R3M: A Universal Visual Representation for Robot Manipulation

Presenter: Cole Smylie

Paper: November 18, 2022
Presentation: September 5, 2023

Main Problem

Standard Robot Learning Procedure - Inefficient, High Human Cost

Collect MANY Demonstrations → Train Policy From Images → Deploy Policy with Image Observations

Reusable Representation Robot Learning Procedure - Efficient, Low Human Cost

Download Pre-Trained Representation → Collect FEW Demonstrations → Train Policy From Pre-Trained Representation → Deploy Policy with Representation

Motivation

Reusable, Pre-Trained Representations

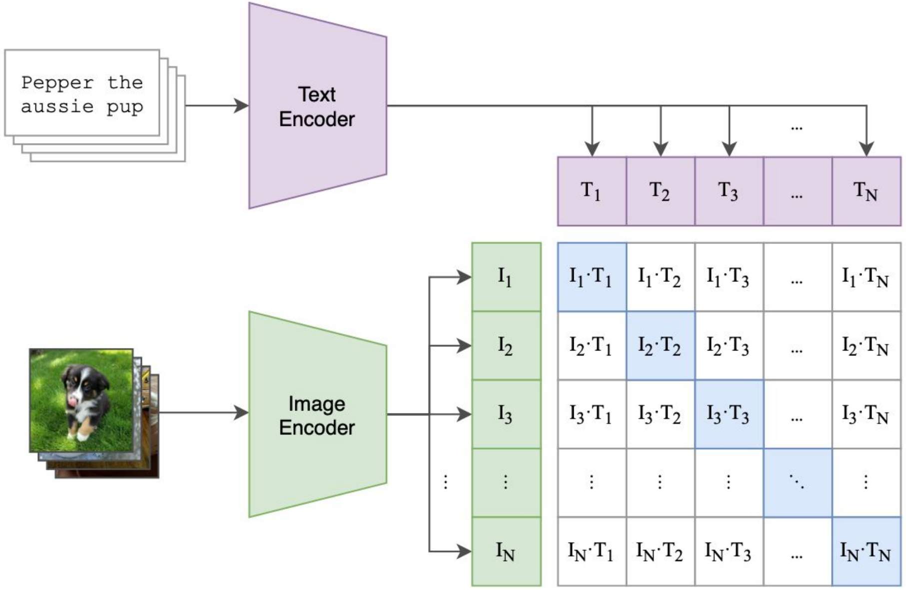
- Lower Data Requirement
- Faster Training
- Generalization

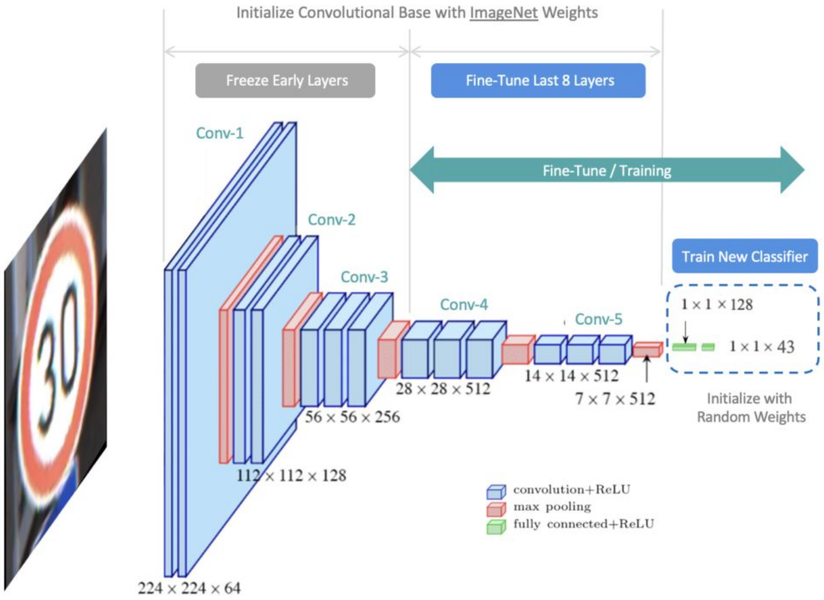
Motivation

Visual Representations

CLIP

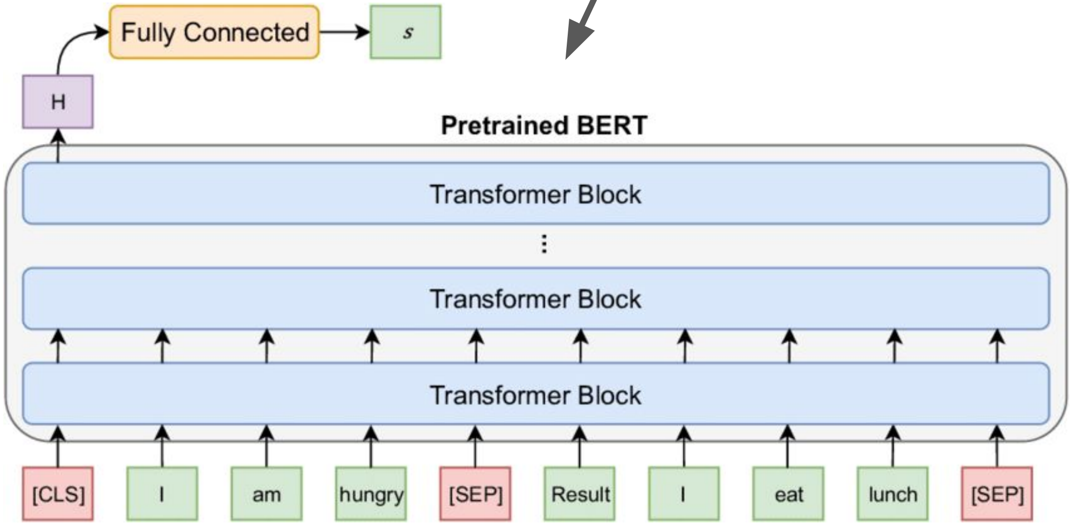
ImageNet





Motivation

Language Representations



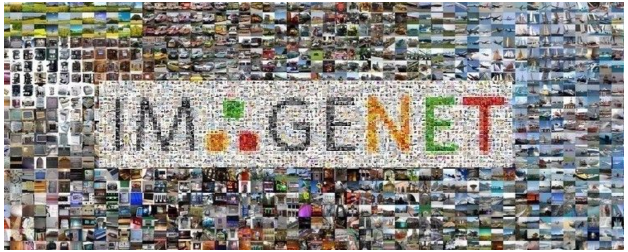
Robotics

Main Problem

Requires Data thats

- Scalable - cheaply or self supervised objectives
- Diverse

Visual Representations



Language Representations

Robotics

Key Insights

Access to videos of humans interacting in semantically interesting ways with their environments - Ego4D

- Large and diverse, spanning scenes across the globe
- Labeled, tasks ranging from folding clothes to cooking a meal

Problem Setting

Can visual representations pre-trained on diverse human videos enable efficient downstream learning of robotic manipulation skills?

- **(1) Temporal Dynamics** - the agent will be sequentially interacting in the environment to accomplish tasks
- **(2) Language** - it should capture semantically relevant features like objects and their relationships
- **(3) Sparsity** - it should be compact, not include features irrelevant to the above criteria

Problem Setting

Input:
dataset D of N videos
(RGB frames $[I_0, I_1, ..., I_T]$)
natural language
description l

→

Single Image Encoder
$F_\phi(I)$

→

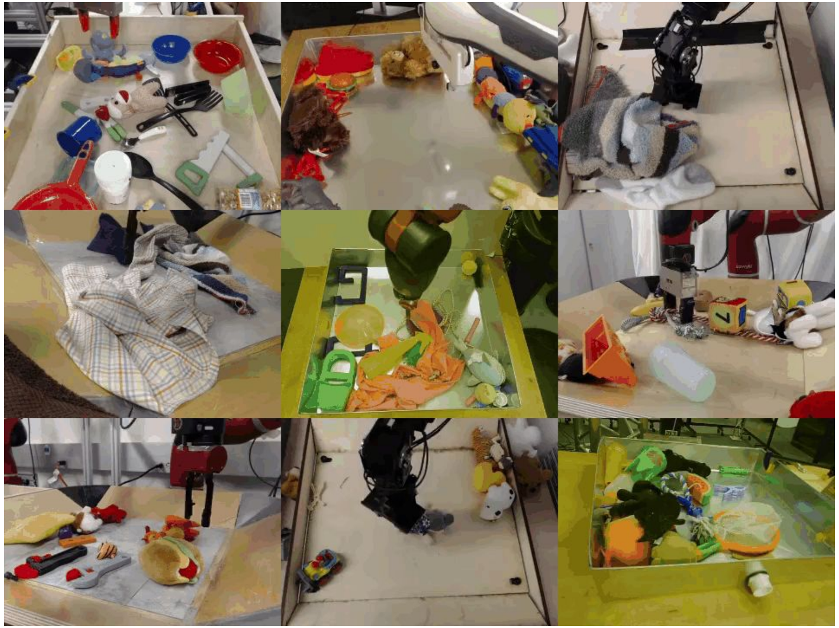
Language Prediction
Module
$G_\theta\,(F_\phi(I_0), F_\phi(I_i), l)$

ϕ and θ trained to optimize Loss Functions:

- L_{tcn}
- L_{lang}
- L_{reg}

Prior Approaches

RoboNet

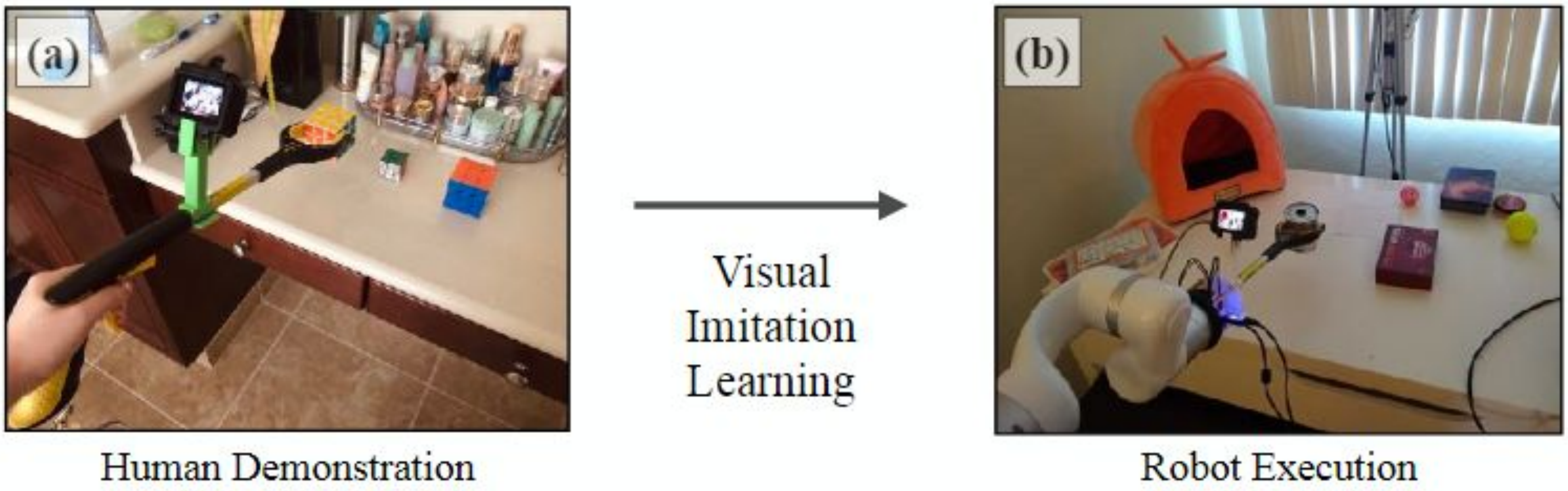


S. Dasari, F. Ebert, S. Tian, S. Nair, B. Bucher, K. Schmeckpeper, S. Singh, S. Levine, and C. Finn. Robonet: Large-scale multi-robot learning. In Conference on Robot Learning, 2019.

RoboTurk

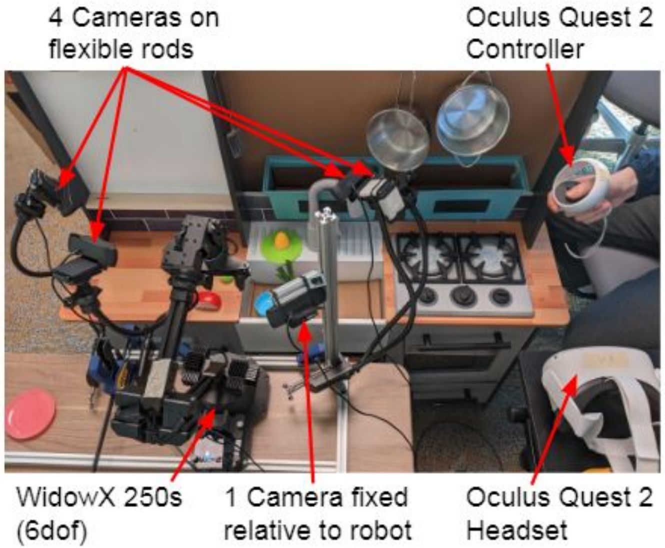
A. Mandlekar, J. Booher, M. Spero, A. Tung, A. Gupta, Y. Zhu, A. Garg, S. Savarese, and L. Fei-Fei. Scaling robot supervision to hundreds of hours with roboturk: Robotic manipulation dataset through human reasoning and dexterity. In 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages 1048–1055. IEEE, 2019.

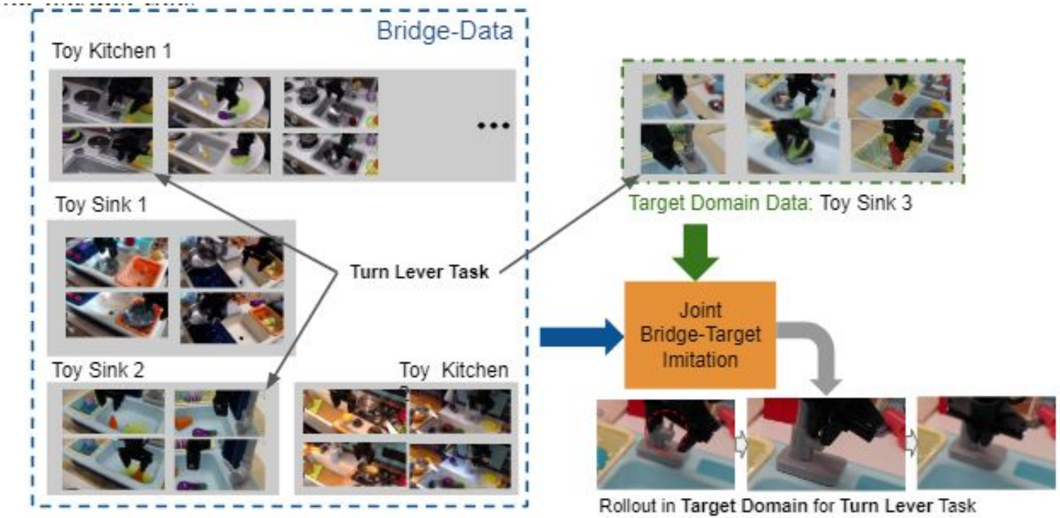
Prior Approaches



S. Young, D. Gandhi, S. Tulsiani, A. Gupta, P. Abbeel, and L. Pinto. Visual imitation made easy. In CoRL, 2020.

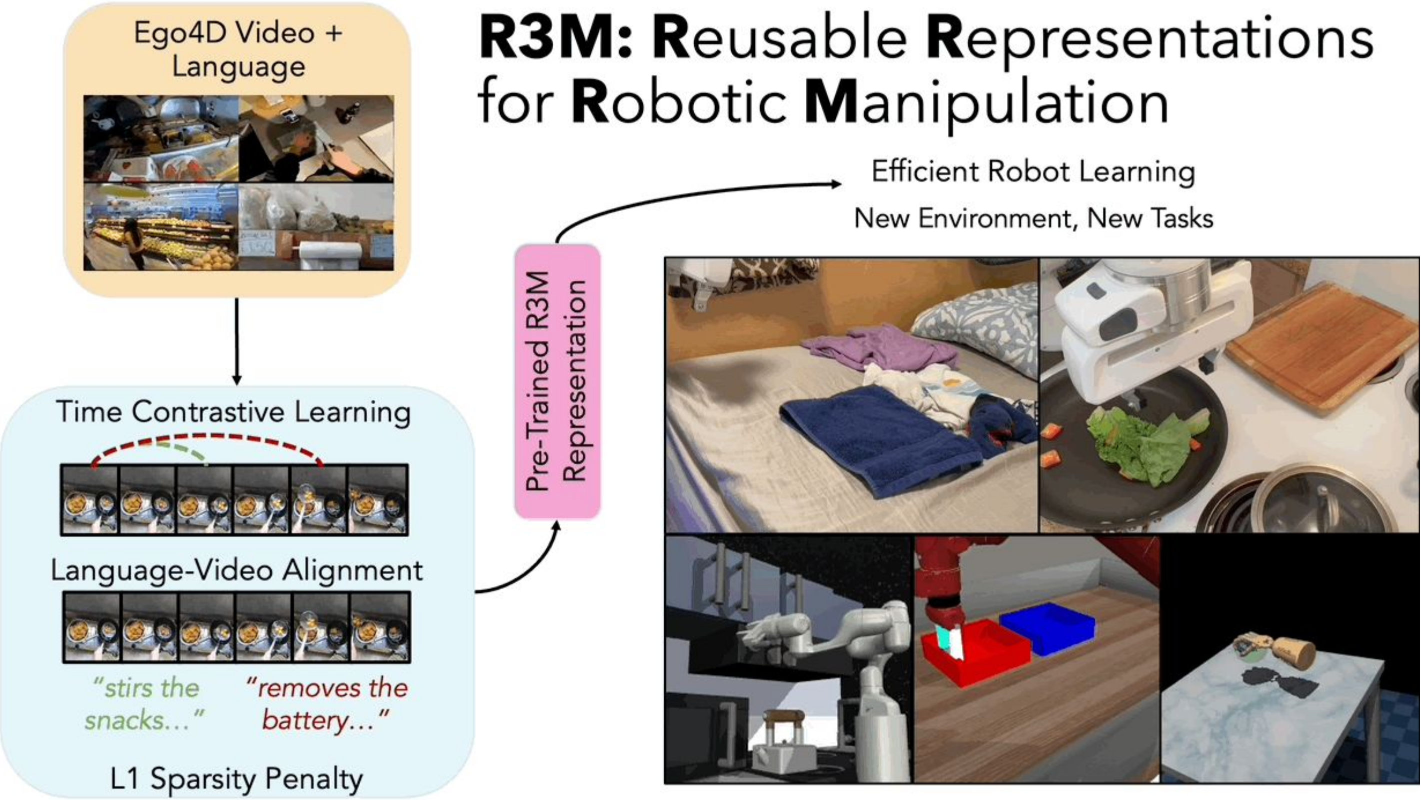
Prior Approaches





F. Ebert, Y. Yang, K. Schmeckpeper, B. Bucher, G. Georgakis, K. Daniilidis, C. Finn, and S. Levine. Bridge data: Boosting generalization of robotic skills with cross-domain datasets.

Proposed Approach



Time Contrastive Learning Algorithm

- Encourages F_ϕ to capture features relevant to physical interaction and sequential decision making
- Produce a representation such that the distance between images closer in time is smaller than for images farther in time or from different video

$$\mathcal{L}_{tcn} = -\sum_{b \in B} \log \frac{e^{\mathcal{S}(z_i^b, z_j^b)}}{e^{\mathcal{S}(z_i^b, z_j^b)} + e^{\mathcal{S}(z_i^b, z_k^b)} + e^{\mathcal{S}(z_i^b, z_i^{\neq b})}}$$

where $z = \mathcal{F}_\phi(I)$, and $z_i^{\neq b}$ is a negative example sampled from a *different video* in the batch. $\mathcal{S}$ denotes a measure of similarity, which in our case is implemented as the negative L2 distance.

Video-Language Alignment Algorithm

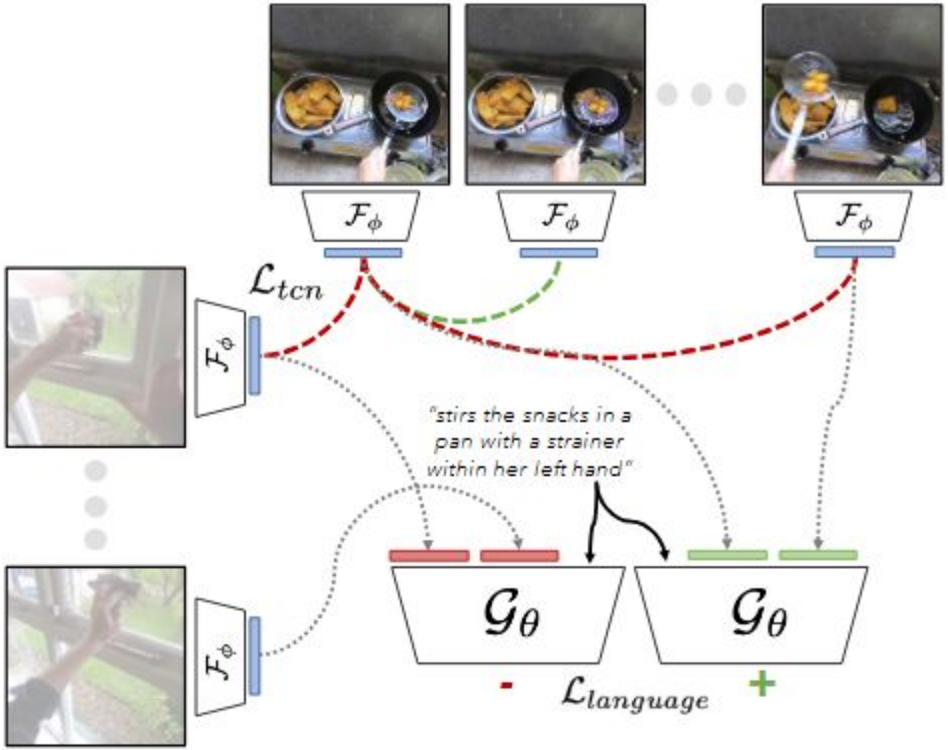
- Encourage F_ϕ to capture semantically relevant features, train a language prediction module from the embedding outputted by F_ϕ
- By capturing features predictive of language, like “putting the apple on the plate”, the learned representation should capture semantically relevant parts of the scene like the plate and apple state
- Train a model $G_\theta(F_\phi(I_0), F_\phi(I_i), l)$ that takes in an initial image I_0, a future image I_i, language l and outputs a score corresponding to if transitioning from I_0 to I_i completes the language l

$$\mathcal{L}_{language} = -\sum_{b \in B} \log \frac{e^{\mathcal{G}_\theta(z_0^b, z_{j>i}^b, l^b)}}{e^{\mathcal{G}_\theta(z_0^b, z_{j>i}^b, l^b)} + e^{\mathcal{G}_\theta(z_0^b, z_i^b, l^b)} + e^{\mathcal{G}_\theta(z_0^{\neq b}, z_{j>i}^{\neq b}, l^b)}}$$

where again $z = \mathcal{F}_\phi(I)$, and $z^{\neq b}$ is a negative example sampled from a *different video* in the batch (that does not match the language instruction l^b).

Regularization Algorithm

- Sparse and compact representations benefit control, particularly in low data imitation learning [68]
- Reducing the effective dimensionality of the state space (implemented as a simple L1 and L2 penalty) can help mitigate State-distribution shift

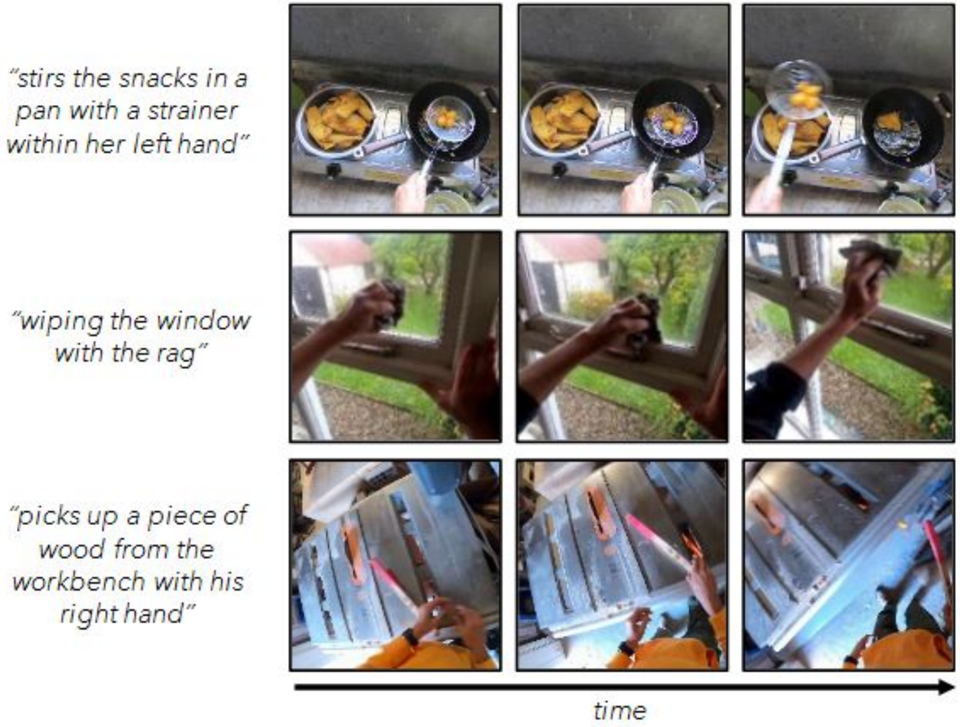


$$\mathcal{L}(\phi,\theta) = \mathbb{E}_{I^{1:B}_{0,i,j,k}\sim\mathcal{D}}[\lambda_1\mathcal{L}_{tcn} + \lambda_2\mathcal{L}_{language} + \lambda_3||\mathcal{F}_\phi(I_i)||_1 + \lambda_4||\mathcal{F}_\phi(I_i)||_2]$$

Experimental Setup - Dataset

Ego4D

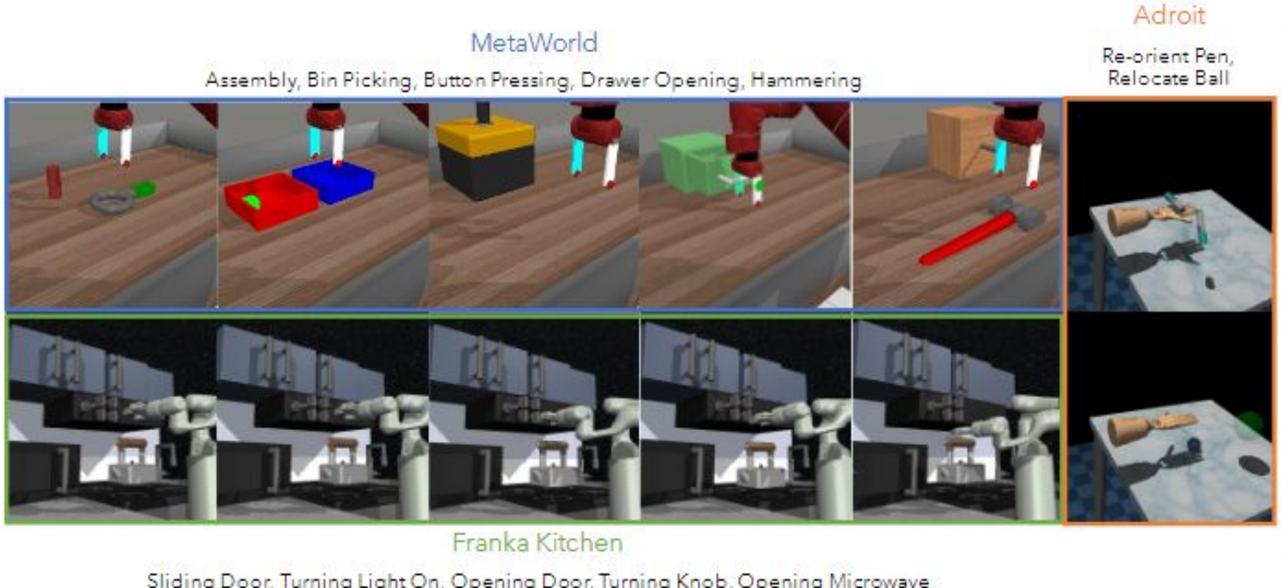
- Videos of people engaging in a wide range of tasks from cooking to socializing to assembling objects
- More than 70 locations across the globe
- More than 3500 hours of data
- A natural language annotation describing the behavior of the person in the video

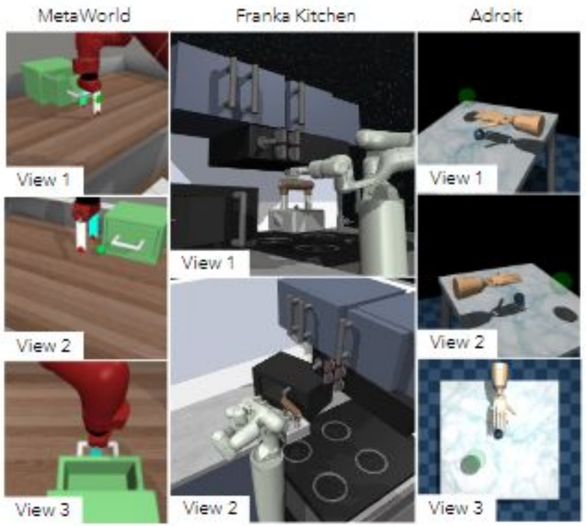


Experimental Setup - Domains

Three robot manipulation domains

- **MetaWorld** - assembling a ring onto a peg, picking and placing a block between bins, pushing a button, opening a drawer, and hammering a nail
- **Franka Kitchen** - sliding the right door open, opening the left door, turning on the light, turning the stove top knob, and opening the microwave
- **Adroit** - reorienting the pen to the specified position, and picking and moving the ball to specified position





Experimental Setup - Domains

Real World Cluttered Environments

- Franka Emika Panda robot into a real graduate student apartment
- (1) closing a dresser drawer, (2) picking a face mask placed randomly on a desk and placing it in the dresser drawer, (3) picking up lettuce randomly placed on a cutting board and putting in a cooking pan, (4) pushing a mug to a goal

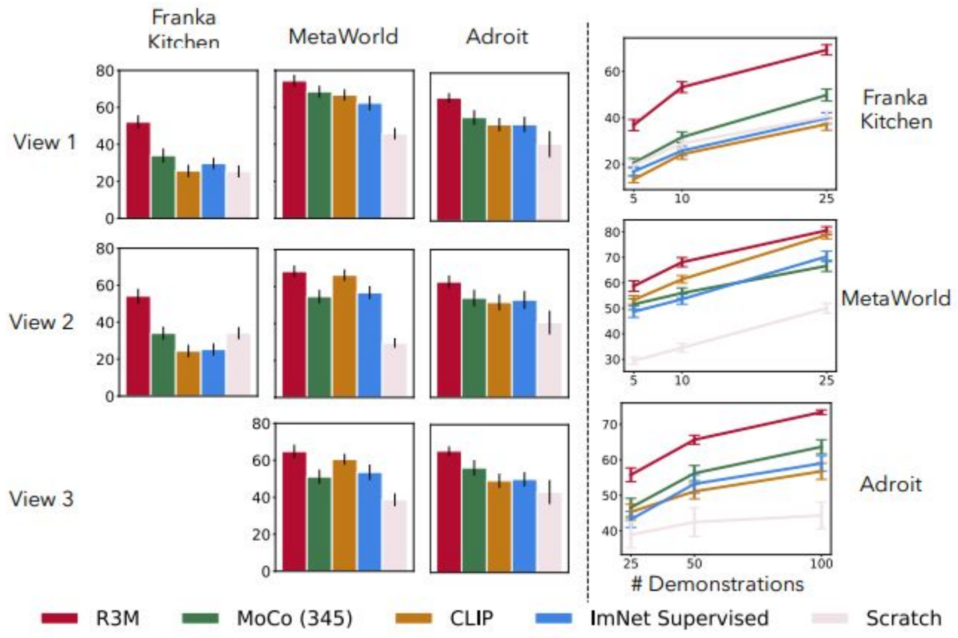
Experimental Setup - Baselines

Compared to

- CLIP which trains image representations to be aligned with paired natural language through contrastive learning and has been shown to be useful for some manipulation and navigation tasks
- ImNet Supervised which uses features pre-trained for ImageNet classification task and has been shown to be effective for reinforcement learning
- MoCo (345) (PVR) which compresses and fuses the third, fourth, and fifth convolutional layers of a ResNet-50 model trained with MoCo on ImageNet, and has been shown to be effective for imitation learning
- Learning from Scratch with default ResNet architecture
- R3M without various components
- Various other models on identical data

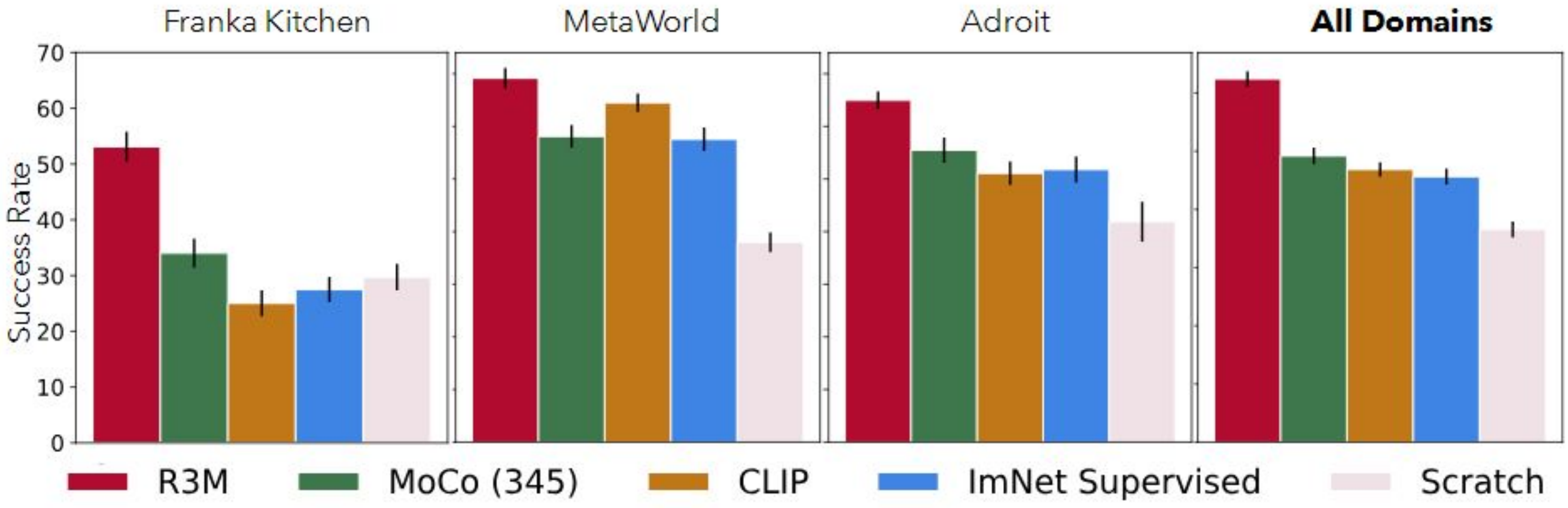
Experimental Setup - Metrics

- Evaluate visual representations as frozen perception modules for downstream policy learning with behavior cloning.
- Train the policy for 20,000 steps, evaluate it online in the environment every 1000 steps, and report the best success rate achieved.
- Final success rate reported on a task is the average over 3 seeds, viewpoints, and demo dataset sizes.



Experimental Results - Simulation

- R3M enables imitation learning with ~62% average success vs 42% for best baseline (MoCo)
- Outperforms all baselines by over 10% on average across 12 simulation tasks
- R3M is the best performing method in all 3 environments, and on 11/12 of the tasks



Experimental Results - Component Analysis

- On average across the three environments, performance drop of ≈ 2% from removing crop augmentation or from removing the L1 regularization
- Impact of removing the sparsity regularization depends on the environment - not Adroit
- Removing language loss causes biggest drop in performance (-9%) average
- Even in the fully self-supervised regime, our R3M model still outperforms prior state of the art visual representations like Image Net trained MoCo (345) (PVR) and CLIP by a significant margin

Environment		*Supervised*		*Self-Supervised*
	R3M	R3M(-Aug)	R3M(-L1)	R3M(-Lang)
Franka Kitchen	**53.1** ±2.7%	**51.1** ±2.7%	46.7 ±2.7%	47.2±2.9%
MetaWorld	**69.2** ±2.0%	**68.9** ±2.1%	65.0 ±2.4%	67.0±2.0%
Adroit	65.0 ±1.7%	61.3 ±2.1%	**66.5** ±1.6%	45.6 ±3.3%
All Domains	**62.4** ±1.3%	60.4 ±1.4%	59.4 ±1.5%	53.2 ±1.5%

Experimental Results - Data Relevance

- MoCo-Ego4D model, which uses the same data and compute asR3M, gets an average success rate ~ 10% lower than R3M in both environments
- MVP models performs ~ 20% worse than R3M
- While there is indeed a large benefit coming from diverse human video data compared to static ImageNet images (34% vs 42% on Franka), the data is not the only source of improvement, and the R3M objective provides an additional~10% boost in success rate

	Franka	Adroit
R3M	**53.1**(2.7)	**65.0** (1.7)
MoCo-Ego4D	42.0 (2.8)	54.9 (2.7)
MVP ([70])	27.0 (2.6)	51.4 (2.7)

Experimental Results - Real World

- The two perform similarly on the easier task of closing the drawer
- R3M consistently performs better on the other four tasks
  - which require more precise visual representations, yielding nearly double the success rate on average

Success out of 10 trials	R3M	CLIP
Closing Drawer	**80%**	70%
Putting Mask in Dresser	**30%**	10%
Putting Lettuce in Pan	**60%**	0%
Pushing Mug to Goal	**70%**	40%
Folding Towel	**40%**	0%
Average	**56%**	24%

Discussion of Results

Ultimately, conclude that pre-trained visual representations are essential to good performance in the low-data imitation learning regime, and using R3M with diverse human video data is especially effective for learning representations useful for robotic manipulation

Result Insights

- Unsurprisingly, learning from scratch performs poorly in this low-data regime
- Suspect the negative effect of sparsity is partly due to the Adroit environment using more demonstrations, mitigating the state distribution shift issue
- Language alignment plays an important role in better capturing semantic features that might be predictive of objects and useful for object manipulation
- Pre-training representations on video data outperforms static images

Putting Lettuce in Pan

Closing Drawer

Folding Towel

Results

Strengths

- State-of-the-art results on 12 simulation tasks and 5 real robot tasks
- Positive ablation results validate design choices
- The conclusions are well supported by the extensive experiments which consistently demonstrate the superiority of R3M across environments

Weakness

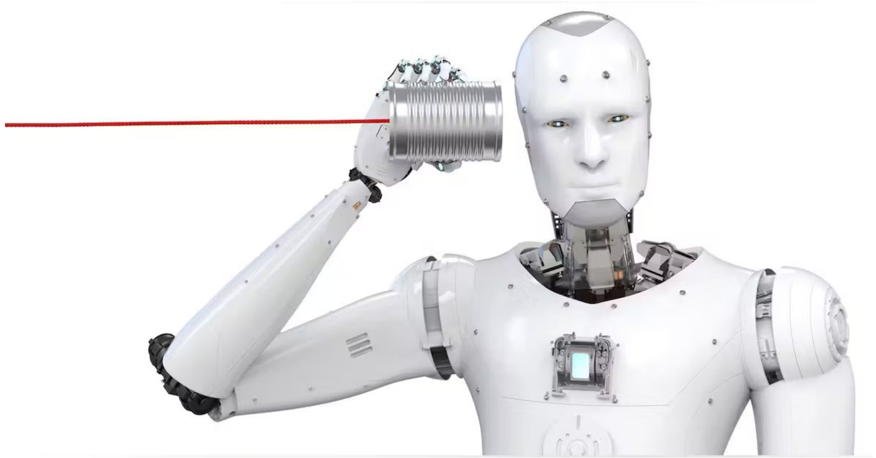
- Not demonstrated in a full reinforcement learning setting
- Safety and scalability to completely open worlds not addressed

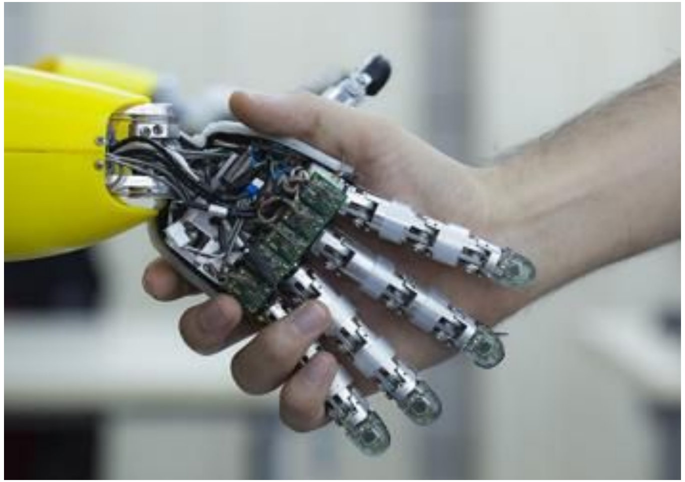
Limitations

- Main evaluation is on imitation learning, specifically behavior cloning, with a small number of task demonstrations, unclear if benefits extend to reinforcement learning
  - Due to reward focus, a good pre-trained representation for RL is not the same as a good pre-trained representation for imitation
- The current R3M model only provides a single-frame state representation
  - Reward learning and task specification
- Requires large labeled datasets of human videos
- Safety concerns with learned policies failing in real world situations
- Scaling to more complex environments as found in the real world remains challenging

Future Work

- How does R3M performs in RL settings and what changes could be made to improve its performance?
- How well does R3M transfer to tasks requiring intricate manipulation skills or tool use?
- Does grounding to other modalities like audio or proprioception confer benefits?





Extended Readings

- R3M: A Universal Visual Representation for Robot Manipulation
  - https://github.com/facebookresearch/r3m

```
from r3m import load_r3m
r3m = load_r3m("resnet50") # resnet18, resnet34
r3m.eval()
```

Extended Readings

- Ego4D: Around the World in 3,000 Hours of Egocentric Video by Kristen Grauman, et al.
  - dataset
- The unsurprising effectiveness of pre-trained vision models for control by S. Parisi, et al.
  - pre-trained methodology
- Masked visual pre-training for motor control by T. Xiao, et al.
  - pre-trained representation from human demos
- Learning Transferable Visual Models From Natural Language Supervision by A. Radford, et al.
  - CLIP
- Scaling robot supervision to hundreds of hours with roboturk: Robotic manipulation dataset through human reasoning and dexterity by A. Mandlekar, et al.
  - Large scale robot manipulation dataset

Summary

Problem: Developing reusable visual representations for efficient robot learning

Importance: Enable robot learning from less task-specific data

Challenge: Capturing semantics from human videos in a way that transfers to robots

Limitation of prior work: Existing methods ineffective for generalizing to low-data learning

Key insight: Pre-train on diverse human videos with temporal and language objectives

Demonstrated: R3M enables state-of-the-art imitation learning success across 12 simulation tasks and 5 real robot tasks with very little data

September 2023

TEXAS
The University of Texas at Austin

THANK YOU

Questions?

Cole Smylie