Learning Vision-Guided Quadrupedal Locomotion End-to-End with Cross-Modal Transformers

Ruihan Yang* UC San Diego
Minghao Zhang* Tsinghua University
Nicklas Hansen UC San Diego
Huazhe Xu UC Berkeley
Xiaolong Wang UC San Diego

***Abstract*—We propose to solve quadrupedal locomotion tasks using Reinforcement Learning (RL) with a Transformer-based model that learns to combine proprioceptive information and high-dimensional depth sensor inputs. While learning-based locomotion has made great advances using RL, most methods still rely on domain randomization for training blind agents that generalize to challenging terrains. Our key insight is that proprioceptive states only offer contact measurements for immediate reaction, whereas an agent equipped with visual sensory observations can learn to proactively maneuver environments with obstacles by anticipating changes in the environment many steps ahead. In this paper, we introduce LocoTransformer, an end-to-end RL method for quadrupedal locomotion that leverages a Transformer-based model for fusing proprioceptive states and visual observations. We evaluate our method in challenging simulated environments with different obstacles. We show that our method obtains significant improvements over policies with only proprioceptive state inputs, and that Transformer-based models further improve generalization across environments. Our project page with videos is at https://LocoTransformer.github.io/.**

I. Introduction

Legged locomotion is one of the core problems in robotics research. It expands the reach of robots and enables them to solve a wide range of tasks, from daily life delivery to planetary exploration in challenging, uneven terrain [15], [2]. Recently, with the success of deep Reinforcement Learning (RL) in navigation [54], [26], [81], [41] and robotic manipulation tasks [73], [39], we have also witnessed tremendous improvement of locomotion skills for quadruped robots, allowing them to walk on uneven terrain [80], [79], and even generalize to real world environments with mud, snow, and running water [44].

While these results are encouraging, most RL methods learn a robust controller for *blind* quadrupedal locomotion, using only the proprioceptive measurements. Lee et al. [44] train a robust RL quadrupedal locomotion policy that can be applied to challenging terrains with domain randomization and large-scale training in simulation. However, is domain randomization with blind agents sufficient for general legged locomotion?

By studying eye movement during human locomotion, Matthis et al. [52] show that humans rely heavily on eye-body coordination when walking, and the gaze depends on the environment, e.g. whether humans walk in flat or rough terrain. This finding motivates the use of visual input to improve quadrupedal locomotion in complicate environment. While handling uneven terrain is still possible without vision, a blind agent is unable to e.g. consistently avoid obstacles in Figure 1. To maneuver around such obstacles, the agent needs to perceive the obstacles at a distance and dynamically make adjustments to its trajectory to avoid any collision, visual observations can therefore play an important role in improving locomotion skills.

In this paper, we propose to combine proprioceptive states and forward-facing depth camera inputs with a cross-modal Transformer for learning RL locomotion policies. Our key insight is that proprioceptive states (i.e. robot pose, Inertial Measurement Unit (IMU) readings, and local joint rotations) give a precise measurement of the current interaction between the agent and the ground for *immediate* reaction, while visual inputs from a depth sensor can help the agent plan to maneuver large obstacles in its path. Inspired by the recent development of multi-modal reasoning with Transformers [77], [75], [22], we propose to fuse two streams of inputs, namely proprioceptive states and depth images, for RL using Transformers, which enables the model to reason using complementary information from both modalities. Transformers also offer a mechanism for agents to attend to certain visual regions (e.g. objects) that are critical for its short-term decision making, which may in turn lead to a more generalizable and interpretable policy.

Our proposed Transformer-based model for locomotion, *LocoTransformer*, consists of the following two encoders: an MLP for proprioceptive states, and a ConvNet for depth image inputs. We obtain a feature embedding from the proprioceptive states and multiple image patch embeddings from the depth images, which are used jointly as token inputs for the Transformer encoders. Features for both modalities are then fused with information propagation among all the tokens using self-attention. Finally, we combine both features for policy action prediction. The resulting model is trained end-to-end directly using rewards, *without* hierarchical RL [59], [40], [30], [37] nor pre-defined controllers [14], [20].

We evaluate our proposed method on challenging simulated environments as shown in Figure 1, including tasks such as maneuvering around obstacles of different sizes and shapes. We show that jointly learning policies with both proprioceptive states and vision significantly improves locomotion in challenging environments, and policies further benefit from adopting our cross-modal Transformer. We also show that *LocoTransformer* generalizes much better to unseen environments. Lastly, we qualitatively show our method learns to anticipate changes in the environment using vision as guidance.

*Equal Contribution

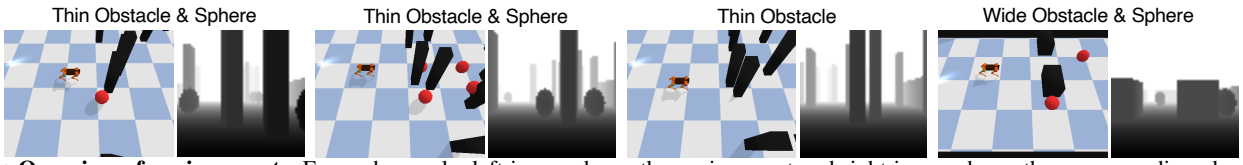


Fig. 1: **Overview of environments.** For each sample, left image shows the environment and right image shows the corresponding observation. The agent is tasked to move forward while avoiding black obstacles and collecting red spheres. Environments are randomized in each episode.

II. Related Work

Learning Legged Locomotion. Developing legged locomotion controllers has been a long standing problem in robotics [55], [63], [74], [24], [83], [4]. While encouraging results have been achieved using Model Predictive Control (MPC) and trajectory optimization [23], [9], [17], [8], [18], [25], [3], these methods require in-depth knowledge of the environment and substantial efforts in manual parameter tuning, which makes these methods challenging to apply to complex environments. Alternatively, model-free RL can learn general policies for tasks with challenging terrain [42], [84], [51], [60], [61], [72], [34], [44], [80], [35], [37], [79]. Xie et al. [80] use dynamics randomization to generalize RL locomotion policy in different environments, and Peng et al. [61] use animal videos to provide demonstrations for imitation learning. Most approaches currently rely only on proprioceptive states without other sensory signals.

Vision-based Reinforcement Learning. To generalize RL to real-world applications beyond state inputs, a lot of effort has been made in RL with visual inputs [64], [36], [46], [47], [58], [39], [21], [50], [82], [43], [67], [68], [66]. Srinivas et al. [67] apply contrastive self-supervised representation learning [29] together with the RL objective to improve the sample efficiency in vision-based RL. Hansen et al. [28] further extend the joint representation learning and RL for better generalization to out-of-distribution environments. Instead of using a single modality input in RL, researchers have also looked into combining multi-modalities for manipulation tasks [45], [6] and locomotion control [30], [53], [20], [38]. Escontrela et al. [20] combine proprioceptive states and LiDAR inputs for learning quadrupedal locomotion using RL using MLPs. Jain et al. [38] use Hierarchical RL (HRL) for locomotion, which learns high-level policies under visual guidance and low-level motor control policies with IMU inputs. Different from previous work, we provide a simple yet effective method to combine proprioceptive states and depth image inputs with a Transformer model, which allows end-to-end training without HRL.

Transformers and Multi-modal Learning. The Transformer model has been widely applied in the fields of language processing [77], [16], [5] and visual recognition and synthesis [78], [57], [12], [19], [7], [10]. Besides achieving impressive performance in a variety of language and vision tasks, the Transformer also provides an effective mechanism for multi-modal reasoning by taking different modality inputs as tokens for self-attention [69], [71], [48], [70], [11], [49], [62], [33], [32], [1], [31]. For example, Sun et al. [70] propose to use a Transformer to jointly model video frames and their corresponding captions from instructional videos for representation learning. Going beyond language and vision, we propose to utilize cross-modal Transformers to fuse proprioceptive states and visual inputs. To our knowledge, this is the first work using cross-modal Transformers for locomotion.

III. Method

We propose to incorporate proprioceptive and visual information for locomotion tasks using a novel Transformer model, *LocoTransformer*. Figure 2 provides an overview of our architecture. Our model consists of two components: (i) separate modality encoders for proprioceptive and visual inputs that project both modalities into a latent feature space; (ii) a shared transformer encoder performing spatial attention over visual tokens, and cross-modality attention over proprioceptive and visual features to predict the actions and values.

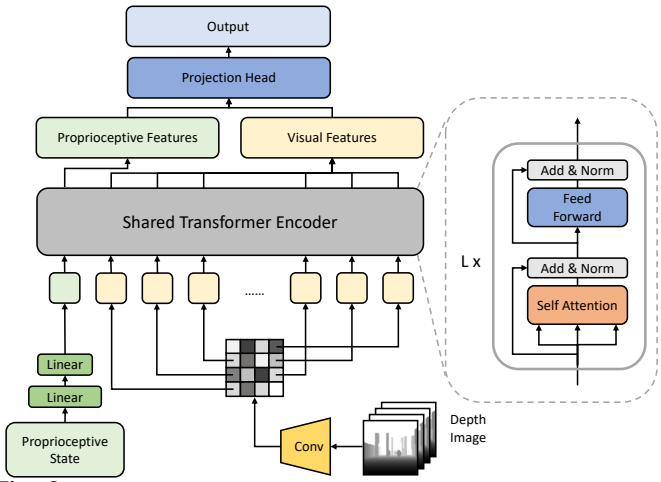


Fig. 2: **Architectural overview**. We process proprioceptive states with a MLP and depth image inputs with a ConvNet. We take proprioceptive feature as a single token, split the spatial visual feature representation into $N \times N$ tokens and feed all tokens into our Transformer encoder. The output of the Transformer are further processed by the projection head to predict value or action distribution.

A. Separate Modality Encoders

Proprioceptive states and visual observations are distinctively different: the proprioceptive input is a 93-D vector and we use depth images to encode the visual observations. To facilitate domain-specific characteristics from both modalities, we use two separate, domain-specific encoders for proprioceptive and visual data, and unify the representation in a latent space.

We use an MLP to encode the proprioceptive input vector into proprioceptive features $E^{\text{prop}} \in \mathbb{R}^{C^{\text{prop}}}$, where C^{prop} is the proprioceptive feature dimension. We provide the policy with visual information using first-person view depth images. In comparison to commonly used third-person view in previous visual reinforcement learning problems [27], [67], [43], first-person view captures the obstacles from the perspective of

the acting robot, and it better reflects potential real-world applications. For visual observations, we stack 4 depth images as input, and encode the stacked depth images using a ConvNet. The ConvNet encodes the depth map inputs into a spatial feature representations E^{visual} with shape $C \times N \times N$ in the latent space, where C is the channel number, and N is the width and height dimension of the feature. While the first-person view is more realistic, the moving camera and the limited field-of-view make learning visual policies significantly more challenging. This makes it essential to leverage proprioceptive information to improve visual understanding. In the following, we present our proposed method for fusing the two modalities and improving their joint representation using a Transformer.

B. Transformer Encoder

Locomotion in unstructured environments requires the agent to be aware of its surroundings. As in Figure 1, the agent should be aware of local information like nearby obstacles, as well as global information such as overall layout, in order to traverse the environment effectively. To do so, the agent needs a mechanism for effectively fusing visual observations containing mainly global information, and proprioceptive states containining local information. Given a spatial, visual feature map with shape $C \times N \times N$ from the ConvNet encoder, we split the spatial features into $N \times N$ different C-dimensional token embeddings $t^{\text{visual}} \in \mathbb{R}^C$ (yellow tokens in Figure 1), each corresponding to a local visual region. We use a fully-connected layer to project the proprioceptive features into a C-dimensional token embedding $t^{\text{prop}} \in \mathbb{R}^C$ (the green token in Figure 1), such that we have $N \times N + 1$ tokens in total. Formally, the tokens are obtained by $t^{\text{prop}} = W^{\text{prop}}(E^{\text{prop}}) + b^{\text{prop}}$, $T_0 = [t^{\text{prop}}, t^{\text{visual}}_{0,0}, t^{\text{visual}}_{0,1}, ..., t^{\text{visual}}_{N-1,N-1}]$, $t^{\text{prop}} \in \mathbb{R}^C$, $t^{\text{visual}}_{i,j} \in \mathbb{R}^C$, where $t^{\text{visual}}_{i,j}$ is the token at spatial position (i, j) of the visual features E^{visual}, and $W^{\text{prop}}, b^{\text{prop}}$ are the weights and biases, respectively, of the linear proprioceptive token embedding. We denote T_m as the sequence of tokens after m Transformer encoder layers, and define T_0 as the input token sequence.

We adopt a stack of Transformer encoder layers [77] to fuse information from the proprioceptive and visual tokens. Specifically, we formulate the Self-Attention (SA) mechanism of the Transformer encoder as a scaled dot-product attention mechanism, omitting subscripts for brevity:

$$T^q, T^k, T^v = TU^q, TU^k, TU^v; W^{\text{sum}} = \text{Softmax}(T^q {T^k}^\top / \sqrt{D})$$

$\text{SA}(T) = W^{\text{sum}} T^v U^{\text{SA}}$, where $U^q, U^k, U^v, U^{\text{SA}} \in \mathbb{R}^{C \times C}$, $W^{\text{sum}} \in \mathbb{R}^{[(N^2+1)]^2}$, D is the dimensionality of the SA layer. SA applies three linear transformations on tokens to produce embeddings T^q, T^k, T^v, then compute a weighted sum over input tokens, where the weight $W^{sum}_{i,j}$ for each token pair (t_i, t_j) is computed as the dot-product of elements t_i and t_j scaled by $1/\sqrt{D}$ and normalized by a Softmax operation. After a matrix multiplication between weights W^{sum} and values T^v, we forward the result to a linear layer with parameters U^{SA}, and denote this as the output $\text{SA}(T)$.

Each Transformer encoder layer consists of a SA layer, two LayerNorm (LN) layers with residual connections, and an MLP as shown in Figure 2 (right). This is formally expressed as,

$$T'_m = \text{LN}(\text{SA}(T_m) + T_m), T_{m+1} = \text{LN}(\text{MLP}(T'_m) + T'_m)$$

where $T_m, T_{m+1} \in \mathbb{R}^{(N^2+1) \times C}$, T'_m is normalized SA. For SA is computed across multiple visual tokens and a single proprioceptive token, proprioceptive information may gradually vanish in transformer encoder layers; the residual connections allow it to propagate more easily through the network.

We stack L Transformer encoder layers. Performing multi-layer self-attention on proprioceptive and visual features enables our model to fuse tokens from both modalities at multiple levels of abstraction. Further, we emphasize that a Transformer-based fusion allows for spatial reasoning, as each token has a separate regional perceptive field, therefore self-attention enables the agent to explicitly attend to relevant visual regions. For modality-level fusion, direct application of a pooling operation across all tokens would easily dilute proprioceptive information since the number of visual tokens far exceed that of the proprioceptive information. To balance information from both modalities, we pool information separately for each modality. We compute the mean of all tokens from the same modality to get a single feature vector for each modality. We then concatenate the proprioceptive and the visual feature vector, and project the concatenated vector into a final output vector using an MLP, which we denote the *projection head*.

Implementation Details. We use the same observation space across all environments which is defined as follows: (i) **proprioceptive data:** a 93-D vector including IMU readings, local joint rotations, actions taken by agent, and the displacement of the base of the robot; and (ii) **visual data:** a stack of the 4 most recent depth images with shape $(64, 64)$. For the proprioceptive encoder and projection head, we use a 2-layer MLP with hidden dimensions 256. Our visual encoder outputs a 4×4 spatial feature map with 128 channels, following [56]. Our shared Transformer consists of 2 Transformer encoder layers with hidden dimension 256.

IV. Experiments

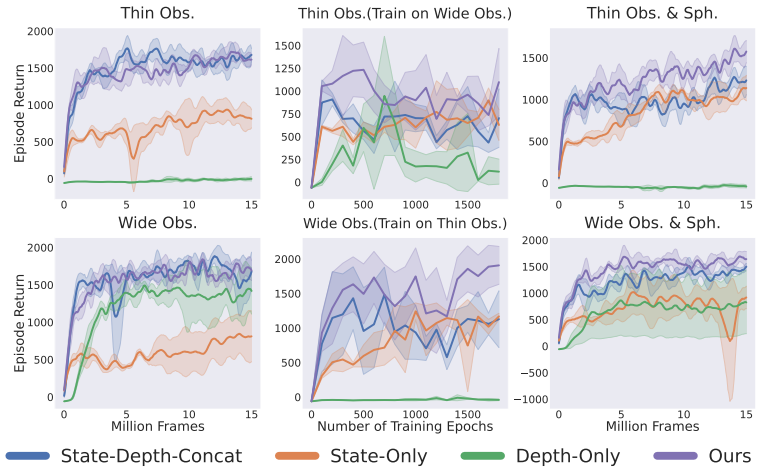
We evaluate our method using a simulated quadruped A1 Robot [76] in challenging environments in PyBullet [13].

A. Environments

We evaluate all methods in 6 distinct environments with varying obstacles to avoid, and spheres to collect for reward bonuses which are designed to evaluate high-level decision making, e.g. avoiding obstacles and collecting spheres. The environments are shown in Figure 1. We consider the following environments: **Wide Obstacle** (Wide Obs.): wide cuboid obstacles on a flat terrain, *without* spheres; **Wide Obstacle & Sphere** (Wide Obs. & Sph.): wide cuboid obstacles on a flat terrain, including spheres that give a reward bonus when collected; **Thin Obstacle** (Thin Obs.): numerous thin cuboid obstacles on a flat terrain, *without* spheres; **Thin Obstacle & Sphere** (Thin Obs. & Sph.): same obstacles setting as Thin Obs, but with spheres that give a reward bonus when collected;

B. Baseline and Experiment Setting

We train all agents using PPO [65], and compare our method to both a state-only baseline that only uses proprioceptive states, a depth-only baseline that only uses visual observation, and a baseline that uses proprioceptive states and vision without our



(a) **Training and evaluation curves** on environments with Obstacles and Sphere

Distance Moved ↑

	Thin Obs.(Train on Wide Obs.)	Wide Obs.(Train on Thin Obs.)
State-Only	$3.6_{\pm 1.3}$	$5.9_{\pm 0.9}$
Depth-Only	$1.1_{\pm 1.1}$	$0.1_{\pm 0.0}$
State-Depth-Concat	$5.57_{\pm 2.1}$	$7.14_{\pm 2.00}$
Ours	$\mathbf{8.2}_{\pm \mathbf{2.5}}$	$\mathbf{14.2}_{\pm \mathbf{2.8}}$

Collision Happened ↓

	Thin Obs.(Train on Wide Obs.)	Wide Obs.(Train on Thin Obs.)
State-Only	$456.3_{\pm 262.2}$	$545.1_{\pm 57.7}$
Depth-Only	$\mathbf{0.0}_{\pm \mathbf{0.0}}$	$\mathbf{0.0}_{\pm \mathbf{0.0}}$
State-Depth-Concat	$406.8_{\pm 89.5}$	$331.1_{\pm 192.8}$
Ours	$\mathbf{310.4}_{\pm \mathbf{131.3}}$	$\mathbf{82.2}_{\pm \mathbf{103.8}}$

(b) **Generalization.**

Fig. 3: **(a)** For environment without sphere, our method achieve comparable training performance (the first column) but much better evaluation performance on unseen environments (the second column). For environment with sphere (the third column), our method achieve better performance and sample efficiency. **(b)** We evaluate the generalization ability of all three methods by evaluating the policy on unseen environment. Our method moved longer distance, and less Collision happened with our method.

proposed Transformer; we denote it as the *State-Depth-Concat* baseline. For the State-Depth-Concat baseline, it use the exact same proprioceptive and visual encoder as our method. Instead of using a Transformer to fuse the multi-modality features, the State-Depth-Concat baseline uses a linear projection to project visual features into a feature vector that has the same dimensions as the proprioceptive features, and feed it into the value and policy networks. For all methods, the value and policy network share the same proprioceptive and visual encoder.

Evaluation Metric. We evaluate policies by (i) mean episode return (ii) the distance an agent moved along its target direction; and (iii) the number of time steps in which there is collision between the robot and an obstacle within an episode.

C. Training & Quantitative Evaluation Results

Navigation. We train all methods on navigation tasks with obstacles to evaluate the effectiveness of modal fusion and stability of locomotion. Results are shown in Figure 3 (first column). Both our method and the State-Depth-Concat baseline significantly outperforms the State-Only baseline in both the *Thin Obstacle* and *Wide Obstacle* environment, demonstrating the clear benefit of vision for locomotion in complex environments. We observe that the simpler State-Depth-Concat baseline performs just as well as our Transformer-based model in these environments. We conjecture that this is because the task of differentiating obstacles from flat terrain is not perceptually complex, and a simple concatenation is therefore sufficient for policy learning. Surprisingly, though Depth-Only baseline have no access to proprioceptive states, when the environment is relatively simple (like Wide Obs. environment), agent can learn a policy..

We further evaluate the generalization ability of methods by evaluating agents trained with thin obstacles to environments with wide obstacles, and vice versa. Figure 3 (second column) shows generalization measured by episode return, and Table 3b shows average distance moved and number of collisions. While the State-Depth-Concat baseline is sufficient for policy learning, Our Transformer-based method improves episode return in transfer by $\mathbf{69}\%$ and $\mathbf{56}\%$ in the *wide* and *thin* obstacle environments, respectively, over the State-Depth-Concat baseline. We observe that our method moves significantly farther on average, and reduces the number of collisions by $\mathbf{402}\%$ and $\mathbf{663}\%$ over the State-Depth-Concat and State-Only baselines, respectively, when trained on thin obstacles and evaluated on wide obstacles. Interestingly, we observe that the generalization ability of the State-Depth-Concat *decreases* as training progresses, whereas it for our method either plateaus or *increases* over time. This indicates that our method is more effective at capturing essential information in the visual and proprioceptive information during training, and is less prone to overfit to training environments.

Navigation with Spheres. We now consider a perceptually more challenging setting with the addition of spheres in the environment; results are shown in Figure 3 (third column). We observe that with the addition of spheres, the sample efficiency of both the State-Depth-Concat baseline and our method decreases. While spheres that provide positive reward provide the possibility for higher episode return, spheres increase complexity in two ways: (i) spheres may lure agents into areas where it is prone to get stuck; (ii) although spheres do not block the agent physically, they may occlude the agent's vision and can be visually difficult to distinguish from obstacles.

V. Conclusion

We propose to incorporate the proprioceptive and visual information with the proposed LocoTransformer model for locomotion control. By borrowing the visual inputs, we show that the robot can plan to walk through different sizes of obstacles both in seen and unseen environments. This shows our Transformer model provides an effective fusion mechanism between proprioceptive and visual information and new possibilities on RL with information from multi-modality.

REFERENCES

[1] Hassan Akbari, Linagzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text, 2021.

[2] P. Arena, L. Fortuna, M. Frasca, L. Patané, and M. Pavone. Realization of a cnn-driven cockroach-inspired robot. *2006 IEEE International Symposium on Circuits and Systems*, pages 4 pp.–, 2006.

[3] Gerardo Bledt and Sangbae Kim. Extracting legged locomotion heuristics with regularized predictive control. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 406–412. IEEE, 2020.

[4] Gerardo Bledt, Matthew J Powell, Benjamin Katz, Jared Di Carlo, Patrick M Wensing, and Sangbae Kim. Mit cheetah 3: Design and control of a robust, dynamic quadruped robot. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2245–2252. IEEE, 2018.

[5] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[6] R. Calandra, Andrew Owens, Dinesh Jayaraman, Justin Lin, Wenzhen Yuan, Jitendra Malik, E. Adelson, and Sergey Levine. More than a feeling: Learning to grasp and regrasp using vision and touch. *IEEE Robotics and Automation Letters*, 3:3300–3307, 2018.

[7] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020.

[8] Jan Carius, René Ranftl, Vladlen Koltun, and Marco Hutter. Trajectory optimization for legged robots with slipping motions. *IEEE Robotics and Automation Letters*, 4(3):3013–3020, 2019.

[9] Jared Di Carlo, Patrick M. Wensing, Benjamin Katz, Gerardo Bledt, and Sangbae Kim. Dynamic locomotion in the MIT cheetah 3 through convex model-predictive control. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2018, Madrid, Spain, October 1-5, 2018*, pages 1–9. IEEE, 2018.

[10] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *International Conference on Machine Learning*, pages 1691–1703. PMLR, 2020.

[11] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European Conference on Computer Vision*, pages 104–120. Springer, 2020.

[12] Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*, 2019.

[13] Erwin Coumans and Yunfei Bai. Pybullet, a python module for physics simulation for games, robotics and machine learning. http://pybullet.org, 2016–2021.

[14] Xingye Da, Zhaoming Xie, David Hoeller, Byron Boots, Animashree Anandkumar, Yuke Zhu, Buck Babich, and Animesh Garg. Learning a contact-adaptive controller for robust, efficient legged locomotion. *ArXiv*, abs/2009.10019, 2020.

[15] F. Delcomyn and M. Nelson. Architectures for a biomimetic hexapod robot. *Robotics Auton. Syst.*, 30:5–15, 2000.

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[17] Jared Di Carlo, Patrick M Wensing, Benjamin Katz, Gerardo Bledt, and Sangbae Kim. Dynamic locomotion in the mit cheetah 3 through convex model-predictive control. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1–9. IEEE, 2018.

[18] Yanran Ding, Abhishek Pandala, and Hae-Won Park. Real-time model predictive control for versatile dynamic motions in quadrupedal robots. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 8484–8490. IEEE, 2019.

[19] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[20] Alejandro Escontrela, George Yu, Peng Xu, Atil Iscen, and Jie Tan. Zero-shot terrain generalization for visual locomotion policies, 2020.

[21] Lasse Espeholt, Hubert Soyer, Remi Munos, Karen Simonyan, Volodymir Mnih, Tom Ward, Yotam Doron, Vlad Firoiu, Tim Harley, Iain Dunning, Shane Legg, and Koray Kavukcuoglu. Impala: Scalable distributed deep-rl with importance weighted actor-learner architectures, 2018.

[22] Valentin Gabeur, Chen Sun, Alahari Karteek, and C. Schmid. Multi-modal transformer for video retrieval. In *ECCV*, 2020.

[23] Christian Gehring, Stelian Coros, Marco Hutter, Michael Blösch, Mark A. Hoepflinger, and Roland Siegwart. Control of dynamic gaits for a quadrupedal robot. In *2013 IEEE International Conference on Robotics and Automation, Karlsruhe, Germany, May 6-10, 2013*, pages 3287–3292. IEEE, 2013.

[24] Hartmut Geyer, Andre Seyfarth, and Reinhard Blickhan. Positive force feedback in bouncing gaits? *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 270(1529):2173–2183, 2003.

[25] Ruben Grandia, Farbod Farshidian, Alexey Dosovitskiy, René Ranftl, and Marco Hutter. Frequency-aware model predictive control. *IEEE Robotics and Automation Letters*, 4(2):1517–1524, 2019.

[26] Saurabh Gupta, Varun Tolani, James Davidson, Sergey Levine, R. Sukthankar, and J. Malik. Cognitive mapping and planning for visual navigation. *International Journal of Computer Vision*, 128:1311–1330, 2019.

[27] Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*, 2019.

[28] Nicklas Hansen and Xiaolong Wang. Generalization in reinforcement learning by soft data augmentation. In *International Conference on Robotics and Automation*, 2021.

[29] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735, 2020.

[30] Nicolas Heess, Dhruva TB, Srinivasan Sriram, Jay Lemmon, Josh Merel, Greg Wayne, Yuval Tassa, Tom Erez, Ziyu Wang, S. M. Ali Eslami, Martin A. Riedmiller, and David Silver. Emergence of locomotion behaviours in rich environments. *CoRR*, abs/1707.02286, 2017.

[31] Lisa Anne Hendricks, John Mellor, Rosalia Schneider, Jean-Baptiste Alayrac, and Aida Nematzadeh. Decoupling the role of data, attention, and losses in multimodal transformers. *arXiv preprint arXiv:2102.00529*, 2021.

[32] Ronghang Hu and Amanpreet Singh. Unit: Multimodal multitask learning with a unified transformer, 2021.

[33] Zhicheng Huang, Zhaoyang Zeng, Yupan Huang, Bei Liu, Dongmei Fu, and Jianlong Fu. Seeing out of the box: End-to-end pre-training for vision-language representation learning, 2021.

[34] Jemin Hwangbo, J. Lee, A. Dosovitskiy, Dario Bellicoso, Vassilios Tsounis, V. Koltun, and M. Hutter. Learning agile and dynamic motor skills for legged robots. *Science Robotics*, 4, 2019.

[35] Atil Iscen, Ken Caluwaerts, Jie Tan, Tingnan Zhang, Erwin Coumans, Vikas Sindhwani, and Vincent Vanhoucke. Policies modulating trajectory generators. In *2nd Annual Conference on Robot Learning, CoRL 2018, Zürich, Switzerland, 29-31 October 2018, Proceedings*, volume 87 of *Proceedings of Machine Learning Research*, pages 916–926. PMLR, 2018.

[36] Max Jaderberg, Volodymyr Mnih, Wojciech Marian Czarnecki, Tom Schaul, Joel Z. Leibo, David Silver, and Koray Kavukcuoglu. Reinforcement learning with unsupervised auxiliary tasks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

[37] Deepali Jain, Atil Iscen, and Ken Caluwaerts. Hierarchical reinforcement learning for quadruped locomotion. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2019, Macau, SAR, China, November 3-8, 2019*, pages 7551–7557. IEEE, 2019.

[38] Deepali Jain, Atil Iscen, and Ken Caluwaerts. From pixels to legs: Hierarchical learning of quadruped locomotion, 2020.

[39] Divye Jain, Andrew Li, Shivam Singhal, Aravind Rajeswaran, Vikash Kumar, and Emanuel Todorov. Learning deep visuomotor policies for dexterous hand manipulation. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 3636–3643, 2019.

[40] Yiding Jiang, Shixiang Gu, Kevin Murphy, and Chelsea Finn. Language as an abstraction for hierarchical deep reinforcement learning. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances*

in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada, pages 9414–9426, 2019.

[41] G. Kahn, P. Abbeel, and Sergey Levine. Badgr: An autonomous self-supervised learning-based navigation system. *IEEE Robotics and Automation Letters*, 6:1312–1319, 2021.

[42] Nate Kohl and Peter Stone. Policy gradient reinforcement learning for fast quadrupedal locomotion. In *Proceedings of the 2004 IEEE International Conference on Robotics and Automation, ICRA 2004, April 26 - May 1, 2004, New Orleans, LA, USA*, pages 2619–2624. IEEE, 2004.

[43] Michael Laskin, Kimin Lee, Adam Stooke, Lerrel Pinto, Pieter Abbeel, and Aravind Srinivas. Reinforcement learning with augmented data. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[44] Joonho Lee, Jemin Hwangbo, Lorenz Wellhausen, Vladlen Koltun, and Marco Hutter. Learning quadrupedal locomotion over challenging terrain. *Science robotics*, 5(47), 2020.

[45] Michelle A Lee, Yuke Zhu, Peter Zachares, Matthew Tan, Krishnan Srinivasan, Silvio Savarese, Li Fei-Fei, Animesh Garg, and Jeannette Bohg. Making sense of vision and touch: Learning multimodal representations for contact-rich tasks. *IEEE Transactions on Robotics*, 36(3):582–596, 2020.

[46] Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. *J. Mach. Learn. Res.*, 17:39:1–39:40, 2016.

[47] Sergey Levine, Peter Pastor, Alex Krizhevsky, Julian Ibarz, and Deirdre Quillen. Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. *Int. J. Robotics Res.*, 37(4-5):421–436, 2018.

[48] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

[49] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020.

[50] Xingyu Lin, Harjatin Singh Baweja, George Kantor, and David Held. Adaptive auxiliary task weighting for reinforcement learning. *Advances in neural information processing systems*, 32, 2019.

[51] Y. Luo, Jonathan Hans Soeseno, T. Chen, and Wei-Chao Chen. Carl: Controllable agent with reinforcement learning for quadruped locomotion. *ArXiv*, abs/2005.03288, 2020.

[52] Jonathan Samir Matthis, Jacob L Yates, and Mary M Hayhoe. Gaze and the control of foot placement when walking in natural terrain. *Current Biology*, 28(8):1224–1233, 2018.

[53] Josh Merel, Saran Tunyasuvunakool, Arun Ahuja, Yuval Tassa, Leonard Hasenclever, Vu Pham, Tom Erez, Greg Wayne, and Nicolas Heess. Catch & carry: reusable neural controllers for vision-guided whole-body tasks. *ACM Trans. Graph.*, 39(4):39, 2020.

[54] P. Mirowski, Razvan Pascanu, Fabio Viola, Hubert Soyer, Andy Ballard, Andrea Banino, Misha Denil, R. Goroshin, L. Sifre, K. Kavukcuoglu, D. Kumaran, and R. Hadsell. Learning to navigate in complex environments. *ArXiv*, abs/1611.03673, 2017.

[55] Hirofumi Miura and Isao Shimoyama. Dynamic walk of a biped. *The International Journal of Robotics Research*, 3(2):60–74, 1984.

[56] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, February 2015.

[57] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In *International Conference on Machine Learning*, pages 4055–4064. PMLR, 2018.

[58] Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *ICML*, 2017.

[59] X. Peng, G. Berseth, KangKang Yin, and M. V. D. Panne. Deeploco: dynamic locomotion skills using hierarchical deep reinforcement learning. *ACM Trans. Graph.*, 36:41:1–41:13, 2017.

[60] Xue Bin Peng, Pieter Abbeel, Sergey Levine, and Michiel van de Panne. Deepmimic: Example-guided deep reinforcement learning of physics-based character skills. *ACM Trans. Graph.*, 37(4):143:1–143:14, July 2018.

[61] Xue Bin Peng, Erwin Coumans, Tingnan Zhang, Tsang-Wei Edward Lee, Jie Tan, and Sergey Levine. Learning agile robotic locomotion skills by imitating animals. In *Robotics: Science and Systems*, 07 2020.

[62] Aditya Prakash, Kashyap Chitta, and Andreas Geiger. Multi-modal fusion transformer for end-to-end autonomous driving. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[63] Marc H Raibert. Hopping in legged systems—modeling and simulation for the two-dimensional one-legged case. *IEEE Transactions on Systems, Man, and Cybernetics*, (3):451–463, 1984.

[64] Alexander Sax, Bradley Emi, Amir R. Zamir, Leonidas J. Guibas, Silvio Savarese, and Jitendra Malik. Mid-level visual representations improve generalization and sample efficiency for learning visuomotor policies. 2018.

[65] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[66] Max Schwarzer, Ankesh Anand, Rishab Goel, R Devon Hjelm, Aaron Courville, and Philip Bachman. Data-efficient reinforcement learning with self-predictive representations. *arXiv preprint arXiv:2007.05929*, 2020.

[67] Aravind Srinivas, Michael Laskin, and Pieter Abbeel. Curl: Contrastive unsupervised representations for reinforcement learning. *arXiv preprint arXiv:2004.04136*, 2020.

[68] Adam Stooke, Kimin Lee, Pieter Abbeel, and Michael Laskin. Decoupling representation learning from reinforcement learning. arXiv:2004.14990.

[69] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019.

[70] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7464–7473, 2019.

[71] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.

[72] Jie Tan, Tingnan Zhang, Erwin Coumans, Atil Iscen, Yunfei Bai, Danijar Hafner, Steven Bohez, and Vincent Vanhoucke. Sim-to-real: Learning agile locomotion for quadruped robots. In Hadas Kress-Gazit, Siddhartha S. Srinivasa, Tom Howard, and Nikolay Atanasov, editors, *Robotics: Science and Systems XIV, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA, June 26-30, 2018*, 2018.

[73] Stephen Tian, Frederik Ebert, Dinesh Jayaraman, Mayur Mudigonda, Chelsea Finn, Roberto Calandra, and Sergey Levine. Manipulation by feel: Touch-based control with deep predictive models. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 818–824. IEEE, 2019.

[74] Nick Torkos and Michiel van de Panne. Footprint–based quadruped motion synthesis. In *Proceedings of the Graphics Interface 1998 Conference, June 18-20, 1998, Vancouver, BC, Canada*, pages 151–160, June 1998.

[75] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Z. Kolter, Louis-Philippe Morency, and R. Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. *Proceedings of the conference. Association for Computational Linguistics. Meeting*, 2019:6558–6569, 2019.

[76] Unitree. A1: More dexterity, more posibility, 2018.

[77] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

[78] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018.

[79] Zhaoming Xie, Xingye Da, Buck Babich, Animesh Garg, and Michiel van de Panne. Glide: Generalizable quadrupedal locomotion in diverse environments with a centroidal model. *CoRR*, abs/2104.09771, 2021.

[80] Zhaoming Xie, Xingye Da, Michiel van de Panne, Buck Babich, and Animesh Garg. Dynamics randomization revisited: A case study for quadrupedal locomotion. *CoRR*, abs/2011.02404, 2020.

[81] Wei Yang, X. Wang, Ali Farhadi, A. Gupta, and R. Mottaghi. Visual semantic navigation using scene priors. *ArXiv*, abs/1810.06543, 2019.

[82] Denis Yarats, Amy Zhang, Ilya Kostrikov, Brandon Amos, Joelle Pineau, and Rob Fergus. Improving sample efficiency in model-free reinforcement learning from images, 2019.

[83] KangKang Yin, Kevin Loken, and Michiel Van de Panne. Simbicon: Simple biped locomotion control. *ACM Transactions on Graphics (TOG)*, 26(3):105–es, 2007.

[84] H. Zhang, S. Starke, T. Komura, and Jun Saito. Mode-adaptive neural networks for quadruped motion control. *ACM Transactions on Graphics (TOG)*, 37:1 – 11, 2018.