Temporally Consistent Object-Centric Learning by Contrasting Slots

Anna Manasyan[1] Maximilian Seitzer[1] Filip Radovic[1] Georg Martius[1,3] Andrii Zadaianchuk[2]
[1]University of Tübingen, Tübingen, Germany [2]University of Amsterdam, Amsterdam, Netherlands
[3]Max Planck Institute for Intelligent Systems, Tübingen, Germany
anna.manasyan@student.uni-tuebingen.de a.zadaianchuk@uva.nl

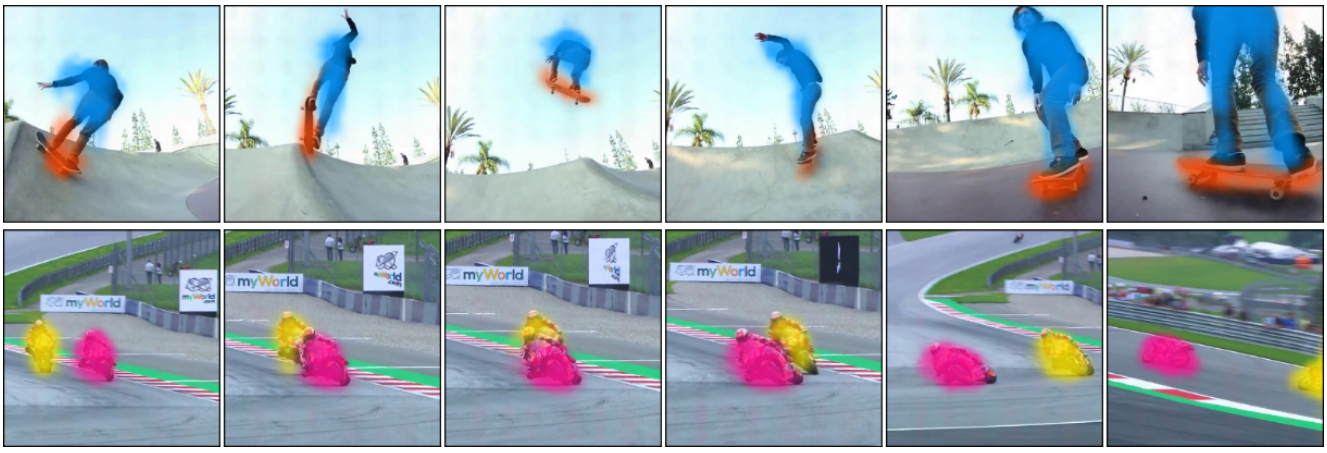

Figure 1. SLOT CONTRAST: Our method introduces a temporal contrastive loss that enhances temporal consistency in unsupervised video object-centric models. It stabilizes representations across frames, achieving state-of-the-art object discovery on complex real-world videos.

Abstract

Unsupervised object-centric learning from videos is a promising approach to extract structured representations from large, unlabeled collections of videos. To support downstream tasks like autonomous control, these representations must be both compositional and temporally consistent. Existing approaches based on recurrent processing often lack long-term stability across frames because their training objective does not enforce temporal consistency. In this work, we introduce a novel object-level temporal contrastive loss for video object-centric models that explicitly promotes temporal consistency. Our method significantly improves the temporal consistency of the learned object-centric representations, yielding more reliable video decompositions that facilitate challenging downstream tasks such as unsupervised object dynamics prediction. Furthermore, the inductive bias added by our loss strongly improves object discovery, leading to state-of-the-art results on both synthetic and real-world datasets, outperforming even weakly-supervised methods that leverage motion masks as additional cues.
Visit slotcontrast.github.io for videos and further details.

1. Introduction

Object-centric learning (OCL) [4, 14, 32, 44] is a rapidly advancing area of visual representation learning that enables autonomous systems to represent, understand, and model high-dimensional data directly in terms of its constituent entities. Structured object-centric representations (often referred to as *slots* [32]) facilitate generalization and robustness [7, 9] of scene representations across diverse downstream tasks, from visual question answering [1, 8, 34, 58] to control [10, 18, 61, 63]. Of particular interest are video-based object-centric methods [2, 11, 20, 28, 45, 57, 64] that learn to represent objects that evolve and interact over time. These representations make the methods powerful tools for applications such as unsupervised online object tracking [35, 53] and structured world modeling [27, 52, 56]. Unsupervised object-centric learning on videos has seen significant progress in recent years [2, 57, 64], mainly due to the use of pre-trained representations from self-supervised foundational models [5, 38] coupled with diverse training datasets like YouTube-VIS [59, 60]. Nevertheless, these methods still face significant challenges, especially maintaining consistent object-centric representations across time and uniquely representing each object—critical factors for

successful multi-object tracking and modeling of dynamic scenes [28, 52, 56].

Temporal consistency [15, 31, 62] in object-centric representations refers to maintaining the same representation placeholder, called slot, for an object throughout a video sequence, effectively serving as a stable object-specific identifier over time. Existing unsupervised object-centric methods [27, 36, 48] aiming to discover consistent representations have primarily been studied on toy datasets with limited complexity [22, 26, 46]. In contrast, real-world video sequences present numerous challenges, including object occlusions, reappearances, and complex multi-object interactions, which complicate maintaining consistent object representations.

In this paper, we introduce a novel method to address the challenge of maintaining consistent temporal representations in object-centric models, extending the line of research on slot-based unsupervised video models [11, 64]. Our approach (named SLOT CONTRAST) scales to real-world video data and produces consistent object-centric representations. Notably, it achieves these results without requiring any human annotations. In particular, we propose a novel self-supervised contrastive learning objective, which contrasts slot representations throughout the batch while ensuring temporal coherence across consecutive frames. In addition, we modify the slot's initialization strategy [32] to promote distinct, contrastive representations. This combination leads to improved temporal consistency of learned representations, which we show to be highly effective for challenging downstream tasks such as unsupervised object tracking and latent object dynamics learning.

Overall, our contributions are as follows:

- We propose the novel slot-slot contrastive loss that sets the state-of-the-art in temporal consistency when integrated into slot-based video processing methods.
- We develop SLOT CONTRAST, a simple and effective OCL architecture using the slot-slot contrastive loss paired with learned initialization that scales to real-world data, such as YouTube videos.
- We extensively study the usefulness of our learned object-centric representations for challenging downstream tasks, including unsupervised online tracking with complete occlusions and latent object dynamics modeling.
- We show that SLOT CONTRAST does not only improve the temporal consistency of the representations, but also achieves state-of-the-art on the object discovery task, outperforming weakly-supervised models using motion cues.

2. Related Work

Unsupervised video object-centric learning There exists an extensive body of research [2, 11, 14, 20, 23, 28, 30, 42, 45, 49, 51, 52, 64] on discovering objects from video without any human annotations, primarily through tracking either object bounding boxes or masks. To achieve this, most of these works combine an auto-encoder framework with a simple reconstruction objective, adding inductive biases for object discovery through structured encoders [4, 32] and decoders [55]. In particular, many modern object-centric image models [8, 21, 24, 44, 57] use a latent slot attention module [32] to extract object representations and corresponding object masks. For video data, most current methods [11, 28, 45, 49, 64, 66] connect slots across frames, with slots from the previous frame initializing those in the current frame. Notably, recent approaches [2, 40, 64] have successfully scaled object discovery to real-world unconstrained videos. To achieve this, SOLV [2] introduces temporal consistency via agglomerative clustering and prediction of middle-frame features, whereas VideoSAUR [64] learns object-centric representations by predicting temporal similarity of self-supervised features [5, 38]. While such methods can decompose short videos, they still struggle with long-term temporal consistency. In contrast, we show that learning representations that are both informative and contrastive can significantly enhance both object discovery and temporal consistency on longer videos.

Temporal Consistency Achieving temporal consistency is essential for any computer vision task involving video data, whether it is tracking points, bounding boxes, segmentation masks, optical flow, or representations [29, 33, 41, 47, 50, 52, 54]. In object-centric learning for videos, a range of different approaches have been proposed. For example, Yu and Xu [62] apply an object-wise sequential VAE to achieve consistency; Zhao et al. [65] and Li et al. [31] use an explicit memory buffer to maintain historical slot information and a transformer as a predictor using the memory buffer to predict the future; Qian et al. [40] achieve temporal consistency by employing student-teacher distillation to establish semantic and instance correspondence over time; and Traub et al. [48] use a recurrent network with a constancy prior [17].

3. Method

Our approach builds upon the existing input reconstruction-based video object-centric framework [28, 64] by introducing a consistency loss that contrasts the slots across consecutive frames and thereby adapting the model to discover consistent representations. See Fig. 2 for an overview of the SLOT CONTRAST architecture.

3.1. Semantic Recurrent Slot Attention Module

Our model is an encoder-decoder object-centric architecture based on Slot Attention module (SA) [13] with additional adaptations for sequential inputs similar to SAVi [28], while leveraging pre-trained semantic features as proposed by DINOSAUR [44]. The model consists of three main components: a pre-trained self-supervised dense feature encoder (e.g., DINOv2 [38]), a Recurrent Slot Attention module that

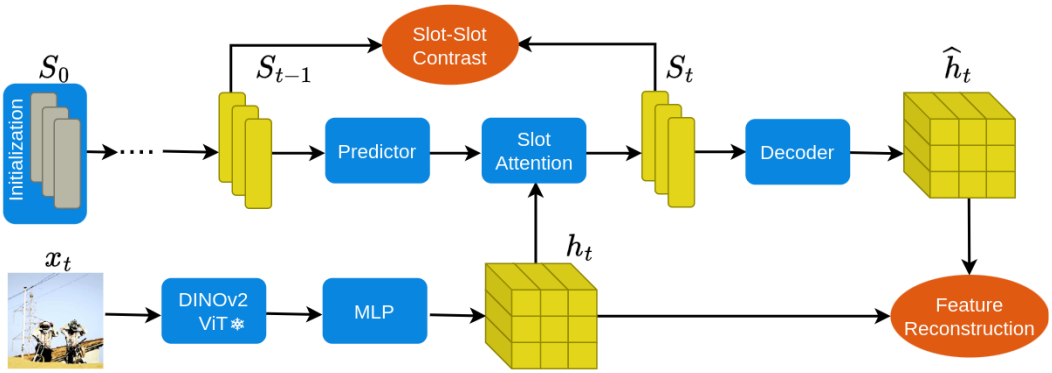

Figure 2. SLOT CONTRAST model architecture overview. For each frame, we extract patch features h_t using DINOv2 ViT. These features are then used to update the previously initialized or predicted slots, resulting in new slots S_t. The model is trained by contrasting the current frame's slots S_t with the slots from the previous frame S_{t-1}, and by reconstructing the patch features h_t.

groups the encoder features into slots and models temporal slot updates, and a decoder that maps slots from each frame to reconstructions of the dense self-supervised features used as inputs. Next, we describe those components in more detail while explaining how to adapt them to the task of consistent object-centric representation learning.

Given a video frame $x_t,\ t \in \{1, 2, \ldots, T\}$ and a pre-trained, frozen self-supervised DINO model f we first extract N patch features g_t,

$$g_t = f(x_t), \quad g_t \in \mathbb{R}^{N \times D}. \tag{1}$$

As those frozen features are mostly semantic and are trained only on images, we further adapt them to the task of temporally consistent object discovery . Specifically, each feature vector g_t is passed through a MLP g_ψ,

$$h_t = g_\psi(g_t), \tag{2}$$

to adapt the frozen dense features for object-centric grouping (see App. I for more details and visualizations). Based on the transformed encoder features h_t and a set of slot representations of the previous timestep S^p_{t-1}, with K slots $s^{k,p}_{t-1} \in S^p_{t-1}$, we use a recurrent grouping module to extract slot representations. The Recurrent Slot Attention module comprises a grouping module C_θ and a predictor module P_ω. The former updates slot representations using the standard Slot Attention module [32] on visual features h_t from the encoder, while the latter captures temporal and spatial interactions between slots:

$$S^c_t = \mathrm{C}_\theta(h_t, S^p_{t-1}), \quad S^p_t = \mathrm{P}_\omega(S^c_t). \tag{3}$$

Both slot-level representations, generated either by the grouping module S^c_t or the predictor S^p_t, can be utilized for subsequent decoding or downstream task processing. In our implementation, the slot-level representations from the grouping module S^c_t are employed for the decoding stage. From now on, we will refer to S^c_t as S_t.

Temporal Slot Attention Initialization Importantly, we found that our setup benefits considerably from a learned initialization S_0, which can influence the efficiency of training across various objectives. Originally, Locatello et al. [32] proposed a randomly sampled query initialization, where slots are sampled from the same Gaussian distribution with learned mean and variance. While such initialization allows different numbers of slots during inference, sampling from the same Gaussian distribution does not create a particularly favorable structure in slot-space. In this work, we use a straightforward learned initialization [19, 43] where a fixed set of initial slot vectors S_0 is learned for the entire dataset. Such initialization allows for learning dissimilar initialization queries that consistently attend to different objects.

Finally, for the reconstruction loss objective, we decode reconstructions $\hat{g}_t$ from all slots using the MLP decoder [44].

3.2. Temporal Consistency through Slot Contrast

Contrastive learning is flexible in supporting diverse data sources and loss function designs. By carefully defining positive and negative examples, we can craft robust loss objectives that effectively guide self-supervised representation learning [6]. For instance, video contrastive methods like CVRL [39] leverage augmented video chunks to define positive (from the same video) and negative (from different videos) examples. In object-centric learning Didolkar et al. [8] employed a contrastive loss function to gain controllability over slot representations guided by language. We propose a novel application of a contrastive loss for temporal consistency in object-centric slot representations. In particular, we define positive samples as the representations of the same slot from two consecutive time steps within a video, while negative samples comprise all other slots across the batch between these time steps. An overview of the proposed loss is presented in Fig. 3

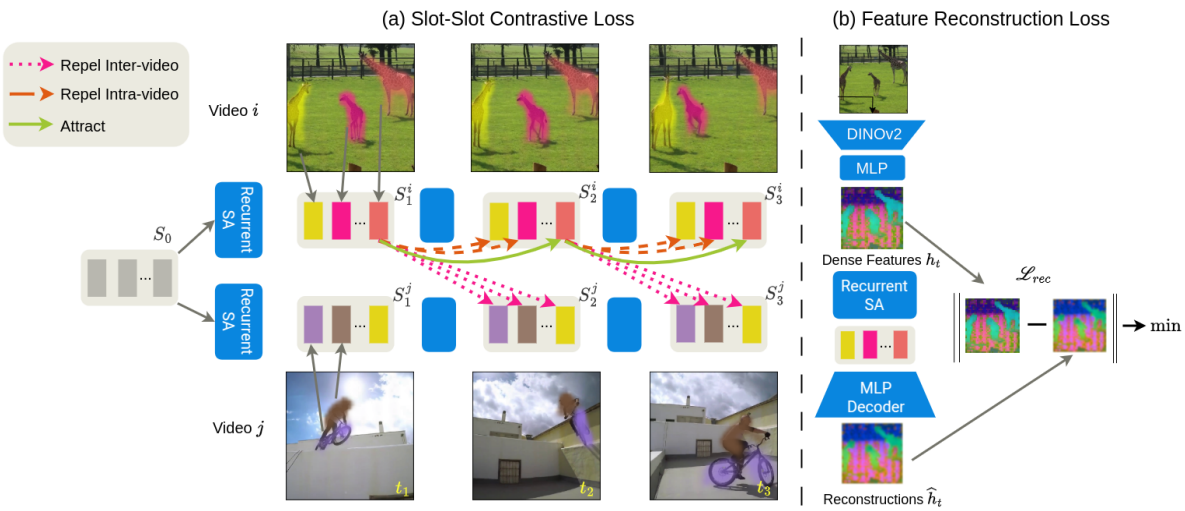

Figure 3. Overview of the losses used in SLOT CONTRAST. (a) Our proposed temporal consistency objective, *slot-slot contrastive loss*, operates on a batch of video sequences by enforcing temporal alignment across object slots. For each frame in the sequence, the model groups object features into specific slot representations S_t^i. The slot-slot contrastive loss then enforces temporal consistency by drawing the corresponding slot representations from adjacent frames closer, while simultaneously pushing apart all other slot representations in the batch—whether they come from different objects within the same video or from objects in other videos. (b) The feature reconstruction loss ensures informativeness of the learned slots by using them to reconstruct original DINOv2 features with an MLP decoder.

Intra-Video Slot-Slot Contrastive Loss To force each slot to be consistent in time, we aim to learn slots that are similar in time while being maximally dissimilar to other slots. Given the sets of slot representations S_{t-1} and S_t at time steps $t-1$ and t, we want elements $s_{t-1}^i \in S_{t-1}$ to be close to the next-frame slots s_t^i corresponding to the same object, while having maximal distance to the next-frame slots $s_t^k, k \neq i$ corresponding to other objects in the video. The corresponding InfoNCE contrastive loss [37] is defined as $\mathcal{L}_{\text{intra}} = \frac{1}{K}\sum_{i=1}^{K} \ell_i^{\text{intra}}$ with

$$\ell_i^{\text{intra}} = -\log \frac{\exp(\text{sim}(s_{t-1}^i, s_t^i)/\tau)}{\sum_{k=1}^{K} \mathbb{1}_{[k\neq i]} \exp(\text{sim}(s_{t-1}^i, s_t^k)/\tau)}, \quad (4)$$

where $K = |S_t|$ is a number of slots per frame, $\text{sim}(u, v) = \frac{u^\top v}{\|u\|_2 \|v\|_2}$ is the cosine similarity, $\mathbb{1}_{[.]}$ is an indicator excluding the self-similarity of the slot s_i from the denominator, and $\tau > 0$ is a temperature parameter.

While being a desirable property, intra-video slot contrast can be achieved simply by amplifying the differences between slots in the SA module's first frame initialization S_0. To encourage a stronger focus on video content and instance specificity of the representations, we propose a further improvement over this loss by extending the negative contrast set.

Batch Video Slot-Slot Contrastive Loss To leverage the benefits of larger contrast sets and prevent degenerate solutions relying solely on the initialization of slots, we exploit the fact that the whole batch of videos can be considered a large set of primarily unique object representations. Consequently, we enhance contrast within a video and between videos by including negative slots from the current and subsequent frames of all videos in the batch. Correspondingly, we define our slot-slot contrastive loss as $\mathcal{L}_{\text{ssc}} = \frac{1}{B \cdot K}\sum_{j=1}^{B}\sum_{i=1}^{K} \ell_{i,j}^{\text{ssc}}$ and

$$\ell_{i,j}^{\text{ssc}} = -\log \frac{\exp(\text{sim}(s_{t-1}^{i,j}, s_t^{i,j})/\tau)}{\sum_{b=1}^{B}\sum_{k=1}^{K} \mathbb{1}_{[k,b\neq i,j]} \exp(\text{sim}(s_{t-1}^{i,j}, s_t^{k,b})/\tau)}, \quad (5)$$

where B is a number of videos in the batch and $s_t^{i,j}$ denotes the i-th slot of the j-th video at time t. For more details on slot-slot contrastive loss implementation, see App. C.

We find that this approach significantly enhances the effectiveness of the slot-slot contrastive loss. Furthermore, since all videos in the batch are processed with the same initial state S_0, this loss function avoids suboptimal solutions that rely solely on the uniqueness of the initialization, instead encouraging object discovery as the basis for contrast.

Final Loss To encourage scene decomposition we use a feature reconstruction loss, similar to DINOSAUR [44] and VideoSAUR [64]. Our final loss function combines the reconstruction loss with our proposed contrastive loss $\mathcal{L}_{\text{ssc}}$, weighted by the hyperparameter α (see Table S1 for details on how the hyperparameters are set):

$$\mathcal{L} = \sum_{t=1}^{T-1} \mathcal{L}_{\text{rec}}(\text{h}_\text{t}, \hat{\text{h}}_\text{t}) + \alpha \mathcal{L}_{\text{ssc}}(S_{t-1}, S_t). \quad (6)$$

Table 1. Consistent object-discovery performance of SLOT CONTRAST in comparison with SAVi, STEVE, VideoSAUR on MOVi-C, MOVi-E, and YouTube-VIS datasets. VideoSAURv2 is an improved version of the VideoSAUR trained on DINOv2 features. Both metrics are computed for the whole video (24 frames for MOVi, up to 76 frames for YouTube-VIS).

	MOVi-C		MOVi-E		YouTube-VIS	
	FG-ARI ↑	mBO ↑	FG-ARI ↑	mBO ↑	FG-ARI ↑	mBO ↑
SAVi [28]	22.2	13.6	42.8	16.0	-	-
STEVE [45]	36.1	26.5	50.6	26.6	15	19.1
VideoSAUR [64]	64.8	**38.9**	73.9	**35.6**	28.9	26.3
VideoSAURv2	-	-	77.1	34.4	31.2	29.7
SLOT CONTRAST	**69.3**	32.7	**82.9**	29.2	**38.0**	**33.7**

4. Experiments

We evaluate our method's temporal consistency on two downstream tasks: object discovery and latent object dynamics prediction. Our experiments address three main questions: (1) How does our model compare to state-of-the-art methods in both temporal consistency and scene decomposition? (2) How effective are our model's learned representations for the challenging downstream task of object dynamics prediction and for object tracking under full occlusions? (3) How important are the different components of our model and loss function for temporal consistency?

Datasets To evaluate our method in the controlled setting, we use MOVi-C and MOVi-E synthetic datasets generated by Kubric [16]. MOVi-C includes richly textured everyday objects, featuring up to 11 objects per scene, while MOVi-E expands this to 23 objects and introduces basic linear camera motion. In addition, to study the scalability of our method to real-world data, we evaluate our method on the real-world YouTube-VIS 2021 (YTVIS21) video dataset [60]. YTVIS21 is an unconstrained, real-world dataset sourced from YouTube, capturing a diverse range of scenes (for more details, see App. E).

Metrics Similar to other object-centric video methods [11, 28, 45, 64], to evaluate consistent object discovery, we use the video foreground adjusted rand index (FG-ARI) [14], measuring how well objects are split. In addition, we evaluate the sharpness of masks using the video intersection over union with mean best overlap matching (mBO) metric [44, 64]. Both metrics are computed *over the full video* and thus reflect how consistent object discovery is. In addition, to investigate the effects of the temporal consistency inductive bias on the per-frame object discovery itself, we use per-frame FG-ARI (image FG-ARI), which we independently compute for each frame and average afterwards. More details can be found in App. F.

Finally, when evaluating how well object-centric representations perform for object dynamics prediction (see Sec. 4.2), we employ the same evaluation metrics as in the object discovery task: FG-ARI and mBO. This time, however, these metrics are computed by comparing the predicted masks (obtained by decoding the predicted slots [56]) with the ground-truth future masks.

4.1. Object Discovery

Implementation Details We employ the DINOv2 model as our feature encoder, using ViT-S/14 for the MOVi-C dataset and ViT-B/14 for MOVi-E and YTVIS21. The slot dimension is set to 128 for MOVi-E and 64 for both MOVi-C and YTVIS21. For the MOVi datasets, we use a resolution of $(336, 336)$, generating 24×24 patches yielding 576 ViT tokens, while for YTVIS21, a resolution of $(518, 518)$ yields 1369 tokens. Full details are provided in App. A.

Baselines We compare SLOT CONTRAST against the previously proposed SAVi [28] and STEVE [45] that employ an image reconstruction objective and with the state-of-the-art method VideoSAUR [64] that uses self-supervised feature reconstructions. Additionally, for a fair comparison, we trained a modification of VideoSAUR with DINOv2 features (referred to as VideoSAURv2). The implementation details are provided in App. G. In addition, to assess how closely SLOT CONTRAST approaches supervised methods, we compared it with SAM2 [41] as a supervised zero-shot baseline for temporal consistency and to weakly-supervised by depth SAVi++ [11] method. The results are in App. G.

Temporally Consistent Object Discovery (Table 1 & Figure 4) SLOT CONTRAST significantly outperforms both SAVi and STEVE by a wide margin. When compared to VideoSAUR using its default parameters, our approach demonstrates higher consistency in terms of video FG-ARI scores. Compared to VideoSAUR and VideoSAURv2, SLOT CONTRAST achieves superior video scene decomposition (measured by FG-ARI scores). However, on synthetic datasets SLOT CONTRAST's masks are less sharp (as reflected by mBO). Notably, on the most challenging real-world YouTube-VIS data, our method surpasses both versions of VideoSAUR, achieving better performance on FG-ARI (+6.8) and mBO (+4). This shows that, given a large enough resolution and natural data inputs well aligned with DINOv2, SLOT CONTRAST can decompose unconstrained videos into consistent object representations. More examples are illustrated in App. M.

Per-Frame Scene Decomposition (Table 2) Previous research has shown the effectiveness of specific inductive biases and training objectives for unsupervised object discovery, such as reconstructing in semantic space or leveraging motion cues in video data. Building on this, we demonstrate that the contrastive nature of our temporal consistency loss

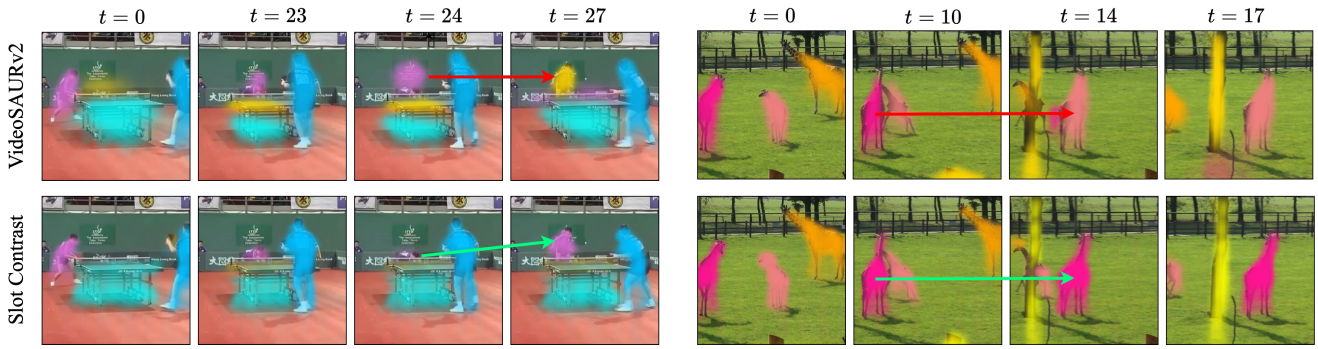

Figure 4. Qualitative comparison with VideoSAURv2 on YouTube-VIS dataset. In challenging situations (e.g., almost full occlusions at $t = 24$ of the 1st video and $t = 14$ of the 2nd video), VideoSAURv2 reassigns slots to different objects (pink arrows), whereas SLOT CONTRAST consistently assigns slots to the same object (green arrows). Note that the colors of the masks are matched manually for better visual comparison.

Table 2. Quantitative Results on MOVi-E in terms of per-frame Image FG-ARI. The methods are grouped by the target data they train on: only images ($\mathcal{I}$), videos with motion segmentation annotations ($\mathcal{V} + \mathcal{M}$), and only videos ($\mathcal{V}$).

	Model	Objective	Image FG-ARI ↑
$\mathcal{I}$	LSD [21]	Image Rec.	53.4
	SlotDiffusion [57]	Image Rec.	60.0
	DINOSAUR [44]	Image Rec.	65.1
$\mathcal{V} + \mathcal{M}$	MoToK [3]	+Mot. Seg.	66.7
	Safadoust et al. [42]	+GT Flow	78.3
	DIOD [25]	+Mot. Seg.	82.2
$\mathcal{V}$	STEVE [45]	Video Rec.	54.1
	VideoSAUR [64]	Temp. Sim.	78.4
	SOLV [2]	Mid. Fr. Pred.	80.8
	SLOT CONTRAST	Slot Contrast	**84.8**

function yields improved scene decomposition as a byproduct. This occurs because our loss function encourages the model to learn consistent feature representations for objects across frames, leading to an adaptive process where dense features become more contrastive, thereby enhancing object discovery in individual frames.

We compare our method with prior approaches in terms of per-frame object discovery, using the image FG-ARI metric for evaluation. Specifically, we compare three categories of methods: image-based, video-based, and methods that use videos with additional motion cues. Image-based methods use only images as a target (feature reconstruction based DINOSAUR [44] and diffusion-based LSD [21] and SlotDiffusion [57] methods). Video-based methods use only videos as targets: STEVE [45] reconstructs current frame features, SOLV [2] predicts middle frame features, and VideoSAUR [64] predicts temporal feature-similarities. Finally, we also compare with weakly-supervised methods using motion masks [3, 25] or ground truth (GT) optical flow [42].

The results on MOVi-E dataset are presented in Table 2, with comparisons across additional datasets provided in App. H. Using temporal signals from the video using feature reconstruction is better than object discovery based on images. Next, additional objectives that exploit the temporal structure of the videos allow even better scene decomposition. Notably, our method, which combines a feature reconstruction objective with a simple contrastive objective, leads to state-of-the-art performance reaching 84.8 per-frame FG-ARI, outperforming methods [3, 25] that use motion segmentation masks for object discovery.

Robustness to Full Occlusions To evaluate our method's robustness in handling complete object occlusions—a challenging scenario for maintaining consistency—we conduct experiments using a targeted subset of the MOVi-C dataset that contains sequences where objects are fully occluded. For evaluation, we retain only the ground-truth masks for the objects that experience occlusion.

We find that the feature reconstruction baseline achieves only 16% mBO vs. our method obtains 21% mBO on fully occluded objects. Our results suggest that SLOT CONTRAST significantly enhances consistency during object disappearances and reappearances. We refer the reader to Fig. 5 and App. K for visual examples and more details.

4.2. Object Dynamics Prediction

Setup To evaluate performance on the task of predicting object dynamics, we train a dynamics module using the object-centric representations inferred by a pretrained object-centric model. For this dynamics module, we select SlotFormer [56], which predicts the slots autoregressively for K rollout steps based on the slots inferred from T burn-in frames preceding the prediction horizon. In our setup, we use 14 burn-in frames and 10 rollout steps. Since SlotFormer

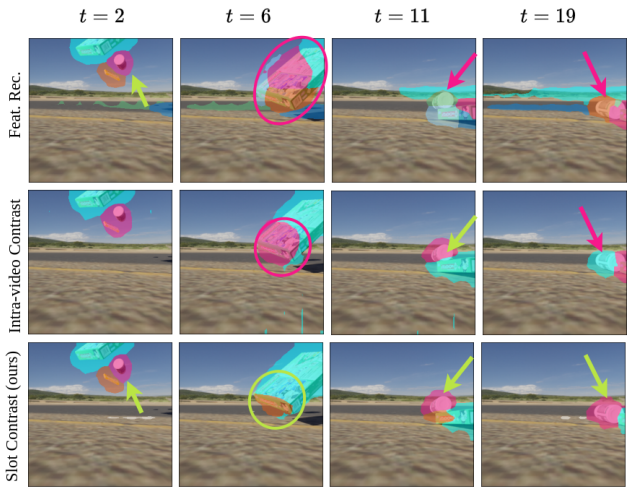

Figure 5. Comparison of the Feature Reconstruction (Feat. Rec.) baseline, the slot-slot contrastive loss using only slots from the same video as the contrast set (Intra-video Contrast), and SLOT CONTRAST on the MOVi-C dataset.

is trained independently from the object-centric model, we first train the latter, subsequently extending the datasets with the inferred slots for each frame. This approach avoids the computational complexity of training SlotFormer by removing the necessity to encode frames into the slot space at each training step. A brief introduction to SlotFormer and the implementation details can be found in App. J.

Baselines We compare SlotFormer [56] trained on object-centric representations derived from a model trained using only feature reconstruction loss with SlotFormer trained on representations from SLOT CONTRAST. Both models perform reconstruction in feature space rather than pixel space, so we use only the slot reconstruction loss for training.

Quality of Predicted Masks (Table 3 and Figure 6) We note that our model has a significantly better FG-ARI on MOVi-C, while mBO is comparable to that of the baseline. On MOVi-E, the performance of our model is comparable to that of the baseline, which highlights the difficulty of adapting to videos with camera motion. There is also a

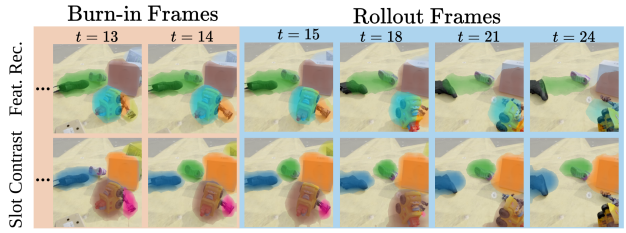

Figure 6. Object dynamics prediction task on MOVi-C using SLOT CONTRAST slots using SlotFormer [56].

Table 3. Downstream task of predicting object dynamics. Comparison of predictions made by SlotFormer based on representations obtained from SLOT CONTRAST and from Feature Reconstruction.

	MOVi-C		MOVi-E		YouTube-VIS	
	FG-ARI ↑	mBO ↑	FG-ARI ↑	mBO ↑	FG-ARI ↑	mBO ↑
Feat. Rec. + SF	50.7	25.9	70.6	24.3	27.4	28.9
Ours + SF	63.8	26.1	70.5	24.9	29.2	29.6

slight improvement in FG-ARI on YTVIS21, while mBO remains comparable. It is worth noting that predicting the motion in this dataset is especially challenging, given the large diversity of possible scenarios.

4.3. Analysis

In this section, we investigate key components of our approach, including the impact of the contrastive loss and the type of slot initialization. In addition, we study how effective SLOT CONTRAST is in automatically shutting down slots in correspondence to the scene's complexity.

Ablation of Loss Components (Table 4 and Figure 5) To demonstrate the value of the proposed slot-slot contrastive loss, we carry out an ablation study, comparing it with the feature reconstruction loss [44] and the intra-video contrastive loss, which contrasts slot representations in a single video. Using the intra-video contrastive loss yields improvements over the feature reconstruction baseline (+5.1 FG-ARI and +1.5 mBO on MOVi-C). However, we observe that in more challenging situations, the intra-video contrastive loss leads to failure cases such as shutting down too many slots (see Fig. 5). Next, we observe that by extending the contrast to the full batch of videos, SLOT CONTRAST learns more consistent representations (+19.6 FG-ARI and +5.3 mBO). This change increases the difficulty of the learning task, which prevents the model from relying on superficial patterns like slot initializations or object positions.

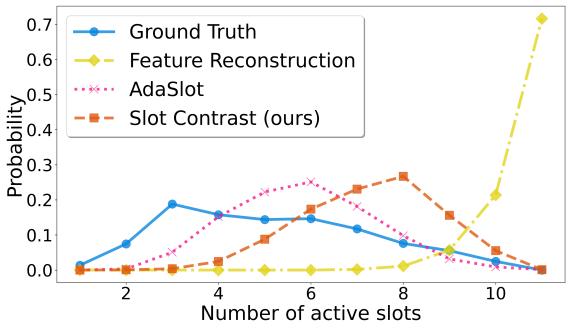

Figure 7. Distribution of ground truth and predicted object numbers (i.e., number of active slots) on the MOVi-C dataset.

Table 4. Ablation of loss components used by SLOT CONTRAST on MOVi-C, MOVi-E, and YouTube-VIS Datasets.

Feat. Rec.	Intra-video Contrast	Slot-Slot Contrast	MOVi-C		MOVi-E		YouTube-VIS	
$\mathcal{L}_{\text{rec}}$	$\mathcal{L}_{\text{intra}}$	$\mathcal{L}_{\text{ssc}}$	FG-ARI ↑	mBO ↑	FG-ARI ↑	mBO ↑	FG-ARI ↑	mBO ↑
✓			49.7	27.4	79.8	28.4	35.3	31.4
✓	✓		54.8	28.9	78.7	29.1	35.7	33.6
✓		✓	69.3	32.7	82.9	29.2	38.0	33.7

Table 5. Comparison of the random initialization (RI) and learned initialization (LI) techniques.

	MOVi-C		MOVi-E		YouTube-VIS	
	FG-ARI ↑	mBO ↑	FG-ARI ↑	mBO ↑	FG-ARI ↑	mBO ↑
Feat. Rec. (RI)	45.3	27.2	71.1	28.3	35.2	30.2
Feat. Rec. (LI)	49.4	27.8	79.8	28.4	35.3	31.4
SLOT CONTRAST (RI)	62.9	32.4	75.3	28.4	36.1	30.8
SLOT CONTRAST (LI)	69.3	32.7	82.9	29.2	38.0	33.7

Choice of the First Frame Initialization (Table 5) In video object-centric learning, slots are typically initialized based on those from the previous time step [28], while the first frame is initialized from learnable parameters. Previous real-world object-centric methods mostly used random initialization samples from Gaussian distribution [2, 64], in this work, we study the impact of the type of initialization under our contrastive objective. The findings are outlined in Table 5. Our experiments indicate that when combined with slot-slot contrastive loss, learned initialization significantly outperforms random initialization. We hypothesize that this improvement stems from the ability of learned initializations to shape the initial state in a way that enhances contrastiveness, a benefit not achievable with random initialization. For similarity visualizations and more details, see App. B.

Number of Active Slots per Video (Figure 7) In models based on the Slot Attention mechanism, all available slots are typically utilized [44], leading to a mismatch between the predicted number of components in scene decomposition and the ground-truth number of objects in the scene. This can cause the random splitting of the objects between slots and non-consistent scene representations when slots are reassigned from one object to a part of another object. To address this challenge, it is important to study whether redundant slots can be effectively deactivated. Recently, AdaSlot [12] introduced a discrete slot sampling module, coupled with a complexity-aware prior, to penalize redundant slots explicitly. Similarly, SOLV [2] used agglomerative clustering to merge redundant slots. In this work, we investigate whether SLOT CONTRAST is capable of accurately determining the number of objects in a scene without relying on explicit priors to minimize the number of active slots.

We compare ground truth and predicted object density on MOVI-C dataset, as shown in Fig. 7. While the feature reconstruction model yields predictions within a narrow range—creating a sharp peak near a predefined number of slots—our model, similarly to AdaSlot [12], achieves a smoother prediction distribution that aligns more closely with the ground truth (note that the consistent shift is because 2–3 slots are used for the background, while the ground truth density is computed only for foreground objects). Interestingly, SLOT CONTRAST achieves this without requiring an explicit prior toward sparsity.

5. Conclusion

SLOT CONTRAST advances unsupervised video object-centric learning by significantly improving the temporal consistency of object representations. Our method explicitly incentivizes temporal consistency by adding a self-supervised contrastive loss. We showed that this loss is not only beneficial for consistency, but also enhances object discovery: SLOT CONTRAST achieves state-of-the-art results on challenging synthetic datasets with many objects and the unconstrained real-world YouTube-VIS dataset. Furthermore, consistent representations directly support temporal downstream tasks such as unsupervised object dynamics prediction and allow for tracking of objects through full occlusions. Finally, SLOT CONTRAST effectively shuts down non-unique slots, leading to a sparser representation that captures the true object distribution more faithfully. Taken together, we expect these improvements to pave the way for broader adoption of video object-centric representations, for instance in applications like word modeling, autonomous control, or video question answering.

Limitations of our work include the fixed number of slots during initialization. Additionally, we cannot directly control the segmentation granularity of entities. Further limitations and failure cases are discussed in App. L.

Future work could explore several promising directions. First, one could use SLOT CONTRAST's robust and consistent representations for learning compositional world models from real-world robotics data to enable object-centric planning and control. Second, investigating the compatibility of our contrastive loss with other object-centric learning approaches with different inductive biases, such as SlotDiffusion [57]. Finally, improving the compactness of object masks [24] to achieve more precise object segmentation masks could also benefit unsupervised class-agnostic video object segmentation applications.

Acknowledgments

We thank Christian Gumbsch for useful suggestions and Ke Fan for providing additional details and results for AdaSlot method. Andrii Zadaianchuk is funded by the European Union (ERC, EVA, 950086). The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Maximilian Seitzer. Georg Martius is a member of the Machine Learning Cluster of Excellence, EXC number 2064/1 – Project number 390727645. This work was supported by the ERC - 101045454 REAL-RL. We acknowledge the support from the German Federal Ministry of Education and Research (BMBF) through the Tübingen AI Center (FKZ: 01IS18039B).

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *International Conference on Computer Vision (ICCV)*, 2015. 1

[2] Görkay Aydemir, Weidi Xie, and Fatma Güney. Self-supervised object-centric learning for videos. In *NeurIPS*, 2023. 1, 2, 6, 8

[3] Zhipeng Bao, Pavel Tokmakov, Yu-Xiong Wang, Adrien Gaidon, and Martial Hebert. Object discovery from motion-guided tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22972–22981, 2023. 6

[4] Christopher P. Burgess, Loic Matthey, Nicholas Watters, Rishabh Kabra, Irina Higgins, Matt Botvinick, and Alexander Lerchner. MONet: Unsupervised Scene Decomposition and Representation. *arXiv:1901.11390*, 2019. 1, 2

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging Properties in Self-Supervised Vision Transformers. *ICCV*, 2021. 1, 2

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations. In *ICML*, 2020. 3

[7] Aniket Didolkar, Andrii Zadaianchuk, Anirudh Goyal, Mike Mozer, Yoshua Bengio, Georg Martius, and Maximilian Seitzer. Zero-shot object-centric representation learning. *arXiv preprint arXiv:2408.09162*, 2024. 1

[8] Aniket Rajiv Didolkar, Andrii Zadaianchuk, Rabiul Awal, Maximilian Seitzer, Efstratios Gavves, and Aishwarya Agrawal. Ctrl-o: Language-controllable object-centric visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 1, 2, 3

[9] Andrea Dittadi, Samuele Papa, Michele De Vita, Bernhard Schölkopf, Ole Winther, and Francesco Locatello. Generalization and Robustness Implications in Object-Centric Learning. In *ICML*, 2022. 1

[10] Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. PaLM-E: An embodied multimodal language model. In *ICML*, 2023. 1

[11] Gamaleldin Fathy Elsayed, Aravindh Mahendran, Sjoerd van Steenkiste, Klaus Greff, Michael Curtis Mozer, and Thomas Kipf. SAVi++: Towards End-to-End Object-Centric Learning from Real-World Videos. In *NeurIPS*, 2022. 1, 2, 5

[12] Ke Fan, Zechen Bai, Tianjun Xiao, Tong He, Max Horn, Yanwei Fu, Francesco Locatello, and Zheng Zhang. Adaptive slot attention: Object discovery with dynamic slot number. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23062–23071, 2024. 8

[13] Locatello Francesco, Weissenborn Dirk, Unterthiner Thomas, Mahendran Aravindh, Heigold Georg, Uszkoreit Jakob, Dosovitskiy Alexey, and Kipf Thomas. Object-centric learning with slot attention. *Advances in Neural Information Processing Systems*, 33:11525–11538, 2020. 2

[14] Klaus Greff, Raphaël Lopez Kaufman, Rishabh Kabra, Nick Watters, Christopher Burgess, Daniel Zoran, Loic Matthey, Matthew Botvinick, and Alexander Lerchner. Multi-Object Representation Learning with Iterative Variational Inference. In *ICML*, 2019. 1, 2, 5

[15] Klaus Greff, Sjoerd Van Steenkiste, and Jürgen Schmidhuber. On the Binding Problem in Artificial Neural Networks. *arXiv:2012.05208*, 2020. 2

[16] Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J. Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, Thomas Kipf, Abhijit Kundu, Dmitry Lagun, Issam Laradji, Hsueh-Ti (Derek) Liu, Henning Meyer, Yishu Miao, Derek Nowrouzezahrai, Cengiz Oztireli, Etienne Pot, Noha Radwan, Daniel Rebain, Sara Sabour, Mehdi S. M. Sajjadi, Matan Sela, Vincent Sitzmann, Austin Stone, Deqing Sun, Suhani Vora, Ziyu Wang, Tianhao Wu, Kwang Moo Yi, Fangcheng Zhong, and Andrea Tagliasacchi. Kubric: A Scalable Dataset Generator. In *CVPR*, 2022. 5

[17] Christian Gumbsch, Martin V Butz, and Georg Martius. Sparsely changing latent states for prediction and planning in partially observable domains. In *NeurIPS*, 2021. 2

[18] Dan Haramati, Tal Daniel, and Aviv Tamar. Entity-centric reinforcement learning for object manipulation from pixels. In *ICLR*, 2024. 1

[19] Baoxiong Jia, Yu Liu, and Siyuan Huang. Improving object-centric learning with query optimization. *arXiv preprint arXiv:2210.08990*, 2022. 3

[20] Jindong Jiang, Sepehr Janghorbani, Gerard de Melo, and Sungjin Ahn. SCALOR: Generative World Models with Scalable Object Representations. In *ICLR*, 2020. 1, 2

[21] Jindong Jiang, Fei Deng, Gautam Singh, and Sungjin Ahn. Object-centric slot diffusion. In *NeurIPS*, 2023. 2, 6

[22] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017. 2

[23] Rishabh Kabra, Daniel Zoran, Goker Erdogan, Loic Matthey, Antonia Creswell, Matthew Botvinick, Alexander Lerchner, and Christopher P. Burgess. SIMONe: View-invariant, temporally-abstracted object representations via unsupervised video decomposition. In *NeurIPS*, 2021. 2

[24] Ioannis Kakogeorgiou, Spyros Gidaris, Konstantinos Karantzalos, and Nikos Komodakis. Spot: Self-training with patch-order permutation for object-centric learning with autoregressive transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22776–22786, 2024. 2, 8

[25] Sandra Kara, Hejer Ammar, Julien Denize, Florian Chabot, and Quoc-Cuong Pham. Diod: Self-distillation meets object discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3975–3985, 2024. 6

[26] Laurynas Karazija, Iro Laina, and Christian Rupprecht. ClevrTex: A Texture-Rich Benchmark for Unsupervised Multi-Object Segmentation. In *NeurIPS Track on Datasets and Benchmarks*, 2021. 2

[27] Thomas Kipf, Elise Van der Pol, and Max Welling. Contrastive learning of structured world models. *arXiv preprint arXiv:1911.12247*, 2019. 1, 2

[28] Thomas Kipf, Gamaleldin Fathy Elsayed, Aravindh Mahendran, Austin Stone, Sara Sabour, Georg Heigold, Rico Jonschkowski, Alexey Dosovitskiy, and Klaus Greff. Conditional Object-centric Learning from Video. In *ICLR*, 2022. 1, 2, 5, 8

[29] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 2

[30] Adam Kosiorek, Hyunjik Kim, Yee Whye Teh, and Ingmar Posner. Sequential Attend, Infer, Repeat: Generative Modelling of Moving Objects. In *NeurIPS*, 2018. 2

[31] Jian Li, Pu Ren, Yang Liu, and Hao Sun. Reasoning-enhanced object-centric learning for videos. *arXiv preprint arXiv:2403.15245*, 2024. 2

[32] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-Centric Learning with Slot Attention. In *NeurIPS*, 2020. 1, 2, 3

[33] Zijia Lu, Bing Shuai, Yanbei Chen, Zhenlin Xu, and Davide Modolo. Self-supervised multi-object tracking with path consistency. *arXiv preprint*, 2024. 2

[34] Amir Mohammad Karimi Mamaghan, Samuele Papa, Karl Henrik Johansson, Stefan Bauer, and Andrea Dittadi. Exploring the effectiveness of object-centric representations in visual question answering: Comparative insights with foundation models. *arXiv preprint arXiv:2407.15589*, 2024. 1

[35] Sha Meng, Dian Shao, Jiacheng Guo, and Shan Gao. Tracking without label: Unsupervised multiple object tracking via contrastive similarity learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16264–16273, 2023. 1

[36] Cristian Meo, Akihiro Nakano, Mircea Lică, Aniket Didolkar, Masahiro Suzuki, Anirudh Goyal, Mengmi Zhang, Justin Dauwels, Yutaka Matsuo, and Yoshua Bengio. Object-centric temporal consistency via conditional autoregressive inductive biases. *arXiv preprint arXiv:2410.15728*, 2024. 2

[37] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 4

[38] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision, 2023. 1, 2

[39] Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6964–6974, 2021. 3

[40] Rui Qian, Shuangrui Ding, Xian Liu, and Dahua Lin. Semantics meets temporal correspondence: Self-supervised object-centric learning in videos. In *ICCV*, 2023. 2

[41] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 2, 5

[42] Sadra Safadoust and Fatma Güney. Multi-object discovery by low-dimensional object motion. In *ICCV*, 2023. 2, 6

[43] Mehdi SM Sajjadi, Daniel Duckworth, Aravindh Mahendran, Sjoerd Van Steenkiste, Filip Pavetic, Mario Lucic, Leonidas J Guibas, Klaus Greff, and Thomas Kipf. Object scene representation transformer. *Advances in Neural Information Processing Systems*, 35:9512–9524, 2022. 3

[44] Maximilian Seitzer, Max Horn, Andrii Zadaianchuk, Dominik Zietlow, Tianjun Xiao, Carl-Johann Simon-Gabriel, Tong He, Zheng Zhang, Bernhard Schölkopf, Thomas Brox, and Francesco Locatello. Bridging the gap to real-world object-centric learning. In *ICLR*, 2023. 1, 2, 3, 4, 5, 6, 7, 8

[45] Gautam Singh, Yi-Fu Wu, and Sungjin Ahn. Simple Unsupervised Object-Centric Learning for Complex and Naturalistic Videos. In *NeurIPS*, 2022. 1, 2, 5, 6

[46] Matthias Tangemann, Steffen Schneider, Julius Von Kügelgen, Francesco Locatello, Peter Gehler, Thomas Brox, Matthias Kümmerer, Matthias Bethge, and Bernhard Schölkopf. Unsupervised object learning via common fate. *arXiv preprint arXiv:2110.06562*, 2021. 2

[47] Pavel Tokmakov, Allan Jabri, Jie Li, and Adrien Gaidon. Object permanence emerges in a random walk along memory. *arXiv preprint arXiv:2204.01784*, 2022. 2

[48] Manuel Traub, Frederic Becker, Sebastian Otte, and Martin V Butz. Looping loci: Developing object permanence from videos. *arXiv preprint arXiv:2310.10372*, 2023. 2

[49] Manuel Traub, Sebastian Otte, Tobias Menge, Matthias Karlbauer, Jannik Thuemmel, and Martin V. Butz. Learning What

and Where: Disentangling Location and Identity Tracking Without Supervision. In *ICLR*, 2023. 2

[50] Basile Van Hoorick, Pavel Tokmakov, Simon Stent, Jie Li, and Carl Vondrick. Tracking through containers and occluders in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13802–13812, 2023. 2

[51] Sjoerd van Steenkiste, Michael Chang, Klaus Greff, and Jürgen Schmidhuber. Relational neural expectation maximization: Unsupervised discovery of objects and their interactions. In *ICLR*, 2018. 2

[52] Rishi Veerapaneni, John D. Co-Reyes, Michael Chang, Michael Janner, Chelsea Finn, Jiajun Wu, Joshua Tenenbaum, and Sergey Levine. Entity Abstraction in Visual Model-based Reinforcement Learning. In *Conference on Robot Learning*, 2019. 1, 2

[53] Ning Wang, Yibing Song, Chao Ma, Wengang Zhou, Wei Liu, and Houqiang Li. Unsupervised deep tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1308–1317, 2019. 1

[54] Xiaolong Wang, Allan Jabri, and Alexei A Efros. Learning correspondence from the cycle-consistency of time. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2566–2576, 2019. 2

[55] Nicholas Watters, Loic Matthey, Christopher P Burgess, and Alexander Lerchner. Spatial broadcast decoder: A simple architecture for learning disentangled representations in vaes. *arXiv preprint arXiv:1901.07017*, 2019. 2

[56] Ziyi Wu, Nikita Dvornik, Klaus Greff, Thomas Kipf, and Animesh Garg. Slotformer: Unsupervised visual dynamics simulation with object-centric models. In *ICLR*, 2023. 1, 2, 5, 6, 7

[57] Ziyi Wu, Jingyu Hu, Wuyue Lu, Igor Gilitschenski, and Animesh Garg. Slotdiffusion: Object-centric generative modeling with diffusion models. In *NeurIPS*, 2023. 1, 2, 6, 8

[58] Jiaqi Xu, Cuiling Lan, Wenxuan Xie, Xuejin Chen, and Yan Lu. Slot-VLM: Object-event slots for video-language modeling. In *NeurIPS*, 2024. 1

[59] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *ICCV*, 2019. 1

[60] Linjie Yang, Yuchen Fan, Yang Fu, and Ning Xu. The 3rd large-scale video object segmentation challenge - video instance segmentation track, 2021. 1, 5

[61] Jaesik Yoon, Yi-Fu Wu, Heechul Bae, and Sungjin Ahn. An investigation into pre-training object-centric representations for reinforcement learning. In *ICML*, 2023. 1

[62] Haonan Yu and Wei Xu. Vonet: Unsupervised video object learning with parallel u-net attention and object-wise sequential vae. *arXiv preprint arXiv:2401.11110*, 2024. 2

[63] Andrii Zadaianchuk, Maximilian Seitzer, and Georg Martius. Self-supervised Visual Reinforcement Learning with Object-centric Representations. In *ICLR*, 2020. 1

[64] Andrii Zadaianchuk, Maximilian Seitzer, and Georg Martius. Object-centric learning for real-world videos by predicting temporal feature similarities. In *Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS 2023)*, 2023. 1, 2, 4, 5, 6, 8

[65] Zixu Zhao, Jiaze Wang, Max Horn, Yizhuo Ding, Tong He, Zechen Bai, Dominik Zietlow, Carl-Johann Simon-Gabriel, Bing Shuai, Zhuowen Tu, Thomas Brox, Bernt Schiele, Yanwei Fu, Francesco Locatello, Zheng Zhang, and Tianjun Xiao. Object-centric multiple object tracking. *ICCV*, 2023. 2

[66] Daniel Zoran, Rishabh Kabra, Alexander Lerchner, and Danilo J. Rezende. PARTS: Unsupervised segmentation with slots, attention and independence maximization. In *ICCV*, 2021. 2