Zero-Shot Context Identification through Clustering and Foundation Modeling for Friction Estimation

Renukanandan Tumu* Ahmad Amine* Lee Milburn* Rajnish Gupta Urara Kono Rahul Mangharam

***Abstract*—Off-road autonomous navigation demands accurate estimation of terrain-dependent parameters, particularly tire-ground friction, which directly impacts control performance and safety. Traditional methods for friction estimation—whether proprioceptive, vision-based, or hybrid—struggle to adapt to abrupt terrain transitions and lack generalization to previously unseen environments. This paper introduces Physics-Constrained and Vision-Informed Friction Estimation (PC-VFE), a framework that combines semantic visual understanding through the use of foundation models with physics-based dynamics modeling to estimate friction in real time. PC-VFE first identifies terrain contexts using a vision-language model and unsupervised clustering, then estimates context-specific friction parameters via a constrained optimization process. Our approach requires no prior knowledge of terrain types, adapts in a zero-shot manner, and enables rapid re-identification of known surfaces.**

I. Introduction

Robust state estimation and control remain central challenges in autonomous mobile robotics, particularly under uncertain and dynamic operating conditions. Classical approaches in structured environments, such as urban driving or warehouse navigation, benefit from high-definition maps, semantic priors, and relatively consistent surface conditions [1], [2], [3], [4]. However, these assumptions break down in off-road contexts, where terrain geometry and physical properties vary rapidly, and exteroceptive sensing is often ambiguous or degraded [5], [6], [7].

Off-road autonomous navigation introduces a distinct set of challenges stemming from the heterogeneity of terrain surfaces, unstructured environmental features, and the absence of prior knowledge [8], [9], [10]. Effective motion planning and control in such settings depend critically on accurate estimation of terrain-dependent dynamics parameters, particularly the tire-ground friction coefficient, which directly impacts vehicle stability, path tracking, and safety [11], [12], [13]. Existing model-based control strategies rely on accurate vehicle dynamics models parameterized by friction and cornering stiffness terms [14], [15], and their performance degrades significantly in the presence of incorrect or outdated parameter estimates.

Prior approaches to friction estimation span a wide spectrum—from purely proprioceptive techniques [16] to purely vision-based learning systems [17]. In wheeled robots, proprioceptive estimation methods such as slip detection or contact-based estimation carry inherent safety risks with late detections [18], unlike their use in legged systems where controlled contact is feasible [19]. Vision-based friction estimation methods often treat terrain type as a proxy for friction and assume friction is invariant across a class (e.g., all "grass" is assigned a fixed coefficient), restriction the problem to a finite class of labels [20], [21]. Additionally, end-to-end learning approaches train on fixed condition in simulation or labels available in the dataset, thus requiring retraining to adapt to changes in the environment [22], [23].

Fig. 1: Autonomous vehicles driving off-road may encounter different terrains, shown here as $T_1, \ldots, T_3$.

To the best of our knowledge, there does not exist a framework that can simultaneously (i) detect and adapt to abrupt transitions in terrain conditions similar to what is shown in Section I, (ii) re-identify terrain contexts without requiring extensive look ahead or buffer windows, and (iii) infer friction parameters in a zero-shot or few-shot fashion for unseen terrain classes. Importantly, off-road systems must operate in settings where the number and identity of terrain types is unknown a-priori and cannot rely on exhaustive prior data collection.

This paper introduces Physics-Constrained and Vision-Informed Friction Estimation (PC-VFE), a framework for accurate and adaptive friction estimation in off-road environments. By combining physics-based modeling with vision foundation models, PC-VFE addresses challenges such as terrain variability and unseen conditions. The approach is divided into two main components: terrain context identification and context-specific friction parameter estimation. It leverages semantic and low-level visual information to dynamically cluster terrain contexts and adapt to new terrains without prior knowledge. A physics-constrained optimization process ensures compatibility with vehicle dynamics, enabling real-time adaptation to changing and unseen terrain conditions.

We highlight three key aspects of PC-VFE:

1) It achieves faster response times to terrain transitions compared to buffer-based methods,
2) It adapts effectively to an unknown and expanding set of unseen terrain conditions, and
3) It improves control performance, demonstrated by higher average speeds and fewer catastrophic failures.

PC-VFE showcases the benefits of integrating vision-driven context identification with physics-based parameter estimation for robust and adaptive friction estimation in off-road environments.

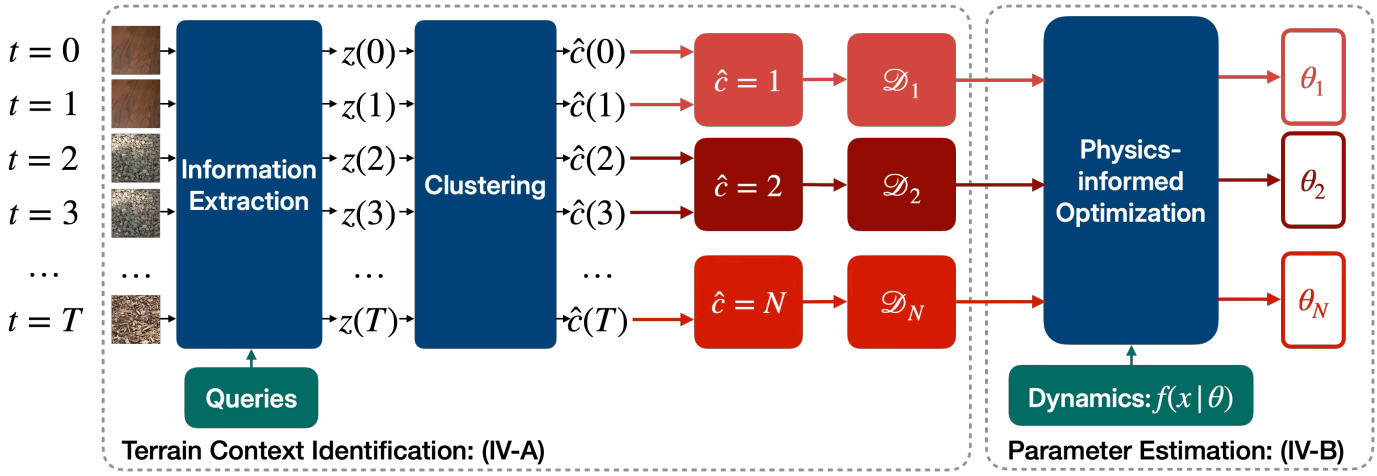

Fig. 2: Overview of the proposed PC-VFE approach.

II. Background

This section will cover the background on related work in off-road and on-road friction estimation, and terrain understanding. Previous approaches can be broadly classified into three categories [24], [25]: (1) machine learning-based methods, (2) dynamics-based methods, and (3) hybrid approaches that combine both. We will also discuss the limitations of these approaches in the context of off-road environments.

A. Machine Learning Based Approaches

Machine learning based approaches to friction estimation have been widely studied in the literature. These methods typically rely on supervised learning techniques, where a model is trained on labeled data to predict friction coefficients based on various features extracted from sensor data. Ribeiro et al. use time-delayed neural networks to generate estimates [26]. Other approaches use camera information paired with a dataset of friction estimates to generate friction estimates [27], [28]. Guo et al. run two parallel estimation techniques, one based on images and the other using a filtering approach, to generate estimates [29]. [30] rely on end-to-end learning of three models to combine image data with IMU data to control their vehicle. This approach relies on training the model on human demonstrations on a specific track, which works in practice but does not guarantee the model's ability to adapt to new environments. For use in legged robotics, recent work predicts the friction and stiffness of the terrain using data collected in simulation [31]. While this can handle changing terrain, its predictions are not guaranteed to be physically accurate or to generate control responses that make sense for the surface.

B. Dynamics Based Approaches

Dynamics-based approaches to friction estimation use state information collected from the robot combined with a vehicle model to identify surface friction [32]. Some approaches use filtering or observer approaches to identify friction [33], [34], [35], while others use gradient descent methods [36]. These methods provide a method for estimating friction using onboard sensors, in or near real time [25]. They often struggle with non-linear parameter estimation outside specific ranges [37], and do not address the problem of terrain switching.

C. Hybrid Approaches

Hybrid approaches to friction estimation combine the strengths of both machine learning and dynamics-based methods. [38] use Bayesian neural networks to learn a probabilistic estimate of the friction parameters of an autonomous vehicle, but rely on a fixed set of classes to differentiate between, which limits the number of terrain contexts that can be identified. [39] train a Gaussian process to learn the friction curve of the vehicle, but are limited to slow changing friction changes due to exponential forgetting. Some approaches have used Physics-Informed Neural Networks (PINNs) to perform state and parameter estimation in manipulators [40], [41]. These approaches are fast and can run in real-time, but do not take into account changing terrain, or the physical correctness of the predictions. These disadvantages stem from the same key factor: that they are trained offline on historically collected data, but do not ensure that estimates found are consistent with data collected online.

While prior work has addressed various aspects of friction estimation, these methods do not jointly estimate terrain-specific friction in real time or adapt to previously unseen terrain contexts. We now formalize this joint estimation problem.

III. Problem Formulation

We address the problem of simultaneously estimating the friction parameters for an off-road vehicle's dynamics model and identifying the terrain context in which the vehicle operates. We assume vehicle dynamics are described by the discrete-time nonlinear model in Equation (1) where $x(t)$ is the vehicle state, $u(t)$ is the control input, and $\theta(t)$ represents the friction parameters at time t.

$$x(t+1) = f(x(t), u(t); \theta(t)) \tag{1}$$

Specifically, we will consider the single-track dynamic bicycle model described in [42], where p_x, p_y, ψ are the pose of the vehicle in world coordinates, δ is the steering angle of the front wheels, v is the velocity, and β is the side-slip angle. θ in this case is the set $[C_{s,f}, C_{s,r}, \mu]$ denoting the cornering stiffness of the front and rear tires and the tire-road friction coefficient respectively. A common approach to solving this problem is to optimize a single friction parameter set across all observed data:

$$\min_{\theta} \sum_{t=0}^{T} \mathcal{L}(x(t), u(t), x(t+1); \theta) \tag{2}$$

where the cost function $\mathcal{L}$ measures the error between the predicted and actual states. For example, the cost function could be the squared error:

$$\mathcal{L}(x(t), u(t), x(t+1); \theta) = \|f(x(t), u(t); \theta) - x(t+1)\|_2^2$$

However, this assumes friction parameters are homogeneous across terrain contexts, which does not hold for off-road environments due to terrain variability (e.g., differences between hard-packed dirt, loose gravel, and mud).

Thus, we propose a joint optimization problem that simultaneously estimates friction parameters θ_k for each terrain context, predicts the current terrain context $c(t)$, and simultaneously identifies the number of terrain contexts K:

$$\begin{aligned} \min_{\{\theta_k\}_{k=1}^{K}, \{\hat{c}(t)\}_{t=1}^{T}, K} \quad & \sum_{t=1}^{T} \|f(x(t), u(t); \theta_{\hat{c}(t)}) - x(t+1)\|_2^2 \\ & + \lambda_P P_{\text{context}}. \end{aligned} \tag{3}$$

P_{context} is a tuneable cost that penalizes the number of contexts identified K, promoting meaningful terrain clusters, with λ_P some tuneable multiplier on P_{context}.

This formulation poses several challenges: the terrain context $c(t)$ is not directly observable, the number of terrain contexts K is unknown a priori, and friction parameters must be estimated reliably with limited data per context. We address these challenges by integrating visual semantic information with clustering techniques to infer terrain contexts effectively, detailed in the following sections.

IV. Proposed Approach (PC-VFE)

Our approach, PC-VFE addresses the problem in Equation (3) by decomposing the friction estimation problem into two sub-components: (1) Terrain context Identification and (2) Context-Specific Friction Parameter Estimation. The Terrain Context Identification task is that of identifying the context $\hat{c}(t)$, given state information and an image. The Context-Specific Parameter Identification task uses the identified context along with the state information to produce a set of dynamics-specific parameters for use in a controller.

A. Terrain Context Identification

This subproblem is that of identifying the terrain context $\hat{c}(t)$ given our observation of the vehicle state $x(t)$ and image input $i(t)$. The image sensor faces directly downwards in order to ensure that the image at time t is representative of the current state of the vehicle. This identification task is often ambiguous because of the variation in observed images in the same terrain. An example of this is intermittent patches of dead grass on a lawn. While we wish to capture terrain context which is specific, we must balance this with the need for enough data per terrain context to find the friction parameters we seek to identify. In order to complete this task, we extract semantic and low-level information from the image, and compose a condensed information vector $z(t)$. We perform unsupervised clustering on $z(t)$ to obtain $\hat{c}(t)$.

Information Extraction: When humans complete this task, we use semantic interpretations of the terrain we see, and classify terrain with similar semantic interpretations together. This has motivated the family of approaches to friction parameter estimation in Section II-A. While classification of terrain might be sufficient for vehicles which operate in known terrain, or a known set of terrains, it often does not perform well when vehicles operate in unknown terrain.

Our approach mimics the human extraction of semantic information through the use of a Vision Foundation Model (VFM), specifically a Vision-Language Model (VLM), which is trained on internet scale data to relate the semantic meaning of a set of captions Q to an image. The VLM accomplishes this task by projecting the image i using the captions Q into a related latent space. The VLM produces a semantic information vector $\text{VLM}(i(t), Q) = z_s \in \mathbb{R}^{|Q|}$ which contains the normalized similarity of the image to each of the captions in the set Q. Each element in the vector represents the softmax of the similarity between the image and an individual caption, therefore the semantic information vector has interpretable components rather than being purely latent. In this paper, we use CLIP [43] as our VLM. The caption set we use contains terms like "grass","gravel", and "snow" prefixed by "This image contains". A full caption would read "This image contains grass". The complete set of captions is presented in Appendix A.

While the VLM can be used to effectively extract the semantic meaning of the vector, some low-level information is valuable to identify our terrain context. We denote the function that extracts this information as $LL(i)$, which produces a low-level information vector $z_l \in \mathbb{R}^n$. The size of the vector is dependent on the specific information extracted. In our implementation, we calculate the average RGB values of the image, resulting in a vector z_l with size $n = 3$.

The fully assembled information vector is the composition of the semantic and low-level information. The information vector at a specific time $z(t)$ is the weighted concatenation of the individual information vectors weighted by a parameter λ_c in order to compensate for scaling issues in the next step. The scaling factor λ_c is selected empirically, and is 0.25 in all presented experiments.

The final information extraction process which maps images $i(t)$ from image-space $\mathbb{I}$ to the space of information vectors

$\mathbb{Z} \in \mathbb{R}^{|Q|+n}$ is presented in Equation (4).

$$z(t) = \text{concat}\left(\text{VLM}(i(t), Q), \lambda_c \text{LL}(i(t))\right) \tag{4}$$

Clustering: Given a set of information vectors $\{z(t), t \in [0, 1, \ldots, T]\}$, we seek to generate a set of predicted contexts $\hat{c}(t)$. We use a clustering algorithm which does not require apriori specification of the number of clusters in order to support an indefinite number of terrain contexts. This allows our robot to autonomously navigate in and adapt to previously unseen terrains. Clustering algorithms that support this behavior include DBSCAN [44], OPTICS [45], and HDBSCAN [46]. In our implementation, we use HDBSCAN, evaluating the clustering algorithm at 0.1Hz, clustering all information vectors provided up until the time of execution.

We implement a context persistence mechanism on top of HDBSCAN to maintain consistent cluster centroids over time, a property typically not guaranteed by offline clustering algorithms. The proposed method, detailed in Algorithm 1, uses cluster centroids in consecutive runs to preserve continuity in $\hat{c}(t)$. We denote the number of clusters at time t by K_t, and the centroid of cluster k at time t by $m_k(t)$.

B. Context Specific Friction Parameter Estimation

Given a dynamics function $x(t+1) = f(x(t), u(t); \theta)$, we seek to find θ that best fits the data collected. Given that the vehicle autonomously navigates different terrains, each with a unique context ID $\hat{c}(t)$ identified in section Section IV-A, we split our data into k different datasets $\mathcal{D}_k$. Each $\mathcal{D}_k$ consists of the tuple $\{\mathcal{X}_k, \mathcal{U}_k, \mathcal{P}_k\}$, where $\mathcal{X}_k$ is the set of states $[x_k(t), x_k(t+1), \ldots, x_k(t+H-1)]$, $\mathcal{U}_k$ is the set of control inputs applied to the system at those states $[u_k(t), u_k(t+1), \ldots, u_k(t+H-1)]$, and $\mathcal{P}_k$ is the set of future states $[x_k(t+1), x_k(t+2), \ldots, x_k(t+H)]$ that we wish our dynamics model to accurately predict using the estimated friction parameters for that cluster $\hat{\theta}_k$.

For each identified terrain context k, we estimate the friction parameters θ_k by solving the following problem:

$$\min_{\hat{\theta}_k} \sum_{t \in \mathcal{I}_k} \|(\hat{x} \backslash \psi)(t) - (x \backslash \psi)(t)\|_2^2 + \left| \tan^{-1}\left(\frac{\sin\left(\hat{\psi}(t) - \psi(t)\right)}{\cos\left(\hat{\psi}(t) - \psi(t)\right)} \right) \right| \tag{5a}$$

$$\text{subject to: } \hat{x}(t) = x(t) \tag{5b}$$

$$\hat{x}(t+1) = f(\hat{x}(t), u(t); \hat{\theta}_k) \tag{5c}$$

$$\mathcal{I}_k = \left\{ t \,\middle|\, \bigwedge_{t \in \mathcal{I}_k} (\hat{c}(t) = k) \wedge (\phi(x(t))) \right\} \tag{5d}$$

where $x \backslash \psi$ denotes the state vector without the heading ψ. ψ is left out of the L_2 error as ψ is a 2π-periodic variable, and takes values in the interval $[0, 2\pi)$. Instead, the loss $\left| \tan^{-1}\left(\frac{\sin(x-y)}{\cos(x-y)} \right) \right|$ is used to handle this periodicity.

$\mathcal{I}_k$ denotes the union of time intervals where the vehicle remains within the same terrain context k. This constraint ensures that we only use data from stable terrain segments for parameter estimation, avoiding transition regions that could contaminate our estimates. Some dynamics functions have numeric instability in some portions of the state space. We use the the filter $\phi(x(t)) \mapsto \{0, 1\}$ to exclude these regions. In our case this filter is described as follows:

$$\phi(x(t)) := \{v \geq 1\text{m/s}\}. \tag{6}$$

This problem is non-convex, and we solve it by back-propagating the loss on the predictions through the dynamics. For each dataset $\mathcal{D}_k$, we set our initial condition as per Equation (5b), and then integrate the state with the control inputs $\mathcal{U}_k$ using the dynamics Equation (5c). These dynamics are parameterized by our current estimates of the friction parameters $\hat{\theta_k}$. Assuming that the dynamics are differentiable as in our case, we can then calculate the gradient on the loss with respect to $\hat{\theta_k}$ to update $\hat{\theta_k}$ using gradient descent. We use Adam optimizer [47] to update our estimate of θ_k, with a learning rate of $0.0.1$ and a batch size of 1024.

V. Conclusion

We have presented Physics-Constrained and Vision-Informed Friction Estimation (PC-VFE), a framework that integrates physics-based modeling with vision-driven semantic information to address the challenges of friction estimation in off-road environments. By decomposing the friction estimation task into two interlinked subproblems — terrain context identification and context-specific friction parameter estimation — PC-VFE effectively captures the dynamic and heterogeneous nature of off-road terrain.

Our preliminary experimental evaluations, conducted in simulation reinforce the potential of this approach. Specifically, PC-VFE demonstrates quick response time to abrupt terrain transitions and adapts to an unknown and potentially expanding set of terrain conditions.

These results underscore the potential of PC-VFE to bridge the gap between visual perception and physics-based modeling, paving the way for more resilient and adaptive autonomous navigation systems. Future work will explore further refinements to the estimation process, additional sensor integration, and hardware experiments.

References

[1] R. Liu, J. Wang, and B. Zhang, "High definition map for automated driving: Overview and analysis," *The Journal of Navigation*, vol. 73, no. 2, pp. 324–341, 2020.

[2] N. V. Kumar and C. S. Kumar, "Development of collision free path planning algorithm for warehouse mobile robot," *Procedia computer science*, vol. 133, pp. 456–463, 2018.

[3] S. L. Bowman, N. Atanasov, K. Daniilidis, and G. J. Pappas, "Probabilistic data association for semantic slam," in *2017 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2017, pp. 1722–1729.

[4] S. Yang, W. Wang, C. Liu, and W. Deng, "Scene understanding in deep learning-based end-to-end controllers for autonomous vehicles," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 49, no. 1, pp. 53–63, 2018.

[5] R. Gonzalez, F. Rodriguez, J. L. Guzman, C. Pradalier, and R. Siegwart, "Combined visual odometry and visual compass for off-road mobile robots localization," *Robotica*, vol. 30, no. 6, pp. 865–878, 2012.

[6] R. Ren, H. Fu, H. Xue, X. Li, X. Hu, and M. Wu, "Lidar-based robust localization for field autonomous vehicles in off-road environments," *Journal of Field Robotics*, vol. 38, no. 8, pp. 1059–1077, 2021.

[7] M.-Y. Yu, R. Vasudevan, and M. Johnson-Roberson, "Lisnownet: Real-time snow removal for lidar point clouds," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 6820–6826.

[8] A. Shaban, X. Meng, J. Lee, B. Boots, and D. Fox, "Semantic terrain classification for off-road autonomous driving," in *Conference on Robot Learning*. PMLR, 2022, pp. 619–629.

[9] X. Meng, N. Hatch, A. Lambert, A. Li, N. Wagener, M. Schmittle, J. Lee, W. Yuan, Z. Chen, S. Deng, G. Okopal, D. Fox, B. Boots, and A. Shaban, "TerrainNet: Visual Modeling of Complex Terrain for High-speed, Off-road Navigation," in *Proceedings of Robotics: Science and Systems*, Daegu, Republic of Korea, July 2023.

[10] J. Frey, M. Patel, D. Atha, J. Nubert, D. Fan, A. Agha, C. Padgett, P. Spieler, M. Hutter, and S. Khattak, "Roadrunner-learning traversability estimation for autonomous off-road driving," *IEEE Transactions on Field Robotics*, 2024.

[11] M. Deremetz, R. Lenain, B. Thuilot, and V. Rousseau, "Adaptive trajectory control of off-road mobile robots: A multi-model observer approach," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 4407–4413.

[12] T. Nagy, A. Amine, T. X. Nghiem, U. Rosolia, Z. Zang, and R. Mangharam, "Ensemble gaussian processes for adaptive autonomous driving on multi-friction surfaces," *IFAC-PapersOnLine*, vol. 56, no. 2, pp. 494–500, 2023.

[13] J. Knaup, K. Okamoto, and P. Tsiotras, "Safe high-performance autonomous off-road driving using covariance steering stochastic model predictive control," *IEEE Transactions on Control Systems Technology*, vol. 31, no. 5, pp. 2066–2081, 2023.

[14] G. Reina, M. Paiano, and J.-L. Blanco-Claraco, "Vehicle parameter estimation using a model-based estimator," *Mechanical Systems and Signal Processing*, vol. 87, pp. 227–241, 2017.

[15] A. Onat, "A novel and computationally efficient joint unscented kalman filtering scheme for parameter estimation of a class of nonlinear systems," *Ieee Access*, vol. 7, pp. 31 634–31 655, 2019.

[16] X. Yu, S. Teng, T. Chakhachiro, W. Tong, T. Li, T.-Y. Lin, S. Koehler, M. Ahumada, J. M. Walls, and M. Ghaffari, "Fully proprioceptive slip-velocity-aware state estimation for mobile robots via invariant kalman filtering and disturbance observer," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 8096–8103.

[17] S. Roychowdhury, M. Zhao, A. Wallin, N. Ohlsson, and M. Jonasson, "Machine learning models for road surface and friction estimation using front-camera images," in *2018 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2018, pp. 1–8.

[18] S. Yang, M. Black, G. Fainekos, B. Hoxha, H. Okamoto, and R. Mangharam, "Safe control synthesis for hybrid systems through local control barrier functions," in *2024 American Control Conference (ACC)*. IEEE, 2024, pp. 344–351.

[19] M. Camurri, M. Fallon, S. Bazeille, A. Radulescu, V. Barasuol, D. G. Caldwell, and C. Semini, "Probabilistic contact estimation and impact detection for state estimation of quadruped robots," *IEEE Robotics and Automation Letters*, vol. 2, no. 2, pp. 1023–1030, 2017.

[20] T. Zhao, P. Guo, and Y. Wei, "Road friction estimation based on vision for safe autonomous driving," *Mechanical Systems and Signal Processing*, vol. 208, p. 111019, 2024.

[21] E. Šabanovič, V. Žuraulis, O. Prentkovskis, and V. Skrickij, "Identification of road-surface type using deep neural networks for friction coefficient estimation," *Sensors*, vol. 20, no. 3, p. 612, 2020.

[22] E. Salvato, G. Fenu, E. Medvet, and F. A. Pellegrino, "Crossing the reality gap: A survey on sim-to-real transferability of robot controllers in reinforcement learning," *IEEE Access*, vol. 9, pp. 153 171–153 187, 2021.

[23] Y. Pan, C.-A. Cheng, K. Saigol, K. Lee, X. Yan, E. Theodorou, and B. Boots, "Learning deep neural network control policies for agile off-road autonomous driving," in *The NIPS Deep Rienforcement Learning Symposium*, vol. 134, 2017.

[24] S. Khaleghian, A. Emami, and S. Taheri, "A technical survey on tire-road friction estimation," *Friction*, vol. 5, no. 2, pp. 123–146, Jun. 2017. [Online]. Available: https://doi.org/10.1007/s40544-017-0151-0

[25] Y. Wang, J. Hu, F. Wang, H. Dong, Y. Yan, Y. Ren, C. Zhou, and G. Yin, "Tire Road Friction Coefficient Estimation: Review and Research Perspectives," *Chinese Journal of Mechanical Engineering*, vol. 35, no. 1, p. 6, Jan. 2022. [Online]. Available: https://doi.org/10.1186/s10033-021-00675-z

[26] A. M. Ribeiro, A. Moutinho, A. R. Fioravanti, and E. C. de Paiva, "Estimation of tire–road friction for road vehicles: a time delay neural network approach," *Journal of the Brazilian Society of Mechanical Sciences and Engineering*, vol. 42, no. 1, p. 4, Nov. 2019. [Online]. Available: https://doi.org/10.1007/s40430-019-2079-y

[27] E. Šabanovič, V. Žuraulis, O. Prentkovskis, and V. Skrickij, "Identification of Road-Surface Type Using Deep Neural Networks for Friction Coefficient Estimation," *Sensors*, vol. 20, no. 3, p. 612, Jan. 2020, number: 3 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/1424-8220/20/3/612

[28] B. Leng, D. Jin, L. Xiong, X. Yang, and Z. Yu, "Estimation of tire-road peak adhesion coefficient for intelligent electric vehicles based on camera and tire dynamics information fusion," *Mechanical Systems and Signal Processing*, vol. 150, p. 107275, Mar. 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0888327020306610

[29] H. Guo, X. Zhao, J. Liu, Q. Dai, H. Liu, and H. Chen, "A fusion estimation of the peak tire–road friction coefficient based on road images and dynamic information," *Mechanical Systems and Signal Processing*, vol. 189, p. 110029, Apr. 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0888327022010974

[30] H. Karnan, K. S. Sikand, P. Atreya, S. Rabiee, X. Xiao, G. Warnell, P. Stone, and J. Biswas, "Vi-ikd: High-speed accurate off-road navigation using learned visual-inertial inverse kinodynamics," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 3294–3301.

[31] J. Chen, J. Frey, R. Zhou, T. Miki, G. Martius, and M. Hutter, "Identifying Terrain Physical Parameters From Vision - Towards Physical-Parameter-Aware Locomotion and Navigation," *IEEE Robotics and Automation Letters*, vol. 9, no. 11, pp. 9279–9286, Nov. 2024, conference Name: IEEE Robotics and Automation Letters. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10669238

[32] F. Gustafsson, "Slip-based tire-road friction estimation," *Automatica*, vol. 33, no. 6, pp. 1087–1099, Jun. 1997. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0005109897000034

[33] X. Ping, S. Cheng, W. Yue, Y. Du, X. Wang, and L. Li, "Adaptive estimations of tyre–road friction coefficient and body's sideslip angle based on strong tracking and interactive multiple model theories," *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering*, vol. 234, no. 14, pp. 3224–3238, Dec. 2020. [Online]. Available: https://journals.sagepub.com/doi/10.1177/0954407020941410

[34] T. Nakatsuji, I. Hayashi, P. Ranjitkar, T. Shirakawa, and A. Kawamura, "Online Estimation of Friction Coefficients of Winter Road Surfaces Using the Unscented Kalman Filter," *Transportation Research Record*, vol. 2015, no. 1, pp. 113–122, Jan. 2007, publisher: SAGE Publications Inc. [Online]. Available: https://doi.org/10.3141/2015-13

[35] S. Solmaz and S. Başlamışlı, "Simultaneous estimation of road friction and sideslip angle based on switched multiple nonlinear observers," *IET Control Theory & Applications*, vol. 6, no. 14, pp. 2235–2247, Sep. 2012. [Online]. Available: http://digital-library.theiet.org/doi/10.1049/iet-cta.2011.0533

[36] W. Chen, D. Tan, and L. Zhao, "Vehicle Sideslip Angle and Road Friction Estimation Using Online Gradient Descent Algorithm," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 12, pp. 11 475–11 485, Dec. 2018, conference Name: IEEE Transactions on Vehicular Technology. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8489902?casa_token=-h0GORZykWEAAAAA:qQv3JnHISb6G7EPt42vdKdtKBSA7BPcNUJK1lpyyi2RZjdysHSmEj2BbQm8P9tl3EVA

[37] L. Shao, C. Jin, C. Lex, and A. Eichberger, "Robust road friction estimation during vehicle steering," *Vehicle System Dynamics*, vol. 57, no. 4, pp. 493–519, Apr. 2019, publisher: Taylor & Francis _eprint: https://doi.org/10.1080/00423114.2018.1475678. [Online]. Available: https://doi.org/10.1080/00423114.2018.1475678

[38] B. Volkmann and K.-P. Kortmann, "Friction and road condition estimation using bayesian networks," *IFAC-PapersOnLine*, vol. 56, no. 2, pp. 854–861, 2023.

[39] K. Berntorp, "Online bayesian tire-friction learning by gaussian-process state-space models," *IFAC-PapersOnLine*, vol. 53, no. 2, pp. 13 939–13 944, 2020.

[40] X. Yang, Y. Du, L. Li, Z. Zhou, and X. Zhang, "Physics-Informed Neural Network for Model Prediction and Dynamics Parameter Identification of Collaborative Robot Joints," *IEEE Robotics and Automation Letters*, vol. 8, no. 12, pp. 8462–8469, Dec. 2023, conference Name: IEEE Robotics and Automation Letters. [Online]. Available: https://ieeexplore.ieee.org/document/10305255
[41] Q. Le Lidec, I. Kalevatykh, I. Laptev, C. Schmid, and J. Carpentier, "Differentiable Simulation for Physical System Identification," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 3413–3420, Apr. 2021, conference Name: IEEE Robotics and Automation Letters. [Online]. Available: https://ieeexplore.ieee.org/document/9363565
[42] M. Althoff, M. Koschi, and S. Manzinger, "CommonRoad: Composable benchmarks for motion planning on roads," in *2017 IEEE Intelligent Vehicles Symposium (IV)*, Jun. 2017, pp. 719–726. [Online]. Available: https://ieeexplore.ieee.org/document/7995802
[43] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning Transferable Visual Models From Natural Language Supervision," Feb. 2021, arXiv:2103.00020 [cs]. [Online]. Available: http://arxiv.org/abs/2103.00020
[44] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, ser. KDD'96. Portland, Oregon: AAAI Press, Aug. 1996, pp. 226–231.
[45] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander, "OPTICS: ordering points to identify the clustering structure," *SIGMOD Rec.*, vol. 28, no. 2, pp. 49–60, Jun. 1999. [Online]. Available: https://dl.acm.org/doi/10.1145/304181.304187
[46] R. J. G. B. Campello, D. Moulavi, and J. Sander, "Density-Based Clustering Based on Hierarchical Density Estimates," in *Advances in Knowledge Discovery and Data Mining*, J. Pei, V. S. Tseng, L. Cao, H. Motoda, and G. Xu, Eds. Berlin, Heidelberg: Springer, 2013, pp. 160–172.
[47] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," Jan. 2017, arXiv:1412.6980 [cs]. [Online]. Available: http://arxiv.org/abs/1412.6980

APPENDIX

A. CLIP Prompts

["asphalt", "concrete", "brick", "cobblestone", "tile", "hardwood", "grass", "dirt", "gravel", "sand", "mulch", "leaves", "snow", "ice", "metal", "sidewalk", "plastic", "ceramic", "granite", "slate", "pavement", "crosswalk", "train track", "boardwalk", "astroturf", "wood chips", "paver stones", "carpet", "linoleum", "speckled", "striped", "wood", "mud"]

B. Cluster Persistence Algorithm

Algorithm 1 Centroid-Based Context Persistence

Require: Previous cluster centroids $\{m_j(t-1)\}_{j=1}^{K_{t-1}}$ with IDs $\{\hat{c}_j^{(t-1)}\}$, current cluster centroids $\{m_k(t)\}_{k=1}^{K_t}$, threshold ϵ
Ensure: Consistent cluster IDs for current centroids
for each previous centroid $m_j(t-1)$ **do**
Find the closest new centroid:

$$k^* \leftarrow \arg\min_k \|m_j(t-1) - m_k(t)\|_2$$

if $\|m_j(t-1) - m_{k^*}(t)\|_2 < \epsilon$ **then**
Assign: $\hat{c}_{k^*}^{(t)} \leftarrow \hat{c}_j^{(t-1)}$
end if
end for
for each new centroid $m_k(t)$ without an assigned ID **do**
Assign a new cluster ID to $m_k(t)$
end for

C. Results

Our experiment was designed to prove the concept that our method can identify an unknown and expanding set of terrains in an unseen environment, and identify and re-identify terrain contexts quickly.

Our future experiments will evaluate these claims and that PC-VFE enables faster, safer control in unknown environments through comparisons to baseline parameter estimators. We will compare PC-VFE to a buffer-based method akin to a UKF, a lookup method with friction values derived from a table, and a Kinematic Model Predictive Controller (KMPC). All controllers which will take friction into account use a Nonlinear Model Predictive Controller (NMPC). The controllers will all use the Single Track Dynamic Model, defined in Althoff et. al.[42]. Our approach will provide friction estimates to the NMPC, as shown in Figure 3.

We evaluate our approach in simulation experiments, conducted in a multi-body physics simulator. Real-world experiments will be conducted using a 1/5th scale autonomous vehicle platform.

D. Simulation Experiments

We evaluate the performance of our proposed PC-VFE framework in a simulated off-road environment using the Commonroad vehicle models [42]. A set of terrain textures are used to simulate the downward facing camera required for our method. The vehicle parameters we use correspond to a Ford Escort. We use a slalom course with two straight sides and two sides composed of s-curves as our reference trajectory, which can be seen in Figure 4, alongside the textures used in the simulation.

We compare the spatial accuracy of our approach with the ground truth in Figure 4. This figure shows the capability of PC-VFE to capture changes in terrain context that align with changes in the environment. These results were collected using only the data from a single lap, without any previous knowledge of the environment. These zero-shot results show the usefulness of VLMs in understanding unseen terrain.

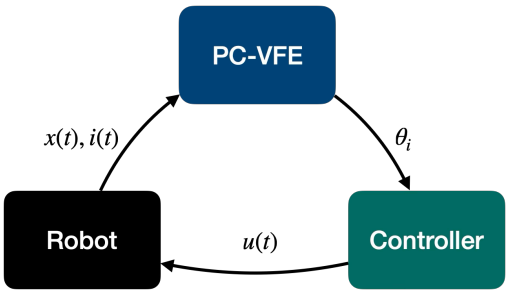

Fig. 3: This figure shows the integration of the PC-VFE method in a robotic control setting. PC-VFE provides model parameter estimates to the controller, which are then used to control the robot and produce states, which are used to refine parameter and terrain context estimates.

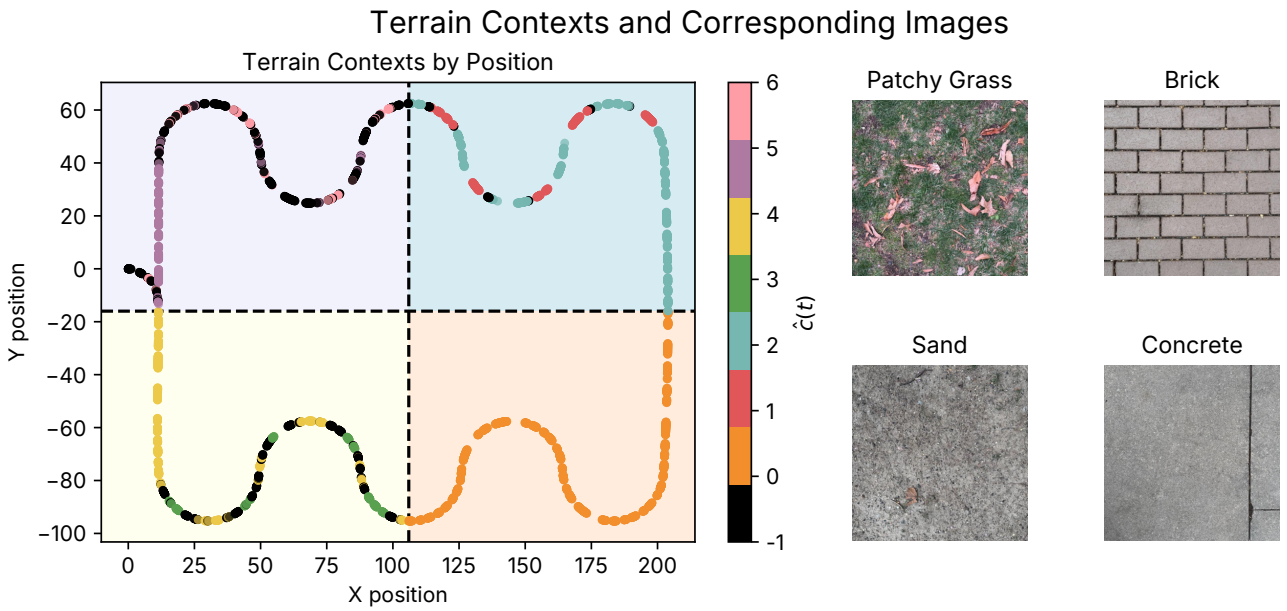

Fig. 4: The figure shows the discovered terrain contexts in the simulation experiments described in Appendix D. The yellow, orange, light blue, and light purple correspond to patchy grass, dry soil, concrete, and brick respectively. The dots indicate the position and cluster. Black is used to show noise clusters, where no terrain context is identified. The clusters above were categorized in a single lap of the track, without any prior information about the environment. The textures used in the simulation are shown on the right hand side.

Parameter	Description	Unit	RoboRacer
l	Wheelbase	m	0.8
m	Mass	kg	15.32
I_z	Moment of inertia	kg · m^2	0.643
l_f	Distance from CoG to front axle	m	0.2735
l_r	Distance from CoG to rear axle	m	0.2585
h_{cg}	Height of center of gravity	m	0.1875
C_{sf}	Cornering stiffness of front wheels	N/rad	2.0
C_{sr}	Cornering stiffness of rear wheels	N/rad	2.0

TABLE I: Single Track Parameters for Experiment Vehicles.

The friction estimate is shown compared to the time elapsed from the start of the experiment in Figure 5. The plot shows the response of the friction estimate to changes in the underlying terrain. While the predicted friction itself is preliminary, and will be the focus of further research, the figure shows the speed of PC-VFE in recognizing and adapting to changing terrain quickly. From around 160 to 220 seconds we see PC-VFE performing a reidentification of the first terrain it encountered. In Figure 7, we can see this reidentification happens in less than 2 seconds.

Our simulation experiments show that PC-VFE is able to recognize and adapt to terrain quickly without apriori knowledge. Future simulation experiments will analyze the performance of friction-aware controllers when presented with updated friction estimates.

E. Real-World Experiments

Real-world experiments will be conducted with the R5 Roboracer, whose parameters can be found in Table I. We

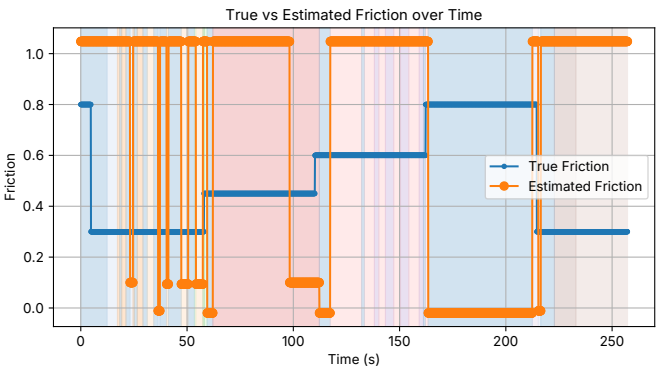

Fig. 5: This plot shows the friction estimate as time elapses from the start of the experiment. The ground truth friction is shown in blue, and the PC-VFE Estimate is shown in orange. The shading behind the plot shows changes in the detected cluster. At 160 seconds elapsed, we can see an example of PC-VFE's ability to quickly detect terrain changes.

will conduct our experiments in a park with three different terrains: sand, leaves, and gravel. This area is depicted from an overhead perspective in Figure 6.

All controllers, localization, and our proposed approach will be computed on-board the platform, using an NVIDIA Jetson AGX Orin. Localization information will be captured by a Fixposition RTK GNSS module, and an onboard VESC which will provide odometry. Our image data will be captured by

Fig. 6: This area will be used to evaluate our method in the real world experiments. The area contains gravel, leaf-covered soil, and mud. Transitional zones contain brick pavers, or splotches of different terrain.

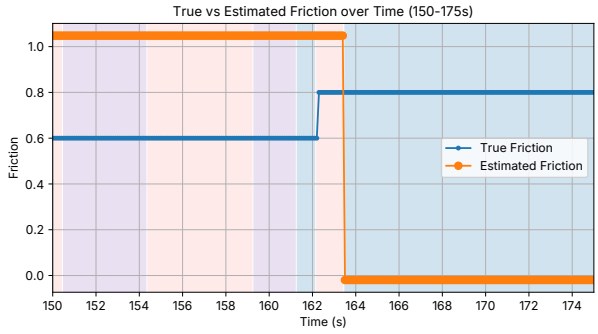

Fig. 7: This figure shows the response time of PC-VFE when re-identifying terrain it has encountered before. This re-identification happens in less than 2 seconds

a GoPro Hero 11 Black, which can achieve the fast shutter speeds required to capture terrain images without blur. We are currently testing the platform to ensure consistency and quality of results.

These results will examine the effectiveness of PC-VFE in real-world settings.

F. PC-VFE Response Time