# Real World Robot Learning with Masked Visual Pre-training

**Anonymous Author(s)**
Affiliation
Address
email

**Abstract:** Learning end-to-end visuomotor control directly within the real world remains a challenge. In this work, we utilize self-supervised visual pre-training on images from diverse, in-the-wild videos for real world robotic tasks. Like prior work, our visual representations are pre-trained via a masked autoencoder (MAE), are frozen, and then passed into a learnable control module. Unlike prior work, we show that these pre-trained representations are effective across a range of real word robotic tasks. We find that our encoder consistently outperforms CLIP (up to 75%), supervised ImageNet pre-training (up to 56%), and training from scratch (up to 81%). Finally, we train a 307M parameter vision encoder on a massive collection of images from the Internet and egocentric videos, and demonstrate the benefits of scaling visual pre-training for robot learning.

**Keywords:** Self-Supervised Learning, Visual Representations, Robot Learning

## 1 Introduction

Learning representations with large neural networks has been effective in computer vision [1, 2], natural language processing [3, 4, 5], and audio [6]. How can we transfer the success stories of modern deep learning to robotics? We can approach this from two ends: shared representations on the perception side or shared representations on the action side. Our focus in this paper is on the shared *visual* representations.

Of course, the devil is in the details. Recent developments in the field of visual learning have made this more feasible: (1) use of diverse, real-world data from the Internet and egocentric videos, (2) self-supervised objective that does not overly rely on data augmentations or other forms of strong human-designed priors, (3) scalable transformer models [7, 8] with high capacity, and (4) training of control policies on top of frozen visual representations. In recent work, Xiao et al. [9] have shown that self-supervised visual pre-training is effective for learning motor control tasks in simulation.

In this paper, we show that this framework is effective for real-world robotic tasks (Figure 1). We take inspiration from their work, but make significant advances in terms of data scale and diversity ($7\times$ larger), model size ($15\times$ bigger), and real world experiments (extensive real robot evaluations).

Specifically, we pre-train *self-supervised* visual representations on real-world images and videos from the Internet [10, 11, 12] and egocentric video datasets [13, 14]. We leverage the masked autoencoders (MAE) [15] that learn representations by masked prediction. The hope is that if the model can predict the missing content in real-world images it has learned useful properties of the visual world that can enable it to accomplish real-world robotic tasks. Given the pre-trained visual representations, we *freeze* the encoder and learn control policies on top. We focus on efficient real-world learning using behavior cloning with a handful of human provided demos per task (20-80).

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

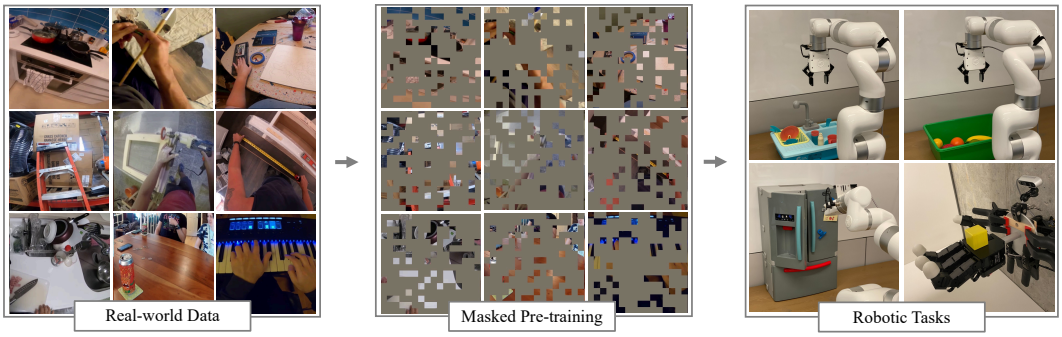


Figure 1: **Real world robot learning with masked visual pre-training.** We learn visual representations from massive collections of Internet and egocentric data. We pre-train representations with masked image modeling, freeze the encoder, and learn control policies for robotic tasks on top.

We evaluate our approach in an extensive real-world study and report results from 672 real-world experiments. We consider tasks varying in motion types, objects, and scenes. We find that our approach achieves considerably higher performance than CLIP (up to 75%), supervised pre-training (up to 56%), and training from scratch (up to 81%). Furthermore, we observe that our representations lead to large improvements in sample complexity, reaching the strongest baseline performance with half the number of demos.

In addition to showing that self-supervised visual pre-training is effective for real-world robotic tasks, we demonstrate the benefits of scale for robotics. In particular, we train a 307M parameter vision encoder [8] on a massive collection of images coming from ImageNet [10], Epic Kitchens [16], Something Something [11], 100 Days of Hands [12], and Ego4D [14]. Importantly, we observe that it is not sufficient to scale the model alone and that larger models require bigger datasets. To the best of our knowledge, ours is the largest vision model deployed for robotics, and demonstrates clearly the benefits of visual pre-training scale for robot learning.

## 2 Related Work

**End-to-end control** is concerned with learning to predict robot actions (e.g., joint velocities, end-effector poses, etc) directly from observations [17, 18, 19], without the need to perform explicit 3D scene understanding [20], grasp planning [21], and motion planning [22, 23]. However these end-to-end approaches tend to be too sample inefficient for real-world training. Some works have tried to find a balance between these explicitly pipelines approaches and end-to-end approaches [24, 25, 26]. Despite these hybrid approaches can accomplish more impressive tasks than purely end-to-end ones, they tend to be engineered for their specific domain.

**Supervised pre-training for robotics** learns one or more pretext tasks through strong supervision and then transfers the representations to downstream robotic tasks. Yen-Chen et al. [27] shows that representations learned from semantic tasks such as detection and segmentation correlates with affordance maps for object manipulation. Shridhar et al. [28] uses language-supervised CLIP model [29] for learning language-conditioned imitation policy. In concurrent work, Nair et al. [30] explore pre-training visual representations using time contrastive learning and language descriptions from human annotators. These methods all require expert labels or cross-domain supervision.

**Self-supervised learning in robotics** has been explored to improve sample efficiency. Pinto and Gupta [31] collects a large-scale dataset through self-supervision for supervising a CNN policy. Sermanet et al. [32] learns representations from videos through a metric loss that maximizes the similarity of temporally close frames over far-away frames. Pari et al. [33] demonstrates that it is feasible to construct a control policy by non-parametric nearest-neighbor retrieval with pre-trained visual representations. These approaches use context and/or task-specific data rather than in-the-wild images and videos. It has recently been shown that MAE on real images acts as a powerful

pre-training mechanism for continuous control RL in simulation [9], however, whether and how this extends to the low-data, real-world setups has not been shown yet.

**Self-supervised learning in vision** has recently focused on techniques that rely on a set of pre-defined data augmentations. These approaches either learn pretext tasks that distinguish views from the augmentations [34, 35, 36, 37], or learn to be invariant to the augmentations by modeling similarities [38, 39, 40, 41]. These approaches implicitly presume downstream task invariances, as pre-defining augmentations introduces inductive biases that may be beneficial for some tasks but harmful for others, and thus in turn limits the scalability of models. Masked autoencoding [42, 15] overcomes this issue and has shown superior performance on visual recognition tasks.

# 3 Real World Robot Learning with Masked Visual Pre-training

We describe our approach for real world robot learning with masked visual pre-traning.

## 3.1 Diverse Large-scale Data

We first compile a large-scale dataset for learning visual representations. We primarily use Ego4D [14], a massive scale, egocentric dataset from 9 countries using portable devices, covering over 3,670 hours of daily-life activities. We combine the Ego4D data with ImageNet [10], as well as the Hand-object Interaction (HoI) data used in [9], which comprises of the egocentric Epic Kitchens [13] dataset, the YouTube 100 Days of Hands dataset [12], and the crowd-sourced Something-Something dataset [11]. Our training data totals 4.5 million frames, 6.5x of the HoI data. We find that a sufficiently large pre-training dataset to perform the mask image modeling self-supervisory task is critical to scale up the vision backbone for real robot tasks.

## 3.2 Masked Visual Pre-training

At the core of our self-supervised visual representation learning approach is masked image modeling via the masked autoencoders (MAE) [15]. MAE randomly masks out patches in an image and reconstructs the missing pixels with a vision transformer (ViT) [8]. A high masking ratio, e.g., 75%, and asymmetrical heavy-encoder light-decoder design, are important for learning good visual representations. Simple and free from dataset or task-specific augmentations [43], MAE is the state-of-the-art self-supervised framework in computer vision [44, 45, 46, 47], and has been demonstrated to work well for motor control tasks in simulation [9].

Despite ViT models trained via the MAE framework yield improving performance in vision tasks as model sizes grow [8, 15, 48], previous work [9] does not show improvement from switching a ViT-Small model to the ViT-Base counterpart of 4x as many parameters. In this study, we scale up to the ViT-Large model and deploy it on the real robot. The model contains 307M parameters and runs at $\sim$64 gigaflops at input size 224$\times$224, approximately 15x as many as a commonly adopted ResNet-50 [49], the largest vision model ever deployed for robotics. As we will show in the experimental section, large models yield consistently better performance on downstream robotic tasks.

## 3.3 Efficient Robot Learning

We learn to perform real robot tasks through behavior cloning. We collect demos containing trajectories of RGB images from a wrist-mounted camera and the robot's joint state at each time step. We primarily use the motion-tracked HTC Vive VR system to control the end-effector, or kinematics teaching mode of the robot arm for tasks whose trajectories are difficult to be specified by the motion controller, e.g., closing a door. We train a control policy that takes in the input image features and prioperceptive states (joint positions) at time step $t$ and predicts the action at time step $t + 1$. We perform joint position control; we do not use any end-effector information. We adopt the MVP pipeline [9] in which image encoder is frozen throughout the policy learning, which prevents large pre-trained encoders from overfitting to a specific setting or task, and greatly reduces GPU memory footprint and training time.

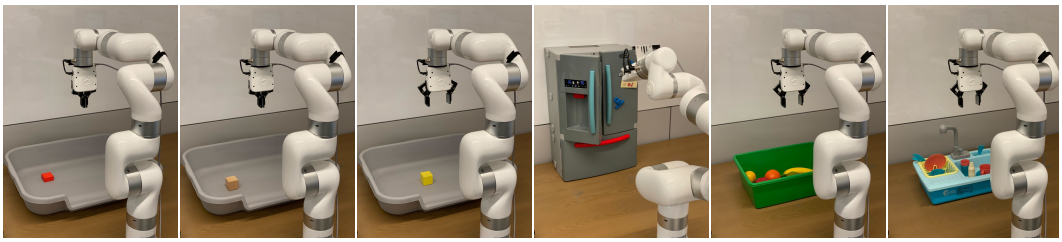

Figure 2: **Real-world robotic tasks.** We perform extensive real robot evaluations using a 7 DoF robot arm with a parallel jaw gripper. Our tasks include basic motor control skills (reaching a red block, pushing a wooden cube, and picking a yellow cube), variations in scenes (closing a fridge), objects (picking fruits), and scenes and objects (picking a detergent bottle from a cluttered sink).

# 4 Experimental Setup

**Data.** We extract frames from Ego4D, Epic Kitchens, and Something-Something at 0.2 fps, 1fps and 0.3fps, respectively. We then combine the Ego4d with ImageNet and the YouTube 100 Days of Hands dataset. This process yields 2.6M frames from Ego4D, 1.2M images from ImageNet, and 700k HoI images from the rest, a total of 4.5M images. We term the combined dataset as "Ego" for abbreviation. Note that [9] only uses the 700k HoI images, excluding the Ego4D and ImageNet.

**Encoders.** We use the standard Vision Transformer (ViT) architecture as the image encoder. We use three models of various sizes: ViT-Small [50], ViT-Base, and ViT-Large models, the smallest among which is approximately the same size as the ResNet-50 model, while the largest among which is ∼15x as many as the ResNet-50 model.

We pre-train the models via the MAE framework [15]. The training recipe closely follows [15], with dataset specific settings from [9]. We use the auxiliary dummy classification token in the MAE for transferring to downstream robotics tasks. We train the MAE models for 400 epochs for the combined Ego dataset; 1600 epochs for the HOI dataset; and 1600 epochs for ImageNet dataset. We use the pre-training recipe in [51] for the study that involves ImageNet supervised models.

**Controllers.** The controller takes in both image features and proprioceptive state of the robot. We use joint positions as the proprioceptive state *without* explicitly appending the end-effector pose to the state. The controller outputs *delta joint angles*. The architecture of the controller closely follows [52], i.e., a simple four-layer MLP with a SeLU [53] activation following each hidden layer. The hidden size is [256, 128, 64] for most tasks and [512, 256, 128] for the `PickFromSink` task. We linearly project the image features and the proprioceptive states to a joint embedding space before feeding them into the controller.

**Robot.** We use the low-cost UFACTORY xArm 7 robot with a 7-DoF arm and a 1-DoF parallel jaw gripper. We use the arm's maximum control frequency of 5 Hz for both demo collection and control. We use a first-person wrist mounted RealSense camera for all tasks. We do *not* use depth information from the camera.

**Tasks.** We consider basic motor control tasks, i.e., `ReachBlock`, `PushCube`, and `PickCube`; more challenging in-context tasks, i.e., `CloseFridgeDoor`, `PickAnyFruit`, and `PickFromSink`. More details are described in the experimental section.

**Demonstrations.** We collect 80 demos per task. We use the motion-tracked HTC Vive VR system for most tasks, except for `CloseFridgeDoor` we use kinematics teaching. We use trajectory replay on the robot to prune the collected demos. We do *not* use key frame information for the learning.

**Evaluation.** We systematically sweep across 16 variations of the environment, e.g., shifting the target object. For maximum consistency and reliability of the study, we use the exact *same* 16 variations for all models in a single task, and benchmark models *sequentially* at each variation, in order for similar lighting conditions, precise object initial locations, etc.

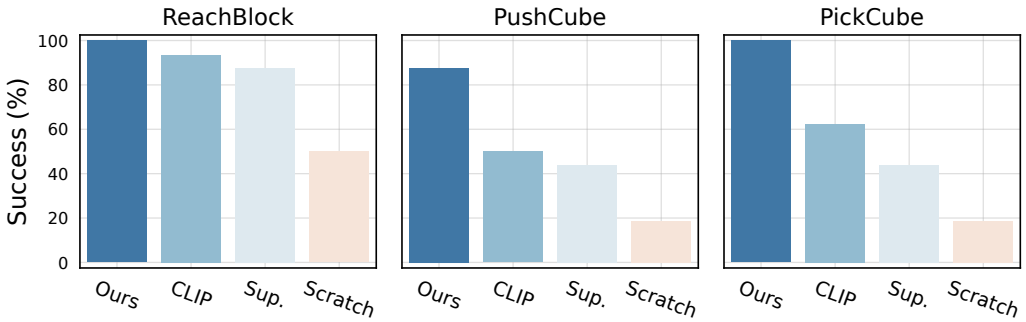


Figure 3: **Comparison to vision encoders.** We compare our approach to visual encoders trained with CLIP and supervised learning on ImageNet, and training from scratch on the task at hand. In all cases, we observe that our approach outperforms the baselines by a considerable margin.

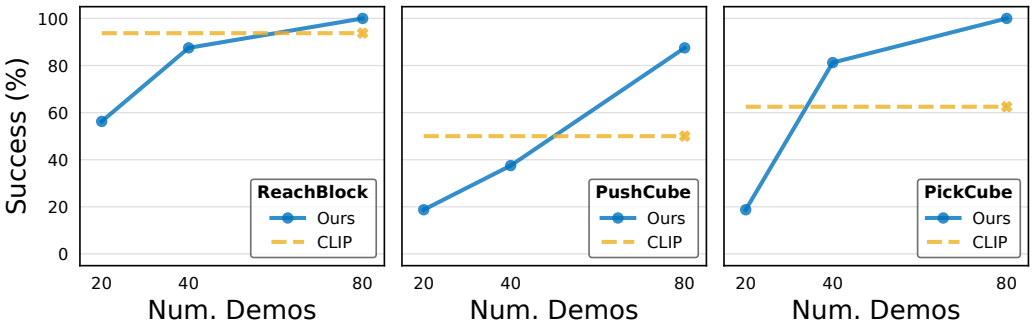


Figure 4: **Sample complexity.** We show the performance of our approach as the number of demos varies from 20 to 80. CLIP performance at 80 demos is shown with a dashed lined for reference. We observe that our approach is comparable to CLIP with using only half the number of demos.

# 5 Experiments

We perform extensive evaluation across a range of visual backbones, real-world robotic tasks, objects, and environments (Figure 2). In total, we report results from 672 real-world experiments.

## 5.1 Basic Motor Control

We begin by evaluating basic motor control tasks in visually simple tabletop contexts. These tasks serve as a foundation for more complex tasks and the visual representations that can afford to learn them efficiently is necessary for robotics. Specifically, we consider three tasks: reaching a red block, pushing a wooden cube, and picking a yellow cube (see Figure 2 for visuals). In all cases, we randomize the initial object and robot positions (see Supplement).

**Comparison to vision encoders.** In Figure 3 we compare our approach to the state of the art vision backbones: CLIP [3] trained on 400M text-image pairs, supervised model trained on ImageNet, and a model trained from scratch with in-domain demonstration data. For fair comparisons, we use the ViT-Base [8] vision encoder for all methods. We observe that our approach consistently outperforms the baselines by a considerable margin. Furthermore, we see that the strong CLIP encoder is the most competitive model. For the remaining experiments in the paper, we compare with CLIP.

**Sample complexity.** In Figure 4 we show the performance of our approach as the number of demos varies from 20 to 80. We observe that in aggregate our approach reaches the performance of CLIP with 80 demos (dashed line) when using 50% fewer demos. This is a promising signal for using our approach for solving more complex robotic tasks, as discussed next.

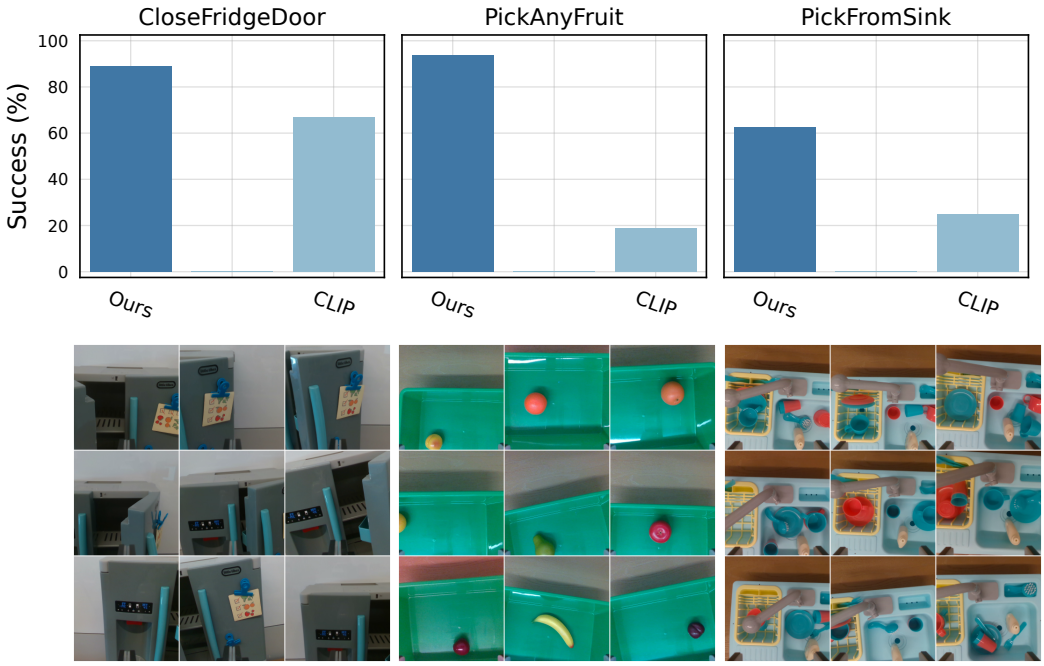


Figure 5: **Variations in scenes and objects.** We compare our approach to CLIP on tasks with variations in scenes (closing the fridge), objects (picking fruits), and scenes and objects (picking an object from a cluttered sink). The models are ViT-Base. Our approach considerably outperforms CLIP and the gap is larger than in simpler settings (see Figure 3). This may suggest that our representations capture more precise spatial structure that is helpful for robotic tasks in more realistic contexts.

### 5.2 Visually Diverse Scenes and Objects

In the previous section, we have seen that our approach enables learning of basic motor control tasks, such as reaching, pushing, and picking, with higher success rate and better sample complexity than the alternate vision backbones. To focus on the motor control aspect, we used a visually simple environment and basic objects. However, one of the main potential benefits of our approach is that learning visual representations from real-world, diverse data may enable solving robotic tasks that involve interaction with everyday objects in visually complex environments. In this section, we evaluate our visual representations on tasks with variations in scenes and objects.

**Everyday scenes.** We first evaluate our approach on more realistic scenes. We consider the common household task of closing a fridge door. As shown in Figure 5, bottom-left, the initial configuration of the fridge and the robot vary considerably, which is quite common in everyday settings. Figure 5, top-left, we see that our approach outperforms CLIP.

**Different objects.** Next, we evaluate our approach on a task with a variety of objects. In particular, we consider grasping 8 different fruits that vary in color, shape, and size. In each trial a fruit is selected at random, and both the fruit and the robot positions are randomized. In Figure 5, middle, we show the results (top) and example starting configurations (bottom). We see that the CLIP encoder struggles in this setting while our approach achieves nearly perfect score.

**Objects in context.** Finally, we evaluate our approach on a task that features interacting with objects in everyday contexts. Specifically, we task the robot with picking a detergent bottle from a cluttered sink, shown in Figure 5 right. The initial configuration of the scene varies considerably making this a challenging task. Like before, we observe that our approach considerably outperforms CLIP using the same ViT-B encoder (see also the next section).

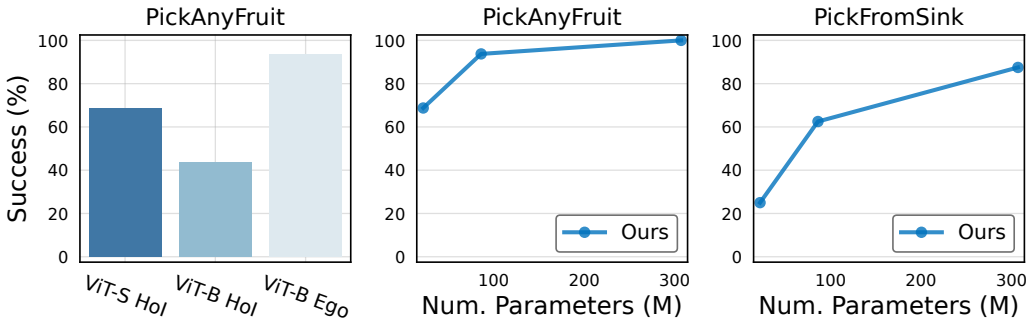


Figure 6: **Scaling model and data.** We study the scaling properties of our approach. We observe that scaling the model size alone from ViT-S to ViT-B while keeping the dataset fixed (HoI image collection; see text for details) does not improve the performance and even hurts (left). However, when we scale both the model and data (our massive Ego image collection; see text for details) we see clear benefits from a larger model. The trend continues when going further from 86 M ViT-B to 307 M ViT-L model (middle & right). Moreover, the gains are larger for harder tasks (right).

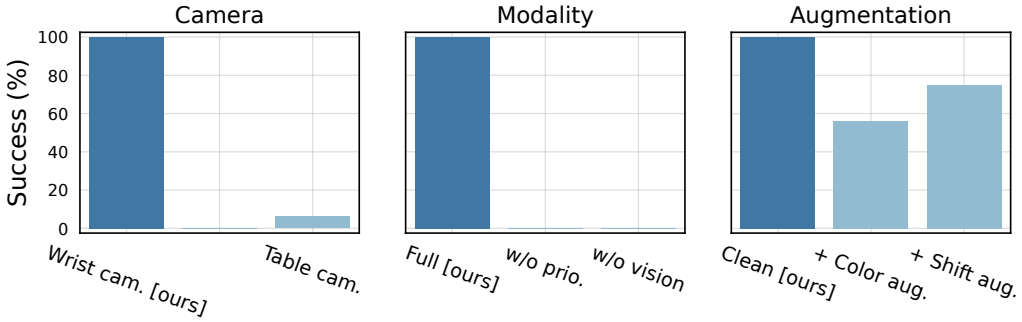


Figure 7: **Ablation studies.** We observe that using wrist-mounted camera works better than third-person camera (left), that using both vision and proprioception is required (middle), and that we do not get benefits from using common data augmentations with frozen vision encoders (right).

### 5.3 Scaling Data and Model Size

An important property of our visual pre-training approach is that uses a self-supervised objective [15] that makes minimal assumptions about the data distribution and does not rely on human-designed proxy tasks, like data augmentations. Therefore the framework is well suited for pre-training from massive collections of unlabeled data. Here we study scaling model and data size.

We first consider increasing the model capacity. In Figure 6, left, we see that increasing the model size from ViT-S to ViT-B, while keeping the data size fixed (HOI image collection from [9]) does not increase performance and even hurts. This is consistent with the simulation results reported in [9]. However, if we also scale the data size from HOI to our massive Ego data collection we see that ViT-B brings considerable gains. This suggests that we must scale both the model and the data.

In Figure 6, middle & right, we show the performance as a function of model size. We see that additionally increasing the model size from 86M parameter ViT-B to 307M parameter ViT-L results in further improvements. We also see that the gains are larger for harder tasks (pick from sink). To the best of our knowledge, ours is the largest vision model deployed to real robot tasks, and clearly demonstrates the benefits of scaling visual pre-training for real world robot learning.

### 5.4 Ablation Studies

We conduct ablation studies on basic environment, model and training settings. We start by comparing using the first-person wrist camera with the third-person table camera (Figure 7, left). The wrist mounted camera yields significantly better results than the third-person camera. During the evalu-

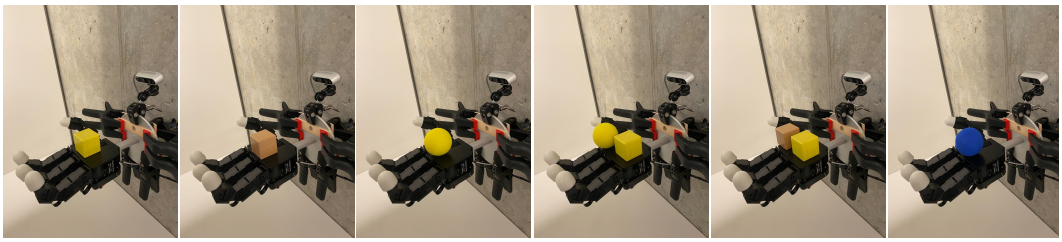

Figure 8: **Multi-finger hand experiments.** We show that our approach readily generalizes to a real robot with different morphology. In particular, we consider a finger reaching task with a multi-finger Allegro hand. We further use this setting as a vehicle to study representations learnt by our model and observe that our representations disentangle shape and color information (see text for details).

ation trials we observe that the third-person camera model struggles with fine-grained localization, and the farther the object is from the camera, the worse the predicted trajectory is. This suggests that the model is confused by the scale ambiguity in the setting because we do not use depth. Next, we experiment with removing proprioception states or images from the model's input (Figure 7, middle); both yield zero success rate. In the third ablation we add commonly used data augmentation for training the task policy (Figure 7, right). We do not observe any benefits from these augmentations likely since our vision encoder is frozen and overfits significantly less during the training.

### 5.5 Case Study: Multi-finger Hand

Our pre-training approach makes no assumptions about the downstream robotic tasks or embodiments. Furthermore, our policies predict joint angle positions and make no domain-specific assumptions in the action space. In this section we test the generality of our approach by using it for a downstream task with a multi-finger Allegro hand (please see the Supplement for more details on the setup). We consider a reaching task where the goal is to place the index finger on top of the yellow cube (Figure 8, left). The position of the cube is randomized across the palm of the hand.

A nice property of training controllers on top of frozen visual representations is that it enables us to perform studies to understand what the pre-trained visual representations are capturing. Specifically, we first train the policy to reach a yellow cube. Then, at test time, we present it with objects of different shape and color (Figure 8, columns 2-6). First, we find that when given the same shape of different color (wooden cube) or same color and different shape (yellow ball) it reaches for the object. Next, when given both the yellow cube and a distractor it reaches for the yellow cube. Finally, when given an object of different shape and color (blue cube) the hand stays still. Overall, these result suggest that our visual representations are able to disentangle between shape and color.

## 6 Discussion

**Conclusion.** We explore learning visual representations from a massive collection of real-world data and using them for downstream robotic tasks. We pre-train representations with masked modeling, freeze the encoder, and learn control policies on top. We perform an extensive evaluation in the real world and show that, across various robotic tasks, our approach leads to higher success rate and better sample complexity than CLIP, supervised ImageNet pre-training, and training from scratch. We further demonstrate the benefits of scaling the model and data size for real world robot learning.

**Limitations.** Our study of different visual backbones and design choices is performed entirely in the real world. We find that there are a number of factors that may influence the results and we do our best to control for them (*e.g.*, object position, environment conditions, robot up time, *etc.*; please see Supplement for details). We believe that coming up with a strict evaluation methodology and a set of guidelines would be beneficial for our future progress, especially as we move toward developing, benchmarking, and widely deploying pre-trained models in real world robotics.

# References


[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 2012.

[2] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.

[3] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. Improving language understanding by generative pre-training. 2018.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HCT*, 2019.

[5] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *NeurIPS*, 2020.

[6] A. Van Den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu. Wavenet: A generative model for raw audio. *SSW*, 2016.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

[8] 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. In *ICLR*, 2020.

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

[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[11] R. Goyal, S. Ebrahimi Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Fruend, P. Yianilos, M. Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *ICCV*, 2017.

[12] D. Shan, J. Geng, M. Shu, and D. F. Fouhey. Understanding human hands in contact at internet scale. In *CVPR*, 2020.

[13] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *ECCV*, 2018.

[14] 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. *arXiv:2110.07058*, 2021.

[15] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked autoencoders are scalable vision learners. *arXiv:2111.06377*, 2021.

[16] D. Damen, H. Doughty, G. M. Farinella, , A. Furnari, J. Ma, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, and M. Wray. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *IJCV*, 2021.

[17] S. Levine, C. Finn, T. Darrell, and P. Abbeel. End-to-end training of deep visuomotor policies. *JMLR*, 2016.

[18] S. James and E. Johns. 3D simulation for robot arm control with deep Q-learning. *Conference on Neural Information Processing Systems Workshop (Deep Learning for Action and Interaction)*, 2016.

[19] S. James, A. J. Davison, and E. Johns. Transferring end-to-end visuomotor control from simulation to real world for a multi-stage task. *CoRL*, 2017.

[20] K. Wada, E. Sucar, S. James, D. Lenton, and A. J. Davison. MoreFusion: Multi-object reasoning for 6D pose estimation from volumetric fusion. In *CVPR*, pages 14540–14549, 2020.

[21] A. T. Miller, S. Knoop, H. I. Christensen, and P. K. Allen. Automatic grasp planning using shape primitives. In *2003 IEEE International Conference on Robotics and Automation (Cat. No. 03CH37422)*, volume 2, pages 1824–1829. IEEE, 2003.

[22] J. J. Kuffner and S. M. LaValle. Rrt-connect: An efficient approach to single-query path planning. In *Proceedings 2000 ICRA. Millennium Conference. IEEE International Conference on Robotics and Automation. Symposia Proceedings (Cat. No. 00CH37065)*, volume 2, pages 995–1001. IEEE, 2000.

[23] K. Wada, S. James, and A. J. Davison. ReorientBot: Learning object reorientation for specific-posed placement. *ICRA*, 2022.

[24] S. James and A. J. Davison. Q-attention: Enabling Efficient Learning for Vision-based Robotic Manipulation. *RA-L*, 2022.

[25] S. James, K. Wada, T. Laidlow, and A. J. Davison. Coarse-to-Fine Q-attention: Efficient Learning for Visual Robotic Manipulation via Discretisation. *CVPR*, 2022.

[26] A. Zeng, P. Florence, J. Tompson, S. Welker, J. Chien, M. Attarian, T. Armstrong, I. Krasin, D. Duong, V. Sindhwani, et al. Transporter networks: Rearranging the visual world for robotic manipulation. *CoRL*, 2020.

[27] L. Yen-Chen, A. Zeng, S. Song, P. Isola, and T.-Y. Lin. Learning to see before learning to act: Visual pre-training for manipulation. In *ICRA*, 2020.

[28] M. Shridhar, L. Manuelli, and D. Fox. Cliport: What and where pathways for robotic manipulation. In *CoRL*, 2022.

[29] 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 *ICML*, 2021.

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

[31] L. Pinto and A. Gupta. Supersizing self-supervision: Learning to grasp from 50k tries and 700 robot hours. In *ICRA*, 2016.

[32] P. Sermanet, C. Lynch, Y. Chebotar, J. Hsu, E. Jang, S. Schaal, S. Levine, and G. Brain. Time-contrastive networks: Self-supervised learning from video. In *ICRA*, 2018.

[33] J. Pari, N. Muhammad, S. P. Arunachalam, L. Pinto, et al. The surprising effectiveness of representation learning for visual imitation. *arXiv:2112.01511*, 2021.

[34] X. Wang and A. Gupta. Unsupervised learning of visual representations using videos. In *ICCV*, 2015.

[35] C. Doersch, A. Gupta, and A. A. Efros. Unsupervised visual representation learning by context prediction. In *ICCV*, 2015.

[36] R. Zhang, P. Isola, and A. A. Efros. Colorful image colorization. In *ECCV*, 2016.

[37] N. Komodakis and S. Gidaris. Unsupervised representation learning by predicting image rotations. In *ICLR*, 2018.

[38] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, 2018.

[39] A. v. d. Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding. *arXiv:1807.03748*, 2018.

[40] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020.

[41] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020.

[42] H. Bao, L. Dong, and F. Wei. Beit: Bert pre-training of image transformers. In *ICLR*, 2022.

[43] T. Xiao, X. Wang, A. A. Efros, and T. Darrell. What should not be contrastive in contrastive learning. In *ICLR*, 2021.

[44] Y. Li, S. Xie, X. Chen, P. Dollar, K. He, and R. Girshick. Benchmarking detection transfer learning with vision transformers. *arXiv preprint arXiv:2111.11429*, 2021.

[45] Z. Tong, Y. Song, J. Wang, and L. Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *arXiv:2203.12602*, 2022.

[46] C. Feichtenhofer, H. Fan, Y. Li, and K. He. Masked autoencoders as spatiotemporal learners. *arXiv preprint arXiv:2205.09113*, 2022.

[47] R. Bachmann, D. Mizrahi, A. Atanov, and A. Zamir. Multimae: Multi-modal multi-task masked autoencoders. *arXiv preprint arXiv:2204.01678*, 2022.

[48] X. Zhai, A. Kolesnikov, N. Houlsby, and L. Beyer. Scaling vision transformers. In *CVPR*, 2022.

[49] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[50] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou. Training data-efficient image transformers & distillation through attention. In *ICML*, 2021.

[51] T. Xiao, P. Dollar, M. Singh, E. Mintun, T. Darrell, and R. Girshick. Early convolutions help transformers see better. In *NeurIPS*, 2021.

[52] V. Makoviychuk, L. Wawrzyniak, Y. Guo, M. Lu, K. Storey, M. Macklin, D. Hoeller, N. Rudin, A. Allshire, A. Handa, et al. Isaac gym: High performance gpu-based physics simulation for robot learning. In *NeurIPS*, 2021.

[53] G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter. Self-normalizing neural networks. In *NeurIPS*, 2017.