Exploring Visual Pre-training for Robot Manipulation: Datasets, Models and Methods

Ya Jing[1,*], **Xuelin Zhu**[1,2,*†], **Xingbin Liu**[1,*†], **Qie Sima**[1,3,†],
Taozheng Yang[1], **Yunhai Feng**[1,†], **Tao Kong**[1]
[1]ByteDance AI Lab, [2]Southeast University, [3]Tsinghua University

Abstract: Visual pre-training with large-scale real-world data has made great progress in recent years, showing great potential in robot learning with pixel observations. However, the recipes of visual pre-training for robot manipulation tasks are yet to be built. In this paper, we first thoroughly investigate the effects of pre-training from three fundamental perspectives: datasets, model architectures and training methods. Several important observations are given that are beneficial for robot manipulation learning. Then, we propose a visual pre-training scheme for robot manipulation termed Vi-PRoM, which combines self-supervised learning and multi-task supervised learning. Concretely, the former employs contrastive learning to acquire underlying patterns from large-scale unlabeled data, while the latter allows learning visual semantics and temporal dynamics to facilitate robot manipulation tasks. Extensive experiments on robot manipulations in various simulation environments and the real robot demonstrate the superiority of the proposed scheme. We hope our study can motivate people in this topic.

Keywords: Visual Pre-training, Robot Manipulation

1 Introduction

The past years have witnessed substantial progress in visual representation learning based on deep neural networks. After pre-training on large-scale visual data [1, 2], the neural network is subsequently employed as a general-purpose encoder to extract visual representations for many tasks, e.g., image segmentation [3], object detection [4] and autonomous driving [5], showing its strong generalization ability, while also highlighting its potential in robot manipulation.

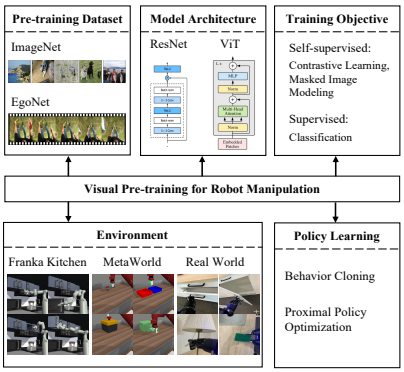


Figure 1: General path of visual pre-training for robot manipulation.

Recently, visual pre-training for robot learning has attracted increasing interest. Prominent performance gains reported on prior works [6, 7, 8] show its great potential in learning robot control from deep representations. However, these works differ in pre-training data, methods and models. So it remains an open question about which types of data, pre-training methods and models can better assist robot manipulation. A system-level benchmark on the profits of visual pre-training is in demand.

In this paper, we first conduct extensive studies on visual pre-training from three fundamental aspects: datasets, models and methods that may influence the performance of robot learning (Figure 1). Hopefully, these findings can facilitate future research in the community. Based on empirical

*Equal contribution. †Interns at ByteDance.
{jingya,kongtao}@bytedance.com, zhuxuelin@seu.edu.cn

findings, we propose a visual pre-training scheme oriented for robot manipulations, which combines self-supervised pre-training and supervised multi-task fine-tuning. Concretely, the visual encoder is first pre-trained based on contrastive learning [2, 9], allowing the model to acquire sequential patterns implicitly for the input data. Then, supervised multi-task learning is applied by constructing pseudo-labels and temporal labels to encourage the visual encoder further to perceive visual semantics and temporal dynamics. Both the self-supervised pre-training and supervised multi-task fine-tuning do not need manual human annotation. In addition, we propose a new dataset named EgoNet based on Ego4d [10] to serve as a benchmark to pre-train visual models for robot manipulations.

2 Exploring and Benchmarking

In this section, we explore key components that affect the pre-training behaviors and the robot manipulation performance, i.e., pre-training datasets, optimization methods, and model architectures. We first pre-train the visual encoder in self-supervised way on the pre-training dataset. Then we adopt typical imitation learning methods on robot manipulation tasks to verify the effectiveness of visual representations, where the encoder parameters are frozen during training. We adopt two robot control simulation environments, i.e., Franka Kitchen [11] and MetaWorld [12], to evaluate the effectiveness.

1) Dataset ImageNet [1] has recently been widely used in self-supervised pre-training for various downstream tasks. However, ImageNet lacks dynamic interaction between objects, making it may be unsuitable to serve as pre-training data for robot manipulation tasks.

We propose a new benchmark, called EgoNet, to pre-train visual encoders for robot manipulation. It comprises nearly 500,000 video clips covering hundreds of scenarios and is rich in human-object interactions. The EgoNet is constructed based on Ego4D [10]. We empirically intercept a short clip with a duration of 1s for each narration. With this strategy, a total of 5.03 million video clips are collected. After a 10-fold down-sampling, EgoNet is obtained that contains about 1.5 million video frames in total, making the training samples number comparable with ImageNet.

Interaction-related dataset is more powerful. We pre-train visual encoders (ResNet-50 and ViT-Base) on different datasets, i.e., ImageNet and EgoNet, using the contrastive learning method (MoCo-v3 [9]), and observe their performance on the robot manipulation tasks. From Table 1, we can see that models pre-trained on EgoNet, whether ResNet-50 or ViT-Base, achieve better performance on robot manipulation tasks. Obviously, the robot favors the interaction-related knowledge and temporal relationships contained in the video in terms of manipulation tasks.

2) Model Architecture The architecture of visual encoder is also important. To explore the effect of model architecture, we choose two typical models, namely convolution-based ResNet-50 [13] and self-attention-based ViT-Base [14]. They both have been the defacto standard for visual representation. In this way, we could provide insight into which architectures are more beneficial for robot manipulation tasks.

Convolution-based network architecture is preferred in retaining visual knowledge for robot manipulation. Recalling Table 1, we can observe that ResNet-50 performs better than ViT-Base on the robot manipulation tasks in both simulation environments, whether pre-trained on ImageNet or EgoNet. This observation shows the advantage of convolution-based model architecture compared to the self-attention-based one on robot manipulation tasks.

3) Pre-training Method The learning objective directly determines the type of representations that the model can learn from a dataset. Contrastive learning [9, 2, 15] and masked image modeling [16, 17], the two most prevalent pre-training methods in self-supervised learning, are naturally the main exploration goals in this work. We choose MoCo-v3 [9] and MAE (Masked AutoEncoder) [16] for contrastive learning and masked image modeling, respectively.

The sequential pattern and semantic information learned by contrastive learning are more effective. We choose ViT-Base as the visual encoder considering its generalization in both contrastive

Model	Dataset	Learning Method	Franka Kitchen	MetaWorld
ResNet-50	ImageNet	MoCo-v3 [9]	31.1	54.1
ResNet-50	EgoNet	MoCo-v3 [9]	**40.5**	**61.2**
ViT-Base	ImageNet	MoCo-v3 [9]	30.5	53.2
ViT-Base	EgoNet	MoCo-v3 [9]	**32.0**	**57.1**
ViT-Base	ImageNet	MAE [16]	11.4	50.3
ViT-Base	EgoNet	MAE [16]	18.0	49.8

Table 1: Effects of pre-training datasets, model architectures and pre-training methods for robot manipulation on Franka Kitchen and MetaWorld, using success rate (%) as the metric.

learning and masked image modeling. As shown in Table 1, MoCo-v3 outperforms MAE on both ImageNet and EgoNet datasets. This result suggests that the visual semantics acquired by contrastive learning are more important for robot manipulation than the structural information learned by masked image modeling.

3 Vi-PRoM

Based on the above explorations, we propose Visual Pre-training scheme for Robot Manipulation (Vi-PRoM). Concretely, we first employ contrastive learning to acquire human-object interaction patterns from the EgoNet dataset in a self-supervised manner. Then two supervised learning objectives, i.e., visual semantics predicting and temporal dynamics predicting, are adopted to further enrich the encoder. Note that we do not need manually annotate the labels to learn both visual semantics and temporal dynamics.

Contrastive Self-supervised Learning We hypothesize a good visual representation should have the ability to distinguish different scenes. Therefore, we use contrastive learning as our self-supervised paradigm to learn rich and general visual representations. Specifically, we sample a minibatch of images and minimize the InfoNCE loss [15] as that in [9].

Supervised Multi-task Learning With the learned representation from contrastive learning, we further enrich the model to learn visual semantics and temporal dynamics to facilitate robot manipulations, inspired by observations in Section 2.

We first introduce a pseudo-label predicting task to fine-tune the learned backbone, encouraging the model to learn better visual semantic representations. Specifically, we employ a ResNet-101 model supervised on ImageNet to generate pseudo labels for each sample x_i in EgoNet. Then, the pseudo label is used to fine-tune our self-supervised learned backbone with the cross-entropy loss:

$$\mathcal{L}_{\text{VS}} = -\mathbb{E}_D \sum_{i=1}^{N} \mathcal{T}(x_i) \log(h_1(f(x_i))), \tag{1}$$

where D is the EgoNet dataset, $\mathcal{T}$ is the ResNet-101 network to generate pseudo labels for each sample, f is the backbone, and h_1 is a classification head.

Further, we design a frame order prediction task to enable the model to learn temporal dynamics for each clip of EgoNet. Given the image set $\mathcal{I} = \{x_0, ..., x_k, ..., x_{N-1}\}$ sampled sequentially from a video clip, we first shuffle these images randomly and then predict the original order for the image x_k. This task is formulated as a classification problem of N classes, which is commonly solved by minimizing the cross-entropy loss.

$$\mathcal{L}_{\text{TD}} = -\mathbb{E}_D \sum_{k=0}^{N-1} \mathbf{y}_k \log(h_2(f(x_k))), \tag{2}$$

where h_2 is a classification head. $\mathbf{y}_k$ denotes the order of the image x_k in original image set $\mathcal{I}$. *We combine the visual semantics (Eq.1) and temporal dynamics (Eq.2) loss for jointly training.*

4 Experiments

To evaluate the pre-trained visual encoder on robot manipulation tasks, we take it as a frozen module for policy learning. Unless otherwise specified, the demonstration dataset size used for imitation learning is set as 5. The average of the best success rates on all manipulation tasks with three different seeds (100, 125, 150) is reported to measure the performance of the visual encoder.

Simulation Environments We compare Vi-PRoM with the state-of-the-art visual pre-training methods for robot manipulation. For fair comparisons, except for the scratch method whose visual encoder parameters are randomly initialized, all other models are pre-trained on our EgoNet dataset and evaluated with the behavior cloning method. Note that the visual encoder for each method is ResNet-50. Experimental results are reported in Table 2. It can be seen that our model achieves the best performance in both simulation environments.

Real Robot We deploy our model on a real robot to demonstrate its performance in the real environment. Figure 2 shows two successful cases of our model in the real robot environment. Overall, benefiting from the powerful representational capability of our model, the robot is competent for manipulation tasks in the real kitchen environment by learning from human demonstrations.

Method	Franka Kitchen		MetaWorld	
	BC	PPO	BC	PPO
Scratch	22.3	15.2	26.5	28.8
R3M [6]	27.4	18.3	61.7	38.6
MoCo-v3 [9]	40.5	36.8	61.2	43.6
Vi-PRoM	**43.8**	**39.5**	**63.5**	**46.8**

Table 2: Main results (success rate in %).

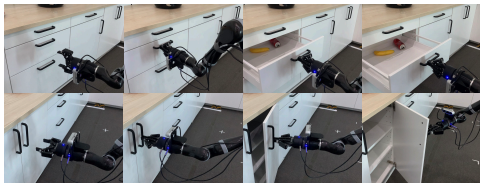
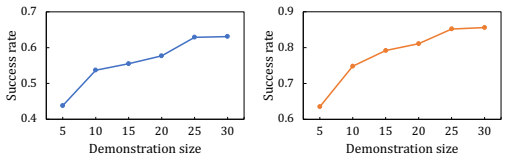


Figure 2: Real robot can open the drawer and the door with the help of Vi-PRoM model.

Visual Semantics	Temporal Dynamics	Franka Kitchen	MetaWorld
		40.5	61.2
✓		43.2	62.0
	✓	40.7	62.6
✓	✓	**43.8**	**63.5**

Table 3: Ablation study on different modules (success rate in %).



(a) Franka Kitchen (b) MetaWorld

Figure 3: Effects of demonstration size on robot manipulation tasks.

Ablation Study Table 3 exhibits the experimental results of ablation studies, which demonstrate the importance of each components to help robot manipulations. We also investigate the scalability of Vi-PRoM as shown in Figure 3. In both Franka Kitchen and MetaWorld environments, the success rate of Vi-PRoM improves steadily as the size of the demonstration data increases.

5 Conclusion and Discussion

We have explored three crucial components that affect the pre-trained model on robot manipulation tasks. Key conclusions are drawn that robot manipulation prefers human-object interaction dataset, convolution-based network, as well as temporal and semantic information. We further propose Vi-PRoM for robot manipulation. Extensive experiments on simulators and the real environment demonstrate its superiority.

There are still many issues to be further explored. First, training visual encoders directly on videos has the potential to learn better temporal dynamics. However, how to construct the representations from a video clip at the initial stage is a problem. Then, currently visual encoders are pre-trained on real-world data while evaluated in simulation environments. This gap can lead to some unexpected results. It inspires us to establish an real-data benchmark to facilitate research.

Acknowledgments

We would like to thank Yifeng Li, Minzhao Zhu, Yuxi Liu, Tao Wang and Yunfei Liu for their help on the robot system, Hang Li for helpful feedback, and other colleagues at ByteDance AI Lab for support throughout this project.

References

[1] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[2] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[3] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[4] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, 2015.

[5] Q. Zhang, Z. Peng, and B. Zhou. Action-conditioned contrastive policy pretraining. *arXiv preprint arXiv:2204.02393*, 2022.

[6] S. Nair, A. Rajeswaran, V. Kumar, C. Finn, and A. Gupta. R3m: A universal visual representation for robot manipulation. In *6th Annual Conference on Robot Learning*, 2022.

[7] T. Xiao, I. Radosavovic, T. Darrell, and J. Malik. Masked visual pre-training for motor control. *arXiv preprint arXiv:2203.06173*, 2022.

[8] I. Radosavovic, T. Xiao, S. James, P. Abbeel, J. Malik, and T. Darrell. Real-world robot learning with masked visual pre-training. *arXiv preprint arXiv:2210.03109*, 2022.

[9] X. Chen, S. Xie, and K. He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

[10] 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 Conference on Computer Vision and Pattern Recognition*, 2022.

[11] A. Gupta, V. Kumar, C. Lynch, S. Levine, and K. Hausman. Relay policy learning: Solving long-horizon tasks via imitation and reinforcement learning. *Conference on Robot Learning (CoRL)*, 2019.

[12] T. Yu, D. Quillen, Z. He, R. Julian, K. Hausman, C. Finn, and S. Levine. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on Robot Learning*, 2020.

[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[14] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[15] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, 2020.

[16] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022.

[17] H. Bao, L. Dong, and F. Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.