On Pre-Training for Visuo-Motor Control: Revisiting a Learning-from-Scratch Baseline

Anonymous Author(s)
Affiliation
Address
email

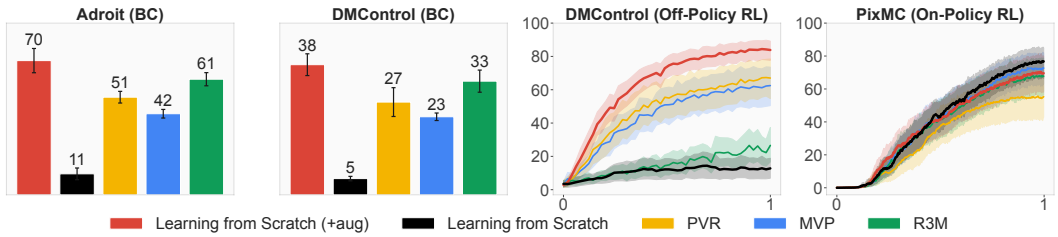


Figure 1: **Pre-training vs. Learning-from-Scratch (LfS).** Success rate (Adroit, PixMC) and normalized return (DMControl) in each of the three task domains that we consider (aggregated across tasks). BC results are averages of top-3 evaluations over 100 epochs [1], and RL results are reported as a function of environment steps [2, 3], normalized to $[0, 1]$ since number of steps differ between tasks. We evaluate strong LfS baselines [2, 4] and find them to be competitive with recent frozen pre-trained representations. We report mean and 95% confidence intervals over 5 seeds.

Abstract: We revisit a simple Learning-from-Scratch baseline for visuo-motor control that uses data augmentation and a shallow ConvNet. We find that this baseline has competitive performance with recent methods that leverage frozen visual representations trained on large-scale vision datasets.

Keywords: to pre-train; not to pre-train

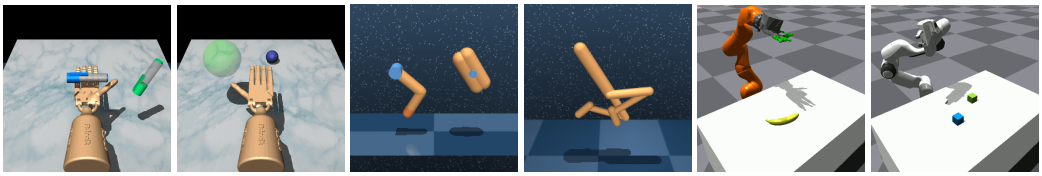
1 Introduction

Large-scale pre-training has delivered promising results in computer vision [5, 6, 7, 8] and natural language processing [9, 10, 11, 12]. Recent works have extended the pre-training paradigm to visuo-motor control by leveraging pre-trained visual representations for policy learning [1, 13, 3, 14, 15]. These works train a visual representation using large out-of-domain vision datasets like ImageNet [16] and Ego4D [17], and freeze the vision model weights for downstream policy learning. When compared to simple Learning-from-Scratch (LfS) methods for visuo-motor control, these works find that frozen pre-trained representations help achieve high sample efficiency and/or asymptotic performance across a variety of domains and algorithms.

However, there exists a rich body of work on strategies to improve performance of LfS methods, such as auxiliary self-supervised representation learning [18, 19] or using carefully curated data augmentations [20, 2, 4]. To gain a sharp understanding of the advantages of visual pre-training for motor control, it is necessary to establish strong LfS baselines. Towards this end, we take the experimental setups of prior works, and implement strong LfS baselines by adopting shallow ConvNet encoders and random shift augmentations. Surprisingly, we find that this modified LfS baseline can achieve results competitive with prior works that leverage frozen pre-trained visual representations. While our contributions are incremental in nature, we believe that our work contains must-know insights for anyone working on pre-trained representations for motor control.

We evaluate our approach across **3** task domains (Adroit [21], DMControl [22], PixMC [3]) and **3** algorithm classes: imitation learning (behavior cloning), on-policy RL (PPO [23]), and off-policy

Submitted to the 6th Conference on Robot Learning (CoRL 2022). Do not distribute.



Adroit DMControl PixMC

Figure 2: **Tasks.** We consider challenging and diverse visuo-motor control tasks spanning **3** domains: Adroit (dexterous manipulation), DMControl (locomotion, manipulation), and PixMC (manipulation).

RL (DrQ-v2 [2]). Our carefully designed LfS baseline is competitive with use of frozen pre-trained representations in all settings, and in some cases even outperforms them. We remain optimistic that pre-trained representations will play an important and increasingly larger role in visuo-motor control. At the same time, we believe that setting a simple yet strong baseline will help benchmark progress in this area. We conjecture that current benchmark tasks are not well suited to reap the benefits of pre-trained representations, since they do not require any visual generalization. As the community builds better benchmarks and harder tasks that require both visual and policy generalization, we conjecture that pre-trained representations will play an increasingly important role. *We are committed to releasing all of our proposed LfS baselines to the public.*

2 Experiments

Comparing two *paradigms* is difficult, and comparing LfS with pre-trainined representations is no exception. To help narrow our scope, we focus on *representative methods* from each paradigm: a simple LfS method using data augmentation and a shallow ConvNet, and three **frozen** visual representations trained on large-scale out-of-domain vision datasets (PVR [1], MVP [3], R3M [13]). We choose to freeze the pre-trained representations to be consistent with prior work. All three pre-trained representations that we consider have been shown to outperform common representations such as supervised learning and MoCo-v2 [6] pretraining on ImageNet [16].

We propose a set of strong LfS baselines that span **3** classes of algorithms: imitation learning (behavior cloning), on-policy RL (PPO [23]), and off-policy RL (DrQ-v2 [2]), and consider a total of **15** tasks across **3** domains: Adroit [21] (dexterous manipulation; 2 tasks), DMControl [22] (locomotion and control; 5 tasks), and PixMC [3] (robotic manipulation; 8 tasks). Figure 2 shows sample tasks from each domain. We base our experiments on the public implementations of PVR, MVP, and DrQ-v2, and *meticulously follow their respective experimental setups*. We summarize our experiments setup as follows:

- **Behavior Cloning (BC).** We consider two domains – Adroit and DMControl – used in PVR. Observations are 256×256 RGB images (center-cropped to 224×224) with no access to proprioceptive information. Policies are trained with BC on 100 demonstrations per task; Adroit demonstrations are generated by oracle DAPG [21] policies as in PVR, and DMControl demonstrations are generated by oracle TD-MPC [24] policies. The original LfS baseline in PVR uses a shallow ConvNet encoder. Our improved LfS baseline additionally uses random shift augmentation [20, 2] during learning, and we refer to this baseline as *LfS (+aug)*. Data augmentation is relatively underexplored in BC literature, but we find that it works surprisingly well. In addition to PVR, we also compare with frozen MVP and R3M representations. Consistent with the experimental setup in PVR, we measure the policy performance with success rate in case of Adroit, and episode returns in case of DMControl. The policies are evaluated every two epochs, and we report the average performance over the three best epochs over the course of learning.

- **On-policy RL.** We reproduce the results of MVP on their proposed PixMC robotic manipulation benchmark. Observations are 224×224 RGB images and also include proprioceptive information. The original LfS baseline uses a small ViT [25] encoder. We propose *two* improved LfS baselines for this setting: *(1)* an LfS baseline that uses a shallow ConvNet encoder and *no* data augmentation, referred to as *LfS*, and *(2)* an LfS baseline that additionally applies random shift augmentation in critic learning, referred to as *LfS (+aug)*. Following prior work [4, 26], we do not augment value

targets. In addition to (frozen) MVP, we also compare with frozen PVR and R3M representations. Following the setup in MVP, we use success rate of the policy as the metric for comparison.

- **Off-policy RL.** We reproduce the results of state-of-the-art LfS method DrQ-v2 on the same DMControl tasks as used in PVR. Observations are 84×84 RGB images with no access to proprioceptive information; we upsample observations to 224×224 when using pre-trained representations. DrQ-v2 uses a shallow ConvNet encoder and random shift augmentation by default, and we refer to this baseline as *LfS (+aug)*. We compare DrQ-v2 to two alternatives: *(1)* not using data augmentation (simply denoted *LfS*), and *(2)* removing data augmentation **and** additionally replacing the LfS encoder with a frozen pre-trained representation, denoted by their representation names (PVR, R3M, MVP) respectively. Following prior work on DMControl, we use (normalized) return as the metric for comparison.

2.1 Results

We summarize our key findings as follows:

- **Performance comparison.** Our proposed Learning-from-Scratch (LfS) baselines are competitive with (and in some cases outperform) recent frozen pre-trained representations for visuo-motor control across a variety of algorithms and domains; see Figure 1 and Table 2. This indicates that, while pre-trained representations have the potential to replace the LfS paradigm in the future, under the set of most widely used metrics, they have yet to exceed the representational power of a *well designed* LfS method on standard benchmarks.
- **No free lunch – yet.** Our results indicate that the efficacy of a frozen pre-trained representation is both *task-dependent* (see Figure 3) and *algorithm-dependent* (see Figure 1): R3M outperforms other pre-trained representations on both Adroit and DMControl using BC, but performs poorly on DMControl using RL. Likewise, MVP performs well on PixelMC, but comparably worse in the two BC domains. Even within a visually consistent benchmark (PixMC), no single representation comes out on top. In contrast, LfS generally produces consistent results across all settings, presumably due to learning from task-specific data; this hypothesis is supported by prior work on in-domain finetuning of pretrained representations [27, 14, 28].

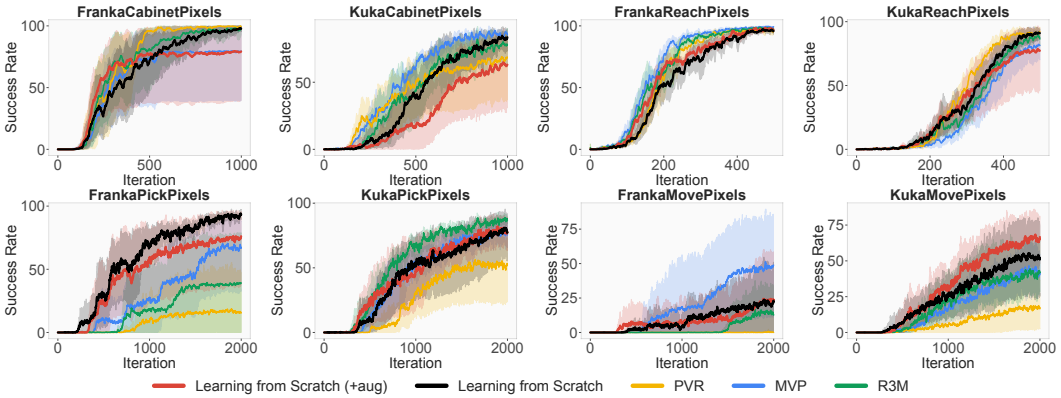


Figure 3: **PixMC benchmark.** PPO learning curves on the 8 robotic manipulation tasks from PixMC [3]. LfS performs comparably to pre-trained representations on most tasks. Averaged across 5 seeds.

Table 1: **Wall-time** of methods learning from scratch vs. using a pre-trained visual representation. For the latter, we report $\min\{\text{PVR}, \text{MVP}, \text{R3M}\}$ for a fair comparison. While LfS generally leads to better downstream task performance, using a **frozen** pre-trained representation can reduce computational cost substantially, especially during the training process. ↓ Lower is better.

	Behavior Cloning				**Reinforcement Learning**	
	Training (s/iteration)		Inference (s/episode)		s/1k frames	s/iteration
Method\Setting	Adroit	DMControl	Adroit	DMControl	DrQ-v2	PPO
LfS (+aug)	0.263	0.270	**1.61**	**3.81**	**10.20**	19.40
Fastest pre-training	**0.003**	**0.006**	2.66	11.00	13.00	**11.90**

- **Computational cost.** Our results so far has focused entirely on downstream task performance, like success rate or return. However, frozen pre-trained representations already demonstrate significant gains along an often-neglected axis: *wall-time*. Training and inference speeds are shown in Table 1. We find that BC policy updates are an order of magnitude faster using frozen pre-trained representations compared to LfS, as we can embed and cache features for the entire dataset in a few forward passes. However, inference speed generally favors LfS due to their smaller visual backbones, which is particularly important for real-robot applications. Since RL training interleaves learning and inference (data collection), wall-times are more balanced in this setting. We do not factor in the cost of learning a pretrained representation, since it is a one-time cost, and the representations can be reused across tasks.

3 Related Work

Pre-training. Representation learning via supervised/self-supervised/unsupervised pre-training on large-scale datasets has emerged as a powerful paradigm in areas such as computer vision [5, 6, 7, 8] and natural language processing [9, 10, 11, 12], where large datasets are available. While pre-trained representations can be finetuned to solve various downstream tasks, it may be prohibitively expensive to do so, and representations are therefore commonly used as-is, *i.e.*, with *frozen* weights. We reflect on recent progress and challenges when leveraging pre-trained visual representations for control, which is an emerging and comparably underexplored application area of such representations.

Table 2: **Imitation Learning.** Success rate (Adroit) and unnormalized return (DMControl) of LfS and our **best** result obtained with a pre-trained representation, *i.e.*, for each task we report $\max\{\text{PVR}, \text{MVP}, \text{R3M}\}$.

Task**Method**	LfS	LfS (+aug)	**Best** pre-training
Pen	14.0±5.6	**85.7±4.3**	**87.0±5.7**
Relocate	8.0±1.4	**55.0±10.1**	34.2±2.2
Finger Spin	6.2±3.0	445.3±12.2	**611.4±18.2**
Reacher Hard	35.6±18.1	**846.2±67.2**	602.0±107.8
Cheetah Run	8.8±5.6	171.0±18.8	**202.2±12.2**
Walker Stand	147.3±7.9	**311.7±55.1**	**309.3±22.7**
Walker Walk	38.4±2.4	**111.4±27.5**	80.6±5.3

Pretrained representations for control. Multiple works have explored learning control policies with visual representations pre-trained on large external datasets [29, 1, 13, 3, 27, 14, 15, 28]. In particular, PVR [1] and R3M [13] propose to learn policies by behavior cloning using pre-trained representations; PVR fuses features from several layers of a ResNet50 learned by MoCo [6], and R3M [13] learn a representation using a time-contrastive objective on ego-centric human videos. MVP [3] learn a policy with PPO and use a pre-trained visual encoder for feature extraction in addition to proprioceptive state information; the pre-trained representation is an MAE [30] trained on frames from diverse human videos. Despite encouraging results in leveraging pre-trained representations for control, we show that LfS remains competitive with (frozen) pre-trained representations at this time.

4 Discussion

We have shown that a carefully designed LfS baseline is competitive with frozen pre-trained representations across a variety of algorithm classes and domains. While this is the current conclusion, we remain optimistic that results will be skewed in favor of pre-trained representations as the paradigm matures. At present, we find that the main benefit of a frozen pretrained representation is the reduced training cost that comes with its *universality* – a single representation can be reused across tasks. Bridging the performance gap while maintaining universality will thus be critical to the adoption of this new paradigm. Recent works show that pretrained representations benefit from finetuning on task-specific data [27, 14, 28], combining elements of pre-training and LfS. However, finetuning large visual backbones present optimization challenges (*e.g.*, instability and catastrophical forgetting), and can be costly. We encourage further research in these directions, and hope that our strong LfS baselines will help accurately benchmark progress in this area.

References

[1] S. Parisi, A. Rajeswaran, S. Purushwalkam, and A. K. Gupta. The unsurprising effectiveness of pre-trained vision models for control. In *ICML*, 2022.

[2] D. Yarats, I. Kostrikov, and R. Fergus. Image augmentation is all you need: Regularizing deep reinforcement learning from pixels. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=GY6-6sTvGaf.

[3] T. Xiao, I. Radosavovic, T. Darrell, and J. Malik. Masked visual pre-training for motor control. *ArXiv*, abs/2203.06173, 2022.

[4] N. Hansen, H. Su, and X. Wang. Stabilizing deep q-learning with convnets and vision transformers under data augmentation. In *NeurIPS*, 2021.

[5] C. Doersch, A. K. Gupta, and A. A. Efros. Unsupervised visual representation learning by context prediction. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1422–1430, 2015.

[6] K. He, H. Fan, Y. Wu, S. Xie, and R. B. Girshick. Momentum contrast for unsupervised visual representation learning. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735, 2020.

[7] A. van den Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding. *ArXiv*, abs/1807.03748, 2018.

[8] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022.

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805, 2019.

[10] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. J. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. *ArXiv*, abs/2005.14165, 2020.

[11] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763, 2021.

[12] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

[13] S. Nair, A. Rajeswaran, V. Kumar, C. Finn, and A. Gupta. R3m: A universal visual representation for robot manipulation. *ArXiv*, abs/2203.12601, 2022.

[14] Y. Ze, N. Hansen, Y. Chen, M. Jain, and X. Wang. Visual reinforcement learning with self-supervised 3d representations. *arXiv preprint arXiv:2210.07241*, 2022.

[15] Z. Yuan, Z. Xue, B. Yuan, X. Wang, Y. Wu, Y. Gao, and H. Xu. Pre-trained image encoder for generalizable visual reinforcement learning. In *First Workshop on Pre-training: Perspectives, Pitfalls, and Paths Forward at ICML 2022*, 2022. URL https://openreview.net/forum?id=E-0zNz5J5BM.

[16] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and L. Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115:211–252, 2015.

[17] K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, M. Liu, X. Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022.

[18] A. Srinivas, M. Laskin, and P. Abbeel. Curl: Contrastive unsupervised representations for reinforcement learning. In *ICML*, 2020.

[19] M. Schwarzer, A. Anand, R. Goel, R. D. Hjelm, A. C. Courville, and P. Bachman. Data-efficient reinforcement learning with self-predictive representations. In *ICLR*, 2021.

[20] I. Kostrikov, D. Yarats, and R. Fergus. Image augmentation is all you need: Regularizing deep reinforcement learning from pixels. *ArXiv*, abs/2004.13649, 2021.

[21] A. Rajeswaran, V. Kumar, A. Gupta, G. Vezzani, J. Schulman, E. Todorov, and S. Levine. Learning Complex Dexterous Manipulation with Deep Reinforcement Learning and Demonstrations. In *Proceedings of Robotics: Science and Systems (RSS)*, 2018.

[22] Y. Tassa, Y. Doron, A. Muldal, T. Erez, Y. Li, D. de Las Casas, D. Budden, A. Abdolmaleki, et al. Deepmind control suite. Technical report, DeepMind, 2018.

[23] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. *ArXiv*, abs/1707.06347, 2017.

[24] N. Hansen, X. Wang, and H. Su. Temporal difference learning for model predictive control. In *ICML*, 2022.

[25] 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. *ArXiv*, abs/2010.11929, 2021.

[26] R. Raileanu, M. Goldstein, D. Yarats, I. Kostrikov, and R. Fergus. Automatic data augmentation for generalization in deep reinforcement learning. *arXiv preprint arXiv:2006.12862*, 2020.

[27] C. Wang, X. Luo, K. W. Ross, and D. Li. Vrl3: A data-driven framework for visual deep reinforcement learning. *ArXiv*, abs/2202.10324, 2022.

[28] Y. Xu, N. Hansen, Z. Wang, Y.-C. Chan, H. Su, and Z. Tu. On the feasibility of cross-task transfer with model-based reinforcement learning. *arXiv preprint arXiv:2210.10763*, 2022.

[29] R. Shah and V. Kumar. Rrl: Resnet as representation for reinforcement learning. *ArXiv*, abs/2107.03380, 2021.

[30] K. He, X. Chen, S. Xie, Y. Li, P. Doll'ar, and R. B. Girshick. Masked autoencoders are scalable vision learners. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15979–15988, 2022.