WHAT MAKES CERTAIN PRE-TRAINED VISUAL REPRESENTATIONS BETTER FOR ROBOTIC LEARNING?

Kyle Hsu[α,†] **Tyler Ga Wei Lum**[α] **Ruohan Gao**[α] **Shixiang Gu**[β] **Jiajun Wu**[α] **Chelsea Finn**[α]

ABSTRACT

Deep learning for robotics is data-intensive, but collecting high-quality robotics data at scale is prohibitively expensive. One approach to mitigate this is to leverage visual representations pre-trained on relatively abundant non-robotic datasets. So far, existing works have focused on proposing pre-training strategies and assessing them via ablation studies, giving high-level knowledge of how pre-training design choices affect downstream performance. However, the significant gap in data and objective between the two stages motivates a more detailed understanding of what properties of better pre-trained visual representations enable their comparative advantage. In this work, we empirically analyze the representations of robotic manipulation data from several standard benchmarks under a variety of pre-trained models, correlating key metrics of the representations with closed-loop task performance after behavior cloning. We find evidence that suggests our proposed metrics have substantive predictive power for downstream robotic learning.

1 INTRODUCTION

A longstanding goal of the application of machine learning to robotics has been to enable robots to handle the myriad, messy conditions of the physical world. Although broad strides towards human-like generalization have recently occurred in other domains of machine learning such as computer vision and natural language processing (Brown et al., 2020; Radford et al., 2021), such progress remains elusive for robotics. This is in no small part due to the confluence of two observations: that scaling up high-quality robotics data collection (whether in sim or real) is prohibitively expensive (Dasari et al., 2019), and, as Moravec's paradox suggests, that the requisite scale for a watershed moment in robotics may be far beyond that required for "higher-level" tasks (Moravec, 1988).

Motivated by this, there has been a flurry of recent work (Shah & Kumar, 2021; Parisi et al., 2022; Nair et al., 2022b) studying to what extent datasets that were designed for, e.g., semantic image understanding (Deng et al., 2009; Radford et al., 2021) or egocentric video analysis (Grauman et al., 2022) rather than robotics can nonetheless be leveraged to improve the outcomes of robotic learning pipelines. A prominent example of this is deploying self-supervised representations obtained via pre-training on this data (He et al., 2020; Radford et al., 2021; Nair et al., 2022b) on downstream imitation learning and reinforcement learning robotics tasks (Rajeswaran et al., 2017; Gupta et al., 2019; Yu et al., 2020).

However, the significant gap in data and objective between the pre-training and downstream stages makes the entire endeavor rather heuristic, and indeed the aforementioned works mainly focus on proposing pre-training strategies and assessing them via their downstream robotics performance. While comparison with regard to (proxy) stakeholder metrics such as success rate are of course the gold standard, we lack a fine-grained understanding of what properties are are enabling the comparative advantage of better representations. This is worth addressing: better representations are somehow ameliorating challenging aspects of the robotic learning problem, including (for the setting we consider) compounding covariate shift and regression to an output space (actions) that the pre-training is entirely agnostic to. Being able to attribute improved downstream performance to

[α]Stanford University [β]Google
[†]Correspondence to kylehsu@cs.stanford.edu.

specific properties of these representations could help us distinguish which challenges may continue to be addressed by further scaling robotics-agnostic visual pre-training, and which need separate solutions.

In this work, we focus on finding properties of pre-trained visual representations that matter for downstream learning of robotic manipulation tasks. We leverage tools from the neural network representational similarity literature and the intrinsic dimensionality literature to facilitate our empirical analysis of 5 popular pre-trained models on 12 downstream robotic manipulation tasks, each instantiated with 3 camera angles. We follow Parisi et al. (2022) and Nair et al. (2022b) in focusing on imitation learning for the robotic learning stage. We find that:

- Representational similarity between pairs of demonstration trajectories for a task under a pre-trained model correlates with closed-loop task success with behavior cloning under that pre-trained model (Section 3.1).
- Representational similarity between a task's human-defined low-dimensional state space and a pre-trained model's state embedding correlates with closed-loop task success with behavior cloning under that pre-trained model (Section 3.2).
- For each pre-trained model, either nonlinear xor linear estimates of representational intrinsic dimensionality correlates with success (Section 4).

2 PRELIMINARIES

In this section, we establish the scope of our investigation, introduce artifacts we use, and ground the process in notation. This is done for the pre-trained visual representations (Section 2.1), the downstream robot learning algorithm and environments (Section 2.2), and the techniques we borrow from the representational similarity analysis (Section 2.3) and intrinsic dimensionality literatures (Section 2.4).

2.1 PRE-TRAINED MODELS

Let $\mathcal{I} = \{0, \ldots, 255\}^{C \times H \times W}$ denote a space of images. We consider pre-trained models of the form $f : \mathcal{I} \to \mathbb{R}^{d_z}$ that take in an image I and produce a dense vector $z = f(I)$. Since we are interested in pre-training on large-scale non-robotics data, we consider the following works from the literature: R3M (Nair et al., 2022b), CLIP (Radford et al., 2021), and ImageNet (He et al., 2016). Since R3M is generally the best-performing model, we also use two of its ablations. These works all provide pre-trained models with the ResNet-50 (He et al., 2016) architecture[3], a standard choice for transfer learning in computer vision. Indeed, ResNet variants remain among the most popular architectures for vision-based robotic learning (Jang et al., 2022; Sundaresan et al., 2022). See Table 1 for the dataset and primary objective terms for each model.

Table 1: Summary of pre-trained visual representations used in our analysis.

model	dataset	objective terms
R3M	Ego4D (Grauman et al., 2022)	temporal contrastive learning (Sermanet et al., 2018) video-language alignment (VLA; Nair et al., 2022a) L^1 representation regularization
R3M w/o VLA	"	temporal contrastive learning L^1 representation regularization
R3M w/o L^1	"	temporal contrastive learning video-language alignment
ImageNet	ILSVRC 2012 (Deng et al., 2009)	cross-entropy using class labels
CLIP	WebImageText (Radford et al., 2021)	image-text contrastive learning (Zhang et al., 2020)

[3]CLIP uses a modified version of this architecture.

2.2 Downstream Robot Learning

For downstream robotic learning, we consider the imitation learning setting. For some task $\mathcal{T}$, a collection of N demonstration trajectories $\mathcal{D}_{\mathcal{T}} = (\tau_i)_{i=1}^{N}$ is provided, each consisting of observation-action tuples $\tau = ((o_t, a_t))_{t=1}^{T}, o \in \mathcal{O}, a \in \mathcal{A}$. This is used to produce a closed-loop policy $\pi : \mathcal{O} \to \mathcal{A}$ that imitates the demonstrations and is successful in generalizing in the environment to accomplish the task. For our analysis, we will additionally leverage the availability of low-dimensional state information given by physical simulation: a tuple (s, o) involves processing information in state $s \in \mathbb{R}^{d_s}$ to render observations.

In the tasks we consider, an observation $o = (I, p), I \in \mathcal{I}, p \in \mathbb{R}^{d_p}$ is itself a tuple comprising of an image and low-dimensional robot proprioception. A frozen pre-trained visual representation is used to produce an embedding of the image $z = f(I)$. This is concatenated with the proprioceptive observation vector before being further processed into actions by a parametric function approximator $g_\phi : \mathbb{R}^{d_z+d_p} \to \mathcal{A}$. For convenience, we will overload $f : \mathcal{O} \to \mathbb{R}^{d_z+d_p}$ to include the concatenation; disambiguation should be clear from context. The policy $\pi = g_\phi \circ f$ is thus a composition of feature extraction and action generation.

We follow Nair et al. (2022b) in using robotic manipulation tasks sourced from three commonly used benchmarks: 5 from Franka Kitchen (Gupta et al., 2019), 5 from MetaWorld (Yu et al., 2020), and 2 from Adroit (Rajeswaran et al., 2017). The tasks differ in state space, environmental assets, visual appearance, robot end-effector and action space, and functional behavior. We use the behavior cloning data and results from Nair et al. (2022b).

2.3 Representational Similarity Analysis

Given a dataset $\mathcal{D} = (x_i)_{i=1}^{n}, x \in \mathcal{X}$, let $A \in \mathbb{R}^{n \times p}$ be the result of applying function $f_A : \mathcal{X} \to \mathbb{R}^p$ to $\mathcal{D}$. We call A the representation of $\mathcal{D}$ under f_A. Let $B \in \mathbb{R}^{n \times q}$ be the representation of $\mathcal{D}$ under f_B.

A rich literature of techniques exists for quantitatively comparing high-dimensional neural network representations. We adopt the convention of doing so in terms of distances (in the colloquial sense; they don't necessarily satisfy the triangle inequality). One family stems from the idea of first computing pairwise similarity between points in each representation, before comparing these similarity structures. This bypasses issues such as neuron correspondence. As an instance of this strategy, we consider centered kernel alignment (CKA) as popularized by Kornblith et al. (2019). Because the authors find that linear and radial basis function CKA perform similarly on realistically trained neural networks, we follow in using linear CKA discrepancy for simplicity:

$$d_{\text{linear CKA}}(A, B) = 1 - \frac{\left\|A^\top B\right\|_F^2}{\left\|A^\top A\right\|_F \left\|B^\top B\right\|_F}, \tag{1}$$

where $\|\cdot\|_F$ is the Frobenius norm.

Ding et al. (2021) find that CKA is undersensitive, specifically to removing all but the largest principal components of a dataset, and propose the use of the orthogonal Procrustes discrepancy (Schönemann, 1966) as a more sensitive alternative:

$$d_{\text{Procrustes}}(A, B) = \|A\|_F^2 + \|B\|_F^2 - 2\left\|A^\top B\right\|_*, \tag{2}$$

where $\|\cdot\|_*$ is the nuclear norm. This quantity is the solution to the orthogonal Procrustes problem

$$\min_R \left\|B - AR^\top\right\|_F^2 \text{ s.t. } R^\top R = I \tag{3}$$

of aligning the representations under an orthogonal transformation. The above is written for $p = q$, but unequal representation dimensionalities can be matched by the use of "dummy" dimensions. Per standard practice, we preprocess each representation by subtracting the mean vector and dividing by the resulting matrix's Frobenius norm.

2.4 Intrinsic Dimensionality

The manifold hypothesis posits that natural data exhibits low-dimensional structure embedded within conventional representations of high ambient dimensionality. Much work has argued that

the generalization performance of machine learning models, including neural networks, is far better explained by intrinsic rather than ambient dimensionality (Nakada & Imaizumi, 2020; Pope et al., 2021).

Various definitions and estimators of intrinsic dimensionality (ID) have been proposed, each with its own characteristics and interpretation. We select two estimators that provide complementary information on the representations. PCA-$\hat{\text{ID}}@\eta^2$ is the minimal number of principal components needed to achieve a cumulative proportional explained variance of $\eta^2 \in [0, 1]$. By definition, PCA-$\hat{\text{ID}}@\eta^2$ is the dimensionality of a linear subspace of the representation. To provide a nonlinear sense of ID, we consider the two nearest neighbors estimator (TwoNN-$\hat{\text{ID}}$; Facco et al., 2017). This estimator assumes only that the dataset has locally uniform density, specifically in the range of the second-nearest neighbor of each datapoint. This is used to derive the following relationship between ID, the ratio μ between a datapoint's distances to its second-nearest and nearest neighbors, and the ratio variable's cumulative distribution function $F(\mu)$:

$$\text{TwoNN-}\hat{\text{ID}} = -\frac{\log(1 - F(\mu))}{\log(\mu)}. \tag{4}$$

In practice, the estimate is determined by fitting $((\log(\mu_i), -\log(1 - F^{\text{emp}}(\mu_i)))_{i=1}^n$ with a straight line through the origin, where the empirical cumulant F^{emp} is determined by sorting the values of $(\mu_i)_{i=1}^n$ in ascending order through a permutation σ, then defining $F^{\text{emp}}(\mu_i) = \sigma(i)/n$.

3 Representational Consistency

One long-identified challenge of offline imitation learning is the phenomenon of *compounding covariate shift*, a vicious cycle of imprudent actions and out-of-distribution observations. We hypothesize that better pre-trained visual representations mitigate this issue by representing observations in a more consistent manner. We view this as a helpful (if not necessary) condition for the policy to generalize well, a stepping stone towards "stabilization" properties that can e.g. bring worst-case regret bounds from quadratic down to linear in horizon (Ross & Bagnell, 2010).

3.1 Intramodel Trajectory Consistency

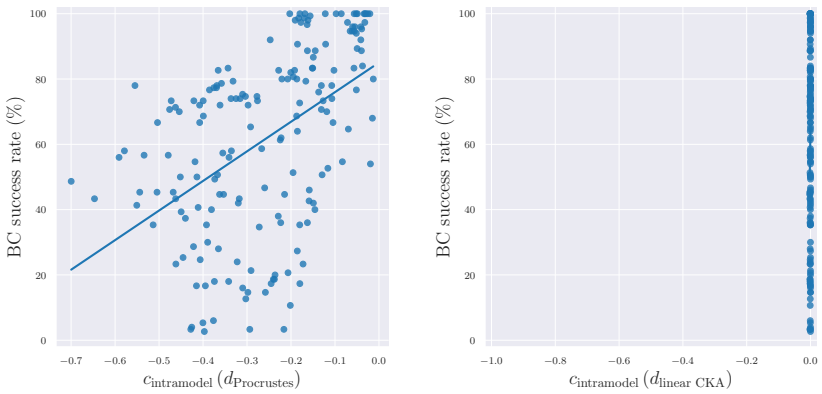

Figure 1: Correlation between intramodel trajectory consistency and behavior cloning (BC) success rate for 36 downstream manipulation tasks under 5 pre-trained models. Left: a moderate positive correlation of $r(175) = 0.49, p = 3.3 \times 10^{-12}$ (two-sided) is measured between intramodel trajectory consistency under the orthogonal Procrustes distance and behavior cloning success rate. Right: consistency measurements with linear CKA exhibited undersensitivity to the variation in the data, with an interquartile range of 0.001 for a codomain of $[-1, 0]$.

To evaluate this hypothesis, we need concrete notions of representational consistency. We first choose to consider the expected representational similarity between pairs of demonstration observation trajectories for the same task under the same pre-trained model. This involves making the assumption that for every task there is correspondence amongst demonstration observations recorded

at the same time step. Concretely, for a task $\mathcal{T}$ and a pre-trained representation f, we evaluate intramodel trajectory consistency

$$c_{\text{intramodel}}(d, \mathcal{T}, f) = -\mathbb{E}_{\tau_i, \tau_j \sim \mathcal{D}_{\mathcal{T}}} \left[d\left(f\left((o_{i,t})_{t=1}^{T} \right), f\left((o_{j,t})_{t=1}^{T} \right) \right) \right] \tag{5}$$

where d is a measure of representational dissimilarity. We use $d_{\text{Procrustes}}$ and $d_{\text{linear CKA}}$ as introduced in Section 2.3.

In Figure 1, we plot $c_{\text{intramodel}}(d, \mathcal{T}, f)$ against the success rate on task $\mathcal{T}$ of a policy trained via behavior cloning with frozen pre-trained representation f. Different points involve a different task, camera angle, or pre-trained representation. For $c_{\text{intramodel}}(d_{\text{Procrustes}})$, we observe a moderate positive correlation with strong statistical significance: $r(175) = 0.49, p = 3.3 \times 10^{-12}$ (two-sided). It is also worth noting that the correlation is stronger in the region where trajectory consistency and success rate are both high. For $c_{\text{intramodel}}(d_{\text{linear CKA}})$, we observe undersensitivity, echoing observations from prior work (Ding et al., 2021). Due to this, we do not view the linear CKA analysis as providing evidence for or against the hypothesis and omit it from the rest of the analysis in this section.

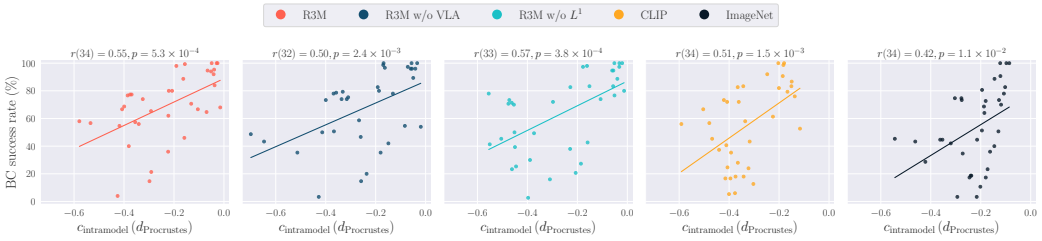

Figure 2: Stratifying $c_{\text{intramodel}}(d_{\text{Procrustes}})$ data in Figure 1 by pre-trained model. Relatively consistent moderate positive correlations between this quantity and downstream behavior cloning success rate appear for every pre-trained model.

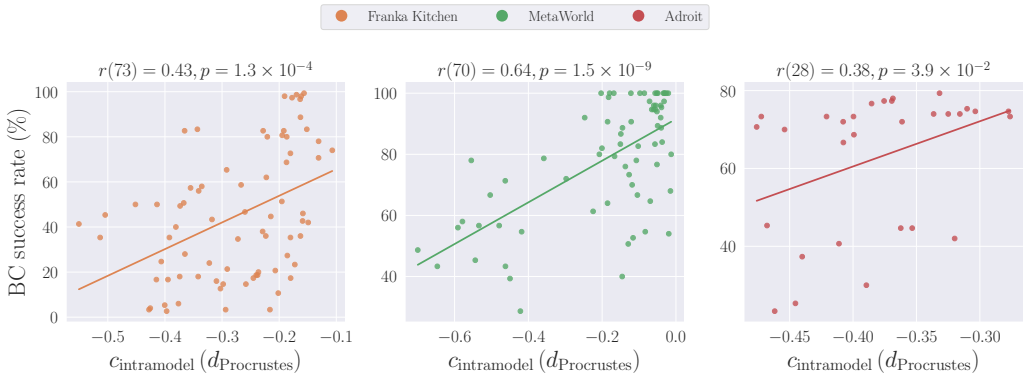

Figure 3: Stratifying $c_{\text{intramodel}}(d_{\text{Procrustes}})$ data in Figure 1 by downstream task domain. The positive correlation is moderate for the Franka Kitchen domain and strong for the MetaWorld domain. It is weaker and has low statistical significance for the Adroit domain.

As our data is heterogeneous, it is important to determine whether this correlation is an artifact of aggregation à la Simpson's paradox. Figure 2 presents the above results stratified by pre-trained model. We observe relatively consistent moderate positive correlations for every pre-trained model with statistical significance. Figure 3 stratifies by downstream task domain. We again see consistent positive correlations for every domain with statistical significance. Interestingly, there is substantial variation in the strength of the correlations amongst domains. Figure 10 further stratifies the data by downstream task. The small sample size per task results in the correlation analysis for most tasks having dubious individual statistical significance. Nevertheless, out of the 12 tasks, 9 suggest moderate or weak positive correlation, and 3 suggest weak negative correlation. Overall, the aggregate,

by-model, and by-domain correlation analyses support the hypothesis. Indeed, because the representations are used for downstream learning, we can say that higher downstream behavior cloning performance is partially explained by intramodel trajectory consistency.

3.2 Model-State Consistency

In a simulated robotics context, there is in some sense a reference representation: the implicitly defined inverse model from observations to low-dimensional state. As this representation is human-engineered to be sufficient yet compact for a given task, we can obtain another notion of representational consistency:

$$c_{\text{model-state}}(d, \mathcal{T}, f) = -d\left(f\left((o_i)_{i=1}^{T|\mathcal{D}_\mathcal{T}|}\right), (s_i)_{i=1}^{T|\mathcal{D}_\mathcal{T}|}\right), \tag{6}$$

that is, for a given task $\mathcal{T}$ and pre-trained visual model f we measure the negative representational distance d between the representation of the task's observations under f and the corresponding low-dimensional states. We leverage this correspondence to concretize the notion of consistency, in the sense that variation in the representation should be commensurate to variation in the state. The assumption of this correspondence is much less objectionable compared to the time-locked correspondence between demonstration trajectories assumed for intramodel trajectory consistency.

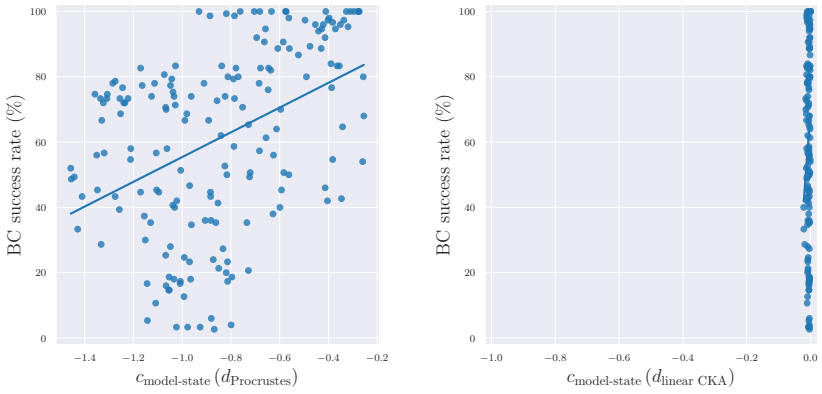

Figure 4: Correlation between model-state consistency and behavior cloning (BC) success rate for 36 downstream manipulation tasks under 5 pre-trained models. Left: a moderate positive correlation of $r(178) = 0.43, p = 2.2 \times 10^{-9}$ (two-sided) is obtained with the orthogonal Procrustes distance. Right: consistency measurements with linear CKA exhibited undersensitivity to the variation in the data, with an interquartile range of 0.008 for a codomain of $[-1, 0]$.

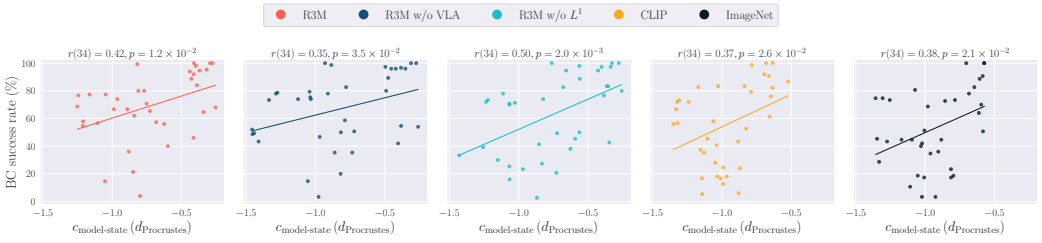

Figure 5: Stratifying $c_{\text{model-state}}(d_{\text{Procrustes}})$ data in Figure 4 by pre-trained model. Relatively consistent moderate positive correlations between this quantity and downstream behavior cloning success rate appear for every pre-trained model.

4 Representational Intrinsic Dimensionality

While we have demonstrated some encouraging evidence for our representation consistency hypothesis, the consistency metrics we use in the previous section are degenerate in the following sense:

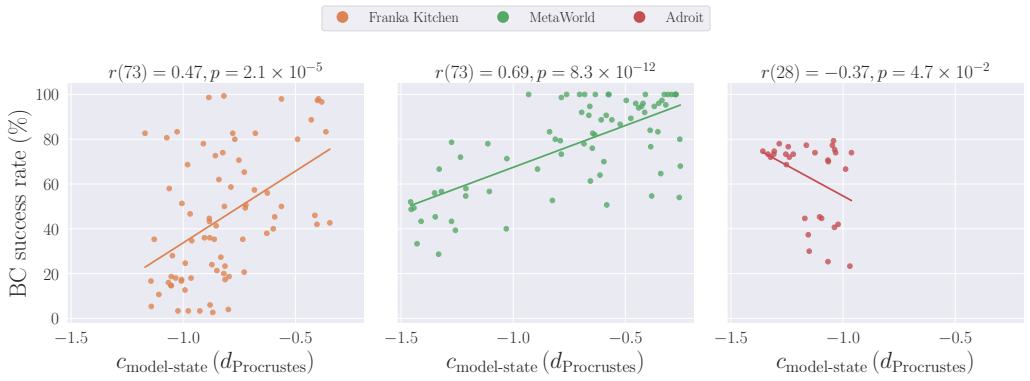

Figure 6: Stratifying $c_{\text{model-state}}(d_{\text{Procrustes}})$ data in Figure 4 by downstream task domain. The positive correlation is moderate for the Franka Kitchen domain and strong for the MetaWorld domain. There is a weak negative correlation with weak statistical significance for the Adroit domain, but this data solely lies in the relatively inconsistent region of the x-axis. Figure 7 reveals that the two constituent tasks each exhibit negligible correlation, further weakening the significance of the domain's negative correlation.

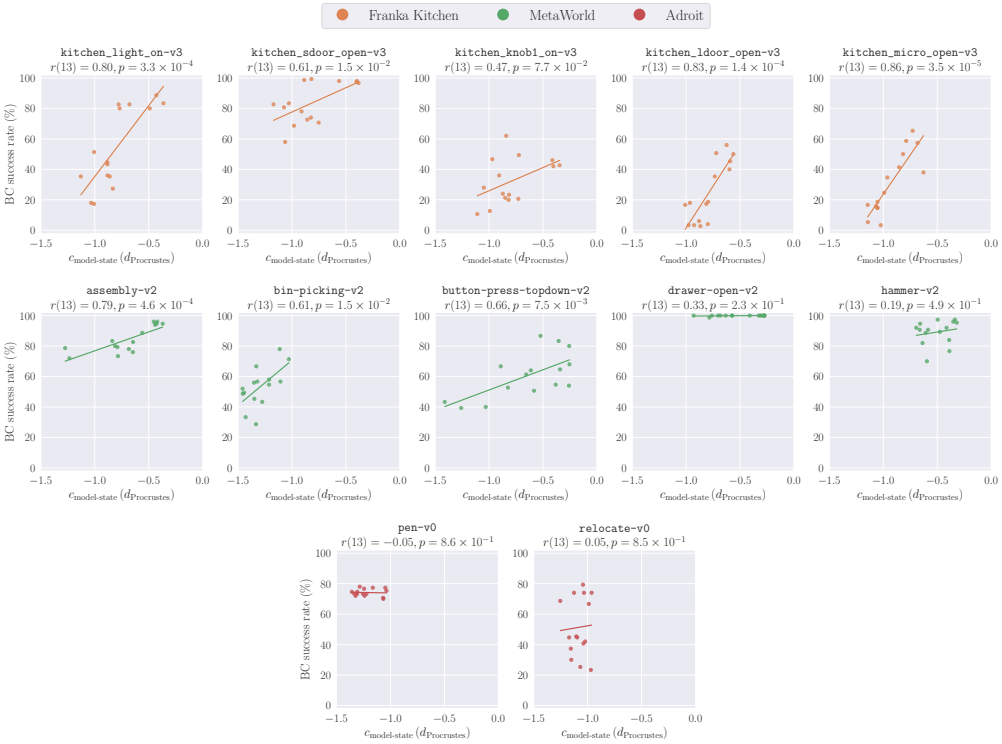

Figure 7: Stratifying $c_{\text{model-state}}(d_{\text{Procrustes}})$ data in Figure 4 by downstream task. Remarkably consistent positive correlations are seen for tasks in the Franka Kitchen and MetaWorld domains. Resolving the Adroit domain into its constituent tasks reveals that the weak negative correlation at the domain level is more accurately a negligible correlation.

trivial representations such as the zero function achieve perfect consistency, yet obviously would not facilitate downstream learning. To complement the previous analysis, we measure the intrinsic dimensionality (ID) of the same downstream demonstration data under pre-trained visual representations; trivial representations are detectable by ID estimates near zero.

We consider the following two ID estimates. Trajectory ID estimates the average ID along a demonstration trajectory:

$$\hat{\text{ID}}_{\text{trajectory}}(\hat{\text{ID}}, \mathcal{T}, f) = \mathbb{E}_{\tau \sim \mathcal{D}_{\mathcal{T}}} \left[\hat{\text{ID}} \left(f \left((o_t)_{t=1}^{T} \right) \right) \right] \tag{7}$$

and task ID estimates the ID of all demonstration data for a task:

$$\hat{\text{ID}}_{\text{task}}(\hat{\text{ID}}, \mathcal{T}, f) = \hat{\text{ID}} \left(f \left((o_i)_{i=1}^{|\mathcal{D}_{\mathcal{T}}|T} \right) \right) \tag{8}$$

To obtain a more meaningful quantity that facilitates comparison between tasks of varying visual and functional complexity, we normalize both quantities by the task ID using low-dimensional state to obtain relative ID:

$$\text{relative } \hat{\text{ID}}_{\text{trajectory/task}}(\hat{\text{ID}}, \mathcal{T}, f) = \frac{\hat{\text{ID}}_{\text{trajectory/task}}(\hat{\text{ID}}, \mathcal{T}, f)}{\hat{\text{ID}}_{\text{trajectory/task}}(\hat{\text{ID}}, \mathcal{T}, \text{state})} \tag{9}$$

For the base ID estimators $\hat{\text{ID}}$ use both TwoNN-$\hat{\text{ID}}$ and PCA-$\hat{\text{ID}}$@0.95. As TwoNN-$\hat{\text{ID}}$ can be sensitive to outliers (Facco et al., 2017), when using it we discard the top 0%, 5%, or 10% of the regression data, chosen by resulting R^2.

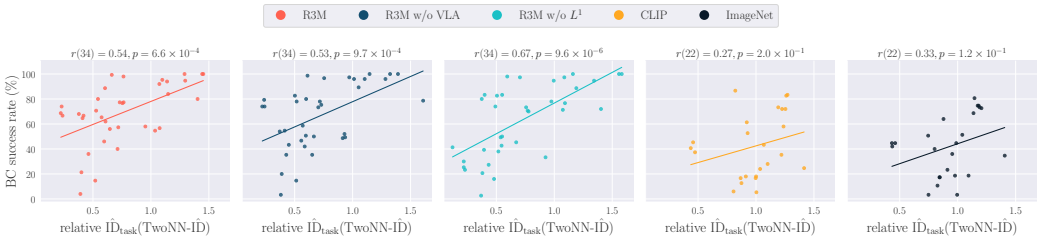

Figure 8: The TwoNN-$\hat{\text{ID}}$ estimator gives a nonlinear notion of ID that results in moderate statistically significant correlations between relative task ID and downstream behavior cloning success rate for R3M and its variants. The correlation is not significant for the other two models.

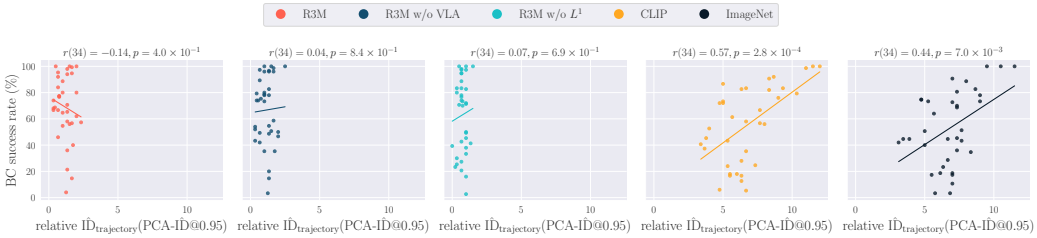

Figure 9: The PCA-$\hat{\text{ID}}$@0.95 estimator gives a linear notion of ID that results in moderate statistically significant correlations between relative trajectory ID and downstream behavior cloning success rate for CLIP and ImageNet. The correlation is not significant for R3M and its variants.

In Figures 8 and 9, we plot relative task ID with TwoNN-$\hat{\text{ID}}$ and relative trajectory ID with PCA-$\hat{\text{ID}}$@@0.95, respectively, against downstream task performance after behavior cloning. We remark that there is a bifurcation between i) R3M and its variants and ii) CLIP and ImageNet for both analyses. Furthermore, the two analyses are complementary: for any pre-trained model, either the nonlinear or the linear estimate of ID yields significant moderate correlation with downstream performance, but not both. Altogether, this suggests that the representations of R3M and its variants are better described by a nonlinear manifold, and that those of CLIP and ImageNet are better described by a linear subspace. This may be related to the fact that CLIP and ImageNet are trained under "higher-level" semantic objectives compared to R3M. The positive correlations between the suitable ID estimate for each model and downstream success suggests that some of the failure cases may be due to undersensitivity to task-critical features in the raw observations. This makes sense considering the significant gap between the pre-training data for each model and downstream domains.

5 Related Work

Most of the pertinent prior work has already been introduced to set context for this work in Sections 1, and 2.

While Zhang et al. (2022) also empirically analyze self-supervised visual representations and present correlations between representation properties and downstream decision-making, our work is distinct in several respects. They consider a single domain (Atari) for both pre-training and downstream learning, whereas we consider a variety of pre-training domains for downstream robotic manipulation. Their focus is on reinforcement learning, whereas ours is on imitation learning. Their empirical analysis of visual representations is based on linear probing, whereas ours does not involve training any additional parameters.

Seo et al. (2022) learn visual world models from action-free robotic manipulation videos and assess transfer to robotic manipulation and locomotion tasks in held-out benchmarks. The breadth of data in their pre-training is relatively limited and their transfer process involves stacking an action-conditional module before model-based reinforcement learning, a pipeline much more complex than simple behavior cloning.

More broadly, we are inspired by the subfield of BERTology (Rogers et al., 2020), in which creative hypotheses about the inner workings of BERT (Devlin et al., 2018) are proposed and assessed; the work of Huh et al. (2016), which investigates the importance of dataset composition for transfer in computer vision; and Naitzat et al. (2020), who investigate the role of geometry and topology in the success of high-dimensional neural network representations .

6 Discussion

In this work, we investigated properties of pre-trained visual representations that explain downstream robotic manipulation performance after imitation learning under the representations. The representational consistency and representational intrinsic dimensionality metrics we proposed appear to correlate moderately and significantly with downstream success rate. R3M and its variants yielded consistent behavior for all of our analyses. This suggests that the success of R3M lies not within any variant's particular cocktail of objective terms, but rather the temporal contrastive learning on a diverse dataset that they all partake in.

Since the representational metrics proposed in this work moderately correlate with downstream success, and since the direction of causality is established by the use of the representations for downstream learning, our analysis could potentially be leveraged as a starting point for predicting downstream learning outcomes and diagnosing failure cases.

There are several limitations to the analysis conducted in this work. Though we took care to interpret the analyses at multiple resolutions, it remains possible that the primary source of the correlations we observe is a confounder not accounted for. The fact that we used models, data, environments, and results from prior work hopefully mitigates concerns regarding inapplicability and cherry-picking, but it could also be fruitful to consider going beyond artifacts from prior work, e.g. designing new probing tasks aimed at further elucidating properties of these representations relevant to robotics.

On that note: this paper is work in progress. We plan to continue by generating data representing various kinds of distribution shifts in the downstream task environments and investigating whether the tools presented above can be used to engineer a transfer procedure that results in a robust policy. We thank the reviewers in advance for any comments and feedback about this work.

ACKNOWLEDGMENTS

We gratefully acknowledge the developers and maintainers of software packages that enabled this research: NumPy (Harris et al., 2020), SciPy (Virtanen et al., 2020), PyTorch (Paszke et al., 2019), matplotlib (Hunter, 2007), and seaborn (Waskom, 2021).

REFERENCES

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Sudeep Dasari, Frederik Ebert, Stephen Tian, Suraj Nair, Bernadette Bucher, Karl Schmeckpeper, Siddharth Singh, Sergey Levine, and Chelsea Finn. Robonet: Large-scale multi-robot learning. *arXiv preprint arXiv:1910.11215*, 2019.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Frances Ding, Jean-Stanislas Denain, and Jacob Steinhardt. Grounding representation similarity through statistical testing. *Advances in Neural Information Processing Systems*, 34:1556–1568, 2021.

Elena Facco, Maria d'Errico, Alex Rodriguez, and Alessandro Laio. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific reports*, 7(1):1–8, 2017.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18995–19012, 2022.

Abhishek Gupta, Vikash Kumar, Corey Lynch, Sergey Levine, and Karol Hausman. Relay policy learning: Solving long-horizon tasks via imitation and reinforcement learning. *arXiv preprint arXiv:1910.11956*, 2019.

Charles R Harris, K Jarrod Millman, Stéfan J Van Der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J Smith, et al. Array programming with numpy. *Nature*, 585(7825):357–362, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

Minyoung Huh, Pulkit Agrawal, and Alexei A Efros. What makes imagenet good for transfer learning? *arXiv preprint arXiv:1608.08614*, 2016.

John D Hunter. Matplotlib: A 2d graphics environment. *Computing in science & engineering*, 9 (03):90–95, 2007.

Eric Jang, Alex Irpan, Mohi Khansari, Daniel Kappler, Frederik Ebert, Corey Lynch, Sergey Levine, and Chelsea Finn. Bc-z: Zero-shot task generalization with robotic imitation learning. In *Conference on Robot Learning*, pp. 991–1002. PMLR, 2022.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, pp. 3519–3529. PMLR, 2019.

Hans Moravec. *Mind children: The future of robot and human intelligence*. Harvard University Press, 1988.

Suraj Nair, Eric Mitchell, Kevin Chen, Silvio Savarese, Chelsea Finn, et al. Learning language-conditioned robot behavior from offline data and crowd-sourced annotation. In *Conference on Robot Learning*, pp. 1303–1315. PMLR, 2022a.

Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3m: A universal visual representation for robot manipulation. *arXiv preprint arXiv:2203.12601*, 2022b.

Gregory Naitzat, Andrey Zhitnikov, and Lek-Heng Lim. Topology of deep neural networks. *J. Mach. Learn. Res.*, 21(184):1–40, 2020.

Ryumei Nakada and Masaaki Imaizumi. Adaptive approximation and generalization of deep neural network with intrinsic dimensionality. *J. Mach. Learn. Res.*, 21(174):1–38, 2020.

Simone Parisi, Aravind Rajeswaran, Senthil Purushwalkam, and Abhinav Gupta. The unsurprising effectiveness of pre-trained vision models for control. *arXiv preprint arXiv:2203.03580*, 2022.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Phillip Pope, Chen Zhu, Ahmed Abdelkader, Micah Goldblum, and Tom Goldstein. The intrinsic dimension of images and its impact on learning. *arXiv preprint arXiv:2104.08894*, 2021.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Aravind Rajeswaran, Vikash Kumar, Abhishek Gupta, Giulia Vezzani, John Schulman, Emanuel Todorov, and Sergey Levine. Learning complex dexterous manipulation with deep reinforcement learning and demonstrations. *arXiv preprint arXiv:1709.10087*, 2017.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866, 2020.

Stéphane Ross and Drew Bagnell. Efficient reductions for imitation learning. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 661–668. JMLR Workshop and Conference Proceedings, 2010.

Peter H Schönemann. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10, 1966.

Younggyo Seo, Kimin Lee, Stephen L James, and Pieter Abbeel. Reinforcement learning with action-free pre-training from videos. In *International Conference on Machine Learning*, pp. 19561–19579. PMLR, 2022.

Pierre Sermanet, Corey Lynch, Yevgen Chebotar, Jasmine Hsu, Eric Jang, Stefan Schaal, Sergey Levine, and Google Brain. Time-contrastive networks: Self-supervised learning from video. In *2018 IEEE international conference on robotics and automation (ICRA)*, pp. 1134–1141. IEEE, 2018.

Rutav Shah and Vikash Kumar. Rrl: Resnet as representation for reinforcement learning. *arXiv preprint arXiv:2107.03380*, 2021.

Priya Sundaresan, Suneel Belkhale, and Dorsa Sadigh. Learning visuo-haptic skewering strategies for robot-assisted feeding. In *6th Annual Conference on Robot Learning*, 2022.

Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, et al. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, 17(3):261–272, 2020.

Michael L Waskom. Seaborn: statistical data visualization. *Journal of Open Source Software*, 6 (60):3021, 2021.

Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Karol Hausman, Chelsea Finn, and Sergey Levine. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on robot learning*, pp. 1094–1100. PMLR, 2020.

Wancong Zhang, Anthony GX-Chen, Vlad Sobal, Yann LeCun, and Nicolas Carion. Light-weight probing of unsupervised representations for reinforcement learning. *arXiv preprint arXiv:2208.12345*, 2022.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. *arXiv preprint arXiv:2010.00747*, 2020.

A REPRESENTATIONAL CONSISTENCY: ADDITIONAL RESULTS

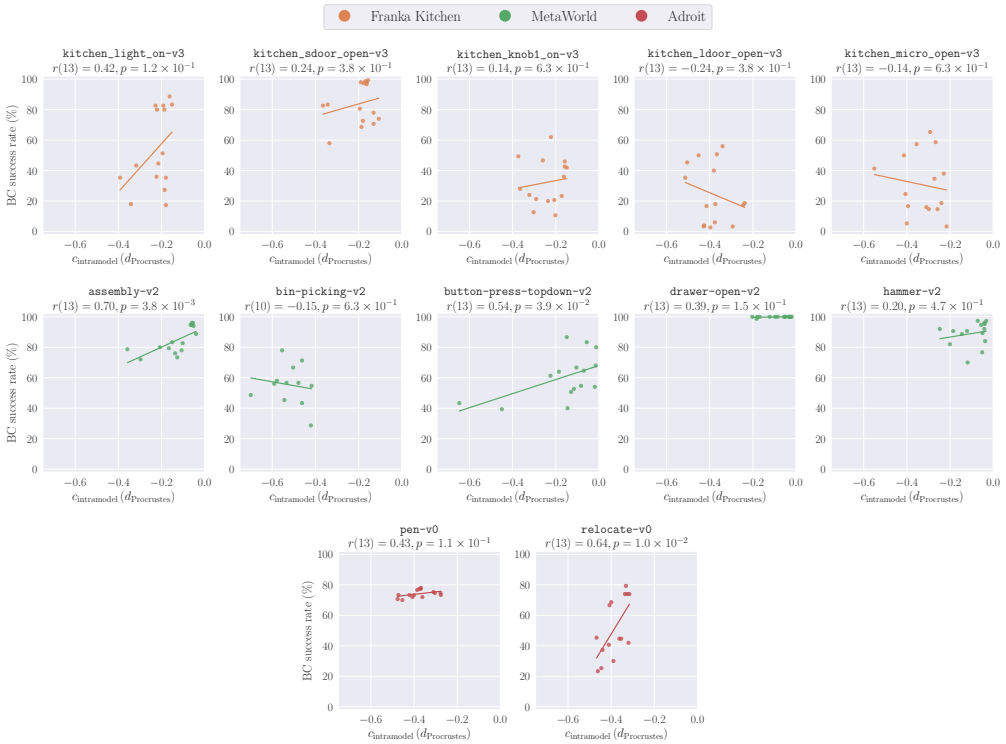

Figure 10: Stratifying $c_{\text{intramodel}}(d_{\text{Procrustes}})$ data in Figure 1 by downstream task.