Leveraging Self-Supervised Learning to enforce Disentangled Representations for Reinforcement Learning

Chia-Hong Hsu [*1] **Jazlyn Lin** [*1] **Shreyas Sundara Raman** [*1] **Vipul Sharma** [*1] **Yichen Wei** [*1]
Randall Balestriero [1]

Abstract

Disentangled representations offer a path to sample-efficient and generalizable reinforcement learning (RL) from rich visual observations. However, traditional disentanglement metrics struggle to uncover factors underlying task dynamics. Current approaches using self-supervised learning (SSL) for disentanglement in RL enforce invariance to visual perturbations or spurious correlations. We propose an SSL representation learning objective that encourages independence between state-features through covariance constraints given only visual observations. Though our learned representations evaluated on DoorKey demonstrate feature disentanglement, they do not lead to sample-efficient policy learning.

1. Introduction

Most data distributions can be explained by a set of causal factors. For example, any chess board can be reduced to 32 factors representing the coordinates of every piece. However, not all representations are equally useful for data-driven decision making. Disentanglement attempts to bridge this gap by learning representations where each element independently captures changes in a different causal factor.

In the context of visual reinforcement learning (RL), disentanglement can distill feature-rich state representations (images) to a subset of factors (latent representation) necessary for the target RL task. These distilled factors could capture the causal factors governing transition or reward dynamics of the target task, which policy learning algorithms can exploit for more sample-efficient learning and more generalized extrapolation to out-of-distributions tasks governed by similar factors. Thus, disentangled representations are crucial for visual RL, where agents need to generalize across continuously shifting feature-rich data distributions as they explore or manipulate their environments.

However, the potentially mutable number of factors and temporal dependence of RL data make traditional disentanglement metrics non-conducive. This makes disentanglement ill-defined for RL, as practitioners often pre-define desirable factors to be isolated and discarding too much information prevents representations from accurately characterizing environment dynamics. Related works learning disentangled representations for RL target invariance to distribution shifts or spuriously correlated features.

To uncover causal factors using disentanglement, a key goal of our work, we turn to factored MDP definitions. We use self-supervised learning (SSL) objectives to enforce desired constraints, whilst eliminating reliance on pre-defined factors. To our knowledge, we are the first to attempt causal disentanglement of visual observations through factored MDP constraints using SSL. We propose a representation learning method that enforces independence between latent features via covaraiance constraints, given only visual observations, with no assumptions on RL tasks.We evaluate our learned representations on the DoorKey-8x8 MiniGrid (across 3 seeds), showing they align with factored MDP disentanglement but are perhaps not useful for policy learning compared to empirical lower/upper bounds.

2. Background

2.1. SSL & Disentanglement for RL

The manifold hypothesis(Meng et al., 2024) posits that high-dimensional data ($\mathcal{X}(\Theta)$) lies on a lower-dimensional manifold, which is defined by a comparatively small set of intrinsic coordinates (Θ). Manifold learning (Chen et al., 2020) can recover these intrinsic coordinates, up to some transfomration e.g. rotation, through self-supervised reconstruction: $\mathcal{X}^{-1}\mathcal{X}(\Theta) = \Theta$ (Misra & van der Maaten, 2019; Chen et al., 2020). For disentanglement in visual RL, Θ represents causal factors (e.g. agent (x, y) or obstacle (x, y)) underlying observation distributions. We aim to learn $\mathcal{X}^{-1} : \mathcal{X}(\Theta) \to \Theta$ a function mapping images to

[*]Equal contribution [1]Brown University, Providence RI, USA. Correspondence to: Chia-Hong Hsu <chia_hong_hsu@brown.edu>, Shreyas Sundara Raman <shreyas_sundara_raman@brown.edu>.

Submitted to the Brown Self Supervised Learning Workshop (BSSL).

disentangled causal factors.

Previous representation learning works (Burgess et al., 2018; Higgins et al., 2018; Wang et al., 2021) show disentanglement improves generalization to visual changes for continuous control – since color/lighting changes only affect a fraction of latent factors with others preserving information. (Oord et al., 2018) extracts useful representations from sequential high-dimensional data by maximizing mutual information between embeddings of future time-steps and embeddings auto-regressively predicted from a context vector — using their proposed InfoNCE loss.

In the field of RL, works like (Dunion et al., 2023a) introduced conditional independence to disentangle spurious correlations between color and transition dynamics. Later, (Dunion et al., 2023b) proposed a self-supervised objective to separate non-stationary, agent-influenced features from static ones, enhancing resilience to adversarial visual changes. However, the consistent advantage of SSL objectives over image augmentations in RL is being actively debated (Li et al., 2022b).

Overall, previous works separate representation learning from policy learning, using mutual-information, conditional mutual-information or contrastive learning for disentanglement focusing on building invariance. Our approach attempts to use disentanglement fo understand the causal factors underlying task dynamics.

2.2. Factored MDPs & the representation gap

Performance disparities between RL with rich, high-dimensional observations and succinct expert states induces a '*representation gap*' (Chandak et al., 2019; He et al., 2022; Allen, 2023), where the latter usually outperforms. Bridging this involves learning lower-dimensional representations that discard irrelevant information (Maćkiewicz & Ratajczak, 1993; Kingma & Welling, 2022; 2019; Su & Wu, 2018; Allen, 2023), yet preserve enough to fully characterize the state space.

We adopt Factored Markov Decision-Process (F-MDP) definitions (Boutilier et al., 1999; Koller & Parr, 2013; Higgins et al., 2016; Degris et al., 2006), where the MDP transition function depends on a set of finite state-variables or *factors* ($\mathcal{K}$) instead of a single monolithic state (s_t). Thus, F-MDP leads to 'focused causes', where changes to each factor $s_{t+1}[i]$ can be explained by changes in a sparse subset of *parent factors* $\rho(s_{t+1}[i])$. The MDP transition function becomes Eq1 in an F-MDP. As the maximum number of *parent factors*, $\mathcal{P} := \forall i, \max \rho(s_t[i])$ reduces relative to the total number of *factors* ($\mathcal{P} << \mathcal{K}$) the representation becomes *more factorized* – until it is fully disentangled i.e. $\rho(s_t[i]) = s_t[i]$. See Figure 1.

$$T(s_{t+1} \mid a_t, s_t) = \prod_{i=1} T(s_{t+1}[i] \mid a_t, \rho(s_{t+1}[i])) \quad (1)$$

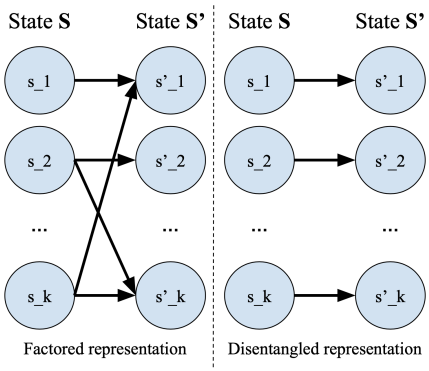

Figure 1. Visualization of a factored and disentangled MDP with $\mathcal{K}$ factors. In this example, $\rho(s_1) = \{s_1, s_2\}$ for factored and $\rho(s_1) = \{s_1\}$ for disentangled representations

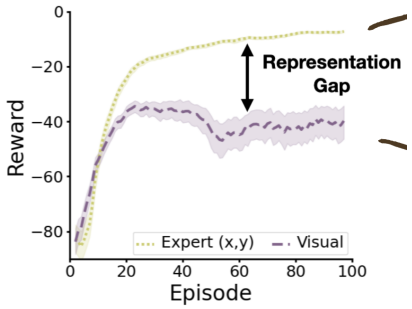

Figure 2. Representation gap (Allen, 2023) captures difference in policy performance between visual observations and expert features – where the latter leads to faster learning. We hope to learn visual encoders ϕ whose latent representation bridges this gap

3. Method

Given visual observations of an RL task, our proposed representation learning method (as pre-text tasks) uses self-supervised objectives for disentanglement via Covariance Constraints. These learned representations are frozen and used for downstream policy learning. To evaluate the degree of disentanglement and improvements to policy learning, we introduce evaluation metrics later in the section.

3.1. Covariance Constraints

Inspired by SSL objectives like Barlow Twins (Zbontar et al., 2021) this method attempts to minimize the Frobenius Norm of the covariance matrix ($\mathcal{C}$) across all pairs of latent factors, bringing the matrix closer to identity ($\mathcal{I} \in \mathcal{R}^{k \times k}$), which motivates learned factors for each batch to be statistically independent of one another. Given a batch of visual observations ($\{o_1, ..., o_t\}, o_i \in \mathbb{R}^{3 \times H \times W}$), this method trains a Nature CNN (Mnih et al., 2015) encoder ($\phi : o_i \rightarrow z_i$)

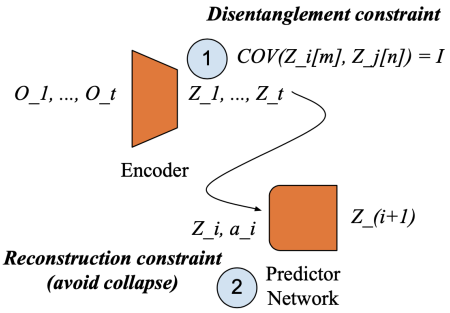

Figure 3. Our covariance constraints method takes a batch of sequential feature-rich visual observations $\{o_1, ..., o_t\}$ from the environment and embeds them into latent representations $z_i = \phi(o_i) \in \mathcal{R}^k$ with k factors. To motivate disentanglement, our loss brings the covariance matrix ($\mathcal{C}$) between every pair of k factors closer to an identity matrix ($\mathcal{I} \in \mathcal{R}^{k \times k}$). To prevent collapse, where $\mathcal{C} = \mathcal{I}$, we also decode successive states (z_{i+1})

whose learned latent representations ($z_i \in \mathcal{R}^k$) are used for downstream policy learning. Here, k determines the number of factors in the disentangled representation.

To prevent representation collapse, where $\mathcal{C} = \mathcal{I}$, we utilize a projection network ($g_\theta(z_i, a_i) = \hat{z_{i+1}}$) to predict successive latent states. Our auxiliary objective attempts to minimize the euclidean-norm between predicted ($\hat{z_{i+1}}$) and the encoded (z_{i+1}) successive latent state. This objective motivates the projection network g_θ to learn transition dynamics ($\mathcal{T}$) of the RL task, under the Markov assumption, hopefully allowing disentangled representations (z_i) to also be relevant for policy learning. A stop-gradient is applied between $\phi(o_i)$ and g_θ so that the projection network remains an effective target during training. See Fig. 3.

3.2. Baselines

Our learned visual encoders $\phi : o_i \rightarrow z_i$ attempt to bridge the representation gap through disentanglement, as discussed in 2.2. Effective evaluation involves comparing against baselines for representation and policy learning.

For policy learning, we train PPO (Schulman et al., 2017) on visual observations jointly for representation/policy learning to measure effects of enforcing disentanglement. We also train PPO using expert states pre-defined to be disentangled – as an empirical upper-bound on sample efficiency. For representation learning, we compare against an encoder trained to predict expert states with full-supervision, to measure quality of representations learned via self-supervision. Experiment details are in Appendix A.3.

3.3. Disentanglement Metrics

To evaluate whether our learned representation capture desired notions of disentanglement, we consider 2 metrics capturing the degree of disentanglement: Frobenius norm of feature correlation and the mutual information gap (MIG) (Chen et al., 2019).

Frobenius norm of feature correlation We measure Pearson correlation (Pearson, 1895), which captures linear inter-feature dependence between variables, across all learned latent feature attributes. We report the Frobenius Norm of this Pearson correlation matrix as a proxy for disentanglement.

Mutual information gap (MIG) Mutual information, MI $I(X;Y)$, quantifies the information shared between two random variables X and Y, quantifying the impact of knowing one variable on reducing uncertainty about the other.

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \tag{2}$$

By measuring how much the joint distribution $p(x,y)$ of 2 variables (X,Y) differs from the product of marginal distributions, $p(x)p(y)$, MI captures dependence between them. MIG (Mutual Information Gap) 3 correlates to disentanglement, by comparing the MI of the most and second-most informative latent dimensions, averaged across all ground-truth (expert) factors. Thus, MIG assesses the extent to which each ground-truth factor of variation is independently captured by one independent/relevant latent dimension. Refer to (Carbonneau et al., 2022) for further detail.

$$\text{MIG}(X,Z) = \frac{1}{n} \sum_{i=1}^{n} \left(I(X_i;Z) - \max_{j \neq i} I(X_j;Z) \right) \tag{3}$$

4. Experiments and Results

We evaluate on the exploration-heavy DoorKey Minigrid (Chevalier-Boisvert et al., 2023) domain. More details of the environment can be found in AppendixA.2.

4.1. Self-supervised Learning

We run two ablation experiments to measure the effectiveness of SSL pre-training objectives. The first experiment compares our covariance constraint method against a Supervised and a Barlow Twins baseline. The second experiment measures the impact of dataset size on disentanglement. For both the experiments, the corresponding downstream RL performance is reported in Table 1.

Impact of Pre-training Objectives Table 1 and Figure 6 shows Barlow Twins method outperforms our Covariance Constraint method for learning disentangled representations. Further, the supervised learning method is outperforming our SSL method, revealing a lot of potential for further disentanglement. One hypothesis for the lack of SSL performance is that the magnitude of pixel-level change representative of factor changes are too small, requiring more data to

Representation Learning	Dataset Size	Disentanglement (Frobenius Norm)	RL Best model Success Rate
Raw Images	N/A	N/A	0.96 ± 0.06
Expert Features	N/A	N/A	$\mathbf{1 \pm 0}$
Supervised	100k	0.0110	0 ± 0
	200k	0.0062	0 ± 0
	300k	**0.0059**	0 ± 0
Barlow Twins	100k	0.0110	0 ± 0
	200k	0.0096	0 ± 0
	300k	0.0101	0 ± 0
Covariance Constraint	100k	0.0159	0 ± 0
	200k	0.0144	0 ± 0
	300k	**0.0130**	0 ± 0

Table 1. The performance of different representation learning methods and dataset size used to train them. It is measured using the Frobenius norm for the disentanglement of the representations and success rate for downstream RL tasks. The disentanglement metrics are averaged across 3 seeds.

disambiguate or that visual augmentations aren't impactful enough to learn good representations.

Impact of Training Data Size It is evident from Table 1 and Figure 6 that a bigger training data size helps to learn disentangled representations. While supervised learning directly enforces disentanglement by reconstructing specific targets, the covariance method decreases slowly since it only loosely enforces disentanglement. Barlow Twins does not force disentanglement, which leads to similar disentanglement performance regardless of data size.

Impact of representation learning on training dynamics Appendix A.4 shows evidence of potential disentanglement between layers 5 to 7 of the encoder. Trend of disentanglement is less clear when measured using Euclidian Distance. Cosine Similarity highlights our method and barlow twins undergo similar representational changes through the encoder; MIG also highlights how this stronger disentanglement metric emerges later in the encoder with our method achieving highest mean MIG – though this is not statistically significant. Overall, there is a weak increase in disentanglement observed over network layers.

4.2. Reinforcement Learning

We compare the success rate of the following RL agents: RL learning visual representation online, with fixed representations trained by BarlowTwins, Covariance, supervised learning, and expert hand-crafted representation. Hyperparameters are recorded in Appendix A.3.1.

RL agent success rate is shown in Fig 4. We find that policies conditioned on an expert representation designed to be disentangled outperform policies trained directly on rich image observation. Other learned representations (covaraince, barlow twins or supervised) are unable to succeed in the RL tasks. One hypothesis for why the learned covariance-based representation perform poorly on RL might be due to insufficient tuning of RL or SSL parameters, due to compute and time constraint. Expert representation performance shows that disentangled representations are clearly useful for RL tasks; however, Appendix A.4 and Table 1 show our representations (and those from supervised baselines) may not be disentangled enough to realize these benefits.

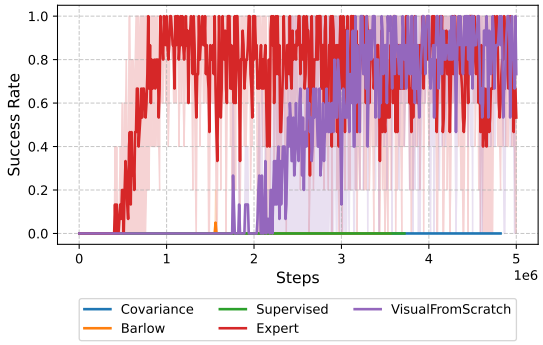

Figure 4. RL task success rates for different representations. SSL representations are trained with 200k samples.

5. Conclusion

We propose a method for self-supervised disentangled representation learning towards higher RL sample efficiency. Our method shows some evidence of disentanglement. However, these disentangled representations do not seem to support policy and RL task learning.

For future work, we plan to explore other domains with greater visual variance from factor changes. We also hope to scale our evaluations (beyond currently limited range) to millions of samples for representation learning to examine scaling laws for representation learning i.e. when does more data have diminishing returns. Finally, we propose an alternate representation learning method that more closely aligns with the factored MDP definitions of disentanglement via masked latent reconstruction, outlined in Appendix A.6.1.

References

Allen, C. S. *Structured Abstractions for General-Purpose Decision Making*. PhD thesis, Brown University, 2023.

Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., LeCun, Y., and Ballas, N. Self-supervised learning from images with a joint-embedding predictive architecture, 2023. URL https://arxiv.org/abs/2301.08243.

Baevski, A., Hsu, W.-N., Xu, Q., Babu, A., Gu, J., and Auli, M. data2vec: A general framework for self-supervised learning in speech, vision and language, 2022. URL https://arxiv.org/abs/2202.03555.

Boutilier, C., Dean, T., and Hanks, S. Decision-theoretic planning: Structural assumptions and computational leverage. *Journal of Artificial Intelligence Research*, 11: 1–94, July 1999. ISSN 1076-9757. doi: 10.1613/jair.575. URL http://dx.doi.org/10.1613/jair.575.

Burgess, C. P., Higgins, I., Pal, A., Matthey, L., Watters, N., Desjardins, G., and Lerchner, A. Understanding disentangling in β-vae, 2018. URL https://arxiv.org/abs/1804.03599.

Carbonneau, M.-A., Zaidi, J., Boilard, J., and Gagnon, G. Measuring disentanglement: A review of metrics, 2022. URL https://arxiv.org/abs/2012.09276.

Chandak, Y., Theocharous, G., Kostas, J., Jordan, S., and Thomas, P. Learning action representations for reinforcement learning. In *International conference on machine learning*, pp. 941–950. PMLR, 2019.

Chen, R. T. Q., Li, X., Grosse, R., and Duvenaud, D. Isolating sources of disentanglement in variational autoencoders, 2019. URL https://arxiv.org/abs/1802.04942.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations, 2020. URL https://arxiv.org/abs/2002.05709.

Chevalier-Boisvert, M., Dai, B., Towers, M., Perez-Vicente, R., Willems, L., Lahlou, S., Pal, S., Castro, P. S., and Terry, J. Minigrid & miniworld: Modular & customizable reinforcement learning environments for goal-oriented tasks. In *Advances in Neural Information Processing Systems 36, New Orleans, LA, USA*, December 2023.

Degris, T., Sigaud, O., and Wuillemin, P.-H. Learning the structure of factored markov decision processes in reinforcement learning problems. In *Proceedings of the 23rd international conference on Machine learning*, pp. 257–264, 2006.

Dunion, M., McInroe, T., Luck, K. S., Hanna, J. P., and Albrecht, S. V. Conditional mutual information for disentangled representations in reinforcement learning, 2023a. URL https://arxiv.org/abs/2305.14133.

Dunion, M., McInroe, T., Luck, K. S., Hanna, J. P., and Albrecht, S. V. Temporal disentanglement of representations for improved generalisation in reinforcement learning, 2023b. URL https://arxiv.org/abs/2207.05480.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. Masked autoencoders are scalable vision learners, 2021. URL https://arxiv.org/abs/2111.06377.

He, Q., Su, H., Zhang, J., and Hou, X. Representation gap in deep reinforcement learning. In *Decision Awareness in Reinforcement Learning Workshop at ICML 2022*, 2022.

Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., Song, D., Steinhardt, J., and Gilmer, J. The many faces of robustness: A critical analysis of out-of-distribution generalization, 2021. URL https://arxiv.org/abs/2006.16241.

Hester, T., Vecerik, M., Pietquin, O., Lanctot, M., Schaul, T., Piot, B., Horgan, D., Quan, J., Sendonaris, A., Osband, I., et al. Deep q-learning from demonstrations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Higgins, I., Matthey, L., Pal, A., Burgess, C. P., Glorot, X., Botvinick, M. M., Mohamed, S., and Lerchner, A. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2016. URL https://api.semanticscholar.org/CorpusID:46798026.

Higgins, I., Amos, D., Pfau, D., Racaniere, S., Matthey, L., Rezende, D., and Lerchner, A. Towards a definition of disentangled representations, 2018. URL https://arxiv.org/abs/1812.02230.

Howard, R. A. Dynamic programming and markov processes. *MIT Press google schola*, 2:39–47, 1960.

Hsu, W.-N., Zhang, Y., and Glass, J. Unsupervised learning of disentangled and interpretable representations from sequential data, 2017. URL https://arxiv.org/abs/1709.07902.

Kingma, D. P. and Welling, M. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019. ISSN 1935-8245. doi: 10.1561/2200000056. URL http://dx.doi.org/10.1561/2200000056.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes, 2022. URL https://arxiv.org/abs/1312.6114.

Klein, L., Carvalho, J. B. S., El-Assady, M., Penna, P., Buhmann, J. M., and Jaeger, P. F. Improving explainability of disentangled representations using multipath-attribution mappings, 2023. URL https://arxiv.org/abs/2306.09035.

Koller, D. and Parr, R. Policy iteration for factored mdps, 2013. URL https://arxiv.org/abs/1301.3869.

Li, F., Zhang, H., xu, H., Liu, S., Zhang, L., Ni, L. M., and Shum, H.-Y. Mask dino: Towards a unified transformer-based framework for object detection and segmentation, 2022a. URL https://arxiv.org/abs/2206.02777.

Li, X., Shang, J., Das, S., and Ryoo, M. Does self-supervised learning really improve reinforcement learning from pixels? *Advances in Neural Information Processing Systems*, 35:30865–30881, 2022b.

Liu, Q., Kuang, Y., and Wang, J. Robust deep reinforcement learning with adaptive adversarial perturbations in action space, 2024. URL https://arxiv.org/abs/2405.11982.

Maćkiewicz, A. and Ratajczak, W. Principal components analysis (pca). *Computers & Geosciences*, 19(3):303–342, 1993.

Meng, L., Goodwin, M., Yazidi, A., and Engelstad, P. Maximum manifold capacity representations in state representation learning, 2024. URL https://arxiv.org/abs/2405.13848.

Misra, I. and van der Maaten, L. Self-supervised learning of pretext-invariant representations, 2019. URL https://arxiv.org/abs/1912.01991.

Mnih, V. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. Human-level control through deep reinforcement learning. *nature*, 518(7540): 529–533, 2015.

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

Pearson, K. Vii. note on regression and inheritance in the case of two parents. *proceedings of the royal society of London*, 58(347-352):240–242, 1895.

Pore, A., Muradore, R., and Dall'Alba, D. Dear: Disentangled environment and agent representations for reinforcement learning without reconstruction, 2024. URL https://arxiv.org/abs/2407.00633.

Puterman, M. L. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

Su, B. and Wu, Y. Learning low-dimensional temporal representations. In *International Conference on Machine Learning*, pp. 4761–4770. PMLR, 2018.

Wang, X., Chen, H., Tang, S., Wu, Z., and Zhu, W. Disentangled representation learning, 2024. URL https://arxiv.org/abs/2211.11695.

Wang, Y., Li, H., Cheng, H., Wen, B., Chau, L.-P., and Kot, A. C. Variational disentanglement for domain generalization. *arXiv preprint arXiv:2109.05826*, 2021.

Yang, R., Xu, H., Wu, Y., and Wang, X. Multi-task reinforcement learning with soft modularization, 2020. URL https://arxiv.org/abs/2003.13661.

Yu, T., Zhang, Z., Lan, C., Lu, Y., and Chen, Z. Mask-based latent reconstruction for reinforcement learning. *Advances in Neural Information Processing Systems*, 35: 25117–25131, 2022.

Zbontar, J., Jing, L., Misra, I., LeCun, Y., and Deny, S. Barlow twins: Self-supervised learning via redundancy reduction, 2021. URL https://arxiv.org/abs/2103.03230.

A. Appendix

A.1. Extended Background: Reinforcement learning & state representations

Reinforcement learning (RL) is a paradigm where an agent takes actions ($a_t \in \mathcal{A}$) to transition between states ($s_t \to s_{t+1}$ according to Eq5) in an environment, receiving rewards (r_t from Eq 6) that guide future decisions, establishing a feedback loop between the agent and environment. Environments are typically formalized as a Markov Decision Process (MDP) (Puterman, 2014) $\mathcal{M} := \{\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \gamma\}$ subject to the Markov assumption(Howard, 1960), which assumes an agent's future state ($s_{t+1} \in \mathcal{S}$) only depends on its current state ($s_t \in \mathcal{S}$) and action ($a_t \in \mathcal{A}$) i.e. environment dynamics are modeled by a transition function ($\mathcal{T}$, Eq5). Agents are formalized as policy learners ($\pi : S \to A$) that aim to maximize expected cumulative reward Eq4.

$$G_t = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k} \,\middle|\, s_t, \pi\right] \tag{4}$$

$$\mathcal{T}(s_t, a_t, s_{t+1}) = P(s_{t+1} \mid s_t, a_t) \tag{5}$$

$$\mathcal{R}(s_t, a_t, s_{t+1}) = \mathbb{E}[r_{t+1} \mid s_t, a_t, s_{t+1}] \tag{6}$$

Thus, choice of state (s_t) representation crucially affects learned transition/reward models and policies. Previous work (Hester et al., 2018)(Liu et al., 2024) has shown Deep RL (Hendrycks et al., 2021) is highly sensitive to perturbations in observations that leave visual features undisturbed. Brittle policies would benefit from disentanglement, that have improved sample complexity(Higgins et al., 2016)(Yang et al., 2020)(Pore et al., 2024)(Wang et al., 2024) and interpretability (Hsu et al., 2017)(Klein et al., 2023) in other domains.

A.2. Reinforcement Learning Tasks

As a benchmark for disentangled representation and long-horizon RL tasks, where sample-efficient is paramount, we learn disentangled representations for DoorKey-8x8 on the MiniGrid environment. This is a goal reaching task, requiring an agent to first `pick_up` a key, then `open` a door before navigating to the goal, as shown in 5. The agent's action space involves left/right rotation, moving forward, pick up and put down.

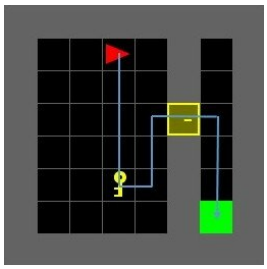

Figure 5. A visual observation from DoorKey-8x8: the red triangle represents the agent, the green square is the goal, grey squares are walls. The agent needs to pick up the yellow key, then open the yellow door whilst holding the key before navigating to the goal

While visual feature exhibit negligible change due to agent actions, the task can be decomposed into 8 interrelated factors i.e. agent position (x_a, y_a), door position (x_d, y_d), key position (x_k, y_k), key holding $\{0, 1\}$, door open $\{0, 1\}$ and goal position (x_g, y_g). Given an 8×8 grid-size, this induces 6 million states in the factored space, making DoorKey-8x8 an appropriately challenging task for disentanglement.

A.3. Experiment Details

A.3.1. Reinforcement Learning

We utilize the `stablebaselines3` package to evaluate policy learning through proximal-policy optimization (PPO). This library uses a Nature CNN (Mnih, 2013) as visual encoder backbone, which we also adopt in our self-supervised methods for fair comparison.

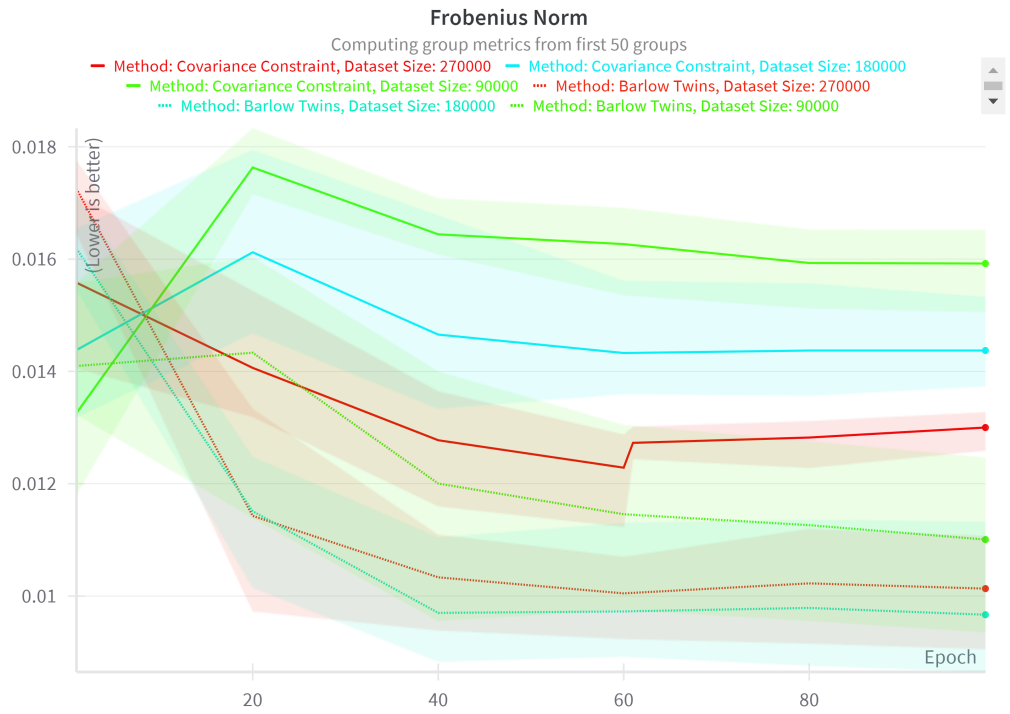

Figure 6. The impact of different pre-training objectives and dataset size on the disentanglement of the learned embeddings across epochs. As we can see, the SSL methods have a limit beyond which they're unable to disentangle the features. The dataset size mentioned here are only the training data size (80% of the total dataset size).

For the expert representation, we used two fully connected linear layers, each with 64 neurons. For the SSL-learned factored representation, we apply two layers of fully connected linear layers, each with 64 neurons, after the learning embedding space, and the weights of the SSL backbone is frozen.

We used the default parameters of PPO, listed below: learning rate $\gamma =$

Parameter	Value
learning rate	3×10^{-4}
γ	0.99
GAE λ	0.95
batch_size	64
clip_range	0.2
n_steps	2048

Table 2. RL PPO parameters

A.3.2. Self-supervised Learning

To demonstrate statistically significant results, we evaluate performance over 3 seeds. Representation learning methods are trained for 100 epochs or until convergence. Downstream policy learning is done over 5×10^6 total steps, which is ablated across 4×10^6,3×10^6 and 2×10^6 policy learning steps for data diet evaluations – with the remainder (1×10^6,2×10^6 and 3×10^6 steps, respectively) going to representation learning.

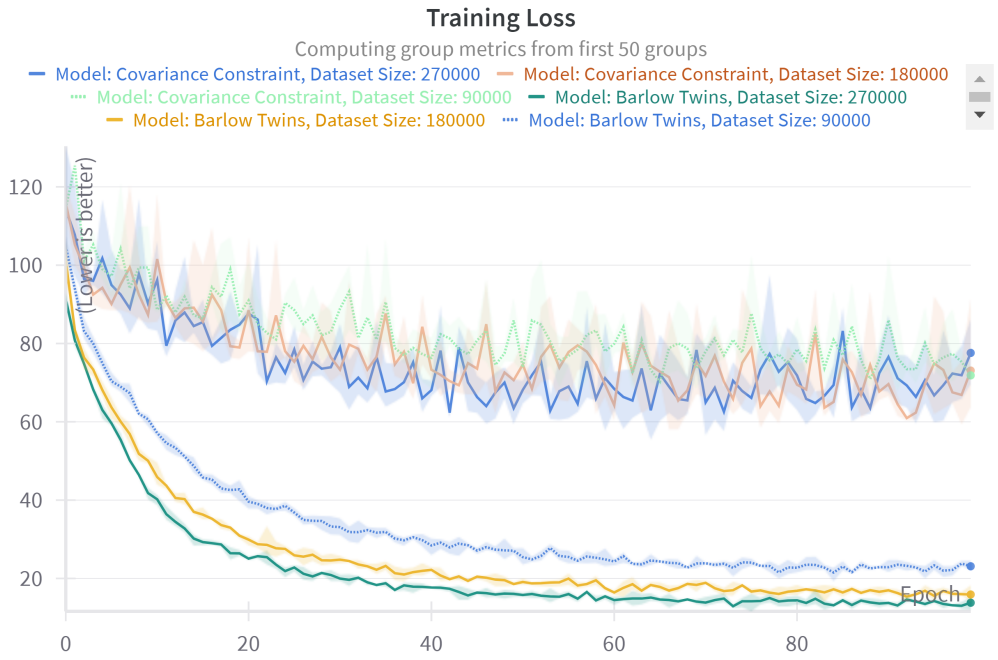

Figure 7. The training loss of different representation learning objectives and dataset sizes. The dataset size mentioned here are only the training data size (80% of the total dataset size).

A.4. Training Dynamics

To causally evaluate where/how disentanglement may emerge in our learned representations, we evaluate a set of disentanglement/distance metrics on layer-specific embeddings within the standard NatureCNN (Mnih et al., 2015) encoder.

We first collect a 2400 sample test-set of image observation pairs, with only 1 expert-factor change between then (e.g. `agent_x`). 100 samples of each factor change, across 12 factors, were uniformly sampled to generate this 2400 test set.

The intermediate embeddings from all convolution and linear layers were extracted and flattened, on which these disentanglement metrics were applied. If stronger disentanglement was to emerge, across representation learning techniques, we hypothesize that a change in 1 factor should induce a comparatively magnified difference in these metrics.

A.5. Author Contributions

* **Shreyas Sundara Raman** initiated the project and conceptualized both method 1 (covariance) and method 2 (masking) after thorough literature review. He also developed a customized dataloader with configurable augmentations and factor changes. Shreyas also implemented the code base for RL training (with Jazlyn) and helped Yichen finetune the RL baselines. He also worked on ablations for learning dynamics, to see where/how disentanglement occurs. He also wrote the introduction, background, methodology (revisions with Chia Hong), abstract (with Jazlyn) results and conclusion.

* **Vipul Sharma** integrated the SSL training framework. He conducted the representation learning experiments for supervised learning, Barlow Twins, and the proposed method 1 (covariance). He also contributed to the ablation experiments measuring the impact of the dataset size used for training and measuring the disentanglement of the learned representations. He also contributed to the writing of the results section.

* **Yichen Wei** implemented the RL training framework and tuned the RL model, ensuring that all baseline models converged. He also wrote the module that loads the trained SSL model and routes the SSL embedding space to the RL agent. He also ran all of the RL-related experiments, edited the background sections, and wrote papers in the results and conclusions section.

* **Chia-Hong Hsu** implemented evaluation metrics to measure disentanglement. He and Shreyas researched the initial methodology to attack this RL problem in the SSL context, which led to our proposed work. He also wrote the methodology

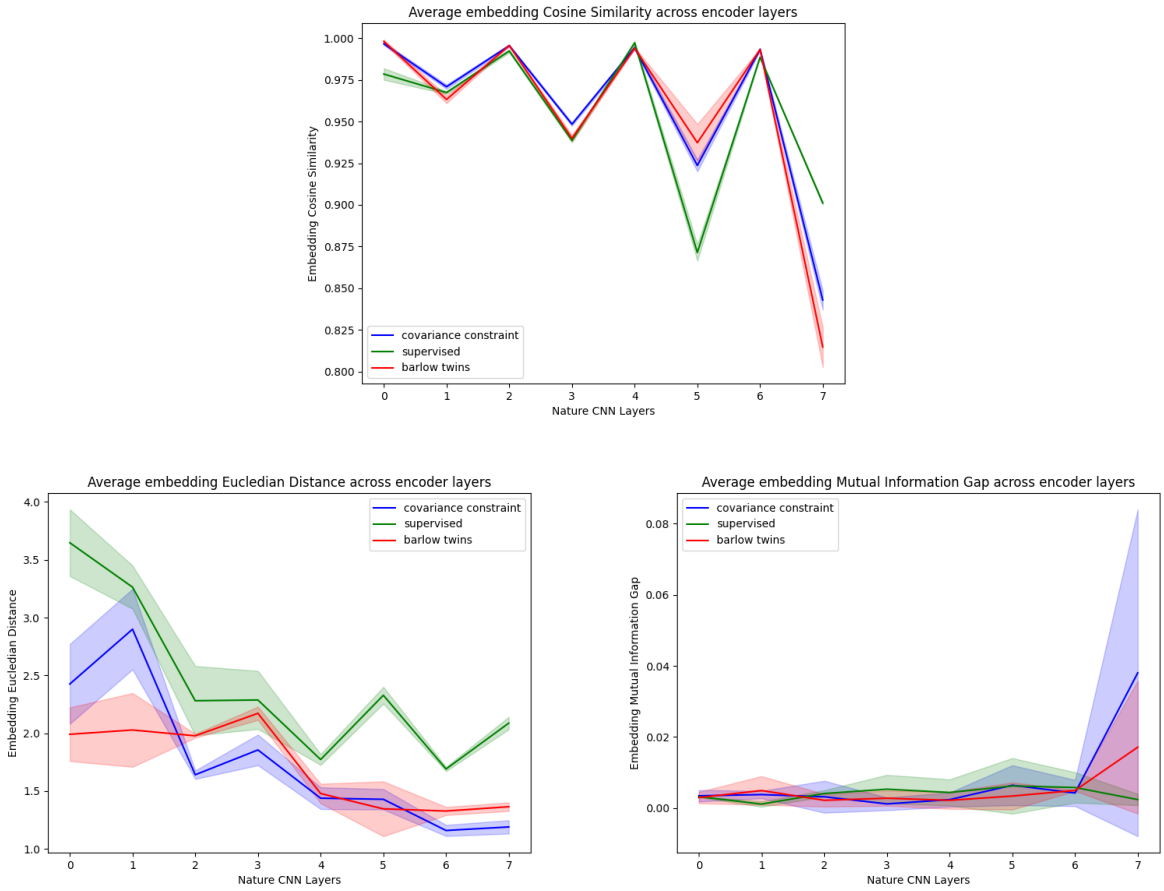

Figure 8. Average measure of disentanglement (cosine similarity, euclidean distance and mutual information gap — counter clockwise) across layers of the shared NatureCNN(Mnih et al., 2015) encoder. Measured using a fixed 2400-sample test set of visual observation pairs that have 1 changing factor between them. More disentangled representations should accentuate differences in the single factor

part and drawn the figures related to the algorithm.

* **Jazlyn Lin** implemented the initial code base for RL training and sample collection code for evaluating disentanglement metric (consulted Shreyas). She also created the presentation slides, wrote the abstract section of the paper and reviewed literatures (with Shreyas).

A.6. Future Proposed Method

A.6.1. Latent Masked Reconstruction

Reviewing approaches that use masked (latent) reconstruction—such as MAE(He et al., 2021), I-JEPA(Assran et al., 2023), Data2Vec(Baevski et al., 2022), DINO(Li et al., 2022a) or others(Yu et al., 2022)—-we find standard paradigms: assymetric student-teacher architectures, use of projector-heads, masking in latent space post-encoding, mean absolute error or euclidean distance losses in latent space. Our method adopts such paradigms, attempting to reconstruct elements in a projected latent state (y_i) using a masked latent state ($\mathcal{M}(x_i)$) with partial information. Masking explicitly enforces sparse inter-dependence between factors, while adhering to transition dynamics defined by factored MDPs, thereby motivating disentanglement.

Given a batch of visual observations ($\{o_1, ..., o_t\}, o_i \in \mathbb{R}^{3\times H\times W}$), a Nature CNN(Mnih et al., 2015) student-encoder ($\phi_s : o_i \to z_i$) generates latent representations ($z_i \in \mathcal{R}^k$) that are used for downstream policy learning. A learned masking ($\mathcal{M}(z_i) \in \mathcal{R}^k$ where $\{0, 1\} \in \mathcal{M}$) is applied to student representations followed by concatenation with one-hot action vectors (a_i). A self-attention head ($h_s(z_i)$) models transition dynamics ($\mathcal{T}$) of the RL task, predicting future latent states

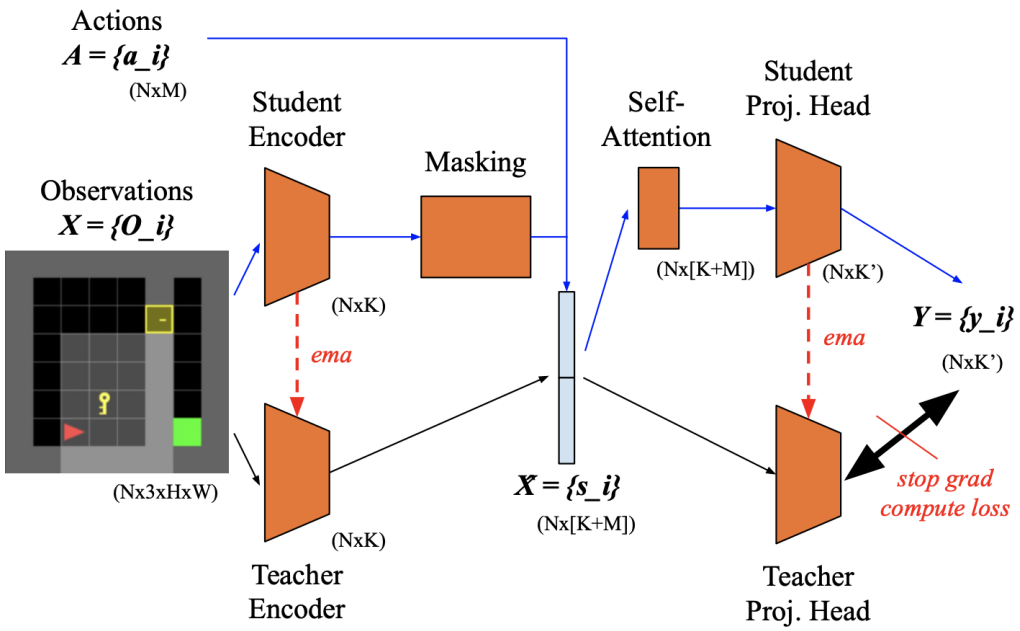

Figure 9. The masked reconstruction method we propose for future work takes a batch of feature-rich observations $\{o_1, ..., o_t\}$ and embeds them into latent representations $z_i = \phi(o_i) \in \mathcal{R}^k$ with k factors. To motivate sparse inter-factor dependence for disentanglement, our loss motivates reconstructing a teacher network's projected state (y_i) from a partially masked student network state representation $\mathcal{M}(z_i)$ – where the masking function is enforced to approximate identity $\mathcal{M} \approx \mathcal{I}$. Stop gradients and momentum updates in the teacher prevent representation collapse.

z_{i+1}. Finally, a linear projection network (g_s) stabilizes gradients and supports generalizable features decoupled from the SSL objective. Masked reconstruction enforces 2 losses: a reconstruction target for l2-loss between $g_s(\hat{z_{i+1}})$ and $g_t(z_{i+1})$ is given by teacher-network encoder (ϕ_t) and projection layer (g_t); disentanglement is motivated by Frobenius Norm of the masking network $\mathcal{M}$. Teacher parameters are updated via momentum i.e. $\theta_{\text{t}} \leftarrow m \cdot \theta_{\text{t}} + (1 - m) \cdot \theta_{\text{s}}$, while stop-gradients on the teacher output prevent representation collapse. See Fig. 9.