GaitNet: Greedy, Dynamic Quadruped Gait Generation

Owen Sullivan

Supervisors:
Prof. Mahdi Agheli

Worcester Polytechnic Institute

November 6, 2025

Contents

List of Figures

List of Tables

List of Abbreviations

CNN	**C**onvolutional **N**eural **N**etwork
DQN	**D**eep **Q** **N**etwork
MCTS	**M**onte **C**arlo **T**ree **S**earch
ML	**M**achine **L**earning
MLP	**M**ulti-**L**ayer **P**erceptron
MPC	**M**odel **P**redictive **C**ontrol
PPO	**P**roximal **P**olicy **O**ptimization
RL	**R**einforcement **L**earning

List of Symbols

$\mathbf{f}_c$	Footstep candidate Vector of *leg index* (or -1 for no leg), *dx*, *dy*, and *cost*. With *dx* and *dy* being offsets from the nominal foot position in the base frame.	$\begin{bmatrix}\text{-} & \text{m} & \text{m} & \text{-}\end{bmatrix}^T$
$\mathbf{f}_a$	Footstep action Vector of *leg index* (or -1 for no leg), *dx*, *dy*, and *swing duration*. With *dx* and *dy* being offsets from the nominal foot position in the base frame.	$\begin{bmatrix}\text{-} & \text{m} & \text{m} & \text{s}\end{bmatrix}^T$
$\mathbf{r}_w$	COM position in world frame	m
$\mathbf{q}_{rw}$	Root orientation (quaternion) in world frame	-
$\mathbf{v}_b$	Root linear velocity in base frame	$\text{m}\,\text{s}^{-1}$
ω_b	Root angular velocity in base frame	$\text{rad}\,\text{s}^{-1}$
$\mathbf{p}_{ib}$	End-effector i position in base frame	m
$\mathbf{g}$	Gravity vector in base frame	$\text{m}\,\text{s}^{-2}$
$\mathbf{c}_i$	End-effector i contact state	-
$\mathbf{q}$	Joint positions	-
$\dot{\mathbf{q}}$	Joint velocities	-
$\ddot{\mathbf{q}}$	Joint accelerations	-
τ	Joint torques	$\text{N}\,\text{m}$
$\mathbf{u}$	Control Input Vector of v_x, v_y, and ω	$\begin{bmatrix}\text{m}\,\text{s}^{-1} & \text{m}\,\text{s}^{-1} & \text{rad}\,\text{s}^{-1}\end{bmatrix}^T$

Note: A symbol defined with a subscript i, but displayed without it indicates the concatenation of all i elements, e.g. $\mathbf{c} = \begin{bmatrix}\mathbf{c}_1 & \mathbf{c}_2 & \mathbf{c}_3 & \mathbf{c}_4\end{bmatrix}^T$

1 Introduction

Quadruped locomotion in unstructured environments remains a significant challenge, particularly when traditional gait cycles prove inadequate [2]. While many existing systems rely on periodic gait patterns or centralized planners [2], recent advances in learning-based methods have enabled more adaptive, non-gaited approaches [3]. This study proposes a novel method for generating non-gaited, dynamic footstep plans using a neural network-based greedy planner, aiming to achieve a balance between computational efficiency and dynamic performance.

1.1 Motivation and Vision

Legged robots hold significant promise for enabling mobility in environments that are inaccessible or hazardous to wheeled and tracked systems. Their ability to traverse uneven, unstructured, and unpredictable terrains—such as disaster zones, staircases, dense vegetation, and rocky landscapes—positions them as key enablers for future exploration, inspection, and rescue missions.

Realizing this potential requires a robot to continuously and intelligently decide where and when to place its feet—a process known as contact planning. Unlike periodic gait patterns suitable for flat or predictable surfaces, real-world locomotion often demands dynamic, acyclic, and adaptive footstep sequences. Even minor errors in contact selection can lead to instability or failure, whereas well-chosen footholds enable efficient, agile, and robust movement.

The generation of such contact plans in real time remains a formidable challenge. It requires the system to perceive its surroundings, anticipate the outcomes of possible actions, and select feasible contact points within tight temporal constraints. Existing methods often trade off between deliberative, computationally intensive planning and fast, reactive control that lacks stability or generalization.

The vision of this research is to develop a control architecture that bridges this divide. Specifically, this work aims to design a system capable of producing non-gaited, dynamic footstep plans that retain both the responsiveness of learning-based approaches and the reliability of model-based control. The objective is to provide quadruped robots with the decision-making agility necessary for navigating complex terrains while maintaining the stability and robustness demanded by real-world operation.

Through this approach, the thesis seeks to contribute toward a new class of hybrid locomotion frameworks that enable real-time, adaptive, and physically consistent quadruped motion—closing the gap between high-level perception and low-level control in legged robotics.

1.2 Research Gap

Quadruped control pipelines can be categorized into several approaches, many of which can be broadly classified as traditional, neural network-based, or hybrid methods.

Traditional approaches typically rely on model predictive control (MPC) frameworks for motion optimization, supported by lower-level controllers to track desired trajectories [4]. Many aspects of quadruped locomotion—such as joint torques and ground reaction forces—are smooth and well-suited to continuous optimization techniques [5]. However, these methods become less tractable when discrete contact states are introduced into the optimization problem [6]. To manage this complexity, most implementations either

assume pre-defined gait sequences [2, 3] or encode discrete contact decisions within larger optimization structures [7]. While effective in structured environments, these strategies constrain the robot's ability to exploit the full versatility of legged locomotion.

Neural network-based methods have emerged as an alternative, particularly with advances in deep reinforcement learning. These approaches often replace the MPC component of traditional control pipelines, learning to directly map observed states to joint torques or motion commands [8]. In doing so, they inherently address the mixed-integer nature of contact planning, allowing the network to learn both continuous and discrete aspects of locomotion simultaneously [5]. Despite their flexibility, such methods generally lack formal guarantees of stability, require substantial training data, and often struggle to generalize beyond their training distribution [8].

Hybrid approaches, which integrate learning-based modules within model-based control frameworks, have recently gained attention as a potential middle ground. In these systems, neural networks are typically employed to solve specific subproblems—such as contact selection or foothold prediction—while the overall motion execution remains governed by a traditional controller [9, 10]. This division of responsibilities enables learning to focus on the non-smooth or combinatorial components of locomotion, while preserving the interpretability, safety, and robustness inherent to model-based systems.

Although hybrid control schemes have demonstrated promise, the challenge of contact and footstep planning continues to limit their performance. Traditional optimization-based planners struggle with the discrete nature of contact transitions, resulting in high computational costs that preclude real-time operation [7]. Alternatively, many rely on pre-specified gait patterns [11, 12, 13, 14], constraining adaptability to irregular or unpredictable terrain. Even more advanced approaches, such as MCTS-based planners, have achieved impressive dynamic behaviors but remain computationally demanding [15, 16].

Despite recent progress, a clear research gap remains. Current hybrid methods do not yet achieve the balance between adaptability and stability that would enable fully dynamic, acyclic locomotion in complex environments. In particular, there is a lack of systems that can match the contact planning agility of neural network-based approaches while retaining the formal guarantees and reliability of traditional control frameworks [17, 5]. Recent work, such as ContactNet [1], has demonstrated the potential of machine learning for footstep planning; however, its design—restricted to single-leg motions with fixed timing—still falls short of the dynamic, non-gaited behaviors exhibited by end-to-end learning systems.

This motivates a central research question:

> Can a hybrid control pipeline, which integrates a greedy, neural-network-based planner with a traditional model-based controller, generate dynamic, acyclic gaits for robust quadruped locomotion in challenging environments?

1.3 Hypothesis

To address the identified research gap, we hypothesize that a hybrid control pipeline, incorporating a greedy, neural network-based footstep planner and a traditional model-based control framework, can effectively generate dynamic, non-gaited locomotion for quadruped robots in challenging environments. By leveraging the advancements introduced by ContactNet and extending its capabilities to support multi-leg motion with dynamic swing durations, we propose that such a planner can produce footstep sequences that enable the robot to traverse complex terrains while maintaining both stability and efficiency.

The proposed approach employs a modified implementation of ContactNet's architecture, adapted specifically to generate footstep candidates. These candidates will be evaluated and ranked by our novel planner, *GaitNet*, which selects the most appropriate action at each time step. By integrating this planner within a traditional control framework, we aim to harness the advantages of both learning-based and model-based methods, yielding a robust and adaptable quadruped locomotion system.

1.4 Contributions

- Development of a novel footstep planner utilizing greedy planning with a neural network, for use in hybrid control of quadruped robots.
- Demonstration that the proposed planner can generate dynamic, acyclic gaits in challenging environments.
- Empirical evaluation showing that the planner achieves comparable diverse expressiveness with fast inference times.

1.5 Thesis Structure

- **Chapter 2: Background** - Covers foundational concepts in quadruped locomotion, gait generation, and ML/RL techniques relevant to this work.
- **Chapter 3: Methodology** - Details the design and implementation of the CNN-based greedy planner, including network architecture and training procedures.
- **Chapter 4: Results and Discussion** - Presents experimental results, evaluates the performance of the proposed planner, and discusses its strengths and limitations.
- **Chapter 5: Conclusions and Future Work** - Summarizes the key findings of the thesis and outlines potential directions for future research.

2 Background

Recent advances in legged locomotion have largely focused on either implicit gait-based policies or perception-driven foothold selection modules. However, these approaches often trade off between agility, expressiveness, and computational cost. In contrast to gaited methods that impose rhythmic structure, non-gaited planning allows for more versatile, terrain-adaptive behaviors—but typically at the expense of increased planning complexity. Bridging this gap, our proposed approach draws on greedy methods for their computational efficiency and on convolutional neural networks (CNNs) for terrain-aware generalization, aiming to achieve dynamic, non-gaited footstep planning in real time. This section reviews the foundations upon which our approach builds: learning-based locomotion strategies, footstep planners, greedy selection mechanisms, and foothold classification techniques. Each informs a different aspect of our planner's design and situates our method within the broader landscape of quadruped control.

2.1 Learning-Based Locomotion Strategies

Deep learning has shown strong potential for generating agile, robust locomotion policies, often without explicit footstep planning. Shi et al. [18] modulate trajectory generator parameters in real time for energy-efficient walking, while Xie et al. [11] train RL policies on centroidal dynamics models to output body accelerations under fixed gait heuristics. Duan et al. [19] learn step-to-step transitions using proprioception, and Siekmann et al. [20] achieve blind stair climbing through LSTM-based proprioceptive policies. Lee et al. [13] infer terrain structure from proprioceptive history using a temporal CNN and automated curriculum learning. Though effective, these methods rely on fixed or implicit gait patterns and lack explicit control over individual footsteps.

Subsequent works extend these ideas through high-level gait selection. Da et al. [21] use a DQN to choose among predefined gait primitives executed by a low-level controller, while Yang et al. [22] learn policies that output gait parameters—frequency, swing ratio, and phase offsets—interpreted by a phase integrator. Both enable efficient gait modulation but assume flat terrain and heuristic foot placement. Sun et al. [23] integrate offline gait optimization with contact-implicit trajectory optimization and high-frequency MPC, achieving dynamic control but limiting adaptability to discrete, speed-dependent gaits.

In contrast, Zhang et al. [24] propose an end-to-end RL approach mapping proprioceptive input directly to joint targets, reaching 2.5 m/s and generalizing across terrains via a terrain curriculum and curiosity rewards. However, such models require slow training, and re-training for new environments. Zhang et al. [25] address multi-gait transitions using heuristically defined tables between similar gaits, improving flexibility but still constraining behaviors to predefined families.

Overall, learning-based locomotion methods trade interpretability and efficiency for expressiveness and adaptability. While gait-conditioned policies achieve reliable control, they lack the flexibility needed for explicit, non-gaited footstep planning.

2.2 Optimization-Based Locomotion Strategies

Optimization-based approaches explicitly plan footstep positions, contact sequences, and timing, offering fine-grained control over locomotion. Deits and Tedrake [26] formulate footstep placement as a mixed-integer quadratic program over convex terrain regions, enabling robust bipedal foothold selection. However, their

method does not optimize contact timing and is limited to sequential foot placement, restricting agility and generalization to more complex contact sequences.

Extensions to this framework incorporate contact scheduling into the optimization itself. Winkler et al. [7] fully optimize contact duration and ordering, allowing multiple steps to be planned simultaneously rather than sequentially. While this increases flexibility, the resulting computation is too slow for real-time control. Similarly, Aceituno-Cabezas et al. [27] integrate footstep positions and gait timings into a mixed-integer convex program based on centroidal dynamics, achieving simultaneous contact and motion planning. Taouil et al. [16] further explore non-gaited footstep planning via MCTS, generating dynamic multi-step sequences in real time, though computation remains high due to the branching factor. Akizhanov et al. [28] improve MCTS efficiency using a learned classifier to prune contact configurations and a "target adjustment network" to compensate for low-level control errors, but the approach is still computationally intensive.

These optimization-based methods provide precise, globally consistent locomotion plans, but high computational costs and limited real-time applicability restrict their use in dynamic, non-gaited scenarios. This motivates approaches that combine optimization principles with learning-based or greedy strategies to achieve fast, terrain-adaptive footstep planning.

2.3 Learning-Based Footstep Planners

Learning-based footstep planners combine perception and control through data-driven models, often coupling learned components with downstream MPC or search-based planning. FootstepNet [29] generates heuristic-free plans via deep RL but is limited to bipeds in simple 2D environments. ContactNet [1] uses a CNN to produce non-gaited footstep sequences with greedy selection, though it is constrained to one-leg-at-a-time motion and coarse foothold discretization. DeepGait [30] separates footstep planning from motion optimization, enabling modularity but relying on static gaits. Omar et al. [31] and Villarreal et al. [14] use CNNs for terrain classification to inform MPC or heuristic-free selection, but both depend on fixed gait sequences. DeepLoco [32] introduces a hierarchical policy for footstep and motion planning, yet its coarse terrain input and biped focus limit dynamic adaptability.

Recent advances further expand terrain generalization and multi-contact reasoning. Tolomei et al. [**?**] create a framework like ContactNet but generalized to 3D terrain and variable foot shapes, ranking footholds based on terrain curvature, foot parameters, and kinematic feasibility, but still follow a fixed gait. Chen et al. [33] apply a DQN to a hexapod, simultaneously selecting next foot positions and gait, with fallback to predefined behaviors outside expected states, but motion is constrained to three-leg-at-a-time gaits. Yao et al. [34] output both swing leg motions and desired robot pose from a deep RL policy, demonstrating accurate control with a small network, yet remain limited to a trotting gait. Meduri et al. [35] use a DQN to learn 3D foothold cost maps, similar to ContactNet, but retain fixed biped gait timing.

Overall, these approaches highlight trade-offs between expressive, non-gaited footstep planning and computational efficiency. CNN-based and RL methods improve terrain generalization and allow multi-step planning, but fixed gaits or sequential leg motions remain common constraints. This motivates hybrid methods that combine learned foothold evaluation with fast, greedy planning to achieve real-time, non-gaited footstep generation—forming the basis of our proposed approach.

2.4 Greedy and Heuristic Footstep Planners

Greedy planners offer computationally efficient alternatives to exhaustive search, often prioritizing goal-directed heuristics or local stability metrics. Gao et al. [36] propose GH-QP, a greedy-heuristic hybrid that incorporates expected robot speed into a footstep planning heuristic, though it is limited to bipedal systems and lacks support for dynamic footstep plans. Zucker et al. [37] use A* with a learned terrain cost and Dubins car heuristic for footstep planning, enabling effective search in SE(2) but restricted to static, gated locomotion. Kalakrishnan et al. [38] combine greedy terrain-cost search with a coarse-to-fine pose optimization pipeline, but rely on fixed gait patterns and struggle with highly constrained environments. While these works demonstrate the potential of greedy methods, none support both dynamically unstable and non-gaited footstep generation. Notably, some prior methods already discussed—such as ContactNet [1], DeepLoco [32], and Villarreal et al. [14]—also incorporate greedy elements in their pipelines.

2.5 Foothold Classification Algorithms

Foothold classification algorithms focus on determining safe stepping regions, often using learned terrain models to support downstream planners that enforce safety constraints. Asselmeier et al. [39] generate steppability maps from visual inputs to guide a trajectory-optimization-based planner, validating their perception-to-control pipeline in simulation. Omar et al. [40] propose a two-stage classifier that first identifies steppable regions before selecting footholds, emphasizing safety over expressiveness by evaluating only candidate swing-foot locations. Similarly, Omar et al. [31] use a CNN to identify safe footholds, which are grouped into convex clusters and passed to an MPC for real-time planning, but the approach relies on fixed gait sequences. These methods provide robust terrain understanding but are typically designed as standalone perception modules, separate from the full footstep generation process, and thus are not directly applicable to dynamic or non-gaited motion planning.

3 Methodology

3.1 System Overview

This work presents a control framework for quadruped robots capable of generating dynamic, acyclic gaits in challenging environments. The framework is hierarchical, processing robot state and terrain data to produce footstep commands for the robot controller.

The first stage of the framework is a footstep evaluation network, similar to ContactNet from [1], which estimates footstep candidates $\mathbf{f_c}$. The second stage, a gait generation network (ContactNet), takes these candidates $\mathbf{f_c}$, ranks them, and outputs the optimal footstep action $\mathbf{f_a}$ to the robot controller.

This hierarchical design enables strict control over the range of possible actions the robot can execute. Between the two stages, candidate actions are filtered to ensure that only valid, prescribed movements are permitted.

TODO: update Figure 3.1 with images of my environment. Also update language of different control blocks and connections between them.

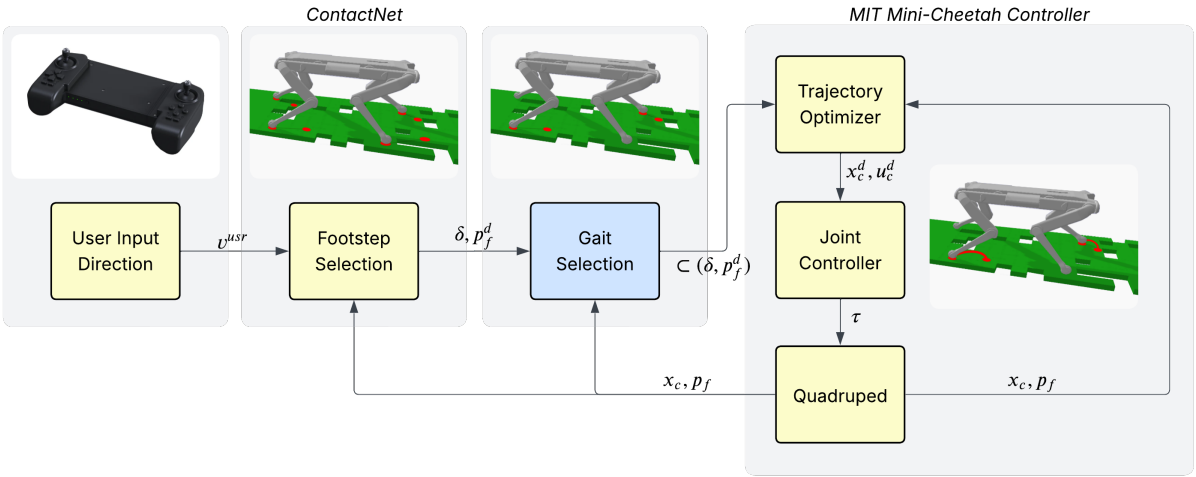


Figure 3.1: A block diagram of the proposed framework. The user defines an input direction v^{usr} which the *footstep selector* [1] uses along with, the robot state x_c, and the current foot positions p_f to generate the swing durations δ and touchdown points p_f^d for all currently grounded feet. The *gait selector* (novel) takes these desired foot movements and selects an appropriate subset $\subset (\delta, p_f^d)$ based on x_c, p_f, and the terrain data. $\subset (\delta, p_f^d)$ is then passed into the MIT Mini-Cheetah Controller to perform lower level control.

3.2 Simulation Environment

The simulation environment used for this project is NVIDIA Isaac Lab [41]. This framework was selected for several reasons. It is a modern platform with a Python interface designed specifically for RL applications and GPU parallelism. The use of GPU parallelism enables significantly faster simulation and data collection, albeit at the cost of increased programming complexity and reduced compatibility with older hardware. Furthermore, the extensive collection of example and community projects provides valuable references for implementing the simulation features required in this work.

Although both simulation and learning processes are executed on the GPU, the robot controllers operate on the CPU. Figure 3.2 illustrates the overall processing flow. The simulation environment runs entirely on the GPU, where multiple robots are simulated in parallel. Meanwhile, the robot MPCs execute on the CPU, with each controller running concurrently. The CPU and GPU communicate at every simulation step to exchange robot states and corresponding control actions.

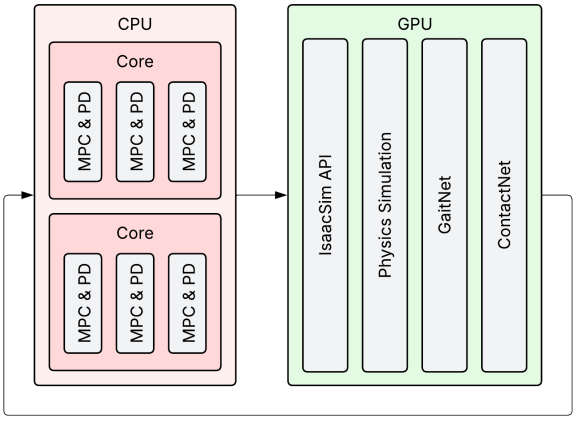


Figure 3.2: A block diagram showing the programming tasks computed on the CPU vs GPU. The full simulation is run in parallel using Nvidia Isaac Lab on the GPU, while the robot MPCs are run in parallel on the CPU.

NVIDIA Isaac Lab uses a declarative system of python dataclasses to define the simulation environment. For this work, a custom environment is used (Figure 3.3), including:

- Unitree Go 1—Configured to use force control for each joint.
- Terrain raycasts—Measure the height of the terrain at each possible footstep location. This emulates the lidar processing steps of a full vision pipeline.
- Custom terrain—A grid of 12x12 sub-terrains (4x4 m each) with increasing difficulty (Figure 3.4). Each sub-terrain consists of a 12 cm square grid pattern with missing sections. The void density varies from ∼0% to 40%. The quantity of terrain is limited by GPU memory.

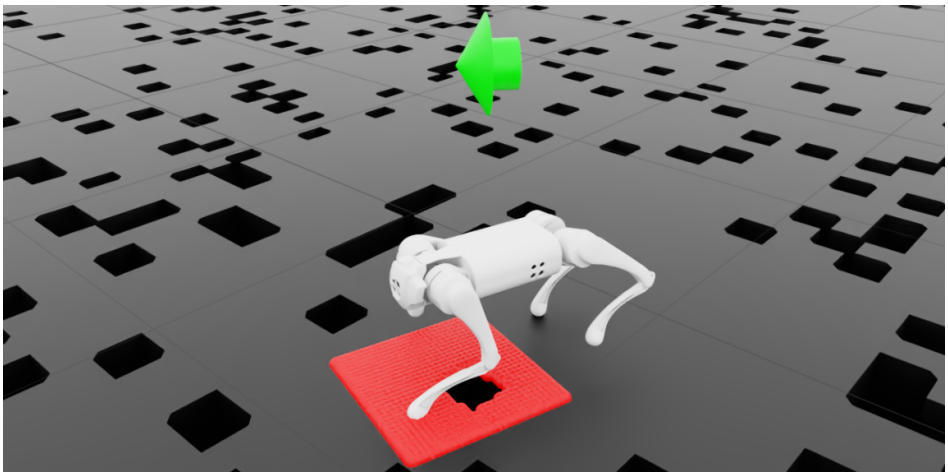


Figure 3.3: Image of a Unitree Go 1 navigating the terrain. Red region shows area converted into heightmap for front left leg. Black regions show voids in the terrain. Green arrow shows desired velocity vector.

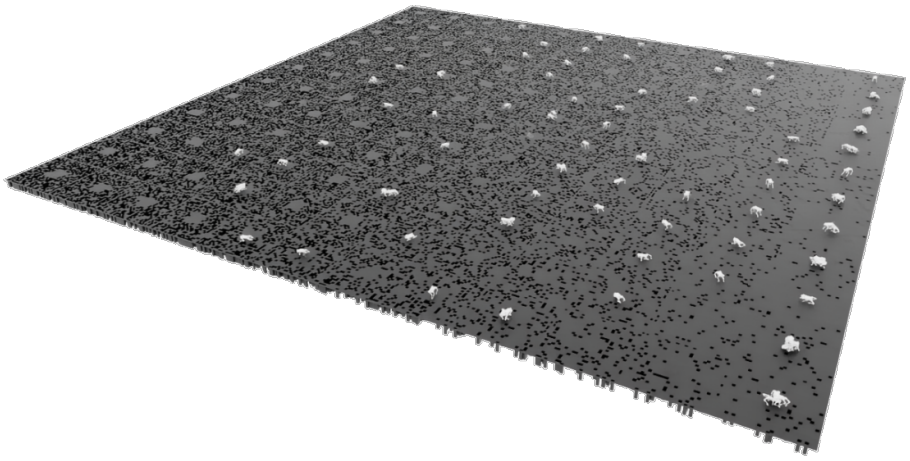


Figure 3.4: Image showing the full terrain used in simulation. Full terrain is a grid of 12x12 sub-terrains with increasing difficulty. Sub-terrains are 4x4 m grids with missing sections.

3.3 Footstep Evaluation Network

The purpose of the footstep evaluation network is to generate footstep candidates for the GaitNet model. Importantly, the precise output of this network is not critical; rather, it is essential that the network provides high-quality candidates when sampled. It achieves this by estimating the cost associated with potential footsteps given the current robot state. The network's architecture and training process are largely based on ContactNet [1], with several key modifications, which are described in detail below.

3.3.1 Architecture

The footstep evaluation network is responsible for generating a set of footstep candidates $\mathbf{f_c}$ based on the robot state $\mathbf{x_c}$:

$$\mathbf{x_c} = \begin{bmatrix} \mathbf{p}_{b,xy} \\ \mathbf{r}_{w,z} \\ \mathbf{v}_b \\ \omega_b \\ \mathbf{u} \end{bmatrix}$$

Here, $\mathbf{p}_{b,xy}$ represents the x and y positions of all end effectors in the base frame, stacked into a single vector, and $\mathbf{r}_{w,z}$ denotes the height of the robot's center of mass in the world frame. The inclusion of ω_b distinguishes this formulation from that in [1]. This was with the intent to improve the model's performance in situations with high rotational velocities.

The network is trained on heuristically computed footstep cost maps (Figure 3.8b), which estimate the cost of moving each foot to candidate positions given the current robot state. These maps are represented as four 5×5 grids, one for each leg. This choice also differs from the implementation in [1]. This grid size was selected to provide more information for each foot, important later when sampling footstep candidates.

The architecture of the footstep evaluation network is shown in Figure 3.5. It consists of a feedforward neural network that initially maps the input through two fully connected layers of 64 nodes each, both with ReLU activations. The resulting 64-dimensional feature vector is reshaped into an 8×8 spatial representation and processed by a convolutional layer with two output channels, a 3×3 kernel, stride 1, and padding 1, followed by a ReLU nonlinearity. The output is flattened and passed through a final fully connected layer before being reshaped into a 5×5 grid. Again, this specific architecture differs from [1], where a larger network without the convolutional layer was used. These specific changes were found to provide better results for the 5×5 output grid.

To produce the four footstep candidate maps, this network is replicated four times—once per leg—with weights not shared between legs. This design allows the network to learn leg-specific behaviors, accommodating potential asymmetries in the robot's dynamics.

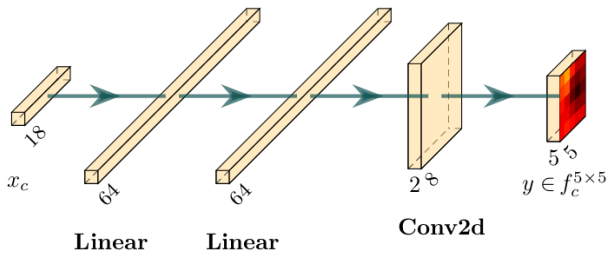


Figure 3.5: Footstep evaluation neural network architecture.

3.3.2 Training

The footstep evaluation network is trained to predict heuristic footstep cost maps, as described in [1]. Training data is generated using the simulation environment outlined in section 3.2, but with planar terrain. As in [1], planar terrain is sufficient for data collection because the cost maps are masked based on terrain after inference (see subsection 3.3.3).

During training, 100 robots are simulated in parallel, divided into four groups of 25. Each group tests a grid of footstep positions for one foot at a time; the front-left group is illustrated in Figure 3.6. An *iteration* is considered complete once all robots have either successfully completed or failed their assigned motions. Importantly, all robots begin each iteration from the same initial state.

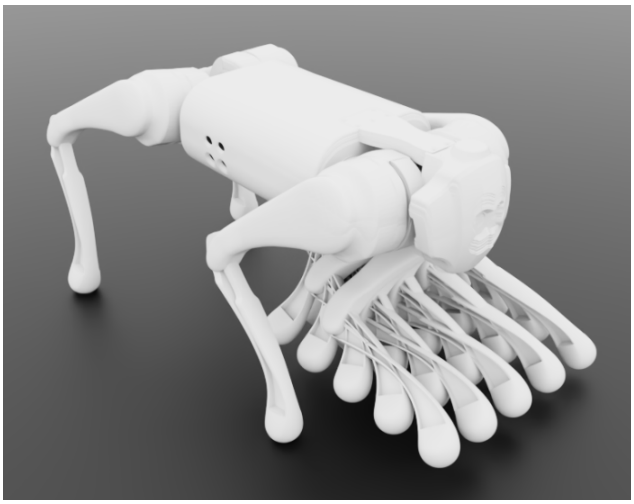


Figure 3.6: Snapshot showing 25 robots testing different footstep positions in parallel. For real data generation, 100 are run in parallel, testing 25 footstep positions for each foot at a time.

Data collection proceeds by chaining multiple iterations together, with control inputs periodically resampled from $\mathbf{u} \in (-0.2, 0.2) \times (-0.2, 0.2) \times (-0.4, 0.4)$ with "$\times$" being the cartesian product. After each iteration, the top 10 actions are inserted into a tree structure as edges, with the resulting state as the leaf node. The tree is explored by randomly selecting leaves for expansion, and this process continues until a predefined maximum number of iterations is reached. To ensure data quality, iterations in which more than 50% of the robots fall are discarded, along with their two parent nodes in the tree. This approach promotes the collection of diverse, yet successful, training data. This implementation differs in the specifics to [1], while still remaining faithful to the overall methodology.

Figure 3.7 illustrates the distribution of foot positions in the training dataset. Darker regions correspond to discrete points on the 5×5 footstep grids; a foot occupies a given position if it was moved there in that iteration.

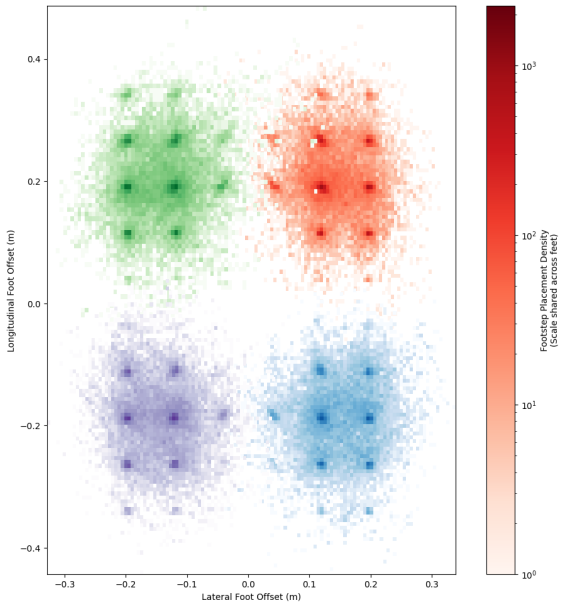
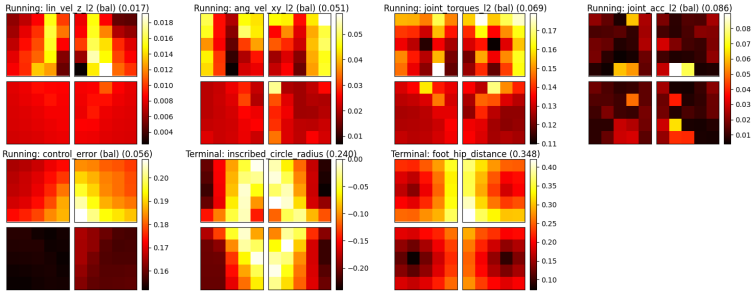


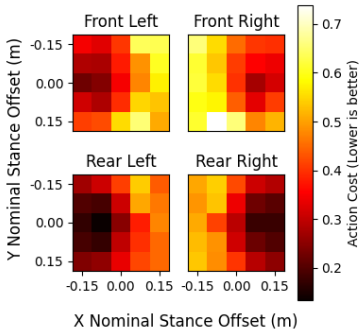
Figure 3.7: Foot placement heatmaps showing distribution of foot positions in the GaitNet training data. Note that the histograms are overlaid in some places, obscuring data underneath.

The purpose of this data collection is to assign a cost to each potential footstep, forming the candidate set $\mathbf{f_c}$. Costs are computed heuristically based on simulation outcomes, balancing stability and efficiency. While similar to the approach in [1], our heuristic differs in several aspects to better suit the output for the footstep candidate sampling. Figure 3.8a shows the individual factors used to construct the footstep candidate maps, and Figure 3.8b displays the resulting combined cost map. These factors are described below:

- *lin_vel_z_l2*—Penalizes high vertical velocity of the trunk.
- *ang_vel_xy_l2*—Penalizes high angular velocity of the trunk in the horizontal axes.
- *joint_torques_l2*—Penalizes high joint torques.
- *joint_acc_l2*—Penalizes high joint accelerations.
- *control_error*—Penalizes errors between the control input and actual robot motion.
- *inscribed_circle_radius*—Measures the distance from the COM to the nearest edge of the support polygon. This encourages the robot to keep its COM in a stable position.
- *foot_hip_distance*—Measures the distance between the hip and foot in the XY plane. This encourages the robot to keep its feet moving along with the body.



(a) Factors influencing footstep candidate maps. *(bal)* indicates that the values for each leg were balanced to have a lower spread, mitigating factors that consistently prefer one leg over another. The last number in parenthesis indicates the total range of the data, the most important factor for the combined cost map.



(b) Combined cost map from factors in (a) (not normalized).

Figure 3.8: Footstep candidate map composition. (a) shows the individual factors that make up the footstep candidate maps, while (b) shows the combined cost map.

As in [1], the cost maps are normalized to enhance training performance. Our approach differs in that the maps are normalized directly to the range $[0, 1]$, rather than preserving only the relative ordering of costs as in [1]. This direct normalization is crucial for providing the upstream GaitNet model with maximal information.

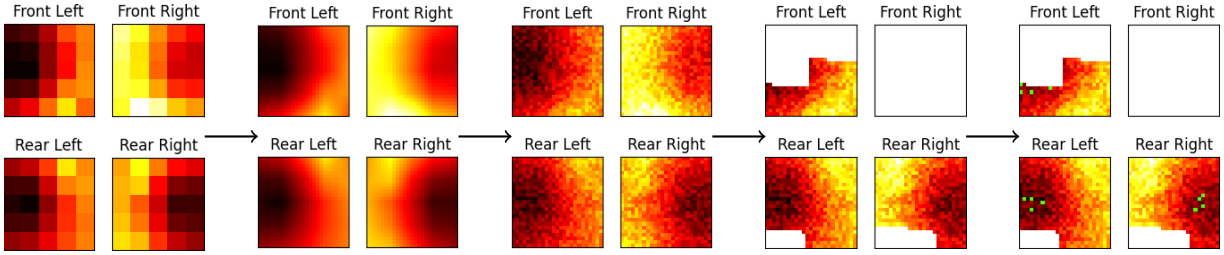
3.3.3 Post-Processing

The primary purpose of the footstep evaluation network is to provide high-quality footstep candidates to the GaitNet model. Figure 3.9 illustrates the processing pipeline.

The raw cost map output from the footstep evaluation network (Figure 3.9a) is first upsampled (Figure 3.9b) to increase the resolution of possible footstep positions and to match the resolution of the terrain data. Next, noise is added (Figure 3.9c) to encourage exploration of a more diverse set of footstep positions; without this noise, all candidates would cluster in close proximity.

TODO: update the number of candidates in text and diagram

The cost map is then filtered based on the robot state and terrain data (Figure 3.9d) to mask out invalid actions. This includes positions that are too close to terrain edges and movements that would attempt to reposition a leg already in the swing phase (as illustrated by the Front Right leg in the figure). Additional masking occurs based on the total number of legs in the swing phase; when 2 legs are already in swing, all remaining options are masked out to prevent unstable motions. Finally, the top 8 candidates from each leg are selected (Figure 3.9e) for processing by GaitNet. If fewer than 8 valid candidates exist for a given leg, no-action candidates are added to maintain consistent tensor dimensions.



(a) Raw (b) Upsample (c) Noise (d) C-space (e) TopK (green)

Figure 3.9: Cost map processing pipeline. Shows how the raw cost map is processed to produce the final footstep candidates.

3.4 GaitNet

GaitNet is designed to select the optimal footstep action from a discrete set of candidates. By restricting selection to a predefined set of continuous actions, the algorithm enforces constraints on the robot's motion, strictly preventing invalid foot placements and limiting the range of possible gaits. The inclusion of a no-action candidate, combined with continuous re-evaluation, distinguishes this approach: it enables the greedy generation of acyclic gaits in which multiple feet can be in the swing phase simultaneously.

3.4.1 Architecture

GaitNet is responsible for ranking footstep candidates based on the one-hot encoded leg index $\mathbf{f_c}'$ and the robot state $\mathbf{x_g}$:

$$\mathbf{x_g} = \begin{bmatrix} \mathbf{p}_{b,xy} \\ \mathbf{r}_{w,z} \\ \mathbf{v}_b \\ \omega_b \\ \mathbf{u} \\ \mathbf{g} \\ \mathbf{c} \end{bmatrix}$$

Descriptions of $\mathbf{p}_{b,xy}$ and $\mathbf{r}_{w,z}$ can be found in subsection 3.3.1. The additional terms, compared to the footstep evaluation network, are $\mathbf{g}$ and $\mathbf{c}$. Using $\mathbf{x_g}$ and $\mathbf{f_c}'$, GaitNet outputs a logit associated with each footstep and the desired swing duration if that step is selected.

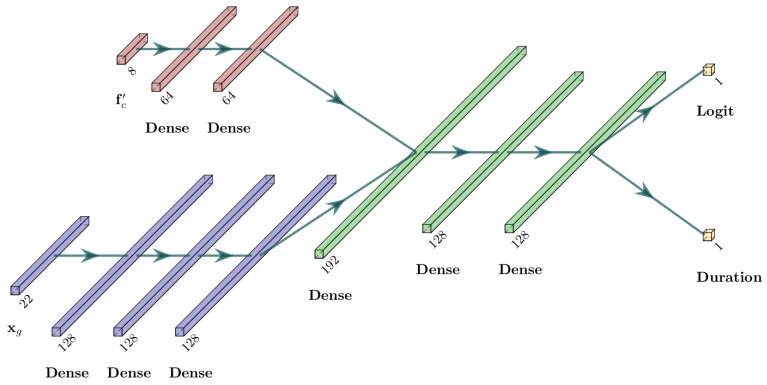


Figure 3.10: Gait net architecture. Sections of the diagram include the footstep candidate encoder in red, robot state encoder in blue, and shared trunk in green. The final output is a logit encoding the value of this option and the desired swing duration if this action is taken.

The model is designed to jointly evaluate robot state and footstep candidates with a middle fusion architecture [42]. The architecture (seen in Figure 3.10) consists of two encoders, a shared trunk, and two task-specific output heads:

- Robot state encoder: The robot state vector is processed by a three-layer feedforward network with intermediate dimensionality of 128 and ReLU activations. This produces a fixed-dimensional latent representation of the current robot state.
- Footstep encoder: Each footstep candidate is encoded by a two-layer feedforward network with hidden size 64 and ReLU activations. No-action candidates are represented through a fixed embedding.
- Shared trunk: The concatenated robot state and footstep embeddings are processed by a three-layer feedforward trunk, reducing dimensionality while applying Layer Normalization and ReLU nonlinearity.
- Output heads: Two parallel prediction heads are applied to the shared trunk:
  - Logit head: A single-layer network outputs the predicted reward value as a logit.
  - Duration head: A single-layer network with sigmoid activation outputs a normalized swing duration, which is then scaled to the range $(0.1, 0.3)$ s.

This design allows the network to sequentially evaluate multiple footstep candidates, assigning each an expected value and a feasible swing duration, while explicitly supporting a "no-action" option via a dedicated embedding.

3.4.2 Training

Due to the inclusion of the $\mathbf{c}$ term in the input, GaitNet cannot be effectively trained using supervised learning because of the high dimensionality of the problem. Instead, it is trained using PPO [43]. In this setup, the actor network is GaitNet (Figure 3.10), while the critic follows a similar architecture but omits the duration head. All logits from the critic are passed through a final MLP to produce a single value estimate. This MLP consists of two hidden layers of size 64, with ReLU activations applied to all layers except the output.

A custom actor-critic implementation was necessary to handle GaitNet's dual outputs. The logits define a categorical distribution over footstep candidates, while the duration output is treated as a separate normal distribution with a fixed standard deviation of 0.01 s for each action. To prevent multiple no-action candidates from skewing the selection, all but one no-action candidate are removed (set to $-\infty$).

The total log probability of an action is computed as the sum of the categorical log probability of the selected footstep candidate and the log probability of the duration under the normal distribution. During action sampling, a footstep candidate is first sampled from the categorical distribution, followed by sampling

the duration from its associated normal distribution. The footstep candidate and duration are then combined to form a complete footstep action.

Following the reward structure, termination conditions, commands, and hyperparameters outlined in Appendix A, GaitNet is trained for ?? environment steps using a batch size of ??.

4 Results and Discussion

4.1 Footstep Evaluation Network

The results of the Footstep Evaluation Network are highly promising, with the model able to predict footstep candidate maps with strong accuracy. Figure 4.1 shows the model output alongside the ground truth for a typical data sample.

Figure 4.2 illustrates a particularly challenging sample, in which the model still successfully identifies the most suitable positions for each leg. Notably, the back-left leg must be positioned far from its nominal location to maintain stability in this scenario, and the model correctly predicts this adjustment.

In the context of this work, the precise accuracy of the model is not critical. Its primary role is to sample the continuous space of foot positions to generate candidate actions for the GaitNet policy.

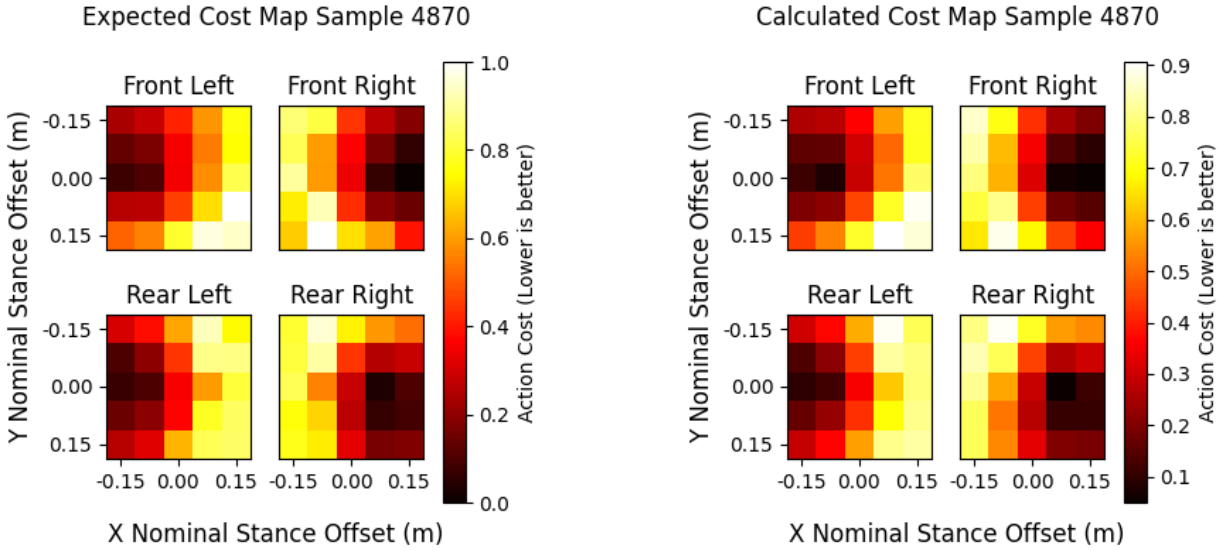


Figure 4.1: Typical data samples showing calculated (left) and expected (right) quadruped images.

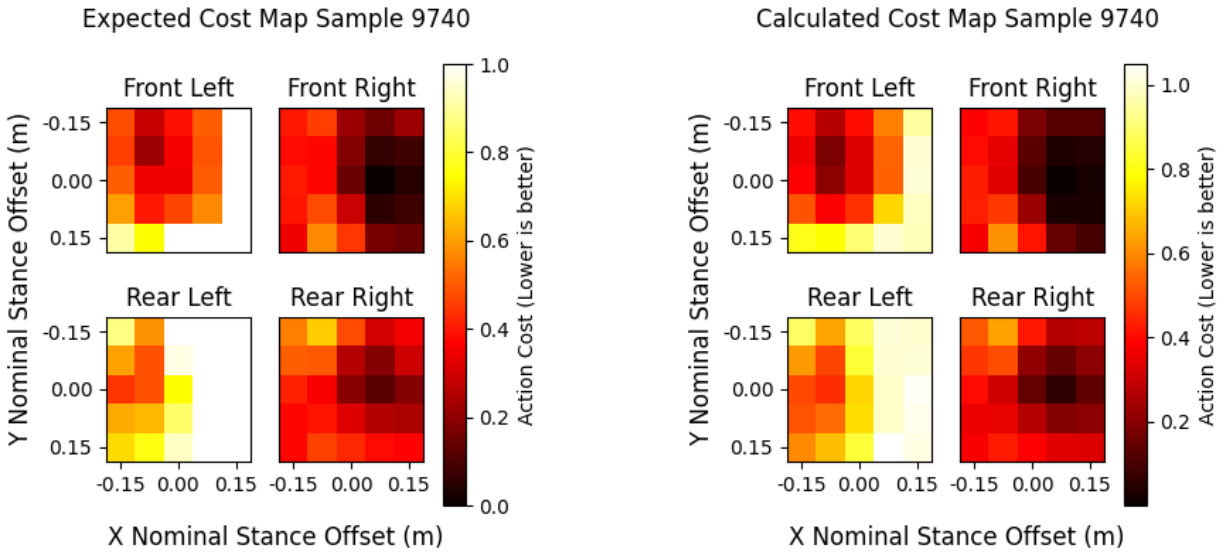


Figure 4.2: Particularly challenging data samples showing calculated (left) and expected (right) quadruped images.

4.2 GaitNet

The primary objective of GaitNet is to generate fully dynamic and acyclic gaits for quadruped robots. An example swing sequence is shown in Figure 4.3, illustrating a non-repetitive, dynamic gait in which the robot performs motions with up to two feet off the ground simultaneously. Swing durations are non-uniform, with some legs remaining in the swing phase longer than others. These results demonstrate that the system can synthesize a wide variety of actions, which are analyzed further below.

Examining Figure 4.3, a short step is observed for the front-left leg (green) at 3 s, while the rear-right leg (red) executes a longer step at 5 s. Although the differences in duration are subtle, they indicate that the system is capable of dynamically varying step timing.

Another important observation from Figure 4.3 is the system's response to changing control inputs. Between 0-7.5 s, the robot is commanded to follow a slow input, and the system takes a cautious approach, moving one foot at a time. After 7.5 s, the input speed increases, prompting a more dynamic gait. During this faster phase, the robot occasionally executes two steps simultaneously and does not follow a strict alternating pattern, further highlighting the system's flexibility in generating acyclic gaits.

TODO: Modify figure Figure 4.3 to show the terrain/robot at multiple time stamps.

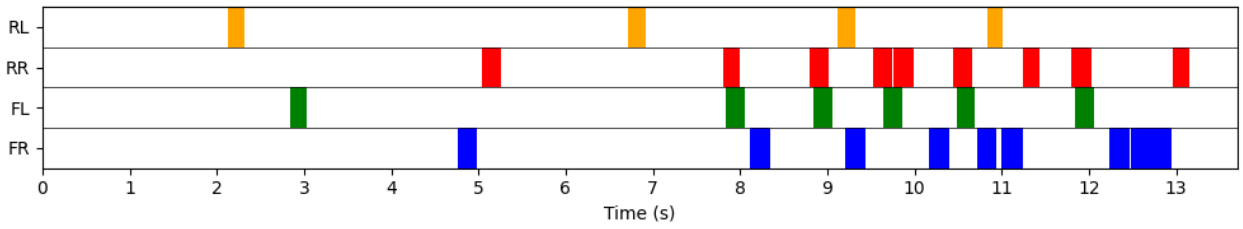


Figure 4.3: Example swing schedule for a single gait cycle. Each row represents a leg, with color indicating the leg is in swing phase.

TODO: Include figure showing the distribution of swing durations.

TODO: Include training figures from tensorboard detailing how the model improves over time.

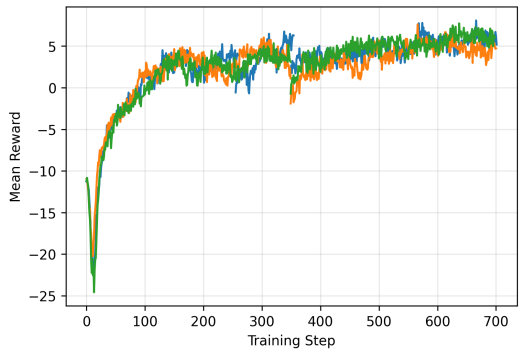


Figure 4.4: Mean episode reward during GaitNet training. Each is a different training instance.

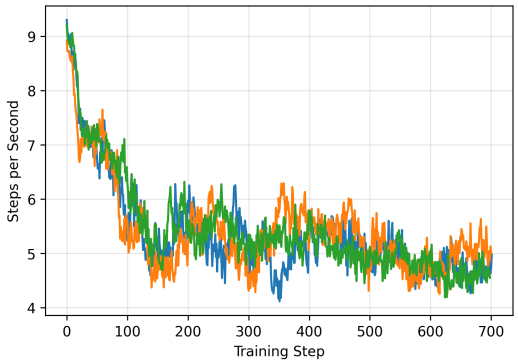


Figure 4.5: Mean steps per second during GaitNet training. Each is a different training instance.

4.3 Baseline Comparison

For the purpose of evaluating GaitNet's performance, a baseline method is established using a single leg motion planner. This planner operates as is described in [1], where in place of ContactNet, we directly use our footstep evaluation network described in section 3.3, excluding the noise post-processing step (Figure 3.9c), and only picking one candidate in the selection step (Figure 3.9e). The lowest cost candidate is then used as the footstep target for that leg, with a 200ms swing duration.

In order to benchmark the two methods, a custom test environment is created (Figure 4.6). This environment has the robots navigate straight forward across narrow strips of terrain with characteristics matching the sub-terrains used in training (section 3.2). The space of terrain difficulties and commanded forward velocities is systematically explored with a grid search, evaluating the success rate of each method over 150 trials of 20 s each. An episode is considered successful if the robot is able to complete the full 20 s without terminating according to the conditions described in section A.2.

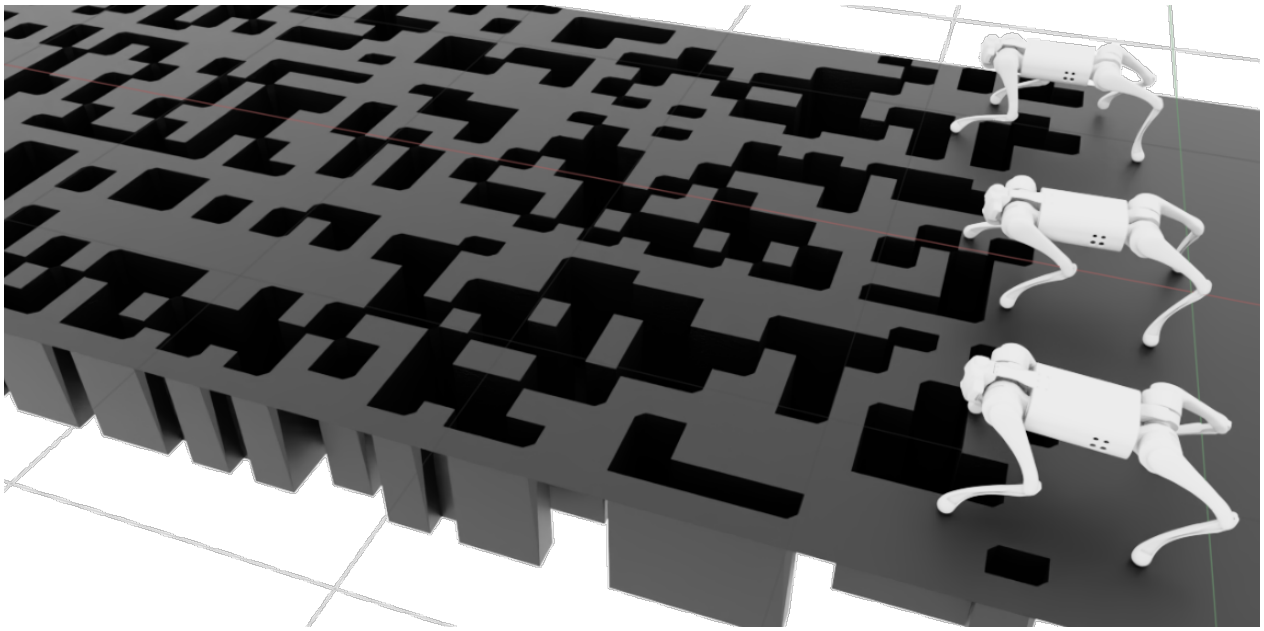


Figure 4.6: Three robots simultaneously navigating a test environment (40% terrain difficulty) used for baseline comparison. The terrain consists of narrow strips of a grid pattern with missing sections. Density of missing sections (difficulty) and commanded forward velocity are varied to evaluate performance.

Figure 4.7 and Figure 4.8 illustrate the performance of GaitNet and the baseline method, respectively, across a range of terrain difficulties and commanded velocities. The results indicate that GaitNet outperforms the baseline in most scenarios, particularly on more challenging terrains and at higher commanded speeds. This demonstrates the effectiveness of GaitNet in generating robust gaits capable of adapting to varying conditions.

A closer examination of the graphs reveals that GaitNet maintains strong performance when the commanded velocity is below 0.15 m/s or terrain difficulty is under 5%. In contrast, ContactNet's performance begins to degrade for commanded velocities above 0.05 m/s and deteriorates further as terrain difficulty increases. GaitNet's superior performance in these scenarios underscores the benefits of its dynamic, acyclic gait generation capabilities.

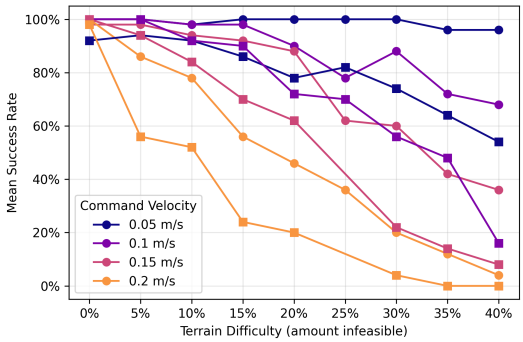


Figure 4.7: GaitNet Evaluation. Overall survival rate of 69.4%. Mean success rate measured as the percentage of 150 trials which completed 20 s without terminating, under the termination conditions described in section A.2. Data point shapes denote different training instances.

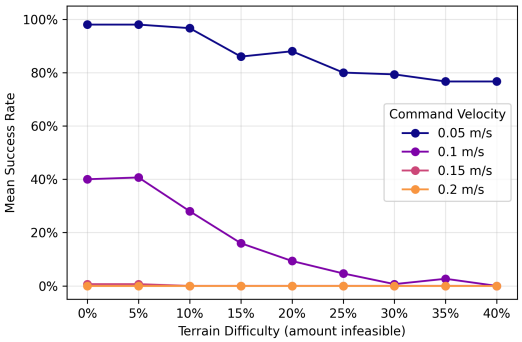


Figure 4.8: Single Leg Motion Planner Evaluation. Overall survival rate of 25.6%. Mean success rate measured as the percentage of 150 trials which completed 20 s without terminating, under the termination conditions described in section A.2.

4.4 Swing Duration Ablation Study

In this section we present an ablation study to assess the impact of dynamic swing duration on GaitNet's performance. We compare two models: the standard GaitNet formulation as described in section 3.4, and a Duration-Ablated-GaitNet trained with a fixed swing duration.

During training, the GaitNet model learns to adjust swing durations based on its input, as illustrated with Figure 4.9 and Figure 4.10. The mean swing duration settles to 0.24 s with a standard deviation of 0.034 s, indicating that the model uses a wide range of its allowable swing durations. These figures highlight how the model learns to effectively utilize dynamic swing durations to adapt to varying conditions.

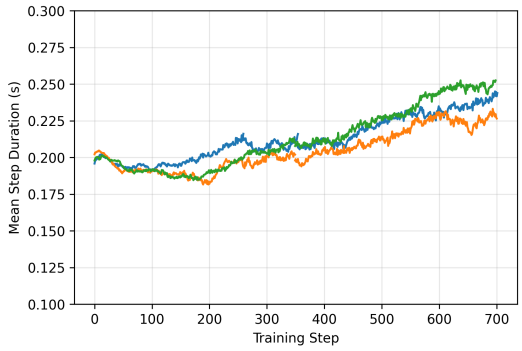


Figure 4.9: Mean swing duration across all environments during GaitNet training. Each color is a different training instance.

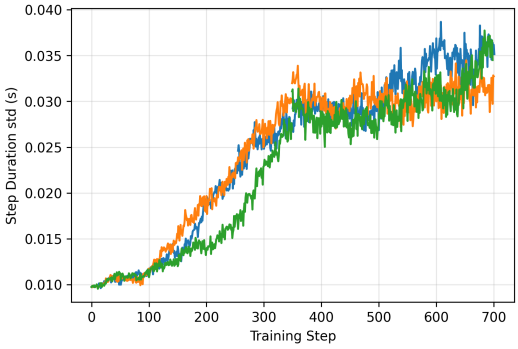


Figure 4.10: Swing duration standard deviation across all environments during GaitNet training. Each color is a different training instance.

Furthermore, a histogram of swing durations selected by GaitNet during training, shown in Figure 4.11, reveals a broad distribution of values. This indicates that the model is not biased towards a narrow range of swing durations, but rather effectively leverages the flexibility provided by dynamic swing durations to optimize its gait.

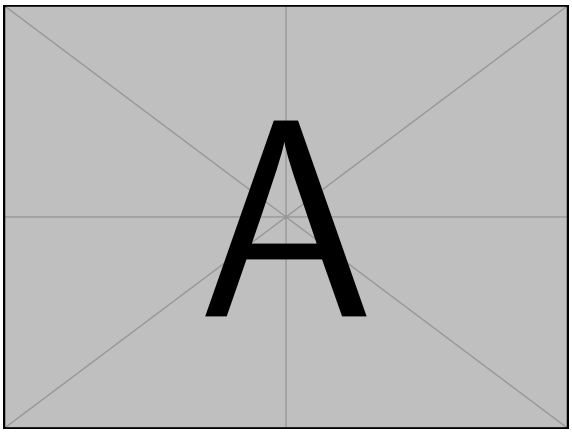


Figure 4.11: Histogram of swing durations selected by GaitNet during evaluation, showing a wide distribution of values.

Now, we train the Duration-Ablated-GaitNet model, setting the swing duration to a constant 0.24 s. The training process is identical to that of the standard GaitNet model, ensuring a fair comparison. The performance of both models is then evaluated using the same metrics outlined in section 4.3.

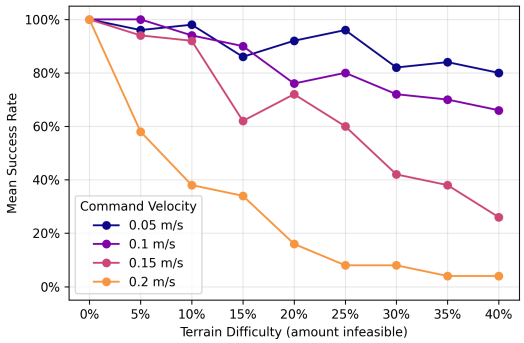


Figure 4.12: Evaluation Duration-Ablated-GaitNet across various terrain difficulties and commanded velocities. Overall survival rate of 67.2%. Mean success rate measured as the percentage of 50 episodes which completed 20 s without terminating, under the termination conditions described in section A.2.

Figure 4.12 presents the performance comparison between the standard GaitNet and the duration-ablated variant. The results indicate that the standard GaitNet outperforms the duration-ablated model across a range of terrain difficulties and commanded velocities. This demonstrates the effectiveness of dynamic swing duration in enhancing GaitNet's ability to adapt to varying conditions and generate robust gaits.

4.5 Action Cost Ablation Study

In this section we present an ablation study to assess the impact of the footstep candidate cost on GaitNet's performance. We compare two models: the standard GaitNet formulation as described in section 3.4, and a Cost-Ablated-GaitNet trained without the footstep candidate (f_c) cost in the input.

During training, the GaitNet model quickly learns to correlate the lower cost candidates with higher value logits. After this initial learning phase, the model learns to refine its predictions based on other features in the input. This is illustrated in Figure 4.13.

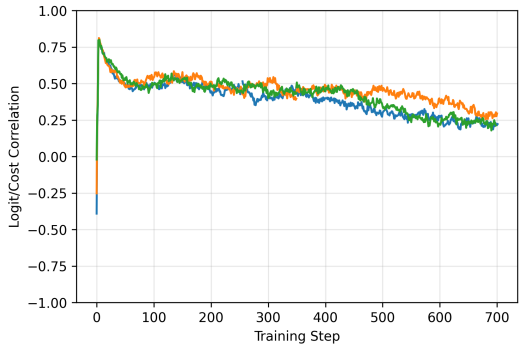


Figure 4.13: Correlation between negative footstep candidate cost and GaitNet output logits during training.

Now, we train the Cost-Ablated-GaitNet model, over-writing all footstep candidate cost values with zero. The training process is identical to that of the standard GaitNet model, ensuring a fair comparison. The performance of both models is then evaluated using the same metrics outlined in section 4.3.

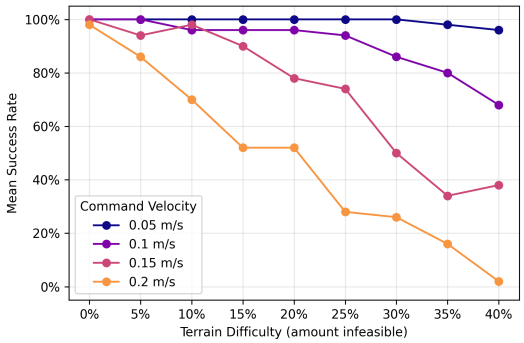


Figure 4.14: Evaluation of Cost-Ablated-GaitNet across various terrain difficulties and commanded velocities. Overall survival rate of 77.7%. Mean success rate measured as the percentage of 50 episodes which completed 20 s without terminating, under the termination conditions described in section A.2.

TODO: if the action cost and logits are very poorly correlated, then there is possible improvement in the candidate action sampling method.

Figure 4.14 presents the performance comparison between the standard GaitNet and the cost-ablated variant. The results indicates that there is minimal difference in performance between the two models. However, looking at the episode reward for living (Figure 4.15), the cost-ablated model can be seen to learn more slowly than the standard GaitNet. This indicates that the footstep candidate cost is useful during the initial stages of training to help the model quickly learn the robot dynamics.

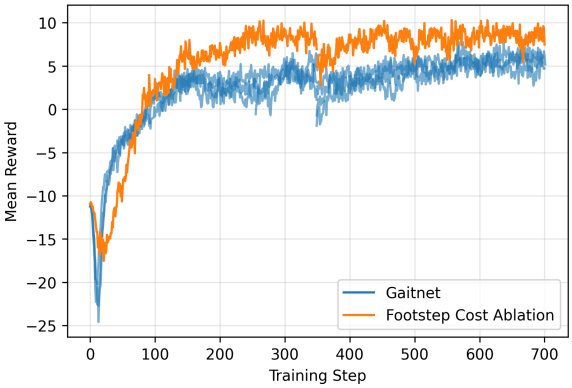


Figure 4.15: Episode reward for living during training for GaitNet and Cost-Ablated-GaitNet.

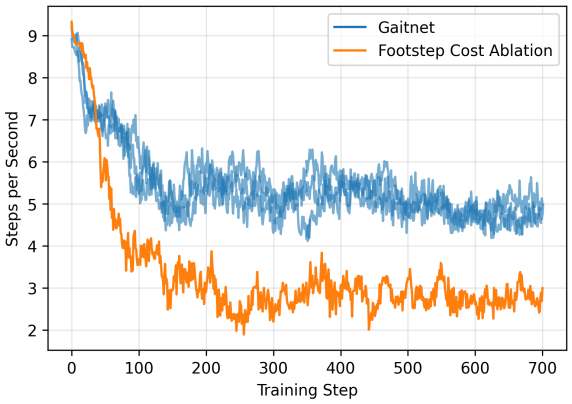


Figure 4.16: Comparison of steps per second during training for GaitNet and Cost-Ablated-GaitNet.

4.6 Discussion

TODO: summarize findings from ablation studies and baseline comparison

5 Conclusions and Future Work

5.1 Summary of Findings

5.2 Limitations

While the results presented in this work are promising, several limitations should be acknowledged. First, the proposed framework has only been validated in simulation. Real-world deployment may introduce additional challenges, such as unmodeled dynamics, sensor noise, and hardware constraints, which could affect performance. Second, GaitNet operates in a greedy, no-lookahead fashion, limiting its effectiveness at higher speeds or in highly dynamic environments where predictive planning could improve stability and efficiency. Additionally, the system's performance is constrained by the accuracy of the low-level controller in executing footstep placements. Errors in tracking or positioning can degrade the overall effectiveness of the generated gaits. Finally, GaitNet's performance is inherently tied to the quality of footstep candidates produced by the footstep evaluation network; suboptimal candidate sampling may restrict both the diversity and effectiveness of the generated actions.

These limitations highlight clear directions for future research, including real-world testing on physical hardware, the incorporation of predictive planning strategies, and improvements to both candidate generation and low-level control precision.

5.3 Future Work

Several avenues exist to extend and improve the current framework. First, real-world deployment remains a crucial next step. Testing GaitNet on physical hardware would expose the system to real-world dynamics, sensor noise, and hardware constraints, providing valuable insights for further refinement.

Swing re-planning is another potential improvement. Currently, GaitNet does not re-plan during the swing phase. Incorporating real-time re-planning would allow the robot to adjust to unexpected disturbances or terrain changes, though this would require a more robust low-level controller capable of accurately tracking the modified footstep trajectories.

Long-horizon planning also presents an opportunity for enhancement. GaitNet is presently trained to generate actions for a single point in time, which limits the MPC's ability to predict future contact states accurately; the MPC expects the robot to revert to a nominal stance after each swing. While this is acceptable at low speeds, it could become problematic at higher velocities, where predictive planning would improve stability and performance.

Currently, the MPC runs on the CPU, limiting the speed of reinforcement learning. Leveraging GPU-accelerated MPCs, as explored in recent works [44, 45], could significantly accelerate training.

GaitNet also selects only one action at a time, leading to slightly staggered swing start times. Initial attempts to select multiple simultaneous actions encountered difficulties with gradient flow and learning stability. Developing a reliable method for multi-action selection could enable more fluid and dynamic gait patterns, though it would require careful network and training design.

Finally, improvements to the action candidate sampler could enhance GaitNet's performance. At present, action selection relies on sampling-based methods, which may not always identify the highest-quality actions. Directly searching the action space using techniques such as projected gradient descent or other optimization

strategies could improve solution quality at the expense of additional computation. Another possibility is to train a candidate selection network jointly with GaitNet to improve the diversity and relevance of sampled actions.

5.4 Final Remarks

References

[1] A. Bratta, A. Meduri, M. Focchi, L. Righetti, and C. Semini, "ContactNet: Online multi-contact planning for acyclic legged robot locomotion,"

[2] H. Chai, Y. Li, R. Song, G. Zhang, Q. Zhang, S. Liu, J. Hou, Y. Xin, M. Yuan, G. Zhang, and Z. Yang, "A survey of the development of quadruped robots: Joint configuration, dynamic locomotion control method and mobile manipulation approach," *Biomimetic Intelligence and Robotics*, vol. 2, p. 100029, Mar. 2022.

[3] Y. Fan, Z. Pei, C. Wang, M. Li, Z. Tang, and Q. Liu, "A Review of Quadruped Robots: Structure, Control, and Autonomous Motion," *Adv. Intell. Syst.*, vol. 6, p. 2300783, June 2024.

[4] D. Kim, J. Di Carlo, B. Katz, G. Bledt, and S. Kim, "Highly Dynamic Quadruped Locomotion via Whole-Body Impulse Control and Model Predictive Control," *arXiv*, Sept. 2019.

[5] P. M. Wensing, M. Posa, Y. Hu, A. Escande, N. Mansard, and A. Del Prete, "Optimization-Based Control for Dynamic Legged Robots," *arXiv*, Nov. 2022.

[6] M. Geisert, T. Yates, A. Orgen, P. Fernbach, and I. Havoutis, "Contact Planning for the ANYmal Quadruped Robot using an Acyclic Reachability-Based Planner," *arXiv*, Apr. 2019.

[7] A. W. Winkler, C. D. Bellicoso, M. Hutter, and J. Buchli, "Gait and trajectory optimization for legged systems through phase-based end-effector parameterization," vol. 3, no. 3, pp. 1560–1567. Publisher: Institute of Electrical and Electronics Engineers (IEEE).

[8] M. Gurram, P. K. Uttam, and S. S. Ohol, "Reinforcement Learning For Quadrupedal Locomotion: Current Advancements And Future Perspectives," *arXiv*, Oct. 2024.

[9] L. Bao, J. Humphreys, T. Peng, and C. Zhou, "Deep Reinforcement Learning for Bipedal Locomotion: A Brief Survey," *arXiv*, Apr. 2024.

[10] Z. Wang, W. Wei, A. Xie, Y. Zhang, J. Wu, and Q. Zhu, "Hybrid Bipedal Locomotion Based on Reinforcement Learning and Heuristics," *Micromachines*, vol. 13, p. 1688, Oct. 2022.

[11] *GLiDE: Generalizable Quadrupedal Locomotion in Diverse Environments with a Centroidal Model*, pp. 523–539. Springer International Publishing. ISSN: 2511-1256, 2511-1264.

[12] R. Grandia, F. Jenelten, S. Yang, F. Farshidian, and M. Hutter, "Perceptive locomotion through nonlinear model predictive control."

[13] J. Lee, J. Hwangbo, L. Wellhausen, V. Koltun, and M. Hutter, "Learning quadrupedal locomotion over challenging terrain," vol. 5, no. 47.

[14] O. Villarreal, V. Barasuol, M. Camurri, L. Franceschi, M. Focchi, M. Pontil, D. G. Caldwell, and C. Semini, "Fast and continuous foothold adaptation for dynamic locomotion through CNNs," vol. 4, no. 2, pp. 2140–2147.

[15] L. Amatucci, J.-H. Kim, J. Hwangbo, and H.-W. Park, "Monte carlo tree search gait planner for non-gaited legged system control."

[16] I. Taouil, L. Amatucci, M. Khadiv, A. Dai, V. Barasuol, G. Turrisi, and C. Semini, "Non-gaited legged locomotion with monte-carlo tree search and supervised learning," vol. 10, no. 2, pp. 1265–1272. Publisher: Institute of Electrical and Electronics Engineers (IEEE).

[17] Y. Meng and C. Fan, "Hybrid Systems Neural Control with Region-of-Attraction Planner," *arXiv*, Mar. 2023.

[18] H. Shi, Q. Zhu, L. Han, W. Chi, T. Li, and M. Q.-H. Meng, "Terrain-aware quadrupedal locomotion via reinforcement learning."

[19] H. Duan, A. Malik, J. Dao, A. Saxena, K. Green, J. Siekmann, A. Fern, and J. Hurst, "Sim-to-real learning of footstep-constrained bipedal dynamic walking," in *2022 International Conference on Robotics and Automation (ICRA)*, pp. 10428–10434, IEEE.

[20] J. Siekmann, K. Green, J. Warila, A. Fern, and J. Hurst, "Blind bipedal stair traversal via sim-to-real reinforcement learning."

[21] X. Da, Z. Xie, D. Hoeller, B. Boots, A. Anandkumar, Y. Zhu, B. Babich, and A. Garg, "Learning a Contact-Adaptive Controller for Robust, Efficient Legged Locomotion," *arXiv*, Sept. 2020.

[22] Y. Yang, T. Zhang, E. Coumans, J. Tan, and B. Boots, "Fast and Efficient Locomotion via Learned Gait Transitions," *arXiv*, Apr. 2021.

[23] H. Sun, J. Yang, Y. Jia, and C. Wang, "Online Hierarchical Planning for Multicontact Locomotion Control of Quadruped Robots," *IEEE/ASME Trans. Mechatron.*, vol. 30, pp. 1718–1728, July 2024.

[24] C. Zhang, N. Rudin, D. Hoeller, and M. Hutter, "Learning Agile Locomotion on Risky Terrains," *arXiv*, Nov. 2023.

[25] D. Zhang, X. Chen, Z. Zhong, M. Xu, Z. Zheng, and H. Lu, "A novel multi-gait strategy for stable and efficient quadruped robot locomotion."

[26] R. Deits and R. Tedrake, "Footstep planning on uneven terrain with mixed-integer convex optimization," in *2014 IEEE-RAS International Conference on Humanoid Robots*, pp. 18–20, IEEE.

[27] B. Aceituno-Cabezas, C. Mastalli, H. Dai, M. Focchi, A. Radulescu, D. G. Caldwell, J. Cappelletto, J. C. Grieco, G. Fernandez-Lopez, and C. Semini, "Simultaneous Contact, Gait and Motion Planning for Robust Multi-Legged Locomotion via Mixed-Integer Convex Optimization," *arXiv*, Apr. 2019.

[28] R. Akizhanov, V. Dhédin, M. Khadiv, and I. Laptev, "Learning feasible transitions for efficient contact planning," *arXiv*, July 2024.

[29] C. Gaspard, G. Passault, M. Daniel, and O. Ly, "FootstepNet: an efficient actor-critic method for fast on-line bipedal footstep planning and forecasting," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 13749–13756, IEEE.

[30] V. Tsounis, M. Alge, J. Lee, F. Farshidian, and M. Hutter, "DeepGait: Planning and control of quadrupedal gaits using deep reinforcement learning."

[31] S. Omar, L. Amatucci, G. Turrisi, V. Barasuol, and C. Semini, "Fast convex visual foothold adaptation for quadrupedal locomotion,"

[32] X. B. Peng, G. Berseth, K. Yin, and M. Van De Panne, "DeepLoco: dynamic locomotion skills using hierarchical deep reinforcement learning," vol. 36, no. 4, pp. 1–13. Publisher: Association for Computing Machinery (ACM).

[33] L. Chen, P. Du, P. Zhan, and B. Xie, "Gait Learning for Hexapod Robot Facing Rough Terrain Based on Dueling-DQN Algorithm," *International Journal of Computer Science and Information Technology*, vol. 2, pp. 408–424, Mar. 2024.

[34] Q. Yao, J. Wang, D. Wang, S. Yang, H. Zhang, and Y. Wang, "Hierarchical Terrain-Aware Control for Quadrupedal Locomotion by Combining Deep Reinforcement Learning and Optimal Control," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 2021–01, IEEE.

[35] A. Meduri, M. Khadiv, and L. Righetti, "DeepQ Stepper: A framework for reactive dynamic walking on uneven terrain," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2021–05, IEEE.

[36] Z. Gao, X. Chen, Z. Yu, C. Li, L. Han, and R. Zhang, "Global footstep planning with greedy and heuristic optimization guided by velocity for biped robot," vol. 238, p. 121798. Publisher: Elsevier BV.

[37] M. Zucker, N. Ratliff, M. Stolle, J. Chestnutt, J. A. Bagnell, C. G. Atkeson, and J. Kuffner, "Optimization and learning for rough terrain legged locomotion," vol. 30, no. 2, pp. 175–191. Publisher: SAGE Publications.

[38] M. Kalakrishnan, J. Buchli, P. Pastor, M. Mistry, and S. Schaal, "Learning, planning, and control for quadruped locomotion over challenging terrain," vol. 30, no. 2, pp. 236–258. Publisher: SAGE Publications.

[39] M. Asselmeier, Y. Zhao, and P. A. Vela, "Steppability-informed quadrupedal contact planning through deep visual search heuristics."

[40] S. Omar, L. Amatucci, V. Barasuol, G. Turrisi, and C. Semini, "SafeSteps: Learning safer footstep planning policies for legged robots via model-based priors," in *2023 IEEE-RAS 22nd International Conference on Humanoid Robots (Humanoids)*, pp. 1–8, IEEE.

[41] M. Mittal, C. Yu, Q. Yu, J. Liu, N. Rudin, D. Hoeller, J. L. Yuan, R. Singh, Y. Guo, H. Mazhar, A. Mandlekar, B. Babich, G. State, M. Hutter, and A. Garg, "Orbit: A unified simulation framework for interactive robot learning environments," vol. 8, no. 6, pp. 3740–3747.

[42] D. Feng, C. Haase-Schütz, L. Rosenbaum, H. Hertlein, C. Gläser, F. Timm, W. Wiesbeck, and K. Dietmayer, "Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 3, pp. 1341–1360, 2021.

[43] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal Policy Optimization Algorithms," *arXiv*, July 2017.

[44] A. L. Bishop, J. Z. Zhang, S. Gurumurthy, K. Tracy, and Z. Manchester, "ReLU-QP: A GPU-Accelerated Quadratic Programming Solver for Model-Predictive Control," *arXiv*, Nov. 2023.

[45] B. Plancher, *GPU Acceleration for Real-time, Whole-body, Nonlinear Model Predictive Control*. PhD thesis, Harvard University, Cambridge, MA, USA, April 2022.

A GaitNet Training Configuration

A.1 Reward Function Analysis

Designing the reward functions for training GaitNet (see Table A.1) proved to be a challenging task. Various combinations of rewards often resulted in undesirable behaviors, such as the robot frequently lifting its feet while idle or dragging feet during locomotion. Achieving a natural gait required careful balancing, particularly between *op_reward* and *mdp.foot_slip_penalty*.

Function[1]	**Weight**	**Description**
mdp.is_alive	0.4	Reward for being alive
mdp.track_lin_vel_xy_exp	0.5	Reward for tracking linear velocity in the XY plane
mdp.track_ang_vel_z_exp	0.5	Reward for tracking angular velocity around the Z axis
op_reward	-0.1	Penalty for performing actions
mdp.is_terminated	-200	Penalty for termination
mdp.lin_vel_z_l2	-2.5	Penalty for linear velocity in the Z direction
mdp.ang_vel_xy_l2	-0.1	Penalty for angular velocity in the XY plane
mdp.flat_orientation_l2	-8	Penalty for non-flat orientation
mdp.foot_slip_penalty	-6	Penalty for foot slip

Table A.1: Reward functions used to train GaitNet.

In addition to the rewards listed in Table A.1, several other reward functions were explored but ultimately not used.

The *a_foot_in_swing* reward was intended to encourage the agent to frequently lift its feet during early learning. However, it proved unnecessary, as the initial network weights already favored foot movement. Furthermore, including this reward caused the agent to move its feet excessively when idle.

The *no_op_reward* was designed to encourage the agent to remain idle when no useful actions were available. In practice, this reward required an excessively high weighting to have any effect, at which point it would overshadow other rewards. The *op_reward* was found to be a more effective alternative for achieving the desired behavior.

A.2 Termination Functions

The termination functions used in training GaitNet are summarized in Table A.2. These functions help define the conditions under which an episode ends, either applying the termination penalty or simply signaling an invalid state.

[1]Functions named "mdp.*" are built-in functions provided by the NVIDIA Isaac Lab framework.
[2]Functions named "mdp.*" are built-in functions provided by the NVIDIA Isaac Lab framework.
[3]Time Out indicates whether the termination applies the mdp.is_terminated penalty.

Function[2]	Time Out[3]	Description
mdp.time_out	True	Terminate at the end of the episode
mdp.bad_orientation	False	Terminate if the robot's orientation is too far from upright
mdp.root_height_below_minimum	False	Terminate if the robot's base height is too low
mdp.terrain_out_of_bounds	True	Terminate if the robot leaves the terrain bounds
foot_in_void	False	Terminate if any foot steps into the void

Table A.2: Termination functions used to train GaitNet.

A.3 Commands

The command used in training GaitNet is summarized below. This command defines the highest level input to the environment, $\mathbf{u}$.

- **UniformVelocityCommand**: This command samples a desired velocity vector $\mathbf{v} = [v_x, v_y, \omega_z]$ from a uniform distribution.
  - resampling time range: (2.5, 10) s
  - v_x range: (-0.2, 0.2) m/s
  - v_y range: (-0.2, 0.2) m/s
  - ω_z range: (-0.4, 0.4) rad/s
  - probability of zero command: 0.05

A.4 PPO Hyperparameters

The PPO hyperparameters used for training GaitNet are summarized in Table A.3. While not all parameters were rigorously tuned, they were found to perform well in practice. The discount factor *gamma* was specifically selected in relation to the agent observation frequency of 25,Hz to provide a reasonable effective horizon. The *learning_rate* was chosen relatively high to compensate for the slower data collection speed.

Hyperparameter	Value
clip_param	0.3
num_learning_epochs	8
num_mini_batches	4
value_loss_coef	0.5
entropy_coef	0.02
learning_rate	3e-4
max_grad_norm	1.0
use_clipped_value_loss	True
gamma	0.99
lam	0.95

Table A.3: Hyperparameters used for PPO training.