# Boosting Object Representation Learning via Motion and Object Continuity

Quentin Delfosse[1(✉)], Wolfgang Stammer[1,2], Thomas Rothenbächer[1], Dwarak Vittal[1], and Kristian Kersting[1,2,3,4]

[1] AI & ML Lab, TU Darmstadt, Darmstadt, Germany
quentin.delfosse@cs.tu-darmstadt.de
[2] Hessian Center for AI (hessian.AI), Darmstadt, Germany
[3] German Research Center for AI (DFKI), Kaiserslautern, Germany
[4] Centre for Cognitive Science, TU Darmstadt, Darmstadt, Germany

**Abstract.** Recent unsupervised multi-object detection models have shown impressive performance improvements, largely attributed to novel architectural inductive biases. Unfortunately, despite their good object localization and segmentation capabilities, their object encodings may still be suboptimal for downstream reasoning tasks, such as reinforcement learning. To overcome this, we propose to exploit object motion and continuity (objects do not pop in and out of existence). This is accomplished through two mechanisms: (i) providing temporal loss-based priors on object locations, and (ii) a contrastive object continuity loss across consecutive frames. Rather than developing an explicit deep architecture, the resulting unsupervised Motion and Object Continuity (MOC) training scheme can be instantiated using any object detection model baseline. Our results show large improvements in the performances of variational and slot-based models in terms of object discovery, convergence speed and overall latent object representations, particularly for playing Atari games. Overall, we show clear benefits of integrating motion and object continuity for downstream reasoning tasks, moving beyond object representation learning based only on reconstruction as well as evaluation based only on instance segmentation quality.

**Keywords:** Object Discovery · Motion Supervision · Object Continuity

## 1 Introduction

Our surroundings largely consist of objects and their relations. In fact, decomposing the world in terms of objects is considered an important property of human perception and reasoning. This insight has recently influenced a surge of research articles in the field of deep learning (DL), resulting in novel neural architectures and inductive biases for unsupervisedly decomposing a visual scene into objects (*e.g.* [6,15,16,20,33,39,41,49]). Integrating these modules into systems for downstream reasoning tasks, *e.g.* playing Atari games in reinforcement


D. Koutra et al. (Eds.): ECML PKDD 2023, LNAI 14172, pp. 610–628, 2023.
https://doi.org/10.1007/978-3-031-43421-1_36

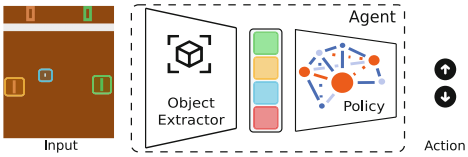


**Fig. 1.** An object-centric reasoner playing Pong. The agent first extracts the object representation and then reasons on them to select an optimal action.

learning (RL) settings is a promising next step for human-centric AI (*cf.* Fig. 1). This integration could provide benefits both in overall performance and in terms of trustworthy human-machine interactions [32,48].

However, although many of the previously mentioned works motivate their object discovery methods with the advantage of object-centric learning for complex downstream tasks, none of these explicitly optimize the object encodings for anything other than reconstruction. In fact, investigations of the object encodings of SPACE [39], SOTA on object detection for Atari games, reveal two shortcomings. First, although object detection does work for some games (*e.g.* Space Invaders), it shows significant issues on other, even as simple as Pong (*cf.* Fig. 2 for qualitative examples (top) and quantitative F-score evaluations (left)). Second, even for such a simple downstream task as object classification, its object encodings appear suboptimal even for games with high detection scores. This is exhibited by the accuracies of a ridge regression model [24,25] trained on SPACE's encodings for object classification (Fig. 2 (left)). The encodings, mapped into a two-dimensional t-SNE [54] embedding and colored by their ground truth class labels (*cf.* Fig. 2 (right)), suggest a cluttered latent space.

These results indicate open issues from two types of errors: (**Type I**) failures in object detection per se and (**Type II**) sub-optimal object representations. Arguably, Type I is somewhat independent of Type II, as an optimal encoding is not a necessity for detecting objects in the first place. However, the Type II error is dependent on Type I, as an object representation can only exist for detected objects. Before integrating such recent modules into more general systems for reasoning tasks, we first need to tackle the remaining issues.

We therefore propose a novel model-agnostic, self-supervised training scheme for improving object representation learning, particularly for integration into downstream reasoning tasks. We refer to this scheme as Motion and Object Continuity supervision (MOC). MOC jointly tackles Type I and II errors by incorporating the inductive biases of object motion (*i.e.* objects tend to move in the real-world) and object continuity (*i.e.* objects tend to still exist over time and do not pop in and out of existence) into object discovery models.

It is based on the notion that visual input for humans is mainly perceived as image sequences. These contain rich object-based information signals, such as that objects tend to move over time, but also that an object identified in one frame will likely be present in the consecutive ones. We refer to the first property as *object motion* (M) and the second property as *object continuity*

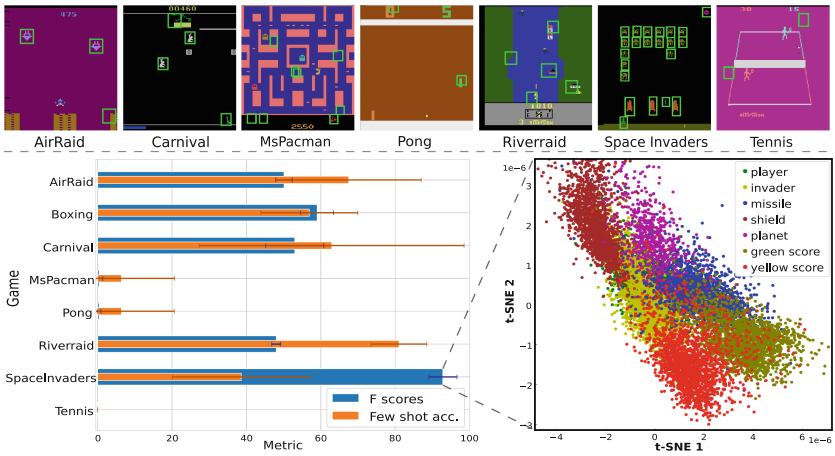


**Fig. 2.** Motivational example: unsupervised object detection models are insufficient for downstream tasks such as classification, exemplified here via SPACE [39] on Atari environments. Top: Example images of SPACE detecting objects on different Atari games. Left: F-score for object detection (blue) and few shot classification accuracy of object encodings via ridge regression (orange, 64 objects per class, 0% accuracy corresponds to no object detected). Right: Two-dimensional t-SNE embedding of object encodings produced by SPACE for Space Invaders. (Color figure online)

(OC). The concept of object continuity can be considered an extension of the Gestalt law of continuity [57] to the temporal dimension and has been identified to be an important property of infant and adult visual perception [51]. MOC specifically makes use of this underlying information via two mechanisms. The first mechanism produces accurate object location priors based on estimated optical flow. The second presents a contrastive loss term on the object encodings of consecutive images. Importantly, due to the model-agnostic approach, MOC can incorporate any object discovery model.

In our experimental evaluations, we show the benefits of the novel MOC training scheme for object-centric learning for downstream reasoning tasks. For this, we integrate the models SPACE [39] and Slot Attention [41] into MOC and quantify the benefits of MOC over the base models through a variety of metrics that highlight important properties of the training scheme. Importantly, and in contrast to previous works, we show the improved quality of MOC trained object encodings for downstream reasoning task performances such as playing Atari games and few-shot object classification. In summary, we show that **inductive biases extracted from motion and object continuity are enough to boost object encodings of a predisposed object discovery model for downstream reasoning tasks**.

Overall, our contributions are the following: (i) We identify two error sources of object discovery models, hindering their current integration into modules for downstream reasoning tasks. (ii) Introduce motion *and* object continuity to

DL as a novel self-supervised training scheme. (iii) Create a novel open-source dataset, Atari-OCTA, to train and evaluate object detection models on Atari games. (iv) Empirically show on the SPACE and Slot Attention architectures that motion and object continuity greatly boosts downstream task performance for object-centric DL and allow for object-centric RL.

We proceed as follows. We first give a brief introduction to object discovery methods, then explain our MOC scheme, detailing both Motion and Object Continuity supervision. We experimentally evaluate improvements on object detection and representations. Before concluding, we touch upon related work.

## 2 Motion and Object Continuity

The Motion and Object Continuity (MOC) training scheme utilizes two object properties for improving object detection and representation, both of which are required for good object representation learning for reasoning tasks and both of which can be discovered in visual observations over time. The first property corresponds to the fact that objects tend to move in an environment. Via optical flow estimations [50,53] this can be translated to a motion detection mechanism that guides object localization and transforms unsupervised object detection into a "motion supervised" one. The second property describes the observation that objects exist at consecutive time points in proximity to their initial location and do not simply disappear. This can be integrated into an object continuity constraint, enforcing encoding alignment of objects over time.

Our approach allows integrating any out-of-the-box object discovery method. For this work we have focused on providing MOC supervision to slot-based models [41] and to models that separately encode the background and the foreground via separate variational autoencoders (VAE) (*e.g.* [20,39]), where we specifically focus on SPACE [39] due to its SOTA object discovery performance on images from Atari games. We provide an overview of these methods in Fig. 7 (*cf.* App. B) and introduce the mathematical modelling of SPACE in the App. B.1 and of Slot Attention in the App. B.2. In the following, we present a high-level description of MOC supervision, illustrated in Fig. 3, independent of the base models implementation. MOC implementation details for each model can be found in App. C.2 and C.3. Let us first provide some details on notations.

In the following we denote `enc` as the object encoding obtained from a base model, *e.g.* the slot encodings for Slot Attention [41], or the $z_{what}$ encodings for SPACE [39]. We further denote `loc` as the positional representation of an object. Specifically, SPACE divides an image into a grid, and explicitly models the presence of an object in each cell using a $z_{pres}$ variable, and its position using a $z_{where}$ variable. For each object, `loc` can thus be obtained using the $z_{where}$ variable of each cell in which an object is present (*e.g.* $z_{pres} > 0.5$). For Slot Attention based models, on the other hand, the position information is implicitly contained within `enc`. However, the corresponding attention masks $\alpha$ also correspond to valid representations of object locations. We therefore denote these masks as the `loc` variables of Slot Attention in the MOC framework.

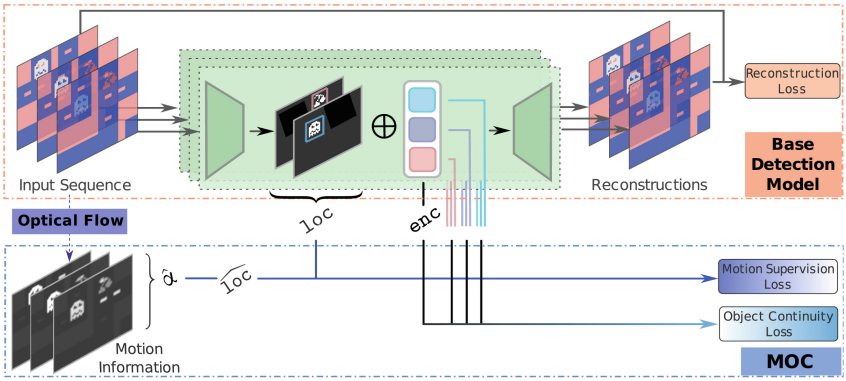


**Fig. 3.** An overview of the MOC training scheme applied to a base object detection model, which provides location and object representations. In our MOC training scheme, (i) motion information (dark blue), is extracted from each frame, allowing to detect objects and directly update the model's latent location variables (`loc`). (ii) Object continuity (black + cyan) aligns the encodings (`enc`) of spatially close objects of consecutive frames using a contrastive loss. (Color figure online)

Furthermore, we denote $\mathcal{L}^{\mathrm{B}}$ to represent the original training loss function of the base model, *e.g.* reconstruction loss.

### 2.1 Motion Supervision

Let us now consider sequences of $T$ frames $\{\mathbf{x}_t\}_{t=1}^{T}$, whereby $\mathbf{x}_t$ corresponds to an RGB image ($\mathbf{x}_t \in \mathbb{R}^{3\times h\times w}$), at time step $t$. Given such a sequence of images, MOC requires preprocessing via any suitable optical flow estimation method (*e.g.* [2,3,28,62]) which accurately detects the motion of objects for the given domain and provides us with sufficient binary masks $\hat{\alpha}$ of the moving objects from which location information, $\widehat{\texttt{loc}}$, can be obtained for each object.

MOC now integrates these masks to provide feedback on the locations of initial object representations. Specifically, for each frame $\mathbf{x}_t$ we compute:

$$\mathcal{L}^{\mathrm{M}}(\mathbf{x}_t) := \sum_{i=1}^{N} \mathcal{L}^{\mathrm{M}^*}(\mathbf{x}_t, \hat{\alpha}_t), \tag{1}$$

where $\mathcal{L}^{\mathrm{M}^*}$ refers to the exact implementation for the base model including a weighting hyperparameter (*cf.* App. C.2 and C.3 for SPACE and Slot Attention, respectively). In short, the masks obtained from optical flow ($\hat{\alpha}_t$) allow for direct supervision on the internal masks representation of Slot Attention based models, and allow us to construct $\widehat{\texttt{loc}}$ variables, to supervise $z_{pres}$ and $z_{where}$ in SPACE.

### 2.2 Object Continuity

Given that an object in consecutive image sequences tends to change its location only gradually and does not pop in and out of existence, we can integrate such

bias by matching objects of similar position and latent encoding over consecutive frames. Particularly, we can make use of this information for explicitly improving the encoding space through a contrastive loss.

For detected entities in consecutive frames, we apply the object continuity loss, $\mathcal{L}^{\text{OC}}$, on the internal representations (*i.e.* `enc`) of the model. This loss makes the representations of the same object depicted over several frames similar, and of different objects heterogeneous. We estimate whether two objects in consecutive frames represent the same entity based on their `loc` and `enc` variables.

In detail, we denote $\Omega_t = \{o_t^j\}_{j=1}^{c_t}$ the set of object representations ($o_t^j = (\texttt{enc}_t^j, \texttt{loc}_t^j)$) that correspond to the $c_t$ detected objects in $\mathbf{x}_t$. Let $\texttt{enc}^*(o_t^j)$ denote a function providing the object in $\Omega_{t+1}$ with the most similar encoding to $o_t^j$ based on the cosine similarity ($S_C$). Correspondingly, $\texttt{loc}^*(o_t^j)$ is a function that provides the nearest object of $\Omega_{t+1}$ to $o_t^j$ based on the locations (*e.g.* via Euclidean distance). We thus introduce the contrastive object continuity loss as:

$$\mathcal{L}^{\text{OC}}(\mathbf{x}_t) = \sum_{o_t^i \in \Omega_t} \sum_{o_{t+1}^j \in \Omega_{t+1}} \lambda_{\text{differ}} \cdot S_C(o_t^i, o_{t+1}^j), \tag{2}$$

where $\lambda_{\text{differ}}$ is defined, using the hyperparameter $\beta \in \mathbb{R}^+$, as:

$$\lambda_{\text{differ}} = \begin{cases} -\beta & \text{if } o_{t+1}^j = \texttt{enc}^*(o_t^i) = \texttt{loc}^*(o_t^i) \\ 1 & \text{else} \end{cases} \tag{3}$$

This loss allows the models to have better internal representation of an object, for which the visual representation can vary across the frames.

### 2.3 General Training Scheme

The MOC scheme thus overall adds motion and object continuity supervision to a base object detection model, resulting in the final loss function:

$$\mathcal{L}^{\text{MOC}} := \mathcal{L}^B + (1 - \lambda_{align}) \cdot \mathcal{L}^{\text{M}} + \lambda_{align} \cdot \lambda_{\text{OC}} \cdot \mathcal{L}^{\text{OC}}. \tag{4}$$

$\mathcal{L}^{\text{M}}$ represents the batch-wise motion supervision loss and $\mathcal{L}^{\text{OC}}$ the batch-wise object continuity loss. We use $\lambda_{align} \in [0, 1]$, to balance the learning of object detection and object representation. We recommend using a scheduling approach for $\lambda_{align}$, described in the App. C.4. An additional $\lambda_{\text{OC}}$ hyperparameter can be used to balance image reconstruction capabilities and encoding improvements.

Concerning our previously suggested error types of object representation learning: the goal of $\mathcal{L}^{\text{OC}}$ is to improve on the Type II error. As previously identified, improvements on Type II also depend on improvements on the Type I error for which $\mathcal{L}^{\text{M}}$ was developed. The ultimate goal of MOC to improve deep object representation learning for downstream reasoning tasks is thus achieved by interdependently improving over both error types.

Lastly, MOC takes advantage of temporal information, while leaving the base architecture unchanged. In this way, it is possible to learn the concept of objects from image sequences, while still allowing for single image inference.

## 3 Experimental Evaluations

Let us now turn to the experimental evaluation of the MOC training scheme. With our experimental evaluations, we wish to investigate whether motion and object continuity biases can benefit object-centric DL for downstream reasoning tasks. Specifically, we investigate the effect of our MOC scheme on object discovery abilities and on object representation learning, such that these can be used for downstream tasks such as object-centric game playing and few-shot object classification. More precisely, we address the following research questions: (**Q1**) Does MOC benefit object discovery? (**Q2**) Does MOC improve latent object representations? (**Q3**) Does MOC improve representations for complex downstream reasoning tasks, such as game playing and few-shot object classification?

**Experimental Setup.** To this end, we consider the SPACE model [39] and a Slot Attention model [41] as base object discovery models, and compare performances of these trained with (+MOC) and without our MOC scheme (baseline, models only optimized for reconstruction). As we are ultimately interested in RL as downstream reasoning task, we train these models on images obtained from different reinforcement learning environments of the Atari 2600 domain [5]. As SPACE is the only object discovery model—to the best of our knowledge—trained on images obtained from different of the mostly used Atari environments, we focus the bulk of our evaluations on this, but provide results also on a Slot Attention model to show the generality and improvements across different base models. Both architectures are visualized in Fig. 7, with further details and training procedures (*cf.* App. B). We specifically focus on a subset of Atari games, namely AirRaid, Boxing, Carnival, MsPacman, Pong, Riverraid, SpaceInvaders and Tennis, many of which were investigated in the original work of SPACE. Note that this subset of Atari games contains a large variance in the complexity of objects, *i.e.* the number, shape, and size of objects, as well as games with both static and dynamic backgrounds, representing a valid test ground.

Overall, we evaluate 3 model settings: the original base models, SPACE and Slot Attention (SLAT), and both incorporated in our full MOC training scheme (SPACE+MOC and SLAT+MOC). We further provide specific ablation experiments in which the base model is only trained via $\mathcal{L}^B$ and $\mathcal{L}^M$, but without $\mathcal{L}^{OC}$. For each experiment, the results are averaged over converged (final) states of multiple seeded runs. We provide details on the hyperparameters in App. B for both models and on the evaluation metrics in each subsection, and further in App. E. Unless specified otherwise, the reported final evaluations are performed on a separated, unseen test set for each game.

As the aim of our work focuses on RL for Atari games, in our experimental evaluations we use a basic optical flow (OF) technique described in App. C.1 which provides us with sufficient masks $\hat{\alpha}$ of moving objects. As previously noted, the exact optical flow implementation, however, is not a core component of MOC and can be replaced with any other out-of-the-box OF approach such as [50,53]. All figures are best viewed in color.

**Atari-OCTA.** As the Atari dataset from [39] is not labelled and thus insufficient particularly for evaluating object encodings, we created a novel version of it.

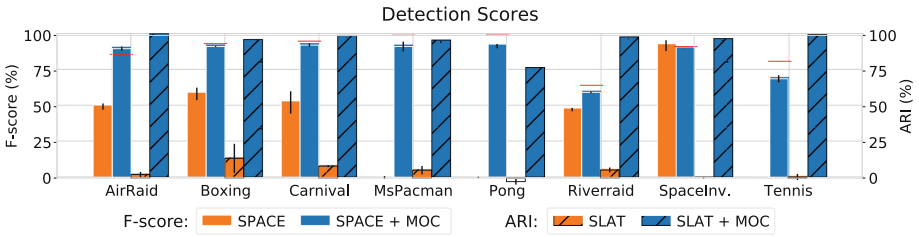


**Fig. 4.** MOC improves object detection. Final F-scores of SPACE models and Adjusted Random Index of Slot Attention (SLAT), both with and without MOC over frames of different Atari-OCTA games. Training via MOC leads to massive improvements over the set of investigated games. Optical flow F-scores are provided in red. They indicate the potential F-score upper-bound obtainable if using Motion supervision only. (Color figure online)

For this, we created one labelled dataset per game, where positions, sizes, and classes of all objects are provided. We used OC-Atari [10] to extract the objects properties. We separate classes of objects that are relevant for successful play from the objects of the HUD (*e.g.* scores, lives), and base our metrics on this first set, as we investigate MOC in the context of RL as reasoning task. Furthermore, we provide image sequences of $T = 4$ for the training data. For details on how Atari-OCTA (the resulting object-centric and time annotated version of Atari) was created, see App. A. Atari-OCTA and the generation code are provided along with this paper.[1] Our evaluations were performed on Atari-OCTA.

**MOC Benefits Object Discovery (Q1).** Let us now move to the actual research question evaluations. First, we evaluate the influence of our training scheme on a base model's performances in the originally intended task of object discovery, thus investigating the potential improvements over the Type I error alone. For this, we compute the foreground F-scores (further denoted as F-score) of SPACE as in the original work of [39] and compare to SPACE+MOC, *i.e.* SPACE incorporated in MOC. As Slot Attention is missing explicit variables for computing the F-scores, we revert to providing foreground Adjusted Rand Index (further denoted as ARI) scores for SLAT as well as SLAT+MOC, as also provided in the original work by Locatello et al. [41].

First, Fig. 4 presents the final foreground object detection scores (i.e. F-scores for versions of SPACE and ARI for versions of SLAT) of the different training setups for images from the eight Atari-OCTA games. Over the set of these, adding motion and object continuity supervision leads to great improvements in object discovery, with an average improvement from 38% to 85% and 4% to 95% over all games for SPACE+MOC and SLAT+MOC, respectively.

In particular, the base SPACE model on average performs quite unsatisfactorily. The base SLAT model performs even more poorly, importantly indicating SPACE's superiority over SLAT for Atari images, possibly due to the object size

[1] https://github.com/k4ntz/MOC.

**Table 1.** Faster convergence through MOC. The average number of steps needed to get an F-score $> 0.5$ ($\infty$ if never reached) on the test set for SPACE, with and without MOC on Atari-OCTA. Lower values are better. For each model, we also provide the average. A complete evolution of the F-scores is given in App. D.

| Game | Airraid | Boxing | Canival | MsPac. | Pong | Riverraid | Sp.Inv. | Tennis | Avg |
|---|---|---|---|---|---|---|---|---|---|
| SPACE | 2600 | 2200 | 420 | $\infty$ | $\infty$ | 3400 | 430 | $\infty$ | 1800 |
| SP.+MOC | **300** | **240** | **350** | **350** | **270** | **270** | **290** | **260** | **280** |

prior of SPACE's and its grid-based approach. On games such as MsPacman and as arguably simple as Pong, SPACE and Slot Attention even appear to fail. Here, using our MOC scheme leads to more than 90% increase in object detection final performances. We also note the reduced performance variance of both +MOC models, suggesting more reliable training via MOC.

Furthermore, object biases via MOC not only improve final object detection performance, but also aid in obtaining such performances with fewer training iterations. Table 1 presents how many training steps are required for SPACE, on average, to obtain 50% of the validation F-scores of the different training schemes on the 8 Atari games. Over all games SPACE+MOC models approach convergence in much fewer number of steps. In fact, SPACE is able to reach an F-score above 50% on only 5 out of 8 games. For these games, it takes on average 1820 steps to reach this threshold, compared to 280 on average for SPACE+MOC on all games, hence leading to more than 7 times faster learning.

We provide detailed F-score progressions on all games, as well as final precision vs recall curves in App. (*cf.* Fig. 9 & 10). We also provide results with SPACE+MOC w/o OC indicating that updating object encodings via the OC loss in MOC on average shows equivalent object detection performances as without. This is an intuitive finding and shows that the motion loss, intentionally developed for tackling the Type I error, achieves its purpose. Our results in summary indicate a strong benefit of object location biases via MOC for object detection, thus affirming **Q1**.

**MOC Improves Latent Object Encodings (Q2).** In the previous section, we presented improvements of base object discovery models, SPACE and SLAT, via MOC, focusing on the task of object discovery for which these models were mainly developed for and evaluated on. However, our investigations in Fig. 2 exhibited that even when the object detection performance was promising (*e.g.* Space Invaders) the object representations themselves proved less useful for a downstream reasoning task as simple as object classification. Thus, apart from improvements in the detection itself, additional improvements need to be achieved in terms of optimal object representations for such models to actually be integrated into a more complex downstream setup. With **Q2**, we thus wish to investigate the effect of the MOC scheme on improving a base model's Type II errors, *i.e.* the usefulness of a model's latent object representations.

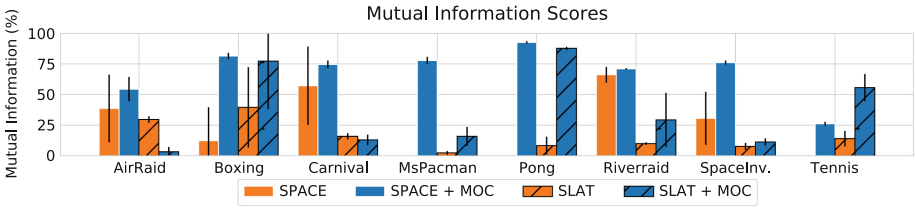


**Fig. 5.** MOC leads to more optimal object encodings as indicated via mutual information score. The adjusted mutual information of object encodings from SPACE and Slot Attention (SLAT), both with and without MOC, of Atari-OCTA are presented (mean ± std). Higher average values are better.

To answer **Q2**, we focus on the adjusted mutual information (AMI) of the encodings. Essentially, one computes a clustering on the latent object encodings and compares the clustering to a ground truth clustering (*cf.* App. F for details). Figure 5 presents the AMI for the two configurations: base model (i.e. SPACE and SLAT) and +MOC for training with MOC. One can observe immense improvements for MOC trained models in comparison to the base models. Averaged over all games we observe an increase in AMI from 26% to 69% and 16% to 37% for SPACE+MOC and SLAT+MOC, respectively.

In addition, one can observe the benefits of leveraging both $\mathcal{L}^{\text{OC}}$ *and* $\mathcal{L}^{\text{M}}$ in an ablation experiment for SPACE in which SPACE+MOC w/o OC is trained only via the motion supervision loss. Results can be found in App. D.5. As the object encodings in the original SPACE and in SPACE+MOC w/o OC are only optimized for reconstruction, there is little supervision to produce distinct object representations. Our results thus indicate that $\mathcal{L}^{\text{OC}}$ provides a strong supervisory signal on the latent encoding space. Lastly, as seen by the reduced cross validation variance, on average MOC produces less sensitive models (supporting the findings of Fig. 4). We provide detailed AMI scores in App. (*cf.* Fig. 13).

These results overall highlight the improvements of MOC, particularly the benefits of integrating both supervisory mechanisms into one training scheme, as well as $\mathcal{L}^{\text{OC}}$'s benefit for improving on Type II errors. Conclusively, we can affirm **Q2**: MOC does improve latent encodings.

**MOC Improves Object Encodings for Downstream Reasoning Tasks (Q3).** AMI is one way of measuring the quality of latent encodings. Ultimately, we are interested in integrating unsupervised object discovery models into more complex downstream tasks, *e.g.* integrating these into RL. In the following, we focus on evaluating object encodings for two specific downstream reasoning tasks, namely few-shot object classification and Atari game playing. The resulting performances in these downstream tasks thus act as important additional evaluations on the quality of learned object encodings and act as the main motivation behind our MOC scheme.

Let us begin with few-shot object classification. For each game, we optimize linear ridge regression models on 1, 4, 16 and 64 object encodings of each object

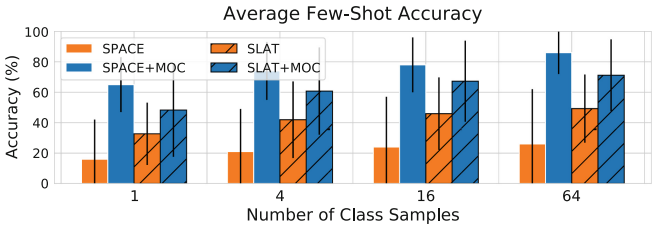


**Fig. 6.** MOC improves object representations for the few-shot classification task. Average few-shot accuracy performance (in %) based on latent object representations. We provide a ridge regression model with 1, 4, 16 and 64 encodings per class and test it on a held-out test set. The values are averaged over the 8 investigated Atari-OCTA games for SPACE and Slot Attention (SLAT), with and without MOC. Detailed, per game results are in App. G.2.

class and measure the classification accuracy on a separate held-out test set. We thus assume useful encodings to correspond to encodings that are similar for objects of the same class. As such, these representations should be easier to be differentiated by a simple (linear) classification method.

The results can be seen in Fig. 6. Specifically, we see a large boost in classification performance via the full MOC training scheme, visible both in SPACE+MOC and SLAT+MOC performances. Note the strong contrast, particularly for SPACE in test accuracy in the very small data regime, *e.g.* when the classifier has only seen 1 object encoding per class. Here, SPACE+MOC reaches 65% average accuracy, 3.5× higher than SPACE and 1.3× higher than SPACE+MOC w/o OC (*cf.* App. D.3). Detailed results on each game are provided in App. (*cf.* Fig. 12) and additional qualitative results can be found in App. D.6 and D.7.

We finally evaluate MOC in the context of RL for game playing as a representative complex downstream reasoning task and investigate in how far MOC can improve a base model's performance such that its learned representations can be used in playing Atari games, as done by [11] in logic oriented settings. We base our evaluations on concept-bottleneck [34] RL agents that are initially trained using perfect information (Info*) with the Dueling Q-learning algorithm [56] and focus here on playing a subset of the games from our previous evaluations (Boxing and Pong). These games are the only two from Atari-OCTA with a fixed number of objects, which allows for a very simple policy architecture, and for which perfect information obtained from the RAM using [1] is available. We note that although many RL agents might possibly provide higher overall game scores than the ones we focus on, the point to make here is that training object encodings via MOC can greatly improve *a* baseline (object-centric) RL agent.

For these experiments we focus only on SPACE as base model (rather than SLAT) due to its superior base performance on object detection in Atari images. We next replace the object extraction from RAM with object extraction performed by SPACE, and SPACE+MOC models but keep the initially learned

**Table 2.** MOC (via SPOC) allows object-centric playing via concept-bottleneck models. Average scores among all agents (avg) and of the best agent (best) of object-centric agents that detect objects based on SPACE and SPOC. Scores for each seed are in App. G.3. Additional base comparisons are agents with perfect information (Info*), random agents and human (from [55]).

| | Method | SPACE | SP.+MOC | Info* | Random | Human |
|---|---|---|---|---|---|---|
| avg | Boxing | $-3.5 \pm 4.3$ | $\mathbf{8.4} \pm 9.9$ | $36 \pm 17$ | $-0.5 \pm 2.$ | 4.3 |
| | Pong | $-21 \pm 0.5$ | $\mathbf{-11} \pm 12$ | $20 \pm 1.3$ | $-21 \pm 0.3$ | 9.3 |
| best | Boxing | $3.8 \pm 6.5$ | $\mathbf{22} \pm 14$ | $52 \pm 3.5$ | $2.6 \pm 3.3$ | - |
| | Pong | $-20 \pm 0.8$ | $\mathbf{4.8} \pm 11$ | $21 \pm 0.$ | $-20 \pm 0.7$ | - |

policies. For this, we transform the `loc` variables of these models into $(x, y)$ coordinates (ignore sizes). For object classes, we use the previously evaluated (same seeded) few-shot classification models on the object encodings (with 16 samples per class, *cf.* Fig. 12). The object coordinates are given to a 3 fully connected layers network, which predicts the Q-value for each action. We provide additional comparisons to human-level performance (H) and random agents (R).

As can be seen in Table 2, MOC greatly empowers the evaluated object-centric agents. On average, SPACE+MOC-based agents largely outperform SPACE-based agents, and even obtain higher average scores than humans on Boxing. For both games, the benefits of our approach is even more remarkable on mean scores obtained by best playing agents, due to the fact that a slightly more accurate object detector results in more reliable states, and thus better performing agent. Our experimental evaluations thus allow us to validate that MOC benefits object-centric reasoning tasks spanning from RL agents game playing capabilities to few-shot object classification (**Q3**).

## 4 Related Work

Our work touches upon several topics of ML research. In the following, we discuss related works from these areas.

**Unsupervised Object Detection.** Object-centric DL has recently brought forth several exciting avenues of unsupervised and self-supervised object discovery research by introducing inductive biases to neural networks to extract objects from visual scenes in an unsupervised manner [6,9,15,16,20,29,30,35,39,41,47,49,60]. We refer to [21] for a detailed overview. However, until recently [45,46] these approaches have shown limited performances on more complex images. All of these mentioned works focus on object discovery from independent images. Importantly, although most of these approaches are motivated with the benefits of object representations for more complex downstream tasks, none of these apply additional constraints onto the latent object representations and only optimize object encodings for reconstruction. Lastly, among these works

are more recent approaches that could provide improved baseline performances than SPACE and the vanilla Slot Attention model (*e.g.* [45,46]), however MOC can also be applied to these and these should be seen as orthogonal to our work.

**Unsupervised Object Detection from Videos.** Leveraging the supervisory signal within video data for object detection goes back several years *e.g.* to approaches based on conditional random fields [44]. More recently, DL approaches, *e.g.* [61] explicitly factorize images into background, foreground and segmentation masks. Also, [13] argue for the value of motion cues and showcasing this on physical reasoning tasks via a dynamics-based generative approach. [64] do multi-object segmentation via a carefully designed architecture of multiple submodules that handle certain aspects of the overall task. This idea is also taken up by [52] who introduce an approach for generative object-centric learning via the property of *common fate*. The recent works SAVi [33] and SAVI++ [14] improve the object discovery abilities of Slot Attention by using optical flow and depth signals, respectively, as a training target. Where many of these works present specifically designed architectures for video processing, MOC represents a model-agnostic scheme which, in principle, can incorporate any base object detection model and in this sense also be applicable to such tasks as semantic world modeling and object anchoring [43]. Additionally, although many mentioned works identify the importance and benefits of motion for object discovery, only [13] also focus on the advantages of learning good object representations for downstream tasks.

**Optical Flow.** Optical flow methods aim at finding translation vectors for each pixel in an input image to most accurately warp the current frame to the next frame. In other words, optical flow techniques try to find pixel groups with a common fate. Estimating optical flow is a fundamental problem of computer vision, starting with such approaches as the Gunnar-Farneback Algorithm [17] (for other traditional approaches, please refer to [18]). By now it is widely used in a variety of applications such as action recognition [7,37], object tracking [31,63], video segmentation [38,40]. Most recent works are based on DL approaches [12, 26–28,42,50,53]. Recently optical flow has also been applied to RL settings [19, 59] which hint the downstream RL agent at possibly important object locations, but do not perform object representation learning. As previously mentioned, optical flow is considered as a preprocessing step in MOC, thus any suitable OF method can in principle be used for a specific MOC implementation.

**Self-supervised Learning.** The idea of motion supervised learning is related to the large field of self-supervised learning. In fact the motion supervision aspect of MOC stands in line with other approaches [14,33,58] and notably [4] and confirms their findings on the general benefit of motion priors for object detection, however these works do not specifically improve or otherwise evaluate obtained object encodings for anything other than object detection. Aside from this, the $\mathcal{L}^{\text{OC}}$ loss of MOC can be considered as a form of contrastive loss, popularly used in self-supervised learning. Among other things, recent works apply patch-based contrasting [23], augmentations [8], cropping [22] or RL based contrasting

[36]. In comparison, MOC contrasts encodings between consecutive time steps. These approaches do not perform object representation learning, but rather try to improve the full encoding of an image.

## 5 Limitations and Future Work

As we focus evaluations on Atari environments for classification and game playing tasks, we make use of a very simple optical flow technique. A more advanced optical flow approach is necessary for more complex environments. The application of MOC to other types of object discovery models is a necessary next step in future work. Additionally, the $\mathcal{L}^{\text{OC}}$ of our SLAT+MOC implementation produces larger variances in AMI than that of SPACE+MOC (*cf.* Fig. 5). Future work should investigate ways of stabilizing this. MOC contains a bias towards moving objects, however it is desirable for a model to also detect stationary objects. Although the scheduling approach tackles this issue, additional investigations are necessary. The bottleneck-based [34] RL agents of **Q3** are somewhat limited, as additional techniques to increase detection robustness (*e.g.* Kalman filters) were not used. Certainly, more complicated models and RL algorithms would be worth investigating in future research. Lastly, certain environments also include static objects that are important to interact with, we would thus like to investigate object recognition based on agent-environment interactions. Following this idea, we think that performing representation learning based on the RL signal (*i.e.* cumulative reward) is another interesting line of research.

## 6 Conclusions

Already young children quickly develop a notion of object continuity: they know that objects do not pop in and out of existence. Here, we have demonstrated that deep networks also benefit from this knowledge. Specifically, we have shown that properties of objects that are observed from time can be leveraged to improve object representation learning for downstream tasks. This MOC training scheme, consisting of a motion supervision and object continuity contrastive loss, greatly improves a base model's object discovery and representation performances. In this way MOC collectively tackles two interdependent error sources of current object discovery models. Finally, along with this paper, we provide the novel dataset Atari-OCTA, an important step for evaluating object discovery in complex reasoning tasks such as game playing.

Apart from performance benefits that object-centric approaches can bring to deep learning, they particularly play an important role in human-centric AI, as such machines can perform inference and provide explanations on the same object-level as human-human communication [32]. Additional avenues for future work include investigating interactive object learning, such as agent-environment interactions, but also human-machine interactions for developing more reliable and trustworthy AI models.

**Acknowledgements.** The authors thank the anonymous reviewers for their valuable feedback. This research work has been funded by the German Federal Ministry of Education and Research and the Hessian Ministry of Higher Education, Research, Science and the Arts (HMWK) within their joint support of the National Research Center for Applied Cybersecurity ATHENE, via the "SenPai: XReLeaS" project. It also benefited from the HMWK cluster projects "The Third Wave of AI" and "The Adaptive Mind" as well as the Hessian research priority program LOEWE within the project "WhiteBox".

**Ethical Statement.** Our work aims to improve the object representations of object discovery models, specifically targeting the improvements of their use in additional modules in downstream reasoning tasks. With the improvements of our training scheme, it is feasible to integrate the findings of unsupervised object discovery methods into practical use-cases. A main motivation, as stated in our introduction, is that such an integration of high-quality object-centric representations is beneficial for more human-centric AI. Arguably, it seems beneficial for humans to perceive, communicate and explain the world on the level of objects. Integrating such level of abstraction and representation to AI agents is a necessary step for fruitful and reliable human-AI interactions.

Obviously, our work is not unaffected from the dual-use dilemma of foundational (AI) research. And a watchful eye should be kept, particularly on object detection research which can easily be misused, *e.g.* for involuntary human surveillance. However, our work or implications thereof do not, to the best of our knowledge, pose an obvious direct threat to any individuals or society in general.

## References

1. Anand, A., Racah, E., Ozair, S., Bengio, Y., Côté, M.-A., Hjelm, R.D.: Unsupervised state representation learning in atari. In: Wallach, H.M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E.B., Garnett, R. (eds.) Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019 (NeurIPS) (2019)
2. Anthwal, S., Ganotra, D.: An overview of optical flow-based approaches for motion segmentation. Imaging Sci. J. **67**(5), 284–294 (2019)
3. Bai, S., Geng, Z., Savani, Y., Kolter, J.Z.: Deep equilibrium optical flow estimation. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
4. Bao, Z., Tokmakov, P., Jabri, A., Wang, Y.-X., Gaidon, A., Hebert, M.: Discovering objects that can move. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
5. Brockman, G., et al.: OpenAI gym. CoRR (2016)
6. Burgess, C.P., et al.: MONet: unsupervised scene decomposition and representation. CoRR (2019)
7. Cai, Z., Neher, H., Vats, K., Clausi, D.A., Zelek, J.S.: Temporal hockey action recognition via pose and optical flows. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2019)

8. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.E.: A simple framework for contrastive learning of visual representations. In: Proceedings of the 37th International Conference on Machine Learning (ICML) (2020)
9. Crawford, E., Pineau, J.: Exploiting spatial invariance for scalable unsupervised object tracking. In: The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI) (2020)
10. Delfosse, Q., Blüml, J., Gregori, B., Sztwiertnia, S., Kersting, K.: OCAtari: object-centric atari 2600 reinforcement learning environments. CoRR (2021)
11. Delfosse, Q., Shindo, H., Dhami, D., Kersting, K.: Interpretable and explainable logical policies via neurally guided symbolic abstraction, CoRR (2023)
12. Dosovitskiy, A., et al.: Flownet: learning optical flow with convolutional networks. In: 2015 IEEE International Conference on Computer Vision (ICCV) (2015)
13. Du, Y., Smith, K., Ulman, T., Tenenbaum, J., Wu, J.: Unsupervised discovery of 3D physical objects from video. In: 9th International Conference on Learning Representations (ICLR) (2021)
14. Elsayed, G.F., Mahendran, A., van Steenkiste, S., Greff, K., Mozer, M.C., Kipf, T.: SAVi++: towards end-to-end object-centric learning from real-world videos. CoRR (2022)
15. Engelcke, M., Kosiorek, A.R., Jones, O.P., Posner, I.: GENESIS: generative scene inference and sampling with object-centric latent representations. In: 8th International Conference on Learning Representations (ICLR) (2020)
16. Eslami, S.M., Heess, N., Weber, T., Tassa, Y., Szepesvari, D., Hinton, G.E.: Attend, infer, repeat: fast scene understanding with generative models. In: Lee, D.D., Sugiyama, M., von Luxburg, U., Guyon, I., Garnett, R. (eds.) Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016 (NeurIPS) (2016)
17. Farnebäck, G.: Two-frame motion estimation based on polynomial expansion. In: Bigün, J., Gustavsson, T. (eds.) 13th Scandinavian Conference on Image Analysis 2003 (SCIA) (2003)
18. Fleet, D.J., Weiss, Y.: Optical flow estimation. In: Paragios, N., Chen, Y., Faugeras, O.D. (eds.) Handbook of Mathematical Models in Computer Vision (2006)
19. Goel, V., Weng, J., Poupart, P.: Unsupervised video object segmentation for deep reinforcement learning. In: Bengio, S., Wallach, H.M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018 (NeurIPS) (2018)
20. Greff, K., et al.: Multi-object representation learning with iterative variational inference. In: Chaudhuri, K., Salakhutdinov, R. (eds.) Proceedings of the 36th International Conference on Machine Learning (ICML) (2019)
21. Greff, K., van Steenkiste, S., Schmidhuber, J.: On the binding problem in artificial neural networks. CoRR (2020)
22. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.B.: Momentum contrast for unsupervised visual representation learning. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
23. Hénaff, O.J.: Data-efficient image recognition with contrastive predictive coding. In: Proceedings of the 37th International Conference on Machine Learning (ICML) (2020)
24. Hoerl, A.E., Kennard, R.W.: Ridge regression: applications to nonorthogonal problems. Technometrics **12**(1), 69–82 (1970)
25. Hoerl, A.E., Kennard, R.W.: Ridge regression: biased estimation for nonorthogonal problems. Technometrics **12**(1), 55–67 (2000)

26. Hur, J., Roth, S.: Optical flow estimation in the deep learning age. CoRR (2020)
27. Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., Brox, T.: Flownet 2.0: evolution of optical flow estimation with deep networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
28. Jeong, J., Lin, J.M., Porikli, F., Kwak, N.: Imposing consistency for optical flow estimation. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
29. Jiang, J., Ahn, S.: Generative neurosymbolic machines. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.-F., Lin, H.-T. (eds.) Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020 (NeurIPS) (2020)
30. Jiang, J., Janghorbani, S., De Melo, G., Ahn, S.: SCALOR: generative world models with scalable object representations. In: 8th International Conference on Learning Representations (ICLR) (2020)
31. Kale, K., Pawar, S., Dhulekar, P.: Moving object tracking using optical flow and motion vector estimation. In: 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions) (2015)
32. Kambhampati, S., Sreedharan, S., Verma, M., Zha, Y., Guan, L.: Symbols as a lingua franca for bridging human-AI chasm for explainable and advisable AI systems. In: Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI) (2022)
33. Kipf, T., et al.: Conditional object-centric learning from video. In: The Tenth International Conference on Learning Representations (ICLR) (2022)
34. Koh, P.W., et al.: Concept bottleneck models. In: Proceedings of the 37th International Conference on Machine Learning (ICML) (2020)
35. Kosiorek, A., Kim, H., Teh, Y.W., Posner, I.: Sequential attend, infer, repeat: generative modelling of moving objects. In: Bengio, S., Wallach, H.M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018 (NeurIPS) (2018)
36. Laskin, M., Srinivas, A., Abbeel, P.: CURL: contrastive unsupervised representations for reinforcement learning. In: Proceedings of the 37th International Conference on Machine Learning (ICML) (2020)
37. Lee, M., Lee, S., Son, S., Park, G., Kwak, N.: Motion feature network: fixed motion filter for action recognition. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) 15th European Conference on Computer Vision 2018 (ECCV) (2018)
38. Li, J., Zhao, Y., He, X., Zhu, X., Liu, J.: Dynamic warping network for semantic video segmentation. Complexity (2021)
39. Lin, Z., et al.: SPACE: unsupervised object-oriented scene representation via spatial attention and decomposition. In: 8th International Conference on Learning Representations (ICLR) (2020)
40. Liu, Y., Shen, C., Yu, C., Wang, J.: Efficient semantic video segmentation with per-frame inference. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12355, pp. 352–368. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58607-2_21
41. Locatello, F., et al.: Object-centric learning with slot attention. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.-F., Lin, H.-T. (eds.) Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, 6–12 December 2020, virtual (2020)
42. Luo, K., Wang, C., Liu, S., Fan, H., Wang, J., Sun, J.: Upflow: upsampling pyramid for unsupervised optical flow learning. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021)

43. Persson, A., Martires, P.Z.D., De Raedt, L., Loutfi, A.: Semantic relational object tracking. IEEE Trans. Cogn. Develop. Syst. **12**(1), 84–97 (2020)
44. Schulter, S., Leistner, C., Roth, P.M., Bischof, H.: Unsupervised object discovery and segmentation in videos. In: Burghardt, T., Damen, D., Mayol-Cuevas, W.W., Mirmehdi, M. (eds.) British Machine Vision Conference (BMVC) (2013)
45. Seitzer, M., et al.: Bridging the gap to real-world object-centric learning. CoRR (2022)
46. Singh, G., Deng, F., Ahn, S.: Illiterate DALL-E learns to compose. In: 10th International Conference on Learning Representations (ICLR) (2022)
47. Smirnov, D., Gharbi, M., Fisher, M., Guizilini, V., Efros, A., Solomon, J.M.: Marionette: self-supervised sprite learning. In: Ranzato, M., Beygelzimer, A., Dauphin, Y.N., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021 (NeurIPS) (2021)
48. Stammer, W., Schramowski, P., Kersting, K.: Right for the right concept: revising neuro-symbolic concepts by interacting with their explanations. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
49. Stelzner, K., Peharz, R., Kersting, K.: Faster attend-infer-repeat with tractable probabilistic models. In: Chaudhuri, K., Salakhutdinov, R. (eds.) Proceedings of the 36th International Conference on Machine Learning (ICML) (2019)
50. Stone, A., Maurer, D., Ayvaci, A., Angelova, A., Jonschkowski, R.: Smurf: self-teaching multi-frame unsupervised raft with full-image warping. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
51. Strickland, B., Wertz, A., Labouret, G., Keil, F., Izard, V.: The principles of object continuity and solidity in adult vision: some discrepancies in performance. J. Vis. **15**(12), 122 (2015)
52. Tangemann, M., et al.: Unsupervised object learning via common fate. CoRR (2021)
53. Teed, Z., Deng, J.: RAFT: recurrent all-pairs field transforms for optical flow. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12347, pp. 402–419. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58536-5_24
54. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. J. Mach. Learn. Res. (2008)
55. Van Hasselt, H., Guez, A., Silver, D.: Deep reinforcement learning with double q-learning. In: Schuurmans, D., Wellman, M.P. (eds.) Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (2016)
56. Wang, Z., Schaul, T., Hessel, M., van Hasselt, H., Lanctot, M., de Freitas, N.: Dueling network architectures for deep reinforcement learning. In: Balcan, M.-F., Weinberger, K.Q. (eds.) Proceedings of the 33nd International Conference on Machine Learning (ICML) (2016)
57. Wertheimer, M.: Untersuchungen zur lehre von der gestalt. ii. Psychologische forschung (1923)
58. Yang, C., Lamdouar, H., Lu, E., Zisserman, A., Xie, W.: Self-supervised video object segmentation by motion grouping. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV) (2021)
59. Yuezhang, L., Zhang, R., Ballard, D.H.: An initial attempt of combining visual selective attention with deep reinforcement learning. CoRR (2018)

60. Zhang, Y., Hare, J., Prugel-Bennett, A.: Deep set prediction networks. In Wallach, H.M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E.B., Garnett, R. (eds.) Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019 (NeurIPS) (2019)
61. Zhao, D., Ding, B., Yulin, W., Chen, L., Zhou, H.: Unsupervised learning from videos for object discovery in single images. Symmetry **13**(1), 38 (2021)
62. Zheng, Z., et al.: DIP: deep inverse patchmatch for high-resolution optical flow. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
63. Zhou, H., Ummenhofer, B., Brox, T.: Deeptam: deep tracking and mapping. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) 15th European Conference on Computer Vision 2018 (ECCV) (2018)
64. Zhou, T., Li, J., Li, X., Shao, L.: Target-aware object discovery and association for unsupervised video multi-object segmentation. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021)