WHAT MAKES PRE-TRAINED VISUAL REPRESENTATIONS SUCCESSFUL FOR ROBUST MANIPULATION?

Kaylee Burns[1] **Zach Witzel**[1] **Jubayer Ibn Hamid**[1] **Tianhe Yu**[2]
Chelsea Finn[1] **Karol Hausman**[1,2]
[1]Stanford University [2]Google DeepMind

ABSTRACT

Inspired by the success of transfer learning in computer vision, roboticists have investigated visual pre-training as a means to improve the learning efficiency and generalization ability of policies learned from pixels. To that end, past work has favored large object interaction datasets, such as first-person videos of humans completing diverse tasks, in pursuit of manipulation-relevant features. Although this approach improves the efficiency of policy learning, it remains unclear how reliable these representations are in the presence of distribution shifts that arise commonly in robotic applications. Surprisingly, we find that visual representations designed for manipulation and control tasks do not necessarily generalize under subtle changes in lighting and scene texture or the introduction of distractor objects. To understand what properties *do* lead to robust representations, we compare the performance of 15 pre-trained vision models under different visual appearances. We find that emergent segmentation ability is a strong predictor of out-of-distribution generalization among ViT models. The rank order induced by this metric is more predictive than metrics that have previously guided generalization research within computer vision and machine learning, such as downstream ImageNet accuracy, in-domain accuracy, or shape-bias as evaluated by cue-conflict performance. We test this finding extensively on a suite of distribution shifts in ten tasks across two simulated manipulation environments. On the ALOHA setup, segmentation score predicts real-world performance after offline training with 50 demonstrations. Code and more information are available at: `https://kayburns.github.io/segmentingfeatures/`.

1 INTRODUCTION

In spite of vast progress in computer vision, the question of how to learn a good visual representation for robotics remains open (Chen* et al., 2021). Elsewhere in computer vision, internet datasets are retrofit to new tasks with transfer learning, which promises both generalization and fast adaptation to downstream tasks in exchange for large-scale pre-training. But in the field of robotics, this promise has yet to be fulfilled even though policies learned from pixels struggle substantially with data efficiency (Cobbe et al., 2018) and especially generalization under visual changes in a scene (Cobbe et al., 2019a).

Recent work (Damen et al., 2018; Grauman et al., 2022) posits that the missing piece is a large pre-training dataset of object interactions across diverse environments — the ImageNet (Deng et al., 2009) or CommonCrawl (Raffel et al., 2020) of manipulation. That is, if we want to improve the visual generalization ability of pre-trained models we simply need to collect datasets of this kind at scale. Indeed, training on large datasets of first-person human interaction data increases policy performance and learning efficiency downstream (Nair et al., 2022; Xiao et al., 2022), but these evaluations occur in environments that are very similar to those used for policy learning. Robotic applications commonly contain environments with varying lighting conditions, scene textures, and background objects, and we want pre-trained representations to allow the robot to handle such variability. Yet we have few concrete measures of how well pre-trained representations generalize out-of-distribution. To take a step towards understanding these problems, our goal in this paper is to

thoroughly answer the questions "which models generalize?" and "how can we predict how well a pre-trained model will generalize?"

Our first key finding is that, when evaluated under visual distribution shifts, models that are designed for manipulation and control do not outperform standard visual pre-training methods. This finding violates our intuitions about what is needed to scale up robot learning and brings into question what constitutes relevant data, how to quantify useful features, and the importance of design choices such as model architecture. In other words, we need more guiding principles to help us understand what representations are good for manipulation and make the problem of iterating on pre-training strategies much more straightforward. Currently, evaluating a pre-trained policy requires training and rolling out downstream policies across multiple environments and experimental conditions. Instead, we can take inspiration from computer vision, which has developed proxies for robust performance on vast out-of-distribution datasets (Geirhos et al., 2021).

Our second key finding is that the emergent segmentation ability of a ViT model is a strong predictor of out-of-distribution generalization performance. We visualize this phenomenon, which we refer to as "segmenting-features," in Figure 1. Other metrics of model quality, such as linear probes on ImageNet (Chen et al., 2020), and metrics of out-of-distribution performance, such as in-domain accuracy (Miller et al., 2021) and shape-bias (Geirhos et al., 2019), are not predictive for this model class, despite their predictive power in other commonly-studied domains like image classification. This hints at the possibility that the transfer setting of manipulation differs from computer vision tasks typically studied within the robustness literature.

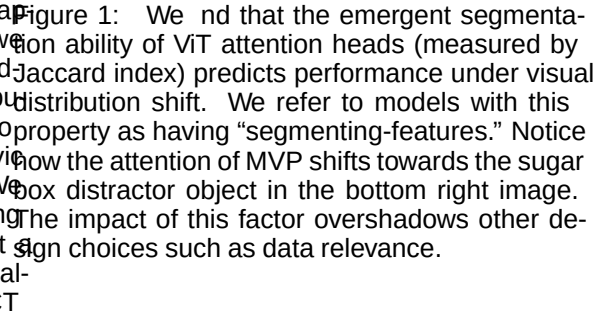

Figure 1: We find that the emergent segmentation ability of ViT attention heads (measured by Jaccard index) predicts performance under visual distribution shift. We refer to models with this property as having "segmenting-features." Notice how the attention of MVP shifts towards the sugar box distractor object in the bottom right image. The impact of this factor overshadows other design choices such as data relevance.

To reach the conclusions above, we run 9,000 different simulated evaluations. Our simulated environments are adapted from two different existing visual distribution shift benchmarks (Xing et al., 2021; Xie* et al., 2023) to capture the shifts that arise commonly in robotics applications: changes in lighting, background and object texture, and the appearance of distractors. More specifically, we train policies on top of 15 pre-trained models, including 4 models designed for manipulation or control: R3M (Nair et al., 2022), two MVP variants (Xiao et al., 2022; Radosavovic et al., 2022), and VIP (Ma et al., 2022). We further validate these findings by comparing a model designed for manipulation against a model with a similar parameter count on a real-world screwdriver pick-up task using the ACT training framework (Zhao et al., 2023). Through these experiments, we make two striking findings: (1) pre-trained visual models designed for control do not necessarily generalize better than models pre-trained on more standard computer vision datasets and (2) the emergent segmentation performance of a ViT model is a strong predictor of the out-of-distribution generalization of a downstream policy.

Figure 2:Evaluation Scheme.We begin our evaluation procedure by training a policy with behavior cloning on top of frozen features. In every experimental setting, we ablate the encoder used to extract features from the image observation. The learned policy is then evaluated in each of the visual shift environments to attain a zero-shot success value.

2 Related Work

Representation learning for manipulation. The correct approach to visual representation learning for robotics is still an open question. There is evidence that separating visual representation learning from policy learning can further improve performance (Pari et al., 2022; Parisi et al., 2022). Recent works have shown that models pre-trained on large manipulation-relevant datasets (Goyal et al., 2017; Damen et al., 2018; Shan et al., 2020; Grauman et al., 2022) or learned with visual affordances from RGBD data (Yen-Chen et al., 2020) can improve the ef ciency and performance of policy learning (Karamcheti et al., 2023) in comparison to standard vision datasets such as ImageNet (Deng et al., 2009), but they do not focus on performance under visual distribution shift. We evaluate the performance of R3M (Nair et al., 2022), MVP (Xiao et al., 2022; Radosavovic et al., 2022), and VIP (Ma et al., 2022). Other work has studied generalization of pre-trained representations to new reinforcement learning tasks for manipulation (Ma et al., 2022) and navigation (Sax et al., 2018) where the agent is able to train on visual data from the new environment. Separate from the question of pre-training visual representations is the question of how to best train policies on top of pixel observations (Laskin et al., 2020b; Yarats et al., 2021). Majumdar et al. (2023) benchmarks the performance of pre-trained visual representations on a handful of manipulation environments, but they focus on in-domain performance and also investigate navigation environments. Hu et al. (2023) shows that model performance is highly sensitive to evaluation. We use imitation learning for our evaluation protocol, which they nd to be a more stable measure of performance. Concurrently with our work, Dasari et al. (2023) demonstrates that the importance of proper data balancing supersedes the content of any one pre-training dataset. We focus on benchmarking visual generalization speci cally and focus on advancing metrics that are predictive of generalization.

Robustness in computer vision.There is extensive work studying the impact of design choices, such as architecture, loss, and data, on the performance of visual models under distribution shift. See Geirhos et al. (2021) for a comprehensive comparison. Most relevant to our paper are studies of shape-bias and architecture. While shape-biased models tend to be more robust than texture-biased ones (Geirhos et al., 2019), the impact of architecture on robustness is less straightforward. For example, vision transformers exhibit better robustness to universal adversarial attacks (Shao et al., 2022), but they are more susceptible to patch-level attacks (Fu et al., 2022). When compared on natural distribution shifts (Hendrycks & Dietterich, 2019; Hendrycks et al., 2021a;b), vision transformers and convolutional networks achieve comparable performance when provided with enough data (Bhojanapalli et al., 2021). But for occlusions speci cally, vision transformers appear to have an edge (Naseer et al., 2021). Miller et al. (2021) studies the predictive power of in-domain performance for out-of-distribution generalization. Unlike all of these prior works, we focus on how pre-trained representations affect robustness in downstream robotics tasks, instead of downstream vision tasks.