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 Research, [2]Southeast University, [3]Tsinghua University
https://explore-pretrain-robot.github.io

***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 thoroughly investigate the effects of visual pre-training strategies on robot manipulation tasks from three fundamental perspectives: pre-training datasets, model architectures and training methods. Several significant experimental findings are provided that are beneficial for robot learning. Further, we propose a visual pre-training scheme for robot manipulation termed Vi-PRoM, which combines self-supervised learning and supervised learning. Concretely, the former employs contrastive learning to acquire underlying patterns from large-scale unlabeled data, while the latter aims learning visual semantics and temporal dynamics. Extensive experiments on robot manipulations in various simulation environments and the real robot demonstrate the superiority of the proposed scheme. Videos and more details can be found on https://explore-pretrain-robot.github.io.**

I. 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, the neural network is subsequently employed as a general-purpose encoder to extract visual representations for many tasks, e.g., image segmentation [1], object detection [2] and autonomous driving [3], showing its strong generalization ability, while also highlighting its potential in robot manipulation.

Learning from visual observations for robot manipulation is known as a challenging task that requires a thorough understanding of both visual semantics and sequential patterns of observations. A common method is to train the visual encoder and model-based policy from scratch in an end-to-end manner with in-domain data [4], [5]. Despite its effectiveness to some degree, such a method requires training on a large number of observation-action samples, which may limit its wide applications. Therefore, pre-training the visual encoder with large-scale off-the-shelf data from the real world can serve as an alternative. Benefiting from its strong generalization ability, the pre-trained visual encoder is expected to generalize across a range of robot manipulation tasks and enable data-efficient learning.

Recently, visual pre-training on large-scale real-world data for robot learning has attracted increasing interest. Prominent performance gains reported on prior works [6], [7] show its great potential in learning robot control from pixels. Despite the claimed advantage, 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.

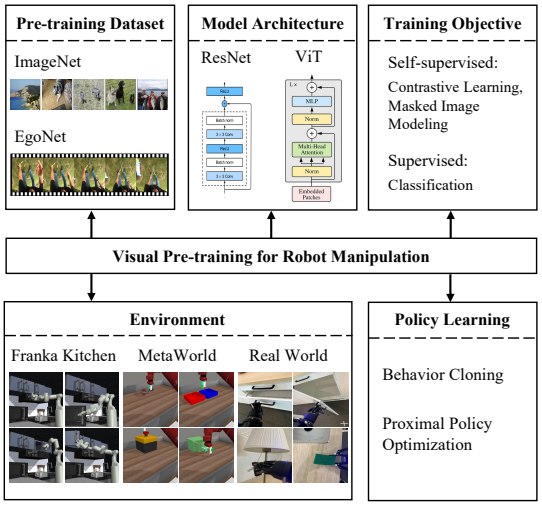


Fig. 1. General path of visual pre-training for robot manipulation.

In this paper, as shown in Figure 1, 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. Hopefully, these can facilitate future research in the community. Based on empirical findings, we propose a visual pre-training scheme oriented for robot manipulations, which sequentially trains a visual encoder using self-supervised learning and supervised fine-tuning. Concretely, the visual encoder is first pre-trained based on contrastive learning [8], allowing the trained model to acquire sequential patterns implicitly for the input data. Then, supervised learning is applied by constructing pseudo-labels and temporal labels to encourage the visual encoder further to perceive visual semantics and temporal dynamics. In addition, we propose a new dataset named EgoNet, which is created based on Ego4d [9] and contains a large-scale egocentric video clips rich in human-object interactions. EgoNet has the potential to serve as a benchmark to pre-train visual models for robot manipulations.

*Equal contribution.
‡Corresponding author: Tao Kong (kongtao@bytedance.com).

Our main contributions are summarized as: (1) We create the EgoNet dataset, a new benchmark enriched with diverse scenarios and human-object interactions for robotic visual pre-training. (2) We fully explore the visual pre-training in terms of datasets, methods and models, and provide several key suggestions for robot manipulation tasks. (3) We propose a novel cascade visual pre-training scheme that enables the visual encoder to learn sequential patterns, visual semantics and temporal dynamics from the large-scale real-world data, and achieves remarkable performance improvement on robot manipulation tasks.

II. RELATED WORK

A. Vision-Based Robot Learning

The robotic community has long focused on vision-based learning methods for various robot tasks in the past decade. Currently, the most prevailing paradigm of vision-based robot learning is the end-to-end method [5]. With the surge of deep learning in the last decade, many CNN-based models have been proposed to enable the visual modality of robots in manipulation tasks [10], [11]. Furthermore, CNN-RNN methods [12], [13] are widely adopted to solve the task of human instruction in natural language. Recently, many methods [6], [7], [14] based on pre-trained models have been proposed for robot learning. Several previous methods investigated the self-supervised pre-training in robot manipulation, e.g., R3M [6], MVP [7], and MaskViT [14]. These works focus on one side of visual pre-training, thus calling for a systematic study.

B. Representation Learning

Self-supervised visual pre-training has been an active research topic recently, and can learn universal visual representations. Visual pre-training aims to learn visual representations by masked image modeling [15], [16] and contrastive learning [8], [17]. While the vision-language pre-training aims to learn the semantic correspondence between different modalities [18], [19]. Pre-training datasets are significant for the representation learning. To learn reusable representations that can generalize well to robotic manipulation tasks, the interaction between humans and objects needs to be captured. Recently, a diverse and large-size dataset Ego4D [9] has been proposed, which contains daily-life activity videos spanning hundreds of scenarios.

C. Robot Manipulation Benchmarks

With the recent progress in exploiting the pre-trained models in robotic tasks, a number of robotic manipulation benchmarks have been introduced to evaluate the performance of the pre-trained model. The off-shelf robotic manipulation benchmarks can be categorized into two main kinds by simulators: RL (Reinforcement Learning) benchmarks and embodied benchmarks. The RL benchmarks focus on the training and evaluation of reinforcement learning agents where a simulated environment with several robot models and scenarios in limited space is usually provided. Recent RL benchmarks explore the training and evaluation of robotic manipulation method in aspects of multi-task training [20], more realistic scenarios with clutter [21], tasks in higher complex level [22], more kinds of manipulation forms [23] and manipulations with linguistic instructions [24], [25]. Meanwhile, the pre-trained models are widely introduced as solutions to robot manipulation tasks.

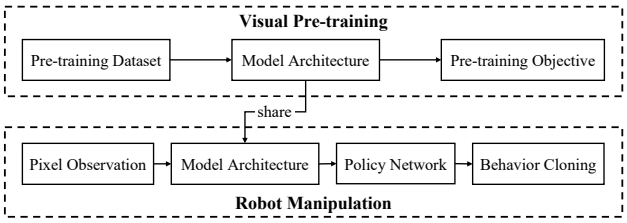


Fig. 2. The study pipeline of visual pre-training for robot manipulation.

III. 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. The study pipeline is shown in Figure 2. We first pre-train the visual encoder 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. In this way, we could give system-level studies of each component.

A. Benchmarking Setup

To evaluate the effectiveness of the pre-trained visual encoder, we adopt two robot control simulation environments, i.e., Franka Kitchen [26] and MetaWorld [20], for robot learning. As shown in the right part of Figure 3, we choose the same tasks as [6]. Please refer to Section V-A for the pre-training details and evaluation metrics.

1) Pre-training Dataset: ImageNet [27] 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 [9]. We experimentally intercept a short clip with a duration of 1s for each narration. With this strategy, a total of 0.503 million video clips rich in human-object interactions are collected. Note that the video in Ego4D has a frame rate of 30 fps. After a 10-fold uniform downsampling, EgoNet is obtained that contains about 1.5 million video frames in total, making the training samples number comparable with ImageNet.

2) Model Architecture: The architecture of visual encoder is also an important element in determining the performance of robot manipulation tasks. To explore its effect, we choose

TABLE I

EFFECTS OF PRE-TRAINING DATASETS ON ROBOT MANIPULATION ON TWO SIMULATORS, I.E., FRANKA KITCHEN AND METAWORLD, USING SUCCESS RATE (%) AS THE METRIC.

Model	Dataset	Franka Kitchen	MetaWorld
ResNet-50	ImageNet	31.1	54.1
ResNet-50	EgoNet	**40.5**	**61.2**

TABLE II

EFFECTS OF MODEL ARCHITECTURES ON ROBOT MANIPULATION.

Model	Dataset	Franka Kitchen	MetaWorld
ResNet-34	EgoNet	22.6	52.4
ResNet-50	EgoNet	**40.5**	61.2
ResNet-101	EgoNet	40.0	**61.6**

three typical models, namely convolution-based ResNet-34 [28], ResNet-50 [28], and ResNet-101 [28], which have been the defacto standard for visual representation extraction. In this way, we could provide insight into which architectures are more beneficial for 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 and masked image modeling, the two most prevalent pre-training methods in self-supervised learning, are naturally the main exploration goals in this work. Contrastive learning aims to encourage the feature similarity between two different augmented views of the same image but suppress the similarity between different images. Masked image modeling resorts to reconstructing the randomly masked patches of the input image. In this work, we choose MoCo-v3 [8] and MAE (Masked AutoEncoder) [15] for contrastive learning and masked image modeling, respectively.

B. Main Observations

1) Pre-training Dataset: **EgoNet is more powerful than ImageNet.** We pre-train visual encoder (i.e., ResNet-50) on different datasets, i.e., ImageNet and EgoNet, using the contrastive learning method (MoCo-v3), and observe their performance on the robot manipulation tasks. From Table I, we can see that the model pre-trained on EgoNet 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. In addition, the egocentric natural images in EgoNet have much more global context about the world, which means richer visual features can be learned.

2) Model Architecture: **ResNet-50 performs better.** From Table II, we can observe that ResNet-50 and ResNet-101 perform better than ResNet-34 on the robot manipulation tasks in both simulation environments pre-trained on EgoNet. In addition, there is no performance improvement as the model increases from ResNet-50 to ResNet-101. Furthermore, recent work suggests that pre-training ViT [29] models with larger pre-trained datasets can achieve better results.

TABLE III

EFFECTS OF PRE-TRAINING METHODS ON ROBOT MANIPULATION.

Learning Method	Dataset	Franka Kitchen	MetaWorld
MAE [15]	ImageNet	11.4	50.3
MoCo-v3 [8]	ImageNet	**31.1**	**54.1**
MAE [15]	EgoNet	18.0	49.8
MoCo-v3 [8]	EgoNet	**40.5**	**61.2**

3) Pre-training Method: **Contrastive learning is preferred.** As shown in Table III, MoCo-v3 outperforms MAE on both ImageNet and EgoNet datasets, demonstrating the effectiveness of contrastive learning compared to masked image modeling for manipulation. This result also suggests that the visual semantics acquired by contrastive learning are more important for robot manipulation than the structural information learned by masked image modeling.

C. Summary

Through the aforementioned explorations on various pre-training datasets, model architectures and pre-training methods, three key conclusions could be drawn:

- Visual pre-training with human-object interaction data is of great importance for robot manipulation.
- Convolution-based ResNet-50 is preferred in retaining visual knowledge for robot manipulation.
- The sequential pattern and semantic information learned by contrastive learning are more effective.

IV. PROPOSED APPROACH

Based on the above explorations, we propose Visual Pre-training scheme for Robot Manipulation (Vi-PRoM), which pre-trains ResNet-50 on the EgoNet dataset to extract comprehensive visual representations for robot manipulation. Specifically, we first employ contrastive learning to acquire human-object interaction patterns from the EgoNet dataset in a self-supervised manner. Then two additional learning objectives, i.e., visual semantics predicting and temporal dynamics predicting, are proposed to further enrich the encoder. Figure 3 shows a basic pipeline of the proposed Vi-PRoM. Note that we do not need manually annotate the labels to learn both visual semantics and temporal dynamics.

A. 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 let the model learn rich and general visual representations. The contrastive objective function pulls features generated by similar images together and pushes the features generated by different images away. Specifically, we sample a minibatch of images and minimize the InfoNCE loss [30].

B. Supervised Learning

With the learned representation from contrastive learning, it is imperative to learn visual semantics and temporal dynamics to generalize well for robot manipulation.

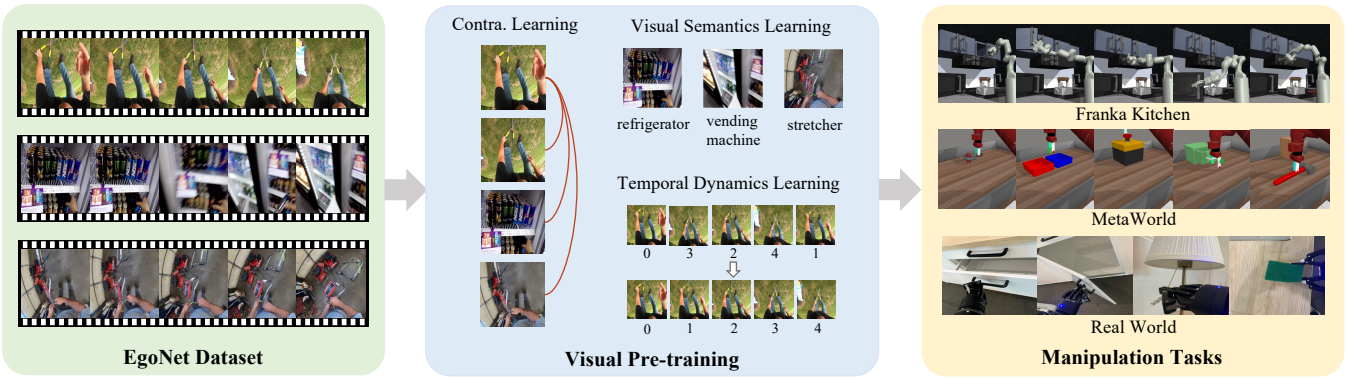


Fig. 3. The pipeline of our Vi-PRoM. The EgoNet dataset is first constructed to serve as pre-training data. We first pre-train the ResNet-50 with contrastive learning, enabling the model to learn universal visual representations. Then, frame-order and pseudo-label predicting tasks are jointly applied to encourage the model to capture temporal and semantic visual representations. Note that the pseudo-labels are automatically generated by the ResNet-101 model pre-trained on ImageNet without manual labeling. Finally, the pre-trained model is utilized to extract visual representations for robot manipulation tasks.

1) Learning Visual Semantics: We 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 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 x_i, f is the backbone, and h_1 is a classification head.

2) Predicting Temporal Dynamics: Robot manipulation tasks require predicting the next actions based on current and historical observations. Thus they are sensitive to temporal dynamics. 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 scramble these images 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 $\mathcal{I}$.

3) Joint Training: We combine the visual semantics and temporal dynamics loss for jointly training:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{VS}} + \lambda \mathcal{L}_{\text{TD}}, \tag{3}$$

where λ is the balance coefficient set as 0.33 in practice. In principle, visual semantics and temporal dynamics predicting together guide the learning, enabling the model to learn semantic and temporal visual representations.

C. Robot Imitation Learning

Given the well-trained visual encoder f, the robot utilizes it to encode visual features of pixel observations for policy learning. In this work, we employ the typical behavior cloning (BC) [31] method to imitate expert demonstrations, where the policy network is parameterized as a two-layer perceptron.

V. EXPERIMENTS

A. Experimental Setup

To evaluate the pre-trained visual encoder on robot manipulation tasks, we take it as a frozen module for policy learning. We train the policy network for 20,000 steps using a batch size of 32 and an Adam optimizer with a learning rate of 0.001. Unless otherwise specified, the demonstration dataset size used for imitation learning is set as 5. In the PPO experiments, we train for 20 iterations with 10 epochs per iteration. The reward function we use is similar to [32]. 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.

In the real environment, our robot hardware is mainly formed by a differential-drive mobile base equipped with a 2d LiDAR and IMU and a 6-DoF arm. A 2-finger parallel gripper is equipped for contact-rich interactions. Between the end-effector and the arm, a force torque sensor is installed to measure the forces and torques experienced by the robot, which is utilized to stop the robot if any large forces or torques appear. The robot's wrist is equipped with an RGBD camera as its perception unit. The Intel core i7 CPU is chosen as the computing unit.

B. Main Results

1) Simulation Environments: To demonstrate the effectiveness of our Vi-PRoM, we compare it 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

TABLE IV

COMPARISON RESULTS WITH THE STATE-OF-THE-ART METHODS.

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 [8]	40.5	36.8	61.2	43.6
Vi-PRoM	**43.8**	**39.5**	**63.5**	**46.8**

TABLE V

ABLATION STUDY ON DIFFERENT MODULES.

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

with the behavior cloning method. Note that the visual encoder for each method is ResNet-50. Experimental results are reported in Table IV. It can be seen that our model achieves the best performance in both simulation environments. In addition, the performance gains of our Vi-PRoM over the MoCo-v3, reaching 3.3% and 2.3% in success rate in Franka Kitchen and MetaWorld, respectively, indicate the value of explicitly learning visual semantics and temporal dynamics.

To learn the temporal dynamics and visual semantics, R3M resorts to the time contrastive learning and video-language alignment. Compared with R3M, our Vi-PRoM shows considerable performance gains, especially in the Franka Kitchen environment. Notably, in terms of the capacity to learn visual semantics and temporal dynamics, our pseudo-label predicting and frame order modeling outperform the time contrastive learning and video-language alignment.

To further verify the effectiveness of our Vi-PRoM, we choose the proximal policy optimization (PPO) algorithm [33] as an alternative to behavior cloning. Experimental results are provided in Table IV. Our Vi-PRoM consistently outperforms all competitors on both learning algorithms.

2) Real Robot: We deploy our model on a real robot to demonstrate its performance in the real environment. In practice, we test our pre-trained representations on four tasks, i.e., `opening the door`, `closing the door`, `opening the drawer` and `closing the drawer`. We collect 30 demonstrations for each task. Figure 4 shows two successful cases of our model in the real robot environment. Overall, benefiting from the powerful representational capability of Vi-PRoM, the robot is competent for various manipulation tasks in the real kitchen environment by learning from demonstrations.

C. Ablation Study

1) Pre-training Components: Table V exhibits the experimental results. When visual semantic learning is absent, the success rate decreases by 3.1% and 0.9% on Franka Kitchen and MetaWorld, respectively. Analogously, a drop in success rate of 0.6% and 1.5% on Franka Kitchen and MetaWorld can

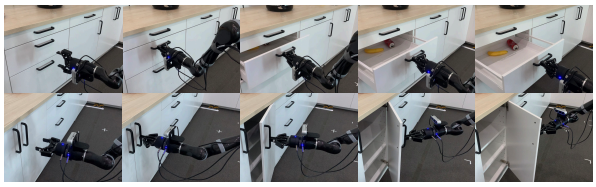
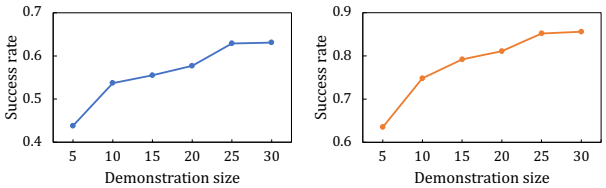


Fig. 4. The real robot is able to successfully open the drawer and the door with the help of our Vi-PRoM model in a kitchen environment.



(a) Franka Kitchen (b) MetaWorld

Fig. 5. Effects of demonstration size on robot manipulation tasks.

be observed in the absence of temporal dynamics learning. These two experimental results demonstrate the importance of visual semantics learning and temporal dynamics learning for robot manipulation. Moreover, when both learning objectives are absent, the success rate of Vi-PRoM suffers from considerable performance degradation. Therefore, the effectiveness of the collaboration between visual semantic learning and temporal dynamics learning is proved.

2) Model Scalability: We also investigate the scalability of Vi-PRoM. As shown in Figure 5, in both the Franka Kitchen and MetaWorld simulation environments, the success rate of Vi-PRoM improves steadily as the size of the demonstration data increases. After training on the larger expert demonstration dataset, our proposed Vi-PRoM model shows its scalability on robot manipulation tasks.

3) Other Models: We also report experimental results of directly taking the popular pre-trained models as visual encoders for robot manipulation, as shown in Table VI. ImageNet Supervised [27] is the ResNet-50 pre-trained for ImageNet classification task. MDETR [34] is the ResNet-101 pre-trained on large-scale image-text pairs. CLIP [35] is the ResNet-50 trained to align the image representation with the paired text. MAE is the ViT-Base trained on ImageNet. MVP is the ViT-large trained on Ego4D. It can be seen that all these models largely lag behind our Vi-PRoM model.

VI. DISCUSSION AND LIMITATION

In this paper, 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 ResNet-50 network, as well as temporal and semantic information. We further propose the Vi-PRoM for robot manipulation. Extensive experiments on simulators and the real environment demonstrate its superiority.

Although our pipeline is effective, there are still many issues to be further explored. First, training visual encoders

TABLE VI

PERFORMANCE OF OFFICIAL MODELS ON ROBOT MANIPULATION TASKS.

Method	Franka Kitchen	MetaWorld
ImageNet Supervised [27]	16.5	51.9
MDETR [34]	19.8	59.5
CLIP [35]	11.0	53.0
MAE [15]	11.4	50.3
MVP [7]	11.4	53.2
MoCo-v3 [17]	31.1	54.1
Vi-PRoM	**43.8**	**63.5**

directly on video clips has the potential to learn better temporal dynamics. Then using larger pre-training datasets is also worth exploring in the future. Finally, currently visual encoders are pre-trained on real-world data but evaluated in simulation environments. The significant gap can lead to some unexpected results, and also inspires us to consider establishing an evaluation benchmark from the real environment to facilitate research.

REFERENCES

[1] 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. 1

[2] 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. 1

[3] Q. Zhang, Z. Peng, and B. Zhou, "Action-conditioned contrastive policy pretraining," *arXiv preprint arXiv:2204.02393*, 2022. 1

[4] S. Levine, C. Finn, T. Darrell, and P. Abbeel, "End-to-end training of deep visuomotor policies," *The Journal of Machine Learning Research*, vol. 17, pp. 1334–1373, 2016. 1

[5] D. Kalashnikov, A. Irpan, P. Pastor, J. Ibarz, A. Herzog, E. Jang, D. Quillen, E. Holly, M. Kalakrishnan, V. Vanhoucke *et al.*, "Scalable deep reinforcement learning for vision-based robotic manipulation," in *Conference on Robot Learning*, 2018. 1, 2

[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. 1, 2, 5

[7] I. Radosavovic, T. Xiao, S. James, P. Abbeel, J. Malik, and T. Darrell, "Real-world robot learning with masked visual pre-training," in *Conference on Robot Learning*. 1, 2, 6

[8] 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. 1, 2, 3, 5

[9] 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. 1, 2

[10] J. Redmon and A. Angelova, "Real-time grasp detection using convolutional neural networks," in *2015 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2015. 2

[11] S. Kumra and C. Kanan, "Robotic grasp detection using deep convolutional neural networks," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2017. 2

[12] M. Shridhar, D. Mittal, and D. Hsu, "Ingress: Interactive visual grounding of referring expressions," *The International Journal of Robotics Research*, vol. 39, pp. 217–232, 2020. 2

[13] H. Zhang, Y. Lu, C. Yu, D. Hsu, X. Lan, and N. Zheng, "INVIGORATE: Interactive Visual Grounding and Grasping in Clutter," in *Robotics: Science and Systems*, 2021. 2

[14] A. Gupta, S. Tian, Y. Zhang, J. Wu, R. Martín-Martín, and L. Fei-Fei, "Maskvit: Masked visual pre-training for video prediction," *arXiv preprint arXiv:2206.11894*, 2022. 2

[15] 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. 2, 3, 6

[16] Z. Tong, Y. Song, J. Wang, and L. Wang, "Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training," *Advances in Neural Information Processing Systems*, 2022. 2

[17] 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. 2, 6

[18] Y.-C. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu, "Uniter: Universal image-text representation learning," in *European Conference on Computer Vision*, 2020. 2

[19] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi, "Align before fuse: Vision and language representation learning with momentum distillation," in *Advances in Neural Information Processing Systems*, 2021. 2

[20] 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. 2

[21] A. Gupta, V. Kumar, C. Lynch, S. Levine, and K. Hausman, "Relay policy learning: Solving long-horizon tasks via imitation and reinforcement learning," in *Conference on Robot Learning*, 2020. 2

[22] 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," *arXiv preprint arXiv:1709.10087*, 2017. 2

[23] S. James, Z. Ma, D. R. Arrojo, and A. J. Davison, "Rlbench: The robot learning benchmark & learning environment," *IEEE Robotics and Automation Letters*, vol. 5, pp. 3019–3026, 2020. 2

[24] K. Zheng, X. Chen, O. C. Jenkins, and X. E. Wang, "Vlmbench: A compositional benchmark for vision-and-language manipulation," *arXiv preprint arXiv:2206.08522*, 2022. 2

[25] O. Mees, L. Hermann, E. Rosete-Beas, and W. Burgard, "Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks," *IEEE Robotics and Automation Letters*, 2022. 2

[26] 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. 2

[27] 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, 5, 6

[28] 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. 3

[29] 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. 3

[30] 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. 3

[31] F. Torabi, G. Warnell, and P. Stone, "Behavioral cloning from observation," *arXiv preprint arXiv:1805.01954*, 2018. 4

[32] T. Xiao, I. Radosavovic, T. Darrell, and J. Malik, "Masked visual pre-training for motor control," *arXiv preprint arXiv:2203.06173*, 2022. 4

[33] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 5

[34] A. Kamath, M. Singh, Y. LeCun, G. Synnaeve, I. Misra, and N. Carion, "Mdetr-modulated detection for end-to-end multi-modal understanding," in *Proceedings of the IEEE International Conference on Computer Vision*, 2021. 5, 6

[35] 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*, 2021. 5, 6