Masters Thesis

Learning Disentangled Representations for Deep Reinforcement Learning using Self-Supervised Learning

By

Shreyas Sundara Raman
Sc.B., Brown University, 2024

Thesis
Submitted in partial fulfillment of the requirements for the
Degree of Master of Science in the Department of Computer Science at Brown University

PROVIDENCE, RHODE ISLAND
MAY 2025

This thesis by **Shreyas Sundara Raman** is accepted in its present form
by the Department of Computer Science as satisfying the

thesis requirements for the degree of Master of Computer Science

Date _____ _____

Prof. George D. Konidaris,

Thesis Advisor

Approved by the Director of Graduate Studies in Computer Science

Date _____ _____

Nikos Triandopoulos,

Director of Graduate Studies in Computer Science, Brown University

LEARNING FACTORED REPRESENTATIONS FOR REINFORCEMENT LEARNING USING SELF-SUPERVISED LEARNING

Shreyas S. Raman, Vipul Sharma, Chia-Hong Hsu, Jazlyn Lin, Yichen Wei, Dan Haramati
Department of Computer Science
Brown University
Rhode Island, USA
shreyas_sundara_raman@brown.edu

Randall Balestriero, Stefanie Tellex & George Konidaris
Department of Computer Science
Brown University
Rhode Island, USA

ABSTRACT

Disentangled representations offer a path to sample-efficient and generalizable reinforcement learning (RL) from rich visual observations by exploiting the compositional structure of tasks in robotics or embodied AI—such as goal reaching or object manipulation—requiring agents to learn the 'rules' of complex environments. However, disentanglement is poorly defined for sequential data and struggles to uncover factors underlying task dynamics. Current approaches using self-supervised learning (SSL) for disentanglement in RL promote invariance to visual perturbations or spurious correlations. We propose 2 novel SSL representation learning objectives using the language of Factored MDPs that encourage independence between state-factors through covariance constraints and partially masked forward dynamics, given only visual observations. Our learned representations are evaluated on 3 exploration heavy and multi-factor environments (DoorKey 6x6, FourRooms and BlockedUnlockPickup) demonstrating strong feature disentanglement, compositionality and greater policy generalization.

1 INTRODUCTION

Reinforcement learning (RL) is formalized by a closed-loop agent-environment interaction, where an agent perceives observations, selects optimal actions using its policy, and receives reward feedback. Agents typically rely on high-dimensional observations from general-purpose sensors to interact with their environment, which directly influences data distributions used for learning policies. However, not all visual representations are equally useful for decision making. The distribution of rich observations can be reduced to a set of task-relevant causal variables or 'factors'. For instance, all 13^{64} positions on a chess-board with any visual variation can be reduced to 64 factors representing the piece on every (x, y) square. Rather than making decisions using an appropriately distilled representation, the agent is forced to use a problem-agnostic sensor.

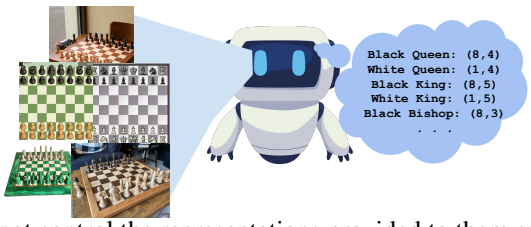


Figure 1: RL agents cannot control the representations provided to them and not all representations are useful for decision making. Though large distributions of rich visual observations can be reduced to a set of causal factors useful for decision making.

Disentanglement attempts to tackle this by recovering distilled representations from feature-rich observations, where each component captures factors that change independently of one another and are relevant to the target task. Within RL, these such representations could capture causal factors governing transition or reward dynamics of a task, which policy learning algorithms can exploit for more sample-efficient learning, more generalized extrapolation to out-of-distributions tasks governed by similar factors or generalized interpolations to variations of factors within a target task. Thus, disentangled representations are crucial for consistently extracting the most-relevant information for tasks in visual RL, where an agents' data distributions are feature-rich and continuously shifting based on exploration of the environment. focus on a narrow set of high-reward states Traditional disentanglement metrics, however, are non-conducive to RL. Factors to recover in RL are often unknown a-priori or task-dependent (non-global). The agent's action history introduces non-stationarity/correlation to the distribution of factors i.e. the data distribution is no longer i.i.d. and dependent on an evolving policy. Also, as policy learning evolves, data distributions gradually skew towards a subset of high-reward states. This makes disentanglement ill-defined and challenging in RL, with practitioners often pre-defining desirable factors Dunion et al. (2023a)Träuble et al. (2021)Locatello et al. (2019). At the same time, useful disentanglement makes policy learning more effective and requires the right level of abstraction or information removal, delicately balancing representation learning with policy learning Dunion et al. (2023b). Remove too little and noise irrelevant to the task is preserved; remove too much and the representation cannot accurately capture environment dynamics. This underscores critical challenges for disentanglement in RL:

- How can we extract task-relevant factors not previously specified from visual observations?
- To what extent can such factors generalize across tasks?
- What is the optimal degree of information to discard when learning useful abstractions?

Most related works learning disentangled representations for RL target invariance to visual distribution shifts and spuriously correlated features, or decompose environment rewards across factors for approximately factored value functions. We focus on uncovering independent factors governing dynamics using disentanglement, and turn to factored MDP definitions of sparse inter-dependence. Both proposed approaches leverage self-supervised learning (SSL) to motivate disentanglement, given only visual observations with no assumptions on RL tasks, thereby eliminating reliance on pre-defined factors. One method sparsifies the covariance between latent features, while the other encourages consistent state transitions by applying a sparse mask that models inter-factor dependence. To our knowledge, we are the first to approach disentanglement of visual observations using SSL for RL by explicitly enforcing factored MDP constraints for transition dynamics. We evaluate our learned representations on 3 MiniGrid environments (DoorKey-6x6, FourRooms and BlockedUnlockPickup across 4 seeds), showing they align with factored MDP definitions for disentanglement and [to be written...].

2 BACKGROUND

2.1 FACTORED MDPS & THE REPRESENTATION GAP

The disparity in RL performance between rich, high-dimensional observations and distilled expert states induces a '*representation gap*' Chandak et al. (2019); He et al. (2022); Allen (2023), where the latter leads to $5 - 20\times$ more sample-efficient policy learning Zhang et al. (2018). Bridging the gap requires learning lower-dimensional representations that discard irrelevant information but preserve enough to fully characterize state environment dynamics. Maćkiewicz & Ratajczak (1993); Kingma & Welling (2022; 2019); Su & Wu (2018); Allen (2023). Rajeswaran et al. (2017) have also found relative improvements using structured low-dimensional representations than entangled deep network representations.

Our work adopts Factored Markov Decision-Process (F-MDP) definitions Boutilier et al. (1999); Koller & Parr (2013); Higgins et al. (2016), where a standard MDP transition function on a a single monolithic state (s_t) instead depends on a set of finite state-variables or *factors* ($\mathcal{K}$) – modeled as a *dynamic Bayesian network* Murphy (1998). This induces a *factored representation* with 'focused causes', where changes to a factor $s_{t+1}[i]$ are fully explained by a sparse subset of *parent factors* $\rho(s_{t+1}[i]) \in \{s_t[j]\}$ for $j \in [1, k]$. An F-MDP transition function follows (Eq1); as the maximum number of *parent factors*, $\mathcal{P} := \max \rho(s_{t+1}[i]), \forall i$ reduces relative to *total number* of factors

($\mathcal{P} << \mathcal{K}$) the representation becomes *more factored* – until it is disentangled i.e. $\rho(s_{t+1}[i]) = s_t[i]$. See Fig 2.

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

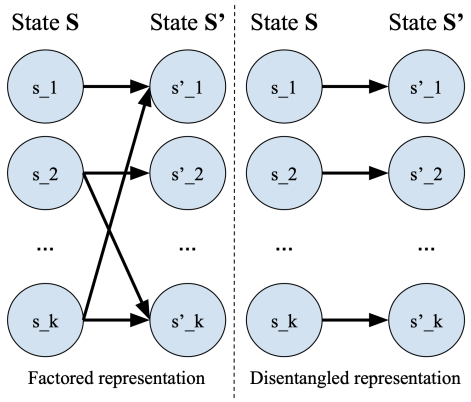


Figure 2: Visualization of a factored and disentangled MDP with $\mathcal{K}$ factors. In this example, $\rho(s_i)$ are parent factors; $\rho(s_1) = \{s_1, s_2\}$ for factored and $\rho(s_1) = \{s_1\}$ for disentangled representations

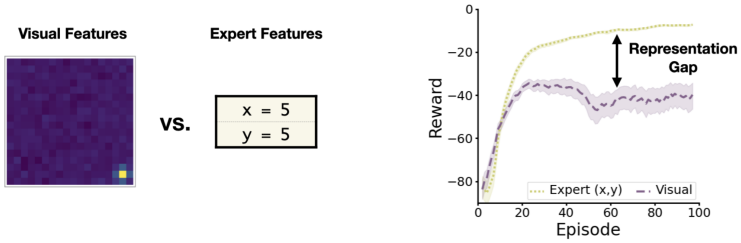


Figure 3: The representation gap (Allen, 2023) captures difference in policy learning observed between using raw visual observations and pre-distilled expert features – where the latter leads to faster learning. Our methods hope to learn visual encoders ϕ whose latent representation bridges this gap

2.2 Disentangled Representations in RL & Other Domains

The pursuit of disentangled latent representations has been central across both model-based Oord et al. (2018); Hafner et al. (2022) and model-free Dunion et al. (2023b;a); Anand et al. (2019) policy learning frameworks, and is typically approached by minimizing mutual information between latent variables through InfoNCE Oord et al. (2018) or Kullback–Leibler divergence losses. Similar definitions of disentanglement have been used to separate static/factors for time series modeling Fotiadis et al. (2023), leading to 33% smaller absolute error within/outside training distribution.

Other task-relevant abstractions, including object centric Feng & Magliacane (2023), graphical Balaji et al. (2021) and mixture-of-expert Eigen et al. (2014) representations, have been explored as alternatives to disentangled latent representations. However, these approaches assume prior knowledge on factor inter-dependencies and object-level attributes or leverage specific architectures for static factorization that may not be semantically related to dynamics.

2.3 SSL for Disentanglement in RL

The manifold hypothesis Meng et al. (2024) posits that high-dimensional data distributions ($\mathcal{X}(\Theta)$) lie on a lower-dimensional manifold, which is defined by a comparatively small set of intrinsic coordinates (Θ). Manifold learning 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 the problem setting of disentanglement in RL with visual observations, Θ represents causal factors (e.g. agent (x, y) or obstacle (x, y)) underlying rich-visual observation distributions and state dynamics. Thus our approach broadly aims to learn $\mathcal{X}^{-1} : \mathcal{X}(\Theta) \rightarrow \Theta$ a function mapping feature-rich images to their disentangled causal factors.

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

Specifically within RL, works like Dunion et al. (2023a) introduced conditional independence (CI) to disentangle spurious correlations between color and transition dynamics. Also, 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 offered by image augmentation invariance through SSL objectives in RL is being debated Li et al. (2022). Most related to our work, AFaR Sodhani et al. (2023) enforces additive decomposition of a state's value $V^{\pi}(s) = \sum V_i^{\pi}(s_i)$ across learned factors using A2C loss without ground-truth factor supervision. MoCoDA Pitis et al. (2022) achieves provable improvements in sample complexity by leveraging either learned or pre-specified masks $M(s, a)$, defining local inter-factor dependencies over known entity-structured factors. These masks enable targeted sampling of counterfactual transitions for unseen combinations of local factors. DARLA Higgins et al. (2017) attempts policy generalization across visual domains by separating representation from policy learning; the approach uses a β-VAE for unsupervised disentanglement of static generative factors (e.g. color, objects, texture), learning factorized encodings that generalize across domains but do not account for temporal or dynamic dependencies.

Overall, previous works achieve disentanglement through mutual-information, conditional mutual-information, or contrastive learning. Generally, disentanglement is used to build invariance to perturbations; though some works disentangle auxilliary value functions/masks supporting towards policy generalization or sample-efficiency. Our methods align more strongly with Factored-MDP definition of disentaglement by sparsifying inter-factor dependencies during latent state dynamics.

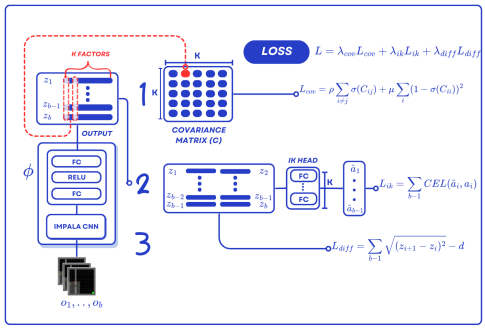
3 METHODOLOGY

Given a feature-rich visual observation, our proposed representation learning methods use self-supervision to disentangle factors underlying state dynamics. Representation and policy learning occur in parallel using an ImpalaCNN Espeholt et al. (2018) and Proximal Policy Learning (PPO) Schulman et al. (2017), respectively. A batch of visual observations ($\{o_1, ..., o_t\}$) are encoded into latent representations ($\phi : o_i \in \mathbb{R}^{3\times H\times W} \rightarrow z_i \in \mathbb{R}^k$) then used for downstream policy learning – where k defines the number of factors in the representation. To evaluate the degree of disentanglement and the policy compositionality/generalization offered by the learned representations, we overview experimental setup and evaluation metrics in Section 4.

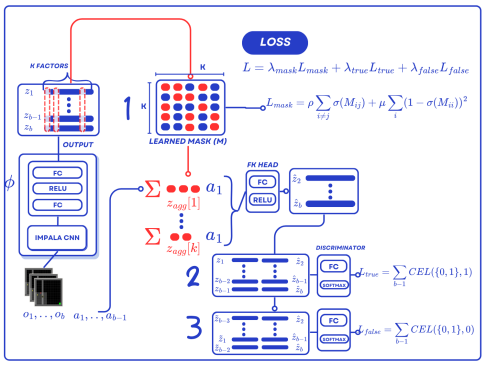
3.1 COVIK: COVARIANCE & INVERSE KINEMATICS

Following a Barlow Twins Zbontar et al. (2021) objective, this method attempts to minimize the Frobenius Norm loss of the covariance matrix ($\mathcal{C}$) across all pairs of latent embeddings of visual observations $z_i := \phi(o_i)$, by bringing the matrix closer to identity ($C \approx \mathcal{I} \in \mathcal{R}^{k\times k}$). This motivates learned factors to be statistically independent of one another within a batch.

To prevent representation collapse, where $\mathcal{C} = \mathcal{I}$, we utilize an inverse-kinematic network ($I(z_i, z_{i+1}) = \hat{a}_i$) to predict the discrete action used to transition between successive latent states, updated by a cross-entropy loss. This auxiliary objective motivates the inverse-kinematic network ik_θ to encode transition dynamics ($\mathcal{T}$) of the RL task, under Markov assumption, motivating disentangled representations (z_i) to also be relevant for policy learning. Additionally, a factor-specific smoothness loss following Allen (2023) promotes *focused effects* by penalizing when multiple factors change together. We also allow the PPO A.1 Q-function loss to shape the representation z_i. See Fig. 4a and loss function in Eq2.



(a) CovIK Method: disentanglement in $\phi(o_i)$ is motivated by aligning the covariance matrix ($\mathcal{C}$) between each pair of k factors with an identity matrix ($\mathcal{I} \in \mathbb{R}^{k \times k}$); mode collapse ($\mathcal{C} = \mathcal{I}$) is prevented by decoding actions between successive state-pairs using an inverse-kinematics projector $\hat{a_i} := ik_\theta(z_i, z_{i+1})$



(b) MaskedFD Method: disentanglement in $\phi(o_i)$ is motivated by sparsifying a learned mask M that defines the subset of current state factors $M(z_i)$ used to estimate factors in future latent states z_{i+1}; mode collapse is prevented by differentiating valid future latent states (z_{i+1}) from shuffled future latent states ($z_{\tilde{i+1}}$)

Figure 4: Overview of proposed methods. A batch of sequential visual observations $\{o_1, ..., o_t\} \in \mathbb{R}^n$ are embedded to latent representations $z_i = \phi(o_i) \in \mathbb{R}^k, k \ll n$ with k factors.

3.2 MaskedFD: Masked Forward Dynamics

Following locally modular dynamics in MoCoDA Pitis et al. (2022), this method attempts factored latent transitions ($F : z_{agg}[j], a_i \rightarrow \hat{z_{i+1}}[j]$) between aggregated latent features in the current latent state ($z_{agg}[j] = M[z_i] = \Sigma_m z_i[m] M_{j,m}$) and actions ($a_i$) to the j^{th} factor in the future latent state ($\hat{z_{i+1}}[j]$). A learned mask $M \in [0, 1]^{k \times k}$ explicitly models inter-factor dependencies by controlling the weightages of each *parent factor* as they are aggregated into z_{agg}. A Frobenius Norm loss on the learned mask (M) enforces disentanglement following FactoredMDP (Eq 1) transitions.

To prevent representation collapse, where $M = \mathcal{I}$, a binary discriminator network ($D(\hat{z_{i+1}}, z_{i+1}) = \{0, 1\}$) predicts whether transitions are real ($\hat{z_{i+1}}, z_{i+1}$) or shuffled ($\hat{z_{i+1}}, \tilde{z_{i+1}}$), updated by a cross-entropy loss. This auxiliary objective on motivates the factor-wise representation ($\hat{z_{i+1}}$) to capture identical information encoded directly in representation z_{i+1}. We also allow the PPO A.1 Q-function loss to shape the representation z_i. See Fig. 4a and loss function in Eq3.

$$\begin{aligned} \mathcal{L}_{\text{covik}} = \lambda_{\text{cov}} & \left(\rho \sum_{i \neq j} \sigma(C_{ij}) + \mu \sum_i (1 - \sigma(C_{ii}))^2 \right) \\ & + \lambda_{\text{ik}} \, CE(IK(\phi(o_{i+1}), \phi(o_i)), a_i) \\ & + \lambda_{\text{diff}} \left(\sqrt{(\phi(o_{i+1}) - \phi(o_i))^2} - d \right) \end{aligned} \tag{2}$$

where C is covariance across factor pairs, d is a margin penalizing large factor changes, and CE is the cross-entropy loss.

$$\begin{aligned} \mathcal{L}_{\text{maskfd}} = \lambda_{\text{mask}} & \left(\rho \sum_{i \neq j} \sigma(M_{ij}) + \mu \sum_i (1 - \sigma(M_{ii}))^2 \right) \\ & + \lambda_{\text{true}} \, CE(D(F(M[\phi(o_i)]), z_{i+1}), 1) \\ & + \lambda_{\text{false}} \, CE(D(F(M[\phi(o_i)]), \tilde{z}_{i+1}), 0) \end{aligned} \tag{3}$$

where M is the learned mask capturing inter-factor dependence and CE is cross-entropy loss

3.3 Baselines

As defined in 2.1, our encoder $\phi : o_i \rightarrow z_i$ uses *self-supervision* to learn a disentangled *representation (z_i)* that bridges the representation gap.

To measure improvements from enforcing latent disentanglement, we learn PPO policies directly on rich visual observations (o_i). As an empirical upper-bound, we also learn PPO policies directly on expert representations (z_i^*) that are disentangled by construction. To quantify benefits from self-supervision, we design a supervised but factored baseline where d-dimensions of a latent representation $\phi(o_i)$ are split across k factors, with each $\frac{d}{k}$ subspace trained to predict a specific ground-truth expert factor in z_i^*; this encourages modular disentanglement with supervision – gauging the benefits of our self-supervised approach. Details on architecture design and hyperparameters can be found in Appendix A.4 and A.5.

4 Experimental Setup

4.1 Reinforcement Learning Tasks

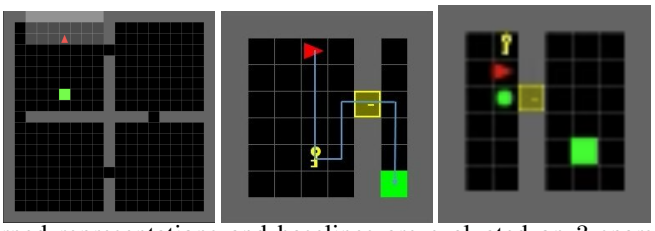


Figure 5: The learned representations and baselines are evaluated on 3 sparse-reward MiniGrid environments i.e. FourRooms-19×19, DoorKey-6×6, BlockedUnlockPickup-6×10

Our learned representations and those of baseline methods are evaluated on 3 sparse-reward MiniGrid Chevalier-Boisvert et al. (2023) environments (Fig. 5). These environments have discrete states and actions (turn left, turn right, forward) but exhibit compositional structure –requiring navigation, object interaction (e.g. with keys, doors, obstacles), and conditional subtask execution – making them ideal environments for the simplest forms of discrete factored representations. Each environment is chosen for a specific evaluative purpose

- FourRooms: has the largest number of factors (13 with 2 agent pos, 1 agent direction, 2 goal pos, 2×4 wall gap pos) yet has the smallest state space i.e. $\approx 1,400$ states. Serves as a standard baseline for recovering several factors with limited interdependence and for sample-efficiency of learning disentangled representations
- DoorKey: has the least number of factors (9 with 2 agent pos, 1 agent direction, 2 goal pos, 2 key pos, 1 door open, 1 door locked, 1 holding key) but has significantly larger state space i.e. $\approx 20k$ states. Serves as a benchmark for direct generalization (without fine-tuning) across starting states and scale to DoorKey-8×8
- BlockedUnlockPickup: has both a large number of factors (11 with DoorKey factors plus 2 obstacle pos) and the largest state space i.e. $\approx 1bn$ states. Serves as a challenging benchmark for recovering multiple heavily interdependent factors; it also captures whether the learned factors are compositional

4.2 Compositionality Across Number of Factors

Given 2 tasks (T_1 and T_2) with identical objectives but a different number of underlying factors and their inter-dependencies, specifically where T_2 has $\subseteq$ of factors in T_1, a disentangled representation should support *compositional* generalization across T_1 and T_2. For instance, in BlockedUnlockPickup, disentangled representations learned across a curriculum of subgoals (e.g., pick up key, open door) could be composed to solve longer-horizon tasks (e.g., pick up key *and* open door) without retraining. Conversely, a policy trained on the full compound task in BlockedUnlockPickup could zero-shot transfer to simpler subtasks (e.g., only pick up key) or to environments (e.g., DoorKey-6×6) that omit some factors (2 factors for obstacle position). This transfer is possible because disentangled representations isolate the dynamics of specific factors into subspaces of the global latent representation, $\mathbb{R}^f \subseteq \mathbb{R}^k$, such that each factor's dynamics can be interpreted and combined independent of the other factors.

We evaluate **(i) bottom-up compositionality:** learning representations for a curriculum of tasks (i.e., unblock door, pick up key, open door, reach goal) then evaluating zero-shot representation

transfer to BlockedUnlockPickup; **(ii) top-down compositionality:** learning representations for BlockedUnlockPickup then evaluating zero-shot policy transfer to sub-goals (i.e., unblock door, pick up key, open door, reach goal); **(iii) task generalization:** learning representations for BlockedUnlockPickup and transferring representation to DoorKey 6×6. We report the following metrics to assess whether representations support localized credit assignment and generalize to subsets of latent features:

- *zero-shot success rate*: the fraction of 20 episodes achieving inference-time task
- *step-weighted reward*: $\frac{1}{T}\sum_{t=1}^{T} r_t$, that accounts for solution efficiency, penalizing inefficient policies, with a $0-1$ reward

4.3 Compositionality Across Scale & State

Given 2 tasks (T_1 and T_2) that share identical number of factors and factor inter-dependencies, but differ in observation scale (factors take on larger range of values) or the initial configuration of those factors, a disentangled representation should generalize across T_1 and T_2. For example, a disentangled representation for DoorKey-6×6 should transfer zero-shot to DoorKey-8×8 provided the visual structure and factor semantics (e.g., position of agent, key, door, goal) are preserved. Similarly in DoorKey-6×6, once a disentangled representation captures the dynamics of individual factors, it should generalize to out-of-distribution configurations of these factors (e.g., goal in a novel location or agent starting behind a locked door) during test time. Given the factors and their dynamics remain constant, the policy should be able to reuse learned factor-dynamics without having to re-learn the task.

We evaluate **(i) scale compositionality:** learning representations for DoorKey-6×6 with padding (for identical visual observation dimensions) then evaluating zero-shot policy transfer to DoorKey-8×8; **(ii) state compositionality:** learning representations for DoorKey-6×6 then evaluating zero-shot policy transfer to 500 randomly sampled initial factor configurations. Similar to compositionality in Sec 4.2, we report *zero-shot success rate* and *step-weighted reward* to assess modularity of learned disentangled representations.

4.4 Disentanglement Metrics

Finally, to evaluate fundamental properties of latent representation disentanglement, we consider 2 metrics: 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.

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \tag{4}$$

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) Eq5 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 is independently captured by one independent/relevant latent dimension. Refer to Carbonneau et al. (2022) for further detail.

$$\text{MIG}(X,Z) = \frac{1}{n}\sum_{i=1}^{n}\left(I(X_i;Z) - \max_{j \neq i} I(X_j;Z)\right) \tag{5}$$

REFERENCES

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

Ankesh Anand, Evan Racah, Sherjil Ozair, Yoshua Bengio, Marc-Alexandre Côté, and R Devon Hjelm. Unsupervised state representation learning in atari. *arXiv preprint arXiv:1906.08226*, 2019.

Bharathan Balaji, Petros Christodoulou, Xiaoyu Lu, Byungsoo Jeon, and Jordan Bell-Masterson. Factored{rl}: Leveraging factored graphs for deep reinforcement learning, 2021. URL https://openreview.net/forum?id=wE-3ly4eT5G.

C. Boutilier, T. Dean, and S. Hanks. 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.

Christopher P. Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β-vae, 2018. URL https://arxiv.org/abs/1804.03599.

Marc-André Carbonneau, Julian Zaidi, Jonathan Boilard, and Ghyslain Gagnon. Measuring disentanglement: A review of metrics, 2022. URL https://arxiv.org/abs/2012.09276.

Yash Chandak, Georgios Theocharous, James Kostas, Scott Jordan, and Philip Thomas. Learning action representations for reinforcement learning. In *International conference on machine learning*, pp. 941–950. PMLR, 2019.

Ricky T. Q. Chen, Xuechen Li, Roger Grosse, and David Duvenaud. Isolating sources of disentanglement in variational autoencoders, 2019. URL https://arxiv.org/abs/1802.04942.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020. URL https://arxiv.org/abs/2002.05709.

Maxime Chevalier-Boisvert, Bolun Dai, Mark Towers, Rodrigo Perez-Vicente, Lucas Willems, Salem Lahlou, Suman Pal, Pablo Samuel Castro, and Jordan Terry. 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.

Mhairi Dunion, Trevor McInroe, Kevin Sebastian Luck, Josiah P. Hanna, and Stefano V. Albrecht. Conditional mutual information for disentangled representations in reinforcement learning, 2023a. URL https://arxiv.org/abs/2305.14133.

Mhairi Dunion, Trevor McInroe, Kevin Sebastian Luck, Josiah P. Hanna, and Stefano V. Albrecht. Temporal disentanglement of representations for improved generalisation in reinforcement learning, 2023b. URL https://arxiv.org/abs/2207.05480.

David Eigen, Marc'Aurelio Ranzato, and Ilya Sutskever. Learning factored representations in a deep mixture of experts, 2014. URL https://arxiv.org/abs/1312.4314.

Lasse Espeholt, Hubert Soyer, Remi Munos, Karen Simonyan, Volodymir Mnih, Tom Ward, Yotam Doron, Vlad Firoiu, Tim Harley, Iain Dunning, Shane Legg, and Koray Kavukcuoglu. Impala: Scalable distributed deep-rl with importance weighted actor-learner architectures, 2018. URL https://arxiv.org/abs/1802.01561.

Fan Feng and Sara Magliacane. Learning dynamic attribute-factored world models for efficient multi-object reinforcement learning, 2023. URL https://arxiv.org/abs/2307.09205.

Stathi Fotiadis, Mario Lino, Shunlong Hu, Stef Garasto, Chris D Cantwell, and Anil Anthony Bharath. Disentangled generative models for robust prediction of system dynamics, 2023. URL https://arxiv.org/abs/2108.11684.

Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering atari with discrete world models, 2022. URL https://arxiv.org/abs/2010.02193.

Qiang He, Huangyuan Su, Jieyu Zhang, and Xinwen Hou. Representation gap in deep reinforcement learning. In *Decision Awareness in Reinforcement Learning Workshop at ICML 2022*, 2022.

Irina Higgins, Loïc Matthey, Arka Pal, Christopher P. Burgess, Xavier Glorot, Matthew M. Botvinick, Shakir Mohamed, and Alexander Lerchner. 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.

Irina Higgins, Arka Pal, Andrei Rusu, Loic Matthey, Christopher Burgess, Alexander Pritzel, Matthew Botvinick, Charles Blundell, and Alexander Lerchner. Darla: Improving zero-shot transfer in reinforcement learning. In *International conference on machine learning*, pp. 1480–1490. PMLR, 2017.

Irina Higgins, David Amos, David Pfau, Sebastien Racaniere, Loic Matthey, Danilo Rezende, and Alexander Lerchner. Towards a definition of disentangled representations, 2018. URL https://arxiv.org/abs/1812.02230.

Diederik P. Kingma and Max Welling. 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.

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

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

Xiang Li, Jinghuan Shang, Srijan Das, and Michael Ryoo. Does self-supervised learning really improve reinforcement learning from pixels? *Advances in Neural Information Processing Systems*, 35:30865–30881, 2022.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations, 2019. URL https://arxiv.org/abs/1811.12359.

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

Li Meng, Morten Goodwin, Anis Yazidi, and Paal Engelstad. Maximum manifold capacity representations in state representation learning, 2024. URL https://arxiv.org/abs/2405.13848.

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

Kevin P. Murphy. Dynamic bayesian networks: Representation, inference and learning. Technical Report Technical Report, University of California, Berkeley, 1998. URL https://www.cs.ubc.ca/~murphyk/Papers/dbnintro.pdf.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Karl Pearson. 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.

Silviu Pitis, Elliot Creager, Ajay Mandlekar, and Animesh Garg. MocoDA: Model-based counterfactual data augmentation. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=w6tBOjPCrIO.

Aravind Rajeswaran, Kendall Lowrey, Emanuel V. Todorov, and Sham M Kakade. Towards generalization and simplicity in continuous control. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/9ddb9dd5d8aee9a76bf217a2a3c54833-Paper.pdf.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

Shagun Sodhani, Sergey Levine, and Amy Zhang. Improving generalization with approximate factored value functions. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=LwEWrrKyja.

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

Frederik Träuble, Elliot Creager, Niki Kilbertus, Francesco Locatello, Andrea Dittadi, Anirudh Goyal, Bernhard Schölkopf, and Stefan Bauer. On disentangled representations learned from correlated data. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 10401–10412. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/trauble21a.html.

Yufei Wang, Haoliang Li, Hao Cheng, Bihan Wen, Lap-Pui Chau, and Alex C Kot. Variational disentanglement for domain generalization. *arXiv preprint arXiv:2109.05826*, 2021.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction, 2021. URL https://arxiv.org/abs/2103.03230.

Chiyuan Zhang, Oriol Vinyals, Remi Munos, and Samy Bengio. A study on overfitting in deep reinforcement learning, 2018. URL https://arxiv.org/abs/1804.06893.

A Appendix