Artificial Intelligence and Robotics Research 人工智能与机器人研究, 2025, 14(4), 1052-1063
Published Online July 2025 in Hans. https://www.hanspub.org/journal/airr
https://doi.org/10.12677/airr.2025.144100

Hans汉斯

四足机器人复杂地形行走与摔倒恢复的统一控制方法

——基于深度强化学习的研究

孔德然，范　永*，赵荣华

山东交通学院轨道交通学院，山东　济南

收稿日期：2025年5月1日；录用日期：2025年7月22日；发布日期：2025年7月31日

摘　要

随着四足机器人在非结构化环境中的应用需求日益增加，如何应对外部扰动和复杂地形引发的失衡与摔倒，成为了机器人自主性和任务连续性面临的重要挑战。传统的摔倒恢复方法依赖外部干预，限制了机器人的自主性，难以满足复杂应用中的需求。尽管深度强化学习在运动控制方面取得了一定进展，但在摔倒后的自主恢复和基于地形特征的动态恢复策略方面的研究仍显不足。本文提出了一种基于深度强化学习的统一运动与恢复控制方法，使四足机器人能够在复杂地形中行走，并自主恢复摔倒。该方法结合了摔倒恢复因子、动态增长策略和安全约束优化，解决了现有方法中的不足。实验表明，机器人能够在不同地形条件下快速恢复并稳定过渡到行走状态，表现出较强的适应性和鲁棒性。本研究为四足机器人在高风险应用中的自主执行能力提供了有效的解决方案。

关键词

四足机器人，深度强化学习，复杂地形，摔倒恢复，NP3O

Unified Control Method for Complex Terrain Locomotion and Fall Recovery of Quadruped Robots

—A Study Based on Deep Reinforcement Learning

Deran Kong, Yong Fan*, Ronghua Zhao

School of Rail Transportation, Shandong Jiaotong University, Jinan Shandong

Received: May 1st, 2025; accepted: Jul. 22nd, 2025; published: Jul. 31st, 2025

*通讯作者。

文章引用: 孔德然, 范永, 赵荣华. 四足机器人复杂地形行走与摔倒恢复的统一控制方法[J]. 人工智能与机器人研究, 2025, 14(4): 1052-1063. DOI: 10.12677/airr.2025.144100

Abstract

With the growing demand for quadruped robots operating in unstructured environments, addressing instability and falls caused by external disturbances and complex terrains has become a critical challenge for ensuring robot autonomy and mission continuity. Traditional fall recovery methods often rely on external interventions, which limit the autonomy of the robot and fail to meet the needs of complex, real-world applications. Although deep reinforcement learning (DRL) has made notable progress in motion control, research on autonomous post-fall recovery and dynamic recovery strategies based on terrain features remains limited. In this paper, we propose a unified locomotion and recovery control framework based on deep reinforcement learning, enabling quadruped robots to walk over complex terrains and autonomously recover from falls. The framework integrates a fall recovery factor, a dynamic scheduling strategy, and safety-constrained optimization to address the limitations of existing approaches. Specifically, a non-symmetric actor-critic architecture is adopted, enhanced with a context-aided estimator to improve terrain-aware decision-making. Additionally, a dynamic β-VAE latent constraint strategy is introduced to facilitate stable training, while the NP3O algorithm ensures safe and efficient policy optimization under torque and stability constraints. Extensive experiments demonstrate that the proposed method enables quadruped robots to quickly recover from falls under various terrain conditions and transition smoothly back into locomotion. The robots exhibit strong adaptability and robustness, significantly improving their operational autonomy in high-risk environments. This study provides an effective solution for enhancing the autonomous capabilities of quadruped robots in real-world applications involving challenging and hazardous terrains.

Keywords

Quadruped Robot, Deep Reinforcement Learning, Complex Terrain, Fall Recovery, NP3O

1. 引言

随着四足机器人在非结构化环境(如陡坡、瓦砾、楼梯)中的需求不断增加，如何应对外部干扰和复杂地形引发的失衡或摔倒问题成为了一个重要挑战。传统的摔倒恢复方法通常依赖人工干预，这限制了机器人的自主性和任务执行的连续性。尽管深度强化学习在提高四足机器人在复杂地形中的运动能力和鲁棒性方面取得了显著进展[1]-[6]，但在摔倒后的自主恢复方面仍然存在不足。现有方法缺乏一种统一的控制策略，既能在不同地形下实现行走，又能确保可靠地恢复。此外，在规划基于地形条件的最优恢复动作时，还需要平衡速度、能量效率和稳定性，这一问题依然具有较大挑战。

近年来，关于四足机器人摔倒恢复的研究取得了进展。2019 年，Laura Smith 等人[7]提出了一个模拟到现实的框架，将摔倒恢复任务分解为站立、自我恢复和行走等子行为。然而，该方法仅适用于平坦地形，并且由于需要手动进行状态切换，任务执行效率较低，且在摔倒相关场景中的响应速度较慢。2021 年，Smith 等人[8]提出了一个结合模拟预训练与现实世界微调的方法，使机器人能够在多种平坦地形上快速恢复，并表现出较强的现实世界性能。然而，该方法在复杂地形中的泛化能力尚未得到充分验证。2023 年，I Made Aswin Nahrendra 等人[9]提出了一个鲁棒学习框架，在多样化地形中取得了有希望的结果，但其仍缺乏综合的运动和恢复策略。

因此，开发一个统一的学习框架，使四足机器人能够在复杂环境中实现行走与摔倒恢复，对于推动机器人自主性尤其是在高风险应用(如搜索与救援、军事侦察和野外探测)中的任务执行至关重要。为此，本研究提出了基于深度强化学习的四足机器人复杂地形行走与摔倒恢复统一控制方法，该方法在[4] [9]基础上整合了摔倒恢复因子、动态增长策略(KL Warm-up) [10]以及 Normalized Penalized Proximal Policy Optimization (NP3O) [11]。通过这一统一方法，机器人不仅能够在多样复杂地形中保持稳定行走，还能在遭遇摔倒时实现快速自我恢复，从而显著增强了在实际应用中的可靠性和任务完成能力。

2. 方法

2.1. 训练环境

训练环境涵盖完整的操作流程，包括动作执行、观测获取、奖励与代价评估以及状态重置。同时，环境集成了训练课程，引导智能体循序渐进地提升运动能力，并在此基础上增添了四足机器人摔倒恢复训练设施，从而提高其在复杂环境中的适应性。

2.1.1. 动作空间

动作空间为一个 12×1 维向量 a_t，表示机器人期望的关节角度。为了提升学习效率，训练策略可以推理相对于机器人静止站立姿态的关节角偏移量 θ_{stand}，从而定义期望关节角：

$$\theta_{\text{des}} = \theta_{\text{stand}} + a_t \tag{1}$$

每个关节的期望角度由比例-微分(PD)控制器精确跟踪。

2.1.2. 状态空间

从训练环境中获取训练模型输入的观测信息，包括 o_t、o_t^H 和 s_t，并对这些信息进行了归一化处理，以便更好地训练模型。

$$o_t = \begin{bmatrix} \omega_t & g_t & c_t & \theta_t & \dot{\theta}_t & a_{t-1} \end{bmatrix} \tag{2}$$

o_t 是本体观测信息，ω_t、g_t、c_t、θ_t、$\dot{\theta}_t$ 和 a_{t-1} 分别表示身体角速度、机器人框架中的重力向量、期望速度命令、关节角度、关节角速度和上一个动作，这些信息均为本体观测信息。o_t^H 是过去 H 个时刻的本体观测信息，表示为

$$o_t^H = \begin{bmatrix} o_t & o_{t-1} & \cdots & o_{t-H} \end{bmatrix}^{\mathrm{T}} \tag{3}$$

s_t 是特权信息，它除了包含本体观测信息 o_t 之外，还包含身体线速度 v_t、周围地面高度 h_t、脚部接触状态 c_{foot}、控制参数(Kp，Kd，电机参数)、摩擦系数、恢复系数、质量质心和系统延时。

2.1.3. 奖励与代价函数

本研究的奖励函数主要参考了其他研究[1] [3] [9] [12]，本研究提出的奖励函数设计涵盖了机器人摔倒恢复、行走控制以及安全约束三个核心方面。所有奖励项均通过加权求和的方式共同作用，使机器人能够在复杂环境中实现高效、稳定且安全地运动策略在每个状态下执行动作所获得的总奖励为：

$$r_t(s_t, a_t) = \sum r_i w_i \tag{4}$$

为确保策略在摔倒恢复和行走控制任务中的合理学习，设计了自恢复因子，根据机器人基座的朝向动态调整不同奖励项的影响程度。

具体而言，不同奖励项的设计体现了策略在摔倒恢复与行走控制中的不同侧重点，如表 1 所示。部分奖励项乘以自恢复因子 1：$\text{clamp}(-g_z, 0, 1)$，用于确保机器人在正常行走时触发这些奖励，而在摔倒时

不会影响策略学习；部分奖励项乘以自恢复因子 2：$1-g_z$，用于站立与摔倒之间的平滑过渡，保证机器人在行走时受到更强的激励，而在摔倒时减少影响，从而引导策略自然调整姿态，增强恢复能力；某些奖励项未乘自恢复因子，是因为它们在所有状态下都应发挥作用，例如能耗优化和动作平滑性等关键目标，不论机器人是否摔倒，都需要持续优化。

Table 1. Reward functions and their weights
表 1. 奖励函数及其权重

奖励	公式(ri)	权重(wi)
Lin velocity tracking	$\exp\left(-\frac{\left\|v_{\text{cmd\_}xy}-v_{xy}\right\|^2}{\sigma}\right)$	2.0
Ang velocity tracking	$\exp\left(-\frac{\left(\omega_{\text{cmd\_yaw}}-\omega_{\text{yaw}}\right)^2}{\sigma}\right)$	1.0
Linear velocity(z)	$\text{clamp}\left(-g_z,0,1\right)\cdot v_z^2$	−4.0
Angular velocity(xy)	$\text{clamp}\left(-g_z,0,1\right)\cdot\left\|\omega_{xy}\right\|^2$	−0.1
Orientation	$\text{clamp}\left(-g_z,0,1\right)\cdot\left\|g_{xy}\right\|^2$	−0.2
collision	$\text{clamp}\left(-g_z,0,1\right)\cdot\sum_{i\in P}1\left(\left\|\boldsymbol{F}_i\right\|>0.1\right)$	−1.0
Body height	$\text{clamp}\left(-g_z,0,1\right)\cdot\left(h_{\text{target}}-h\right)^2$	−10.0
Foot clearance	$\text{clamp}\left(-g_z,0,1\right)\cdot\sum_{i\in\text{feet}}\left(p_{i,z}^{\text{body}}-p_z^{\text{target}}\right)\cdot\left\|v_{i,xy}^{\text{body}}\right\|$	−0.5
Foot mirror	$\text{clamp}\left(-g_z,0,1\right)\cdot\left(\left\|\theta^{\text{LF}}-\theta^{\text{RR}}\right\|^2+\left\|\theta^{\text{RF}}-\theta^{\text{LR}}\right\|^2\right)$	−0.05
Foot slide	$\text{clamp}\left(-g_z,0,1\right)\cdot\sum_{i\in\text{feet}}c_i\cdot\left\|v_{i,xy}^{\text{body}}\right\|$	−0.05
Stumble	$\text{clamp}\left(-g_z,0,1\right)\cdot1\left(\left\|F_{\text{foot},xy}\right\|>5\times\left\|F_{\text{foot},z}\right\|\right)$	−0.05
Base uprightness	$1-g_z$	0.6
Foot contact	$\left(\left\|v_{xy}^{\text{cmd}}\right\|<0.1\right)\cdot c_{\text{foot}}$	0.6
Stand nice	$\left(\left\|v_{xy}^{\text{cmd}}\right\|<0.1\right)\cdot\left(1-g_z\right)\cdot\left\|\theta-\theta^{\text{default}}\right\|$	−0.1
Action rate	$\left\|a_t-a_{t-1}\right\|^2$	−0.01
Smoothness	$\left\|a_t-2a_{t-1}+a_{t-2}\right\|^2$	−0.01
Joint accelerations	$\left\|\ddot{\theta}\right\|^2$	−2.5e−7
Joint power	$\left\|\tau\right\|\left\|\dot{\theta}\right\|$	−2e−5
Feet contact forces	$\sum_{i\in\text{feet}}\max\left(\left\|\boldsymbol{F}_i\right\|-F_{i\_\max},0\right)$	−0.00015 0.2

为了增强四足机器人在实际部署中的运动安全性，本研究引入了三种代价评估机制，分别对关节位置、速度和扭矩施加额外约束。代价函数对超出物理极限的行为进行惩罚，从而引导策略在训练过程中避免越界动作。为了进一步加强约束效果，代价函数中加入了物理极限缩放因子，记作 κ_{pos}、κ_{vel} 和 κ_{torque}。这些因子通常在 0.6 到 1.0 之间变化，适度地收紧关节位置、速度和扭矩的阈值。这样有效地增加了安全余量，并鼓励策略在更为保守和安全的动作空间内操作。代价函数设计如表 2 所示。

Table 2. Cost functions and their equations
表 2. 代价函数及其方程

代价函数	方程(c_i)
joint position limits	$\sum\left[\max\left(0,\lvert\theta\rvert-\theta_{\max}\cdot\kappa_{\text{pos}}\right)+\max\left(0,\lvert\theta\rvert\theta_{\min}\cdot\kappa_{\text{pos}}\right)\right]$
joint velocity limits	$\sum\max\left(0,\lvert\dot{\theta}\rvert-\dot{\theta}_{\text{limit}}\cdot\kappa_{\text{vel}}\right)$
torque limits	$\sum\max\left(0,\lvert\tau\rvert-\tau_{\text{limit}}\cdot\kappa_{\text{torque}}\right)$

2.1.4. 训练课程

为训练四足机器人从任意姿态恢复至可行走状态，每个训练 episode 开始时，将四足机器人从各种姿态随机投放到复杂地形上[9]。为模拟其从高空坠落的情况，可适当扩大初始根部高度的范围，以提升其恢复至可行走状态的能力。为确保机器人能够逐步适应不同地形，本研究采用了游戏式地形课程[12]。

2.2. 算法框架与优化方法设计

2.2.1. 不对称 Actor-Critic 架构与 Context-aided Estimator 设计

如图 1 所示，本研究采用了不对称的 Actor-Critic 架构[13]，策略的优化采用了 NP3O 算法，为满足 NP3O 算法的需求，额外引入了用于计算代价的 Cost Critic 网络。在非对称的框架基础上融合了一个 Context-aided Estimator 网络[5]，Context-aided Estimator 以过去 H 个时刻的本体观测信息 o_t^H 作为输入，Context-aided Estimator 网络内部编码出 $\tilde{v}_t$ 和 z_t。Actor 网络以 $\tilde{v}_t$、z_t 和 o_t 作为输入，输出动作 a_t。Value Critic 网络和 Cost Critic 网络以特权信息 s_t 作为输入，允许在训练阶段访问特权信息，从而可以提供更准确的状态值和代价估计。本研究的训练模型所涉及的所有网络均基于多层感知机(MLP)构建，整体架构简洁高效，有助于降低训练开销并提升策略优化的稳定性。

在 Context-aided Estimator 网络内部，Encoder 网络将历史观测 o_t^H 编码为 $\tilde{v}_t$ 和 z_t。第一个解码头用于估计 $\tilde{v}_t$，而第二个解码头用于重建 z_t。采用了 β 变分自编码器[14]-[16]作为自编码器的基本架构。Context-aided Estimator 网络通过一个混合损失函数进行优化，定义如下：

$$\mathcal{L}_{\text{CE}}=\mathcal{L}_{\text{vel}}+\mathcal{L}_{\text{VAE}} \tag{5}$$

其中，$\mathcal{L}_{\text{vel}}$ 和 $\mathcal{L}_{\text{VAE}}$ 分别表示机体速度估计损失和 VAE 损失。为了实现显式状态估计，使用均方误差(MSE)损失函数，度量估计的机体速度 v_t 与来自仿真器的真实机体速度 $\tilde{v}_t$ 之间的差异，具体如下：

$$\mathcal{L}_{\text{vel}}=\text{MSE}\left(\tilde{v}_t,v_t\right) \tag{6}$$

VAE 网络使用标准的 β-VAE 损失进行训练，该损失由重建损失和潜变量损失组成。在重建损失中使用均方误差(MSE)，在潜变量损失中使用 Kullback-Leibler(KL)散度[17]。VAE 损失公式如下：

$$\mathcal{L}_{\text{VAE}}=\text{MSE}\left(\tilde{o}_{t+1},o_{t+1}\right)+\beta D_{\text{KL}}\left(q\left(z_t\mid o_t^H\right)\Vert\, p\left(z_t\right)\right) \tag{7}$$

其中，$\tilde{o}_{t+1}$是重建的下一时刻观测，$q\left(z_t \mid o_t^H\right)$是在给定历史观测$o_t^H$条件下的潜变量$z_t$的后验分布，$p\left(z_t\right)$是上下文的先验分布，由高斯分布参数化。选择标准正态分布作为先验分布，因为所有观测值在输入前均已归一化为零均值和单位方差。

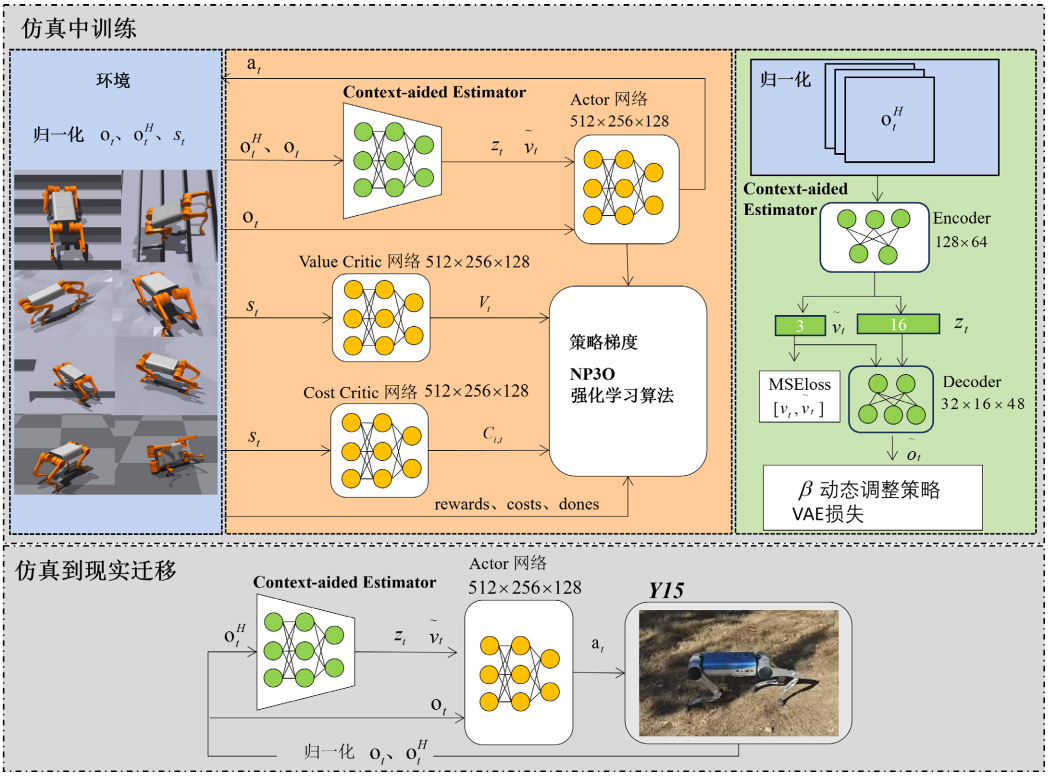

Figure 1. The proposed training and transfer framework for quadruped locomotion and fall recovery over complex terrains
图 1. 本研究提出的四足机器人复杂地形行走与摔倒恢复训练与迁移框架

2.2.2. 动态调度的 β-VAE 潜变量约束策略

在变分自编码器(VAE)以及 β-VAE 等结构中，KL 散度项用于对潜在变量的分布进行约束，从而提升潜在空间的可解释性与生成能力。然而，固定的 β 系数往往会带来两个极端问题：若 β 过小，模型可能无法学习到具有结构性的潜在表示；若 β 较大，训练初期容易抑制重构能力，甚至导致“后验塌缩”现象，即编码器放弃从输入中提取信息，仅输出近似先验的无用分布。

为了解决上述问题，本研究在 β-VAE 结构中引入了一种简单而有效的 β 动态增长策略[10]。该策略采用两阶段调整机制：在训练的前 N 步内固定使用一个极小值 β_0，以保证模型优先优化重构目标；随后 β 以指数函数的形式逐步增长，最终趋近于上限 $\beta_{\max}$，从而逐步加强对潜在空间分布的正则约束。该策略可形式化为：

$$\beta(i)=\begin{cases}\beta_0, & i<N\\ \min\left(\beta_{\max}, \beta_0\cdot r^{i-N}\right), & i\geq N\end{cases} \tag{8}$$

2.2.3. 基于归一化惩罚机制的 NP3O 强化学习算法

如果仅依赖 Proximal Policy Optimization (PPO) [18]及其奖励函数进行限制，难以有效约束力矩，可能导致实际部署时的安全风险。利用 NP3O 算法，训练出策略能够达到与传统 PPO 算法相当的性能指标，

并且违反约束的情况更少[19]。NP3O 算法将 Actor 网络、Reward Critic 网络和 Cost Critic 网络整合到统一的优化框架中，结合了策略优化、价值评估和代价评估的目标，确保在满足安全约束的条件下最大化累积奖励。NP3O 的最终优化目标可以表示为以下公式：

$$L_{\mathrm{NP3O}}(\theta)=L_R^{\mathrm{CLIP,N}}(\theta)+\kappa\sum_i \max\left\{0,L_{C_i}^{\mathrm{CLIP,N}}(\theta)\right\} \tag{9}$$

其中：

$$L_R^{\mathrm{CLIP,N}}(\theta)=\mathop{\mathbb{E}}_{\substack{s\sim d^{\pi_k}\\ a\sim\pi_k}}\left[-\min\left\{r(\theta)\tilde{A}_R^{\pi_k}(s,a),\mathrm{clip}\left(r(\theta),1-\epsilon,1+\epsilon\right)\tilde{A}_R^{\pi_k}(s,a)\right\}\right] \tag{10}$$

$$L_{C_i}^{\mathrm{CLIP,N}}(\theta)=\mathop{\mathbb{E}}_{\substack{s\sim d^{\pi_k}\\ a\sim\pi_k}}\left[-\min\left\{r(\theta)\tilde{A}_{C_i}^{\pi_k}(s,a),\mathrm{clip}\left(r(\theta),1-\epsilon,1+\epsilon\right)\tilde{A}_{C_i}^{\pi_k}(s,a)\right\}+\frac{(1-\gamma)\left(J_{C_i(\pi_k)}-d_i\right)+\mu_{C_i}}{\sigma_{C_i}}\right] \tag{11}$$

优势归一化是一种广泛使用的启发式算法，用于提高策略梯度算法的稳定性[14]。$L_R^{\mathrm{CLIP,N}}$ 和 $L_{C_i}^{\mathrm{CLIP,N}}$ 上标 N 表示使用归一化优势估计，$\tilde{A}_R^{\pi_k}$ 是 Critic 网络提供的经过归一化的优势函数，$\tilde{A}_{C_i}^{\pi_k}$ 是 Cost 网络提供的经过归一化的代价优势函数。κ 是惩罚因子，用于控制约束违反的惩罚强度，定义式如下

$$\kappa=\min\left(1,\kappa_{current}\cdot 1.0004^{i}\right) \tag{12}$$

i 表示迭代次数，通过渐进式调整约束强度，既保证了训练初期的探索能力，又确保了训练后期的稳定性。

$r(\theta)=\dfrac{\pi_\theta(a\mid s)}{\pi_{\theta_{\mathrm{old}}}(a\mid s)}$ 是新旧策略的概率比，$J_{C_i(\pi_k)}$ 是当前策略的期望代价回报，d_i 是对应的代价约束的上限，μ_{C_i} 和 σ_{C_i} 分别为代价优势函数的均值和标准差，ϵ 是一个超参数，用于控制策略更新的幅度。

2.3. 面向 Sim-to-Real 的性能提升方案

为了使训练策略适用于仿真到现实的迁移，本研究随机化了地面摩擦和恢复系数、机器人的质量质心、电机强度、关节 PD 增益和系统延迟，在观察值中添加了噪声，并在训练期间随机推动机器人，以教它们更稳定的站立姿态。每个参数的随机化范围如表 3 所示。

Table 3. Domain randomizations and their ranges
表 3. 域随机化和它们各自的范围

参数	范围	单位
Ground Friction	[0.5, 2.0]	-
Ground Restitution	[0.0, 1.0]	-
Body Mass	[-2.5, 2.5] × nominal value	Kg
CoM	[−0.1, 0.1] × [−0.1, 0.1] × [−0.1, 0.1]	m
Motor Strength	[0.9, 1.1] × motor torque	Nm
Joint Kp	[0.9, 1.1] × 30	-
Joint Kd	[0.9, 1.1] × 0.5	-
System Delay	[0, 3Δt]	s
Externa velocity(xy)	[−5, 5] × [−5, 5]	m/s

3. 实验

3.1. 仿真设置

本研究采用 Isaac Gym 仿真器[20]，基于[12]提出的公开实现版本。在训练过程中，我们联合训练了 Context-aided Estimator、Actor 网络、Value Critic 网络和 Cost Critic 网络，进行了 10,000 次迭代。总共并行训练了 4096 个智能体，并引入了域随机化技术。训练过程中使用了 NP3O 算法，剪切参数、广义优势估计(GAE)系数和折扣因子分别设置为 0.2、0.95 和 0.99，网络优化采用 Adam 优化器[21]，学习率设置为 10-3。所有实验均在配置为 Intel Core i7-13650 CPU、8GB RAM 和 NVIDIA RTX 4060 GPU 的笔记本电脑上进行。

3.2. 实体部署

本研究基于整个训练框架导出了用于部署的训练模型，如图 1 中仿真到实体迁移。真实世界实验使用由山东优宝特智能机器人有限公司推出的电动四足机器人 Y15 作为实验平台。该机器人高 37 cm，重量 13.8 kg，配备 12 个驱动执行器，最大输出扭矩可达 48Nm。机器人传感器包括关节位置编码器和惯性测量单元(IMU)。期望关节角由 PD 控制器进行跟踪，比例增益(Kp = 30)，微分增益(Kd = 0.5)。

3.3. 消融实验

为了评估框架中各模块的贡献，进行了几组消融实验。具体而言，本研究的训练框架(Ours)与以下三种变体进行了对比：

(1) w/o NP3O：在该变体中，移除了 NP3O 动态调整机制，采用标准的 PPO 算法进行训练；

(2) w/o KL Warm-up：在该变体中，β 动态调度策略移除，使用固定的 β 值；

(3) w/o self-recovery：在该变体中，奖励函数中不包含摔倒恢复因子，策略仅依据基础行走奖励和恢复身体直立的奖励进行优化。

如图 2 所示，不同方法在训练过程中的平均奖励曲线表明，本研究的训练方法整体表现最佳。其平均奖励随着迭代次数的增加稳步上升，并最终收敛到较高水平。这表明本研究的训练框架的优越性能依赖于 NP3O 模块、动态调度策略和自恢复因子的协同作用，移除其中任一组件都会导致性能下降。

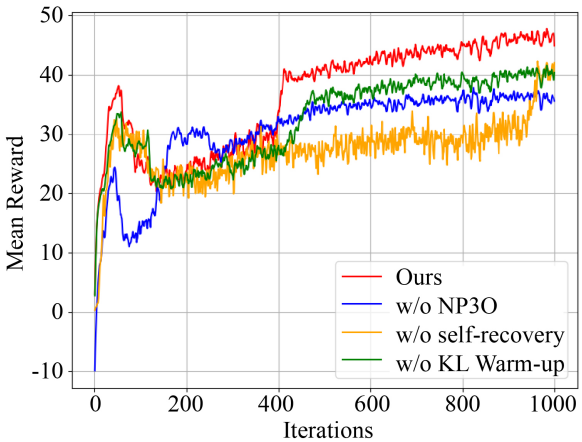

Figure 2. Average reward curves during training for different methods

图 2. 不同方法在训练过程中的平均奖励曲线

如图 3 所示，图中展示了本研究训练方法与去除 KL Warm-up 变体(w/o KL Warm-up)训练过程中重建损失与体速度估计损失的曲线。从图中可以看出，本研究训练方法的重建损失与体速度估计损失均较低，而去除 KL Warm-up 变体的损失曲线较高。这表明，在 VAE 框架下，采用 KL Warm-up 策略能够有效降低损失，从而提升模型的性能。

如图 4 所示，图中对比了本研究训练方法与去除 NP3O 变体(w/o NP3O)在平地环境下执行摔倒恢复动作时 12 个关节状态的输出结果。对比分析表明，w/o NP3O 变体在摔倒恢复过程中存在较高风险，关节角变化幅度可达到 3.5 rad，关节速度峰值达到 26.31 rad/s，关节力矩峰值高达 53.63 Nm。相比之下，本研究的训练方法训练出的将关节角波动控制在 2.5 rad 以内，关节速度变化控制在 10 rad/s 以内，力矩波动限制在 30 Nm 以内，展现出显著更平稳且更易控制的动态响应特性。需要指出的是，Y15 机器人的电机力矩物理限制为 48 Nm，而 w/o NP3O 变体已出现力矩超限现象，存在损伤电机及无法在真实系统中部署的风险。而本研究的训练框架训练出的策略更适合长期部署和稳定运行于复杂环境中。

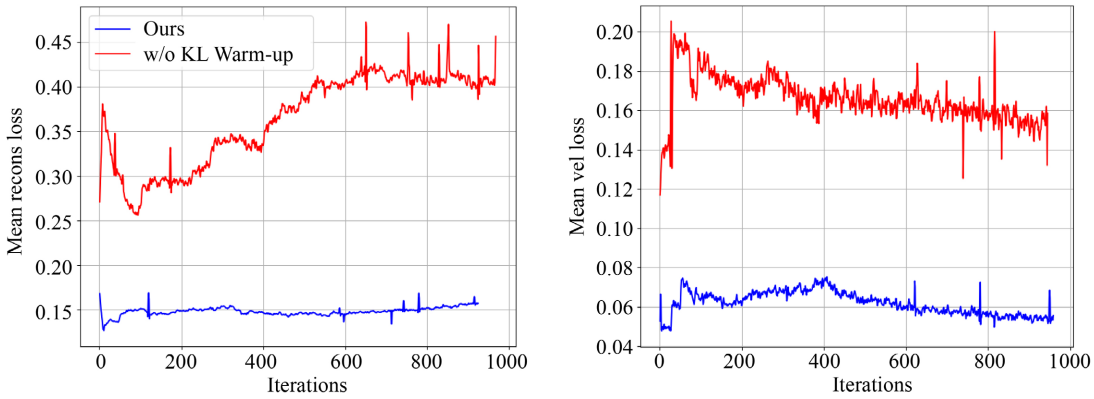

Figure 3. Comparison between the proposed training method and variants without KL Warm-up, reconstruction loss, and body velocity estimation loss
图 3. 本研究训练方法与去除 KL Warm-up 变体重建损失与体速度估计损失对比

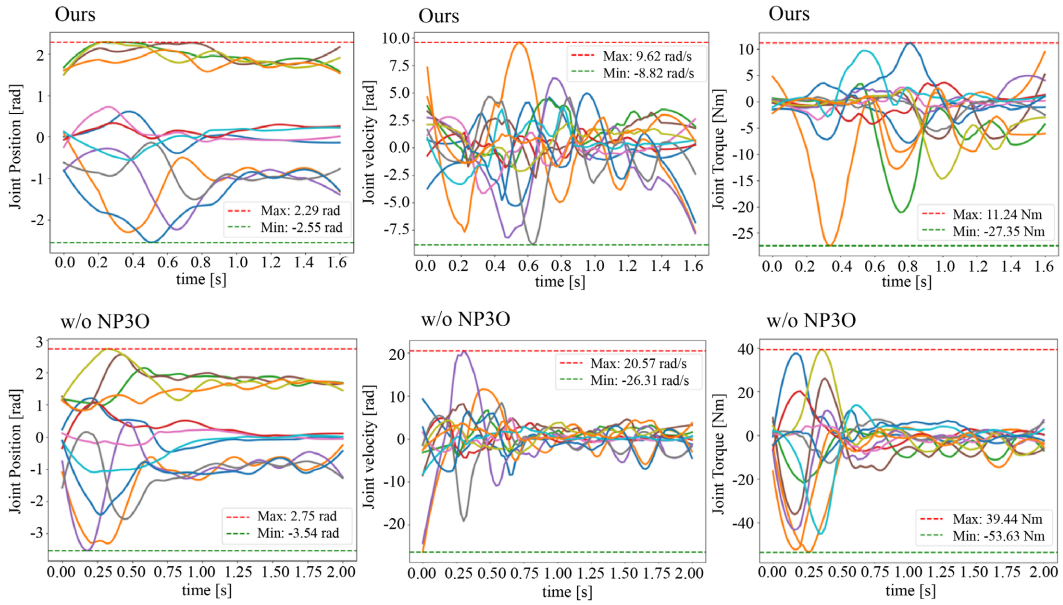

Figure 4. Comparison of joint states between the proposed training method and the variant without NP3O
图 4. 本研究训练方法与去除 NP3O 变体的关节状态对比

3.4. 真实环境下摔倒恢复与运动能力的验证

在本研究中，在楼梯、平地和草地等典型地形上，对 Y15 四足机器人从摔倒状态到恢复站立的全过程进行了实验验证。图 5 给出了机器人在不同地形上执行摔倒恢复任务的动作分解示意图，清晰呈现了其在各阶段的关键姿态与支撑策略。图 6 则展示了恢复过程中各关节的速度与力矩变化曲线，进一步证明了 Y15 机器人在复杂地形下能够高效、平稳地完成摔倒恢复。

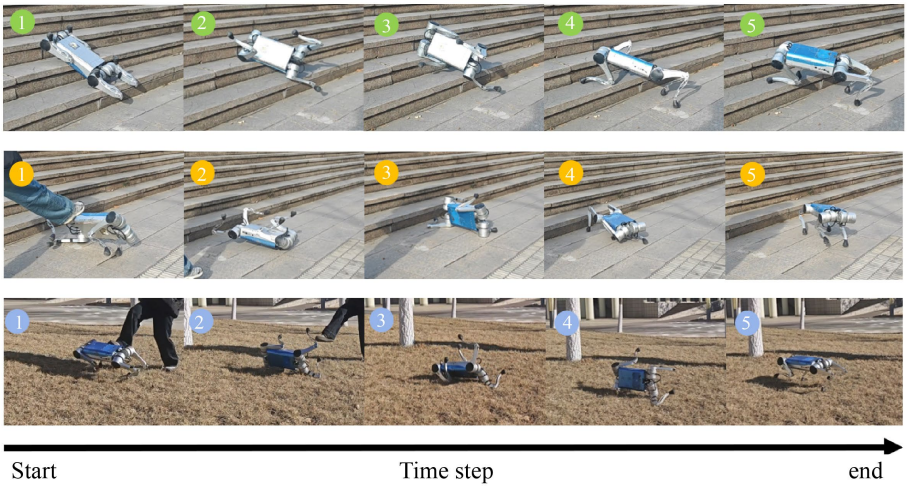

Figure 5. Motion decomposition diagrams of the Y15 robot performing fall recovery tasks on different terrains
图 5. Y15 机器人在不同地形上执行摔倒恢复任务时的动作分解示意图

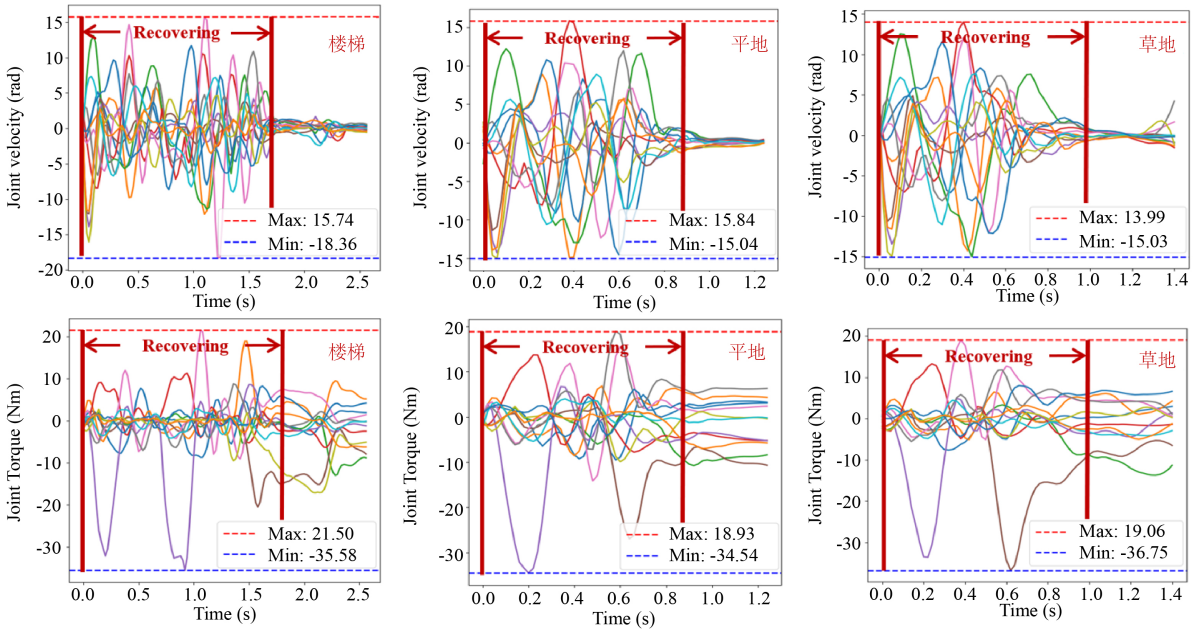

Figure 6. The variations in joint velocities and torques of the motors during the fall recovery tasks performed by the Y15 robot on different terrains
图 6. Y15 机器人在不同地形上执行摔倒恢复任务时的电机的关节速度与力矩变化

从图 5 可以看出，Y15 机器人在摔倒后能够根据地形的几何特征和接触条件灵活调整其恢复策略，

例如重新排序支撑腿的着地顺序、调整身体姿态，或优化关节动作轨迹，从而显著提升恢复过程的稳定性与效率。这种对环境的自适应能力，使其具备在非结构化复杂地形中实现自主摔倒恢复的潜力。

图 6 所示的实验结果进一步表明，Y15 机器人在不同地形上始终能够保持动作平稳、响应迅速，顺利实现从摔倒到站立再到行走的无缝衔接。在所有实验中，恢复时间均控制在 3 秒以内，充分展现了机器人控制系统的高响应性与鲁棒性。具体来看，Y15 机器人的最大关节速度为 40 rad/s，最大关节力矩为 48 Nm，整个恢复过程中各关节的输出始终处于安全范围内，未出现过载或剧烈振荡，表明其控制系统在功率分配与稳定性管理方面表现可靠。从力矩与速度的变化曲线可以观察到，在恢复初期，机器人会产生较高的关节输出以克服重力和不利姿态的影响，随后快速趋于平稳，顺利完成站立与步态初始化。该过程充分体现了所采用控制策略在动力调度、动作协调和稳定性控制等方面的有效性与先进性。

为了进一步评估所提出策略的运动能力和在复杂环境中的适应性，进行了真实环境下的实地实验。如图 7 所示，图中展示了机器人在两种具有代表性的复杂地形上的测试情况：一是攀爬 16 cm 高的楼梯，二是从 80 cm 高的平台顺利下落并恢复平衡。实验结果表明，机器人在无外部辅助的情况下，能够稳定完成楼梯攀爬任务，展现出良好的步态协调性和动态稳定性；在面对大高度差的平台下降过程中，机器人能够有效吸收冲击并快速完成身体姿态调整，避免了失衡摔倒现象。

Figure 7. The Y15 robot climbing 16 cm high stairs and descending from an 80 cm platform
图 7. Y15 爬 16 cm 高的楼梯以及下 80 cm 高台

4. 结论

本研究提出了一种新的四足机器人摔倒恢复与行走策略优化方法。通过高效的训练策略，机器人能够在多种地形上自主恢复摔倒并平稳过渡至行走状态，体现了其卓越的自适应能力。为解决固定系数可能带来的极端问题，本文引入了动态调度策略，并通过两阶段调整机制逐步增大系数，从而有效平衡了潜在空间的重构能力与正则化约束，提升了策略的稳定性与泛化能力。此外，本研究通过摔倒恢复因子和统一奖励设计的引入，成功实现了运动与恢复过程的有机整合，使机器人能够从摔倒状态自然过渡到行走状态，确保了恢复过程的平滑和稳定。在训练过程中，结合 NP3O 算法进行安全约束训练，通过动态收紧约束机制增强了策略的遵从性并提高了训练的稳定性。这种方法不仅有效防止了约束违背，还大大提升了机器人在实际部署中的安全性。总体而言，本研究显著提高了四足机器人的运动恢复能力，展现了良好的应用前景和实际价值。

参考文献

[1] Lee, J., Hwangbo, J., Wellhausen, L., Koltun, V. and Hutter, M. (2020) Learning Quadrupedal Locomotion over Challenging Terrain. *Science Robotics*, **5**, eabc5986. https://doi.org/10.1126/scirobotics.abc5986

[2] Miki, T., Lee, J., Hwangbo, J., Wellhausen, L., Koltun, V. and Hutter, M. (2022) Learning Robust Perceptive Locomotion

for Quadrupedal Robots in the Wild. *Science Robotics*, **7**, eabk2822. https://doi.org/10.1126/scirobotics.abk2822

[3] Kumar, A., Fu, Z., Pathak, D. and Malik, J. (2021) RMA: Rapid Motor Adaptation for Legged Robots. *Robotics*: *Science and Systems XVII*, 12-16 July 2021, 1-12. https://doi.org/10.15607/rss.2021.xvii.011

[4] Aswin Nahrendra, I.M., Yu, B. and Myung, H. (2023) DreamWaQ: Learning Robust Quadrupedal Locomotion with Implicit Terrain Imagination via Deep Reinforcement Learning. 2023 *IEEE International Conference on Robotics and Automation* (*ICRA*), London, 29 May-2 June 2023, 5078-5084. https://doi.org/10.1109/icra48891.2023.10161144

[5] Long, J., Wang, Z., Li, Q., *et al.* (2023) Hybrid Internal Model: Learning Agile Legged Locomotion with Simulated Robot Response. arXiv: 2312.11460.

[6] Long, J., Yu, W., Li, Q., *et al.* (2024) Learning H-Infinity Locomotion Control. arXiv: 2404.14405.

[7] Lee, J., Hwangbo, J. and Hutter, M. (2019) Robust Recovery Controller for a Quadrupedal Robot Using Deep Reinforcement Learning. arXiv:1901.07517.

[8] Smith, L., Kew, J.C., Bin Peng, X., Ha, S., Tan, J. and Levine, S. (2022) Legged Robots That Keep on Learning: Fine-Tuning Locomotion Policies in the Real World. 2022 *International Conference on Robotics and Automation* (*ICRA*), Philadelphia, 23-27 May 2022, 1593-1599. https://doi.org/10.1109/icra46639.2022.9812166

[9] Nahrendra, I.M.A., Oh, M., Yu, B., *et al.* (2023) Robust Recovery Motion Control for Quadrupedal Robots via Learned Terrain Imagination. arXiv: 2306.12712.

[10] Bowman, S.R., Vilnis, L., Vinyals, O., Dai, A., Jozefowicz, R. and Bengio, S. (2016) Generating Sentences from a Continuous Space. *Proceedings of the* 20*th SIGNLL Conference on Computational Natural Language Learning*, Berlin, August 2016, 10-21. https://doi.org/10.18653/v1/k16-1002

[11] Shen, L., Yang, L., Chen, S., Yuan, B., Wang, X., Tao, D., *et al.* (2022) Penalized Proximal Policy Optimization for Safe Reinforcement Learning. arXiv: 2205.11814.

[12] Rudin, N., Hoeller, D., Reist, P. and Hutter, M. (2022) Learning to Walk in Minutes Using Massively Parallel Deep Reinforcement Learning. arXiv: 2109.11978.

[13] Pinto, L., Andrychowicz, M., Welinder, P., Zaremba, W. and Abbeel, P. (2018) Asymmetric Actor Critic for Image-Based Robot Learning. *Robotics*: *Science and Systems XIV*, Pittsburgh, 26-30 June 2018, 1-10. https://doi.org/10.15607/rss.2018.xiv.008

[14] Kingma, D.P. and Welling, M. (2013) Auto-Encoding Variational Bayes. arXiv: 1312.6114.

[15] Higgins, I., Matthey, L., Pal, A., *et al.* (2017) *β*-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. *Proceeding of International Conference on Learning Representations* (*ICLR*) 2017, Toulon, 24-26 April 2017, 1-13.

[16] Burgess, C.P., Higgins, I., Pal, A., *et al.* (2017) Understanding Disentangling in *β*-VAE. arXiv: 1804.03599.

[17] Kullback, S. and Leibler, R.A. (1951) On Information and Sufficiency. *The Annals of Mathematical Statistics*, **22**, 79-86. https://doi.org/10.1214/aoms/1177729694

[18] Schulman, J., Wolski, F., Dhariwal, P., *et al.* (2017) Proximal Policy Optimization Algorithms. arXiv: 1707.06347.

[19] Lee, J., Schro, K.V., *et al.* (2023) Evaluation of Constrained Reinforcement Learning Algorithms for Legged Locomotion. arXiv: 2309.15430.

[20] Makoviychuk, V., Wawrzyniak, L., Guo, Y., *et al.* (2021) Isaac Gym: High Performance GPU-Based Physics Simulation for Robot Learning. arXiv: 2108.10470.

[21] Kingma, D.P. and Ba, J. (2015) Adam: A Method for Stochastic Optimization. arXiv: 1412.6980.