Extended Abstract Track

# Homomorphism AutoEncoder — Learning Group Structured Representations from Observed Transitions

**Hamza Keurti**[1,2,3] HAMZA.KEURTI@TUEBINGEN.MPG.DE
**Hsiao-Ru Pan**[1] HPAN@TUEBINGEN.MPG.DE
**Michel Besserve**[1] MICHEL.BESSERVE@TUEBINGEN.MPG.DE
**Benjamin Grewe**[2] BGREWE@ETHZ.CH
**Bernhard Schölkopf**[1] BS@TUEBINGEN.MPG.DE
[1] *Max Planck Institute for Intelligent Systems, Tübingen, Germany*
[2] *Institute of Neuroinformatics, ETH Zürich, Switzerland*
[3] *Max Planck ETH Center for Learning Systems*

**Editors:** Sophia Sanborn, Christian Shewmake, Simone Azeglio, Arianna Di Bernardo, Nina Miolane

## Abstract

It is crucial for agents, both biological and artificial, to acquire world models that veridically represent the external world and how it transforms under the agent's own actions. We consider the case where such transformations are a subset of the group structuring the world state space. We use tools from representation learning and group theory to learn latent representations that account for both sensory information and the actions that alters it during interactions. We introduce the Homomorphism AutoEncoder (HAE), an autoencoder equipped with a learned group representation linearly acting on its latent space trained on 2-step transitions to implicitly enforce the group homomorphism property on the action representation. Compared to existing work, our approach makes fewer assumptions on the group representation and on which transformations the agent can sample from. We motivate our method theoretically, and demonstrate empirically that it can learn the correct representation of the groups and the topology of the environment. We also compare its performance in trajectory prediction with previous methods.

**Keywords:** representation learning, equivariance, homomorphism.

## 1. Introduction

Developmental psychologist Piaget (1964) states that a representation on an object is an understanding of how it transforms under different interactions instead of a copy of it. These representations of the world are learned from interaction starting from the first stage of development: the sensorimotor phase, and that these representations of the world are an understanding of how the world transforms under performed and imagined interventions.

The concept of group, as a mathematical structure, is pervasive in the description of the states and properties of the world. In physics, Noether's theorem shows the continuous symmetries of a system correspond to conserved quantities (Thompson and Cook, 1995). Conversely, conserved quantities are associated with symmetries that only act on them while keeping everything else constant. With the aim of representation learning being to recover these quantities from observed stimuli (Bengio et al., 2012), it is appropriate for the representation to transport the existing symmetries of the world states into the model.

© 2022 H. Keurti, H.-R. Pan, M. Besserve, B. Grewe & B. Schölkopf.

Extended Abstract Track

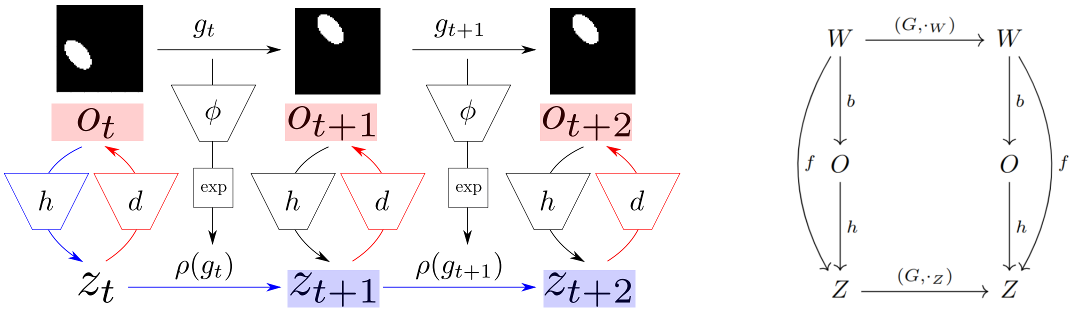


Figure 1: Left: The proposed HAE consisting of $h$, $d$ and $\rho = \exp \circ \phi$, relies on 2-step latent prediction to jointly learn the group representation $\rho$ and the observation representation $h$. Right: Commutative diagram for a group-structured representation.

Such a group-structured representation satisfies an equivariance property between the true generating factors and the learned representations. To observe these symmetries, an agent can perform motor interactions to intervene on the world's state. These interactions also organize in a group which can be used to parametrize the symmetries of the real world.

While one can mathematically create infinitely many such representations, an agent with bounded computational abilities needs to choose one allowing efficient manipulation, interpretation and prediction of changes in its environment. A representational property compatible with this desideratum is disentanglement (Bengio et al., 2012; Kulkarni et al., 2015) which states the latent representation decomposes into parts reflecting the different interpretable properties of the environment that the agent can modify independently.

While a group theoretic account of disentangled representations has been proposed by Higgins et al. (2018) without proposing a learning algorithm, few attempts have followed (Caselles-Dupré et al., 2019; Quessard et al., 2020), all assuming that the agent is able to separately intervene on independent factors of variation.

In this work, we propose the Homomorphism Autoencoder (HAE) framework to jointly learn a group representation of *transitions* between world states, as well as a group-structured disentangled representation of observations with minimal assumptions on the group or the actions the agent can sample from. We show theoretically and experimentally that the HAE learns the group structure of the set of transitions. In addition, the HAE learns to separate the pose of an object from the identity of acted-on objects, which can be identified as orbits of the non-transitive group action.

## 2. Background

### 2.1. Group-Structured Representation Learning

Following the group theoretic formalism introduced by Higgins et al. (2018) (we refer to Appendix A for background on groups), we assume the set of observations $O \subset \mathbb{R}^m$ is obtained from a set of world states $W$ through an unknown *generative process* $b : W \to O$, that we assume to be injective for the scope of this work. An *inference process* $h : O \to Z$ maps observations to their vector representations in $Z \subset \mathbb{R}^k$.

A group of symmetries $G$ structures the world states set by its action $\cdot_W : G \times W \to W$. The different semantic properties of the world can each be thought of as induced by the action of a subgroup in the direct product decomposition of $G = G_1 \times ... \times G_n$. Where, each subgroup only transforms a specific latent property while keeping all others constant. We focus in this manuscript on developing how we learn group structured representation and we leave discussion around disentangling the representation to the Appendices A.2 and C. A representation is a group-structured representation if the diagram in Figure 1 commutes. It then satisfies:

1. There is a (non-trivial) action of $G$ on $Z$, i.e., $\cdot_Z : G \times Z \to Z$.
2. The composition $f = h \circ b : W \to Z$ is equivariant, meaning that transformations of $W$ are reflected on $Z$, i.e., $\forall g \in G, w \in W, \quad f(g \cdot_W w) = g \cdot_Z f(w)$.

A group action $\cdot_Z : G \times \ Z \to \ Z$ induces a group homomorphism $\rho : G \to Sym(Z)$ where $Sym(Z)$ is the group of invertible mappings from $Z$ to itself (more in Appendix A.1). We require the group action $\cdot_Z$ on $Z$ to be linear, in which case, the induced homomorphism $\rho : G \to GL(Z)$ is called a group representation.

## 3. HAE Architecture and Derivation

### 3.1. The Homomorphism Autoencoder (HAE)

To jointly learn the latent representation $h$ of the observations and the group representation $\rho$, we introduce the HAE, described in Figure 1. An autoencoder equipped with a learnable mapping $\rho : G \to GL(Z)$ which maps an observed transition $g$ to an invertible matrix $\rho(g)$. The obtained matrix transforms encoding vectors of observations $z_i = h(o_i)$ to predict the encoding of future images. The latent prediction is evaluated on both the latent space through the latent prediction loss and on the image space through the reconstruction loss. $\gamma$ is a scalar hyperparameter.

$$\mathcal{L} = \mathcal{L}^2_{rec} + \gamma * \mathcal{L}^2_{pred}$$

We obtain the mapping $\rho$ by first mapping $g$ to the linear algebra $\mathfrak{g} = \mathbb{R}^{k\times k}$ through a neural network $\phi$ which is then composed with the matrix exponential to obtain invertible matrices (See Appendix D.2.6).

We define the losses used throughout.

**The latent prediction loss** uses the group representation action to predict the evolution of stimuli encodings.

$$\mathcal{L}^N_{pred}(\rho, h) = \sum_t \sum_{j=1}^{N} ||h(o_{t+j}) - \prod_{i=0}^{j-1} \rho(g_{t+i}) h(o_t)||^2_2$$

**The reconstruction loss** estimates the evolution of the observations from the predicted evolution of encodings. The reconstruction loss also evaluates the reconstruction of the initial observation like a standard autoencoder.

$$\mathcal{L}^N_{rec}(\rho, h, d) = \sum_t \sum_{j=0}^{N} ||o_{t+j} - d\big(\prod_{i=0}^{j-1} \rho(g_{t+i}) h(o_t)\big)||^2_2$$

We theoretically show in 3.2 how the proposed architecture leads to learning a group-structured representation $(\rho, h)$.

### 3.2. The HAE Learns Group-Structured Representations

In this section, we provide theoretical insights on learning group-structured representations and how the two-step HAE architecture achieves that with minimal assumptions. The one step latent prediction loss is simply enforcing the commutative diagram in Figure 1. With the assumption that the group $G$ is a compact Lie Group, it admits a *faithful* group representation $\rho^*$ that we can assume is the one acting on the world states $W$. If we assume $\rho^*$ is given on $Z$, then minimizing $\mathcal{L}^1_{pred}(\rho^*, h)$ is enough to learn a group-structured representation (See Appendix B Proposition 10).

However, when $\rho^*$ is not known and a group representation $\rho$ of $G$ needs to be learned over a space of arbitrary mappings, minimizing $\mathcal{L}^1_{pred}(\rho, h)$ can lead to the trivial representation (See Appendix B Proposition 11).

The reconstruction loss of the initial observation helps avoid the representation collapse into a trivial solution by ensuring $h$ is not constant for a given fixed group representation $\rho^0$ (See Appendix B Proposition 12). Although we found using $\mathcal{L}^2_{rec}(\rho, h)$ works better than $\mathcal{L}^0_{rec}(\rho, h)$ when jointly learning $(\rho, h)$.

We now present the main theoretical result of the paper: The HAE through enforcing the 2-step latent prediction loss and the observations reconstruction (enforcing $h$ is injective) learns a group-structured representation.

**Proposition 1** *Assume $(\rho, h)$ minimizes $\mathcal{L}^2_{pred}(\rho, h)$ and $h$ is injective, then $\rho$ is a non-trivial group representation and $(\rho, h)$ is a group-structured representation.*

## 4. Results and Discussion

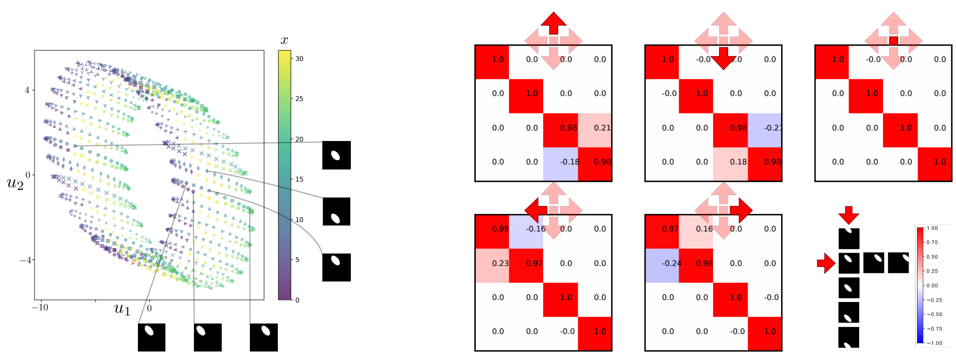

Figure 2: Left: Random 2D projection of the learned 4D latent encodings of the translated ellipse dataset. Markers and colors indicate the $x$ and $y$ positions. Right: Learned $\rho(g)$ for the UP, DOWN, LEFT and RIGHT translations. See Figure 4 for a visualization of a wider neighbourhood of the identity.

We consider a toy dataset (See Appendix D.1) of shapes translated cyclically on a black background. The first experiment (detailed in Appendix D.2) looks at the learned representation when the dataset is generated by the transitive action of translations on the ellipse shape. The resulting manifold can be seen Figure 2 to match the expected 2-Torus $S^1 \times S^1$ and the group representation learned corresponds to the independent rotation matrices Figure 2. An advantage of our proposed group representation is it gives access to the group's Lie Algebra where composition is a simple addition (See Appendix D.2.6). Enforcing the disentangled structure is detailed in Appendix C. We also show in Figure 10 how the consistency obtained in the learned group structured representations are useful for the rollout prediction of states after a long succession of actions (detailed in Appendix D.3). We compare the performance of our model to another group structured approach (Quessard et al., 2020) (labelled 'Rotations') and to an unstructured joint encoder of state and action. Finally, we show how in the case of multiple shapes being acted on by the same translation group, the HAE learns to separate the different shapes into orbits along a new representational unit, resulting in 3 parallel identical manifolds Figure 11.

## Acknowledgments

Hamza Keurti is grateful to CLS for generous funding support. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Hsiao-Ru Pan. Further this work was supported by the Swiss National Science Foundation (B.F.G. CRSII5-173721 and 315230_189251), ETH project funding (B.F.G. ETH-20 19-01), the Human Frontiers Science Program (RGY0072/2019).

## References

Francis Bach, Rodolphe Jenatton, Julien Mairal, and Guillaume Obozinski. Optimization with sparsity-inducing penalties. *Foundations and Trends in Machine Learning*, 4:–, 2011. ISSN 19358237. doi: 10.1561/2200000015. URL https://hal.archives-ouvertes.fr/hal-00613125https://hal.archives-ouvertes.fr/hal-00613125/document.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation Learning : A Review and New Perspectives. (1993):1–30, 2012.

Hugo Caselles-Dupré, Michael Garcia-Ortiz, and David Filliat. Symmetry-based disentangled representation learning requires interaction with environments. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Rishidev Chaudhuri, Berk Gerçek, Biraj Pandey, Adrien Peyrache, and Ila Fiete. The intrinsic attractor manifold and population dynamics of a canonical cognitive circuit across waking and sleep. *Nature Neuroscience*, 22, 2019.

Melvyn A. Goodale and A. David Milner. Separate visual pathways for perception and action, 1992. ISSN 01662236.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

Irina Higgins, David Amos, David Pfau, Sebastien Racaniere, Loic Matthey, Danilo Rezende, and Alexander Lerchner Deepmind. Towards a Definition of Disentangled Representations. dec 2018. doi: 10.48550/arxiv.1812.02230.

Tejas D Kulkarni, William F Whitney, Pushmeet Kohli, and Josh Tenenbaum. Deep convolutional inverse graphics network. *Advances in neural information processing systems*, 28, 2015.

Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. dsprites: Disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset/, 2017.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

Jean Piaget. Part I: Cognitive development in children: Piaget development and learning. *Journal of Research in Science Teaching*, 2(3):176–186, 1964. ISSN 10982736. doi: 10.1002/tea.3660020306.

Robin Quessard, Thomas D. Barrett, and William R. Clements. Learning disentangled representations and group structure of dynamical environments. In *Advances in Neural Information Processing Systems*, volume 2020-December, 2020.

Extended Abstract Track

Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, et al. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588(7839): 604–609, 2020.

William J. Thompson and LeRoy F. Cook. Angular momentum: An illustrated guide to rotational symmetries for physical systems. *American Journal of Physics*, 63, 1995. ISSN 0002-9505. doi: 10.1119/1.17835.

Extended Abstract Track

# Appendix A. Background

## A.1. Group Theory

In this section, we provide an overview of group theory concepts exploited in this work.

**Definition 2 (Group)** *A set $G$ is a group if it is equipped with a binary operation $\cdot : G \times G \to G$ and if the group axioms are satisfied*

1. *Associativity:* $\forall a, b, c \in G,\ (a \cdot b) \cdot c = a \cdot (b \cdot c)$
2. *Identity: There exists $e \in G$ such that $\forall a \in G$, $a \cdot e = e \cdot a = a$.*
3. *Inverse: $\forall a \in G$, there exists $b \in G$ such that $a \cdot b = b \cdot a = e$. This inverse is denoted $a^{-1}$.*

We are often interested in sets of transformations, which respect a group structure, but are applied to objects that are not necessarily group elements. This can be studied through group actions, which describe how groups *act* on other mathematical entities.

**Definition 3 (Group Action)** *Given a group $G$ and a set $X$, a group action is a function $\cdot_X : G \times X \to X$ such that the following conditions are satisfied.*

1. *Identity: If $e \in G$ is the identity element, then $e \cdot_X x = x$, $\forall x \in X$.*
2. *Compatibility: $\forall g, h \in G$ and $\forall x \in X$, $g \cdot_X (h \cdot_X x) = ((g \cdot h) \cdot_X x)$*

**Induced homomorphism.** The group action $\cdot_X : G \times X \to X$ induces a group homomorphism $\rho_{\cdot_X} : G \to Sym(X)$. (where $Sym(X)$ is the group of all invertible transformations of $X$) through:

$$\forall (g, x) \in G \times X, \quad g \cdot_X x = \rho_{\cdot_X}(g)(x)$$

The group homomorphism property of $\rho_{\cdot_X}$ comes from the group action axioms of $\cdot_X$:

$$\rho_{\cdot_X}(id)(x) = id \cdot_X x = x \quad \text{(identity)}$$

$$= id_X(x)$$

So $\rho_{\cdot_X}(id) = id_X$. and

$$\rho_{\cdot_X}(g_1 \cdot g_2)(x) = (g_1 \cdot g_2) \cdot_X x = g_1 \cdot_X (g_2 \cdot_X x) \quad \text{(compatibility)}$$

$$= \rho_{\cdot_X}(g_1) \circ \rho_{\cdot_X}(g_2)(x)$$

Equality over all of $X$ leads to equality of the functions: $\rho_{\cdot_X}(g_1 \cdot g_2) = \rho_{\cdot_X}(g_1) \circ \rho_{\cdot_X}(g_2)$.

In what follows, we are interested in *linear group actions* in which case the acted on space is a vector space $V$ and the induced homomorphism $\rho$ maps $G$ to the group $GL(V)$ of invertible linear transformations of $V$. This mapping is called a group representation. Actions of this type have been studied extensively in representation theory.

**Definition 4 (Group Representation)** *Let $G$ be a group and $V$ a vector space. A representation is a function $\rho : G \to GL(V)$ such that $\forall g, h \in G$, one has $\rho(g)\rho(h) = \rho(g \cdot h)$.*

Note that such definition is not restricted to finite dimensional vector spaces, however we will limit our study to this case, such that representations are appropriately described by mappings from $G$ to a space of square matrices.

**Definition 5 (Lie Group)** *A Lie Group $G$ is a nonempty set satisfying the following conditions:*

- *$G$ is a group.*
- *$G$ is a smooth manifold.*
- *The group operation $\cdot : G \times G \to G$ and the inverse map $.^{-1} : G \to G$ are smooth.*

We limit ourselves to the study of linear Lie Groups, Lie groups that are matrix groups. The tangent space to a Lie Group at the identity forms a Lie Algebra. A Lie Algebra $\mathfrak{g}$ is a vector space equipped with a bilinear map $[.,.] : \mathfrak{g} \times \mathfrak{g} \to \mathfrak{g}$ called the Lie Bracket. We will not introduce the Lie Bracket as we do not make use of it. The Lie Algebra somehow describes most of everything happening in its Lie Group. This connection is established through the exponential map.

**Definition 6 (Exponential Map)** *The exponential map $exp : \mathfrak{g} \to G$ is defined for matrix Lie Groups by the series:*

$$e^A = \sum_{k=0}^{\infty} \frac{1}{k} A^k. \quad \forall A \in \mathfrak{g}$$

The exponential map is not always surjective. However if we only consider groups that are connected and compact, the exponential is surjective, which justifies our parametrization of the group representation through:

$$\rho : G \xrightarrow{\phi} \mathfrak{g} = M_n(\mathbb{R}) \xrightarrow{exp} GL_n(\mathbb{R})$$

Where $\phi$ is a trainable arbitrary mapping.

**Group action types** The effect of a group action on a base space $X$ varies according to the properties of the homomorphism defined by the group action

$$\begin{aligned} \tau : & G \to Sym(X) \\ & g \mapsto g \cdot_X \square \end{aligned}$$

We introduce two types of actions:

**Definition 7 (Transitive Group Action)** *The action of $G$ on $X$ is* transitive *if $X$ forms a single* orbit.

*in other words, $\forall x, y \in X, \exists g \in G; g \cdot x = y$.*

Extended Abstract Track

**Definition 8 (Faithful Group Action)** *The action of $G$ on $X$ is* faithful *if the homomorphism $G \to Sym(X)$ corresponding to the action is bijective (an isomorphism).*

*In that case, $\forall g_1 \neq g_2 \in G, \exists x \in X; \quad g_1 \cdot x \neq g_2 \cdot x$ .*

We also define the *orbits* by a group action:

**Definition 9 (Orbit by a Group Action)** *The orbit of an element $x \in X$ by the action $\cdot_X$ of a group $G$ is the set*

$$G \cdot_X x = \{g \cdot_X x : g \in G\}$$

When the action of $G$ is transitive on $X$, then $X$ is the single orbit by the action of $G$:

$$\forall x \in X, G \cdot_X x = X$$

Such is the case for our experiments using a single shape. We also explore the case where the action is not transitive in the multi shape experiment visualized in Figure 11.

### A.2. Group-Structured Disentangled Representation Learning

The group-structured representation is defined in Section 2.1. We proceed to define group-structured *disentangled* representations. The group-structured representation is disentangled with regard to the group decomposition $G = G_1 \times ... \times G_n$ if it satisfies this additional condition:

3. $Z$ can be written as a product of spaces $Z = Z_1 \times ... \times Z_n$ or as a direct sum of subspaces $Z = Z_1 \oplus ... \oplus Z_n$ such that each subgroup $G_i$ acts non trivially on $Z_i$ and acts trivially on $Z_j$ for $j \neq i$.

If we require the group action $\cdot_Z$ on $Z$ to be linear, then the disentanglement of the group action on $Z$ is expressed as:

3. (Linear) There exists a decomposition

$$Z = Z_1 \oplus ... \oplus Z_n$$

and a decomposition of the group representation

$$\rho = \rho_1 \oplus ... \oplus \rho_n$$

where each $\rho_i : G_i \to GL(Z_i)$ is a subrepresentation.

The action on $Z$ can then be written

$$\begin{aligned} g \cdot_Z z &= \rho(g_1, ..., g_n)(z_1 \oplus ... \oplus z_n) \\ &= \rho_1(g_1)z_1 \oplus ... \oplus \rho_n(g_n)z_n \end{aligned} \tag{1}$$

for $g = (g_1, ..., g_n) \in G$ and $z = z_1 \oplus ... \oplus z_n \in Z$. Clearly each subgroup $G_i$ acts trivially on $Z_j$, $j \neq i$.

### A.3. Observing Symmetries

Caselles-Dupré et al. (2019) proved that learning such group-structured representations requires observing the group elements that transition one state of the world into another instead of isolated observations. Similarly to past works (Quessard et al., 2020; Caselles-Dupré et al., 2019), we suggest using the interventions of an agent on its environment as a means to probe the symmetries of the environment. However, we do not assume that these interventions only act on one latent at a time. Instead, the agent intervenes through motor signals which are parametrized by the joints positions and it is not known a priori when an intervention only modifies one true latent while keeping others constant. The parametrization of the observed group elements is also expressed in the joints space instead of along the true latents. For a given group element, we will denote $\tilde{g}$, its parametrization in the latents space, and $g$ its parametrization in the joints space. And we assume there exists a deterministic mapping between parametrizations $\varphi : \tilde{g} \mapsto g$. For instance, when a person moves a chalk along a blackboard, the chalk describes a 2D movement overparametrized by the rotation angles of the joints of the arm.

We propose to learn the inference process $h$ and a disentangled group representation $\rho$ using the HAE, described in section 3.1.

### A.4. Lie Groups and the Exponential Map

We assume the group $G$ is a connected compact Lie group, the observed interventions are a discrete subset sampled from $G$. When a group $G$ is also a differentiable manifold, it is called a Lie Group. The tangent space to the group $G$ at the identity forms a Lie Algebra $\mathfrak{g}$: A vector space equipped with a bilinear product, the Lie bracket. We will leverage a convenient property of Lie Groups and their Lie Algebras that is the group can be studied through its tangent space. Indeed, the matrix exponential, called the exponential map in this setting, transports elements from the Lie Algebra to the Lie Group. Under certain assumptions, for instance if $G$ is connected compact, which we assume, the exponential map is surjective and therefore the whole group $G$ can be described from the tangent space at its identity. The exponential of an arbitrary matrix $A$ is given by the series $e^A = \sum_{k=0}^{\infty} \frac{1}{k} A^k$.

We construct the learnable group representation $\rho$ as the composition of the exponential map with an arbitrary mapping to the Lie Algebra.

$$\rho : G \xrightarrow{\phi} \mathfrak{g} \xrightarrow{exp} GL(Z)$$

We show how access to the group algebra can be leveraged in appendix D.2.6.

## Appendix B. Theoretical results

### B.1. Propositions and Proofs

First we prove the main theoretical result of the paper, then proceed to provide other propositions and their proofs.

**Proof** [Proposition 1] Given that the group $G$ is supposed compact, it admits a group representation by the *Peter-Weyl theorem*. We can therefore assume that the true state space $W$ is acted on linearly by $G$ through its representation $\rho^*$. As such the inverse of the generating process $b^{-1}$ and $\rho^*$ verify $\mathcal{L}^2_{pred}(\rho^*, b^{-1}) = 0$.

Assume $h$ is injective, guaranteed by a minimization of the 0-step reconstruction loss.

Assume $(\rho, h)$ minimizes $\mathcal{L}^2_{pred}(\rho, h)$ therefore $\mathcal{L}^2_{pred}(\rho, h) = 0$.

Then for all observed 2-step transitions $(o_t, g_t, o_{t+1}, g_{t+1}, o_{t+2})$ — note that observed transitions $(o_t, g_t, o_{t+1})$ correspond to an action on the true world states $w_{t+1} = g_t \cdot_W w_t$ — $(\rho, h)$ verifies:

$$\rho(g_t)h(o_t) = h(o_{t+1})$$

and

$$\rho(g_{t+1})\rho(g_t)h(o_t) = h(o_{t+2})$$

Let us prove $\rho$ is a group representation, meaning it verifies $\rho(g_2 g_1) = \rho(g_2)\rho(g_1), \forall g_1, g_2 \in G$.

Let $g_1, g_2, g_3 \in G$ such that $g_3 = g_2 g_1$.

Let $w_1, w_2, w_3$ which verify $w_2 = g_1 \cdot_W w_1$, $w_3 = g_2 \cdot_W w_2$, $w_3 = g_3 \cdot_W w_1$. Therefore, the associated transitions $(o_1, g_1, o_2)$, $(o_2, g_2, o_3)$ and $(o_1, g_3, o_3)$ can be observed.

$$\begin{tikzcd} o_1 \arrow[r, "g_1"] \arrow[rr, bend right, "g_2 g_1"'] & o_2 \arrow[r, "g_2"] & o_3 \end{tikzcd}$$

Assume, we observe the 2-step transitions $(o_t, g_t, o_{t+1}, g_{t+1}, o_{t+2})$ and $(o_{t'}, g_{t'}, o_{t'+1}, g_{t'+1}, o_{t'+2})$ such that:

$$\begin{cases} o_t = o_{t'} = o_1 \\ o_{t+1} = o_2 \\ o_{t+2} = o_{t'+1} = o_3 \end{cases}$$

and

$$\begin{cases} g_t = g_1 \\ g_{t+1} = g_2 \\ g_{t'+1} = g_3 = g_2 g_1 \end{cases}$$

$\mathcal{L}^2_{pred} = 0$ gives for the transition at $t$: $\rho(g_{t+1})\rho(g_t)h(o_t) = h(o_{t+2})$ therefore $\rho(g_2)\rho(g_1)h(o_1) = h(o_3)$.

And for the transition at $t'$: $\rho(g_{t'})h(o_{t'}) = h(o_{t'+1})$ therefore $\rho(g_3)h(o_1) = h(o_3)$

Therefore we have $\rho(g_2)\rho(g_1)h(o_1) = \rho(g_3)h(o_1)$ or $\rho(g_2)\rho(g_1)h(o_1) = \rho(g_2 g_1)h(o_1)$

With this equality verified over the set $O_{tr}$ of first observations $o_1$ in the training set — meaning we observe two successions of the action of group elements $(g_1, g_2)$ and $g_3$ for different values of the starting observation $o_1$— by assuming $h(O) \subseteq span(h(O_{tr}))$, we get that $\rho(g_2)\rho(g_1) = \rho(g_2 g_1)$ over $h(O)$ (equality of linear mappings over vectors that span a vector subspace).

Therefore the subrepresentation of $\rho$ over $span(h(O))$ is a group representation of $G$.

The injectivity assumption ensure $h(O)$ does not collapse to a single element.

Let us show $h$ is a group-structured representation.

We have for every observed transition $(o_t, g_t, o_{t+1})$:

Extended Abstract Track

$$h(o_{t+1}) = \rho(g_t)h(o_t)$$

by the generative model assumptions

$$o_{t+1} = b(w_{t+1}) = b(g_t \cdot_W w_t)$$

Also $o_t = b(w_t)$, finally

$$h \circ b(g_t \cdot_W w_t) = \rho(g_t) h \circ b(w_t)$$

■

**Proposition 10** *Assume we observed the action of the group $G$ on each point of the observation space. Assume $h$ minimizes $\mathcal{L}^1_{pred}(\rho^*, h)$ then $h$ is a group-structured representation, meaning $h \circ b$ is equivariant.*

**Proof** [Proposition 10] Given that the group $G$ is supposed compact, it admits a group representation by the *Peter-Weyl theorem*. We can therefore assume that the true state space $W$ is acted on linearly by $G$ through its representation $\rho^*$. As such the inverse of the generating process $b^{-1}$ and $\rho^*$ verify $\mathcal{L}^2_{pred}(\rho^*, b^{-1}) = 0$.

As a consequence, $h$ also achieves zero loss such that

$$\sum_t \sum_{j=1}^{N} ||h(o_{t+j}) - \prod_{i=0}^{j-1} \rho^*(g_{t+i})h(o_t)||_2^2$$

such that for all $(j, t)$

$$h(o_{t+j}) = \prod_{i=0}^{j-1} \rho^*(g_{t+i})h(o_t)$$

In particular for $j = 1$

$$h(o_{t+1}) = \rho^*(g_t)h(o_t)$$

by the generative model assumptions

$$o_{t+1} = b(w_{t+1}) = b(\rho^*(g_t)w_t)$$

Also $o_t = b(w_t)$ such that

$$h \circ b(\rho^*(g_t)w_t) = \rho^*(g_t) h \circ b(w_t)$$

■

**Proposition 11** *The trivial group representation $\rho = I$ (that always maps to the identity matrix) combined with a constant $h$ is a zero of the prediction loss $\mathcal{L}^1_{pred}(\rho, h)$.*

**Proposition 12** *Assume the data samples at least once all points of the observation space. If $h$ minimizes $\mathcal{L}^N_{rec}(\rho^0, h)$ then $h$ is injective.*

**Proof** [Proposition 12] Assume the encoder $h$ and the decoder $d$ minimize the 0-step reconstruction.

$$\forall o, o' \in O, \text{such that } o \neq o'.$$
$$d(h(o)) = o \text{ and } d(h(o')) = o'$$

Therefore

$$d(h(o)) \neq d(h(o')).$$

Finally

$$h(o) \neq h(o')$$

Therefore $h$ is injective.

■

## Appendix C. Disentanglement

### C.1. Disentanglement through Structural Constraints

As expressed in Appendix A.2 the disentanglement condition for a linear action on $Z$ defined through its group representation $\rho$, is a decomposition of both the representation space $Z = \bigoplus_1^n Z_i$ and the group representation $\rho = \bigoplus_{i=1}^n \rho_i$. Where the subgroup representations $\rho_i$ are representations of the subgroups $G_i$ on the subspaces $Z_i$.

Following that the group G is decomposed in the true latent's parametrization, the observed group representation is disentangled with regard to the group decomposition if in matrix form, the group representation of any group element $g = \varphi(\tilde{g})$ is a block-diagonal matrix of the subgroups representations:

$$\rho\big(g = \varphi(\tilde{g}^1, ..., \tilde{g}^n)\big) = \begin{pmatrix} \rho_1(\tilde{g}^1) & 0 & \dots & 0 \\ 0 & \rho_2(\tilde{g}^2) & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \rho_n(\tilde{g}^n) \end{pmatrix} \tag{2}$$

We can therefore constrain our trainable group representation in the space of matrices of the block diagonal form given in equation 2. This requires prior knowledge of:

- The number of groups in the decomposition.
- The dimension of each subgroup representation $dim(Z_i)$.

We start by assuming knowledge of this information, however we hope that in future works, one would be able to search for the best decomposition. However, we do not assume prior knowledge of the decomposition. Meaning that given an observed transition $g$, we do not have access to the decomposition $\varphi^{-1}(g) = \tilde{g} = (\tilde{g}^1, ..., \tilde{g}^n)$ along the group decomposition $G = G_1 \times ... \times G_n$. Which means we are learning representations of the form given in equation 3.

$$\rho(g) = \begin{pmatrix} \rho_1(g) & 0 & \dots & 0 \\ 0 & \rho_2(g) & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \rho_n(g) \end{pmatrix} \tag{3}$$

Extended Abstract Track

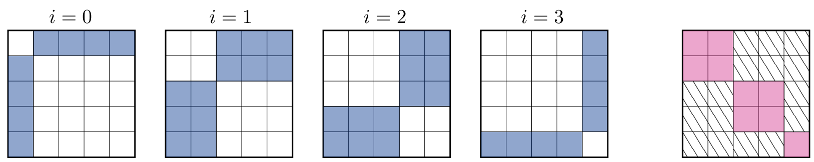


Figure 3: Left: Non allowed non-zero regions induced by each value of the index $i$ on a representation of dimension 5. Right: an example pattern resulting from the activations of the terms corresponding to $i \in \{1, 3\}$.

While we do not prove that this block diagonal constraint leads to disentanglement, we show through experiments that HAE learns a group-structured representation $(\rho, h)$ that is disentangled with regard to the true factors and takes the form in equation 2.

### C.2. Disentanglement through Sparsity Loss

To drop the assumption of prior knowledge of the number of subgroups in the representation as well as the dimension for their representation, we propose a sparsity loss (Equation 4) on the group representation that prefers block-diagonal patterns (Example in Figure 3) by minimizing the terms of $(\rho_{ij}(g_t))_{ij}$ that do not belong to allowed non-zero terms. See Appendix C.2 for a brief discussion. The loss is inspired from works on sparsity inducing losses (Bach et al., 2011).

$$\mathcal{L}_{sparse}(\rho) = \sum_t \sum_{i \geq 0} \sqrt{\sum_{j \geq i+1} \rho_{ij}(g_t)^2 + \rho_{ji}(g_t)^2} \tag{4}$$

The model is then trained on the composite loss:

$$\mathcal{L} = \mathcal{L}_{rec} + \gamma * \mathcal{L}_{pred} + \delta * \mathcal{L}_{sparse}$$

We can then choose a representation space $Z = \mathbb{R}^d$ with $d$ large enough to accommodate the total dimension of the group representation, extra dimensions are then trivialized by $\mathcal{L}_{sparse}$.

Figure 3 shows how the term associated with each value of $i$ induces a direct sum of 2 subrepresentations $\rho = \rho_1 \oplus \rho_2$ with a different split of the total representation dimension. Combining patterns produces any possible decomposition into a direct sum of subrepresentations. See Appendix D.2.7 for associated results.

## Appendix D. Experiments

### D.1. Data

We use a subset of the dSprites dataset consisting only of the ellipse at a fixed scale and orientation with varying $x$ and $y$ positions, there are 32 equally spaced positions for each. We consider that the ellipse is acted on by the group $G = G_x \times G_y$ of cyclic translations, where the sprite warps to the opposite extremity when it reaches an extremal position. For each image $o_1$ we sample group elements $\tilde{g}_1 = (\tilde{g}_x, \tilde{g}_y)$ uniformly from a square around identity spanning the range $[\![-10, 10]\!]$. We assume the agent observes $g_1 = \varphi_{45}(\tilde{g}_1) = \frac{\sqrt{2}}{2}(\tilde{g}_x - \tilde{g}_y, \tilde{g}_x + \tilde{g}_y)$. We obtain the first transition $(o_1, g_1, o_2)$, we do the same for $o_2$ to get 2-step transitions $(o_1, g_1, o_2, g_2, o_3)$.

## D.2. Learning a Group-Structured Disentangled Representation

### D.2.1. Setup

We consider a subset of the dsprites dataset (Matthey et al., 2017) where a fixed scale and orientation ellipse is acted on by the group of 2D cyclic translations $G = C_x \times C_y$. The corresponding transition dataset contains tuples $(o_1, g_1, o_2, ..., g_{n-1}, o_n)$, where the observations $o_i$ are $64 \times 64$ pixels and the transitions are given by $g_i = \varphi_{\pi/4}(\tilde{g}_i)$ where $\tilde{g}_i = (\tilde{g}_i^x, \tilde{g}_i^y)$ parametrizes the displacement along $x$ and $y$, and $\varphi_{\pi/4}$ is the rotation of the 2D plane by 45 deg.

### D.2.2. Hyperparameters

**Model architecture** We use a symmetrical architecture for the encoder and decoder, which we summarize in Table 1. The network was trained on the combined loss:

$$\mathcal{L} = \mathcal{L}^2_{rec}(\rho, h) + \gamma \mathcal{L}^2_{pred}(\rho, h)$$

Where we use the Binary Cross Entropy loss for the reconstruction term instead of the Mean Squared Error as it is better behaved during training.

Table 1: Network architecture.

| Parameter | Value |
|---|---|
| Conv. Channels | [64, 64, 64, 64] |
| Kernel Sizes | [6, 4, 4, 4] |
| Strides | [2, 2, 1, 1] |
| Linear Layer Size | 1024 |
| Activation | ReLU |
| Latent space | 4 |
| $\gamma$ | 400 |
| Group representation dimensions | [2,2] |

**Training hyperparameters** We trained the network using the hyperparameters summarized in Table 2.

### D.2.3. Visualization

We obtained the manifold in Figure 2 by projecting the 4D representation vector for each image in the dataset on a random 2D plane through Random Matrix Projection. We chose the projection with the most explainable visualization.

Extended Abstract Track

Table 2: Training hyperparameters.

| Parameter | Value |
|---|---|
| Optimizer | Adam |
| Learning rate | 0.001 |
| Number of training sequences | 10000 |
| Batch size | 500 |
| Epochs | 101 |

#### D.2.4. Results

**Learned data representation** We visualize the learned 4-dimensional encodings of the whole dataset through 2-dimensional random matrix projections, and selected a projection matrix showing a discernable projected manifold (Figure 2 Left). The learned manifold corresponds to the expected latent space topology $S^1 \times S^1$.

**Learned Group Representation** $\rho$ We then evaluate the learned matrices for the identity $id = (0,0)$, the generating elements of each subgroup $\tilde{1}_x = (1,0), \tilde{1}_y = (0,1)$ and their inverses $\tilde{g}_x^{-1} = (-1,0), \tilde{g}_y^{-1} = (0,-1)$, Figure 2. Note that the corresponding observed actions $g$ presented to the group representation $\rho$ are in the same order: $(0,0), \frac{\sqrt{2}}{2}(1,1), \frac{\sqrt{2}}{2}(-1,1), \frac{\sqrt{2}}{2}(-1,-1), \frac{\sqrt{2}}{2}(1,-1)$.

The matrices obtained, for example, actions in Figure 2, show that $\rho(0) = I_4$ and that representations for elements belonging to subgroups of the decomposition $G = C_x \times C_y$ follow the disentangled group representation predicted in subsection C.1 where actions on the same subgroup have representations that act on the same subspace while fixing the other subspace. In addition, the blocks correspond to 2D rotation matrices which are of the form:

$$R(\theta) = \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix}$$

The angle of rotation of an elementary step corresponds to the number of equally spaced true latent values for each subgroup: $2\pi/32$, with $\cos(2\pi/32) \approx 0.981$ and $\sin(2\pi/32) \approx 0.195$.

#### D.2.5. Representation of a Neighbourhood of Identity

We show in Figure 4 the group representation $\rho(g)$ for $g$ in a grid around identity.

#### D.2.6. Lie Algebra and Latent Traversal

We show how learning a mapping to the group algebra can be leveraged to navigate the group and the data manifold. We remind that $\rho = \exp \circ \phi$, where $\phi$ maps to the algebra $\mathfrak{g}$ of the group $G$, and exp is the matrix exponential which gives a connection between the algebra and the lie group.

The mapping $\phi = \phi_1 \oplus \phi_2$ and the group representation $\rho = \rho_1 \oplus \rho_2$ are constrained on the space of block diagonal matrices of the form $M = M_1 \oplus M_2$ each of dimension $2 \times 2$. However since each block is made of elements from a $1D$ subgroup of $GL_2(\mathbb{R})$: $SO(2)$, its algebra is the $1D$ subalgebra of $M_2(\mathbb{R})$ of skew-symmetric matrices. We find this subspace by performing a PCA over each of the sets $\{\phi_1(g_t)\}_t$ and $\{\phi_2(g_t)\}_t$ for a batch $\{g_t\}_t$ of

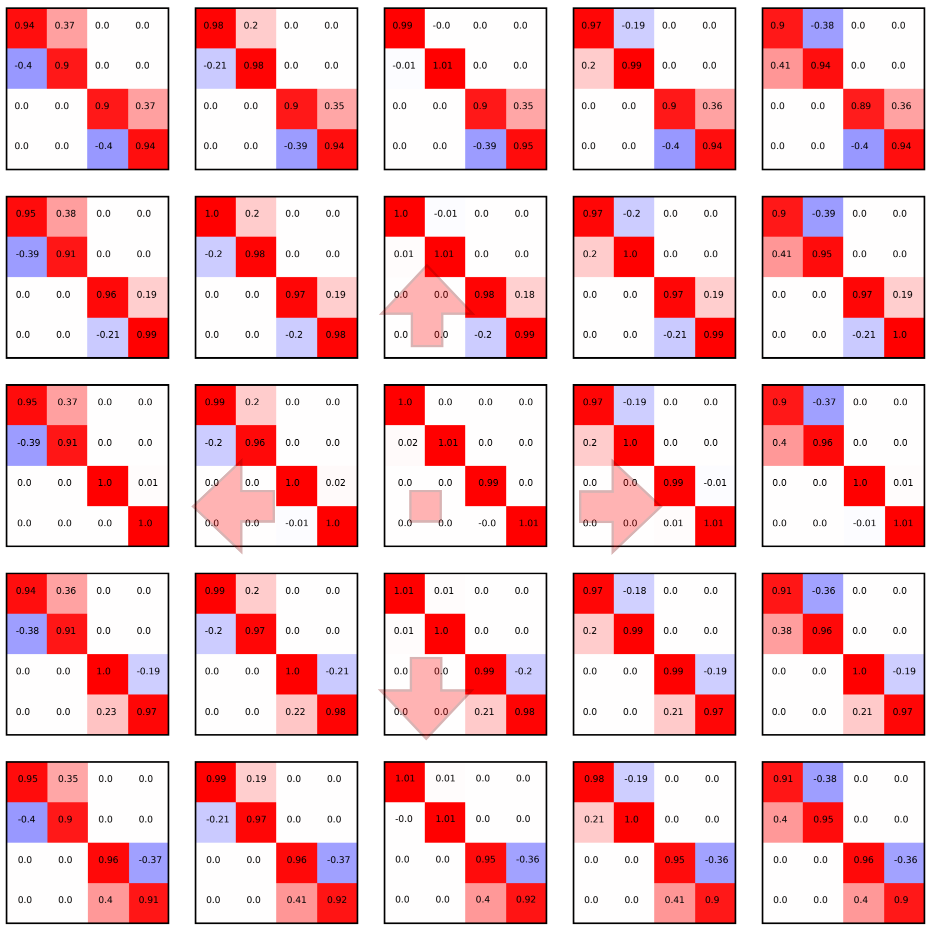

Figure 4: The representation of actions in a square $[\![-2, 2]\!] \times [\![-2, 2]\!]$ around the identity show a homomorphism structure.

Extended Abstract Track

observed transitions. The first component for each block $E_1$ and $E_2$ corresponds to the only base vector for that subalgebra. We find:

$$E_1 \oplus 0_{2,2} = \begin{bmatrix} -0.04 & -0.65 & 0 & 0 \\ 0.76 & -0.04 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

$$0_{2,2} \oplus E_2 \approx \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0.01 & -0.65 \\ 0 & 0 & 0.76 & 0.01 \end{bmatrix}$$

We obtain the figure 5 by linearly traversing the subalgebras through $t(E_1 \oplus 0_{2,2})$ and $t(0_{2,2} \oplus E_2)$ for equally spaced values of $t \in [\![0, 9]\!]$ and passing it to the matrix exponential which yields invertible matrices of the form $R_{1,t} = e^{tE_1} \oplus I_2$ and $R_{2,t} = I_2 \oplus e^{tE_2}$. We encode an arbitrary initial observation to obtain its representation vector $z$, and traverse the latent space through $R_{i,t}z$. We decode the obtained vectors to obtain the predicted images.

The group algebra offers a smooth parametrization of the group and consequently of the data manifold and enables the prediction of observations evolution in the absence of performed actions. Indeed, in the above example, all transformations can be obtained in the form $\exp(t_1 E_1) \oplus \exp(t_2 E_2)$ for $t_1, t_2 \in \mathbb{R}$.

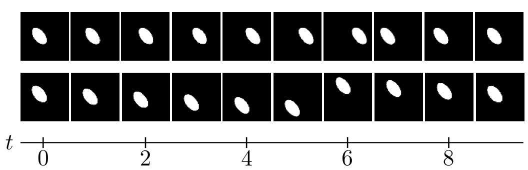


Figure 5: We visualize the linear traversal of the group algebra and its effect on the predicted image reconstruction. The first line corresponds to a traversal $tE_1 \oplus 0_{2,2}$, while the second line corresponds to the traversal $0_{2,2} \oplus tE_2$.

### D.2.7. Sparsity Based Disentanglement

We show in Figure 6 the learned group representation for the same dataset used in the experiment described in Appendix D.2.1 where instead of enforcing a block diagonal structure, it emerges under the influence of the sparsity inducing loss described in Appendix C.2. We do not assume knowledge of the representation dimension and we set it to a value high enough (8 in the example) and we find the HAE learns the disentangled representation found in the main experiment and trivializes the extra dimensions.

### D.2.8. Additional Experiment

We consider a subset of the dataset consisting of all variations of the heart under a fixed scale. As such the heart is acted on by the group $G = G_\theta \times G_x \times G_y = C_{39} \times C_{32} \times C_{32}$.

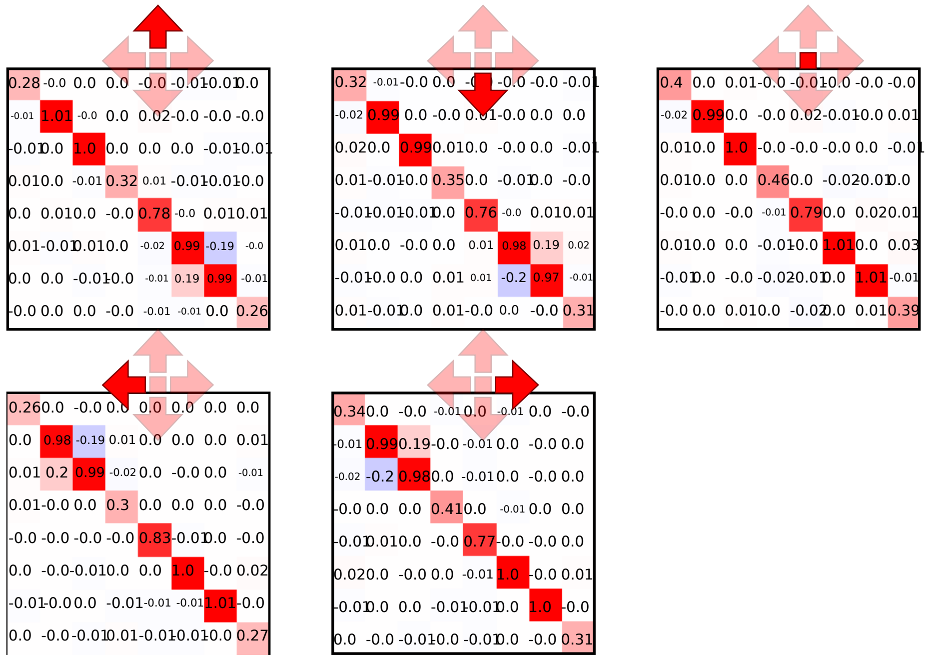

Figure 6: Representation learned through sparsity loss with no prior knowledge on the number of subgroups or the dimension of their representation. Displayed representation matrices for example actions.

We train a similar model to the one described in the section D.2.2 by changing the latent space to 6 dimensions, the group representation is fixed of the form $\rho = \rho_1 \oplus \rho_2 \oplus \rho_3$ each of dimension 2. We expect a disentangled representation space $Z = Z_1 \oplus Z_2 \oplus Z_3$. For the visualization of the learned representation manifold Figure 7, we visualize each subspace $Z_i$ separately by only varying one generative factor and keeping all else fixed. We also visualized the learned representations for a subset of transitions corresponding to the elementary generative transitions for each subgroup in Figure 8.

### D.3. Rollouts prediction

One important application of learning structured representation is to predict how the observations would change given sequences of actions. We compare HAE to two other approaches of modeling the dynamics in the latent space: (1) *Unstructured*: $z_{t+1} = h(z_t, g_t)$, where $h$ is a learnable function. Similar approaches have been widely adapted in recent model-based deep RL methods (Ha and Schmidhuber, 2018; Schrittwieser et al., 2020). (2) *Rotations*: $z_{t+1} = R_g z_t$, where $R_g = \prod_{i,j} G(i, j, \theta_{ij,g})$ and $G(i, j, \theta_{ij,g})$ are the Givens rotation matrices. This approach was proposed by Quessard et al. (2020) and was shown to be capable of learning group-structured representations when the actions are sampled along the true latents.

We evaluate the methods in an offline setting, where we train each method on a given set of 2-step trajectories and test their generalization ability on a hold-out set of 128-

Extended Abstract Track

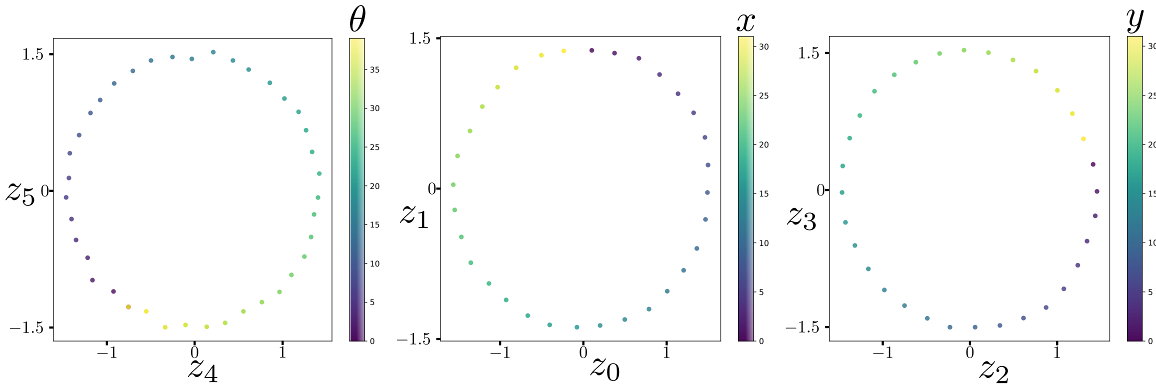

Figure 7: Visualizaion of the 6D embedding vectors for the heart dataset. For each visualized 2D subspace, we only vary the latent represented by the subspace.

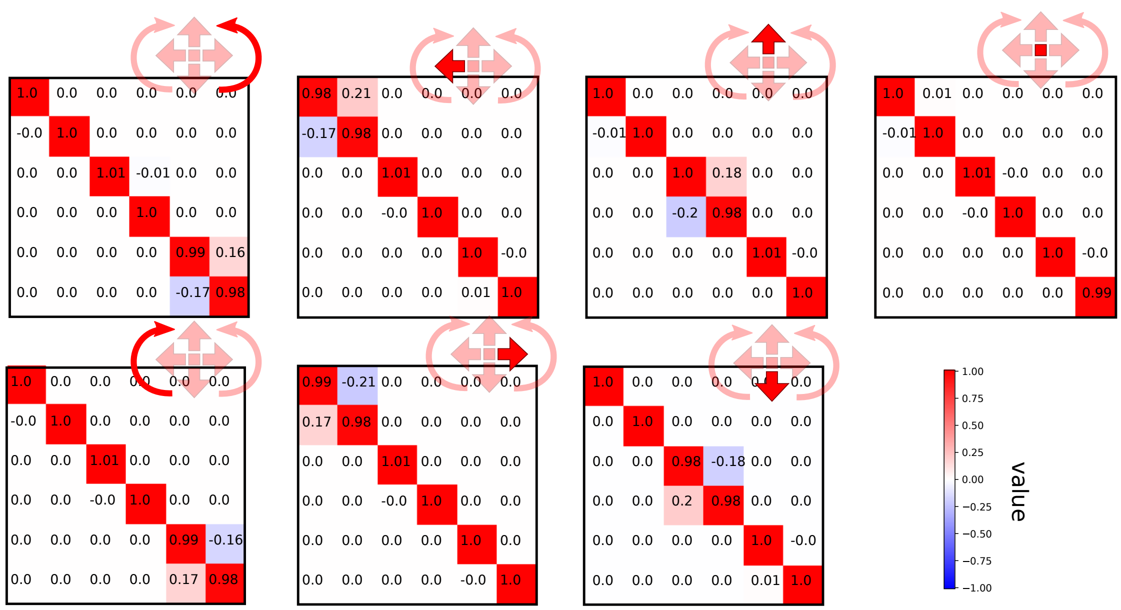

Figure 8: Evaluation of the learned group representation for the identity (upper left) and generative transitions of each subgroup yields disentangled block rotation matrices.

step trajectories. Figures 10 and 9 shows the reconstruction loss for each method on the test trajectories. Figure 10 considers rollouts of actions sampled in a discrete grid around the identity while Figure 9 considers sparse actions sampled along a single property as was performed in Quessard et al. (2020). Our result suggests that when the actions sampled are not sparse (i.e., each action may involve changes in multiple generating factors),

Extended Abstract Track

the Rotations method may perform worse than the Unstructured method while HAE can outperform them significantly.

In our experiment, we allow the agent to control the object in the dSprite dataset with 7 actions. Namely, translation in the x-y axes, rotation in both directions (clockwise, counter-clockwise), and idle. Each action corresponds to an increment/decrement in one of the generating factors of the dataset, except for the identity, which does nothing. Additionally, we use the heart shape from the dataset to fully utilize the orientation latent factor as other shapes present rotational symmetry.

We train each method with a set of pre-generated set of 2-step trajectories and evaluate on a hold-out pre-generated set of 128-step trajectories. For each trajectory, we begin by sampling a random initial state (x,y position and orientation) from all possible states. After that, we sample actions uniformly from the 7 possible actions at each step until the number of steps is satisfied.

For the Rotations method (Quessard et al., 2020) and HAE, each action is represented using a matrix $\rho(g)$, and the transition in the latent space is simply $z_{t+1} = \rho(g_t)z_t$. For the Unstructured method, we use a 2 layer MLP of size [128, 128] to model the transition by $z_{t+1} = f_\theta(z_t, g_t)$, where we concatenate the latent vector $z_t$ and the one-hot encoding of the action $a_t$.

The reconstruction loss is the same for all three methods, as described in Section 3.2. For HAE, we additionally add the latent prediction loss $\mathcal{L}_{pred}$ as described in Section 3.2. We increase $\gamma$ to 1600 which we found to be more stable when matrices are directly parameterized instead of mapped from MLPs. For the Rotations method, an additional entanglement loss $\mathcal{L}_{ent}$ is required to encourage each matrix to act on the fewest dimensions of the latent space, which is equal to

$$\mathcal{L}_{ent} = \sum_g \sum_{(i,j)\neq(\alpha,\beta)} |\theta^g_{i,j}|^2 \quad \text{with} \quad \theta^g_{\alpha,\beta} = \max_{i,j} |\theta^g_{i,j}|.$$

For the Unstructured method, we only use the reconstruction loss and no additional terms.

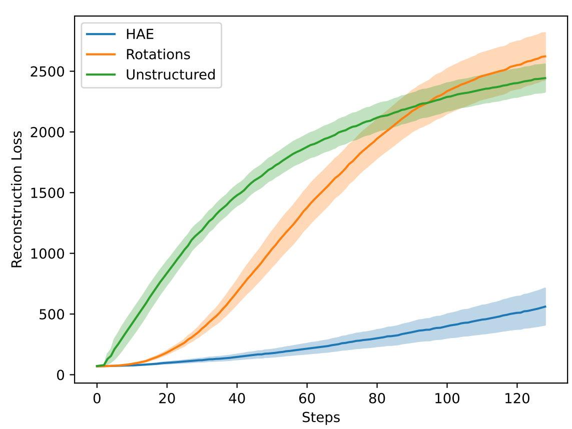


Figure 9: Step-wise reconstruction loss. Lines and shadings represent the mean and one standard error over 15 seeds.

Extended Abstract Track

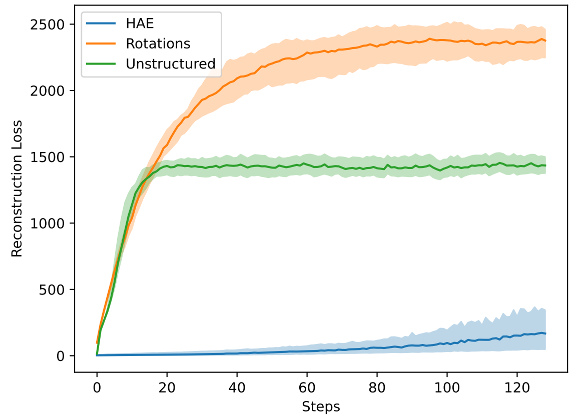


Figure 10: Step-wise reconstruction loss on the test dataset. Lines and shadings represent median and interquartile range over 50 random seeds.

### D.4. Multi-objects

We consider the subset of the dSprites dataset consisting of three shapes (heart, square and ellipse) acted on by the previous group of 2D cyclic translations $G = C_x \times C_y$. The action of the group is not *transitive*, as it describes a separate *orbit* for each shape (see appendix A). Indeed, no intervention transforms a shape into another: the model is trained on sequences of 3 images containing the same shape. In this experiment, we augment the representation space by one dimension on which the group representation acts trivially. We therefore learn 5-dimensional encodings of the observations $2 \times 2D$ subspaces acted on by the subrepresentations of the group $G$ and 1 representation unit trivially acted on.

Our results in Figure 11 show that the model not only learns the representation of the cyclic translation group action shared among shapes, but also learns to separate the representation of observations by shape along the last $G$-invariant representation unit, giving rise to three identical manifolds. This is reminiscent of the two-streams hypothesis of visual processing (Goodale and Milner, 1992), the "What" pathway processes information related to object identity, while the "Where" pathway processes information related to the object pose, or the required motor action for manipulation.

Figure 11 is obtained by projecting the 5D encodings of the dataset on a 2D space through Random Matrix Projection. Note that although the three different tori appear of different sizes, it is only an artifact of the projection while the tori are identical.

### D.5. Unsupervised Identity Separation from Intervention

A more thorough analysis of the experiment can be found in Appendix D.4.

## Appendix E. Link to Neuroscience

It is interesting to note that similar cyclic embeddings have been reported in neuroscience, for example, in the hippocampus of mice where toroid-like embeddings encode the animals' head-direction (Chaudhuri et al., 2019). In fact, our cyclic shape translation task restricted

Extended Abstract Track

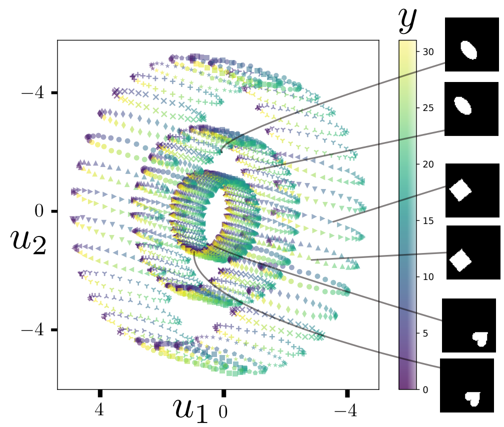


Figure 11: $5D$ representation space of a trained homomorphism autoencoder with two $2D$ group subrepresentations and a $1D$ representation space acted on trivially. The representation vectors for the whole dataset are projected to $2D$ via Random Projection. Colors indicate $y$ position, markers indicate $x$ position and each torus corresponds to a separate shape.

to a single dimension could be viewed as an agent horizontally rotating its head while observing its environment.

## Appendix F. Limitations

One limitation of our theoretical analysis is the deterministic nature of our model, the intrinsically non-linear nature of the problem would make the theoretical analysis more challenging in the stochastic setting. On the experimental side, we solely assessed the extraction and internal representation of geometric action structure, we leave it to future work to test if the same principled HAE approach generalizes to learn other, non cyclic, facets of the structure of the external world.

Extended Abstract Track

### F.1. Computational resources

The experiments were performed on an NVIDIA GeForce RTX 3090 and A100 GPUs. The training of our optimal models run for approximately 20 mins.

## Appendix G. Third-Party Software

### G.1. Deep Learning Framework

To implement our architecture we used the deep learning framework PyTorch. Paszke et al. (2019)

### G.2. Hyperparameter Search

We used the hyperparameter search utility provided in the hypnettorch project `https://github.com/chrhenning/hypnettorch/tree/master/hypnettorch/hpsearch` to perform a random grid search.

### G.3. Dataset

In the presented experiments, we used the dSprites dataset Matthey et al. (2017). The dSprites dataset is an image dataset of white sprites on a black background, varying in shape (heart, ellipse, square), in scale (6 values), in orientation (39 values, cyclic), in $x$ and $y$ position (32 values each). We consider all factors besides shape to be cyclic, in particular for the $x$ and $y$ positions, we "glued" opposite borders of images into a torus. The resolution of the images is $64 \times 64$ pixels.

## Appendix H. Societal Impact

This work proposes new findings in basic research. To the best of our knowledge, this work does not have immediate applications with a negative societal impact.