Learning Geometric Representations of Objects via Interaction

Alfredo Reichlin (✉)*[1], Giovanni Luca Marchetti (✉)*[1], Hang Yin[2], Anastasiia Varava[1], and Danica Kragic[1]

[1] KTH Royal Institute of Technology, Stockholm, Sweden
{alfrei,glma}@kth.se
[2] University of Copenhagen, Copenhagen, Denmark

Abstract. We address the problem of learning representations from observations of a scene involving an agent and an external object the agent interacts with. To this end, we propose a representation learning framework extracting the location in physical space of both the agent and the object from unstructured observations of arbitrary nature. Our framework relies on the actions performed by the agent as the only source of supervision, while assuming that the object is displaced by the agent via unknown dynamics. We provide a theoretical foundation and formally prove that an ideal learner is guaranteed to infer an isometric representation, disentangling the agent from the object and correctly extracting their locations. We evaluate empirically our framework on a variety of scenarios, showing that it outperforms vision-based approaches such as a state-of-the-art keypoint extractor. We moreover demonstrate how the extracted representations enable the agent to solve downstream tasks via reinforcement learning in an efficient manner.

Keywords: Representation Learning · Equivariance · Interaction

1 Introduction

A fundamental aspect of intelligent behavior by part of an agent is building rich and structured *representations* of the surrounding world [10]. Through structure, in fact, a representation potentially leads to semantic understanding, efficient reasoning and generalization [17]. However, in a realistic scenario an agent perceives observations of the world that are high-dimensional and unstructured e.g., images. Therefore, the ultimate goal of inferring a representation consists of extracting structure from the observed data [3]. This is challenging and in some instances requires supervision or biases. For example, it is known that *disentangling* factors of variation in data is mathematically impossible in a completely unsupervised way [18]. In order to extract structure, it is therefore necessary to design methods and paradigms relying on additional information and specific assumptions.

*Equal Contribution

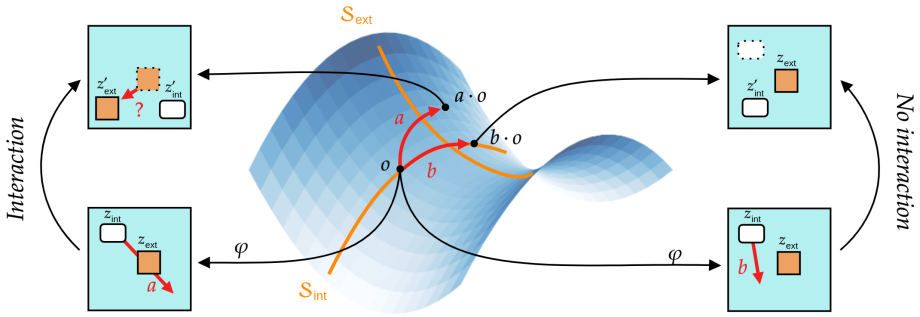

Fig. 1: Our framework enables to learn a representation φ recovering the geometric and disentangled state of both an agent (z_{int}, white) and an interactable object (z_{ext}, brown) from unstructured observations o (e.g., images). The only form of supervision comes from actions a, b performed by the agent, while the transition of the object (question mark) in case of interaction is unknown. In case of no interaction, the object stays invariant.

In the context of an agent interacting with the world, a fruitful source of information is provided by the *actions* performed and collected together with the observations. Based on this, several recent works have explored the role of actions in representation learning and proposed methods to extract structure from interaction [15, 22, 25]. The common principle underlying this line of research is encouraging the representation to replicate the effect of actions in a structured space – a property referred to as *equivariance* [3]. In particular, it has been shown in [20] that equivariance enables to extract the location of the agent in physical space, resulting in a lossless and geometric representation. The question of how to represent features of the world which are extrinsic to the agent (e.g., objects) has been left open. Such features are dynamic since they change as a consequence of interaction. They are thus challenging to capture in the representation but are essential for understanding and reasoning by part of the agent.

In this work we consider the problem of learning representations of a scene involving an agent and an external rigid object the agent interacts with (see Figure 1). We aim for a representation disentangling the agent from the object and extracting the locations of both of them in physical space. In order words, we aim for representations that are isometric w.r.t. to the geometry of the world. To this end, we focus on a scenario where the object displaces only when it comes in contact with the agent, which is realistic and practical. We make no additional assumption on the complexity of the interaction: the object is allowed to displace arbitrarily and its dynamics is unknown. Our assumption around the interaction enables to separate the problem of representing the agent – whose actions are known and available as a supervisory signal – from the problem of representing

[3] Alternative terminologies from the literature are *World Model* [15] and *Markov Decision Process Homomorphism* [26].

the object – whose displacement is unknown. Following this principle, we design an optimization objective relying on actions as the only form of supervision. This makes the framework general and in principle applicable to observations of arbitrary nature. We moreover provide a formalization of the problem and theoretical grounding for the method. Our core theoretical result guarantees that the representation inferred by an ideal learner recovers isometric representations as desired. We complement the theoretical analysis with an empirical investigation. Results show that our proposed representations outperform in quality of structure a state-of-the-art keypoint extractor and can be leveraged by the agent in order to solve control tasks efficiently by reinforcement learning. In summary, our contributions include:

- A representation learning framework extracting representations from observations of a scene involving an agent interacting with an object.
- A theoretical result guaranteeing that the above learning framework, when implemented by an ideal learner, infers an isometric representation for data of arbitrary nature.
- An empirical investigation of the framework on a variety of environments with comparisons to computer vision approaches (i.e., keypoint extraction) and applications to a control task.

We provide Python code implementing our framework together with all the experiments at the following public repository: `https://github.com/reichlin/GeomRepObj`.

2 Related Work

Equivariant Representation Learning. Several recent works have explored the idea of incorporating interactions into representation learning. The common principle is to infer a representation which is equivariant i.e., such that transitions in observations are replicated as transitions in the latent space. One option is to learn the latent transition end-to-end together with the representation [15,26,33]. This approach is however non-interpretable and the resulting representations are not guaranteed to extract any structure. Alternatively, the latent transition can be designed a priori. Linear and affine latent transitions have been considered in [9], [22] and [25] while transitions defined by (the multiplication of) a Lie group have been discussed in [20], [21]. As shown in [20], for static scenarios (i.e., with no interactive external objects) the resulting representations are structured and completely recover the geometry of the underlying state of the agent. Our framework adheres to this line of research by modelling the latent transitions via the additive Lie group $\mathbb{R}^n$. We however further extend the representation to include external objects. Our framework thus applies to more general scenarios and dynamics while still benefiting from the geometrical guarantees.

Keypoint Extraction. When observations are images, computer vision offers a spectrum of classical approaches to extract geometric structure. In particular, extracting keypoints enables to identify any object appearing in the

observed images. Popular keypoint extractors include classical non-parametric methods [19], [2] as well as modern self-supervised learning approaches [16], [8]. However, keypoints from an image provide a representation based on the geometry of the field of view or, equivalently, of the pixel plane. This means that the intrinsic three-dimensional geometry of states of objects is not preserved since the representation differs from it by an unknown projective transformation. In specific situations such transformation can still be recovered by processing the extracted keypoints. This is the case when images are in first person view w.r.t. the observer: the keypoints can then be converted into three-dimensional landmarks via methods such as bundle adjustment [31], [29]. Differently from computer vision approaches, our framework is data-agnostic and does not rely on specific priors tied to the nature of observations. It instead extracts representations based on the actions performed by the agent, which is possible due to the dynamical assumptions described in Section 3.

Interactive Perception. The role of interaction in perception has been extensively studied in cognitive sciences and neuroscience [7, 12, 23]. Inspired by those, the field of interactive perception from robotics aims to enhance the understanding of the world by part of an artificial system via interactions [5]. Applications include active control of cameras [1] and manipulators [32] in order to improve the perception of objects [4, 13, 28]. Our work fits into the program of interactive perception since we crucially rely on performed actions as a self-supervisory signal to learn the representation. We show that the location of objects can be extracted from actions alone, albeit in a particular dynamical setting. Without interaction, this would require strong assumptions and knowledge around the data and the environment as discussed in Section 2.

3 Formalism and Assumptions

In this section we introduce the relevant mathematical formalism together with the assumptions necessary for our framework. We consider the following scenario: an agent navigates in a Euclidean space and interacts in an unknown way with an external object. This means that the space of states $\mathcal{S}$ is decomposed as

$$\mathcal{S} = \mathcal{S}_{\text{int}} \times \mathcal{S}_{\text{ext}} \tag{1}$$

where $\mathcal{S}_{\text{int}}$ is the space of states of the agent (*internal* states) and $\mathcal{S}_{\text{ext}}$ is the space of states of the object (*external* states). We identify both the agent and the object with their location in the ambient space, meaning that $\mathcal{S}_{\text{int}} \subseteq \mathbb{R}^n \supseteq \mathcal{S}_{\text{ext}}$, where n is the ambient dimension. The actions that the agent performs are displacements of its state i.e., the space of actions consists of translations $\mathcal{A} = \mathbb{R}^n$. In our formalism we thus abstract objects as material points for simplicity of the theoretical analysis. The practical extension to volumetric objects together with their orientation is discussed in Section 4.3 while the extension of agent's actions to arbitrary Lie groups is briefly discussed in Section 6.

Our first assumption is that the agent can reach any position from any other via a sequence of actions. This translates in the following connectivity condition:

Assumption 1. (Connectedness) *The space* $\mathcal{S}_{\text{int}}$ *is connected and open.*

When the agent performs an action $a \in \mathcal{A}$ the state $s = (s_{\text{int}}, s_{\text{ext}})$ transitions into a novel one denoted by $a \cdot s = (s'_{\text{int}}, s'_{\text{ext}})$. Since the actions displace the agent, the internal state gets translated as $s'_{\text{int}} = s_{\text{int}} + a$.[4] However, the law governing the transition of the object $s'_{\text{ext}} = T(s, a)$ is assumed to be unknown and can be arbitrarily complex and stochastic. We stick to deterministic transitions for simplicity of explanation. Crucially, the agent does not have access to the ground-truth state s. Instead it perceives unstructured and potentially high-dimensional observations $o \in \mathcal{O}$ (e.g., images) via an unknown emission map $\omega : \mathcal{S} \to \mathcal{O}$. We assume that ω is injective so that actions induce deterministic transitions of observations, which we denote as $o' = a \cdot o$. This assumption is equivalent to total observability of the scenario and again simplifies the forthcoming discussions by avoiding the need to model stochasticity in $\mathcal{O}$.

The fundamental assumption of this work is that the dynamics of the external object revolves around *contact* i.e., the object does not displace unless it is touched by the agent. This is natural and often satisfied in practice. In order to formalize it, note that when the agent in state s_{int} performs an action $a \in \mathcal{A}$ we can imagine it moving along the open segment $\lfloor s_{\text{int}},\ s_{\text{int}} + a \rfloor = \{s_{\text{int}} + ta\}_{0<t<1}$. Our assumption then translates into (see Figure 1 for a graphical depiction):

Assumption 2. (Interaction Occurs at Contact) *For all agent states* $s_{\text{int}} \in S$ *and actions* $a \in \mathcal{A}$ *it holds that* $s'_{\text{ext}} = s_{\text{ext}}$ *if and only if* $s_{\text{ext}} \notin \lfloor s_{\text{int}},\ s_{\text{int}} + a \rfloor$.

As such, the dynamics of the external object can be summarized as follows:

$$s'_{\text{ext}} = \begin{cases} s_{\text{ext}} & \text{if } \ s_{\text{ext}} \notin \lfloor s_{\text{int}},\ s_{\text{int}} + a \rfloor, \\ T(s, a) & \text{otherwise.} \end{cases} \tag{2}$$

Finally, we need to assume that interaction is possible for every state of the object i.e., the latter has to be always reachable by the agent. This is formalized via the following inclusion:

Assumption 3. (Reachability) *It holds that* $\mathcal{S}_{\text{ext}} \subseteq \mathcal{S}_{\text{int}}$.

4 Method

4.1 Representations and Equivariance

We now outline the inference problem addressed in the present work. Given the setting introduced in Section 3, the overall goal is to infer a *representation* of observations $\varphi : \mathcal{O} \to \mathcal{Z} = \mathcal{Z}_{\text{int}} \times \mathcal{Z}_{\text{ext}}$, where $\mathcal{Z}_{\text{int}} = \mathcal{Z}_{\text{ext}} = \mathbb{R}^n$. Ideally φ recovers the underlying inaccessible state in $\mathcal{S} \subseteq \mathcal{Z}$ and disentangles $\mathcal{S}_{\text{int}}$ from $\mathcal{S}_{\text{ext}}$. In order to achieve this, our central idea is to split the problem of representing the

[4] Whenever we write $a \cdot s$ we implicitly assume that the action is valid i.e., that $s_{\text{int}} + a \in \mathcal{S}_{\text{int}}$.

agent and the object. Since the actions of the agent are available, $z_{\text{int}} \in \mathcal{Z}_{\text{int}}$ can be inferred geometrically by existing representation learning methods. The representation of the object $z_{\text{ext}} \in \mathcal{Z}_{\text{ext}}$ can then be inferred based on the one of the agent by exploiting the relation between the dynamics of the two (Equation 2). In order to represent the agent, we consider the fundamental concept of (translational) *equivariance*:

Definition 1. *The representation φ is said to be* equivariant *(on internal states) if for all $a \in \mathcal{A}$ and $o \in \mathcal{O}$ it holds that $z'_{\text{int}} = z_{\text{int}} + a$ where $(z_{\text{int}}, z_{\text{ext}}) = \varphi(o)$ and $(z'_{\text{int}}, z'_{\text{ext}}) = \varphi(a \cdot o)$.*

We remark that Definition 1 refers to internal states only, making our terminology around equivariance unconventional. As observed in previous work [20], equivariance guarantees a faithful representation of internal states. Indeed if φ is equivariant then z_{int} differs from s_{int} by a constant vector. This means that the representation of internal states is a translation of ground-truth ones and as such is lossless (i.e., bijective) and isometrically recovers the geometry of $\mathcal{S}_{\text{int}}$.

The above principle can be leveraged in order to learn a representation of external states with the same benefits as the representation of internal ones. Since the external object displaces only when it comes in contact with the agent (Assumption 2), the intuition is that z_{ext} can be inferred by aligning it with z_{int}. The following theoretical result formalizes the possibility of learning such representations and traces the foundation of our learning framework.

Theorem 4. *Suppose that the representation $\varphi : \mathcal{O} \to \mathcal{Z}$ satisfies:*

1. *φ is equivariant (Definition 1),*
2. *φ is injective,*
3. *for all $o \in \mathcal{O}$ and $a \in \mathcal{A}$ it holds that either $z'_{\text{ext}} = z_{\text{ext}}$ or $z_{\text{ext}} \in \lfloor z_{\text{int}}, z_{\text{int}} + a \rfloor$ where $(z_{\text{int}}, z_{\text{ext}}) = \varphi(o)$ and $(z'_{\text{int}}, z'_{\text{ext}}) = \varphi(a \cdot o)$.*

Then $\varphi \circ \omega$ is a translation i.e., there is a constant vector $h \in \mathbb{R}^n$ such that for all $s \in \mathcal{S}$ it holds that $\varphi(\omega(s)) = s + h$. In particular, $\varphi \circ \omega$ is an isometry w.r.t. the Euclidean metric on both $\mathcal{S}$ and $\mathcal{Z}$.

We refer to the Appendix for a proof. Theorem 4 states that if the conditions $1. - 3.$ are satisfied (together with the assumptions stated in Section 3) then the representation recovers the inaccessible state up to a translation and thus isometrically preserves the geometry of the environment. All the conditions from Theorem 4 refer to properties of φ depending on observations and the effect of actions on them, which are accessible in practice. The goal of the forthcoming section is to describe how these conditions can be enforced on φ by optimizing a system of losses.

4.2 Learning the Representation

In this section we describe a viable implementation of a representation learning framework adhering to the conditions of Theorem 4. We model the representation

learner $\varphi = (\varphi_{\text{int}}, \varphi_{\text{ext}})$ as two parameterized functions $\varphi_{\text{int}} : \mathcal{O} \to \mathcal{Z}_{\text{int}}$, $\varphi_{\text{ext}} : \mathcal{O} \to \mathcal{Z}_{\text{ext}}$ e.g., two deep neural network models. In order to train the models, we assume that the dataset $\mathcal{D}$ consists of transitions observed by the agent in the form of $\mathcal{D} = \{(o, a, o' = a \cdot o)\} \subseteq \mathcal{O} \times \mathcal{A} \times \mathcal{O}$. Such data can be collected by the agent autonomously exploring its environment and randomly interacting with the external object. This implies that the only form of supervision required consists of the actions performed by the agent together with their effect on the observations.

First, we propose to enforce equivariance, condition 1 from Theorem 4, by minimizing the loss:

$$\mathcal{L}_{\text{int}}(o, a, o') = d(z'_{\text{int}}, z_{\text{int}} + a) \tag{3}$$

where d is a measure of similarity on $\mathcal{Z}_{\text{int}} = \mathbb{R}^n$ and the notation is in accordance with Definition 1. Typically d is chosen as the squared Euclidean distance as described in previous work [15, 22].

Next, we focus on the representation of the external object. As stated before, the dataset consists of transitions either with or without interaction. When an interaction occurs, z_{ext} should belong to the segment $\lfloor z_{\text{int}}, z_{\text{int}} + a \rfloor$. When it doesn't, the representation should be invariant i.e., $z_{\text{ext}} = z'_{\text{ext}}$. These two cases are outlined in condition 2 of Theorem 4 and can be enforced via the following losses:

$$\mathcal{L}_-(o, a, o') = d(z_{\text{ext}}, z'_{\text{ext}}) \qquad \mathcal{L}_+(o, a, o') = d(z_{\text{ext}}, \lfloor z_{\text{int}}, z_{\text{int}} + a \rfloor). \tag{4}$$

The distance involved in $\mathcal{L}_+$ represents a point-to-set metric and is typically set as $d(z, E) = \inf_{x \in E} d(z, x)$. The latter has a simple explicit expression in the case E is a segment.

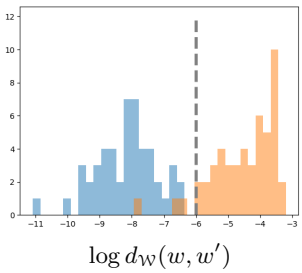

Fig. 2: Histograms of the log-distances in $\mathcal{W}$. Colors indicate whether interaction occurs (orange) or not (blue). The dotted line represents the threshold from Otsu's algorithm.

However, the data contains no information on whether interaction occurs or not. It is, therefore, necessary to design a procedure determining when to optimize $\mathcal{L}_+$ and $\mathcal{L}_-$. To this end, we propose to train a parallel model φ_{cont} :

$\mathcal{O} \to \mathcal{W}$ with latent *contrastive representation* $\mathcal{W}$ (potentially different from $\mathcal{Z}$). This is trained to attract $w = \varphi_{\text{cont}}(o)$ to $w' = \varphi_{\text{cont}}(o')$ while forcing injectivity of φ (condition 2 from Theorem 4). To this end, we stick to the popular *InfoNCE* loss from contrastive learning literature [6]:

$$\mathcal{L}_{\text{cont}}(o, o') = d_{\mathcal{W}}(w, w') + \log \mathbb{E}_{o''} \left[e^{-d_{\mathcal{W}}(w', w'') - d(z'_{\text{int}}, z''_{\text{int}})} \right] \tag{5}$$

where o'' is marginalized from $\mathcal{D}$. The second summand of Equation 5 encourages the joint encodings (z_{int}, w) to spread apart and thus encourages φ to be injective. Since subsequent observations where interaction does not occur share the same external state, these will lie closer in $\mathcal{W}$ than the ones where interaction does not occur. This enables to exploit distances in $\mathcal{W}$ in order to choose whether to optimize $\mathcal{L}_-$ or $\mathcal{L}_+$. We propose to partition (the given batch of) the dataset in two disjoint classes $\mathcal{D} = C_- \sqcup C_+$ by applying a natural thresholding algorithm to the quantities $d_{\mathcal{W}}(w, w')$. This can be achieved via one-dimensional 2-means clustering, which is equivalent to Otsu's algorithm [24] (see Figure 2 for an illustration). We then optimize:

$$\mathcal{L}_{\text{ext}}(o, a, o') = \begin{cases} \mathcal{L}_-(o, a, o') & \text{if } (o, a, o') \in C_-, \\ \mathcal{L}_+(o, a, o') & \text{if } (o, a, o') \in C_+. \end{cases} \tag{6}$$

In summary, the total loss minimized by the models $(\varphi_{\text{int}}, \varphi_{\text{ext}}, \varphi_{\text{cont}})$ w.r.t. the respective parameters is (see the pseudocode included in the Appendix):

$$\mathcal{L} = \mathbb{E}_{(o,a,o') \sim \mathcal{D}} [\mathcal{L}_{\text{int}}(o, a, o') + \mathcal{L}_{\text{ext}}(o, a, o') + \mathcal{L}_{\text{cont}}(o, o')]. \tag{7}$$

4.3 Incorporating Volumes of Objects

So far we have abstracted the external object as a point in Euclidean space. However, the object typically manifests with a body and thus occupies a volume. Interaction and consequent displacement (Assumption 3) occur when the agent comes in contact with the boundary of the object's body. The representation thus needs to take volumetric features into account in order to faithfully extract the geometry of states.

In order to incorporate volumetric objects into our framework we propose to rely on *stochastic* outputs i.e., to design z_{ext} as a probability density over $\mathcal{Z}_{\text{ext}}$ representing (a fuzzy approximation of) the body of the object. More concretely, the output of φ_{ext} consists of (parameters of) a Gaussian distribution whose covariance matrix represents the inertia ellipsoid of the object i.e., the ellipsoidal approximation of its shape. By diagonalizing the covariance matrix via an orthonormal frame, the orientation of the object can be extracted in the form of a rotation matrix in $\text{SO}(n)$. The losses of our model are naturally adapted to the stochastic setting as follows. The distance d appearing in Equation 4 is replaced with Kullback-Leibler divergence. The latter has an explicit simple expression for Gaussian densities which allows to compute $\mathcal{L}_-$ directly. In order to compute $\mathcal{L}_+$ we rely on a Monte Carlo approximation, meaning that we sample a point uniformly from the interval and set $\mathcal{L}^+$ as the negative log-likelihood of the point w.r.t. the density defining z_{ext}.

5 Experiments

We empirically investigate the performance of our framework in correctly identifying the position of an agent and of an interactive object. The overall goal of the experimental evaluation is to show that our representation is capable of extracting the geometry of states without relying on any prior knowledge of observations e.g., depth information. All the scenarios are normalized so that states lie in the unit cube. Observations are RGB images of resolution 100×100 in all the cases considered. We implement each of φ_{int}, φ_{ext} and φ_{cont} as a ResNet-18 [11] and train them for 100 epochs via the Adam optimizer with learning rate 0.001 and batch-size 128. We compare our framework with two baselines:

- *Transporter Network* [16]: a vision-based state-of-the-art unsupervised keypoint extractor. The approach heavily relies on image manipulation in order to infer regions of the pixel plane that are persistent between pairs of images. We train the model in order to extract two (normalized) keypoints representing z_{int} and z_{ext} respectively.
- *Variational AutoEncoder* (VAE) [14, 27]: a popular representation learner with a standard Gaussian prior on its latent space. We impose the prior on $\mathcal{Z}_{\text{ext}}$ only, while φ_{int} is still trained via the equivariance loss (Equation 3). The decoder takes the joint latent space $\mathcal{Z}$ in input. We set $\dim(\mathcal{Z}_{\text{ext}}) = 32$. This makes the representations disentangled, so that z_{int} and z_{ext} are well-defined. The resulting representation of the object is generic and is not designed to extract any specific structure from observations.

In order to evaluate the preservation of geometry we rely on the following evaluation metric $\mathcal{L}_{\text{test}}$. Given a trained representation $\varphi : \mathcal{O} \to \mathcal{Z}$ and a test set $\mathcal{D}_{\text{test}}$ of observations with known ground-truth states, we define:

$$\mathcal{L}_{\text{test}} = \mathbb{E}_{o \sim \mathcal{D}_{\text{test}}} [\ d(z_{\text{int}} - z_{\text{ext}}, s_{\text{int}} - s_{\text{ext}})\ ] \tag{8}$$

where d is the squared Euclidean distance. Since both our framework and (the encoder of) VAE have stochastic outputs (see Section 4.3), we set z_{ext} as the mean of the corresponding Gaussian distribution. Equation 8 measures the quality of preservation of the relative position between the agent and the object by part of the representation. When $\mathcal{L}_{\text{test}} = 0$, φ is an isometry (w.r.t. the Euclidean metric) and thus recovers the geometry of states. The translational invariance of $\mathcal{L}_{\text{test}}$ makes the comparison agnostic to any reference frame eventually inferred by the given learner.

5.1 Sprites

For the first experiment we procedurally generate images of two sprites (the agent and the object) moving on a black background (see Figure 3, top-left). Between images, the agent (red figure) moves according to a known action. If the agent comes in contact with the object (green diamond) during the execution of the action (see Assumption 2) the object is randomly displaced on the next

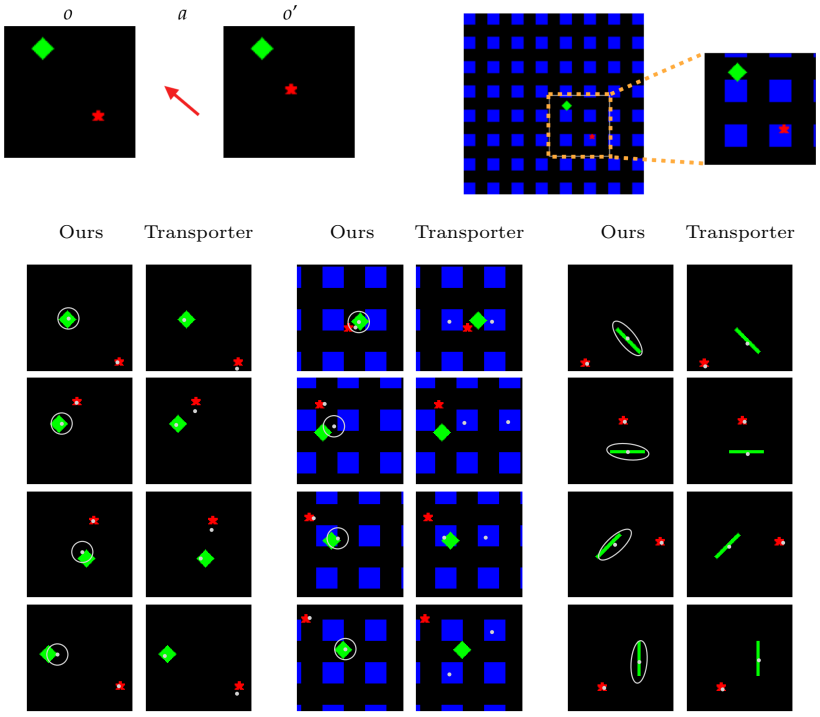

Fig. 3: **Top:** Visualization of the dataset from the Sprites experiment. On the left, an example of a datapoint $(o, a, o') \in \mathcal{D}$. On the right, an example of an observation from the second version of the dataset where a dynamic background is added as a visual distractor. **Bottom:** Comparison of z_{int}, z_{ext} (gray dots, with the ellipse representing the learned std) extracted via our model and the Transporter network on the three versions of the Sprites dataset: vanilla version (left), with dynamic background (middle) and with anisotropic object (right).

image. In other words, the object's transition function $T(s, a)$ is stochastic with a uniform distribution. Such a completely stochastic dynamics highlights the independence of the displacement of the agent w.r.t. the one of the object. We generate the following two additional versions of the dataset:

- A version with *dynamic background*. Images are now overlaid on top of a nine-times larger second image (blue squares in Figure 3, top-right). The field of view and thus the background moves together with the agent. The background behaves as a visual distractor and makes it challenging to extract structure (e.g., keypoints) via computer vision.
- A version with *anisotropic object*. The latter is now a rectangle with one significantly longer side. Besides translating, the object rotates as well when

interaction occurs. The goal here is showcasing the ability of our model in inferring the orientation of the object as described in Section 4.3.

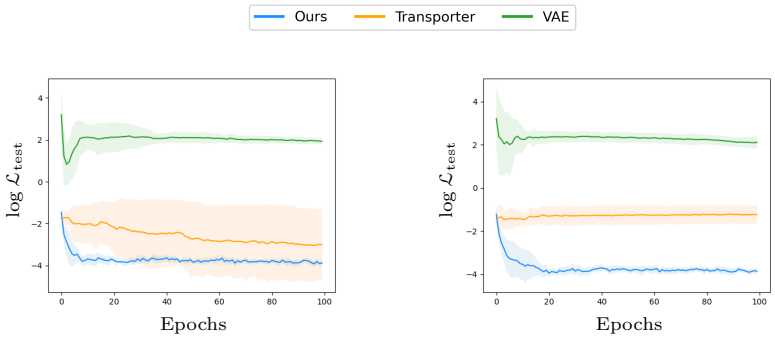

Fig. 4: Log-scale plots of the evaluation metric (Equation 8) as the training progresses for the Sprite experiment. The curves display mean and std (for 10 experimental runs). **Left**: vanilla version of the dataset. **Right:** version with a dynamic background.

Figure 4 displays the analytic comparison of the performances between our model and the baselines in terms of the evaluation metric (Equation 8). The plot is in log-scale for visualization purposes. Moreover, Figure 3 (bottom) reports a qualitative comparison between our model and the Transporter network. As can be seen, for the simpler version of the experiment (plot on the left) both our model and the Transporter network successfully achieve low error and recover the geometry of both the agent and the object. Note that the Transporter network converges slowly and with high variance (Figure 4, left). This is probably due to the presence of a decoder in its architecture. Our framework instead involves losses designed directly in the latent space, avoiding an additional model to decode observations. As expected, VAE achieves significantly worse performances because of the lack of structure in its representation. As can be seen from Figure 3 (bottom-right), when the object is anisotropic our model correctly infers its orientation by encoding it into the covariance of the learned Gaussian distribution. The Transporter network instead places a keypoint on the barycenter of the object and is therefore unable to recover the orientation.

For the more challenging version of the experiment with dynamic background, the transporter is not able to extract the expected keypoints. As can be seen from Figure 3 (bottom-middle), the distracting background causes the model to focus on regions of the image not corresponding to the agent and the object. This is reflected by a significantly higher error (and variance) w.r.t. our framework (Figure 4, right). The latter still infers the correct representation and preserves geometry. This empirically confirms that our model is robust to visual distractors since it does not rely on any data-specific feature or structure.

5.2 Soccer

For the second experiment we test our framework on an environment consisting of an agent on a soccer field colliding with a ball (see Figure 5, left). The scene is generated and rendered via the Unity engine. The physics of the ball is simulated realistically: in case of contact, rolling takes gravity and friction into account. Note that even though the scene is generated via three-dimensional rendering, the (inaccessible) state space is still two-dimensional since the agent navigates on the field. We generate two datasets of 10000 triples $(o, a, o' = a \cdot o)$ with observations of different nature. The first one consists of views in third-person perspective from a fixed external camera. In the second one, observations are four views in first-person perspective from four cameras attached on top of the agent and pointing in the 4 cardinal directions. We refer to Figure 5 (left) for a visualization of the two types of observations. In Figure 5 (right), we report visualizations of the learned representations. The extracted representation of our proposed method depends solely on the geometry of the problem at hand rather than the nature of the observation. The learned representation is thus identical when learned from the third-person dataset or the first-person one, as shown in 5 (right).

Figure 6 (left) displays the comparison of the performances between our model and the baselines in terms of the evaluation metric (Equation 8). The Transporter network is trained on observations in third person and as can be seen, correctly extracts the keypoints on the *pixel plane*. As discussed in Section 2, such a plane differs from $\mathcal{S}_{\text{int}}$ by an unknown projective (and thus non-isometric) transformation. This means that despite the successful keypoint extraction, the geometry of the state space is not preserved, which is reflected by the high error on the plot. This is a general limitation of vision-based approaches: they are unable to recover the intrinsic geometry due to perspective in the case of a three-dimensional scene. Differently from that, our framework extracts an isometric representation and achieves low error independently from the type of observations.

5.3 Control Task

In our last experiment we showcase the benefits of our representations in solving downstream control tasks. The motivation is that a geometric and low-dimensional representation improves efficiency and generalization compared to solving the task directly from observations. To this end we design a control task for the Soccer environment consisting in kicking the ball *into the goal*. The reward is given by the negative distance between the (barycenter of the) ball and the (barycenter of the) goal. Observations are views in third person perspective. In each episode the agent and the ball are initially placed in a random location while the ball is placed in the center. The maximum episode length is 20 steps.

We train a number of models via the popular reinforcement learning method *Proximal Policy Optimization* (PPO; [30]). One model (*End-to-End*) receives raw observations as inputs. The others operate on pre-trained representations

$\mathcal{Z}$ given by the Transporter network, the VAE and our method respectively. All the models implement a comparable architecture for a fair comparison.

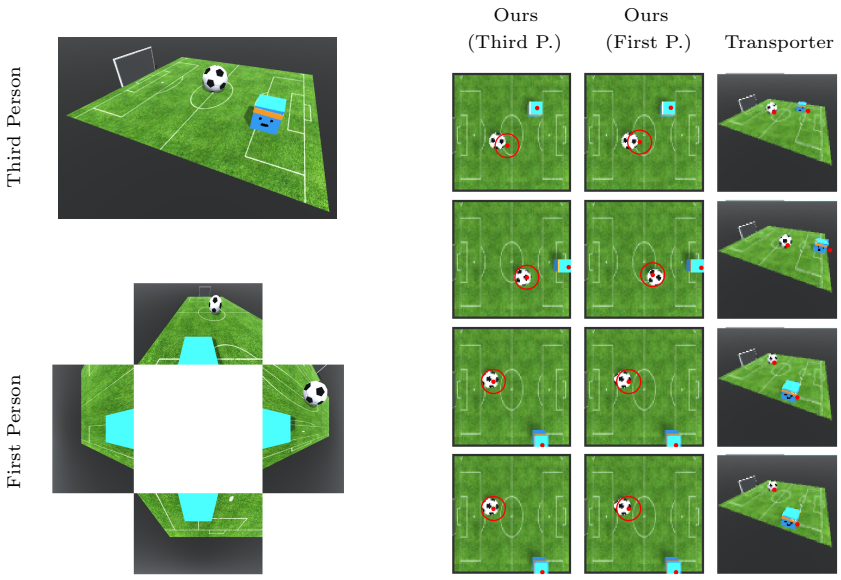

Fig. 5: **Left**: an example of the two types of observations (third and first person respectively) from the Soccer experiment. **Right**: visual comparison of z_{int}, z_{ext} (red dots) extracted via our model (from third-person view and first-person view) and the Transporter network. For our model, we overlap the representation to a view of the scene from the top instead of the original observation.

Figure 6 (right) displays the reward gained on test episodic runs as the training by reinforcement learning progresses. As can be seen, our geometric representation enables to solve the task more efficiently than both the competing representations (Transporter and VAE) and the end-to-end model. Note that the Transporter not only does not preserve the geometry of the state space, but has the additional disadvantage that the keypoint corresponding to the agent and the object can get swapped in the output of φ. This causes indeterminacy in the representation and has a negative impact on solving the task. Due to this, the Transporter performs similarly to the end-to-end model and is outperformed by the generic and non-geometric representation given by the VAE. In conclusion, the results show that a downstream learner can significantly benefit from geometric representations of observations in order to solve downstream control tasks.

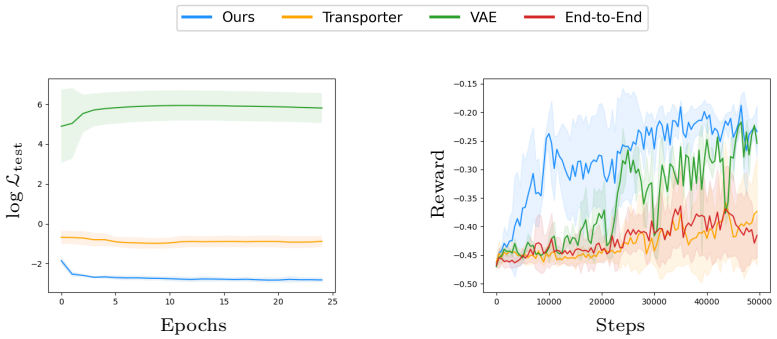

Fig. 6: **Left**: log-scale plot of the evaluation metric as the training progresses for the Soccer experiment. Observations are in third person. **Right**: plot of the reward gained via reinforcement learning on top of different representations.

6 Conclusions and Future Work

In this work we proposed a novel framework for learning representations of both an agent and an object the agent interacts with. We designed a system of losses based on a theoretical principle that guarantees isometric representations independently from the nature of observations and relying on supervision from performed actions alone. We empirically investigated our framework on multiple scenarios showcasing advantages over computer vision approaches.

Throughout the work we assumed that the agent interacts with a single object. An interesting line of future investigation is extending the framework to take multiple objects into account. In the stochastic context (see Section 4.3) an option is to model z_{ext} via multi-modal densities, with each mode corresponding to an object. As an additional line for future investigation, our framework can be extended to actions beyond translations in Euclidean space. Lie groups other than $\mathbb{R}^n$ often arise in practice. For example, if the agent is able to rotate its body then (a factor of) the space of actions has to contain the group of rotations $\mathrm{SO}(n)$, $n = 2, 3$. Thus, a framework where actions (and consequently states) are represented in general Lie groups defines a useful and interesting extension.

Acknowledgements

This work was supported by the Swedish Research Council, the Knut and Alice Wallenberg Foundation, the European Research Council (ERC-BIRD-884807) and the European Horizon 2020 CANOPIES project. Hang Yin would like to acknolwedge the support by the Pioneer Centre for AI, DNRF grant number P1.

Ethical Statement

We believe that the present work does not raise specific ethical concerns. Generally speaking, however, any system endowing artificial agents with intelligent behavior may be misused e.g., for military applications. Since we propose a representation learning method enabling an agent to locate objects in an environment, this can be potentially embedded into intelligent harmful systems and deployed for unethical applications.

References

1. Bajcsy, R.: Active perception. Proceedings of the IEEE **76**(8), 966–1005 (1988)
2. Bay, H., Tuytelaars, T., Gool, L.V.: Surf: Speeded up robust features. In: European conference on computer vision. pp. 404–417. Springer (2006)
3. Bengio, Y., Courville, A., Vincent, P.: Representation learning: A review and new perspectives. IEEE transactions on pattern analysis and machine intelligence **35**(8), 1798–1828 (2013)
4. Björkman, M., Bekiroglu, Y., Högman, V., Kragic, D.: Enhancing visual perception of shape through tactile glances. In: 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 3180–3186 (2013)
5. Bohg, J., Hausman, K., Sankaran, B., Brock, O., Kragic, D., Schaal, S., Sukhatme, G.S.: Interactive perception: Leveraging action in perception and perception in action. IEEE Transactions on Robotics **33**(6), 1273–1291 (2017)
6. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
7. Gibson, J.J., Carmichael, L.: The senses considered as perceptual systems, vol. 2. Houghton Mifflin Boston (1966)
8. Gopalakrishnan, A., van Steenkiste, S., Schmidhuber, J.: Unsupervised object keypoint learning using local spatial predictability. ICLR 2021 (2020)
9. Guo, X., Zhu, E., Liu, X., Yin, J.: Affine equivariant autoencoder. In: IJCAI. pp. 2413–2419 (2019)
10. Ha, D., Schmidhuber, J.: World models. arXiv preprint arXiv:1803.10122 (2018)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
12. Held, R., Hein, A.: Movement-produced stimulation in the development of visually guided behavior. Journal of comparative and physiological psychology **56**(5), 872 (1963)
13. Ilonen, J., Bohg, J., Kyrki, V.: Three-dimensional object reconstruction of symmetric objects by fusing visual and tactile sensing. The International Journal of Robotics Research **33**(2), 321–341 (2014)
14. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
15. Kipf, T., van der Pol, E., Welling, M.: Contrastive learning of structured world models. arXiv preprint arXiv:1911.12247 (2019)
16. Kulkarni, T.D., Gupta, A., Ionescu, C., Borgeaud, S., Reynolds, M., Zisserman, A., Mnih, V.: Unsupervised learning of object keypoints for perception and control. Advances in neural information processing systems **32** (2019)

17. Lake, B.M., Ullman, T.D., Tenenbaum, J.B., Gershman, S.J.: Building machines that learn and think like people. Behavioral and brain sciences **40** (2017)
18. Locatello, F., Bauer, S., Lucic, M., Raetsch, G., Gelly, S., Schölkopf, B., Bachem, O.: Challenging common assumptions in the unsupervised learning of disentangled representations. In: international conference on machine learning. pp. 4114–4124. PMLR (2019)
19. Lowe, D.G.: Object recognition from local scale-invariant features. In: Proceedings of the seventh IEEE international conference on computer vision. vol. 2, pp. 1150–1157. Ieee (1999)
20. Marchetti, G.L., Tegnér, G., Varava, A., Kragic, D.: Equivariant representation learning via class-pose decomposition. arXiv preprint arXiv:2207.03116 (2022)
21. Mondal, A.K., Jain, V., Siddiqi, K., Ravanbakhsh, S.: Eqr: Equivariant representations for data-efficient reinforcement learning. In: International Conference on Machine Learning. pp. 15908–15926. PMLR (2022)
22. Mondal, A.K., Nair, P., Siddiqi, K.: Group equivariant deep reinforcement learning. arXiv preprint arXiv:2007.03437 (2020)
23. Noë, A., Noë, A., et al.: Action in perception. MIT press (2004)
24. Otsu, N.: A threshold selection method from gray-level histograms. IEEE transactions on systems, man, and cybernetics **9**(1), 62–66 (1979)
25. Park, J.Y., Biza, O., Zhao, L., van de Meent, J.W., Walters, R.: Learning symmetric embeddings for equivariant world models. arXiv preprint arXiv:2204.11371 (2022)
26. van der Pol, E., Kipf, T., Oliehoek, F.A., Welling, M.: Plannable approximations to mdp homomorphisms: Equivariance under actions. arXiv preprint arXiv:2002.11963 (2020)
27. Rezende, D.J., Mohamed, S., Wierstra, D.: Stochastic backpropagation and approximate inference in deep generative models. In: International conference on machine learning. pp. 1278–1286. PMLR (2014)
28. Schiebener, D., Morimoto, J., Asfour, T., Ude, A.: Integrating visual perception and manipulation for autonomous learning of object representations. Adaptive Behavior **21**(5), 328–345 (2013)
29. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4104–4113 (2016)
30. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017)
31. Triggs, B., McLauchlan, P.F., Hartley, R.I., Fitzgibbon, A.W.: Bundle adjustment—a modern synthesis. In: International workshop on vision algorithms. pp. 298–372. Springer (1999)
32. Tsikos, C., Bajcsy, R.: Segmentation via manipulation. IEEE Transactions on Robotics and Automation **7**(3), 306–319 (1991)
33. Watter, M., Springenberg, J., Boedecker, J., Riedmiller, M.: Embed to control: A locally linear latent dynamics model for control from raw images. Advances in neural information processing systems **28** (2015)