sampled frames to illustrate the interpretability of our method. For each sample, left to right: camera, lidar, bird-view image. First row: original sensor inputs and ground truth mask. Second row: reconstructed images. Only the raw camera and lidar images are observed.

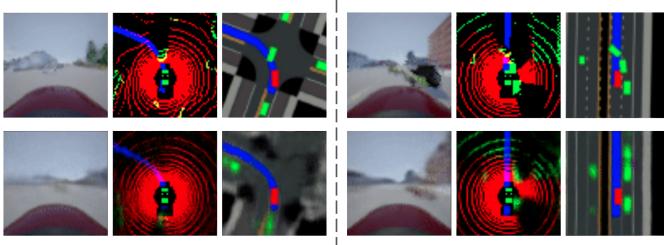


Fig. 10: Examples of failure cases interpretation.

driving scenarios. The driving policy was learned jointly with a sequential environment model using latent state space. The learned driving policy took camera and lidar images as input, and generated control commands to navigate the autonomous car through urban driving scenarios. The learned environment model provided an interpretable explanation of how the autonomous car understands the driving situation by generating a bird-view semantic mask. The mask was enforced to connect with a certain sub-module in traditional autonomous driving framework, thus providing an explanation of how the learned policy behave in response to current-time environment. The method was implemented and evaluated in CARLA simulator, which was shown to have significantly better performance over classic RL baselines.

Although our framework is able to provide interpretable explanations about how the model understand the environment, it does not provide any intuition about how it makes the decisions, because the driving policy is obtained in a model-free way. In the future, model-based method will be investigated within in this framework to further improve the performance and interpretability.

REFERENCES

- [1] S. Thrun, M. Montemerlo, H. Dahlkamp, D. Stavens, A. Aron, J. Diebel, P. Fong, J. Gale, M. Halpenny, G. Hoffmann, *et al.*, “Stanley: The robot that won the darpa grand challenge,” *Journal of field Robotics*, vol. 23, no. 9, pp. 661–692, 2006.
- [2] C. Urmson, J. Anhalt, D. Bagnell, C. Baker, R. Bittner, M. Clark, J. Dolan, D. Duggins, T. Galatali, C. Geyer, *et al.*, “Autonomous driving in urban environments: Boss and the urban challenge,” *Journal of Field Robotics*, vol. 25, no. 8, pp. 425–466, 2008.
- [3] M. Bojarski, D. Del Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang, *et al.*, “End to end learning for self-driving cars,” *arXiv preprint arXiv:1604.07316*, 2016.
- [4] F. Codevilla, M. Miiller, A. López, V. Koltun, and A. Dosovitskiy, “End-to-end driving via conditional imitation learning,” in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1–9, IEEE, 2018.
- [5] H. Xu, Y. Gao, F. Yu, and T. Darrell, “End-to-end learning of driving models from large-scale video datasets,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2174–2182, 2017.
- [6] M. Bansal, A. Krizhevsky, and A. Ogale, “Chauffeurnet: Learning to drive by imitating the best and synthesizing the worst,” *arXiv preprint arXiv:1812.03079*, 2018.
- [7] J. Chen, B. Yuan, and M. Tomizuka, “Deep imitation learning for autonomous driving in generic urban scenarios with enhanced safety,” *arXiv preprint arXiv:1903.00640*, 2019.
- [8] P. Wolf, C. Hubschneider, M. Weber, A. Bauer, J. Härtl, F. Dürr, and J. M. Zöllner, “Learning how to drive in a real world simulation with deep q-networks,” in *2017 IEEE Intelligent Vehicles Symposium (IV)*, pp. 244–250, IEEE, 2017.
- [9] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, “Continuous control with deep reinforcement learning,” *arXiv preprint arXiv:1509.02971*, 2015.
- [10] A. Kendall, J. Hawke, D. Janz, P. Mazur, D. Reda, J.-M. Allen, V.-D. Lam, A. Bewley, and A. Shah, “Learning to drive in a day,” in *2019 International Conference on Robotics and Automation (ICRA)*, pp. 8248–8254, IEEE, 2019.
- [11] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, “Playing atari with deep reinforcement learning,” *arXiv preprint arXiv:1312.5602*, 2013.
- [12] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, *et al.*, “Human-level control through deep reinforcement learning,” *Nature*, vol. 518, no. 7540, p. 529, 2015.
- [13] O. Vinyals, I. Babuschkin, J. Chung, M. Mathieu, M. Jaderberg, W. M. Czarnecki, A. Dudzik, A. Huang, P. Georgiev, R. Powell, *et al.*, “AlphaStar: Mastering the real-time strategy game starcraft ii,” *DeepMind Blog*, 2019.
- [14] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, *et al.*, “Mastering the game of go with deep neural networks and tree search,” *nature*, vol. 529, no. 7587, p. 484, 2016.
- [15] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, *et al.*, “Mastering the game of go without human knowledge,” *nature*, vol. 550, no. 7676, pp. 354–359, 2017.
- [16] S. Levine, C. Finn, T. Darrell, and P. Abbeel, “End-to-end training of deep visuomotor policies,” *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 1334–1373, 2016.
- [17] D. Kalashnikov, A. Irpan, P. Pastor, J. Ibarz, A. Herzog, E. Jang, D. Quillen, E. Holly, M. Kalakrishnan, V. Vanhoucke, *et al.*, “Qt-opt: Scalable deep reinforcement learning for vision-based robotic manipulation,” *arXiv preprint arXiv:1806.10293*, 2018.
- [18] H. Van Hasselt, A. Guez, and D. Silver, “Deep reinforcement learning with double q-learning,” in *Thirtieth AAAI conference on artificial intelligence*, 2016.
- [19] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, “Asynchronous methods for deep reinforcement learning,” in *International conference on machine learning*, pp. 1928–1937, 2016.

- [20] S. Fujimoto, H. van Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," *arXiv preprint arXiv:1802.09477*, 2018.
- [21] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, "Trust region policy optimization," in *International conference on machine learning*, pp. 1889–1897, 2015.
- [22] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.
- [23] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," *arXiv preprint arXiv:1801.01290*, 2018.
- [24] T. Haarnoja, A. Zhou, K. Hartikainen, G. Tucker, S. Ha, J. Tan, V. Kumar, H. Zhu, A. Gupta, P. Abbeel, et al., "Soft actor-critic algorithms and applications," *arXiv preprint arXiv:1812.05905*, 2018.
- [25] J. Chen, Z. Wang, and M. Tomizuka, "Deep hierarchical reinforcement learning for autonomous driving with distinct behaviors," in *2018 IEEE Intelligent Vehicles Symposium (IV)*, pp. 1239–1244, IEEE, 2018.
- [26] M. Bojarski, P. Yeres, A. Choromanska, K. Choromanski, B. Firner, L. Jackel, and U. Muller, "Explaining how a deep neural network trained with end-to-end learning steers a car," *arXiv preprint arXiv:1704.07911*, 2017.
- [27] J. Kim and J. Canny, "Interpretable learning for self-driving cars by visualizing causal attention," in *Proceedings of the IEEE international conference on computer vision*, pp. 2942–2950, 2017.
- [28] A. Sauer, N. Savinov, and A. Geiger, "Conditional affordance learning for driving in urban environments," *arXiv preprint arXiv:1806.06498*, 2018.
- [29] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.
- [30] R. G. Krishnan, U. Shalit, and D. Sontag, "Deep kalman filters," *arXiv preprint arXiv:1511.05121*, 2015.
- [31] M. Karl, M. Soelch, J. Bayer, and P. Van der Smagt, "Deep variational bayes filters: Unsupervised learning of state space models from raw data," *arXiv preprint arXiv:1605.06432*, 2016.
- [32] M. Fraccaro, S. Kamronn, U. Paquet, and O. Winther, "A disentangled recognition and nonlinear dynamics model for unsupervised learning," in *Advances in Neural Information Processing Systems*, pp. 3601–3610, 2017.
- [33] D. Hafner, T. Lillicrap, I. Fischer, R. Villegas, D. Ha, H. Lee, and J. Davidson, "Learning latent dynamics for planning from pixels," *arXiv preprint arXiv:1811.04551*, 2018.
- [34] A. X. Lee, A. Nagabandi, P. Abbeel, and S. Levine, "Stochastic latent actor-critic: Deep reinforcement learning with a latent variable model," *arXiv preprint arXiv:1907.00953*, 2019.
- [35] S. Levine, "Reinforcement learning and control as probabilistic inference: Tutorial and review," *arXiv preprint arXiv:1805.00909*, 2018.
- [36] K. Rawlik, M. Toussaint, and S. Vijayakumar, "On stochastic optimal control and reinforcement learning by approximate inference," in *Twenty-third international joint conference on artificial intelligence*, 2013.
- [37] B. D. Ziebart, "Modeling purposeful adaptive behavior with the principle of maximum causal entropy," 2010.
- [38] D. Ha and J. Schmidhuber, "World models," *arXiv preprint arXiv:1803.10122*, 2018.
- [39] M. Okada, N. Kosaka, and T. Taniguchi, "Planet of the bayesians: Reconsidering and improving deep planning network by incorporating bayesian inference," *arXiv preprint arXiv:2003.00370*, 2020.
- [40] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.
- [41] T. Haarnoja, H. Tang, P. Abbeel, and S. Levine, "Reinforcement learning with deep energy-based policies," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1352–1361, JMLR. org, 2017.
- [42] B. Eysenbach and S. Levine, "If maxent rl is the answer, what is the question?," *arXiv preprint arXiv:1910.01913*, 2019.
- [43] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.
- [44] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "Carla: An open urban driving simulator," *arXiv preprint arXiv:1711.03938*, 2017.
- [45] J. Chen, B. Yuan, and M. Tomizuka, "Model-free deep reinforcement learning for urban autonomous driving," in *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, pp. 2765–2771, IEEE, 2019.
- [46] E. Yurtsever, J. Lambert, A. Carballo, and K. Takeda, "A survey of autonomous driving: Common practices and emerging technologies," *arXiv preprint arXiv:1906.05113*, 2019.



Jianyu Chen received the B.E. degree from Tsinghua University, China, in 2015. He received the Ph.D. degree working with Prof. Masayoshi Tomizuka at the University of California, Berkeley in 2020. He is working at an intersection of robotics and machine learning to build structured learning systems for intelligent robots which can learn safe, interpretable and scalable perceptual-control policies. Applications of his work mainly focus on autonomous driving. His research interests include motion planning, optimal control, reinforcement learning, imitation learning, and representation learning.



Shengbo Eben Li received the M.S. and Ph.D. degrees from Tsinghua University in 2006 and 2009. Before joining Tsinghua University, he has worked at Stanford University, University of Michigan, and UC Berkeley. He is now leading Intelligent Driving Lab (iDLab) at Tsinghua University. His active research interests include intelligent vehicles and driver assistance, reinforcement learning and optimal control, distributed control and estimation, etc. He is the author of over 100 peer-reviewed journal/conference papers, and the co-inventor of over 30 patents. Dr. Li was the recipient of Best Paper Award in 2014 IEEE ITS, Best Paper Award in 14th Asian ITS, National Award for Technological Invention of China (2013), Excellent Young Scholar of NSF China (2016), Young Professorship of Changjiang Scholar Program (2016), Tsinghua University Excellent Professorship Award (2017), National Award for Progress in Science and Technology of China (2018), Distinguished Young Scholar of Beijing NSF (2018), etc. He also serves as Board of Governor of IEEE ITS Society, AEs of IEEE ITSM, IEEE Trans ITS, etc.



Masayoshi Tomizuka (M'86-SM'95-F'97-LF'17) received his Ph. D. degree in Mechanical Engineering from MIT in February 1974. In 1974, he joined the faculty of the Department of Mechanical Engineering at the University of California at Berkeley, where he currently holds the Cheryl and John Neerhout, Jr., Distinguished Professorship Chair. His current research interests are optimal and adaptive control, digital control, signal processing, motion control, and control problems related to robotics, precision motion control and vehicles. He served as Program Director of the Dynamic Systems and Control Program of the Civil and Mechanical Systems Division of NSF (2002- 2004). He served as Technical Editor of the ASME Journal of Dynamic Systems, Measurement and Control, J-DSMC (1988-93), and Editor-in-Chief of the IEEE/ASME Transactions on Mechatronics (1997-99). Prof. Tomizuka is a Fellow of the ASME, IEEE and IFAC. He is the recipient of the Charles Russ Richards Memorial Award (ASME, 1997), the Rufus Oldenburger Medal (ASME, 2002) and the John R. Ragazzini Award (2006).