RESEARCH STATEMENT

Chengzhou Tang (chengzhout@gmail.com)

Introduction

We are witnessing an intelligent computing revolution led by Large Language Models (LLMs). ChatGPT (GPT-3.5&4) acquired 1 million users only five days after the launching, and its applications have a broad spectrum ranging from education and legal compliance to content creation and robotics. The model's versatility stems from not only the inherent nature of language as an interface of human minds, but also the model's ability to learn basic reasoning mechanisms embedded in languages. However, this linguistic and logical intelligence comes at a huge economic and environmental cost [1], and potentially cause social inequality [2].

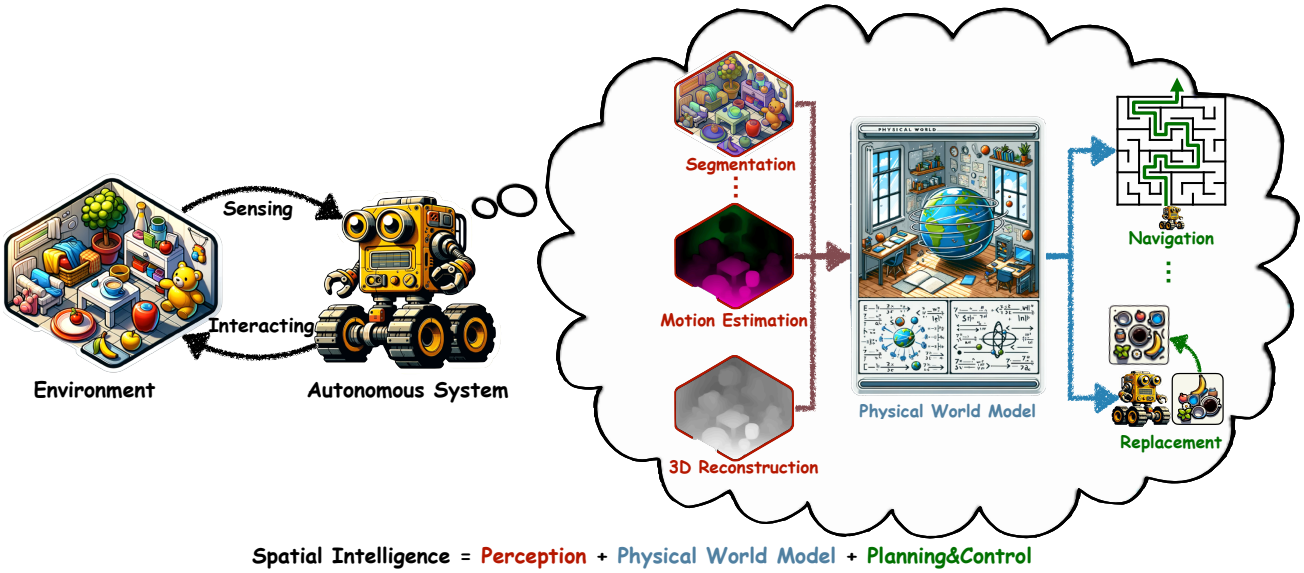


Figure 1: My vision for advancing spatial intelligence is rooted in my prior works on low&mid-level perception tasks. Building on this, I aim to develop a physical world model that transforms perceived information into a representation that contains physic rules explicitly, and integrating it with cutting-edge control algorithms to enhance the capabilities of autonomous systems.

Instead, autonomous systems can already perform various tasks without abstract reasoning. Equipped with classical planning and control algorithms, ground robots can navigate and avoid obstacles by perceiving static scenes and dynamic objects, such as moving cargoes in a factory or vacuuming a room; while industrial robotic arms can pick and sort objects based on only shapes, weights, and positions of objects. Although these systems are typically tailored for specific use cases, they have demonstrated the potential of spatial intelligence, which is orthogonal to linguistic and logical intelligence. Developing generic spatial intelligent systems that interact with the physical world and understand the underlying principles could have significant social impacts. The impacts extend beyond robotics, with other applications such as mixed reality, civil engineering, agriculture and assistants for the blind and visually impaired community.

Previous Researches

In pursuit of this goal, my previous researches focus on low&mid-level perceptions. Perception is the foundation of spatial intelligence that acquires, interprets, and understands sensory information from its environment. The perception tasks, including but not limited to 3D reconstruction, stereo matching, optical flow, and segmentation, have achieved significant progress in the past decade, driven by deep neural networks.

Distinct from the works that attempt to rediscover the physical rules implicitly through training on large amount of data, my researches adopts a first-principles strategy: I made minimization differentiable for training and leveraged neural networks only for representation learning. This innovative methodology embeds physical rules through objective functions to be minimized, enabling the learning of a unified model that is proficient in multiple perception tasks and capable of generalizing to new tasks in a zero-shot manner. In the rest of this section, I will chronologically introduce this unique line of researches:

- **Learning 3D Reconstruction with Multi-View Geometry.** The first challenge for a spatial intelligent system is the precise determination of its positions and understanding the 3D structure of the environments. This challenge was conventionally addressed by Simultaneous Localization and Mapping (SLAM) or Structure-from-Motion (SfM) systems, while more recent advances have explored neural network-based solutions [3, 4].

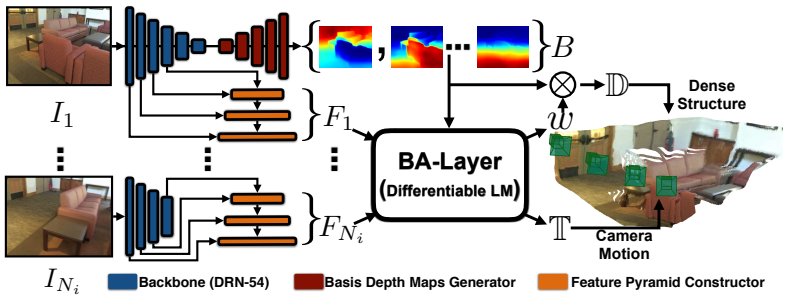
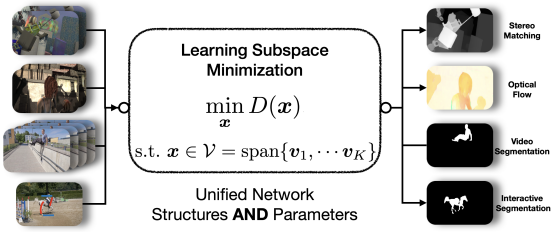


Figure 2: The BA-Net architecture. It contains a backbone network, a basis depth maps generation decoder, a feature pyramid decoder, and a differentiable bundle adjustment [5] Layer.

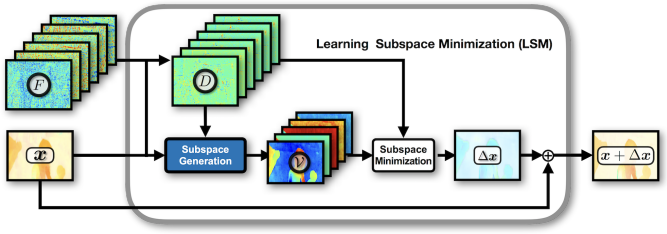
  My research represents a pioneering effort in bridging the gap between these two methodologies. I innovatively combined the principles of multi-view geometry [6] with convolutional neural networks (CNNs). A feature-metric projection error is defined on features extracted by the network, serving as the objective function for 3D reconstruction. The network also generates a basis representation for depth maps, to further regularize the minimization. This novel methodology was highlighted at ICLR 2019 as an oral presentation [7]. Significantly, the concept of minimizing the feature-metric error has set a new standard, influencing many subsequent studies on various topics [8, 9, 10, 11].

- **Generic Low&Mid-Level Perceptions.** Understanding the static environment and its own position is merely the initial stage for a spatial intelligent system to interact with its surroundings. In more common dynamic environments, these systems must advance to perceiving scene dynamics, caused by either the active movement of objects or the passive movement manipulated by external forces, such as robotic arms. To facilitate this dynamic

perception capability, it is critical to not only estimate the structure and movement of objects through stereo and optical flow techniques but also to effectively segment image pixels into distinct objects. All these processes should be executed simultaneously to ensure immediate and appropriate interaction with dynamic environments.



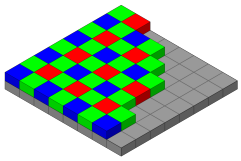
(a) LSM is task generic.



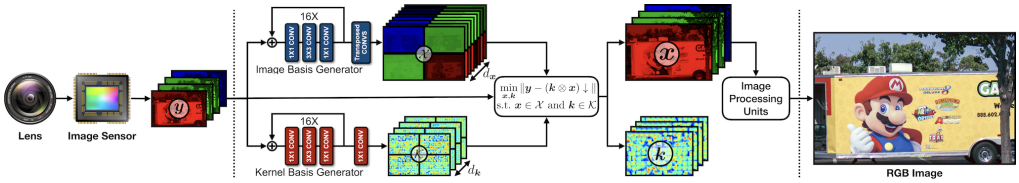
(b) A single minimization iteration of LSM.

In this research, I addressed these tasks in a unified framework. These tasks were generally formulated as a minimization problem of $\min_{\boldsymbol{x}} D(\boldsymbol{x}) + R(\boldsymbol{x})$, where $\boldsymbol{x}$ is the solution, $D(\boldsymbol{x})$ is the data term, and $R(\boldsymbol{x})$ is the regularization term. With stereo matching as the example, $\boldsymbol{x}$ is the disparities, $D(\boldsymbol{x})$ is the consistency between left and right views, and $R(\boldsymbol{x})$ enforce the spatial smoothness of $\boldsymbol{x}$, such as the $L1$ spatial smoothness. To preserve the first principle of a task explicitly within the network, $D(\boldsymbol{x})$ is preserved because it is well designed following the first principle of a task, while $R(\boldsymbol{x})$ is replaced with a learnable subspace constraint, because it is heuristic. The subspace generation incorporates both the single image features and the derivatives of $D(\boldsymbol{x})$, and it is the key to enable a single network for simultaneous multi-task perceptions with fully shared parameters. This Learning Subspace Minimization (LSM) framework not only facilitates instant multi-task perception of dynamic scenes but also enhances the extendability of spatial intelligent systems. For example, new tasks can be integrated in a plug-and-play manner as long as their data terms can be formulated into the objective functions. This research was presented at CVPR 2020 as an oral presentation [12].

- **Image Enhacement.** Spatial intelligent systems are supposed to operate in diverse environments, thus enhancing sensory inputs can substantially improve their perception capabilities. For instance, to enable a patrol robot to function continuously for 24 hours, it is beneficial to denoise and deblur input images during the night, before they are processed by subsequent perception tasks.



(a) Byaer Pattern of RAW images.



(b) Solving image enhancement problems on RAW images.

In this research, I revisited these image enhancement problems that share a unifed formulation of $\min_{\boldsymbol{x},\boldsymbol{k}} \|\boldsymbol{y} - (\boldsymbol{k} \otimes \boldsymbol{x} + \boldsymbol{n})\|$, where $\boldsymbol{y}$ is the input image, $\boldsymbol{x}$ is the enhanced image, $\boldsymbol{n}$ is

the noise, and $\boldsymbol{k}$ is the kernel map which involves degradation kernels for super-resolution or motion blur kernels for motion deblur. The kernel map $\boldsymbol{k}$ is spatially-variant and should be solved jointly with the enhanced output. To this end, I extended the LSM framework to address this dual-variable problem and initialized it with a closed-form solution. This work is also the first that operates solely on RAW images and has achieved significantly better results at a lower computational cost. The RAW images have a distinct Bayer Pattern from RGB images, and are directly exported from image sensors without any post-processing. Therefore, image enhancement on RAW images can serve as the retina of a spatial intelligent system and be connected with the subsequent perception tasks in an end-to-end pipeline. This research was published at CVPR 2022 [13].

Besides the above researches that have set a foundation for my pursuit of a physical world model for spatial intelligence. I also did two non-learning based researches that can also potentially contribute to this goal:

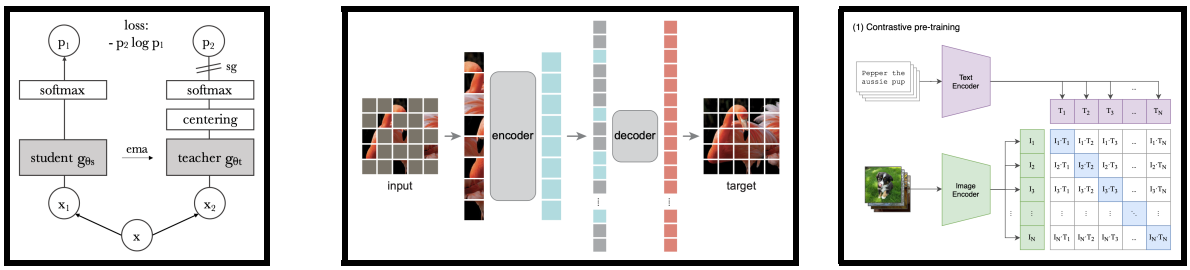
- **Robust Initialization for Monocular SLAM.** Most robotics systems incorporate various depth sensors for localization tasks, yet these sensors, including Direct-Time-of-Flight (dToF) sensors, have inherent challenges such as narrow fields of view and limited sensing ranges. In contrast, using a monocular camera for localization in 3D environments offers benefits like lower power usage and no need to calibrate and synchronize between sensors, which are advantageous particularly for compact robotics and wearable devices. To this end, I have developed an innovative rank-1 factorization-based algorithm for the initialization process within a monocular Simultaneous Localization and Mapping (SLAM) system [14]. This algorithm effectively and robustly converts a set of 2D trajectories into 3D point clouds with given rotations, improving the efficiency and robustness of the overall localization systems in Robotics or Mixed Reality applications.

- **Fish-eye and 360° Cameras.** Fish-eye and 360° cameras are common in Robotics and Mixed Reality. However, algorithms developed for perspective (pinhole) cameras often do not generalize well in these settings. Typically, fish-eye and 360° images are rectified into perspective images for further processing. Instead, I proposed a motion estimation technique tailored for 360-degree videos in [15]. This technique directly minimizes an objective function defined on a sphere, surpassing the performance of the commonly used five-point algorithm designed for perspective cameras. Additionally, I developed a spherical version of the As-Rigid-As-Possible warping technique. This work has provided me with profound insights into ultra-wide-angle cameras which can contribute future researches on Robotics and Ego-centric perception.

Research Plan

Building upon my previous researches in perception tasks, I am interested in pursuing a physical world model that transforms perceived information into a representation explicitly modeling physical rules and equations. This physical world model serves as an abstract medium for planning and control algorithms to interact with the environment. I also plan to continue working on perception topics such as self-supervised pre-training and segmentation. These areas are not only important as standalone tasks but also contributes to the exploration of the physical world model.

- **Self-supervised Pre-training.** Self-supervised pre-training is a milestone approach for computer vision, it leverages unlabeled data to learn feature representations. Existing methods, such as DINO [16] and Masked AutoEncoder [17], are primarily designed for high-level perception tasks like image and video classification. However, there remains an open question regarding the learning of visual feature representations in a self-supervised manner for low&mid-level perception tasks. To address this, we must consider what constitutes a good feature representation for these tasks. Drawing on my previous experience, one potential answer could lie in achieving equivariance [18] to both geometric and photometric transformations. This is crucial for spatial intelligent systems, because it needs to interact with the same environment or objects under varying spatial and illumination conditions.



(a) DINO (b) MAE (c) CLIP

Figure 5: Different self-supervised pre-training strategies: (a) DINO (Self-Distillation), (b) MAE (Masked Image Model) and (c) CLIP (Contrastive Learning). It remains an open question about learning equivariant features for low&mid-level perceptions from pre-training.

- **Segmentation.** Another direction I plan to pursue is a deeper dive into segmentation. Rather than relying on annotations in the first frame as in video object segmentation, which I worked on before, I aim to achieve segmentation through alternative cues, such as motion movement and geometric structures. Although recent research primarily focuses on semantic segmentation or instance segmentation, grouping points into objects without semantic labels can already serve as an intermediate representation for spatial intelligent systems. For instance, recognizing the category of an object is not always necessary for robotic arm manipulation. A potential solution is to reformulate conventional segmentation algorithms, such as the normalized cut [19], as a differentiable component within a neural network. While a previous work has incorporated Normalized cut into the training loss [20], to the best of my knowledge, integrating it into the model remains unexplored.

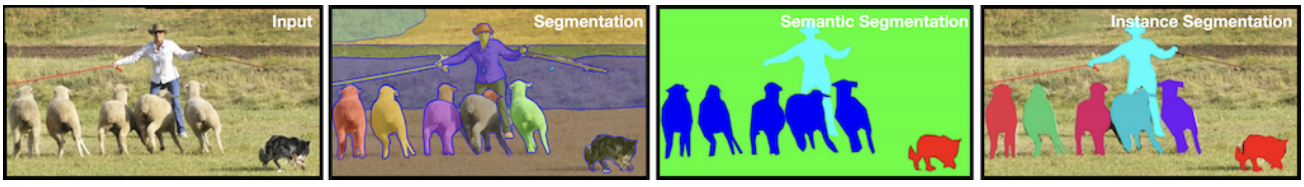


Figure 6: The difference between a) segmentation: grouping pixels into segments, b) semantic segmentation: classifying each pixel to a semantic category, and c) instance segmentation: classifying each pixel to a category and differentiating between instances of the same category.

- **Physical World Model Learning for Robotics Control.** Beyond specific perception tasks, I am committed to the goal of enabling spatial intelligent systems to understand the physical world and its underlying principles. Once equipped with this capability, a spatial intelligent system can interact more efficiently with the environment using both differentiable optimization-based controllers like Motion Predictive Control (MPC) [21] and the more recent RL algorithms [22, 23, 24]. To progress towards this goal step-by-step, **I have devised a three-phase plan with a series of checkpoints and milestones.**

  – In the first phase, I will learn and verify a physical world model on the task of future frame prediction [25], a popular auxiliary task in robotics. Given the recent release of the Open-X-Embodiment dataset [26], being able to predict future frames on diverse sequences could prove the physical world model's effectiveness without involving actual planning and control.

Figure 7: Future frame prediction and the Open-X-Embodiment dataset.

  – Once the physical world model has been validated, the next phase will integrate it with controllers inside a simulator. Recent advancements in simulators, such as Isaac Sim [27] and Habitat [28], have narrowed the gap between synthetic and real-world environments. Therefore, training and testing the physical world model jointly with a controller in a simulator could scale up the experiments and accelerate the research.

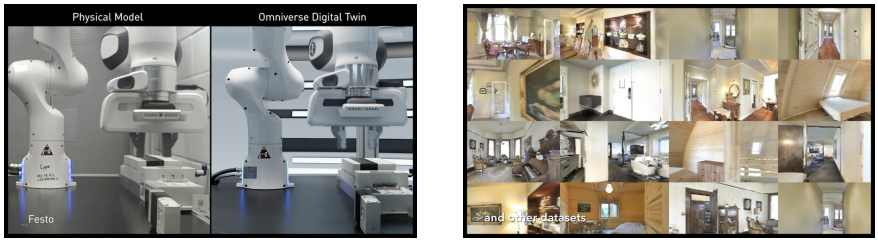


Figure 8: Realistic Simulation from Issac Sim (left) and Habitat(right).

  – Finally, I will transfer knowledge and models learned in the simulator to real-world scenarios for tasks including but not limited to robotic arm manipulation and indoor navigation. To compensate the potential reality gap between the simulator and the real world, I will explore efficient online adaptation algorithms [29, 30]. Besides, imitation learning [31, 32, 33] that replicates human-like actions from demonstration could also be explored for specific tasks such as grasping.

This three-phase plan establishes clear checkpoints and milestones, allowing for review and refinement at each stage, advancing towards the goal in an efficient and solid manner.

Collaborations. I plan to establish broader collaborations with both industrial and academic researchers that align with my research goals. To initiate this, I will leverage my existing network, encompassing professionals from big corporations like Apple, Meta, Google, Microsoft, and NVIDIA, as well as respected academic institutions including SFU, UIUC, HKUST, CUHK, PKU and the University of Birmingham. Specifically, I aim to foster partnerships that will enable the sharing of resources, data, and expertise, contributing to advancements in learning physical world model for spatial intelligence. I am also looking forward to expanding my collaboration networks through conferences, talks, as well as the existing faculty members within the department. Through these collaborative efforts, I can create a synergy that not only elevates my research but also contributes significantly to the research community.

Founding Sources. I plan to apply for NSERC Discovery Grants in my first year based on my previous researches and the research plan. I also plan to seek fundings from industry sources, such as Google Research Scholar Program and Meta Research Awards.

References

[1] Hugo Touvron et al. Llama 2: Open foundation and fine-tuned chat models, 2023. arXiv:2307.09288.

[2] Billy Perrigo. Exclusive: OpenAI used kenyan workers on less than $2 per hour to make ChatGPT less toxic, 2023. The Time Magazine.

[3] Michael Bloesch, Jan Czarnowski, Ronald Clark, Stefan Leutenegger, and Andrew J. Davison. Codeslam — learning a compact, optimisable representation for dense visual slam. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[4] Benjamin Ummenhofer, Huizhong Zhou, Jonas Uhrig, Nikolaus Mayer, Eddy Ilg, Alexey Dosovitskiy, and Thomas Brox. Demon: Depth and motion network for learning monocular stereo. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5622–5631, 2017.

[5] Bill Triggs, Philip F. McLauchlan, Richard I. Hartley, and Andrew W. Fitzgibbon. Bundle adjustment - a modern synthesis. In *Proceedings of the International Workshop on Vision Algorithms: Theory and Practice*, ICCV '99, 2000.

[6] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, second edition, 2004.

[7] Chengzhou Tang and Ping Tan. BA-Net: Dense bundle adjustment networks. In *International Conference on Learning Representations (ICLR)*, 2019.

[8] Paul-Edouard Sarlin et al. Back to the Feature: Learning robust camera localization from pixels to pose. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[9] Chang Shu, Kun Yu, Zhixiang Duan, and Kuiyuan Yang. Feature-metric loss for self-supervised learning of depth and egomotion. In *European Conference on Computer Vision (ECCV)*, 2020.

[10] Lukas von Stumberg, Patrick Wenzel, Qadeer Khan, and Daniel Cremers. GN-Net: The gauss-newton loss for multi-weather relocalization. *IEEE Robotics and Automation Letters (RA-L)*, 2020.

[11] Shun Iwase, Xingyu Liu, Rawal Khirodkar, Rio Yokota, and Kris M. Kitani. RePOSE: Fast 6d object pose refinement via deep texture rendering. In *International Conference on Computer Vision (ICCV)*, 2021.

[12] Chengzhou Tang, Lu Yuan, and Ping Tan. LSM: Learning subspace minimization for low-level vision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[13] Chengzhou Tang, Yuqiang Yang, Bing Zeng, Ping Tan, and Shuaicheng Liu. Learning to zoom inside camera imaging pipeline. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[14] Chengzhou Tang, Oliver Wang, and Ping Tan. GSLAM: Initialization-robust monocular visual SLAM via global structure-from-motion. In *International Conference on 3D Vision (3DV)*, 2017.

[15] Chengzhou Tang, Oliver Wang, Feng Liu, and Ping Tan. Joint stabilization and direction of 360° videos. *ACM Transactions on Graphics (TOG)*, 2019.

[16] Maxime Oquab et al. DINOv2: Learning robust visual features without supervision, 2023. arXiv:2304.07193.

[17] Kaiming He et al. Masked autoencoders are scalable vision learners. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[18] Karel Lenc and Andrea Vedaldi. Understanding image representations by measuring their equivariance and equivalence. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[19] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Trans. on Pattern Analysis and Machine Intelligence (T-PAMI)*, 2000.

[20] Meng Tang et al. Normalized cut loss for weakly-supervised CNN segmentation, 2018. arXiv:1804.01346.

[21] Brandon Amos et al. Differentiable mpc for end-to-end planning and control. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[22] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic algorithms and applications. *arXiv preprint arXiv:1812.05905*, 2019.

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

[24] Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. *arXiv preprint arXiv:1602.01783*, 2016.

[25] Mohammad Babaeizadeh, Chelsea Finn, Dumitru Erhan, Roy H. Campbell, and Sergey Levine. Stochastic variational video prediction. *CoRR*, abs/1710.11252, 2017.

[26] Open X-Embodiment Collaboration et al. Open X-Embodiment: Robotic learning datasets and RT-X models, 2023. arXiv:2310.08864.

[27] NVIDIA Corporation. What is Isaac Sim?, 2023. Online source.

[28] Manolis Savva et al. Habitat: A platform for embodied AI research. In *International Conferences on Computer Vision (ICCV)*, 2019.

[29] Yuqing Du, Olivia Watkins, Trevor Darrell, Pieter Abbeel, and Deepak Pathak. Auto-tuned sim-to-real transfer. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1290–1296, 2021.

[30] Ashish Kumar, Zipeng Fu, Deepak Pathak, and Jitendra Malik. RMA: Rapid motor adaptation for legged robots. *Robotics: Science and Systems*, 2021.

[31] Brenna D Argall, Sonia Chernova, Manuela Veloso, and Brett Browning. A survey of robot learning from demonstration. *Robotics and Autonomous Systems*, 57(5):469–483, 2009.

[32] Tianhao Zhang, Zoe McCarthy, Owen Jow, Dennis Lee, Xi Chen, Ken Goldberg, and Pieter Abbeel. Deep imitation learning for complex manipulation tasks from virtual reality teleoperation. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1–8. IEEE, 2018.

[33] Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. *The Journal of Machine Learning Research*, 17(1):1334–1373, 2016.