CAHSOR: Competence-Aware High-Speed Off-Road Ground Navigation in $\mathbb{SE}(3)$

Anuj Pokhrel, Mohammad Nazeri, Aniket Datar, and Xuesu Xiao

***Abstract*— While the workspace of traditional ground vehicles is usually assumed to be in a 2D plane, i.e., $\mathbb{SE}(2)$, such an assumption may not hold when they drive at high speeds on unstructured off-road terrain: High-speed sharp turns on high-friction surfaces may lead to vehicle rollover; Turning aggressively on loose gravel or grass may violate the non-holonomic constraint and cause significant lateral sliding; Driving quickly on rugged terrain will produce extensive vibration along the vertical axis. Therefore, most offroad vehicles are currently limited to drive only at low speeds to assure vehicle stability and safety. In this work, we aim at empowering high-speed off-road vehicles with competence awareness in $\mathbb{SE}(3)$ so that they can reason about the consequences of taking aggressive maneuvers on different terrain with a 6-DoF forward kinodynamic model. The model is learned from visual and inertial Terrain Representation for Off-road Navigation (TRON) using multimodal, self-supervised vehicle-terrain interactions. We demonstrate the efficacy of our Competence-Aware High-Speed Off-Road (CAHSOR) navigation approach on a physical ground robot in both an autonomous navigation and a human shared-control setup and show that CAHSOR can efficiently reduce vehicle instability by 62% while only compromising 8.6% average speed with the help of TRON.**

I. INTRODUCTION

Autonomous mobile robot navigation has been a research topic in the robotics community for decades. Being equipped with perception, planning, and control techniques, different types of ground robots, e.g., differential-drive or Ackermann-steering, are able to efficiently move toward their goals in their 2D workspaces considering their 3-DoF motion models (x, y, and yaw) without colliding with obstacles, mostly in structured and homogeneous environments.

Bringing those robots into the unstructured real world, researchers have also investigated off-road navigation. Most off-road robots drive at slow speeds to assure vehicle stability and safety. Even when aiming at driving fast, they still assume a simplified 2D workspace and 3-DoF model in $\mathbb{SE}(2)$ despite the highly likely disturbances from the off-road terrain on other dimensions of the state space (e.g., drift along y, roll around x, or bumpiness along z). These realistic kinodynamic effects may be tolerable in some cases, but may lead to catastrophic consequences in others with increasing speed on unstructured terrain (Fig. 1).

To enable safe and robust off-road navigation, high-speed ground robots need to be competence-aware, i.e., knowing what is the consequence of taking an aggressive maneuver on different off-road terrain. For example, a sharp turn on high-friction pavement may lead to vehicle rollover (Fig. 1 top); Blasting through rugged surfaces can generate extensive vertical vibrations and damage onboard components (Fig. 1 bottom left); Aggressive swerving on loose grass or gravel will cause the vehicle to slide sideways and risk collision or falling off a cliff (Fig. 1 bottom right).

All authors are with the Department of Computer Science, George Mason University {apokhre, mnazerir, adatar, xiao}@gmu.edu

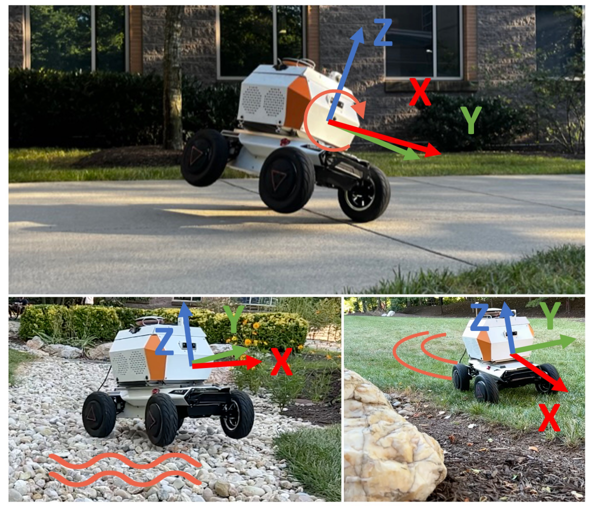

Fig. 1: Challenges of High-Speed Off-Road Ground Navigation in $\mathbb{SE}(3)$.

To this end, we propose a Competence-Aware High-Speed Off-Road (CAHSOR) ground navigation approach based on a 6-DoF forward kinodynamic model in $\mathbb{SE}(3)$. The model is learned as a downstream task of a new Terrain Representation for Off-road Navigation (TRON) approach with multimodal, self-supervised learning using viewpoint-invariant visual terrain patches and underlying Inertia Measurement Unit (IMU) responses during vehicle-terrain interactions. CAHSOR learns to predict potential next states according to different candidate actions and the current visual and/or inertial terrain representation to make competence-aware decisions in order to maximize speed while satisfying 6-DoF vehicle stability constraints in $\mathbb{SE}(3)$, e.g., without excessive sliding and rolling motions or bumpy vibrations. Our contributions can be summarized as:

- a TRON approach with multimodal self supervision that allows onboard visual and inertial observations to augment each other and maximizes the information embedded in the representation of each perceptual modality;
- a comprehensive study of various end-to-end and representation learning techniques with different modalities for different off-road kinodynamic modeling tasks;

- a CAHSOR framework for high-speed off-road vehicles to take aggressive maneuvers with stability and safety; and
- a set of real-world, off-road robot experiments to demonstrate the effectiveness of CAHSOR based on TRON in both an autonomous navigation and a human shared-control setup, exhibiting 62% vehicle instability reduction while only compromising 8.6% average speed.

II. APPROACH

We formulate the problem of forward kinodynamics modeling in $\mathbb{SE}(3)$ for ground robots driving on unstructured off-road terrain at high speeds, present a multimodal self-supervised learning approach to represent off-road conditions using onboard visual and inertial observations, introduce a data-driven approach to learn the forward kinodynamic model from past vehicle-terrain interactions, and develop a competence-aware navigation framework that allows robots to drive at the maximum possible speed while maintaining vehicle stability in $\mathbb{SE}(3)$.

A. Forward Ground Kinodynamics in $\mathbb{SE}(3)$

We adopt a forward kinodynamics formulation where we denote vehicle state as $\mathbf{s}$, which includes 6-DoF vehicle pose in $\mathbb{SE}(3)$ (x, y, z, roll r, pitch p, and yaw ϕ, expressed in the global or robot frame) and their corresponding velocity components. For brevity, only the pose components are included in the following derivation. The vehicle control $\mathbf{u} = [v, \omega]^T$ contains linear velocity and angular velocity (for differential-driven vehicles, or steering curvature for Ackermann-steering vehicles). We use a world state $\mathbf{w}$ to denote all necessary effects from the environment that will affect kinodynamics, in our case, from unstructured off-road terrain. Therefore, in a discrete setting, we have

$$\mathbf{s}_{t+1} = f(\mathbf{s}_t, \mathbf{u}_t, \mathbf{w}_t), \quad \mathbf{o}_t = g(\mathbf{s}_t, \mathbf{w}_t),$$
$$\mathbf{s}_t = [x_t, y_t, z_t, r_t, p_t, \phi_t]^T \in \mathbb{SE}(3), \ \mathbf{u}_t = [v_t, \omega_t]^T \in \mathbb{R}^2,$$

where $f(\cdot)$ is a forward kinodynamic function in $\mathbb{SE}(3)$, while $g(\cdot)$ is an observation function. For off-road driving, the forward kinodynamic function $f(\cdot)$ also takes in the world state $\mathbf{w}$ as input, which aids the robot to navigate the complexities of unstructured terrain. However, world state $\mathbf{w}$ is usually not directly observable and cannot be easily modeled.

B. Visual and Inertial Representation of World State

We use a multimodal self-supervised learning approach to represent the world state $\mathbf{w}$ and approximate the observation function $g(\cdot)$ with onboard visual and inertial sensors. A camera is used to provide the visual signature of the terrain patch λ_t. An IMU is used to sense the underlying kinodynamic responses i_t and GPS is used to observe the vehicle speed. For navigation, the vehicle may need to reason about future kinodynamic responses, for which only the visual observations from forward-facing camera are available. So in this work, we use multimodal self-supervised learning to allow both visual and inertial observations to augment each other by correlating them in effective representation spaces, thus either (or both) can be used to enable competence awareness when available (e.g., manual shared-control using current underlying inertia and autonomous planning with vision of future terrain).

We posit that the visual and inertial observations can provide multimodal self-supervised learning signals to represent different terrain kinodynamics. To achieve such self-supervision, we use a non-contrastive approach to maximize the correlation between visual and inertial embeddings. However, a key difference of high-speed off-road navigation compared to existing terrain representation learning approaches is that the correlation between vision and inertia is also dependent on the (high) vehicle speed: different speeds on grass vs. gravel may coincidentally lead to similar IMU readings. Therefore, CAHSOR extends the vision–inertia correlation to vision & speed–inertia correlation to account for the effect caused by various speeds during high-speed off-road navigation.

We design a viewpoint-invariant visual patch extraction technique to tackle the problem of visual perception being sensitive to changes in viewpoints, and lighting and also the problem of extracting the visual signature of terrain underneath the current robot state $\mathbf{s}_t$

We denote the camera image captured h time steps ahead as c_{t-h} and the transformation from time step $t-h$ to t extracted from vehicle odometry as d^t_{t-h}. By projecting c_{t-h} to an overhead Bird-Eye View (BEV) using the camera homography $h_{t-h} = H(i_{t-h})$, we can extract the terrain patch currently underneath the robot, $\lambda_t = P(c_{t-h}, h_{t-h}, d^t_{t-h})$. λ_t is designed to be slightly larger than the vehicle footprint to consider actuation latency.

C. Terrain Representation for Off-road Navigation

The set of viewpoint-invariant visual terrain patches $\Lambda_t = \{\lambda^j_t\}^J_{j=1}$, the IMU readings i_t, and the current vehicle speed s_t correspond to multimodal perception of the robot at time t and provide self-supervised learning signals for our vision & speed–inertia correlation (Fig. 2 left). To be specific, a vision, speed, and inertia encoder embeds any terrain patch $\lambda_t \in \Lambda_t$, current vehicle speed s_t, and underlying inertial readings i_t into a visual, speed, and inertial representation, ψ^{V}_t, ψ^{S}_t, and ψ^{I}_t, respectively. Considering the causal relation from driving at a particular speed on a certain visual terrain patch to corresponding IMU readings, we concatenate the visual and speed representations, ψ^{V}_t and ψ^{S}_t, and further encode them into a joint vision & speed embedding ψ^{VS}_t. To correlate ψ^{VS}_t and ψ^I_t, we project them independently into a higher dimensional feature space, ρ^{VS}_t and ρ^I_t. We then maximize the correlation between ρ^{VS}_t and ρ^I_t while considering viewpoint invariance using Barlow Twins:

$$\mathcal{L}_{\mathrm{TRON}} = \mathcal{BL}_{\mathrm{V}}(\rho^{\mathrm{V^1S}}_t, \rho^{\mathrm{V^2S}}_t) + \\ (0.5 \times \mathcal{BL}_{\mathrm{V^1I}}(\rho^{\mathrm{V^1S}}_t, \rho^I_t) + 0.5 \times \mathcal{BL}_{\mathrm{V^2I}}(\rho^{\mathrm{V^2S}}_t, \rho^I_t)), \tag{1}$$

where V^1 and V^2 correspond to two views of the same terrain patch to encourage viewpoint invariance, i.e., $\lambda^1_t, \lambda^2_t \sim \Lambda_t$.

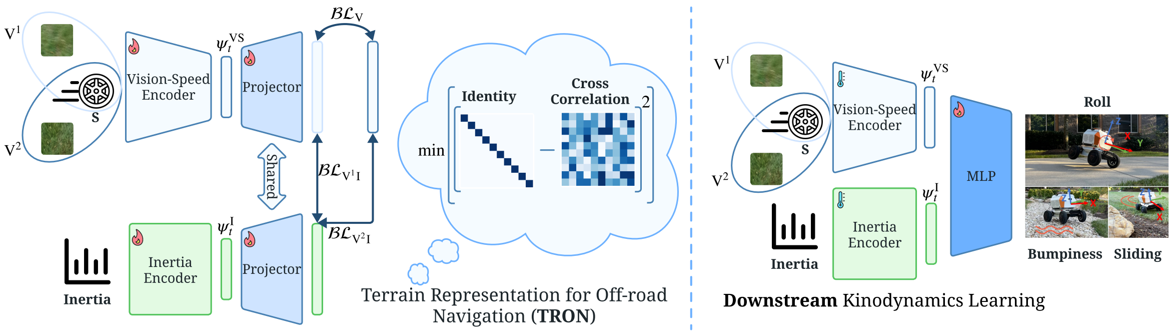

Fig. 2: TRON (Left) and Downstream Kinodynamics Learning (Right) Architecture: Flame and temperature denote training and frozen parameters respectively.

$\mathcal{BL}$ is defined as:

$$\mathcal{BL} = \sum_i (1 - \mathcal{C}_{ii})^2 + \gamma \sum_i \sum_{j \neq i} \mathcal{C}_{ij}^2, \tag{2}$$

where γ is a weight term to trade off the importance between invariance and redundancy reduction. $\mathcal{C}$ is the cross-correlation matrix computed between ρ^1 and ρ^2:

$$\mathcal{C}_{ij} = \frac{\sum_b \rho_{b,i}^1 \rho_{b,j}^2}{\sqrt{\sum_b (\rho_{b,i}^1)^2} \sqrt{\sum_b (\rho_{b,j}^2)^2}},$$

where $\rho_{b,i}^{1(2)}$ denotes the ith dimension of the bth sample in a data batch of $\rho^{1(2)}$, which can be one of $\rho_t^{\text{V}^1\text{S}}$, $\rho_t^{\text{V}^2\text{S}}$, or ρ_t^{I}.

Trained with multimodal self-supervision, the visual, speed, and inertial representation, ψ_t^{V}, ψ_t^{S}, and ψ_t^{I}, can be used to enable downstream kinodynamic modeling tasks. Depending on the scenario, either ψ_t^{V} (predicting multiple future states without terrain interactions to induce inertial responses) or ψ_t^{I} (directly predicting the immediate next state from the induced inertial responses from the underlying terrain), or both, may be available. For simplicity, we denote our visual-speed, inertial, or visual-speed-inertial representation as $\psi_t^{\text{V, S, I}}$.

D. Downstream Kinodynamic Model Learning

After learning the terrain representation and freezing the learned parameters, we also adopt a self-supervised approach to learn the forward kinodynamics due to the difficulty in analytically modeling $f(\cdot)$. We represent the unknown world state $\mathbf{w}_t$ using $\psi_t^{\text{V, S, I}}$ and learn an approximate forward kinodynamic function $f_\theta(\cdot)$ as a downstream task of TRON:

$$\mathbf{s}_{t+1} = f_\theta(\mathbf{s}_t, \mathbf{u}_t, \psi_t^{\text{V, S, I}}). \tag{3}$$

With a self-supervised vehicle-terrain interaction dataset,

$$\mathcal{D} = \{\mathbf{s}_{j+1}, \mathbf{s}_j, \mathbf{u}_j, \psi_j^{\text{V, S, I}}\}_{j=0}^{N-1},$$

of N data points, the optimal parameters θ^* can then be learned by minimizing a supervised loss function (Fig. 2 right),

$$\theta^* = \underset{\theta}{\operatorname{argmin}} \sum_{\substack{(\mathbf{s}_{j+1}, \mathbf{s}_j, \mathbf{u}_j, \psi_j^{\text{V, S, I}}) \\ \in \mathcal{D}}} ||\mathbf{s}_{j+1} - f_\theta(\mathbf{s}_j, \mathbf{u}_j, \psi_j^{\text{V, S, I}})||. \tag{4}$$

E. Competence-Aware High-Speed Off-Road Navigation

The approximate forward kinodynamic function (Eqn. (3)) learned with the self-supervised loss (Eqn. (4)) can be combined with subsequent planners to enable competence-awareness.

By rolling out the forward kinodynamic model, the robot can pick the optimal control command(s) that produces minimal cost or is most similar to human control, without violating vehicle stability constraints. While a motion planner or a human controller needs to consider a variety of costs including obstacle avoidance, goal distance, execution accuracy, etc., for simplicity, we combine all these costs into one general cost term $C(\mathbf{s}_t, \mathbf{s}_{t+1})$ and use only one time-step rollout in our presentation in order to explicitly showcase the high speed and competence awareness aspect of the navigation problem. Otherwise, the robot is solely maximizing navigation speed or following human control. Notice that it is easy to combine it with any other costs when necessary and extend to multiple time steps. Expressing the robot $\mathbb{SE}(3)$ state in the current robot frame (i.e., x forward, y left, and z up), the competence-aware navigation can be formulated as a constrained optimization problem:

$$\begin{gathered} \mathbf{u}_t^* = \underset{\mathbf{u}_t}{\operatorname{argmax}} \left[||x_{t+1} - x_t|| - C(\mathbf{s}_t, \mathbf{s}_{t+1}) \right], \\ \text{s.t. all } \mathbb{SE}(3) \text{ constraints are satisfied,} \\ \mathbf{s}_{t+1} = f_\theta(\mathbf{s}_t, \mathbf{u}_t, \psi_t^{\text{V, S, I}}). \end{gathered} \tag{5}$$

Notice that the objective function in Eqn. (5) can be formulated in other ways when necessary. For example, maximizing the displacement along x can be replaced by minimizing the difference between the control u and a desired manual command. The navigation planner then finds the best control $\mathbf{u}_t^*$ to maximize speed along x (and considers other costs in $C(\cdot, \cdot)$), while in a human shared-control setup $\mathbf{u}_t^*$ aims to minimize the difference compared to human command, both without violating $\mathbb{SE}(3)$ vehicle kinodynamic constraints.

III. IMPLEMENTATION

We implement CAHSOR ground navigation on a 1/6-scale autonomous vehicle, an AgileX Hunter SE, with a top speed of 4.8m/s on different off-road terrain at high speeds to demonstrate the proposed competence awareness.

We collect a dataset of 30-minute vehicle-terrain interactions. The collected GPS-RTK, onboard IMU, front-facing camera, and vehicle control data are synchronized and processed into training data. We integrate the learned TRON and downstream kinodynamic models and the CAHSOR framework with an autonomous navigation planner and a human shared-control setup.

A. CAHSOR *Implementations*

1) TRON*:* The terrain vision encoder is a 4-layer Convolutional Neural Network (CNN) to produce a 512-dimensional viewpoint-invariant visual representation. The speed encoder is a 2-layer neural network, whose 512-dimensional output is combined with the visual representation to construct our vision-speed representation ψ_t^{VS}. The last 2-second accelerometer and gyroscope data are converted into the frequency domain using Power-Spectral Density (PSD) representation before being fed into the 2-layer inertia encoder and producing a 512-dimensional inertial representation ψ_t^{I}. All encoders are trained to minimize $\mathcal{L}_{\text{TRON}}$ (Eqn. (1)).

2) Kinodynamics: Our $\mathbb{SE}(3)$ vehicle state is instantiated in the current robot frame, i.e., $[x_t, y_t, z_t, r_t, p_t, \phi_t]^T = \mathbf{0}$, and therefore omitted from the input of our forward kinodynamic model (Eqn. 3). To explicitly showcase the efficacy of the learned kinodynamic model on state dimensions beyond $\mathbb{SE}(2)$, we limit the model output to three metrics to reflect sliding along y, roll around x, and bumpiness along z, i.e., $[\texttt{sliding}_{t+1}, \texttt{roll}_{t+1}, \texttt{bumpiness}_{t+1}]^T$. While it is not necessary for the human shared-control setting, for autonomous navigation planning, other state dimensions are produced using a simple Ackermann-steering model, whose predicted 3-DoF trajectories are evaluated for competence awareness with the learned kinodynamic model. Such a practice also avoids the computation overhead of sequentially rolling out a large set of multi-step, 6-DoF candidate trajectories, which cannot be efficiently parallelized on GPUs. To be specific, $\texttt{sliding}_{t+1}$ is captured by the ground speed sensed by GPS-RTK projected onto the robot y axis (left); We compute the absolute angular acceleration around the x axis (front) from the gyroscope averaged over 0.1s as $\texttt{roll}_{t+1}$; $\texttt{bumpiness}_{t+1}$ is computed as the absolute jerk along the z axis (up) from the accelerometer averaged over 0.1s. As a downstream task of TRON, the kinodynamic model (Eqn. 3) is learned with three 256-64-1 neural network heads, which take as input the pretrained visual, speed, and/or inertial representation $\psi_t^{\text{V, S, I}}$ and candidate control actions $\mathbf{u}_t = (v, \omega)$ (omitted in Fig. 2 right for simplicity), to produce $\texttt{sliding}_{t+1}$, $\texttt{roll}_{t+1}$, and $\texttt{bumpiness}_{t+1}$.

B. *Autonomous Navigation Planning with* CAHSOR

We integrate our CAHSOR model with a Model Predictive Path Integral (MPPI) planner. Our MPPI planner rolls out a set of candidate 3-DoF state trajectories using sampled action sequences and then combines those samples based on a predefined cost function. The cost function is informed by the prediction of the learned 6-DoF kinodynamic model, assigning infinitely large costs to candidate trajectories that involve sliding, roll, and bumpiness. MPPI then updates the sampling distribution to sample actions that are more likely to lead to low cost trajectories, i.e., moving the robot toward a goal at the fastest possible speed. MPPI rolls out 550 trajectories, each with 8 vehicle states. We select six goals in an outdoor off-road environment for the robot to drive to in a loop. For MPPI rollouts, future terrain inertial responses are not available to the TRON model. Therefore, we only use the visual and speed representation ψ_t^{VS} as $\psi_t^{\text{V, S, I}}$ to represent the world state $\mathbf{w}_t$ associated with each future vehicle state $\mathbf{s}_t$ on the candidate trajectories. For computation efficiency, we divide the current BEV into a 15×51 grid and pick the terrain patch that is closest to $\mathbf{s}_t$ on the candidate trajectories for parallelized model query on GPU during one MPPI cycle.

C. *Human-Autonomy Shared-Control with* CAHSOR

We also demonstrate the use case of CAHSOR in a human-autonomy shared-control setup, in which a human driver aims at driving the robot as fast as possible, while CAHSOR takes care of satisfying all vehicle $\mathbb{SE}(3)$ constraints with the closet possible vehicle control to the human command. In this case, the objective function in Eqn. (5) becomes

$$\mathbf{u}_t^* = \operatorname*{argmin}_{\mathbf{u}_t} ||\mathbf{u}_t - \mathbf{u}_t^H||. \tag{6}$$

$\mathbf{u}_t^H$ is the desired human control input, which will potentially violate the $\mathbb{SE}(3)$ constraints. In this shared-control setup, both inertia and vision (from past camera images) are available, so TRON takes in visual, speed, and inertial representation as $\psi_t^{\text{V, S, I}}$ to represent the current world state $\mathbf{w}_t$.

IV. EXPERIMENTS

We deploy CAHSOR navigation with human-autonomy shared-control setup and autonomous navigation planning using MPPI. Please see our video `https://www.youtube.com/watch?v=vqhIKjabTF4` for the experimental results and more information.

V. CONCLUSIONS

Our CAHSOR ground navigation approach is able to utilize multimodal, self-supervised terrain representation, i.e., TRON, to reason about the consequences of taking aggressive maneuvers on different off-road terrain, i.e., being competence-aware. inertial observations contain the most information to enable efficient kinodynamics learning, but may not be available during planning. augmenting easily available vision combined with speed using inertia with TRON, similar kinodynamics learning performance can be achieved. extensive physical experiments in both an autonomous navigation planning and human shared-control setup demonstrate CAHSOR's superior competence awareness during high-speed off-road navigation.