Self-Supervised Multisensory Pretraining for Contact-Rich Robot Reinforcement Learning

Rickmer Krohn[1,2,3], Vignesh Prasad[1,2,3], Gabriele Tiboni[1] and Georgia Chalvatzaki[1,2,3]

***Abstract*— Effective contact-rich manipulation requires robots to synergistically leverage vision, force, and proprioception. However, Reinforcement Learning agents struggle to learn in such multisensory settings, especially amidst sensory noise and dynamic changes. We propose MultiSensory Dynamic Pretraining (MSDP), a novel framework for learning expressive multisensory representations tailored for task-oriented policy learning. MSDP is based on masked autoencoding and trains a transformer-based encoder by reconstructing multisensory observations from only a subset of sensor embeddings, leading to cross-modal prediction and sensor fusion. For downstream policy learning, we introduce a novel asymmetric architecture, where a cross-attention mechanism allows the critic to extract dynamic, task-specific features from the frozen embeddings, while the actor receives a stable pooled representation to guide its actions. Our method demonstrates accelerated learning and robust performance under diverse perturbations, including sensor noise, and changes in object dynamics. Evaluations in multiple challenging, contact-rich robot manipulation tasks in simulation and the real world showcase the effectiveness of MSDP. Our approach exhibits strong robustness to perturbations and achieves high success rates on the real robot with as few as 6,000 online interactions, offering a simple yet powerful solution for complex multisensory robotic control.**

I. Introduction

Reinforcement Learning (RL) has shown impressive successes in learning complex tasks ranging from Atari [1], locomotion [2], vision-based manipulation [3] to multisensory peg insertion [4], [5]. However, incorporating multiple sensor modalities, especially in complex contact-rich robotic manipulation tasks, remains a challenge for RL, due to the heterogeneous dynamics of different sensor modalities. Additionally, the importance of each input modality changes during the execution of a manipulation task, e.g., coarse scene understanding from visual input to fine-grained force feedback when in contact. Thus, robotic agents need to learn how to dynamically focus on the most relevant sensory information while adapting to perturbations and dynamic changes in the environment. This field of *sensor fusion* is a long-studied problem in the field of robotics for various tasks ranging from control [6], [7], manipulation [8], [9], localization, and navigation [10], but remains underexplored in RL settings.

To this end, Imitation Learning approaches [11], [12], [13] have shown promise in utilizing multisensory data for learning skills, but require experts collecting informative data, thus limiting their usability, robustness and generalization. Self-directed exploration in RL, on the other hand, facilitates the development of adaptable strategies that can be generalized across diverse object properties and contexts. These advantages position RL as a compelling approach for learning multisensory contact-rich manipulation [4], [14], [15]. Recent advances in multimodal learning [16], [17] have shown the advantages of reconstructing masked or noised inputs to learn expressive cross-modal representations for downstream tasks. Such masking-based self-supervision approaches, whether at the sensory or embedding level, have also been shown to improve the network's robustness [18], [19], [20].

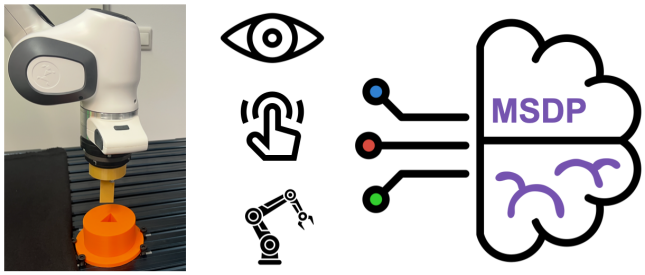


Fig. 1: Multisensory Dynamic Pretraining fuses multiple sensors, like human senses, to solve complex contact-rich manipulation tasks.

In this paper, we propose ***M**ulti**S**ensory **D**ynamic **P**retraining* (MSDP), a novel RL framework to learn expressive multisensory representations for contact-rich manipulation tasks. MSDP first learns to extract and fuse sensor features via an offline, pretraining phase based on masked autoencoding and cross-sensor prediction; then, an online RL agent leverages the features from the pretrained multisensory encoder through a new combination of cross-attention and pooling mechanisms applied to the critic and actor, respectively. In turn, this design enables RL policies to synergistically fuse multisensory inputs and inherently cope with sensor noise or missing modalities.

Our experimental results demonstrate that MSDP yields effective representations that accelerate RL on a variety of contact-rich manipulation tasks while being robust to sensor noise and changing object dynamics. The FT-sensor boosts MSDP's performance in two challenging real robot tasks by 14 % leading to near-optimal performance. Notably, the policy is trained directly in the real world on MSDP's multisensory latent representation, without any sim-to-real transfer. We achieve near-optimal performance in only 6,000 online interactions, which takes less than 55 minutes including data collection and pretraining. Beyond these results, MSDP

Corresponding author: Rickmer Krohn (rickmer.krohn@tu-darmstadt.de).
[1] Interactive Robot Perception & Learning (PEARL) Lab, TU Darmstadt, Germany, [2] Hessian.AI; [3] Robotics Institute Germany (RIG); This research is funded by the German Research Foundation (DFG) Emmy Noether Programme (CH 2676/1-1), the EU's Horizon Europe project "ARISE" (Grant no.: 101135959), the German Federal Ministry of Education and Research (BMBF) project "RiG" (Grant no.: 16ME1001) and the European Research Council (ERC) project "SIREN" (Grant No.: 101163933).

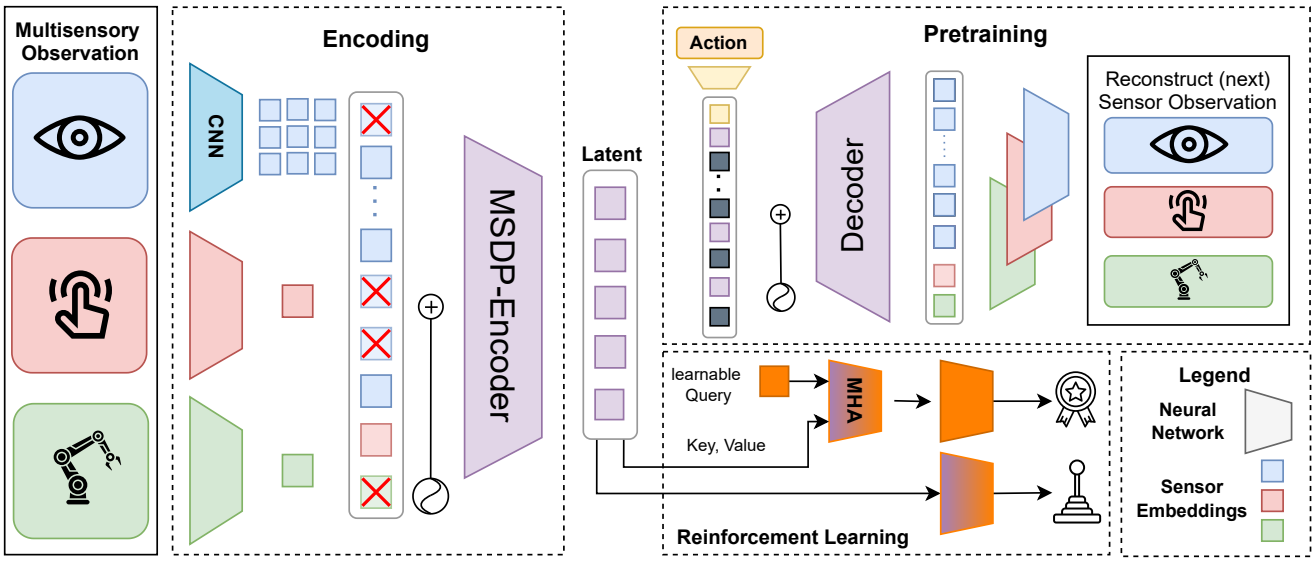


Fig. 2: The MSDP framework with MSDP-Encoder (left), Pretraining (top right) and downstream RL (bottom right): The current multisensory observation gets projected with a CNN-stem and linear layers to the embedding space. The MSDP-encoder fuses all sensor embeddings to form our expressive multisensory latent representation. The encoder is trained via (next) sensor observation reconstruction from a subset of sensor embeddings resulting in dynamic cross-sensor prediction. For downstream RL we extract multisensory task-specific features via a single cross-attention layer for the critic and via pooling for the actor. Our Framework offers an expressive and robust multisensory representation for complex contact-rich manipulation tasks in simulation and the real world.

scales naturally to an increasing number of diverse input modalities and only introduces a few learnable parameters in downstream training of RL agents.

To summarize, the key contributions of our work are twofold: (i) we develop an effective pretraining strategy based on masked autoencoding to form a rich multisensory representation, and (ii) we introduce a novel architecture that allows task-specific feature extraction to achieve efficient robot Reinforcement Learning. Our findings pave the way for adaptive and resilient RL agents to handle multiple input modalities to master complex contact-rich manipulation tasks.

II. Related Work

Tasks where a robotic manipulator has to interact with its environment, either directly or indirectly via a tool, require a good understanding of the interaction forces that shape the task at hand. Solving contact-rich tasks often relies on an accurate estimate of force and dynamics. When such estimates are available, classical control approaches can be adapted to solve the task [21], [22]. To better handle unseen or difficult contact dynamics in unstructured environments, RL presents itself as an ideal candidate that can learn directly via interactions [23], without accurate state estimation [24]. Various works have employed RL to learn contact-rich manipulation policies on a variety of tasks [25], [26], [27].

Lee et al. [4] learns a multimodal representation from vision, force torque, and proprioception via MLP-fusion and multiple self-supervised objectives. The frozen representation leads to a robust representation for RL to solve multiple peg insertion tasks. [14] extended the Vision Transformer [28] with a force torque sensor to solve a variety of contact-rich tasks. They additionally incorporate the self-supervised objectives from [4] to shape a representation using SLAC [29]. [30] pretrains an audio-encoder [31] to combine vision and audio via a transformer decoder for Imitation Learning. The audio signal from the contact microphone provides rich feedback for various manipulation tasks. To also account for different sensor frequencies [13] developed a multi-resolution policy based on pretrained Vision Language Models to improve inference time using proprioception and force.

A robust representation is essential to integrate and process various sensory inputs to ensure stable and efficient learning in multisensory RL. Previous architectures often focused on straightforward latent fusion approaches by concatenating the various representations for downstream tasks [32], [33], [34], [35], [36]. Feng et al. [37] take this a step further by adding a subgoal-aware weighting for learning the stage-wise importance of difference sensors. Alternatively, contrastive learning approaches [38], [39] emphasize feature alignment, rather than fusion, to learn a shared latent representation between multiple modalities. However, distilling task-relevant features while maintaining sensor-specific information can be challenging. In contrast, inspired by the success of masked token prediction [16], recent works have also explored how masked multisensory pretraining can enhance representation learning for contact-rich manipulation [5], [15].

III. Multisensory Dynamic Pretraining

We present **M**ulti**S**ensory **D**ynamic **P**retraining (MSDP), a novel framework for representation learning that builds upon Masked Autoencoders [40], tailored to enhance contact-rich robot reinforcement learning tasks that require perception through multiple sensor modalities. MSDP introduces a modular architecture to seamlessly handle diverse input

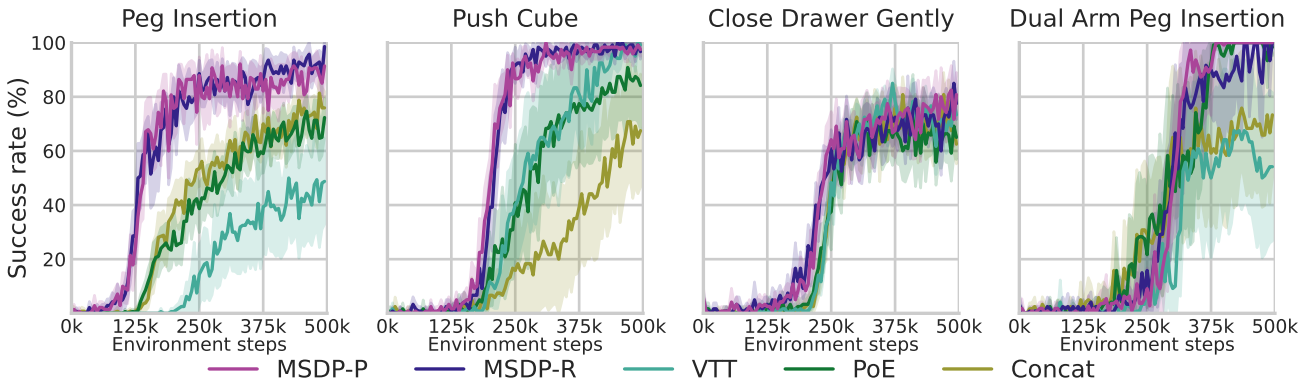


Fig. 3: Performance comparison between MSDP-P and MSDP-R to the baselines in Peg Insertion, Push Cube, Close Drawer Gently and Dual Arm Peg Insertion. Our method significantly accelerates RL training and achieves the highest final success rate across all tasks.

sensors, and a multisensory masking scheme to promote rich cross-modal representations, i.e., to retain knowledge about the task and the environment dynamics even in the absence or perturbation of one or more modalities. Our framework decouples representation learning and downstream RL to ensure rich feature extraction and fusion, while offering stable, robust and compact downstream representations. Notably, our architecture is scalable w.r.t. the number and types of input modalities and also supports pretraining using additional sensors, which may not be available during policy training. As a result, our pretraining phase on limited offline data leads to rich representations that may be directly used to solve complex contact-rich RL tasks from high-dimensional input (including images) in less than $500k$ environment steps, whereas in the real world only $6k$ online samples are needed.

A. MSDP Architecture

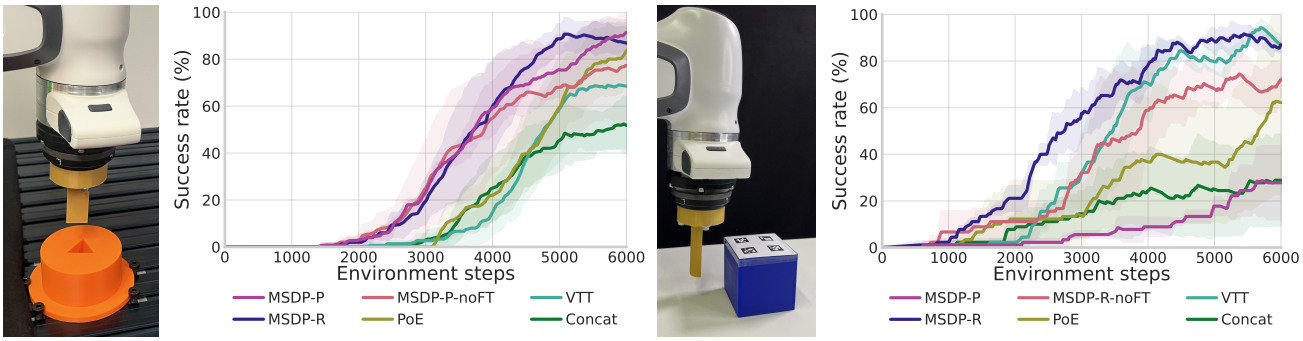
Given the inputs from multiple sensor modalities, such as vision, force torque readings and proprioception, we encode each modality using a separate network referred to as *sensor encoders*. Particularly for the vision input, we use a 4 layer CNN-stem similar to [41], which allows us to mask at embedding- rather than pixel-level, as masked object positions cannot be retrieved in pixel-space. The CNN-stem introduces redundancy as vision embeddings have overlapping receptive fields and, most importantly, stabilizes training [42]. We use a linear projection to encode force torque readings and proprioception to the 128-dim embedding space given their low dimensionalities. Subsequently, we add learnable parameters to each sensor embedding to contextualize the position and modality of it. Once we obtain embedding representations from all sensors, we randomly mask out a subset (70 %) of sensor tokens and feed the remaining to our 2 layer transformer encoder with 4 attention heads, pre-normalization and a mlp ratio of 2. The attention mechanism of the encoder leads to dynamic multisensory fusion promoted by the representation objective described in Section III-B. A learnable mask embedding is added to the randomly masked out embedding positions, next to the multisensory embeddings from the encoder, to generate the original number of embeddings. Separate decoder heads reconstruct the observation following the representation learning objective. We use a shared linear projection for all vision embeddings to reconstruct their corresponding patch.

B. Representation Learning

Our representation learning objective is based on multi-modal masked autoencoding [40], [16], [17]. The objective is to either reconstruct the current $\mathbf{O}_t^{MS}$ or next observation $\mathbf{O}_{t+1}^{MS}$ from a random subset of sensor embeddings. Vision has a high number of embeddings and needs to extract information about other, potentially masked, sensors e.g. identifying contact to estimate force. Non-vision sensors on the other hand, are beneficial to reconstruct the vision observation as e.g. proprioception defines the robot position. Vision as a global sensing modality never gets fully masked out, whereas low-dimensional sensors are not available when masked. This representation objective results in cross-sensor prediction thus leading to fusion of all modalities.

C. Policy Learning

From the pretrained encoder, we receive expressive multi-sensory embeddings given the available sensors. We name the mapping between embeddings and a compact representation *latent bridging*, which has a considerable impact on performance. To address this, we depart from works that naively extract the "CLS"-token from the high-dimensional embeddings [20], [28], [43], [44] and, instead, propose an asymmetric *latent bridging* strategy between actor and critic. The critic uses a single cross-attention layer with a learnable query and the multisensory embeddings from the MSDP encoder as keys and values. It offers dynamic task-specific feature extraction over the task-solving process. Fine-grained understanding of the environment leads to faster convergence compared to a global representation. The policy, on the other hand, does not profit from a cross-attention layer as it may destabilize training. Instead, we mean pool all embeddings originating from the vision sensor before pooling all embeddings, to account for the uneven number of sensor embeddings. Pooling sensor tokens results in a stable and parameter-free latent bridging similar to [5], [41]. This assymmetric representation between actor and critic, follows



(a) Real World Peg Insertion.

(b) Real World Push Cube.

Fig. 4: Real world setup and experimental results. Our MSDP framework enables training RL policies directly in the real world, outperforming various baselines. Force torque readings are essential to consistently push the cube to the goal or to insert the peg under occlusion improving task success by 14 % (cf. MSDP-noFT). Policies are learned directly on the pretrained multisensory representation, without any sim-to-real transfer, with only 6,000 online interactions.

[45], where the actor benefits from a stable representation over task stages and the critic from a dynamic-specific representation. We train with the off-policy algorithm SAC in simulation and RLPD in the real world, where we incorporate offline data for pretraining in our replay buffer.

IV. Experiments and Results

In this section, we present our competitive baselines, simulation environments and experiments as well as our real world results.

A. Baselines

We compare our methods against one transformer and two non transformer-based baselines. The latter models extract sensor-specific features with a CNN for vision and an MLP for proprioception and force torque readings. The **Concat** model fuses the concatenated features with a 2-layer MLP to form the multisensory latent representation [4]. The **PoE** model generates separate means and variances for each sensor and fuses them with a product of experts approach. **VTT** [14] fuses all sensors via a transformer-encoder and compresses the features via multiple linear layers to a compact latent.

B. Multisensory Simulation Environments and Experiments

Our multisensory environments are based on panda-gym [46] and Pybullet [47]. The tasks Peg Insertion, Push Cube, Close Drawer Gently and Dual Arm Peg Insertion are challenging contact-rich manipulation tasks with sensor noise, varying dynamics and number of sensors. Each task has a dense reward function and comes with multisensory observations consisting of proprioception, vision and force torque readings. The observation of the Dual Arm environment has an additional FT-sensor and proprioception of the second robot arm. The action space consists of the 3-dimensional endeffector displacement.

Figure 3 shows the performance of our MSDP-P and MSDP-R models compared to the baselines. We obtain superior performance in Peg Insertion and Push Cube, while also outperforming in Close Drawer Gently and Dual Arm Peg Insertion, indicating an expressive representation for contact-rich tasks. Especially in Peg Insertion, our sensor fusion provides fine-grained features to solve the task around 80 % of the time after only $200,000$ environment steps. The baselines struggle to combine sensor knowledge to consistently insert the peg. Baseline models need an extended amount of training in Push Cube to be able to account for varying object dynamics, like mass and center of mass of the cube, while our representations allow quick adaptation to those changes, leading to a steep learning curve and optimal performance.

C. Real World Experiments

We conduct the Peg Insertion and Push Cube tasks in the real world using a Franka robot arm with a wrist-mounted FT-sensor, 3rd person camera and a custom endeffector. The starting position of the robot is randomized and we use a sparse reward of +1 upon success in both tasks. We use a cartesian impedance controller for safe interaction and pretrain the multisensory encoder with data from 20 demonstrations and ~ 2,000 samples of play data. Downstream tasks are trained directly on the real system for only 6,000 environment interactions with RLPD [48]. The complete training pipeline, including data collection, pretraining, and online RL, learns the final policy in less than 55 minutes for both tasks.

Figures 4a and 4b show setup and results for Peg Insertion (5 runs) and Push Cube (3 runs), respectively. MSDP achieves superior performance in Peg Insertion, whereas baselines suffer from vision noise and occlusion. The Force Torque sensor enables MSDP to make and maintain contact with the cube. The prediction objective is not able to extract a suitable representation due to the complexity of learning forward dynamics in a vision-demanding task on limited real world data. The usage of the FT-sensor improves the performance of MSDP by 14 % in both real-world tasks, underscoring the significance of multi-sensor integration in contact-rich manipulation.

V. Conclusion

In this work, we propose MultiSensory Dynamic Pretraining (MSDP), a novel pretraining framework to learn multisensory representations for contact-rich manipulation tasks. Specifically, MSDP learns to reconstruct varying

sensory information from a subset of input sensor embeddings, leading to cross-sensor predictions and sensor fusion. MSDP captures the interplay between different sensor modalities to learn a rich multisensory representation, which in combination with task-specific feature extraction through cross-attention, leads to a superior performance in challenging contact-rich manipulation tasks in simulation and the real world.

REFERENCES

[1] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing Atari with Deep Reinforcement Learning," 2013, arXiv:1312.5602 [cs]. [Online]. Available: http://arxiv.org/abs/1312.5602

[2] Z. Zhuang, Z. Fu, J. Wang, C. Atkeson, S. Schwertfeger, C. Finn, and H. Zhao, "Robot Parkour Learning," 2023, arXiv:2309.05665 [cs]. [Online]. Available: http://arxiv.org/abs/2309.05665

[3] F. Zhang, J. Leitner, M. Milford, B. Upcroft, and P. Corke, "Towards Vision-Based Deep Reinforcement Learning for Robotic Motion Control," 2015, arXiv:1511.03791 [cs]. [Online]. Available: http://arxiv.org/abs/1511.03791

[4] M. A. Lee, Y. Zhu, K. Srinivasan, P. Shah, S. Savarese, L. Fei-Fei, A. Garg, and J. Bohg, "Making Sense of Vision and Touch: Self-Supervised Learning of Multimodal Representations for Contact-Rich Tasks," *CoRR*, vol. abs/1810.10191, 2018, _eprint: 1810.10191. [Online]. Available: http://arxiv.org/abs/1810.10191

[5] C. Sferrazza, Y. Seo, H. Liu, Y. Lee, and P. Abbeel, "The Power of the Senses: Generalizable Manipulation from Vision and Touch through Masked Multimodal Learning," 2023, arXiv:2311.00924 [cs]. [Online]. Available: http://arxiv.org/abs/2311.00924

[6] J. G. García, A. Robertsson, J. G. Ortega, and R. Johansson, "Sensor fusion for compliant robot motion control," *IEEE Transactions on Robotics*, vol. 24, no. 2, pp. 430–441, 2008.

[7] F. F. Khalil and P. Payeur, "Dexterous robotic manipulation of deformable objects with multi-sensory feedback-a review," *Robot Manipulators Trends and Development*, no. March 2010, 2010.

[8] Y. Hu, Z. Li, G. Li, P. Yuan, C. Yang, and R. Song, "Development of sensory-motor fusion-based manipulation and grasping control for a robotic hand-eye system," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 47, no. 7, pp. 1169–1180, 2016.

[9] Z. Xia, Z. Deng, B. Fang, Y. Yang, and F. Sun, "A review on sensory perception for dexterous robotic manipulation," *International Journal of Advanced Robotic Systems*, 2022.

[10] M. B. Alatise and G. P. Hancke, "A review on challenges of autonomous mobile robot and sensor fusion methods," *IEEE Access*, vol. 8, pp. 39 830–39 846, 2020.

[11] H. Li, Y. Zhang, J. Zhu, S. Wang, M. A. Lee, H. Xu, E. Adelson, L. Fei-Fei, R. Gao, and J. Wu, "See, Hear, and Feel: Smart Sensory Fusion for Robotic Manipulation," 2022, arXiv:2212.03858 [cs]. [Online]. Available: http://arxiv.org/abs/2212.03858

[12] Y. Hao, R. Wang, Z. Cao, Z. Wang, Y. Cui, and D. Sadigh, "Masked Imitation Learning: Discovering Environment-Invariant Modalities in Multimodal Demonstrations," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2023, pp. 1–7, iSSN: 2153-0866. [Online]. Available: https://ieeexplore.ieee.org/document/10341728/?arnumber=10341728

[13] S. Saxena, M. Sharma, and O. Kroemer, "MResT: Multi-Resolution Sensing for Real-Time Control with Vision-Language Models," 2024, arXiv:2401.14502 [cs]. [Online]. Available: http://arxiv.org/abs/2401.14502

[14] Y. Chen, A. Sipos, M. Van der Merwe, and N. Fazeli, "Visuo-Tactile Transformers for Manipulation," 2022, arXiv:2210.00121 [cs]. [Online]. Available: http://arxiv.org/abs/2210.00121April3rd

[15] Q. Liu, Z. Sun, Y. Cui, L. Gaofeng, Q. Ye, and J. Chen, *Masked Visual-Tactile Pre-training for Robot Manipulation*, 2024.

[16] R. Bachmann, D. Mizrahi, A. Atanov, and A. Zamir, "MultiMAE: Multi-modal Multi-task Masked Autoencoders," 2022. [Online]. Available: https://arxiv.org/abs/2204.01678v1

[17] X. Geng, H. Liu, L. Lee, D. Schuurmans, S. Levine, and P. Abbeel, "Multimodal Masked Autoencoders Learn Transferable Representations," 2022. [Online]. Available: https://arxiv.org/abs/2205.14204v3

[18] S. Skand, B. Pandit, C. Kim, L. Fuxin, and S. Lee, "Simple Masked Training Strategies Yield Control Policies That Are Robust to Sensor Failure," 2024. [Online]. Available: https://openreview.net/forum?id=AsbyZRdqPv

[19] G.-H. Liu, A. Siravuru, S. Prabhakar, M. Veloso, and G. Kantor, "Learning End-to-end Multimodal Sensor Policies for Autonomous Navigation," 2017.

[20] I. Radosavovic, T. Xiao, S. James, P. Abbeel, J. Malik, and T. Darrell, "Real-World Robot Learning with Masked Visual Pre-training," 2022, arXiv:2210.03109 [cs]. [Online]. Available: http://arxiv.org/abs/2210.03109

[21] D. E. Whitney, "Historical perspective and state of the art in robot force control," *The International Journal of Robotics Research*, vol. 6, no. 1, pp. 3–14, 1987.

[22] D. E. Whitney *et al.*, "Quasi-static assembly of compliantly supported rigid parts," *Journal of Dynamic Systems, Measurement, and Control*, vol. 104, no. 1, pp. 65–77, 1982.

[23] M. Nuttin and H. Van Brussel, "Learning the peg-into-hole assembly operation with a connectionist reinforcement technique," *Computers in Industry*, vol. 33, no. 1, pp. 101–109, 1997.

[24] S. Levine, C. Finn, T. Darrell, and P. Abbeel, "End-to-end training of deep visuomotor policies," *Journal of Machine Learning Research*, vol. 17, no. 39, pp. 1–40, 2016.

[25] Í. Elguea-Aguinaco, A. Serrano-Muñoz, D. Chrysostomou, I. Inziarte-Hidalgo, S. Bøgh, and N. Arana-Arexolaleiba, "A review on reinforcement learning for contact-rich robotic manipulation tasks," *Robotics and Computer-Integrated Manufacturing*, vol. 81, p. 102517, 2023.

[26] M. Suomalainen, Y. Karayiannidis, and V. Kyrki, "A survey of robot manipulation in contact," *Robotics and Autonomous Systems*, vol. 156, p. 104224, 2022.

[27] R. Liu, F. Nageotte, P. Zanne, M. de Mathelin, and B. Dresp-Langley, "Deep reinforcement learning for the control of robotic manipulation: a focussed mini-review," *Robotics*, vol. 10, no. 1, p. 22, 2021.

[28] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *CoRR*, vol. abs/2010.11929, 2020, _eprint: 2010.11929. [Online]. Available: https://arxiv.org/abs/2010.11929

[29] A. X. Lee, A. Nagabandi, P. Abbeel, and S. Levine, "Stochastic Latent Actor-Critic: Deep Reinforcement Learning with a Latent Variable Model," 2020, arXiv:1907.00953 [cs]. [Online]. Available: http://arxiv.org/abs/1907.00953

[30] J. Mejia, V. Dean, T. Hellebrekers, and A. Gupta, "Hearing Touch: Audio-Visual Pretraining for Contact-Rich Manipulation," 2024, arXiv:2405.08576 [cs]. [Online]. Available: http://arxiv.org/abs/2405.08576

[31] P. Morgado, N. Vasconcelos, and I. Misra, "Audio-Visual Instance Discrimination with Cross-Modal Agreement," 2021, arXiv:2004.12943 [cs]. [Online]. Available: http://arxiv.org/abs/2004.12943

[32] M. A. Lee, Y. Zhu, P. Zachares, M. Tan, K. Srinivasan, S. Savarese, L. Fei-Fei, A. Garg, and J. Bohg, "Making sense of vision and touch: Learning multimodal representations for contact-rich tasks," *IEEE Transactions on Robotics*, vol. 36, no. 3, pp. 582–596, 2020.

[33] H. Li, Y. Zhang, J. Zhu, S. Wang, M. A. Lee, H. Xu, E. Adelson, L. Fei-Fei, R. Gao, and J. Wu, "See, hear, and feel: Smart sensory fusion for robotic manipulation," in *Conference on Robot Learning*. PMLR, 2023, pp. 1368–1378.

[34] A. Li, R. Liu, X. Yang, and Y. Lou, "Reinforcement learning strategy based on multimodal representations for high-precision assembly tasks," in *Intelligent Robotics and Applications: 14th International Conference, ICIRA 2021, Yantai, China, October 22–25, 2021, Proceedings, Part I 14*. Springer, 2021, pp. 56–66.

[35] I. Guzey, B. Evans, S. Chintala, and L. Pinto, "Dexterity from touch: Self-supervised pre-training of tactile representations with robotic play," *arXiv preprint arXiv:2303.12076*, 2023.

[36] Y. Han, K. Yu, R. Batra, N. Boyd, C. Mehta, T. Zhao, Y. She, S. Hutchinson, and Y. Zhao, "Learning generalizable vision-tactile robotic grasping strategy for deformable objects via transformer," *IEEE/ASME Transactions on Mechatronics*, 2024.

[37] R. Feng, D. Hu, W. Ma, and X. Li, "Play to the score: Stage-guided dynamic multi-sensory fusion for robotic manipulation," in *8th Annual Conference on Robot Learning*, 2024.

[38] V. Dave, F. Lygerakis, and E. Rueckert, "Multimodal Visual-Tactile Representation Learning through Self-Supervised Contrastive

Pre-Training," 2024, arXiv:2401.12024 [cs]. [Online]. Available: http://arxiv.org/abs/2401.12024

[39] F. Lygerakis, V. Dave, and E. Rueckert, "M2CURL: Sample-Efficient Multimodal Reinforcement Learning via Self-Supervised Representation Learning for Robotic Manipulation," 2024, arXiv:2401.17032 [cs]. [Online]. Available: http://arxiv.org/abs/2401.17032

[40] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked Autoencoders Are Scalable Vision Learners," 2021, arXiv:2111.06377 [cs]. [Online]. Available: http://arxiv.org/abs/2111.06377

[41] Y. Seo, D. Hafner, H. Liu, F. Liu, S. James, K. Lee, and P. Abbeel, "Masked World Models for Visual Control," 2023, arXiv:2206.14244 [cs]. [Online]. Available: http://arxiv.org/abs/2206.14244

[42] T. Xiao, M. Singh, E. Mintun, T. Darrell, P. Dollár, and R. Girshick, "Early Convolutions Help Transformers See Better," 2021, arXiv:2106.14881 [cs]. [Online]. Available: http://arxiv.org/abs/2106.14881

[43] T. Xiao, I. Radosavovic, T. Darrell, and J. Malik, "Masked Visual Pre-training for Motor Control," 2022, arXiv:2203.06173. [Online]. Available: http://arxiv.org/abs/2203.06173

[44] A. Majumdar, K. Yadav, S. Arnaud, Y. J. Ma, C. Chen, S. Silwal, A. Jain, V.-P. Berges, P. Abbeel, J. Malik, D. Batra, Y. Lin, O. Maksymets, A. Rajeswaran, and F. Meier, "Where are we in the search for an Artificial Visual Cortex for Embodied Intelligence?" 2024, arXiv:2303.18240 [cs]. [Online]. Available: http://arxiv.org/abs/2303.18240

[45] S. Garcin, T. McInroe, P. S. Castro, P. Panangaden, C. G. Lucas, D. Abel, and S. V. Albrecht, "Studying the Interplay Between the Actor and Critic Representations in Reinforcement Learning," 2025, arXiv:2503.06343 [cs]. [Online]. Available: http://arxiv.org/abs/2503.06343

[46] Q. Gallouédec, N. Cazin, E. Dellandréa, and L. Chen, "panda-gym: Open-Source Goal-Conditioned Environments for Robotic Learning," *4th Robot Learning Workshop: Self-Supervised and Lifelong Learning at NeurIPS*, 2021.

[47] E. Coumans and Y. Bai, "PyBullet, a Python module for physics simulation for games, robotics and machine learning," 2016. [Online]. Available: http://pybullet.org

[48] P. J. Ball, L. Smith, I. Kostrikov, and S. Levine, "Efficient Online Reinforcement Learning with Offline Data," 2023, arXiv:2302.02948 [cs]. [Online]. Available: http://arxiv.org/abs/2302.02948