

Learning Interpretable, High-Performing Policies for Autonomous Driving

Rohan Paleja*, Yaru Niu*, Andrew Silva, Chace Ritchie, Sugju Choi, and Matthew Gombolay

Atlanta, Georgia 30332

Georgia Institute of Technology

{rohan.paleja, yaruniu, andrew.silva, critchie7, schoi424, matthew.gombolay}@gatech.edu

Abstract—Gradient-based approaches in reinforcement learning (RL) have achieved tremendous success in learning policies for autonomous vehicles. While the performance of these approaches warrants real-world adoption, these policies lack interpretability, limiting deployability in the safety-critical and legally-regulated domain of autonomous driving (AD). AD requires interpretable and verifiable control policies that maintain high performance. We propose Interpretable Continuous Control Trees (ICCTs), a tree-based model that can be optimized via modern, gradient-based, RL approaches to produce high-performing, interpretable policies. The key to our approach is a procedure for allowing direct optimization in a sparse decision-tree-like representation. We validate ICCTs against baselines across six domains, showing that ICCTs are capable of learning interpretable policy representations that parity or outperform baselines by up to 33% in AD scenarios while achieving a 300x-600x reduction in the number of policy parameters against deep learning baselines. Furthermore, we demonstrate the interpretability and utility of our ICCTs through a 14-car physical robot demonstration.

I. INTRODUCTION

The deployment of autonomous vehicles (AVs) has the potential to increase traffic safety [29], decrease traffic congestion, increase average traffic speed in human-driven traffic [16], reduce CO_2 emissions, and allow for more affordable transportation [1]. Recent success in developing AVs with high-performance, real-time decision-making capabilities has been driven by the generation of continuous control policies produced via reinforcement learning (RL) with deep function approximators. However, while the performance of these controllers opens up the possibility of real-world adoption, the conventional deep-RL policies used in prior work [38, 63, 16] lack *interpretability*, limiting deployability in safety-critical and legally-regulated domains [18, 36, 10, 60].

White-box approaches, as opposed to typical black-box models (e.g., deep neural networks) used in deep-RL, model decision processes in a human-readable representation. Such approaches afford interpretability, allowing users to gain insight into the model’s decision-making behavior. In autonomous driving, such models would provide insurance companies, law enforcement, developers, and passengers with insight into how an autonomous vehicle (AV) reasons about state features and makes decisions. Utilizing such white-box approaches within machine learning is necessary for the

deployment of autonomous vehicles and essential in building trust, ensuring safety, and enabling developers to inspect and verify policies before deploying them to the real world [42, 26, 4]. In this work, we present a novel tree-based architecture that affords gradient-based optimization with modern RL techniques to produce high-performance, interpretable policies for autonomous driving applications. We note that our proposed architecture can be applied to a multitude of continuous control problems in robotics [38], protein folding [28], and traffic regulation [16].

Prior work [42, 31, 26] has attempted to approximate interpretability via explainability, a practice that can have severe consequences [50]. While the explanations produced in prior work can help to partially explain the behavior of a control policy, the explanations are not guaranteed to be accurate or generally applicable across the state-space, leading to erroneous conclusions and a lack of accountability of predictive models [50]. In autonomous driving, where understanding a decision-model is critical to avoiding collisions, local explanations are insufficient. An *interpretable model* provides a transparent *global* representation of a policy’s behavior. This model can be understood directly by its structure and parameters [15] (e.g., linear models, decision trees, and our ICCTs), offering verifiability and guarantees that are not afforded by post-hoc explainability frameworks. Few works have attempted to learn an interpretable model directly; rather, prior work has attempted policy distillation to a decision tree [19, 8, 64] or imitation learning via a decision tree across trajectories generated via a deep model [7], leaving much to be desired. Interpretable RL remains an open challenge [51]. In this work, we directly produce high-performance, interpretable policies represented by a minimalistic tree-based architecture augmented with low-fidelity linear controllers via RL, providing a novel interpretable RL architecture. Our Interpretable Continuous Control Trees are human-readable, allow for closed-form verification (associated with safety guarantees), and parity or outperform baselines by up to 33% in autonomous driving scenarios. In this work:

- 1) We propose Interpretable Continuous Control Trees (ICCTs), a novel tree-based model that can be optimized via gradient descent with modern RL algorithms to produce high-performance, interpretable continuous control policies. We provide several extensions to prior DDT

*Authors contributed equally to this work.

frameworks to increase expressivity and allow for direct optimization on a sparse decision-tree-like representation.

- 2) We empirically validate ICCTs across six continuous control domains, including four autonomous driving scenarios.
- 3) We provide a qualitative example of our ICCTs, displaying its interpretability and utility.
- 4) We demonstrate our algorithm with physical robots in a 14-car driving scenario and provide an online, easy-to-inspect visualization of the ego vehicle control policy.

II. RELATED WORK

Due to recent accidents with autonomous vehicles (c.f. Yurtsever et al. [66]), there has been growing interest in developing Explainable AI (xAI) approaches to understand an AV’s decision-making and ensure robust and safe operation. Explainable AI (xAI) is concerned with understanding and interpreting the behavior of AI systems [39]. In recent years, the necessity for human-understandable models has increased greatly for safety-critical and legally-regulated domains, many of which involve continuous control (e.g., specifying joint torques for a robot arm or the steering angle for an autonomous vehicle) [32, 18]. In such domains, prior work [52, 38, 20, 23] has typically used highly-parameterized deep neural networks in order to learn high-performance policies, completely lacking in model transparency.

Interpretable machine learning approaches refers to a subset of xAI techniques that produce globally transparent policies (i.e., humans can inspect the entire model, as in a decision tree [11, 6, 43] or rule list [3, 61, 36, 13]). Tree-based frameworks are of particular importance to our work. Decision trees [11] represent a hierarchical structure where an input decision can be traced to an output via evaluation of decision nodes (i.e., “test” on an attribute) until arrival at a leaf node. Decision nodes within the tree are able to split the problem space into meaningful subspaces, simplifying the problem as the tree gets deeper [35, 33, 57]. Decision trees provide *global* explanations of a decision-making policy that are valid throughout the input space [5], as opposed to local explanations typically provided via “post-hoc” explainability techniques [48, 54, 45]. Several approaches have attempted to distill trained neural network models into decision trees [65, 8]. While these approaches produce interpretable models, the resulting model is an approximation of the neural network rather than a true representation of the underlying model. Our work, instead, directly learns an interpretable tree-based policy via reinforcement learning, producing a model that can be directly verified without utilizing error-prone post-hoc explainability techniques. We emphasize that explainability stands in contrast to interpretability, as explanations may fail to capture the true decision-making process of a model or may apply only to a local region of the decision-space, thereby preventing a human from building a clear or accurate mental model of the entire policy [50, 40, 2, 44].

Recently, [51] presented a set of grand challenges in interpretable machine learning to guide the field towards solving critical research problems that must be solved before machine

learning can be safely deployed within the real world. In this work, we present a solution to directly assess two challenges: (1) Optimizing sparse logical models such as decision trees and (10) Interpretable reinforcement learning. We propose a novel high-performing, sparse tree-based architecture, Interpretable Continuous Control Trees (ICCTs), which allows end-users to directly inspect the decision-making policy and developers to verify the policy for safety guarantees.

III. PRELIMINARIES

In this section, we review differentiable decision trees (DDTs) and reinforcement learning.

A. Differentiable Decision Trees (DDTs)

Prior work has proposed differentiable decision trees (DDTs) [56, 54] – a neural network architecture that takes the topology of a decision tree (DT). Similar to a decision tree, DDTs contain decision nodes and leaf nodes; however, each decision node within the DDT utilizes a sigmoid activation function (i.e., a “soft” decision) instead of a Boolean decision (i.e., a “hard” decision). Each decision node, i , is represented by a sigmoid function, displayed in Equation 1.

$$y_i = \frac{1}{1 + \exp(-\alpha(\vec{w}_i^T \vec{x} - b_i))} \quad (1)$$

Here, the features vectors describing the current state, \vec{x} , are weighted by \vec{w}_i , and a splitting criterion, b_i , is subtracted to form the splitting rule. y_i is the probability of decision node i evaluating to TRUE, and α governs the steepness of the sigmoid activation, where $\alpha \rightarrow \infty$ results in a step function. Prior work with discrete-action DDTs modeled each leaf node with a probability distribution over possible output classes [54, 45]. Leaf node distributions are then summed weighted by the likelihood of reaching the respective leaf to produce a final action distribution over possible outputs.

1) *Conversion of a DDT to a DT*: DDTs with decision nodes represented in the form of Equation 1 are not interpretable. As DDTs maintain a one-to-one correspondence to DTs with respect to their structure, prior work [54, 45] proposed methodology to convert a DDT into an interpretable decision tree (a process termed “crispification”). To create an interpretable, “crisp” tree from a differentiable form of the tree, prior work adopted a simplistic procedure. Starting with the differentiable form, prior work first converts each decision node from a linear combination of all variables into a single feature check (i.e., a 2-arity predicate with a variable and a threshold). The feature reduction is accomplished by considering the feature dimension corresponding to the weight with the largest magnitude (i.e., most impactful), $k = \arg \max_j |w_i^j|$, resulting in the decision node representation $y_i = \sigma(\alpha(w_i^k x^k - b_i))$. The sigmoid steepness, α , is also set to infinity, resulting in a “hard” decision (branch left OR right) [54, 45]. After applying this procedure to each decision node, decision nodes are represented by $y_i = \mathbb{1}(w_i^k x^k - b_i > 0)$. As each leaf node is represented as a probability mass function over output classes in prior work, leaf nodes, l , must be modified to produce

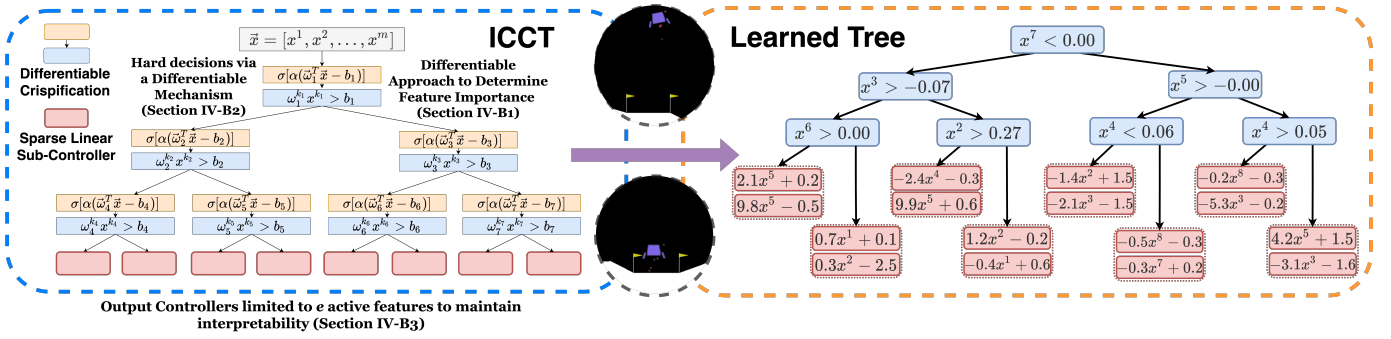


Fig. 1: The ICCT framework (left) displays decision nodes, both in their fuzzy form (orange blocks) and crisp form (blue blocks¹), and sparse linear leaf controllers with pointers to sections discussing our contributions. A learned representation of a high-performing ICCT policy in Lunar Lander (right) displays the interpretability of our ICCTs. Each decision node is conditioned upon only a single feature and the sparse linear controllers (to control the main engine throttle and left/right thrusters) are set to have only **one** active feature.

a single output class, o , during crispification. As such, we can apply an argument max, $o_d = \arg \max_a l_d^a$, to find the maximum valued class within the d^{th} leaf distribution.

Drawbacks: This simplistic crispification procedure results in an interpretable crisp tree that is inconsistent with the original DDT (model differences arise from each *argmax* operation). These inconsistencies can lead to performance degradation of the interpretable model, as we show in Section VI, and results in an interpretable model that is not representative of and inconsistent with the model learned via reinforcement learning.

In our work, we design a novel architecture that updates its parameters via gradient descent while maintaining an interpretable decision-tree-like representation, thereby avoiding any inconsistencies generated through a crispification procedure. To the best of our knowledge, we are the first work to deploy an interpretable tree-based framework for continuous control.

B. Reinforcement Learning (RL)

A Markov Decision Process (MDP) M is defined as a 6-tuple $\langle S, A, R, T, \gamma, \rho_0 \rangle$. S is the state-space, A is the action-space, $R(s, a)$ is the reward received by an agent for executing action, a , in state, s , $T(s'|s, a)$ is the probability of transitioning from state, s , to state, s' , when applying action, a , $\gamma \in [0, 1]$ is the discount factor, and $\rho_0(s)$ is the initial state distribution. A policy, $\pi(a|s)$, gives the probability of an agent taking action, a , in state, s . The goal of RL is to find the optimal policy, $\pi^* = \arg \max_{\pi} \mathbb{E}_{\tau \sim \pi} \left[\sum_{t=0}^T \gamma^t R(s_t, a_t) \right]$ to maximize cumulative discounted reward, where $\tau = \langle s_0, a_0, \dots, s_T, a_T \rangle$ is an agent's trajectory. In this work, while ICCTs are framework-agnostic (i.e., ICCTs will work with any update rule), we proceed with Soft Actor-Critic (SAC) [23] as our RL algorithm. The actor objective within SAC is given in Equation 2, where $Q_w(s_t, a_t)$ is expected, future discounted reward as parameterized by ω .

$$J_{\pi}(\theta) = \mathbb{E}_{s_t \sim \mathcal{D}} [\mathbb{E}_{a_t \sim \pi_{\theta}} [\alpha \log(\pi_{\theta}(a_t|s_t)) - Q_w(s_t, a_t)]] \quad (2)$$

¹For figure simplicity, when displaying the crisp node (blue block), we assume $\alpha > 0$ in the fuzzy node (orange block). If $\alpha < 0$, the sign of the inequality would be flipped (i.e., $w_i^k x^{k_i} < b_i$).

IV. METHOD

In this section, we introduce our ICCTs, a novel interpretable reinforcement learning architecture. ICCTs are able to maintain interpretability while representing high-performance continuous control policies, making them suitable for representing the control policies of AVs. We provide several extensions to prior DDT frameworks within our proposed architecture including 1) a differentiable crispification procedure allowing for optimization in a sparse decision-tree like representation, and 2) the addition of sparse linear leaf controllers to increase expressivity while maintaining legibility.

A. ICCT Architecture

Our ICCTs are initialized to be a symmetric decision tree with l decision leaves and $l - 1$ decision nodes (tree depth by $\log_2(l)$). Each decision leaf is represented by a linear sparse controller that is parameterized by \vec{x} . Decisions are routed via decision nodes towards a leaf controller, which is then used to produce the continuous control output (e.g., acceleration or steering wheel angle). Our ICCT is similar to hierarchical models which maintain a high-level controller over several low-level controllers. Prior work has shown this to be a successful paradigm in continuous control [41].

Each decision node, i , has an activation steepness weight, α , associated weights, \vec{w}_i , with cardinality, m , matching that of the input feature vector, \vec{x} , and a scalar bias term, b_i , similar to that of Equation 1. Each leaf node, l_d , where $d \in \{1, \dots, l\}$, contains per-leaf weights, $\vec{\beta}_d \in \mathbb{R}^m$, per-leaf selector weights that learn the relative importance of candidate splitting criterion, $\vec{\theta}_d \in \mathbb{R}^m$, per-leaf bias terms, $\vec{\phi}_d \in \mathbb{R}^m$, and per-leaf scalar standard deviations, γ_d . We note that if the action space is multi-dimensional, then *only* the leaf controllers (and associated weights) are expanded across $|A|$ dimensions, where $|A|$ is the cardinality of the action space. For each action dimension, the mean of the output action distribution is represented by the linear controller, $\vec{\beta}_d^T \vec{x} - \vec{\phi}_d$, for leaf d , where the output action can be determined via sampling ($a \sim \mathcal{N}(\vec{\beta}_d^T \vec{x} - \vec{\phi}_d, \gamma_d)$) during training and directly

via the mean during runtime. We term decision nodes that are represented as Equation 1 as fuzzy decision nodes, displayed by the orange rectangles within the left-hand side of Figure 1. Similarly, we term leaf nodes that are represented as $\vec{\beta}_d^T \vec{x} - \vec{\phi}_d$ as fuzzy leaf nodes.

Utilizing a novel differentiable crispification procedure to convert fuzzy decision nodes into crisp decision nodes (i.e., 2-arity predicate with a variable and a threshold) and fuzzy leaf nodes into sparse leaf nodes (i.e., linear controller conditioned upon a subset of features), our model representation follows that of a decision tree with sparse linear controllers at the leafs (shown on the right-hand side of Figure 1). We further discuss our differentiable crispification procedure in Sections IV-B1 and IV-B2 (i.e., the mechanism that translates orange blocks to blue within Figure 1).

While decision trees (DT) are generally considered interpretable [36], trees of arbitrarily large depths can be difficult to understand [21] and simulate [40]. A sufficiently sparse DT is desirable and considered interpretable [34]. Utilizing linear controllers at the leaves also allows us to maintain interpretability, as linear controllers are widely used and generally considered interpretable for humans [25]. However, similarly, for large feature spaces typically encountered in real-world problems, such a controller would not be interpretable. As such, in our work, we utilize *sparse* linear controllers at the leaves to balance the trade-off between sparsity/complexity in logic, model depth, and performance.

B. ICCT Key Elements

In this section, we discuss our ICCT’s interpretable procedure for determining an action given an input feature. As our ICCT configuration maintains interpretability both during training via RL and deployment, the inference of an action must allow gradient flow. We present a novel approach that allows for direct optimization of sparse logical models via an online differentiable crispification procedure to determine feature importance (Section IV-B1) and allows for bifurcate decisions (Section IV-B2). In Algorithm 1, we provide general pseudocode representing our ICCT’s decision-making process.

At each timestep, the ICCT model, $I(\cdot)$, receives a state feature, \vec{x} . To determine an action in an interpretable form, in Steps 1 and 2 of Algorithm 1, we start by applying the differentiable crispification approaches of `NODE_CRISP` and `OUTCOME_CRISP` to decision nodes so that each decision node is only conditioned upon a single variable (Section IV-B1), and the evaluation of a decision node results in a Boolean (Section IV-B2). Once the operations are completed, in Step 3, we can utilize the input feature, \vec{x} , and logically evaluate each decision node until arrival at a linear leaf controller (`INTERPRETABLE_NODE_ROUTING`). The linear leaf controller is then modified, in Step 4, to only condition upon e features, where e is a sparsity parameter specified a priori (Section IV-B3). Finally, an action can be determined via sampling from a Gaussian distribution conditioned upon the mean generated via the input-parameterized sparse leaf controller, l_d^* , and scalar variance maintained within the leaf,

γ_d , (Step 6) during training or directly through the outputted mean (Step 8) during runtime.

Algorithm 1 ICCT Action Determination

Input: ICCT $I(\cdot)$, state feature $\vec{x} \in S$, controller sparsity e , training flag $t \in \{True, False\}$

Output: action $a \in \mathbb{R}$

```

1: NODE_CRISP( $\sigma(\alpha(\vec{w}_i^T \vec{x} - b_i))$ )  $\rightarrow \sigma(\alpha(w_i^k x^k - b_i))$ 
2: OUTCOME_CRISP( $\sigma(\alpha(w_i^k x^k - b_i))$ )  $\rightarrow \mathbb{1}(\alpha(w_i^k x^k - b_i) > 0)$ 
3:  $l_d \leftarrow \text{INTERPRETABLE\_NODE\_ROUTING}(\vec{x})$ 
4:  $l_d^* \leftarrow \text{ENFORCE\_CONTROLLER\_SPARSITY}(e, l_d)$ 
5: if  $t$  then
6:    $a \sim \mathcal{N}(l_d^*(\vec{x}), \gamma_d)$ 
7: else
8:    $a \leftarrow l_d^*(\vec{x})$ 
9: end if

```

1) *Decision Node Crispification:* The `NODE_CRISP` procedure in Algorithm 1 recasts each decision node to split upon a single dimension of \vec{x} . Instead of using a non-differentiable argument max function as in Silva and Gombolay [54] to determine the most impactful feature dimension, we utilize a softmax function, also known as softargmax [22], described by Equation 3. In this equation, we denote the softmax function as $f(\cdot)$, which takes as input a set of class weights and produces class probabilities. Here, \vec{w}_i represents a categorical distribution with class weights, individually denoted by w_i^j , and τ is the temperature, determining the steepness of $f(\cdot)$.

$$f(\vec{w}_i)_k = \frac{\exp\left(\frac{w_i^k}{\tau}\right)}{\sum_j^m \exp\left(\frac{w_i^j}{\tau}\right)} \quad (3)$$

While setting the temperature near-zero would satisfy our objective of producing a one-hot vector, where the outputted class probability of the index of the most impactful feature would be one, this operation can lead to vanishing gradients. We therefore set τ equal to 1, and utilize a `one_hot` function which produces a one-hot vector with the element associated with the highest-weighted class set to one and all other elements set to zero. We display the differentiable procedure for determining the weight associated with the largest magnitude in Equation 4, where $g(\cdot)$ is the `one_hot` function.

$$\vec{z}_i = g(f(|\vec{w}_i|)) \quad (4)$$

Here, $|\vec{w}_i|$ represents a vector with absolute elements within \vec{w}_i . The one-hot encoding \vec{z}_i can be element-wise multiplied by the original weights to produce a new set of weights with only one active weight, $\vec{z}_i \odot \vec{w}_i \rightarrow \vec{w}_i^l$. Accordingly, the decision node representation is transferred from $\sigma(\alpha(\vec{w}_i^T \vec{x} - b_i)) \rightarrow \sigma(\alpha(\vec{w}_i^l \vec{x} - b_i)) = \sigma(\alpha(w_i^k x^k - b_i))$, where k is the index of the most impactful feature. *We maintain differentiability in the procedure described in Equation 4 by utilizing the straight-through trick [9].* The differentiable one-hot operation is further described in Algorithm 4 within the Appendix (Section C). We provide an algorithm detailing the `NODE_CRISP` procedure within the Appendix (Section A). Below, we conduct a short example detailing our procedure.

Example: Say we have a two-leaf decision tree (one decision node) with an input feature, $\vec{x} = [2, 3]$ with a cardinality of 2 (i.e. $m = 2$), associated weights of $\vec{w}_1 = [2, 1]^T$, and a bias term b_1 of 1. The sigmoid steepness, α is also set equal to 1 for simplicity. It is easily seen that the most impactful weight within the decision node is $w_1^1 = 2$. Utilizing Equation 4, we can compute $\vec{z}_1 = [1, 0]$. Multiplying \vec{z}_1 to the original weights, \vec{w}_1 and input feature, \vec{x} , subtracting b_1 , and scaling by α , we have an crisp decision node $\sigma(2x^1 - 1)$ or $\sigma(3) = 0.95$. Here, .95 is the probability that the decision node evaluates to *TRUE*.

2) *Decision Outcome Crispification:* Here, we describe the second piece of our online differentiable crispification procedure, noted as `OUTCOME_CRISP` in Algorithm 1. `OUTCOME_CRISP` translates the outcome of a decision node so that the outcome is a Boolean decision rather than a set of probabilities generated via a sigmoid function (i.e., $p = y_i$ for True/Left Branch and $q = 1 - y_i$ for False/Right Branch). We start by creating a soft vector representation of the decision node output $\vec{v}_i = [\alpha(w_i^k x^k - b_i), 0]$, for the i^{th} decision node. Placing \vec{v}_i through a softmax operation, we can produce the probability of tracing down the left branch or right. We can then apply a one-hot operation to produce a hard decision of whether to branch left or right, denoted by y_i and described by Equation 5. Essentially, the decision node will evaluate to True if $\alpha(w_i^k x^k - b_i) > 0$ and right otherwise. This process can be expressed as an indicator function $\mathbb{1}(\alpha(w_i^k x^k - b_i) > 0)$.

$$(y_i, 1 - y_i) = g(f(\vec{v}_i)) \quad (5)$$

We note the procedure of $g(f(\vec{v}_i))$ is highly similar to that in Equation 4, both outputting a one-hot vector, with the former input being the decision node weights, $|\vec{w}_i|$, and the latter input being the soft vector representation of the decision node outcome, \vec{v}_i . We provide an algorithm detailing the `OUTCOME_CRISP` procedure within the Appendix (Section B), including the operations to maintain gradients via the straight-through trick (Algorithm 4).

Example: Continuing the example in Section IV-B1, we can take the outputted crisp decision node and generate a vector $\vec{v}_1 = [2x^1 - 1, 0]$, or by substituting the input feature, $\vec{v}_1 = [3, 0]$. Performing the operations specified in Equation 5, we receive the intermediate output from the softmax $[.95, .05]$ (rounded to two decimal places), the one-hot vector $[1, 0]$ after performing the one_hot operation, and finally $y_1 = 1$, denoting that the decision-tree should follow the left branch.

Conversion to a Simple Form: The above crispification process produces decision-tracing equal to that of a DT. The node representation can be further simplified to that of Figure 1 by algebraically reducing each crisp decision node to $x^k > \frac{b_i}{w_i^k}$ (given $\alpha w_i^k > 0$) or $x^k < \frac{b_i}{w_i^k}$ (given $\alpha w_i^k < 0$).

3) *Sparse Linear Leaf Controllers:* After applying the decision node and outcome crispification to all decision nodes and outcomes, the decision can be routed to leaf node (Step 3). This section describes the procedure to translate a linear leaf controller to condition upon e features

(`ENFORCE_CONTROLLER_SPARSITY` procedure in Algorithm 1), enforcing sparsity within the leaf controller and thereby, enhancing ICCT interpretability. As noted in Section IV-A, our ability to utilize sparse sub-controllers allows us to balance the interpretability-performance tradeoff. The sparsity of the linear sub-controllers ranges from setting $e = 0$ and maintaining static leaf distributions, where each leaf node contains scalar value representing the mean (i.e., ICCT-static in Section VI), to $e = m$, containing a linear controller parameterized by the entire feature space of \vec{x} (i.e., ICCT-complete in Section VI).

Equation 6 displays the procedure for determining a `k_hot` encoding, \vec{u}_d , that represents the k (or in our case, e) most impactful selection weights within a leaf's linear controller. The `k_hot` function, denoted by $h(\cdot)$, takes as input a vector of class weights and returns an equal-dimensional vector with k elements set to one. The indexes associated with the elements set to one match the k highest-weighted elements within the input feature.

$$\vec{u}_d = h(f(|\vec{\theta}_d|)) \quad (6)$$

Here, $|\vec{\theta}_d|$ represents a vector with absolute elements within $\vec{\theta}_d$. We maintain differentiability through the operation in Equation 6 by utilizing the straight-through trick [9]. In Equation 7, we transform a fuzzy leaf node, for leaf d , into a sparse linear sub-controller.

$$l_d^* = (\vec{u}_d \circ \vec{\beta}_d)^T (\vec{u}_d \circ \vec{x}) + \vec{u}_d^T \vec{\phi}_d \quad (7)$$

A depiction of the sparse sub-models can be seen at the bottom of Figure 1, where the sparsity of the sub-controllers, e , is set to 1 and the dimension of the action space is 2.

Summary: In this section, we discuss our novel interpretable reinforcement learning architecture, ICCTs. We present a description of components of ICCTs, including decision nodes and linear leaf controllers, and provide a differentiable crispification procedure allowing for optimization in a sparse decision-tree like representation. To the best of our knowledge, we present the first truly interpretable tree-based framework for continuous control.

V. ENVIRONMENTS

Here, we provide short descriptions across six domains used in our extensive evaluation. We start with two common continuous control problems, Inverted Pendulum and Lunar Lander provided by OpenAI Gym [12]. We then test across four autonomous driving scenarios: Lane-Keeping provided by Leurent [37] and Single-Lane Ring Network, Multi-Lane Ring Network, and Figure-8 Network all provided by the Flow deep reinforcement learning framework for mixed autonomy traffic scenarios [63]. We provide additional details and depictions of each domain within the Appendix (Section H).

Inverted Pendulum: In Inverted Pendulum [58], a control policy must apply throttle (ranging from +3 to move left to -3 to move right) to balance a pole. The observation includes the cart position, velocity, pole angle, and pole angular velocity.

Lunar Lander: In Lunar Lander [46, 12], a policy must throttle main engine and side engine thrusters for a lander

TABLE I: In this table, we display the results of our evaluation. For each evaluation, we report the mean (\pm standard error) and the complexity of the model required to generate such a result. Our table is broken into three segments, the first containing equally interpretable approaches that utilize static distributions at their leaves. The second segment contains interpretable approaches that maintain linear controllers at their leaves. The ordering of methods denotes the relative interpretability. The third segments displays black-box approaches. We bold the highest-performing method in each segment, and break ties in performance by model complexity. We color table elements in association with the number of parameters and performance. Reddish colors relate to a larger number of policy parameters and lower average reward.

Worst to Best:

Method	Common Continuous Control Problems			Autonomous Driving Problems		
	Inverted Pendulum	Lunar Lander	Lane Keeping	Single-Lane Ring	Multi-Lane Ring	Figure-8
DT	155.0 \pm 0.9 256 leaves (766 params)	-285.5 \pm 15.6 256 leaves (1022 params)	-359.0 \pm 11.0 256 leaves (766 params)	123.2 \pm 0.3 32 leaves (94 params)	503.2 \pm 24.8 256 leaves (1022 params)	831.1 \pm 1.1 256 leaves (766 params)
DT w\ DAGGER	793.1 \pm 189.9 32 leaves (94 params)	124.3 \pm 138.7 32 leaves (126 params)	263.3 \pm 194.7 16 leaves (46 params)	121.5 \pm 0.03 16 leaves (46 params)	1263.9 \pm 22.4 31 leaves (122 params)	1047.1 \pm 15.1 16 leaves (46 params)
CDDT-Crisp ²	5.0 \pm 0.0 2 leaves (7 params)	-451.6 \pm 97.3 8 leaves (53 params)	-43526.0 \pm 15905.0 16 leaves (77 params)	68.1 \pm 18.7 16 leaves (77 params)	664.5 \pm 192.6 16 leaves (109 params)	322.9 \pm 47.1 16 leaves (525 params)
ICCT-static	984.0 \pm 10.4 32 leaves (127 params)	192.4 \pm 10.7 32 leaves (221 params)	374.2 \pm 55.8 16 leaves (77 params)	120.5 \pm 0.5 16 leaves (77 params)	1271.7 \pm 4.1 16 leaves (109 params)	1003.8 \pm 27.2 16 leaves (77 params)
ICCT-1-feature	1000.0 \pm 0.0 8 leaves (45 params)	190.1 \pm 13.7 8 leaves (69 params)	437.6 \pm 7.0 16 leaves (77 params)	121.6 \pm 0.5 16 leaves (93 params)	1269.6 \pm 10.7 16 leaves (157 params)	1072.4 \pm 37.1 16 leaves (93 params)
ICCT-2-feature	1000.0 \pm 0.0 4 leaves (29 params)	258.4 \pm 7.0 8 leaves (101 params)	458.5 \pm 6.3 16 leaves (93 params)	121.9 \pm 0.5 16 leaves (125 params)	1280.4 \pm 7.3 16 leaves (237 params)	1088.6 \pm 21.6 16 leaves (125 params)
ICCT-3-feature	1000.0 \pm 0.0 2 leaves (17 params)	275.8 \pm 1.5 8 leaves (133 params)	448.8 \pm 3.0 16 leaves (109 params)	120.8 \pm 0.5 16 leaves (157 params)	1280.8 \pm 7.7 16 leaves (317 params)	1048.7 \pm 46.7 16 leaves (157 params)
ICCT-L1-sparse	1000.0 \pm 0.0 4 leaves (33 params)	265.2 \pm 4.3 8 leaves (181 params)	465.5 \pm 4.3 16 leaves (269 params)	121.5 \pm 0.3 16 leaves (781 params)	1275.3 \pm 6.7 16 leaves (2221 params)	993.2 \pm 14.6 16 leaves (525 params)
ICCT-complete	1000.0 \pm 0.0 2 leaves (15 params)	300.5 \pm 1.2 8 leaves (181 params)	476.6 \pm 3.1 16 leaves (269 params)	120.7 \pm 0.5 16 leaves (781 params)	1248.6 \pm 3.6 16 leaves (2221 params)	994.1 \pm 29.1 16 leaves (525 params)
CDDT-controllers Crisp ²	84.0 \pm 10.4 2 leaves (15 params)	-126.6 \pm 53.5 8 leaves (181 params)	-39826.4 \pm 21230.0 16 leaves (269 params)	97.9 \pm 12.0 16 leaves (781 params)	639.62 \pm 160.4 16 leaves (2221 params)	245.5 \pm 48.5 16 leaves (525 params)
MLP-L	1000.0 \pm 0.0 86 params	231.6 \pm 49.8 124 params	474.7 \pm 5.8 134 params	121.8 \pm 0.6 155 params	646.4 \pm 151.2 229 params	868.4 \pm 100.9 107 params
MLP-U	1000.0 \pm 0.0 130 params	288.7 \pm 2.8 224 params	467.9 \pm 8.5 422 params	121.8 \pm 0.3 722 params	1239.5 \pm 4.2 3332 params	1077.7 \pm 31.1 1042 params
MLP-Max	1000.0 \pm 0.0 67589 params	298.5 \pm 0.7 69124 params	478.2 \pm 6.7 69634 params	121.7 \pm 0.4 77826 params	1011.9 \pm 141.3 83972 params	1104.3 \pm 9.4 73730 params
CDDT	1000.0 \pm 0.0 2 leaves (10 params)	226.4 \pm 44.5 8 leaves (102 params)	464.7 \pm 5.4 16 leaves (242 params)	120.9 \pm 0.5 16 leaves (708 params)	1248.0 \pm 6.4 16 leaves (1084 params)	1033.2 \pm 24.1 16 leaves (482 params)
CDDT-controllers	1000.0 \pm 0.0 2 leaves (18 params)	289.0 \pm 2.4 8 leaves (230 params)	469.7 \pm 11.1 16 leaves (434 params)	120.1 \pm 0.3 16 leaves (1412 params)	1243.8 \pm 3.6 16 leaves (3196 params)	1010.9 \pm 25.7 16 leaves (930 params)

to land on a specified landing pad. The observation is 8-dimensional including the lander’s current position, linear velocity, tilt, angular velocity, and information about ground contact. The continuous action space is two dimensional for controlling the main engine thruster and side thrusters.

Lane-Keeping [37]: A control policy must control a vehicle’s steering angle to stay within a curving lane. The observation is 12-dimensional, which consists of the vehicle’s lateral position, heading, lateral speed, yaw rate, linear, lateral, and angular velocity, and the lane information. The action is the steering angle to control the vehicle.

Flow Single-Lane Ring Network [63]: A control policy must apply acceleration commands to a vehicle agent to stabilize traffic flow consisting of 21 other human-driven (synthetic) vehicles. The observation includes the world position on velocity of all vehicles.

Flow Multi-Lane Ring Network [63]: A control policy must apply acceleration and lane-changing commands to an ego vehicle to stabilize the flow of noisy traffic flow across multiple lanes. The observation includes the world position and velocity of all vehicles.

Flow Figure-8 Network [63]: A control policy must apply acceleration to a vehicle to stabilize the flow in a Figure-8 network (contains a section where the vehicles must cross paths at the center of the 8), requiring the policy to adapt its control input to create a stable flow through this section. The observation is the world position and velocity of all vehicles.

VI. RESULTS

In this section, we present the set of baselines we test our model against. Then, we report the results of our approach versus these baselines across the six continuous control domains, as shown in Table 1. All presented results are across five random seeds and all differentiable frameworks are trained via SAC [23]. Each tree-based framework is trained while maximizing performance and minimizing the complexity required to represent such a policy, thereby emphasizing interpretability. We release our anonymous codebase at <https://github.com/CORE-Robotics-Lab/ICCT>.

²As this model is derived via crispification [54], the complexity associated with this model is directly determined by its non-crisp counterpart.

A. Baselines

We provide a list of baselines alongside abbreviations used for reference and brief definitions below. We compare against interpretable models, black-box models, and models that can be converted post-hoc into an interpretable form.

- Continuous DDTs (CDDT): We translate the framework of Silva and Gombolay [54] to function with continuous action-spaces by modifying the leaf nodes to represent static probability distributions. When converted into an interpretable form post-hoc, this approach is reported as CDDT-crisp.
- Continuous DDTs with controllers (CDDT-controllers): We modify CDDT leaf nodes to utilize linear controllers rather than static distributions. When converted into an interpretable form post-hoc, this approach is reported as CDDT-controllers-crisp.
- ICCTs with static leaf distributions (ICCT-static): We modify the leaf architecture of our ICCTs to utilize static distributions for each leaf (i.e., set $e = 0$). Comparing ICCT and ICCT-static displays the effectiveness of the addition of sparse linear sub-controllers.
- ICCT with complete linear sub-controllers (ICCT-complete): We allow the leaf controllers to maintain weights over all features (no sparsity enforced, i.e., $e = m$). Comparing ICCT-complete and CDDT-controllers displays the effectiveness of the proposed differentiable crispification procedure.
- ICCT with L1-regularized controllers (ICCT-L1-sparse): We achieve sparsity via L1-regularization applied to ICCT-complete rather than enforce sparsity directly via the ENFORCE_CONTROLLER_SPARSITY procedure. While this baseline produces sparse sub-controllers, there are drawbacks limiting its interpretability. L1-regularization enforces weights to be near-zero rather than exactly zero. These small weights must be represented within decision-nodes and thus, the interpretability of the resulting model is limited.
- Multi-layer Perceptron (MLP) - We maintain three variants of an MLP. The first contains a very large number of parameters (MLP-Max), typically utilized in continuous control domains. The second maintains approximately the same number of parameters of our ICCT, including all inactive parameters (MLP-Upper). The last maintains approximately the same number of *active* parameters as our ICCT (MLP-Lower).
- Decision Tree (DT): We train a DT via CART [11] on state-action pairs generated from MLP-Max.
- DT w/ DAGGER: We utilize the DAGGER imitation learning algorithm [49] to train a DT to mimic MLP-Max.

B. Discussion

We present the results of our trained policies in Table I. We provide the performance of each method alongside the associated complexity of each benchmark in Table I across three sections, with the top section representing interpretable approaches that maintain static distributions at their leaves, the middle section containing interpretable approaches that maintain linear controllers at their leaves, and the bottom section containing black-box methods.

Static Leaf Distributions (Top): The frameworks of DT, DT \ DAGGER, CDDT-Crisp, ICCT-static maintain similar representations and are equal in terms of interpretability given that the approaches have the same depth. We see that across four of the six control domains, ICCT-static is able to widely outperform both the DT and CDDT-Crisp model. In complex autonomous driving domains such as Lane Keeping, we are able to widely outperform prior work and outperform DT w/ DAGGER by 42.11% while maintaining less parameters.

Controller Leaf Distributions (Middle): Here, we rank frameworks (top-down) by their relative interpretability. As the sparsity of the sub-controller decreases, the interpretability diminishes. We see that most approaches are able to achieve the maximum performance in the simple domain of Inverted Pendulum. However, CDDT-controllers-crisp encounters an inconsistency issue from the crispification procedure of [54, 45] and achieves very low performance. In regards to interpretability-performance tradeoff, in Inverted Pendulum, we see that as sparsity increases within the sub-controller, a lower-depth ICCT can be used to achieve an equally high-performing policy. We note that across all domains, we do not find such a linear relationship. We provide additional results within Section VIII that provide deeper insight into the interpretability-performance tradeoff.

Black-Box Approaches (Bottom): MLP-based approaches and fuzzy DDTs are not interpretable. While the associated approaches perform well across many of the six domains, the lack of interpretability limits the utility of such frameworks in real-world applications such as autonomous driving. We see that in half the domains, highly-parameterized architectures with over 65,000 parameters are required to learn effective policies (denoted by the dark green shade).

Comparison Across All Approaches: We see that across all continuous control domains, CDDT-Crisp and CDDT-controllers Crisp typically are the lowest-performing models. This displays the drawbacks of the crispification procedure of [54, 45] and the resultant performance inconsistency. Comparing our ICCTs to black-box models, we see that in all domains, we parity or outperform deep highly-parameterized models in performance while reducing the number of parameters required by orders of magnitude. In the difficult Multi-Lane Ring scenario, we see to we can outperform MLPs by 33% on average while achieving a 300x-600x reduction in the number of policy parameters required.

Overall, we find extremely positive support for our Interpretable Continuous Control Trees, displaying the ability to at least parity black-box approaches while maintaining high interpretability. Our novel architecture and training procedure provide a strong step towards providing solutions for two grand challenges in interpretableML: (1) Optimizing sparse logical models such as DTs and (10) Interpretable RL.

VII. QUALITATIVE EXPOSITION OF ICCT INTERPRETABILITY

Here, we provide a display of the utility and interpretability of a learned ICCT model. In Figure 2, we present our learned

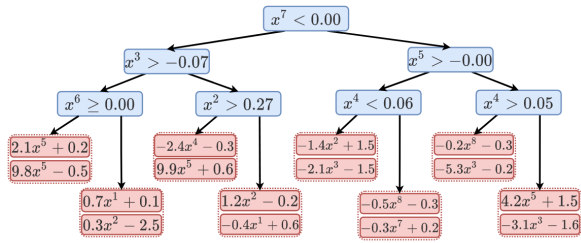


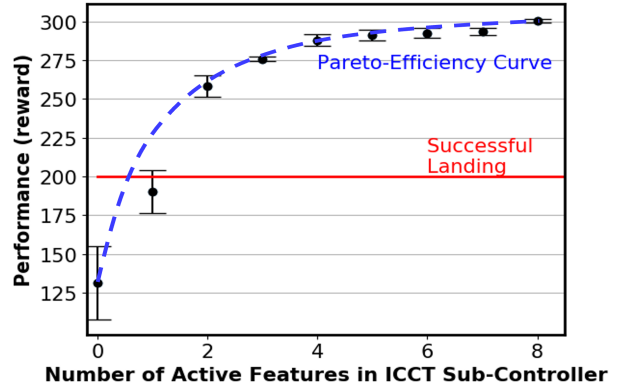
Fig. 2: A Learned ICCT in Lunar Lander

ICCT model in Lunar Lander, rounding each element to two decimal places for brevity. The displayed figure is an ICCT-1-feature model (i.e., only one active feature within the sparse sub-controller). The 8-dimensional input in Lunar Lander is composed of position (x^1, x^2), velocity (x^3, x^4), angle (x^5), angular velocity (x^6), left (x^7) and right (x^8) lander leg-to-ground contact. The action space is two-dimensional: the first (dictated by the top of each pair of the red-colored leaves) controls the main engine thrust, and the second (bottom) controls the net thrust for the side-facing engines. The tree can be interpreted as follows: taking the leftmost path as an example, if the left leg is not touching the ground (≤ 0.00 m), the horizontal velocity is greater than -0.07 m/s, and the angular velocity is greater than 0.00 rad/s, then the main engine action is $2.1 \times (\text{the lander angle}) + 0.2$, and the side engine action is $9.8 \times (\text{the lander angle}) - 0.5$. Such a tree has several use cases: 1) An engineer/developer may pick certain edge cases and verify the behavior of the lander. Tree-based models are amenable to verification [59]. Furthermore, *tree-based models similar to ICCTs can be verified in linear time* [14], while *DNN verification is NP-complete* [30]. 2) An engineer can evaluate the decision-making in the tree and detect anomalies. Furthermore, there are hands-on use-cases of such a model, such as threshold editing (directly modifying nodes to increase affordances), etc.

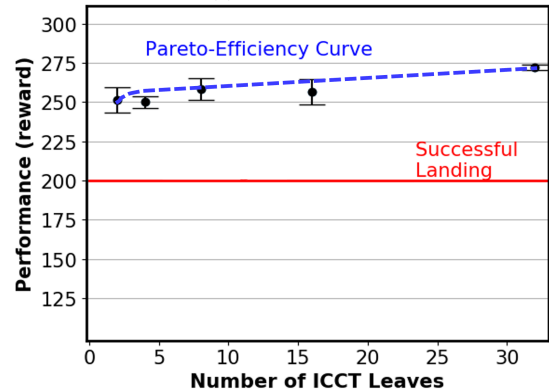
VIII. ABLATION: INTERPRETABILITY-PERFORMANCE TRADEOFF

Here, we provide an ablation study over how ICCT performance changes with respect to the number of active features within our linear sub-controllers and depth of the learned policies. Lakkaraju et al. [34] states that decision trees are interpretable because of their simplicity and that there is a cognitive limit on how complex a model can be while also being understandable. Accordingly, for our ICCTs to maximize interpretability, we emphasize the sparsity of our sub-controllers and attempt to minimize the depth of our ICCTs. Here, we present a deeper analysis by displaying the performance of our ICCTs while varying the number of active features, e , from ICCT-static to ICCT-complete (Figure 3a), and varying the number of leaves maintained within the ICCT from $l = 2$ to $l = 32$. We conduct our ablation study within Lunar Lander.

In Figure 3a, we show how the performance of our ICCTs change as a function of active features in the Sub-Controller.



(a) ICCT Framework



(b) Example Learned Decision Tree in Lunar Lander

Fig. 3: In this figure, we display the interpretability-performance tradeoff of our ICCTs with respect to active features within our linear sub-controllers (Figure 3a) and tree depth (Figure 3b). Within each figure, we display the Pareto-Efficiency Curve and denote the reward required for a successful lunar landing as defined by [12].

Here, we fix the number of ICCT leaves to 8. We see that as the number of active features increase, the performance also increases. However, there is a tradeoff in interpretability. As above 200 reward is considered successful in this domain, a domain expert may determine a point on the Pareto-Efficiency curve that maximizes the interpretability-performance tradeoff. In Figure 3b, we show how the performance of our ICCTs change as a function of tree depth while fixing the number of active features in the ICCT sub-controller to two. We see a similar, albeit weaker, relationship between performance and interpretability. As model complexity increases, there is a slight gain in performance and a large decrease in interpretability. The Pareto-Efficiency curve provides insight into the interpretability-performance tradeoff for ICCT tree depth.

IX. PHYSICAL ROBOT DEMONSTRATION

Here, we demonstrate our algorithm with physical robots in a 14-car figure-8 driving scenario and provide an online, easy-to-inspect visualization of our ICCTs, which controls the ego vehicle. We utilize the Robotarium, a remotely accessible

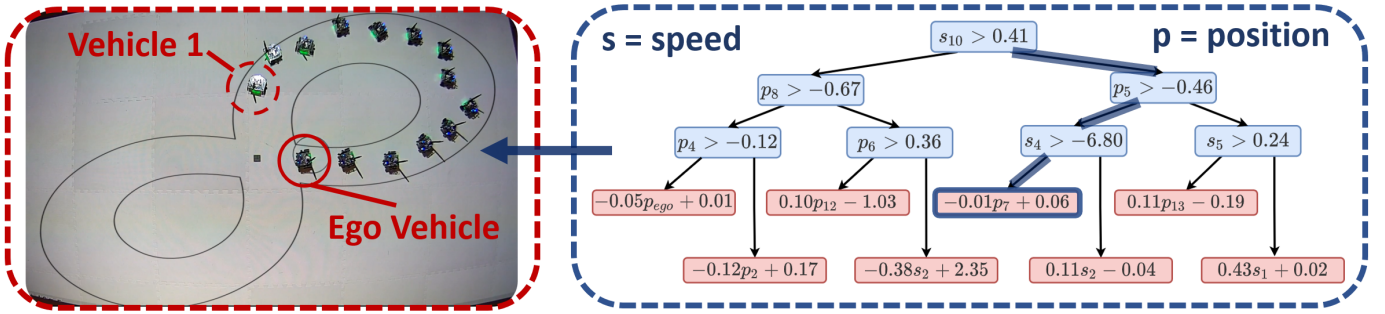


Fig. 4: In this figure, we display our ICCTs controlling a vehicle in a 14-car physical robot demonstration within a Figure-8 traffic scenario. Active nodes and edges are highlighted by the right online visualization, where s_i represents the speed of vehicle i , and p_i represents the position of vehicle i . We include a full video, including an enlarged display of our ICCT within the supplementary material.

swarm robotic research platform [62], to demonstrate the learned ICCT policy. The demonstration displays the feasibility of ego vehicle behavior produced by our ICCT policy and provides an online visualization of our ICCTs. A frame taken from the demonstrated behavior is displayed in Figure 4. We provide a complete video of the demonstrated behavior and the online visualization of the control policy within the supplementary material.

X. CONCLUSION

In this work, we present a novel tree-based model for autonomous vehicle control. Our Interpretable Continuous Control Trees (ICCTs) have competitive performance to that of deep neural networks across six continuous control domains, including four difficult autonomous driving scenarios, while maintaining high interpretability. The maintenance of both high performance and interpretability within an interpretable reinforcement learning architecture provides a paradigm that would be beneficial for the real-world deployment of autonomous systems.

XI. LIMITATIONS AND FUTURE WORK:

In planned work, we propose a deepening algorithm to dynamically grow ICCTs, allowing ICCTs to mitigate covariate shift and represent additional complexities when deployed to real-world settings. In the appendix, we present a sample of our deepening algorithm (Section G, which can currently be used to alleviate the need to set the number of ICCT leaves a priori. We also plan to conduct a complete user-study, assessing if users can simulate ICCTs, verify ICCTs, and gain insight into the model’s decision-making, similar to [55, 45].

Our framework has several limitations. Continuous control outputs (e.g., predicting a steering angle) may not be interpretable to end-users and may require post-processing to enhance a user’s understanding. Also, the relationship between controller sparsity, tree depth, and interpretability is not clear, making controller sparsity and tree depth difficult-to-define hyperparameters. We also note that in more challenging environments, larger ICCTs may be required for representative power. However, in these cases, while ICCTs will be difficult

to interpret by end-users due to their size, our ICCT policies can still be verified by experts and can be interpreted within tree sub-spaces.

XII. ACKNOWLEDGMENTS

This work was supported by a gift award from the Ford Motor Company, NSF 1757401 (SURE Robotics), and a research grant from MIT Lincoln Laboratory (7000437192).

REFERENCES

- [1] Ryosuke Abe. Introducing autonomous buses and taxis: Quantifying the potential benefits in japanese transportation systems. *Transportation Research Part A: Policy and Practice*, 2019.
- [2] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access*, 6:52138–52160, 2018.
- [3] Elaine Angelino, Nicholas Larus-Stone, Daniel Alabi, Margo Seltzer, and Cynthia Rudin. Learning certifiably optimal rule lists for categorical data. *The Journal of Machine Learning Research*, 18(1):8753–8830, 2017.
- [4] Lisa Anne Hendricks, Ronghang Hu, Trevor Darrell, and Zeynep Akata. Grounding visual explanations. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 264–279, 2018.
- [5] Pietro Barbiero, Gabriele Ciravegna, Dobrik Georgiev, and Francesco Giannini. Pytorch, explain! a python library for logic explained networks. 2021.
- [6] J Basak. Online adaptive decision trees. *Neural computation*, 16(9):1959–1981, 2004.
- [7] Osbert Bastani, Yewen Pu, and Armando Solar-Lezama. Verifiable reinforcement learning via policy extraction. In *NeurIPS*, 2018.
- [8] Osbert Bastani, Yewen Pu, and Armando Solar-Lezama. Verifiable reinforcement learning via policy extraction. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

- [9] Yoshua Bengio, Nicholas Léonard, and Aaron C. Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *ArXiv*, abs/1308.3432, 2013.
- [10] Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José M. F. Moura, and Peter Eckersley. Explainable machine learning in deployment, 2019.
- [11] Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and C. J. Stone. Classification and regression trees. 1983.
- [12] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.
- [13] Chaofan Chen and Cynthia Rudin. An optimization approach to learning falling rule lists. *arXiv preprint arXiv:1710.02572*, 2017.
- [14] Hongge Chen, Huan Zhang, Si Si, Yang Li, Duane S. Boning, and Cho-Jui Hsieh. Robustness verification of tree-based models. In *NeurIPS*, 2019.
- [15] Gabriele Ciravegna, Pietro Barbiero, Francesco Giannini, M. Gori, Pietro Li'o, Marco Maggini, and S. Melacci. Logic explained networks. *ArXiv*, abs/2108.05149, 2021.
- [16] Jiayun Cui, William Macke, Harel Yedidsion, Aastha Goyal, Daniel Urielli, and Peter Stone. Scalable multiagent driving policies for reducing traffic congestion. *ArXiv*, abs/2103.00058, 2021.
- [17] George V. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2:303–314, 1989.
- [18] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.
- [19] Nicholas Frosst and Geoffrey E. Hinton. Distilling a neural network into a soft decision tree. *ArXiv*, abs/1711.09784, 2017.
- [20] Scott Fujimoto, Herke van Hoof, and David Meger. Addressing function approximation error in actor-critic methods. *ArXiv*, abs/1802.09477, 2018.
- [21] Abhishek Ghose and Balaraman Ravindran. Interpretability with accurate small models. *Frontiers in Artificial Intelligence*, 3, 2020.
- [22] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press.
- [23] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pages 1861–1870. PMLR, 2018.
- [24] Danijar Hafner, Timothy P. Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering atari with discrete world models. *ArXiv*, abs/2010.02193, 2021.
- [25] Daniel Hein, Steffen Limmer, and Thomas A. Runkler. Interpretable control by reinforcement learning. *ArXiv*, abs/2007.09964, 2020.
- [26] Lisa Anne Hendricks, Ronghang Hu, Trevor Darrell, and Zeynep Akata. Generating counterfactual explanations with natural language. *arXiv preprint arXiv:1806.09809*, 2018.
- [27] Eric Jang, Shixiang Shane Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *ArXiv*, abs/1611.01144, 2017.
- [28] John M Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Zidek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A A Kohl, Andy Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David A. Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with alphafold. *Nature*, 596:583 – 589, 2021.
- [29] Christos Katrakazas, Mohammed A. Qudus, Wen-Hua Chen, and Lipika Deka. Real-time motion planning methods for autonomous on-road driving: State-of-the-art and future research directions. *Transportation Research Part C-emerging Technologies*, 60:416–442, 2015.
- [30] Guy Katz, Clark W. Barrett, David L. Dill, Kyle D. Julian, and Mykel J. Kochenderfer. Reluplex: An efficient smt solver for verifying deep neural networks. *ArXiv*, abs/1702.01135, 2017.
- [31] Been Kim. Interactive and interpretable machine learning models for human machine collaboration. 2015.
- [32] Jinkyu Kim and John F. Canny. Interpretable learning for self-driving cars by visualizing causal attention. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2961–2969, 2017.
- [33] Peter Kontschieder, Madalina Fiterau, Antonio Criminisi, and Samuel Rota Buló. Deep neural decision forests. In *Proceedings of the IEEE international conference on computer vision*, pages 1467–1475, 2015.
- [34] Himabindu Lakkaraju, Stephen H. Bach, and Jure Leskovec. Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1675–1684, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322. doi: 10.1145/2939672.2939874. URL <https://doi.org/10.1145/2939672.2939874>.
- [35] Dmitry Laptev and Joachim M Buhmann. Convolutional decision trees for feature learning and segmentation. In *German Conference on Pattern Recognition*, pages 95–106. Springer, 2014.
- [36] Benjamin Letham, Cynthia Rudin, Tyler H McCormick, David Madigan, et al. Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. *The Annals of Applied Statistics*, 9(3):1350–1371, 2015.

- [37] Edouard Leurent. An environment for autonomous driving decision-making. <https://github.com/eleurent/highway-env>, 2018.
- [38] Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Manfred Otto Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *CoRR*, abs/1509.02971, 2016.
- [39] Pantelis Linardatos, Vasilis Papastefanopoulos, and S. Kotsiantis. Explainable ai: A review of machine learning interpretability methods. *Entropy*, 23, 2021.
- [40] Zachary C Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, 2018.
- [41] Ofir Nachum, Shixiang Shane Gu, Honglak Lee, and Sergey Levine. Data-efficient hierarchical reinforcement learning. In *NeurIPS*, 2018.
- [42] Chris Olah, Arvind Satyanarayan, Ian Johnson, Shan Carter, Ludwig Schubert, Katherine Ye, and Alexander Mordvintsev. The building blocks of interpretability. *Distill*, 3(3):e10, 2018.
- [43] C Olaru and L Wehenkel. A complete fuzzy decision tree technique. *Fuzzy sets and systems*, 138(2):221–254, 2003.
- [44] Rohan Paleja, Muyleng Ghuy, Nadun Ranawaka Arachchige, Reed Jensen, and Matthew Gombolay. The utility of explainable ai in ad hoc human-machine teaming. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 610–623. Curran Associates, Inc., 2021. URL <https://proceedings.neurips.cc/paper/2021/file/05d74c48b5b30514d8e9bd60320fc8f6-Paper.pdf>.
- [45] Rohan R. Paleja, Andrew Silva, Letian Chen, and Matthew Gombolay. Interpretable and personalized apprenticeship scheduling: Learning interpretable scheduling policies from heterogeneous user demonstrations. 2020.
- [46] Ian Parberry. *Introduction to Game Physics with Box2D*. CRC Press, 2017.
- [47] Xue Bin Peng, Marcin Andrychowicz, Wojciech Zaremba, and P. Abbeel. Sim-to-real transfer of robotic control with dynamics randomization. *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1–8, 2018.
- [48] C. Ribeiro and M. Tulio. Model-agnostic explanations and evaluation of machine learning. 2019.
- [49] Stéphane Ross, Geoffrey J. Gordon, and J. Andrew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *AISTATS*, 2011.
- [50] C. Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1:206–215, 2018.
- [51] Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, and Chudi Zhong. Interpretable machine learning: Fundamental principles and 10 grand challenges. *ArXiv*, abs/2103.11251, 2021.
- [52] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *ArXiv*, abs/1707.06347, 2017.
- [53] Rastko R. Selmic and Frank L. Lewis. Neural-network approximation of piecewise continuous functions: application to friction compensation. *IEEE transactions on neural networks*, 13 3:745–51, 2002.
- [54] Andrew Silva and Matthew Craig Gombolay. Encoding human domain knowledge to warm start reinforcement learning. In *AAAI*, 2021.
- [55] Andrew Silva, M. Gombolay, Taylor W. Killian, Ivan Dario Jimenez Jimenez, and Sung-Hyun Son. Optimization methods for interpretable differentiable decision trees applied to reinforcement learning. In *AISTATS*, 2020.
- [56] Alberto Suárez and James F Lutsko. Globally optimal fuzzy decision trees for classification and regression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(12):1297–1311, 1999.
- [57] Ryutaro Tanno, Kai Arulkumaran, Daniel C. Alexander, Antonio Criminisi, and Aditya V. Nori. Adaptive neural trees. *arXiv preprint arXiv:1807.06699*, 2018.
- [58] Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5026–5033. IEEE, 2012.
- [59] Marko Vasic, Andrija Petrović, Kaiyuan Wang, Mladen Nikolic, Rishabh Singh, and Sarfraz Khurshid. Moet: Mixture of expert trees and its application to verifiable reinforcement learning. 2019.
- [60] Paul Voigt and Axel Von dem Bussche. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed.*, Cham: Springer International Publishing, 2017.
- [61] Sholom M. Weiss and Nitin Indurkha. Rule-based machine learning methods for functional prediction. *Journal of Artificial Intelligence Research*, 3:383–403, 1995.
- [62] Sean Wilson, Paul Glotfelter, Li Wang, Siddharth Mayya, Gennaro Notomista, Mark L. Mote, and Magnus Egerstedt. The robotarium: Globally impactful opportunities, challenges, and lessons learned in remote-access, distributed control of multirobot systems. *IEEE Control Systems*, 40:26–44, 2020.
- [63] Cathy Wu, Aboudy Kreidieh, Kanaad Parvate, Eugene Vinitsky, and Alexandre M. Bayen. Flow: Architecture and benchmarking for reinforcement learning in traffic control. *ArXiv*, abs/1710.05465, 2017.
- [64] Mike Wu, Michael C Hughes, Sonali Parbhoo, Maurizio Zazzi, Volker Roth, and Finale Doshi-Velez. Beyond sparsity: Tree regularization of deep models for interpretability. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
- [65] Mike Wu, S. Parbhoo, Michael C. Hughes, Ryan Kindle, Leo Anthony Celi, Maurizio Zazzi, Volker Roth, and

- Finale Doshi-Velez. Regional tree regularization for interpretability in deep neural networks. In *AAAI*, 2020.
- [66] Ekim Yurtsever, Jacob Lambert, Alexander Carballo, and K. Takeda. A survey of autonomous driving: Common practices and emerging technologies. *IEEE Access*, 8: 58443–58469, 2020.

APPENDIX

A. Node Crispification Algorithm

In this section, we provide a description of node crispification, displayed in Algorithm 2 and termed `NODE_CRISP` in the main paper. We display the transformation performed by node crispification by the green arrow in Figure 5. Node crispification recasts each decision node to split upon a single dimension of the input.

Algorithm 2 Node Crispification: `NODE_CRISP`(\cdot)

Input: The original fuzzy decision node $\sigma(\alpha(\vec{w}_i^T \vec{x} - b_i))$, where i is the decision node index, $\vec{w}_i = [w_i^1, w_i^2, \dots, w_i^j, w_i^{j+1}, \dots, w_i^m]^T$, and m is the number of input features

Output: The intermediate decision node representation $\sigma(\alpha(w_i^k x^k - b_i))$ (see the green box in Figure 5)

- 1: $\vec{z}_i = \text{DIFF_ARGMAX}(|\vec{w}_i|)$ ($\text{DIFF_ARGMAX}(\cdot)$ displayed in Algorithm 4)
 - 2: $\vec{w}_i' = \vec{z}_i \circ \vec{w}_i$
 - 3: $\sigma(\alpha(w_i^k x^k - b_i)) = \sigma(\alpha(\vec{w}_i'^T \vec{x} - b_i))$
-

Node crispification takes as input the original fuzzy decision node, $\sigma(\alpha(\vec{w}_i^T \vec{x} - b_i))$, where *all* input features are used in determining the output of decision node i . The output of this function is an intermediate decision node, $\sigma(\alpha(w_i^k x^k - b_i))$, where the output of decision node i is only determined by a *single* feature, x^k . To perform this transformation, in Line 1, we use the differentiable argument max function (in Algorithm 4) to produce a one-hot vector, \vec{z}_i , with the element associated with the most impactful feature set to one and all other elements set to zero. In Line 2, we element-wise multiply the one-hot encoding, \vec{z}_i , by the original weights, \vec{w}_i , to produce a new set of weights with only one active weight, \vec{w}_i' . In Line 3, we show that by multiplying \vec{x} by \vec{w}_i' , we can obtain the intermediate decision node $\sigma(\alpha(w_i^k x^k - b_i))$, where k is the index of the most impactful feature (i.e., $k = \arg \max_j (|w_i^j|)$).

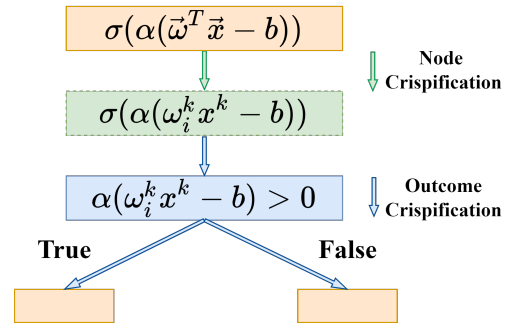


Fig. 5: This figure displays the process of our proposed differentiable crispification, including node crispification (Algorithm 2) and outcome crispification (Algorithm 3).

B. Outcome Crispification Algorithm

In this section, we provide a description of outcome crispification, displayed in Algorithm 3 and termed

OUTCOME_CRISP in the main paper. We display the transformation performed by outcome crispification by the blue arrows in Figure 5. Outcome crispification translates the outcome of a soft decision node to a hard decision node, resulting in a Boolean output from the decision node rather than a set of probabilities.

Algorithm 3 Outcome Crispification: OUTCOME_CRISP(\cdot)

Input: The intermediate decision node $\sigma(\alpha(w_i^k x^k - b_i))$, where i is the decision node index, $k = \arg \max_j (|w_i^j|)$, and w_i^j is the j th element in \vec{w}_i

Output: Crisp decision node $\mathbb{1}(\alpha(w_i^k x^k - b_i) > 0)$ (see the blue box in Figure 5)

- 1: $\vec{v}_i = [\alpha(w_i^k x^k - b_i), 0]$
 - 2: $\vec{z}_i' = \text{DIFF_ARGMAX}(\vec{v}_i)$ (DIFF_ARGMAX(\cdot) displayed in Algorithm 4)
 - 3: $\mathbb{1}(\alpha(w_i^k x^k - b_i) > 0) = \vec{z}_i'[0]$
-

Outcome crispification takes in the intermediate decision node $\sigma(\alpha(w_i^k x^k - b_i))$, which outputs the probability of branching left. The output of OUTCOME_CRISP is the crisp decision node, $\mathbb{1}(\alpha(w_i^k x^k - b_i) > 0)$, a Boolean decision to trace down to the left branch OR right. In Line 1, we construct a soft vector representation of the decision node i 's output, \vec{v}_i , by concatenating $\alpha(w_i^k x^k - b_i)$ with a 0. In Line 2, we use the differentiable argument max function (in Algorithm 4) to produce a one-hot vector, \vec{z}_i' , where the first element represents the Boolean outcome of the decision node. In Line 3, we show that the output of the crisp decision node, $\mathbb{1}(\alpha(w_i^k x^k - b_i) > 0)$, can be obtained by choosing the first element of vector \vec{z}_i' (we use bracket indexing notation here, starting with zero).

C. Differentiable Argument Max Function for Differentiable Crispification

In this section, we provide a description of the differentiable argument max function which is utilized in both decision node crispification and decision outcome crispification.

Algorithm 4 Differentiable Argument Max Function for Crispification: DIFF_ARGMAX(\cdot)

Input: Logits \vec{q}

Output: One-Hot Vector \vec{h}

- 1: $\vec{h}_{soft} \leftarrow f(\vec{q})$
 - 2: $\vec{h}_{hard} \leftarrow g(f(\vec{q}))$
 - 3: $\vec{h} = \vec{h}_{hard} + \vec{h}_{soft} - \text{STOP_GRAD}(\vec{h}_{soft})$
-

Similar to [24], we present a function call (in Algorithm 4) that can be utilized to maintain gradients over a non-differentiable one-hot operation. The function takes in a set of logits, \vec{q} , and applies a softmax operation, denoted by f , to output \vec{h}_{soft} , as shown in Line 1. In Line 2, the logits are transformed using a softmax followed by a one-hot procedure, g , causing the removal of gradient information, producing \vec{h}_{hard} . In line 3, we combine \vec{h}_{soft} , \vec{h}_{hard} , and $\text{STOP_GRAD}(\vec{h}_{soft})$ to output \vec{h} . The outputted value of \vec{h} is equal to that of \vec{h}_{hard} .

However, the gradient maintained within \vec{h} is that associated with \vec{h}_{soft} . Automatic differentiation frameworks can then utilize the outputted term to perform backpropagation.

D. Differentiable Crispification and Enforcing Sparsity with Gumbel-Softmax

In this section, we provide an ablation study where we utilize a Gumbel-Softmax [27] function to perform decision node crispification, perform decision outcome crispification, and enforce sub-controller sparsity. To use a Gumbel-Softmax function in our ICCTs, as opposed to a standard softmax function, requires modifying the function, f , introduced in Section 4 within the main paper to f' as follows:

$$f'(\vec{w}_i)_k = \frac{\exp\left(\frac{w_i^k + g_i^k}{\tau}\right)}{\sum_j^m \exp\left(\frac{w_i^j + g_i^j}{\tau}\right)} \quad (8)$$

Here, \vec{w}_i is a vector composed of m elements $w_i^1, w_i^2, \dots, w_i^m$, and $g_i^1, g_i^2, \dots, g_i^m$ are i.i.d samples from a Gumbel(0, 1) distribution [27]. Here, we compare the performance of ICCT-complete, ICCT-1-feature, and ICCT-2-feature to their variants using Gumbel-Softmax in Lunar Lander and Lane-Keeping. All the methods and their corresponding variants are trained using the same hyperparameters.

From the results shown in Figure 6 and Table II, we find that the addition of Gumbel noise reduces performance by a wide margin. Furthermore, comparing crisp ICCTs utilizing Gumbel-Softmax to ICCTs utilizing Gumbel-Softmax, we see that due to the sampling procedure within the Gumbel-Softmax, an inconsistency issue arises between non-crisp and crisp performance.

E. Universal Function Approximation

In this section, we provide a proof to show our ICCTs are universal function approximators, that is, can represent any decision surface given enough parameters. Our ICCT architecture consists of successive indicator functions, whose decision point lies among a single dimension of the feature space, followed by a linear controller to determine a continuous control output. For simplicity, we assume below that the leaf nodes contain static distributions. However, maintaining a linear controller at the leaves is more expressive and thus, the result below generalizes directly to ICCTs.

The decision-making of our ICCTs can be decomposed as a sum of products. In Equation 9, we display a computed output for a 4-leaf tree, where decision node outputs, y_i , are determined via Equation 1 of the main paper. Here, the sigmoid steepness, α is set to infinity (transforming the sigmoid function into an indicator function) resulting in hard decision points ($y_i \in \{0, 1\}$). Equation 9 shows that the chosen action is determined by computation of probability of reaching a leaf, y , multiplied by static tree weights maintained at the distribution, p .

$$\begin{aligned} ICCT(x) = & p_1(y_1 * y_2) + p_2(y_1 * (1 - y_2)) \\ & + p_3((1 - y_1) * y_3) + p_4 * ((1 - y_1) * (1 - y_3)) \end{aligned} \quad (9)$$

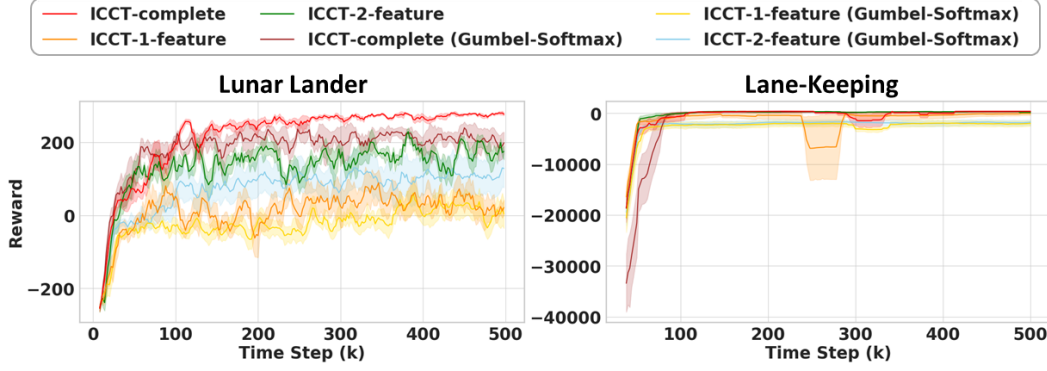


Fig. 6: This figure displays the average running rollout rewards of six methods for the ablation study during training. The results are averaged over 5 seeds, and the shadow region represents the standard error.

TABLE II: Performance Comparison between ICCTs utilizing the differentiable crispification procedure discussed in Section 4 within the main paper, ICCTs utilizing a Gumbel-Softmax differentiable crispification procedure, and crisp ICCTs utilizing a Gumbel-Softmax differentiable crispification procedure. Across each approach, we include ICCTs with fully parameterized sub-controllers (ICCT-complete) and sparse sub-controllers. We present our findings across Lunar Lander and Lane-Keeping.

Method	Lunar Lander	Lane-Keeping
ICCT-complete	300.5 ± 1.2	476.6 ± 3.1
ICCT-complete (Gumbel-Softmax)	276.7 ± 7.0	412.6 ± 31.3
ICCT-complete (Gumbel-Softmax, Crisp)	239.0 ± 18.9	309.1 ± 94.6
ICCT-1-feature	190.1 ± 13.7	437.6 ± 7.0
ICCT-1-feature (Gumbel-Softmax)	113.2 ± 43.1	-853.4 ± 333.2
ICCT-1-feature (Gumbel-Softmax, Crisp)	-20.1 ± 50.0	-658.114 ± 345.3
ICCT-2-feature	258.4 ± 7.0	458.5 ± 6.3
ICCT-2-feature (Gumbel-Softmax)	161.7 ± 54.8	-560.6 ± 251.6
ICCT-2-feature (Gumbel-Softmax, Crisp)	62.3 ± 82.2	-945.0 ± 331.0

Equation 9 can be directly simplified into the form of $G(x) = \sum_{j=1}^N p_j \sigma(w_j^T x + b_j)$, similar to Equation 1 in [17]. [17] demonstrates that finite combination of fixed, univariate functions can approximate any continuous function. The key difference between our architecture is that our univariate function is an indicator function rather than the commonly used sigmoid function. Below, we provide two lemmas to show that indicator functions fall within the space of univariate functions [17].

Lemma 1: An indicator function is sigmoidal.

Proof: This follows from the definition of sigmoidal: $\sigma(t) \rightarrow 1$ as $t \rightarrow \infty$ and $\sigma(t) \rightarrow 0$ as $t \rightarrow -\infty$.

Lemma 2: An indicator function is discriminatory.

Proof: As an indicator function is bounded and measurable, by Lemma 1 of [17], it is discriminatory.

Theorem 1: Let σ be any continuous discriminatory function. ICCTs are universal function approximators, that is, dense in the space of $C(I_n)$. In other words, there is a representation of ICCTs, $I(x)$, for which $|I(x) - f(x)| < \epsilon$ for all $x \in I_n$, for any function, f ($f \in C(I_n)$), where $C(I_n)$ denotes the codomain of an n -dimensional unit cube, I_n .

Proof: As the propositional conditions hold for Theorem 1 in [17], the result that ICCTs are dense in $C(I_n)$ directly follows. We note that as the indicator function jump-continuous, we refer readers to [53] whom extend UFA for

$G(x) = \sum_{j=1}^N p_j \sigma(w_j^T x + b_j)$ to the case when σ is jump-continuous.

F. Learning Curves

In Figure 7, we display the learning curves of the eleven methods across six domains shown in Table 1 of the main paper. In general, ICCT-complete has competitive or better performance with regards to running-average rollout rewards and convergence rate, compared to MLP-Max, MLP-Upper and fuzzy DDTs in Inverted Pendulum, Lunar Lander, Lane-Keeping, Sing-Lane Ring, and Multi-Lane Ring, while maintaining interpretability. We also notice as the sparsity of the linear sub-controller increases, the performance of ICCT gradually drops. However, the interpretable approaches of ICCT-3-feature and ICCT-2-feature still have comparable or better performance with respect to MLP-Lower and ICCT-L1-sparse.

G. Dynamic Depth

In this section, we present a dynamic deepening algorithm here drawing inspiration from [54]. Allowing our ICCTs to automatically deepen and increase in complexity has several advantages. In continuous control, it is typical that a deployed policy may encounter a covariate shift in a real-world setting [47]. As such, our ICCT may need to change to account for features previously thought unimportant. Dynamic deepening

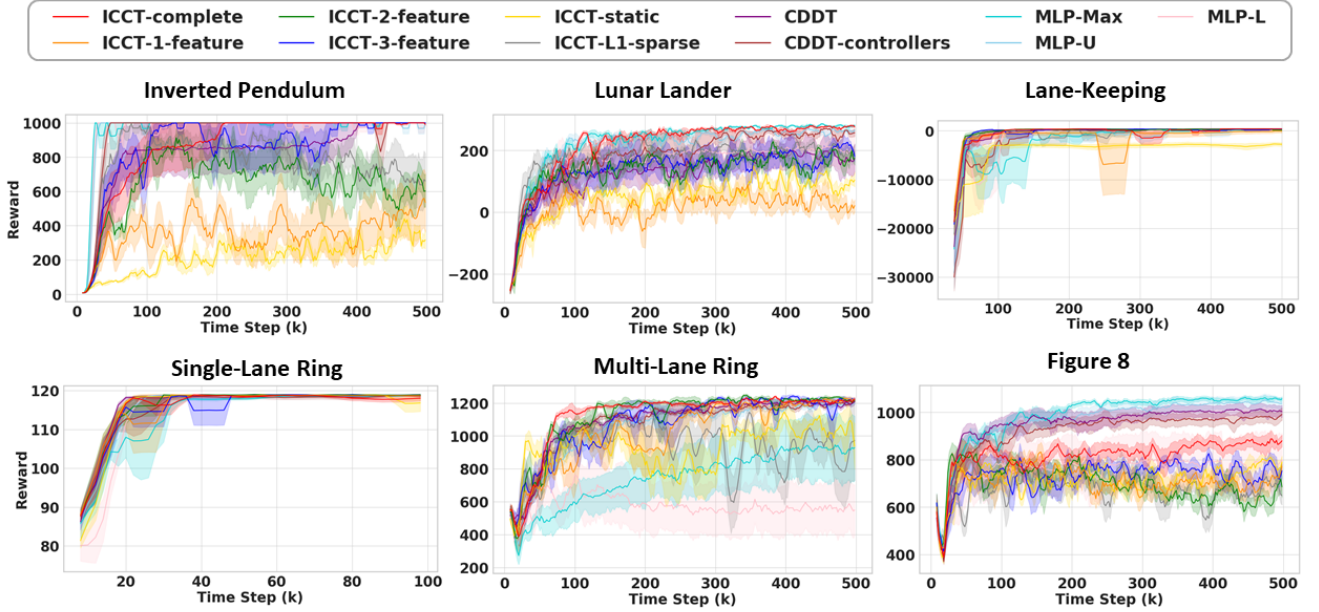


Fig. 7: In this figure, we display the learning curves of eleven methods in six domains shown in the main paper. The rewards are rollout rewards throughout the training process. The curves in 5 domains except Lane-Keeping are smoothed by a sliding window of size 5, the curves in Lane-Keeping are smoothed by a sliding window of size 20. The shadow region depicts the standard error across 5 seeds.

would allow our ICCTs to mitigate the encountered covariate shift and represent additional complexities when deployed. Furthermore, the ability to deepen eliminates the need to set the tree depth a priori.

Algorithm 5 Dynamic Deepening Procedure

Input: Pretrained ICCT P , deepened ICCT P_{deep} , controller sparsity e

```

1: for  $i$  epochs do
2:   Collect trajectory rollouts,  $\tau$ , with  $P$ 
3:    $P, P_{deep} \leftarrow \text{NETWORK\_UPDATE}(P, P_{deep})$ 
4:    $\vec{H}, \vec{H}_{deep} \leftarrow \text{CALCULATE\_LEAF\_ENTROPIES}(P, P_{deep}, \tau)$ 
5:   if  $H(P_{deep}) + \epsilon < H(P)$  then
6:      $P \leftarrow P_{deep}$ 
7:      $P_{deep} \leftarrow \text{DEEPEN}(P_{deep})$ 
8:   end if
9: end for

```

Our proposed procedure for deepening is shown in Algorithm 5. This procedure should be conducted after a ICCT model, P , has been pretrained with an initial dataset (e.g., with simulated data for transfer to the real world). At initialization of the dynamic deepening procedure, two ICCT models are maintained, a shallow pre-trained version, P , and deeper-by-one-depth version P_{deep} . Both models utilize the same controller sparsity e . The deepened ICCT is initialized so that the top-level weights match that of P , and lower-level weights are randomly initialized. During deployment, the pre-trained ICCT model, P is utilized to collect trajectory rollouts, τ , as shown in Line 1 of Algorithm 5. Given these trajectory rollouts and associated rewards, both models are updated via gradient descent, as shown by the function `NETWORK_UPDATE` in Line

3. P is updated via Equation 2 in our main paper. As we do not have access to rollout trajectories for P_{deep} , the model update is simulated by utilizing the likelihood that P_{deep} will take similar actions to P given the states within τ . In Line 4 of Algorithm 5 (`CALCULATE_LEAF_ENTROPIES`), we calculate the entropy across each leaf within P and P_{deep} . As our leaf nodes are input-parameterized (based on state), we utilize the sample of trajectories collected in Line 2 to estimate the leaf entropy. Here, \vec{H} and \vec{H}_{deep} represent a vector of leaf entropies. The deeper ICCT has more leaves (generated via the deepening procedure) and, thus, \vec{H}_{deep} is a higher-dimensional vector. In Line 5 of Algorithm 5, we compare the entropy of adjacent leaf nodes between P and P_{deep} . For example, in the case where we have P as a two-leaf tree and P_{deep} as a four-leaf tree, if the entropy of the left leaf node of P is at least ϵ greater than that of the combined entropy of the two left leaf nodes of P_{deep} , P_{deep} has learned a leaf distribution that is more precise in representing high-performance control behavior. Thus, in Line 6 and 7 of Algorithm 5, the shallow model, P , is updated the additional leaves of the deeper model, P_{deep} , and the deeper model, P_{deep} , is deepened by an additional level for each decision tree path that had lower entropy (determined in Line 5). This procedure continues for a set number of predefined epochs.

H. Environments

We provide detailed descriptions for the two common continuous control problems: Inverted Pendulum and Lunar Lander, and autonomous driving domains: Lane-Keeping and those from Flow, including Single-Lane Ring, Multi-Lane

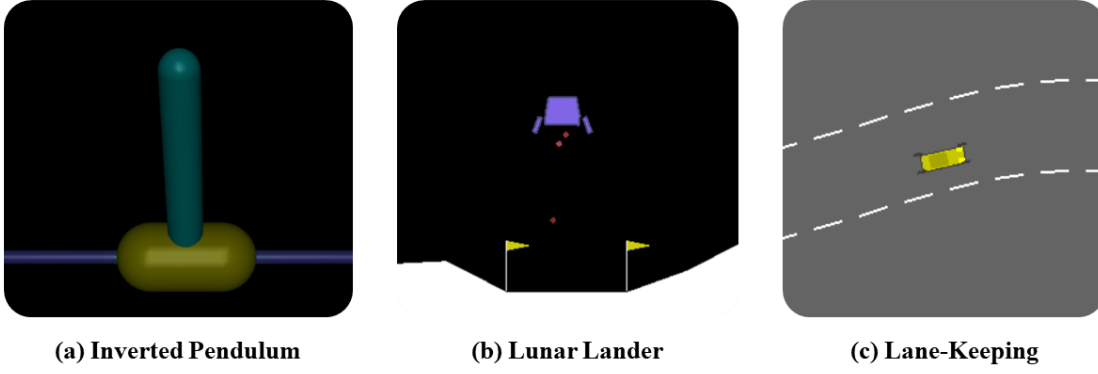


Fig. 8: This figure displays the first three environments we utilized including Inverted Pendulum, Lunar Lander, and Lane-Keeping.

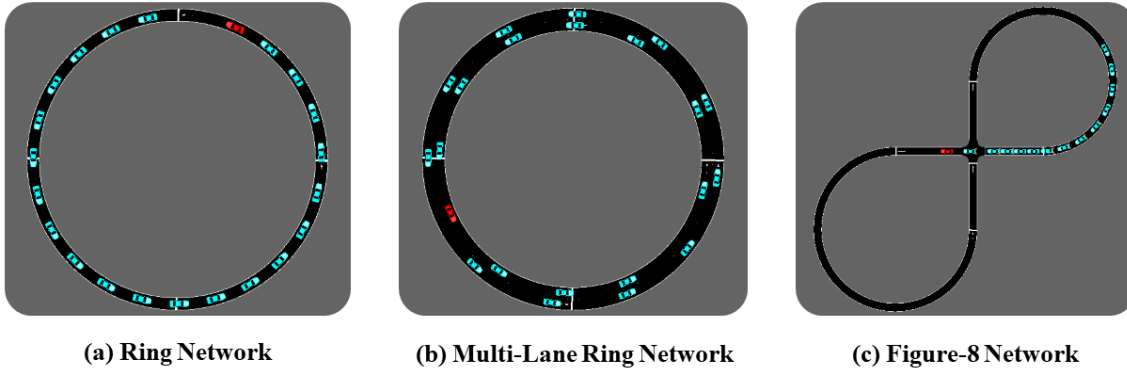


Fig. 9: In this figure, we display the configuration of each Flow network used in our experimentation. In each image, the red vehicle is the autonomous agent that utilizes the learned control policy to stabilize traffic flow. Each cyan vehicle is a simulated human vehicle which contains a noisy acceleration behavior, the severity of which is defined by the user.

Ring, and Figure 8.

Inverted Pendulum: Inverted Pendulum (Figure 8(a)) is provided by MuJoCo [58] and OpenAI Gym [12]. The observation includes the cart position, velocity, pole angle, and pole angular velocity. The goal is to apply a force to the cart to balance a pole on it and prevent the pole from falling. A plus one reward is provided at each timestep provided if the pole keeps upright at each time step.

Lunar Lander: Lunar Lander (Figure 8(b)) is a game provided by Box2D [46] and OpenAI Gym. The goal is to land a lunar lander as close to a landing pad between the flags. The observation is 8-dimensional including the lander’s current position, linear velocity, tilt, angular velocity, and information about ground contact. The continuous action space is two dimensional for controlling the main engine thruster and side thrusters. At each timestep, the lander reward is determined by a proximity to the landing pad, whether each leg is touching the landing pad, and a fuel cost. The episode ends if the lander crashes and a terminal reward of -100 is provided. If the lander successfully lands, a terminal reward of 100 is provided.

Lane-Keeping: Lane-Keeping (Figure 8(c)) is a domain with continuous actions within highway-env [37]. The observation is 12-dimensional, which consists of the vehicle’s lateral

position, heading, lateral speed, yaw rate, linear, lateral, and angular velocity, and the lane information. The action is the steering angle to control the vehicle. At each time step, +1 reward will be provided if the vehicle can keep in the center of the lane. The reward decreases as the vehicle drives away from the lane center. The terminal condition within this domain is a maximum timestep of 500.

Flow Domains [63]: Each of the following domains are custom continuous actions domains provided within the Flow deep reinforcement learning framework for mixed autonomy traffic scenarios. Flow utilizes the SUMO traffic simulator, which allows for both autonomous agents and simulated human agents. The simulated human agents are constructed by adjusting the noise factor present within their acceleration and deceleration control. Stabilization of traffic is defined as the average velocity of all vehicles approaching a set value for velocity. The general observation includes the global positions and velocity of all vehicles in the network. These are normalized based upon the length and max velocity for the network, which is defined by the user, and then combined into a single 1-Dimensional array with a length that is double the number of vehicles present. The reward is measured by how closely the network’s average velocity matches the user-

defined average velocity. The episode is terminated if any collision between two vehicles is detected, or preset time steps are executed.

Flow Single-Lane Ring Network: The Single-Lane Ring Network (Figure 9(a)) is a ring road network, with the objective being to stabilize the flow of all traffic within the network. A control policy must apply acceleration commands to an autonomous agent in order to stabilize the flow. There are 21 human vehicles and 1 ego vehicle, with a maximum time step of 750.

Flow Multi-Lane Ring Network: The Multi-Lane Ring network (Figure 9(b)) is a complex ring road network consisting of multiple lanes of traffic, with the objective being to stabilize the flow of all traffic within the network. The observation for this network includes the general observation for Flow networks, as well as the current lane index for each vehicle. The network holds 21 human vehicle and 1 ego vehicle. A control policy must apply acceleration commands and lane changing commands (a continuous value from -1 to 1) to the ego vehicle in order to stabilize the flow of traffic between both lanes. The maximum step in this domain is set to 1500.

Flow Figure-8 Network: The Figure-8 Network (Figure 9(c)) is a complex Flow domain as it contains a road section where the vehicles will cross-over, which is in the center of the figure-8. The control policy must consider this congestion point when controlling the ego vehicle, applying acceleration commands to the autonomous agent in order to stabilize the flow. The network holds 13 human vehicles besides the ego vehicle. Here the maximum time step is 1500.

I. Hyperparameters

All methods are trained using SAC with the same structure of critic network, which consists of an MLP with 2 hidden layers of 256 units. The buffer size is always set to 1000000, and the discount factor γ is always set to 0.99. In the Single-Lane Ring network, the training steps are 100000, otherwise the training steps are 500000 for all the methods. The soft update coefficient, τ , is set to 0.005 for ICCT-static in Inverted Pendulum, and set to 0.01 in all other domains. We display the learning rates, batch sizes, and network sizes used for the methods discussed in the paper across each domain in Tables III-VIII.

TABLE III: Hyperparameters in Inverted Pendulum

Hyperparameter	Learning Rate	Batch Size	Network Size
ICCT-complete	3×10^{-4}	1024	2 leaves
ICCT-1-feature	6×10^{-4}	1024	8 leaves
ICCT-2-feature	5×10^{-4}	1024	4 leaves
ICCT-3-feature	5×10^{-4}	1024	2 leaves
ICCT-static	5×10^{-4}	1024	32 leaves
ICCT-L1-sparse	3×10^{-4}	256	4 leaves
CDDT	3×10^{-4}	1024	2 leaves
CDDT-controllers	3×10^{-4}	1024	2 leaves
MLP-Max	3×10^{-4}	1024	[256, 256]
MLP-U	3×10^{-4}	1024	[8, 8]
MLP-L	3×10^{-4}	1024	[6, 6]

TABLE IV: Hyperparameters in Lunar Lander

Hyperparameter	Actor Learning Rate	Critic Learning Rate	Batch Size	Network Size
ICCT-complete	5×10^{-4}	5×10^{-4}	256	8 leaves
ICCT-1-feature	5×10^{-4}	5×10^{-4}	256	8 leaves
ICCT-2-feature	5×10^{-4}	5×10^{-4}	256	8 leaves
ICCT-3-feature	5×10^{-4}	5×10^{-4}	256	8 leaves
ICCT-static	5×10^{-4}	3×10^{-4}	256	32 leaves
ICCT-L1-sparse	5×10^{-4}	5×10^{-4}	256	8 leaves
CDDT	5×10^{-4}	3×10^{-4}	256	8 leaves
CDDT-controllers	5×10^{-4}	3×10^{-4}	256	8 leaves
MLP-Max	3×10^{-4}	3×10^{-4}	256	[256, 256]
MLP-U	3×10^{-4}	3×10^{-4}	256	[10, 10]
MLP-L	3×10^{-4}	3×10^{-4}	256	[6, 6]

TABLE V: Hyperparameters in Lane-Keeping

Hyperparameter	Learning Rate	Batch Size	Network Size
ICCT-complete	3×10^{-4}	1024	16 leaves
ICCT-1-feature	3×10^{-4}	1024	16 leaves
ICCT-2-feature	3×10^{-4}	1024	16 leaves
ICCT-3-feature	3×10^{-4}	1024	16 leaves
ICCT-static	2×10^{-4}	1024	16 leaves
ICCT-L1-sparse	3×10^{-4}	1024	16 leaves
CDDT	3×10^{-4}	256	16 leaves
CDDT-controllers	3×10^{-4}	512	16 leaves
MLP-Max	3×10^{-4}	256	[256, 256]
MLP-U	3×10^{-4}	256	[14, 14]
MLP-L	3×10^{-4}	256	[6, 6]

TABLE VI: Hyperparameters in Sing-Lane Ring Network

Hyperparameter	Learning Rate	Batch Size	Network Size
ICCT-complete	5×10^{-4}	1024	16 leaves
ICCT-1-feature	5×10^{-4}	1024	16 leaves
ICCT-2-feature	5×10^{-4}	1024	16 leaves
ICCT-3-feature	5×10^{-4}	1024	16 leaves
ICCT-static	5×10^{-4}	1024	16 leaves
ICCT-L1-sparse	5×10^{-4}	1024	16 leaves
CDDT	5×10^{-4}	1024	16 leaves
CDDT-controllers	5×10^{-4}	1024	16 leaves
MLP-Max	3×10^{-4}	1024	[256, 256]
MLP-U	3×10^{-4}	1024	[12, 12]
MLP-L	3×10^{-4}	1024	[3, 3]

TABLE VII: Hyperparameters in Multi-Lane Ring Network

Hyperparameter	Learning Rate	Batch Size	Network Size
ICCT-complete	5×10^{-4}	1024	16 leaves
ICCT-1-feature	5×10^{-4}	1024	16 leaves
ICCT-2-feature	6×10^{-4}	1024	16 leaves
ICCT-3-feature	5×10^{-4}	1024	16 leaves
ICCT-static	5×10^{-4}	1024	16 leaves
ICCT-L1-sparse	5×10^{-4}	1024	16 leaves
CDDT	5×10^{-4}	1024	16 leaves
CDDT-controllers	5×10^{-4}	1024	16 leaves
MLP-Max	5×10^{-4}	1024	[256, 256]
MLP-U	5×10^{-4}	1024	[32, 32]
MLP-L	5×10^{-4}	1024	[3, 3]

TABLE VIII: Hyperparameters in Figure-8 Network

Hyperparameter	Learning Rate	Batch Size	Network Size
ICCT-complete	5.5×10^{-4}	1024	16 leaves
ICCT-1-feature	6×10^{-4}	1024	16 leaves
ICCT-2-feature	6×10^{-4}	1024	16 leaves
ICCT-3-feature	7×10^{-4}	1024	16 leaves
ICCT-static	5×10^{-4}	1024	16 leaves
ICCT-L1-sparse	5×10^{-4}	1024	16 leaves
CDDT	5×10^{-4}	1024	16 leaves
CDDT-controllers	5×10^{-4}	1024	16 leaves
MLP-Max	5×10^{-4}	1024	[256, 256]
MLP-U	5×10^{-4}	1024	[20, 20]
MLP-L	5×10^{-4}	1024	[3, 3]