Proceedings of Machine Learning Research vol 242:1187–1201, 2024

State-Wise Safe Reinforcement Learning with Pixel Observations

Simon Sinong Zhan SINONGZHAN2028@U.NORTHWESTERN.EDU
Northwestern University, Evanston, IL 60208, USA

Yixuan Wang YIXUANWANG2024@U.NORTHWESTERN.EDU
Northwestern University, Evanston, IL 60208, USA

Qingyuan Wu QINGYUAN.WU@LIVERPOOL.AC.UK
The University Of Liverpool, Liverpool, L69 3BX, UK

Ruochen Jiao RUOCHENJIAO2024@U.NORTHWESTERN.EDU
Northwestern University, Evanston, IL 60208, USA

Chao Huang CHAO.HUANG@SOTON.AC.UK
The University Of Southampton, Southampton, SO17 1BJ, UK

Qi Zhu QZHU@NORTHWESTERN.EDU
Northwestern University, Evanston, IL 60208, USA

Abstract

In the context of safe exploration, Reinforcement Learning (RL) has long grappled with the challenges of balancing the tradeoff between maximizing rewards and minimizing safety violations, particularly in complex environments with contact-rich or non-smooth dynamics, and when dealing with high-dimensional pixel observations. Furthermore, incorporating state-wise safety constraints in the exploration and learning process, where the agent must avoid unsafe regions without prior knowledge, adds another layer of complexity. In this paper, we propose a novel pixel-observation safe RL algorithm that efficiently encodes state-wise safety constraints with unknown hazard regions through a newly introduced latent barrier-like function learning mechanism. As a joint learning framework, our approach begins by constructing a latent dynamics model with low-dimensional latent spaces derived from pixel observations. We then build and learn a latent barrier-like function on top of the latent dynamics and conduct policy optimization simultaneously, thereby improving both safety and the total expected return. Experimental evaluations on the safety-gym benchmark suite demonstrate that our proposed method significantly reduces safety violations throughout the training process, and demonstrates faster safety convergence compared to existing methods while achieving competitive results in reward return. Source code: `https://github.com/SimonZhan-code/Step-Wise_SafeRL_Pixel`.

Keywords: State-wise Safety, Safe Model-based RL, High-dimensional Observations

1. Introduction

Reinforcement Learning (RL) has demonstrated promising achievements in addressing control problems across diverse domains including robotics (Zhao et al., 2020), games (Silver et al., 2016), buildings (Xu et al., 2022; Wei et al., 2017) and various cyber-physical systems (Yu et al., 2021; Liu et al., 2020; Wang et al., 2020; Li et al., 2017). Despite its potential, the occurrence of state-wise safety violations during the learning exploration phases has restrained industries from integrating RL methods into safety-critical applications such as traffic control (Wei et al., 2018), autonomous driving (Kiran et al., 2021), and power grid (Duan et al., 2019).

Conventional safe RL methods are based on the Constrained Markov Decision Process (CMDP) paradigm (Altman, 2021), which encodes the safety constraints through a cost function of safety violation and reduces the exploration space to where the trajectory-level discounted cumulative expected cost below a predefined threshold. However, a fundamental issue arises from the soft nature of the safety constraints in CMDP, which can hardly capture and enforce stringent reachability-based state-wise safety constraints (Wang et al., 2023b). On the other hand, the state-of-the-art theoretical control approaches such as barrier theory (Ames et al., 2019), contraction theory (Tsukamoto et al.), and reachability analysis (Bansal et al., 2017; Xue et al., 2023; Wang et al., 2024) have their advantages in effectively encoding and optimizing state-wise safety. There are attempts to combine the aforementioned methods with model-based RL techniques (Choi et al., 2020; Dawson et al., 2022b; Wang et al., 2023a,b) to train an RL policy with safety guarantee. Nonetheless, those control-theoretical RL methods encounter significant challenges and limitations when operating in unknown environments with pixel observations (Zhu et al., 2020). Overall, we summarize the challenges when dealing with pixel-observation state-wise safe RL problems as follows.

- **Challenge 1:** The existing CMDP problem is too soft to encode the state-wise safety constraint.
- **Challenge 2:** Control-theoretical approaches typically rely on relatively low-dimensional state spaces with clear physical interpretations, making it challenging for them to scale and adapt to the complexities posed by high-dimensional pixel observations.
- **Challenge 3:** Control-theoretical approaches typically rely on prior knowledge of the unsafe regions and require explicit models of the environment dynamics which are often smooth (ODE, SDE, etc.). However, such requirements become impractical in the context of unknown contact-rich and non-smooth environments, with no prior knowledge about the hazard regions, a common scenario in RL setups.

To address these challenges, we introduce a novel state-wise safety-constrained RL framework tailored for image observation in unknown environments. Our framework aims to optimize control policies and minimize safety violations during the training and exploration stages. Specifically, for **Challenge 2**, we learn to compress the pixel observation into a low-dimensional latent space. For **Challenge 3**, with the compressed latent space, we further learn the latent MDP-like dynamics in it to deal with the data from contact-rich and non-smooth unknown environments. We establish a latent barrier-like function on it to encode the state-wise safety constraint for **Challenge 1**.

Our framework jointly conducts the latent modeling, latent barrier-like function learning, and policy optimization in an actor-critic framework, leading to improvements in both safety and overall return. Compared to model-free approaches, our approach is sample-efficient by generating training data from the latent model to avoid unsafe interactions with environmental hazards. In comparison to existing model-based methods, our approach directly enforces safety by the power of the barrier function, resulting in fewer state-wise safety violations during the training process and faster convergence of cost return, while achieving similar total reward return.

The paper is organized as follows. Section 2 introduces related works; Section 3 elaborates on the formulation of our problem; Section 4 presents our joint learning framework, including latent modeling, latent barrier-like function learning, and policy optimization; Section 5 showcases experiments setup and results; Section 6 concludes our paper.

2. Related Works

Safe RL by CMDP with High-Dimensional Input: The primal-dual approaches have been widely adopted to solve the Lagrangian problem of CMDP in a model-free manner, such as PDO (Chow et al., 2017), OPDOP (Ding et al., 2021), CPPO (Stooke et al., 2020), FOCOPS (Zhang et al., 2020b), CRPO (Xu et al., 2021), and P3O (Shen et al., 2022), which typically deal with state input rather than RGB pixels as in our work. The RL community has recently shown a growing interest in the challenge of learning policies from rich, high-dimensional data inputs (Rafailov et al., 2021; Zhang et al., 2020a; Hafner et al., 2019, 2020; Hansen et al., 2022). Recent efforts have been made to address CMDP with pixel observations (Yarats et al., 2021). However, the model-free CMDP approaches are sample-inefficient in dealing with the high dimensional image input (Shang et al., 2021) and thus lead to a large number of safety violations during the exploration. Recent approaches have aimed to mitigate these issues by mapping image observations to low-dimensional spaces to reduce sampling complexity and number of safety violations (As et al., 2022; Hogewind et al., 2022) in a model-based manner for CMDP. Despite this, CMDP approaches still have difficulty enforcing the state-wise safety constraints (Xiong et al., 2023; Wang et al., 2023b).

Safe RL with Sate-wise Constraint and High-Dimensional Input: There are some efforts to combine classic control theory methods to enforce safety in safe RL, such as reach-avoid RL (Hsu et al., 2021), contraction RL (Sun et al., 2021). Among them, barrier function is one of the most powerful approaches. The barrier function is a formal certificate that is affiliated with a control policy to prove state-wise safety of a dynamical system (Ames et al., 2016; Dawson et al., 2023). In classical control theory, a common approach is to relax the conditions of the barrier function into optimization formulations such as linear programs (Wang et al., 2023a; Yang et al., 2016) and quadratic programs (Ames et al., 2019; Choi et al., 2020). Recent work (Wang et al., 2023b,a; Cheng et al., 2019; Qin et al., 2021; Dawson et al., 2022b) proposes to jointly learn control policy and neural barrier function to optimize state-wise safety constraints in RL. However, a major problem of these approaches is their limited scalability to a higher-dimensional system let alone pixel observation. Efforts to encode state-wise safety using barrier functions derived from visual inputs have been made in (Dawson et al., 2022a; Tong et al., 2023; Cui et al., 2022; Abdi et al., 2023). However, these methods typically rely on a known dynamic model or require depth/distance information from the perception beyond the scope of unknown environments with RGB pixel observations as assumed in our work. Additionally, there have been attempts to transform high-dimensional state spaces into lower-dimensional representations and utilize sampling-based in-distribution barrier functions (Castañeda et al., 2023). Nonetheless, this approach is limited to a perfect continuous dynamic model and necessitates the presence of an existing reference controller.

3. Problem Formulation

We consider an RGB-pixel-observation safe RL problem with an unknown environment. Assume the environment can be modeled as a finite-horizon Markov Decision Process (MDP) $\mathcal{M} \sim (\mathcal{S}, \mathcal{O}, \mathcal{A}, \mathcal{P}, r, \gamma)$. $\mathcal{S} \subset \mathbb{R}^n (n \in \mathbb{Z}_+)$ stands for a continuous state space, $\mathcal{A} \subset \mathbb{R}^m (m \in \mathbb{Z}_+)$ stands for a continuous action space, and $s_{t+1} \sim \mathcal{P}(\cdot|s_t, a_t)$ stands for the unknown transition dynamic function of the environment, where $s_t \in \mathcal{S}$ and $a_t \in \mathcal{A}$. $\mathcal{O} \subset \mathbb{R}^{c_o \times h_o \times w_o} (c_o, h_o, w_o \in \mathbb{Z}_+)$ is the observation space captured by the camera module on the agent, $r : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ is the reward function of RL, and $\gamma \in [0, 1]$ is a discount factor. The observation space is dependent on the state space as there exists an unknown function that maps an underlying $s_t \in \mathcal{S}$ to the $o_t \in \mathcal{O}$

captured by the agent. We assume the control policy π_θ by the agent generates control actions a_t by consuming the underlying state o_t, i.e., $a_t \in \mathcal{A} \sim \pi(\cdot|o_t)$. It is a realistic assumption as in the real world, the agent usually is not able to get the ground truth of its state and takes actions based on the observed image. We consider state-wise safety for this MDP formulation. Assume there exists some **unknown** unsafe spaces as a set $S_u \subset \mathbb{R}^n$, the state-wise safety violation is defined as $s_t \in S_u$, *assuming there exists a safety detector κ that can check the safety violation.* Overall, the state-wise safe RL problem with pixel observation is defined as

$$\max_\theta J(\pi_\theta) = \mathbb{E}_{\mathcal{P}(\cdot|s_t,a_t)}\left[\sum_{t=0}^{T} \gamma^t r(s_t, a_t, s_{t+1})\right], \text{s.t.}, \sum_{t=0}^{T} \kappa(s_t) \leq D, a_t \sim \pi_\theta(\cdot|o_t). \tag{1}$$

where $\kappa(s_t) \in \{0, 1\}$ indicates safety violation, $D \in \mathbb{R}$ is a safety violation budget. In real-life safety-critical systems, we would like the safety violations as few as possible, therefore we would like to minimize the number of safety violations towards zero during learning, e.g., $D \to 0$.

4. Our Approach

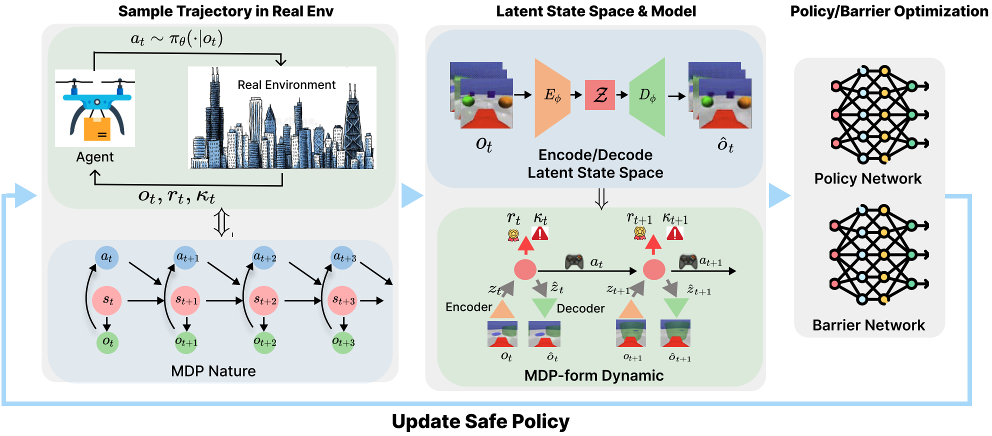

Figure 1: Overview logistic diagram of our joint learning framework. We sample pixel-observation data from real environments, learn to compress the image data to a low-dimensional latent model with MDP-like latent dynamics, and then learn a latent barrier-like function on top of it to encode state-wise safety constraints and conduct policy optimization for safe exploration and performance improvement simultaneously.

We introduce our joint learning framework for state-wise safe RL with pixel observations. Figure 1 shows the high-level overview of our approach including latent modeling, barrier-like function learning, and policy optimization. To alleviate the complexity of the high-dimensional pixel observations (**Challenge 2**), we first design our approach to learn to compress the image observation into a low-dimensional latent vector by reconstructing it in a VAE-like manner and further build forward dynamics within this latent space, i.e., latent dynamics. Due to the power of learning, such a latent modeling approach can consume the data from the non-smooth contact-rich environment observations and is able to learn the unsafe regions by our design of latent safety predictor (**Challenge 3**). Hence, the MDP-like latent model functions as a generative model for producing synthetic data used in training. Consequently, our approach operates in a model-based fashion, minimizing interactions with the actual environment and thereby reducing safety violations. With the foundation of latent dynamics, we then build a latent barrier-like function on top of the latent model to encode state-wise safety constraints to further improve safety (**Challenge 1**) by training from the synthetic data with the safety labels from a learned latent safety predictor. It is worth noting that the training gradient from the barrier-like function can back-propagate to the control policy for safer actions. Meanwhile, we conduct policy optimization to improve the total expected return in a model-based manner. The overview of our approach is in Algorithm 1. Next, we are going to introduce each component by each subsection.

Algorithm 1 State-wise Safe RL with Pixel Observations

```
Data: Unknown environment with an initial policy $\pi_\theta$, generated horizon $H$, action repeat $R$, collect interval $C$, batch size $B$, chunk length $L$, total episodes $E$, episode length $T$
Result: Policy $\pi_\theta$ with barrier-like function $B_\theta$ and the latent model with $\phi$
Initialize dataset $\mathcal{D}$ with random seed episodes, models with parameters $\theta$, $\phi$, and $\omega$.
for epoch in $E$ do
    for update step in $C$ do
        Sample batch of sequence chunks $\{(o_t, a_t, r_t, \kappa_t)_{t=k}^{L+k}\}_{i=1}^{B} \sim \mathcal{D}$.
        Train latent model 4.1 and calculate $\mathcal{L}_m$ from Equation (2).
        Update $\phi \leftarrow \phi + \varphi \nabla_\phi \mathcal{L}_m$.        // Update the latent model in one pass
        Generate trajectories $\{(z_t, a_t, \hat{r}_t, \hat{\kappa}(z_t))_{t=\tau}^{\tau+H}\}_{i=1}^{B\times(L+k)}$ using current policy in latent space.
        Compute $\mathcal{L}_b$ from Equation (4), $\mathcal{L}_p$ from Equation (5).
        Update $\theta \leftarrow \theta + \alpha \nabla_\theta \mathcal{L}_p$.        // Update barrier and policy in one pass
        Update $\omega \leftarrow \omega + \alpha \nabla_\omega \sum \frac{1}{2} \| v_\omega(z_t) - \widehat{V}_\omega^\pi(z_t) \|$.        // Update value network
    end
    for i in $\frac{T}{R}$ do
        Compute $z_t$ and $a_t$ from latent model and $\pi_\theta$, add exploration noise on top $a_t$.
        $r_t, \kappa_{t+1}, o_{t+1} \leftarrow$env.step($a_t$). // Deploy in real env to collect traj
    end
    Add the new trajectory to $\mathcal{D}$.
end
```

4.1. Pixels to Latent State Space with Latent Dynamics

Our framework first learns to transform the environment MDP (as in Section 3) into an MDP-like latent model with a low-dimensional latent space $(\mathcal{Z}, \mathcal{A}, \mathcal{T}, \hat{r}, \hat{\kappa}, \gamma)$. The comprehensive depiction of our latent space is illustrated in Figure 1. To streamline computational complexity, we utilize a VAE-like Encoder(E_ϕ)-Decoder(D_ϕ) structure to encode pixel observations $\mathcal{O}$ into low-dimensional **latent state spaces** $\mathcal{Z}$ ($\mathcal{Z} \sim E_\phi(\mathcal{O})$, $\mathcal{O} \sim D_\phi(\mathcal{Z})$). In addition, we construct a **reward predictor** and **safety predictor** model for predicting the reward $\hat{r}_t(z_t, a_t)$ and safety status $\hat{\kappa}_t(z_t) \in \{0, 1\}$ associated with the respective latent state z_t for a given o_t. It is important to note that in our formulated problem (see Formulation 3), we presume the existence of a safety detector $\kappa(s_t)$, which might be a fusion of different types of sensors. The role of the safety predictor $\hat{\kappa}(z_t)$ is to predict the output of this safety detector $\kappa(s_t)$, such that the latent space can tell the safety violation within it. Furthermore, to emulate the dynamics of the MDP in Section 3, an inference **transition model**

$\mathcal{T}$ is applied to the latent state space. This model, taking a latent state z_t and an action from the policy a_t as input, outputs the Gaussian distribution of the z_{t+1}, i.e., $z_{t+1} \sim \mathcal{T}(\cdot|z_t, a_t)$. We note that this latent model shares the same control policy with the real environment MDP. With these components, we can fully capture the dynamical nature of the environment in the latent space (i.e., $s_{t+1} \sim \mathcal{P}(\cdot|s_t, a_t) \rightarrow z_{t+1} \sim \mathcal{T}(\cdot|z_t, a_t)$) with reward and safety signals. Besides, this latent model can serve as a generative model to synthesize data for training the control policy, i.e., we can sample latent trajectory data $\{(z_t, a_t, \hat{r}_t, \hat{\kappa}_t)\}_{t=0}^{T}$ and thus reduce the agent's interactions with the real environment during the training to avoid unsafe manners. We train our latent model by using trajectories chunk with time length T from the data buffer of the real environment MDP, and we define the loss function as follows.

$$\mathcal{L}_m = \sum_{t=0}^{T} (\mathrm{KL}(z_t || E_\phi(o_t)) + \mathrm{MSE}(\hat{r}_t, r_t) + \mathrm{MSE}(\hat{\kappa}_t, \kappa_t) + \mathrm{MSE}(\hat{o}_t, o_t)) \tag{2}$$

The first *KL-Divergence Loss* measures the distribution difference between inferred latent state z_t and ground-truth compressed from real observation $E_\phi(o_t)$, which is used to update our transition model $\mathcal{T}$; The second and third *MSE Loss* serve to learn reward predictor and safety predictor; And the last *MSE Loss* captures loss of observation compression process through reconstruction from latent space. All the components of the latent model share common parameters ϕ and therefore are updated at the same backward pass. This low-dimensional latent model with assumptions on the MDP nature of dynamics can learn the non-smooth transition to address the aforementioned **Challenge 2, 3**. For implementation details, the probabilistic transition model $\mathcal{T}$ is implemented as an RNN, The observation encoder and decoder are CNN and transposed CNN respectively, and the reward predictor and safety predictor are all constructed as DNNs. We build our latent model on top of existing RSSM structure (Hafner et al., 2019). It is worth noting that we leverage a different notion of latent state space and learn the safety predictor of the environment.

4.2. Latent Barrier-like Function Learning

Based on the previous latent model, we introduce our barrier-like function on latent state space to intuitively enforce the forward invariance for state-wise safety constraints where the safe and unsafe latent states can be separated from the aforementioned safety predictor.

Definition 1 *Given a policy π_θ, B_θ is a barrier-like function of the latent state space if it satisfies the conditions below:*

$$B_\theta(z_s) > 0, \quad B_\theta(\mathbb{E}(z_t|z_{t-1})) - B_\theta(z_{t-1}) + \alpha(B_\theta(\mathbb{E}(z_t|z_{t-1}))) > 0, \quad B_\theta(z_u) < 0 \tag{3}$$

where $z_t \sim \mathcal{T}(\cdot|z_{t-1}, \pi_\theta(z_{t-1}))$, θ stands for the parameters of the barrier-like function and policy network. $z_s \in \mathcal{Z}_s \subset \mathcal{Z}$ stands for state in safe latent state set $\mathcal{Z}_s$, $z_u \in \mathcal{Z}_u \subset \mathcal{Z}$ stands for state in unsafe latent state set $\mathcal{Z}_u$, and α is a class-$\mathcal{K}$ function.

Note that $\mathcal{Z}_s$ and $\mathcal{Z}_u$ are categorized by safety predictor learned in the latent model 4.1, i.e., $\hat{\kappa}(z_t) = 1, z_t \in \mathcal{Z}_u$, otherwise $z_t \in \mathcal{Z}_s$. The latent barrier-like function offers a state-wise definition of safety. The idea is to have the agent start in the $z_t \in \mathcal{Z}_s, B(z_t) > 0$ and, by encoding the positivity of the time derivative (as approximated by the second condition in Equation (3)), ensure that the expected subsequent state maintain this positivity, i.e., $B_\theta(\mathbb{E}[z_{t+1}]) > 0$. This approach results in the invariance of the agent within the safe state set in expectation. However, due to partial observability, there may be instances where the agent unintentionally enters the unsafe state set. When this occurs, the barrier-like function yields a negative value. Still, the positivity of the consecutive state function value difference guides the agent away from unsafe regions and towards states characterized by a positive barrier value. In contrast, CMDP approaches primarily focus on minimizing the total cost over an entire trajectory in expectation without considering this state-wise encoding of safety.

Remark 2 *In this study, we add a stochastic element to the mean of the distribution derived from the transition model $\mathcal{T}$ and denote it as z_t for subsequent computations. This approach is commonly employed in various model-based methods (As et al., 2022; Hogewind et al., 2022). Our experimental findings reveal that without incorporating stochasticity into the mean significantly deteriorates efficacy of the reconstruction process due to the deterministic shortcut from encoder output directly to decoder input (Hafner et al., 2019). This, consequently, results in poor overall quality of latent model learning and policy performance.*

We implement the latent barrier-like function as a dense neural network and derive the following loss vector with inspiration from (Qin et al., 2021).

$$\mathcal{L}_b = [\mathrm{ReLU}(\eta - B_\theta(z_s)), \mathrm{ReLU}(\eta - (B_\theta(z_t) - B_\theta(z_{t-1}) + \alpha(B_\theta(z_t)))), \mathrm{ReLU}(\eta + B_\theta(z_u))] \tag{4}$$

The first term penalizes non-positive safe states; the second term enforces the positivity of the approximated time derivative condition; the third term penalizes the non-negative unsafe states of the function. We apply a nominal positive learning margin $\eta \in \mathbb{R}$ to make the optimization more feasible. Note that since the problem possesses partial observability, the above formulation can only enforce the forward invariance without a formal guarantee.

4.3. Policy Optimization

To optimize the total rewards while considering state-wise safety, we formulate an actor-critic approach with barrier-like function learning in the loop *within the latent model*, by using the trajectories $\{z_t, a_t, \hat{r}_t, \hat{\kappa}_t\}$ generated by the latent model. With the encoder network embedded inside, the policy DNN $\pi_\theta(\cdot|E_\phi(o_t))$ (or equivalently $\pi_\theta(\cdot|z_t), z_t \approx E_\phi(o_t)$) outputs action a_t in a Gaussian distribution, which is randomly sampled for training and provides mean action value for evaluation. The value (critic) function of RL $v_\omega(z_t)$ can be expressed as $v_\omega(z_t) = \mathbb{E}_{\pi_\theta(\cdot|z_t)}\left[\sum_{t=\tau}^{\tau+T} \gamma^t \hat{r}_t\right]$. We define the total expected return as $J_\phi(\pi_\theta) = v_\omega(z_0)$ and aim to reduce the following loss function for the overall RL objective.

$$\mathcal{L}_p = -\max J_\phi(\pi_\theta) + \beta^T \mathcal{L}_b \tag{5}$$

where we add barrier loss function as a regularization term for safety. β here is a coefficient vector corresponding to each term in Equation (5). And we backward $\mathcal{L}_p$ through stochastic backpropagation to update the policy network and barrier-like function in the same pass. Specifically, for the critic network, we use the sampled synthetic latent trajectory $\{z_t, a_t, \hat{r}_t, \hat{\kappa}_t\}$ and Monte Carlo approach to provide the learning target $\widehat{V}_\omega^\pi(z_t) = \frac{1}{N}\sum_{i=1}^N \left[\sum_{\tau=0}^T \gamma^\tau \hat{r}_\tau + \gamma^{t+T} v_\omega(z_{t+T})\right]$ for it to reduce the loss $\left\| v_\omega - \widehat{V} \right\|^2$. For the actor network, the policy gradient from J_ϕ is well-established as in (Sutton and Barto, 2018).

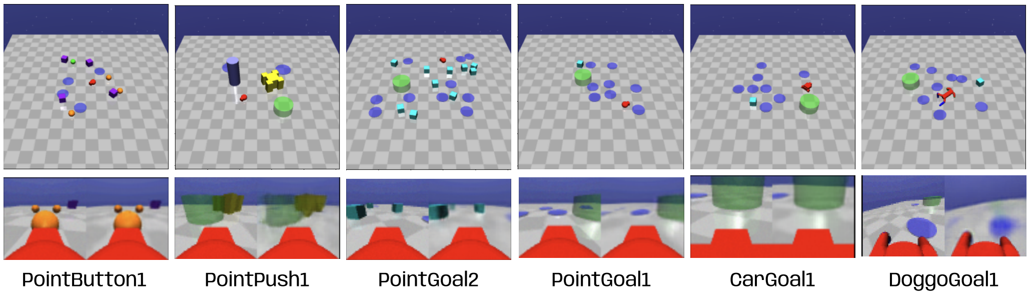

Figure 2: This is a graphic expression of `Safety Gym SG6` (Ray et al., 2019) environments. Figures at the top row are with birdeye views of each benchmark. All images on the left-hand side in the bottom row are pixel observations taken by the agents, and all on the right-hand side are reconstructed images from our learned latent model, which appear similarly to the left observations.

5. Experiments

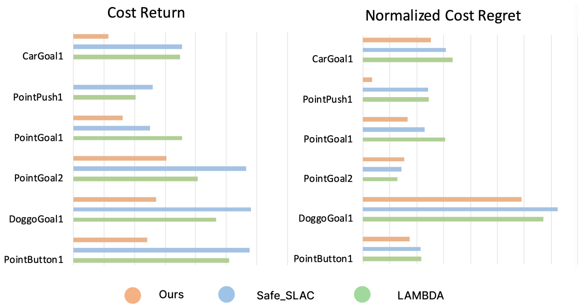

Figure 3: Left shows the *Cost Return* of LAMBDA (green), Safe-SLAC (blue), and our approach (red) on the `Safety Gym` benchmarks; Right illustrates the *Normalized Cost Regret* with respect to PPO's. Overall, we can tell that our approach achieves fewer safety violations.

We adopt the widely used `Safety Gym SG6` tasks (Ray et al., 2019) as test examples. For our approach and baselines, the agents take in $3 \times 64 \times 64$ pixels images from the agents' point of view, as shown in Figure 2. To the best of our knowledge, there is currently no state-wise safe RL approach capable of handling non-smooth environments encountered in the `Safety Gym` framework with pixel observations. Therefore, we compare to popular model-free CMDP methods including CPO, PPO-L, and TRPO-L, and pay more focus on LAMBDA (As et al., 2022) and Safe-SLAC (Hogewind et al., 2022), two state-of-the-art model-based CMDP approaches that study safe RL with pixel observations. We train different algorithms with one million environment steps, except the `DoggoGoal1` with two million steps. We measure the performance of all approaches on the *Cost Return* and *Cost Regret* (Ray et al., 2019), corresponding to state-wise safety in evaluation and training. In addition, we showcase the learning curve of our method compared with others' converged values after training.

- The average *Cost Return* for N episodes is defined as $\hat{J}(\pi) = \frac{\sum_{i=0}^{N} \sum_{t=0}^{T_{ep}} c_t}{N}$, where T_{ep} is the length of a single episode. Since c_t is fixed for each step of safety violation in the `Safety Gym` environment, we can interpret this metric as the average safety violations performance of policy π after each training episode.
- The *Cost Regret* is the sum of costs during the whole training process over the total interaction steps T. We defined as $\rho_c(\pi) = \frac{\sum_{t=0}^{T} c_t}{T}$, T is the total environment interactions, and c_t is the cost

corresponding to each interaction. This metric represents total safety violations accumulated in the entire training process.

- The average *Reward Return* for N episodes is defines as $\hat{R}(\pi) = \frac{\sum_{i=0}^{Tep} r_t}{N}$, where r_t is the reward received at time step t.

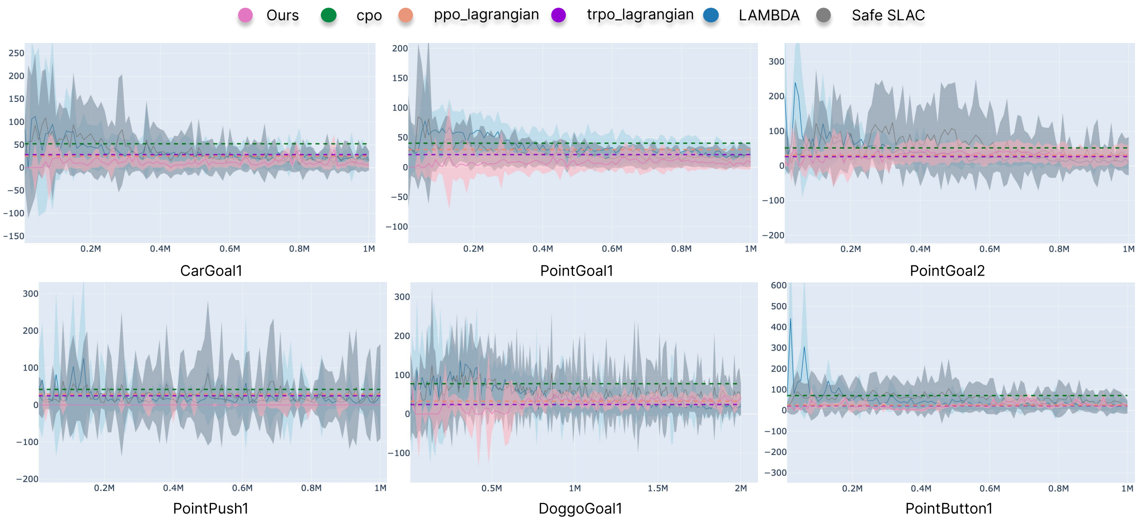

Figure 4: Dash-lines are *Cost Return* of the model-free CPO, PPO-L, TRPO-L trained for 10M steps except `DoggoGoal1` with 50M training steps. Our approach shows faster cost convergence in all benchmarks compared with other approaches. Besides, for the majority of the benchmark, our approach can achieve the lowest converged cost return.

5.1. Safety Evaluation during Learning Explorations

Figure 3 shows *Cost Return* and normalized *Cost Regret* of our approach and baselines, where we normalize the *Cost Regret* by dividing it by the *Cost Regret* achieved by the **PPO** method, i.e., $\hat{\rho}_{ours} = \frac{\rho_{ours}}{\rho_{ppo}}$. It is worth noting that model-free approaches in principle lead to more safety violations (As et al., 2022) and thus we mainly focus on model-based Safe-SLAC and LAMBDA for a fair comparison in Figure 3. Compared to these two baselines, our approach exhibits a notable enhancement for *Cost Return* in all benchmarks and consistently demonstrates lower *Cost Regret*, except the `PointGoal2` environment. The results affirm the advantages of our latent barrier-like function learning for encoding state-wise safety constraints over the CMDP formulation in the baselines. In addition, from Figure 4, we can tell that our approach shows faster convergence in the *Cost Return*. The reason is that during the training, our latent model quickly identifies and captures the majority of unsafe latent states by supervised learning. With more interactions, the latent barrier-like encoding hard state-wise safety constraint progressively forces the agent to take safer actions, leading to significantly lower *Cost Return* compared to CMDP approaches.

Remark 3 *Safety violations are unavoidable in our setup as the agent only receives partial image observations from a single front-view camera in an unknown environment without state information, where zero violation fundamentally is a hard problem to solve. In addition, due to learning errors,* *our latent model may not always accurately differentiate the safe and unsafe images, which could lead to safety violations.*

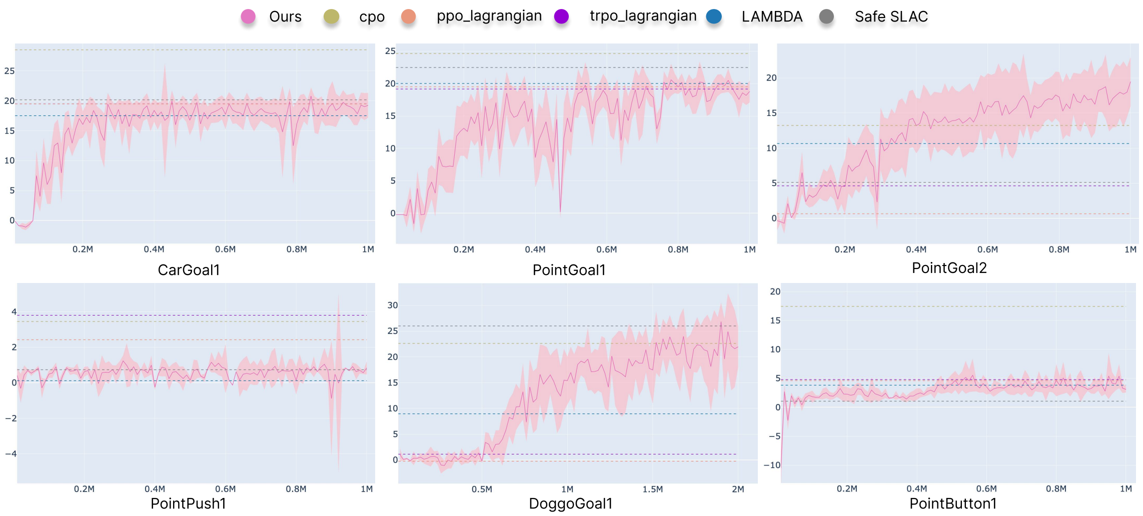

Figure 5: *Reward Return*: The model-free methods CPO, PPO-L, and TRPO-L are trained for 10M steps except on `DoggoGoal1` with 50M steps. The model-based Safe-SLAC and LAMBDA are trained with 1M steps except on `DoggoGoal1` with 2M environment steps. Our approach can achieve similar performance as model-free methods or slightly better as other model-based methods.

5.2. Reward Performance of the Learned Policies

In principle, model-free approaches could obtain higher total expected return after convergence since model-based approaches face model-mismatch errors in the learning process (Altman, 2021). Surprisingly, in Figure 5, our approach can achieve similar performance as model-free methods and slightly surpass other model-based methods across most of the benchmarks, except `PointPush1`. This pheromone indicates our latent model can accurately compress and reconstruct the image observation space, as shown in Figure 2. In `PointPush1`, model-free methods outperform all model-based methods including ours. Our hypothesis is based on the observation in Figure 2, which suggests that, in `PointPush1`, the agent's ability to capture additional visual information is severely limited when attempting to push the yellow box, which results in inaccurate model learning.

6. Conclusion

This paper introduces a model-based state-wise safe RL framework with pixel observations. We first learn to compress the high-dimensional image into a latent model where we establish a latent barrier-like function to encode state-wise safety constraints. We jointly conduct latent modeling, barrier-like function learning, and policy optimization to improve safety and performance simultaneously. Our approach significantly enhances safety while maintaining performance levels comparable to other established model-free and model-based safe RL methods. A possible future direction is to consider the distribution-based barrier function under the MDP setting for safe and unsafe states by encoding the forward invariance within the distributions, rather than the current sampled state sets.

Acknowledgments

Simon Sinong Zhan's work is supported by Northwestern First-Year Graduate Fellowship; Yixuan Wang, Ruochen Jiao, and Qi Zhu's work is partially supported by US National Science Foundation grants 1834701 and 2324936; Qingyuan Wu and Chao Huang's work is supported by the grant EP/Y002644/1 under the EPSRC ECR International Collaboration Grants program, funded by the International Science Partnerships Fund (ISPF) and the UK Research and Innovation.

References

Hossein Abdi, Golnaz Raja, and Reza Ghabcheloo. Safe control using vision-based control barrier function (v-cbf). In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 782–788. IEEE, 2023.

Eitan Altman. *Constrained Markov decision processes*. Routledge, 2021.

Aaron D Ames, Xiangru Xu, Jessy W Grizzle, and Paulo Tabuada. Control barrier function based quadratic programs for safety critical systems. *IEEE Transactions on Automatic Control*, 62(8): 3861–3876, 2016.

Aaron D Ames, Samuel Coogan, Magnus Egerstedt, Gennaro Notomista, Koushil Sreenath, and Paulo Tabuada. Control barrier functions: Theory and applications. In *2019 18th European control conference (ECC)*, pages 3420–3431. IEEE, 2019.

Yarden As, Ilnura Usmanova, Sebastian Curi, and Andreas Krause. Constrained policy optimization via bayesian world models. *arXiv preprint arXiv:2201.09802*, 2022.

Somil Bansal, Mo Chen, Sylvia Herbert, and Claire J Tomlin. Hamilton-jacobi reachability: A brief overview and recent advances. In *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, pages 2242–2253. IEEE, 2017.

Fernando Castañeda, Haruki Nishimura, Rowan Thomas McAllister, Koushil Sreenath, and Adrien Gaidon. In-distribution barrier functions: Self-supervised policy filters that avoid out-of-distribution states. In *Learning for Dynamics and Control Conference*, pages 286–299. PMLR, 2023.

Richard Cheng, Gábor Orosz, Richard M Murray, and Joel W Burdick. End-to-end safe reinforcement learning through barrier functions for safety-critical continuous control tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3387–3395, 2019.

Jason Choi, Fernando Castaneda, Claire J Tomlin, and Koushil Sreenath. Reinforcement learning for safety-critical control under model uncertainty, using control lyapunov functions and control barrier functions. *arXiv preprint arXiv:2004.07584*, 2020.

Yinlam Chow, Mohammad Ghavamzadeh, Lucas Janson, and Marco Pavone. Risk-constrained reinforcement learning with percentile risk criteria. *The Journal of Machine Learning Research*, 18(1):6070–6120, 2017.

Yuxiang Cui, Longzhong Lin, Xiaolong Huang, Dongkun Zhang, Yunkai Wang, Wei Jing, Junbo Chen, Rong Xiong, and Yue Wang. Learning observation-based certifiable safe policy for decentralized multi-robot navigation. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 5518–5524. IEEE, 2022.

Charles Dawson, Bethany Lowenkamp, Dylan Goff, and Chuchu Fan. Learning safe, generalizable perception-based hybrid control with certificates. *IEEE Robotics and Automation Letters*, 7(2): 1904–1911, 2022a.

Charles Dawson, Zengyi Qin, Sicun Gao, and Chuchu Fan. Safe nonlinear control using robust neural lyapunov-barrier functions. In *Conference on Robot Learning*, pages 1724–1735. PMLR, 2022b.

Charles Dawson, Sicun Gao, and Chuchu Fan. Safe control with learned certificates: A survey of neural lyapunov, barrier, and contraction methods for robotics and control. *IEEE Transactions on Robotics*, 2023.

Dongsheng Ding, Xiaohan Wei, Zhuoran Yang, Zhaoran Wang, and Mihailo Jovanovic. Provably efficient safe exploration via primal-dual policy optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 3304–3312. PMLR, 2021.

Jiajun Duan, Di Shi, Ruisheng Diao, Haifeng Li, Zhiwei Wang, Bei Zhang, Desong Bian, and Zhehan Yi. Deep-reinforcement-learning-based autonomous voltage control for power grid operations. *IEEE Transactions on Power Systems*, 35(1):814–817, 2019.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *International conference on machine learning*, pages 2555–2565. PMLR, 2019.

Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering atari with discrete world models. *arXiv preprint arXiv:2010.02193*, 2020.

Nicklas Hansen, Xiaolong Wang, and Hao Su. Temporal difference learning for model predictive control. *arXiv preprint arXiv:2203.04955*, 2022.

Yannick Hogewind, Thiago D Simao, Tal Kachman, and Nils Jansen. Safe reinforcement learning from pixels using a stochastic latent representation. *arXiv preprint arXiv:2210.01801*, 2022.

Kai-Chieh Hsu, Vicenç Rubies-Royo, Claire J Tomlin, and Jaime F Fisac. Safety and liveness guarantees through reach-avoid reinforcement learning. *arXiv preprint arXiv:2112.12288*, 2021.

B Ravi Kiran, Ibrahim Sobh, Victor Talpaert, Patrick Mannion, Ahmad A Al Sallab, Senthil Yogamani, and Patrick Pérez. Deep reinforcement learning for autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 23(6):4909–4926, 2021.

Hongjia Li, Tianshu Wei, Ao Ren, Qi Zhu, and Yanzhi Wang. Deep reinforcement learning: Framework, applications, and embedded implementations: Invited paper. In *2017 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, pages 847–854, 2017. doi: 10.1109/ICCAD.2017.8203866.

Teng Liu, Bin Tian, Yunfeng Ai, and Fei-Yue Wang. Parallel reinforcement learning-based energy efficiency improvement for a cyber-physical system. *IEEE/CAA Journal of Automatica Sinica*, 7 (2):617–626, 2020.

Zengyi Qin, Kaiqing Zhang, Yuxiao Chen, Jingkai Chen, and Chuchu Fan. Learning safe multi-agent control with decentralized neural barrier certificates. *arXiv preprint arXiv:2101.05436*, 2021.

Rafael Rafailov, Tianhe Yu, Aravind Rajeswaran, and Chelsea Finn. Offline reinforcement learning from images with latent space models. In *Learning for Dynamics and Control*, pages 1154–1168. PMLR, 2021.

Alex Ray, Joshua Achiam, and Dario Amodei. Benchmarking Safe Exploration in Deep Reinforcement Learning. 2019.

Wenling Shang, Xiaofei Wang, Aravind Srinivas, Aravind Rajeswaran, Yang Gao, Pieter Abbeel, and Misha Laskin. Reinforcement learning with latent flow. *Advances in Neural Information Processing Systems*, 34:22171–22183, 2021.

Li Shen, Long Yang, Shixiang Chen, Bo Yuan, Xueqian Wang, Dacheng Tao, et al. Penalized proximal policy optimization for safe reinforcement learning. *arXiv preprint arXiv:2205.11814*, 2022.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

Adam Stooke, Joshua Achiam, and Pieter Abbeel. Responsive safety in reinforcement learning by pid lagrangian methods. In *International Conference on Machine Learning*, pages 9133–9143. PMLR, 2020.

Dawei Sun, Susmit Jha, and Chuchu Fan. Learning certified control using contraction metric. In *Conference on Robot Learning*, pages 1519–1539. PMLR, 2021.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Mukun Tong, Charles Dawson, and Chuchu Fan. Enforcing safety for vision-based controllers via control barrier functions and neural radiance fields. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 10511–10517. IEEE, 2023.

Hiroyasu Tsukamoto, Soon-Jo Chung, and Jean-Jacques Slotine. Learning-based adaptive control via contraction theory.

Yixuan Wang, Chao Huang, and Qi Zhu. Energy-efficient control adaptation with safety guarantees for learning-enabled cyber-physical systems. In *Proceedings of the 39th International Conference on Computer-Aided Design*, pages 1–9, 2020.

Yixuan Wang, Simon Zhan, Zhilu Wang, Chao Huang, Zhaoran Wang, Zhuoran Yang, and Qi Zhu. Joint differentiable optimization and verification for certified reinforcement learning. In *Proceedings of the ACM/IEEE 14th International Conference on Cyber-Physical Systems (with CPS-IoT Week 2023)*, pages 132–141, 2023a.

Yixuan Wang, Simon Sinong Zhan, Ruochen Jiao, Zhilu Wang, Wanxin Jin, Zhuoran Yang, Zhaoran Wang, Chao Huang, and Qi Zhu. Enforcing hard constraints with soft barriers: Safe reinforcement learning in unknown stochastic environments. In *International Conference on Machine Learning*, pages 36593–36604. PMLR, 2023b.

Yixuan Wang, Weichao Zhou, Jiameng Fan, Zhilu Wang, Jiajun Li, Xin Chen, Chao Huang, Wenchao Li, and Qi Zhu. Polar-express: Efficient and precise formal reachability analysis of neural-network controlled systems. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 43(3):994–1007, 2024. doi: 10.1109/TCAD.2023.3331215.

Hua Wei, Guanjie Zheng, Huaxiu Yao, and Zhenhui Li. Intellilight: A reinforcement learning approach for intelligent traffic light control. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2496–2505, 2018.

Tianshu Wei, Yanzhi Wang, and Qi Zhu. Deep reinforcement learning for building hvac control. In *2017 54th ACM/EDAC/IEEE Design Automation Conference (DAC)*, June 2017. doi: 10.1145/3061639.3062224.

Nuoya Xiong et al. Provably safe reinforcement learning with step-wise violation constraints. *arXiv preprint arXiv:2302.06064*, 2023.

Shichao Xu, Yangyang Fu, Yixuan Wang, Zhuoran Yang, Zheng O'Neill, Zhaoran Wang, and Qi Zhu. Accelerate online reinforcement learning for building hvac control with heterogeneous expert guidances. In *Proceedings of the 9th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*, BuildSys '22, page 89–98, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450398909. doi: 10.1145/3563357.3564064. URL https://doi.org/10.1145/3563357.3564064.

Tengyu Xu, Yingbin Liang, and Guanghui Lan. Crpo: A new approach for safe reinforcement learning with convergence guarantee. In *International Conference on Machine Learning*, pages 11480–11491. PMLR, 2021.

Bai Xue, Naijun Zhan, Martin Fränzle, Ji Wang, and Wanwei Liu. Reach-avoid verification based on convex optimization. *IEEE Transactions on Automatic Control*, 2023.

Zhengfeng Yang, Chao Huang, Xin Chen, Wang Lin, and Zhiming Liu. A linear programming relaxation based approach for generating barrier certificates of hybrid systems. In *International Symposium on Formal Methods*, pages 721–738. Springer, 2016.

Denis Yarats, Amy Zhang, Ilya Kostrikov, Brandon Amos, Joelle Pineau, and Rob Fergus. Improving sample efficiency in model-free reinforcement learning from images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 10674–10681, 2021.

Liang Yu, Shuqi Qin, Meng Zhang, Chao Shen, Tao Jiang, and Xiaohong Guan. A review of deep reinforcement learning for smart building energy management. *IEEE Internet of Things Journal*, 8(15):12046–12063, 2021.

Amy Zhang, Rowan McAllister, Roberto Calandra, Yarin Gal, and Sergey Levine. Learning invariant representations for reinforcement learning without reconstruction. *arXiv preprint arXiv:2006.10742*, 2020a.

Yiming Zhang, Quan Vuong, and Keith Ross. First order constrained optimization in policy space. *Advances in Neural Information Processing Systems*, 33:15338–15349, 2020b.

Wenshuai Zhao, Jorge Peña Queralta, and Tomi Westerlund. Sim-to-real transfer in deep reinforcement learning for robotics: a survey. In *2020 IEEE symposium series on computational intelligence (SSCI)*, pages 737–744. IEEE, 2020.

Qi Zhu, Wenchao Li, Hyoseung Kim, Yecheng Xiang, Kacper Wardega, Zhilu Wang, Yixuan Wang, Hengyi Liang, Chao Huang, Jiameng Fan, and Hyunjong Choi. Know the unknowns: Addressing disturbances and uncertainties in autonomous systems. In *Proceedings of the 39th International Conference on Computer-Aided Design*, ICCAD '20, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450380263. doi: 10.1145/3400302.3415768. URL https://doi.org/10.1145/3400302.3415768.