Learning Quadruped Locomotion from Casual Videos

Anonymous Author(s)

***Abstract*— Learning-based locomotion policies for quadruped robots have shown great promise in achieving increasingly complex tasks; however, most existing approaches rely on extensive manual reward tuning or expert data derived from expensive motion capture systems or human-designed trajectories. Motivated by recent progress in learning from video for robotic tasks, we propose a framework that leverages casually recorded video data for quadruped locomotion learning. In particular, our method uses task-specific videos from a low-cost camera and sparse keypoint tracking to construct a reward component for subsequent learning of deployable and energy-efficient policies. We extensively evaluated our method in both simulation and real-world experiments, demonstrating that the proposed approach achieves performance competitive with human-engineered baselines and motion-capture-based training, and provides, to our knowledge, the first demonstration of non-trivial quadruped behaviors (e.g., box climbing, stair climbing, and standing up) learned from video data. Overall, our work represents a step toward leveraging accessible video sources to advance locomotion learning and paves the way for scaling such methods to more challenging scenarios. Website: https://sites.google.com/view/locomotion-from-video.**

I. INTRODUCTION

Learning complex locomotion skills for quadruped robots remains challenging, often requiring carefully designed reward functions [1], meticulously engineered environments [2], or manually designed trajectories [3]. While learning from visual data has shown compelling results in manipulation [4, 5] and humanoid control [6], it remains less explored for quadrupeds due to the lack of accurate 3D reconstruction models for animal morphologies.

In this work, we propose a framework for learning quadruped locomotion from recorded RGB(D) videos (Figure 1). Our key insight is to leverage tracking of sparse keypoints in the demonstration to construct an expert dataset for policy learning. Our approach leverages casually recorded animal videos from low-cost cameras to learn quadruped locomotion, achieving task-tracking and energy-efficiency competitive with manually engineered baselines and motion capture datasets. We demonstrate hardware deployment across three challenging scenarios (box climbing, stair climbing, stand-up) requiring only two reward terms rather than 10+ in standard approaches.

II. METHOD

We propose a framework that leverages casual RGB(D) videos to learn deployable quadruped locomotion policies (Fig. 2). A key insight is that tracking a sparse set of keypoints in a demonstration is sufficient to construct a dataset capturing locomotion style. The resulting policies are task-conditioned and output joint position targets.

Fig. 1: We learn locomotion policies using recorded video clips of a dog as references for different scenarios. The videos provide a reference for the locomotion style, and the policies are trained in simulation with a task and style objective. We show deployment on the hardware while maintaining the style from the demonstration.

Our pipeline takes RGB images and corresponding depth maps as input. We track six sparse keypoints (front, back, and four feet) using a TAP model [8]. The 2D keypoints are lifted to 3D using depth maps, camera intrinsics, and 6D camera poses estimated via SLAM [9] adapted for dynamic scenes [10]. For occluded keypoints, we apply linear interpolation. Post-processing includes depth filtering, smoothing, and foot clearance amplification to facilitate policy transfer.

We adapt kinematic retargeting from [11] to map reference trajectories to the robot. The base position and orientation (yaw, pitch) are reconstructed from the front and back keypoints, while joint positions are determined via inverse kinematics from the foot keypoints (Figure 3). Velocities are computed using finite differences.

We use AMP [7] for policy learning, combining a task reward r_t^{task} with a style reward r_t^{style} obtained from the video demonstration, both weighted with a factor w:

$$r = w_{task} r_t^{task} + w_{style} r_t^{style}. \quad (1)$$

While the task component is specified by the user, the style component encourages the agent to exhibit similar state transitions (s, s') observed in the video demonstration.

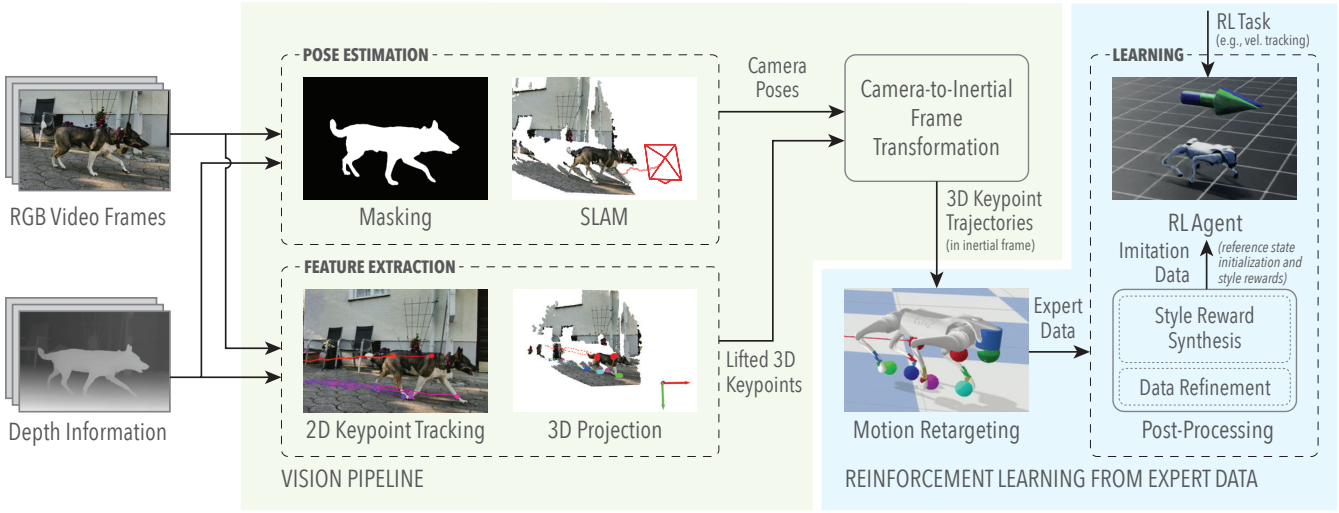

Fig. 2: Visualization of the pipeline of our approach, which takes RGB video frames and depth information as input. We continuously track sparse keypoints in the RGB frames and lift them to 3D space using the depth information. Simultaneously, we reconstruct the camera poses in the video by leveraging masks for moving objects and a SLAM algorithm. The 3D keypoints, along with the reconstructed camera poses, are used to generate a 3D keypoint trajectory in the inertial frame from the demonstration. The keypoints are then retargeted to the robot platform, yielding the expert dataset that defines the style reward for policy learning with AMP [7].

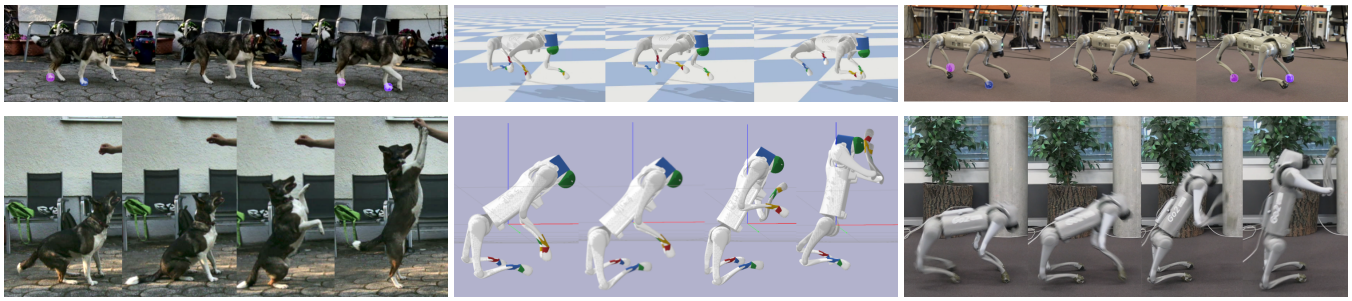

Fig. 3: Successful imitation of video clips for on robot hardware with out proposed method. The dots in the top row highlight the foot contact schedule for flat walking (FR, RL, and FL, RR) that is maintained from the demonstration to hardware deployment. **Left**: recorded video clip. **Middle**: result of kinematic motion retargeting. **Right**: real-world deployment of the learned policies on a Unitree Go2.

III. EXPERIMENTAL RESULTS

Baselines and ablations. We compare learning from video (*Video*) against learning from a common motion capture dataset recorded on flat terrain [12] (*MoCap*), a motion trajectory for flat walking that we manually designed (*Manual Trajectory*), and a state-of-the-art reconstruction model that densely reconstructs the animal from RGB videos [13] (*Animal Avatar*). Additionally, we include baselines with a simple task reward only formulation (*Simple Reward*), and a complex reward that additionally includes 13 regularization terms (*Complex Reward*). We tune the regularizing reward term weights for each locomotion scenario.

For our method, we ablate the depth source, i.e., using the depth from the camera (*Video w. Depth Camera*) and from a state-of-the-art estimation model (*Video w. DepthAnythingV2* [14], as well as two sizes of the visual dataset for walking on flat ground (*Video* vs. *Video (extended)*).

Video data. We record all motion clips outdoors on a hand-held Orbbec Femto Bolt RGBD camera. The camera records with a resolution of 640×360 pixels at 30fps and works with a time-of-flight depth sensor in the near infrared spectrum for depth measurements. For flat terrain walking, we record clips for pacing, trotting, a right turn, and a left turn motion, with a total duration of $6.0\,\mathrm{s}$ similar to data in the MoCap dataset [15]. We record an extended dataset (*Video (extended)*) for flat walking with further clips of walking, left-right-turn, and a start-stop motion fitting our target command distribution, now totalling $12.2\,\mathrm{s}$. For the other scenarios, stair climbing, box jumping, and standing, we record three short clips each. Examples of the recorded clips are shown in Figure 3.

Experimental setup. We conduct extensive quantitative experiments in simulation and deploy policies for each scenario to hardware. All experiments are conducted on a Unitree Go2 quadruped. The simulations are run in IsaacLab [16] with rsl-rl [17]. For Proximal Policy Optimization (PPO), we use the hyperparameters from rsl-rl, and for AMP from [18], respectively. All policies output joint position targets at $50\,\mathrm{Hz}$ that are tracked using a lower-level PD controller with $K_p = 25$ and $K_d = 0.5$ that runs at approximately

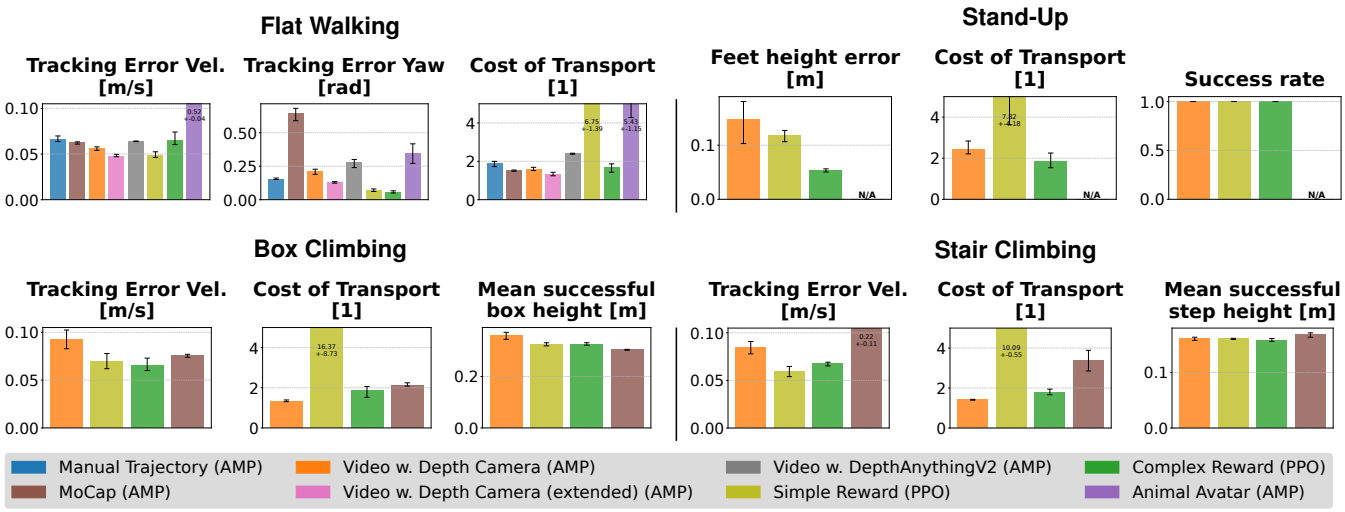

Fig. 4: Our framework leveraging casual videos performs competitively to the human-engineered baselines *Complex Reward* and *Manual Trajectory* as well as policies trained on the *MoCap* dataset. Shown are the mean and range between three seeds. **Flat walking**: Using depth measurements from the camera (*Depth Camera*) results in the best learning outcomes compared to a depth estimation model (*DepthAnythingV2*) and a full 3D reconstruction model (*Animal Avatar*). **Stand-up, Box, Stair**: Video data collected for each scenario offers advantages over MoCap data recorded on flat terrain. The slight variations in the cost of transport can translate to significant differences in the robot's actual movement, as shown in the supplementary video.

200 Hz. All policies receive the user command, base velocity, joint positions and velocities, projected gravity vector in base frame, and the previous actions of the last five timesteps as input. For box and stair, we provide the policies with privileged information about the environment. The policies run on an external workstation connected via Ethernet, and the low-level controllers run on the robot.

We utilize the cost of transport (CoT) to estimate the efficiency of the movement, as done in prior works [18]. We define the mechanical CoT as $\frac{\text{Power}}{\text{Weight} \times \text{Velocity}} = \sum_{\text{actuators}} [\tau\dot{\theta}]^+ / (W\|v\|)$, where τ is the joint torque, $\dot{\theta}$ is the motor velocity, W is the robot's weight, and $\|v\|$ is the velocity.

A. Flat Walking

Policies trained on our vision data achieve comparable task command tracking and locomotion efficiency to those trained on MoCap data and manually engineered baselines (Figure 4). This performance is achieved despite the inherently noisier nature of vision data. We show the deployment and qualitative imitation of the style from the video in Figure 3.

The *Video (extended)* dataset improves the overall command tracking accuracy and energy efficiency across a wide range of target commands (Figure 5). For instance, the additional left-right turn improves the yaw tracking accuracy, and including a start-stop motion improves tracking and imitation for a standing command. Recall that for us, it is comparatively easy to collect that additional data using video recordings

Compared to the complex reward baseline, policies trained with expert demonstration have the largest performance gap in tracking the yaw command. This is likely because the expert data contains a limited number of turning motions, as capturing controlled turning behaviors in animals is more challenging than straight walking. However, when including additional left-right turns in the *Video (extended)* dataset, the yaw tracking error is reduced.

Policies trained on the hand-designed and noise-free expert demonstrations do not achieve the same energy efficiency as policies trained on dog walking data from MoCap or vision demonstrations. We hypothesize that millions of years of evolution have endowed canine species with energy-efficient movements, whose intricacies are difficult to mimic by human designers.

Scaling learning from video would benefit from accessing readily available online videos. To assess the feasibility with our approach, we study the impact of using different sources of depth information. Policies trained with expert data using DepthAnythingV2 [14] or AnimalAvatarr [13] exhibit a significantly higher cost of transport, suggesting unnatural locomotion that does not transfer effectively to real-world applications. In fact, the reconstruction models suffer from physically unrealistic depth estimation of limbs, potentially causing issues such as leg or paw penetrations as shown in Figure 6. These factors significantly hinder the learning and transferability of locomotion policies.

B. Stand-Up, Box, Stairs

We test how the MoCap dataset recorded for walking on flat terrain generalizes to boxes and stairs, and record separate video clips for each scenario with our framework.

Despite climbing similar stair heights, the MoCap dataset fails to learn natural and effective locomotion as indicated by a high cost of transport. This difference in the cost of transport results in highly inefficient locomotion, as demonstrated in the supplementary video. The MoCap dataset generalizes

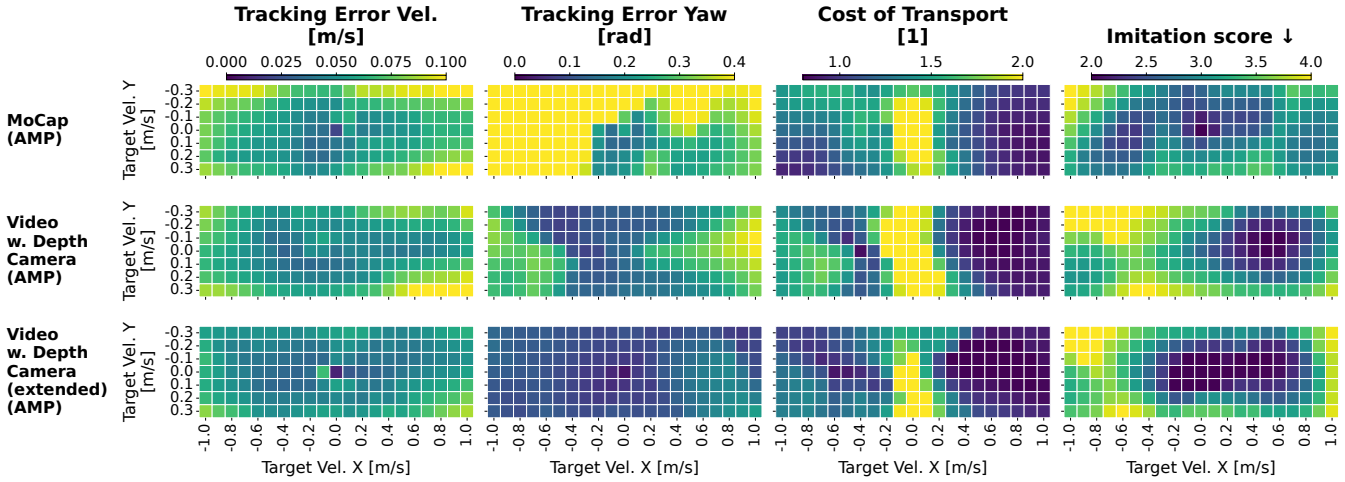

Fig. 5: Extending the size of the visual expert dataset (*Video (extended)*) with more diverse video clips, including left-right turns and start-stop motions, improves task tracking, CoT, and imitation across the distribution of target commands. For instance, if the task includes a standing command ($v_x = v_y = 0$), the demonstrations should contain a standing position, which was only included in the *Video (extended)* dataset. The significant yaw tracking error that we observe for *MoCap* for our setting is explained in **??**. Overall, fitting the expert dataset to the desired task distribution is crucial for improving task fulfillment and imitation.

slightly better to the box-climbing scenario, likely because it consists of a larger proportion of flat walking. However, the MoCap dataset limits the robot's ability to climb stairs as high as our vision dataset, achieving only $30\,\mathrm{cm}$ compared to $39\,\mathrm{cm}$.

For the stand-up task, the task reward is defined by reaching a specified height with the front feet. While the task is solved successfully with our framework, it is the most challenging, in particular for motion retargeting. Embodiment differences between the dog and the robot cause the robot's knees to penetrate the ground plane during kinematic retargeting (Figure 3 (middle)). This tension between imitation data and physical feasibility increases the challenges of policy learning. Despite these we can successfully learn a stand-up behavior from the recorded video data.

Deployment. We found that two additional well-motivated reward terms are necessary for deployment. First, we add a penalty on the applied robot torques. The style reward only encourages imitation but not necessarily in an efficient manner, i.e., the robot might otherwise apply maximum torques to improve imitation. Second, for box and stair climbing stepping close to the edges of a step decreases sim-to-real robustness. Thus, we included a reward term that penalizes stepping close to the edges of the steps as it it not contained in the style reward. Demonstrations are shown in the supplementary video.

IV. CONCLUSION

Data-driven methods, such as learning from video data, become increasingly popular in robot control. We demonstrate an approach to leverage casual video recordings for quadruped policy learning. Our framework lowers barriers to train quadruped locomotion from demonstrations by replacing resource-intensive methods with handheld RGB(D). Generally, we find that learning from animal data achieves energy-efficient locomotions above the levels of manually tuned methods, further strengthening the point of reducing human engineering. We find that accurate depth prediction models are key to accessing further data sources, such as YouTube videos, for policy learning with our approach.

(a) Camera depth measurements

(b) Reconstruction with camera depth

(c) AnimalAvatar reconstruction

(d) Reconstruction with DepthAnythingV2 depth

Fig. 6: Examples of depth reconstructions demonstrating the weaknesses of each approach. (a) + (b) Camera depth measurements and reconstruction, (c) AnimalAvatar [13] (uses only RGB as input), and (d) DepthAnythingV2 [14]. Camera depth measurements provide the most realistic reconstructions, though their accuracy can be limited for fast-moving parts of the scene. Both DepthAnythingV2 and AnimalAvatar can lead to physically unrealistic reconstructions with penetrating limbs circled in the images.

REFERENCES

[1] N. Bohlinger *et al.*, "One policy to run them all: An end-to-end learning approach to multi-embodiment locomotion," *Conference on Robot Learning (CoRL)*, 2024.

[2] N. Heess *et al.*, "Emergence of locomotion behaviours in rich environments," *arxiv.org/abs/1707.02286*, 2017.

[3] J. Lee, J. Hwangbo, L. Wellhausen, V. Koltun, and M. Hutter, "Learning quadrupedal locomotion over challenging terrain," *Science robotics*, 2020.

[4] S. L. Li *et al.*, "Controlling diverse robots by inferring jacobian fields with deep networks," *Nature*, 2025.

[5] K. Black *et al.*, "Pi0: A vision-language-action flow model for general robot control," *arxiv.org/abs/2410.24164*, 2024.

[6] T. He *et al.*, "Asap: Aligning simulation and real-world physics for learning agile humanoid whole-body skills," *Robotics Science and Systems*, 2025.

[7] X. B. Peng, Z. Ma, P. Abbeel, S. Levine, and A. Kanazawa, "Amp: Adversarial motion priors for stylized physics-based character control," *ACM Transactions on Graphics*, 2021.

[8] N. Karaev, I. Rocco, B. Graham, N. Neverova, A. Vedaldi, and C. Rupprecht, "Cotracker: It is better to track together," *European Conference on Computer Vision*, 2023.

[9] Z. Teed and J. Deng, "Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras," *Advances in neural information processing systems*, 2021.

[10] N. Ravi *et al.*, "Sam 2: Segment anything in images and videos," *International Conference on Learning Representations*, 2024.

[11] X. B. Peng, E. Coumans, T. Zhang, T.-W. Lee, J. Tan, and S. Levine, "Learning agile robotic locomotion skills by imitating animals," *Robotics: Science and Systems*, 2020.

[12] H. Zhang, S. Starke, T. Komura, and J. Saito, "Mode-adaptive neural networks for quadruped motion control," *ACM Transactions on Graphics*, 2018.

[13] R. Sabathier, N. J. Mitra, and D. Novotny, "Animal avatars: Reconstructing animatable 3d animals from casual videos," *European Conference on Computer Vision*, 2024.

[14] L. Yang *et al.*, "Depth anything v2," *Computer Vision and Pattern Recognition Conference*, 2024.

[15] Y. Zhang, Q. H. Vuong, K. Song, X.-Y. Gong, and K. W. Ross, "Efficient entropy for policy gradient with multidimensional action space," *arxiv.org/abs/1806.00589*, 2018.

[16] M. Mittal *et al.*, "Orbit: A unified simulation framework for interactive robot learning environments," *IEEE Robotics and Automation Letters*, 2023.

[17] N. Rudin, D. Hoeller, P. Reist, and M. Hutter, "Learning to walk in minutes using massively parallel deep reinforcement learning," *Conference on Robot Learning*, 2021.

[18] A. Escontrela *et al.*, "Adversarial motion priors make good substitutes for complex reward functions," *International Conference on Intelligent Robots and Systems*, 2022.