BENCHMARKING UNSUPERVISED OBJECT REPRESENTATIONS FOR VIDEO SEQUENCES

Anonymous authors
Paper under double-blind review

ABSTRACT

Perceiving the world in terms of objects and tracking them through time is a crucial prerequisite for reasoning and scene understanding. Recently, several methods have been proposed for unsupervised learning of object-centric representations. However, since these models have been evaluated with respect to different downstream tasks, it remains unclear how they compare in terms of basic perceptual abilities such as detection, figure-ground segmentation and tracking of individual objects. To close this gap, we design a benchmark with three datasets of varying complexity and seven additional test sets which feature challenging tracking scenarios relevant for natural videos. Using this benchmark, we compare the perceptual abilities of four unsupervised object-centric learning approaches: VIMON, a video-extension of MONET, based on a recurrent spatial attention mechanism, OP3, which exploits clustering via spatial mixture models, as well as TBA and SCALOR, which use an explicit factorization via spatial transformers. Our results suggest that architectures with unconstrained latent representations and full-image object masks such as VIMON and OP3 are able to learn more powerful representations in terms of object detection, segmentation and tracking than the explicitly parameterized spatial transformer based architecture of TBA and SCALOR. We also observe that none of the methods are able to gracefully handle the most challenging tracking scenarios despite their synthetic nature, suggesting that our benchmark may provide fruitful guidance towards learning more robust object-centric video representations.

1 INTRODUCTION

Humans understand the world in terms of objects. Being able to decompose our environment into independent objects that can interact with each other is an important prerequisite for reasoning and scene understanding. Similarly, an artificial intelligence system would benefit from the ability to both extract objects and their interactions from video streams, and keep track of them over time.

Recently, there has been an increased interest in *unsupervised* learning of *object-centric representations*. The key insight of these methods is that the compositionality of visual scenes can be used to both discover (Eslami et al., 2016; Greff et al., 2019; Burgess et al., 2019) and track objects in videos (Greff et al., 2017; van Steenkiste et al., 2018; Veerapaneni et al., 2019) without supervision. However, it is currently not well understood how the learned visual representations of different models compare to each other quantitatively, since the models have been developed with different downstream tasks in mind and have not been evaluated using a common protocol. Hence, in this work, we propose a benchmark based on procedurally generated video sequences to test basic perceptual abilities of object-centric video models under various challenging tracking scenarios.

Intuitively, an unsupervised object-based video representation should *(1)* effectively identify objects as they enter a scene, *(2)* accurately segment objects, as well as *(3)* maintain a consistent representation for each individual object in a scene over time. These perceptual abilities can be evaluated quantitatively in the established *multi-object tracking* framework (Bernardin & Stiefelhagen, 2008; Milan et al., 2016). We propose to utilize this protocol for analyzing the strengths and weaknesses of different object-centric representation learning methods, independent of any specific downstream task, in order to uncover the different inductive biases hidden in their choice of architecture and loss formulation. We therefore compiled a benchmark consisting of three procedurally generated video

datasets of varying levels of visual complexity and two generalization tests. Using this benchmark, we quantitatively compared three classes of object-centric models, leading to the following insights:

- All of the models have shortcomings handling occlusion, albeit to different extents.
- OP3 (Veerapaneni et al., 2019) performs strongest in terms of quantitative metrics, but exhibits a surprisingly strong dependency on color to separate objects and accumulates false positives when fewer objects than slots are present.
- Spatial transformer models, TBA (He et al., 2019) and SCALOR (Jiang et al., 2020), train most efficiently and feature explicit depth reasoning in combination with amodal masks, but are nevertheless outperformed by the simpler model, VIMON, lacking a depth or interaction model, suggesting that the proposed mechanisms may not yet work as intended.

We will make our code, data, as well as a public leaderboard of results available.

2 RELATED WORK

Several recent lines of work propose to learn object-centric representations from visual inputs for static and dynamic scenes without explicit supervision. Though their results are promising, methods are currently restricted to handling synthetic datasets and as of yet are unable to scale to complex natural scenes. Furthermore, a systematic quantitative comparison of methods is lacking.

Selecting and processing parts of an image via *spatial attention* has been one prominent approach for this task (Mnih et al., 2014; Eslami et al., 2016; Kosiorek et al., 2018; Burgess et al., 2019; Yuan et al., 2019; Crawford & Pineau, 2019; Locatello et al., 2020). As an alternative, *spatial mixture* models decompose scenes by performing image-space clustering of pixels that belong to individual objects (Greff et al., 2016; 2017; 2019; van Steenkiste et al., 2018). While some approaches aim at learning a suitable representation for downstream tasks (Watters et al., 2019a; Veerapaneni et al., 2019), others target scene generation (Engelcke et al., 2019; von Kügelgen et al., 2020). We analyze three classes of models for processing videos, covering three models based on spatial attention and one based on spatial mixture modeling.

Spatial attention models with unconstrained latent representations use per-object variational autoencoders, as introduced by Burgess et al. (2019). von Kügelgen et al. (2020) adapts this approach for scene generation. So far, such methods have been designed for static images, but not for videos. We therefore extend MONET (Burgess et al., 2019) to be able to accumulate evidence over time for tracking, enabling us to include this class of approaches in our evaluation. Recent concurrent work on AlignNet (Creswell et al., 2020) applies MONET frame-by-frame and tracks objects by subsequently ordering the extracted objects consistently.

Spatial attention models with factored latents use an explicit factorization of the latent representation into properties such as position, scale and appearance (Eslami et al., 2016; Crawford & Pineau, 2019). These methods use spatial transformer networks (Jaderberg et al., 2015) to render per-object reconstructions from the factored latents (Kosiorek et al., 2018; He et al., 2019; Jiang et al., 2020). SQAIR (Kosiorek et al., 2018) does not perform segmentation, identifying objects only at the bounding-box level. We select Tracking-by-Animation (TBA) (He et al., 2019) and SCALOR (Jiang et al., 2020) for analyzing spatial transformer methods in our experiments, which explicitly disentangle object shape and appearance, providing access to object masks.

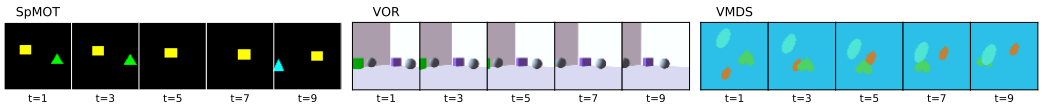
Spatial mixture models cluster pixels using a deep neural network trained with expectation maximization (Greff et al., 2017; van Steenkiste et al., 2018). IODINE (Greff et al., 2019) extends these methods with an iterative amortised variational inference procedure (Marino et al., 2018), improving segmentation quality. SPACE (Lin et al., 2020) combines mixture models with spatial attention to improve scalability. To work with video sequences, OP3 (Veerapaneni et al., 2019) extends IODINE by modeling individual objects' dynamics as well as pairwise interactions. We therefore include OP3 in our analysis as a representative spatial mixture model.

3 OBJECT-CENTRIC REPRESENTATION BENCHMARK

To compare the different object-centric representation learning models on their basic perceptual abilities, we use the well-established multi-object tracking (MOT) protocol (Bernardin & Stiefelhagen,

Table 1: Summary of datasets and example video sequences. See Appendix B for details.

	Objects						Background	
Dataset	Shape	Motion	Count Over Sequence	Size Variation	Orientation	Color	Motion	Color
SpMOT	4 Templates (2D)	Linear	Varies (0-3)	Minimal	Fixed	6 Colors	None	Black
VOR	2 Templates (3D)	Static	Varies (0-4)	Moderate	Random	6 Colors	Moving Camera	Random
VMDS	3 Templates (2D)	Non-Linear	Fixed (1-4)	Moderate	Random	256^3 Colors	None	Random



2008). In this section, we describe the datasets and metrics considered in our benchmark, followed by a brief description of the models evaluated.

3.1 Datasets

Current object-centric models are not capable of modeling complex natural scenes (Burgess et al., 2019; Greff et al., 2019; Lin et al., 2020). Hence, we focus on synthetic datasets that resemble those which state-of-the-art models were designed for. Specifically, we evaluate on three synthetic datasets[1] (see Table 1), which cover multiple levels of visual and motion complexity. Synthetic stimuli enable us to precisely generate challenging scenarios in a controllable manner in order to disentangle sources of difficulty and glean insights on what models specifically struggle with. We design different scenarios that test complexities that would occur in natural videos such as partial or complete occlusion as well as similar object appearances.

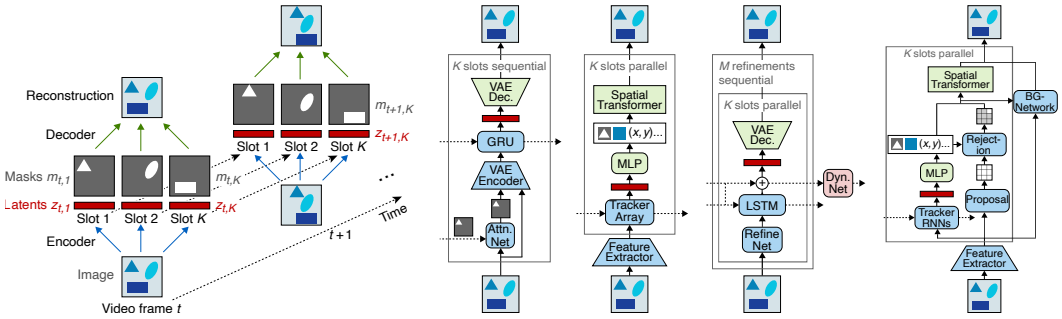
Sprites-MOT (**SpMOT**, Table 1 left), as proposed by He et al. (2019), features simple 2D sprites moving linearly on a black background with objects moving in and out of frame during the sequence. **Video-Multi-dSprites** (**VMDS**, Table 1 right) is a video dataset we generated based on a colored, multi-object version of the dSprites dataset (Matthey et al., 2017). Each video contains one to four sprites that move non-linearly and independently of each other with the possibility of partial or full occlusion. Besides the i.i.d. sampled training, validation and test sets of VMDS, we generate seven additional challenge sets that we use to study specific test situations we observed to be challenging, such as guaranteed occlusion, specific object properties, or out-of-distribution appearance variations. **Video Objects Room** (**VOR**, Table 1 middle) is a video dataset we generated based on the static Objects Room dataset (Greff et al., 2019), which features static objects in a 3D room with a moving camera. For full details on the datasets and their generation, see Appendix B.

3.2 Metrics

Our evaluation protocol follows the multi-object tracking (MOT) challenge, a standard and widely-used benchmark for supervised object tracking (Milan et al., 2016). The MOT challenge uses the CLEAR MOT metrics (Bernardin & Stiefelhagen, 2008), which quantitatively evaluate different performance aspects of object detection, tracking and segmentation. To compute these metrics, predictions have to be matched to ground truth. Unlike Bernardin & Stiefelhagen (2008) and Milan et al. (2016), we use binary segmentation masks for this matching instead of bounding boxes, which helps us better understand the models' segmentation capabilities. We consider an intersection over union (IoU) greater than 0.5 as a match (Voigtlaender et al., 2019). The error metrics used are the fraction of **Misses (Miss)**, **ID switches (ID S.)** and **False Positives (FPs)** relative to the number of ground truth masks. In addition, we report the **Mean Squared Error (MSE)** of the reconstructed image outputs summed over image channels and pixels.

To quantify the overall number of failures, we use the **MOT Accuracy (MOTA)**, which measures the fraction of all failure cases compared to the total number of objects present in all frames. A model with 100% MOTA perfectly tracks all objects without any misses, ID switches or false positives. To quantify the segmentation quality, we define **MOT Precision (MOTP)** as the average IoU of segmentation masks of all matches. A model with 100% MOTP perfectly segments all tracked objects, but does not necessarily track all objects. Further, to quantify detection and tracking performance

[1] Datasets are available at this https URL.



(a) Common Principle (b) ViMON (c) TBA (d) OP3 (e) SCALOR

Figure 1: Common principles of all models: Decomposition of an image into a fixed number of slots, each of which contains an embedding $z_{t,k}$ and a mask $m_{t,k}$ of (ideally) a single object. Dotted lines: temporal connections. Solid lines: information flow within one frame.

independent of false positives, we measure the **Mostly Detected (MD)** and **Mostly Tracked (MT)** metrics, the fraction of ground truth objects that have been detected and tracked for at least 80% of their lifespan, respectively. If an ID switch occurs, an object is considered detected but not tracked. For full details regarding the matching process and the evaluation metrics, refer to Appendix A.

3.3 Models

We consider three classes of unsupervised object-centric representation learning models: *(1)* a spatial attention model with unconstrained latents, ViMON, which is our video extension of MONet (Burgess et al., 2019); *(2)* spatial transformer-based attention models, TBA (He et al., 2019) and SCALOR (Jiang et al., 2020); *(3)* a scene mixture model, OP3 (Veerapaneni et al., 2019). At a high-level, these methods share a common structure which is illustrated in Fig. 1a. They decompose an image into a fixed number of *slots* (Burgess et al., 2019), each of which contains an embedding $\mathbf{z}_{t,k}$ and a mask $\mathbf{m}_{t,k}$ of (ideally) a single object. These slots are then combined in a decoding step to reconstruct the image. Below, we briefly describe each method. Appendix C provides a detailed explanation in a unified mathematical framework.

Video MONet (ViMON) is our video extension of MONet (Burgess et al., 2019). MONet recurrently decomposes a static scene into slots, using an attention network to sequentially extract attention masks $\mathbf{m}_k \in [0,1]^{H \times W}$ of individual objects k. A Variational Autoencoder (VAE) (Kingma & Welling, 2014) encodes each slot into a latent representation $\mathbf{z}_k \in \mathbb{R}^L$ of the corresponding object. We use MONet as a simple frame-by-frame baseline for detection and segmentation that does not employ temporal information. ViMON accumulates evidence about the objects over time to maintain a consistent object-slot assignment throughout the video. This is achieved by *(1)* seeding the attention network the predicted mask $\widehat{\mathbf{m}}_{t,k} \in [0,1]^{H \times W}$ from the previous time step and *(2)* introducing a gated recurrent unit (GRU) (Cho et al., 2014), which aggregates information over time for each slot separately, enabling it to encode motion information. For full details on MONet and ViMON, as well as ablations to provide context for the design decisions, refer to Appendix C.1, C.2 and E.3.

Tracking-by-Animation (TBA) (He et al., 2019) is a spatial transformer-based attention model. Frames are encoded by a convolutional feature extractor f before being passed to a recurrent block g called Reprioritized Attentive Tracking (RAT). RAT re-weights slot input features based on their cosine similarity with the slots from the previous time step and outputs latent representations for all K slots in parallel. Each slot latent is further decoded into a mid-level representation $\mathbf{y}_{t,k}$ consisting of pose and depth parameters, as well as object appearance and shape templates (see Fig. 1c). For rendering, a Spatial Transformer Network (STN) (Jaderberg et al., 2015) is used with an additional occlusion check based on the depth estimate. TBA is trained on frame reconstruction with an additional penalty for large object sizes to encourage compact bounding boxes. TBA can only process scenes with static backgrounds, as it preprocesses sequences using background subtraction (Bloisi & Iocchi, 2012). For full details on TBA, refer to Appendix C.3.

Object-centric Perception, Prediction, and Planning (OP3) (Veerapaneni et al., 2019) extends IODINE (Greff et al., 2019) to operate on videos. IODINE decomposes an image into objects and

Table 2: Analysis of SOTA object-centric representation learning models for MOT. Results shown as mean ± standard deviation of three runs with different random training seeds.

Model	MOTA ↑	MOTP ↑	MD ↑	MT ↑	Match ↑	Miss ↓	ID S. ↓	FPs ↓	MSE ↓
					SpMOT				
MONET	70.2 ± 0.8	89.6 ± 1.0	92.4 ± 0.6	50.4 ± 2.4	75.3 ± 1.3	4.4 ± 0.4	20.3 ± 1.6	5.1 ± 0.5	13.0 ± 2.0
VIMON	92.9 ± 0.2	**91.8 ± 0.2**	87.7 ± 0.8	87.2 ± 0.8	95.0 ± 0.2	4.8 ± 0.2	**0.2 ± 0.0**	2.1 ± 0.1	11.1 ± 0.6
TBA	79.7 ± 15.0	71.2 ± 0.3	83.4 ± 9.7	80.0 ± 13.6	87.8 ± 9.0	9.6 ± 6.0	2.6 ± 3.0	8.1 ± 6.0	11.9 ± 1.9
OP3	89.1 ± 5.1	78.4 ± 2.4	92.4 ± 4.0	91.8 ± 3.8	**95.9 ± 2.2**	3.7 ± 2.2	0.4 ± 0.0	6.8 ± 2.9	13.3 ± 11.9
SCALOR	**94.9 ± 0.5**	80.2 ± 0.1	**96.4 ± 0.1**	**93.2 ± 0.7**	**95.9 ± 0.4**	**2.4 ± 0.0**	1.7 ± 0.4	**1.0 ± 0.1**	**3.4 ± 0.1**
					VOR				
MONET	37.0 ± 6.8	81.7 ± 0.5	76.9 ± 2.2	37.3 ± 7.8	64.4 ± 5.0	15.8 ± 1.6	19.8 ± 3.5	27.4 ± 2.3	12.2 ± 1.4
VIMON	**89.0 ± 0.0**	**89.5 ± 0.5**	**90.4 ± 0.5**	**90.0 ± 0.4**	**93.2 ± 0.4**	**6.5 ± 0.4**	0.3 ± 0.0	4.2 ± 0.4	6.4 ± 0.6
OP3	65.4 ± 0.6	89.0 ± 0.6	88.0 ± 0.6	85.4 ± 0.5	90.7 ± 0.3	8.2 ± 0.4	1.1 ± 0.2	25.3 ± 0.6	**3.0 ± 0.1**
SCALOR	74.6 ± 0.4	86.0 ± 0.2	76.0 ± 0.4	75.9 ± 0.4	77.9 ± 0.4	22.1 ± 0.4	**0.0 ± 0.0**	**3.3 ± 0.2**	6.4 ± 0.1
					VMDS				
MONET	49.4 ± 3.6	78.6 ± 1.8	74.2 ± 1.7	35.7 ± 0.8	66.7 ± 0.7	13.6 ± 1.0	19.7 ± 0.6	17.2 ± 3.1	22.2 ± 2.2
VIMON	86.8 ± 0.3	86.8 ± 0.0	86.2 ± 0.3	85.0 ± 0.3	92.3 ± 0.2	7.0 ± 0.2	0.7 ± 0.0	5.5 ± 0.1	10.7 ± 0.1
TBA	54.5 ± 12.1	75.0 ± 0.9	62.9 ± 5.9	58.3 ± 6.1	75.9 ± 4.3	21.0 ± 4.2	3.2 ± 0.3	21.4 ± 7.8	28.1 ± 2.0
OP3	**91.7 ± 1.7**	**93.6 ± 0.4**	**96.8 ± 0.5**	**96.3 ± 0.4**	**97.8 ± 0.1**	**2.0 ± 0.1**	**0.2 ± 0.0**	6.1 ± 1.5	**4.3 ± 0.2**
SCALOR	74.1 ± 1.2	87.6 ± 0.4	67.9 ± 1.1	66.7 ± 1.1	78.4 ± 1.0	20.7 ± 1.0	0.8 ± 0.0	**4.4 ± 0.4**	14.0 ± 0.1

represents them independently by starting from an initial guess of the segmentation of the entire frame, and subsequently iteratively refines it using the refinement network f (Marino et al., 2018). In each refinement step m, the image is represented by K slots with latent representations $\mathbf{z}_{m,k}$. OP3 applies IODINE to each frame x_t to extract latent representations $\mathbf{z}_{t,m,k}$, which are subsequently processed by a dynamics network d (see Fig. 1e), which models both the individual dynamics of each slot k as well as the pairwise interaction between all combinations of slots, aggregating them into a prediction of the posterior parameters for the next time step $t+1$ for each slot k. For full details on IODINE and OP3, refer to Appendix C.4 and C.5, respectively.

SCALable Object-oriented Representation (SCALOR) (Jiang et al., 2020) is a spatial transformer-based model that factors scenes into background and multiple foreground objects, which are tracked throughout the sequence. Frames are encoded using a convolutional LSTM f. In the proposal-rejection phase, the current frame t is divided into $H \times W$ grid cells. For each grid cell a object latent variable $\mathbf{z}_{t,h,w}$ is proposed, that is factored into existence, pose, depth and appearance parameters. Subsequently, proposed objects that significantly overlap with a propagated object are rejected. In the propagation phase, per object GRUs are updated for all objects present in the scene. Additionally, SCALOR has a background module to encode the background and its dynamics. Frame reconstructions are rendered using a background decoder and foreground STNs for object masks and appearance. For full details on SCALOR, refer to Appendix C.6.

4 RESULTS

We start with a summary of our overall results across the three datasets and four models (Table 2) before analyzing more specific challenging scenarios using variants of the VMDS dataset.

We first ask whether tracking could emerge automatically in an image-based model like MONET, which may produce consistent slot assignments through its learned object-slot assignment. This is not the case: MONET exhibits poor tracking performance (Table 2). While MONET correctly finds and segments objects, it does not assign them to consistent slots over time (Fig. E.2). In the following, we will thus focus on the video models: VIMON, TBA, OP3 and SCALOR.

SpMOT. All models perform tracking well on SpMOT with the exception of one training run of TBA with poor results leading to high standard deviation (cp. best TBA model: 89.8% MT; Table E.1). SCALOR outperforms the other models on the detection and tracking metrics MD and MT, while VIMON exhibits the highest MOTP, highlighting its better segmentation performance on SpMOT.

VOR. TBA is not applicable to VOR due to the dynamic background which cannot be resolved using background subtraction. VIMON and OP3 show similarly good performance on detection (MD) and

segmentation (MOTP), while ViMON outperforms OP3 on the tracking metrics MOTA and MT. OP3 accumulates a high number of false positives leading to a low MOTA due to the splitting of objects into multiple masks as well as randomly segmenting small parts of the background (Fig. E.4). In contrast, SCALOR has almost no false positives or ID switches, but accumulates a high number of misses leading to a poor MOTA. It often segments two objects as one that are occluding each other in the first frame, which is common in VOR due to the geometry of its scenes (Fig. F.11, last row).

VMDS. OP3 outperforms the other models on VMDS, on which TBA performs poorly, followed by SCALOR, which again accumulates a high number of misses. We will analyze the models on VMDS qualitatively and quantitatively in more detail in the following.

Accumulation of evidence over time. Recognition and tracking of objects should improve if models can exploit prior knowledge about the objects in the scene from previous video frames. To test whether the models exploit such knowledge, we evaluate their MOTA performance on VMDS after warm starting with up to 10 frames which are not included in evaluation (Fig. 2). Note that the models were trained on sequences of length 10, but are run for 20 frames in the case of a warm start of 10 frames. The performance of ViMON improves with longer warm starts, showing that the GRU accumulates evidence over time. TBA, in contrast, does not use temporal information beyond 2–3 frames, while SCALOR's performance slightly drops after 3 frames. OP3 appears to most strongly rely on past information and is able to integrate information over longer time scales: its performance does not even saturate with a warm start of 10 frames. However, the effect for all models is rather small.

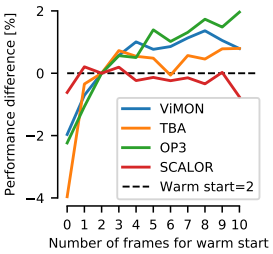


Figure 2: MOTA on frames 11–20 of the VMDS test set with warm starts of 1–10 frames (0 = no warm start). Difference to performance of warm start = 2 shown.

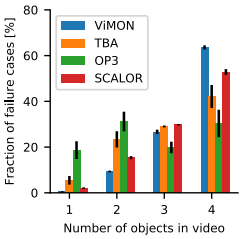


Figure 3: Distribution of failure cases dependent on number of objects in VMDS videos. Mean of three training runs. Error bars: standard deviation.

Challenging scenarios for different models. *The number of objects in the sequence* matters for ViMON, TBA and SCALOR: more objects increase the number of failure cases (Fig. 3). In contrast, OP3 does not exhibit this pattern: it accumulates a higher number of false positives (FPs) in videos with fewer (only one or two) objects (Fig. E.1), as it tends to split objects into multiple slots if fewer objects than slots are present.

Occlusion leads to failure cases for all models (Fig. 4a–b). Partial occlusion can lead to splitting of objects into multiple slots (Fig. 4a). Objects that reappear after full occlusion are often missed when only a small part of them is visible (Fig. 4a). In particular, SCALOR tends to segment two objects as one when they overlap while entering the scene, leading to a high number of misses.

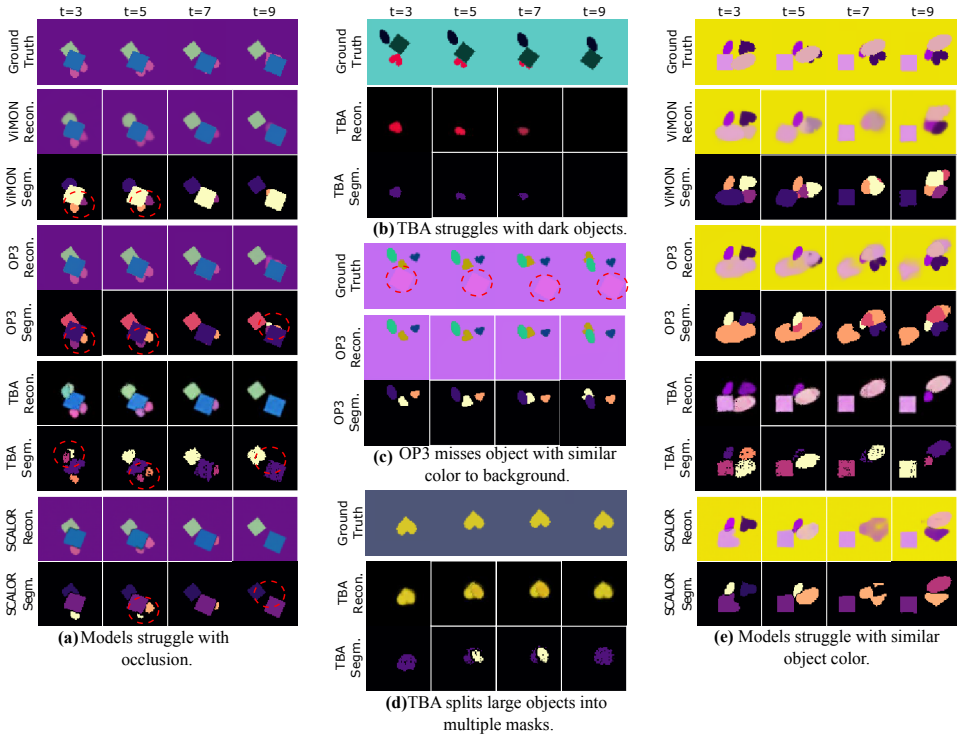
Color of the object is important. TBA often misses dark objects (Fig. 4b). In contrast, ViMON, OP3 and SCALOR struggle with scenes that feature objects of similar colors as well as objects that have similar colors to the background (Fig. 4c,e).

False positives are more prevalent for OP3 and TBA than for ViMON and SCALOR (Table 2). FPs of OP3 are due to objects split in multiple masks (Fig. 4a) and random small parts of the background being individually segmented (Fig. 4e), while TBA tends to compose larger objects using multiple smaller, simpler components (Fig. 4d).

Challenge sets. Based on the challenging scenarios identified above, we design multiple 'challenge sets': videos featuring *(1)* heavy occlusion, *(2)* objects with same colors, *(3)* only small objects and *(4)* only large objects (Fig. 5, top). For details, see Appendix (B.1.1).

Occlusion reduces performance of all models compared with the i.i.d. sampled VMDS test set, albeit to different degrees (Fig. 5; for absolute performance see Table E.2). OP3 is more robust to occlusion than the other models.

Tracking objects with the same color is challenging for all models (Fig. 5). In particular, OP3 appears to rely on object color as a way to separate objects.



(a) Models struggle with occlusion.

(b) TBA struggles with dark objects.

(c) OP3 misses object with similar color to background.

(d) TBA splits large objects into multiple masks.

(e) Models struggle with similar object color.

Figure 4: Example failure cases for all models on VMDS. Segmentation masks are binarized and color-coded to signify slot assignment.

OP3, ViMON and SCALOR are not sensitive to object size (Fig. 5). They exhibit only slightly decreased performance on the large objects test set, presumably because large objects cause more occlusion (Fig. 5). TBA shows increased performance on small objects but performs poorly on the large objects set.

Out-of-distribution test sets. Next, we assess generalization to out-of-distribution (o.o.d.) changes in object appearance that are not encountered during training. In the training set of VMDS, object color, size and orientation are constant throughout a video. To test o.o.d. generalization, we evaluate models trained on VMDS on three datasets that feature unseen object transformations (Fig. 6 and Table E.3): continuous changes in object **color** or **size** as well as continuous **rotation** around the object's centroid while moving. For details, see Appendix B.1.2.

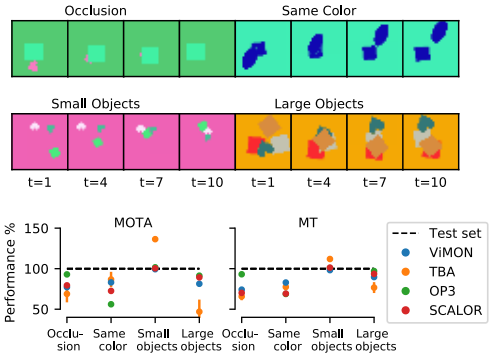


Figure 5: Performance on challenge sets relative to performance on VMDS test set (100%).

Continuous changes in object size do not pose a serious problem to TBA, OP3 and SCALOR, while ViMON's performance drops (Fig. 6). Surprisingly, continuous color changes of objects do not impact the performance of any model. Tracking performance of ViMON drops significantly for rotated objects, while OP3 and SCALOR are affected less. TBA's tracking performance is not as strongly influenced by object rotation (for absolute values, see Table E.3).

Stability of training and runtime. TBA and SCALOR train faster and require less memory than OP3 and ViMON (see Table E.4 for details). However, some training runs converge to suboptimal minima for TBA. Training OP3 is sensitive to the learning rate and unstable, eventually diverging in almost all experiments. Interestingly, it often reached its best performance prior to divergence. ViMON and TBA are less sensitive to hyper-parameter settings in our experiments. For a more detailed analysis of the runtime, see Appendix E.2.

5 DISCUSSION

Our experimental results provide insights into the inductive biases and failure cases of object-centric models that were not apparent from their original publications.

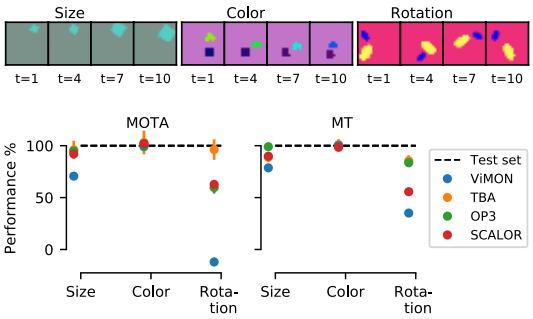


Figure 6: Performance on out-of-distribution sets relative to VMDS test set (100%).

Handling **occlusion** of objects during the video is a key property object-centric representations should be capable of. Qualitatively and quantitatively, OP3 is more robust to occlusion than the other models, suggesting that its dynamics network that models interactions between objects is currently most successful at modeling occluded objects. Surprisingly, TBA and SCALOR, which explicitly encode depth, do not handle occlusion more gracefully than VIMON, whose much simpler architecture has no explicit way of dealing with depth.

TBA has a significantly lower MOTP than the other models on all datasets, suggesting that the simple rendering-based decoder might be less suitable to generate accurate segmentation masks (see also Fig. F.5 and Fig. F.4) compared to the VAE-based approaches of VIMON, OP3 and SCALOR.

All models struggle with detecting objects that have similar **color** as the background (for TBA: dark objects, since background is removed and set to black in a pre-processing step). Color is a reliable cue to identify objects in these datasets. However, the auto-encoding objective incurs little extra loss for missing objects with similar color as the background and, thus, the models appear to not to learn to properly reconstruct them. Furthermore, the models – particularly so OP3 – struggle with separating objects of similar colors. This result hints at a mismatch between the intuitions motivating these models and what the models actually learn: it should be more efficient to decompose two objects – even of similar colors – into two masks with simple shapes, rather than encoding the more complicated shape of two objects simultaneously in one slot. However, since none of the models handle occlusion with amodal segmentation masks successfully, they learn to encode overly complex (modal) mask shapes. As a consequence, they tend to merge similarly colored objects into one slot. This result suggests that resolving the issues surrounding the depth reasoning in combination with amodal segmentation masks would enable much more compact latents and could also resolve the merging of similarly colored objects.

A major difference between models is the **spatial transformer** based model formulation of TBA and SCALOR compared to VIMON and OP3, which operate on image-sized masks. The parallel processing of objects and the processing of smaller bounding boxes renders training TBA and SCALOR significantly faster and more memory efficient, and enables scaling them to a larger number of objects. On the down side, the spatial transformer introduces its own complications. TBA depends strongly on its prior on object size and performs well only when this prior fits the data well and when the data contains little variation in object sizes, as on SpMOT (Table 2). However, it is not able to handle VMDS and its larger variation in object sizes and shapes. SCALOR performs tracking well in scenes where objects are clearly separated, but struggles to separate objects that partially occlude each other when entering the scene. This difficulty is caused by its discovery mechanism which can propose at most one bounding box per grid cell, leading to a high number of misses on datasets which feature significant occlusion (VOR and VMDS). Unfortunately, simply increasing the number of proposals is not an easy solution as SCALOR's performance is very sensitive to properly tweaking the number of proposals.

In conclusion, our analysis shows that even for relatively simple synthetic datasets none of the models shine in all respects. Future work should focus on developing robust mechanisms for reliably handling depth and occlusion, additionally combining the transformer-based efficiency of TBA and SCALOR with the stable training of VIMON and the interaction model of OP3. The key open challenges for scaling these models to natural videos include their computational inefficiency, complex training dynamics, as well as over-dependence on simple appearance cues like color.

REFERENCES

Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: The clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008.

Domenico Daniele Bloisi and Luca Iocchi. Independent multimodal background subtraction. In *CompIMAGE*, 2012.

Christopher P. Burgess, Loïc Matthey, Nicholas Watters, Rishabh Kabra, Irina Higgins, Matthew Botvinick, and Alexander Lerchner. MONet: Unsupervised scene decomposition and representation. *arXiv.org*, 1901.11390, 2019.

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv.org*, 1406.1078, 2014.

Eric Crawford and Joelle Pineau. Spatially invariant unsupervised object detection with convolutional neural networks. *Proc. of the Conf. on Artificial Intelligence (AAAI)*, 2019.

Antonia Creswell, Kyriacos Nikiforou, Oriol Vinyals, Andre Saraiva, Rishabh Kabra, Loic Matthey, Chris Burgess, Malcolm Reynolds, Richard Tanburn, Marta Garnelo, and Murray Shanahan. Alignnet: Unsupervised entity alignment. *arXiv.org*, 2007.08973, 2020.

Martin Engelcke, Adam R. Kosiorek, Oiwi Parker Jones, and Ingmar Posner. GENESIS: Generative scene inference and sampling with object-centric latent representations. *arXiv.org*, 1907.13052, 2019.

S. M. Ali Eslami, Nicolas Heess, Theophane Weber, Yuval Tassa, David Szepesvari, Koray Kavukcuoglu, and Geoffrey E Hinton. Attend, infer, repeat: Fast scene understanding with generative models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

Klaus Greff, Antti Rasmus, Mathias Berglund, Tele Hao, Harri Valpola, and Jürgen Schmidhuber. Tagger: Deep unsupervised perceptual grouping. In *Advances in Neural Information Processing Systems (NeurIPS)*. 2016.

Klaus Greff, Sjoerd van Steenkiste, and Jürgen Schmidhuber. Neural expectation maximization. In *Advances in Neural Information Processing Systems (NeurIPS)*. 2017.

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 *Proc. of the International Conf. on Machine learning (ICML)*, 2019.

Zhen He, Jian Li, Daxue Liu, Hangen He, and David Barber. Tracking by animation: Unsupervised learning of multi-object attentive trackers. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9 (8):1735–1780, November 1997. ISSN 0899-7667. doi: 10.1162/neco.1997.9.8.1735. URL https://doi.org/10.1162/neco.1997.9.8.1735.

Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. In *Advances in Neural Information Processing Systems (NeurIPS)*. 2015.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *Proc. of the International Conf. on Learning Representations (ICLR)*, 2017.

Jindong Jiang, Sepehr Janghorbani, Gerard De Melo, and Sungjin Ahn. Scalor: Generative world models with scalable object representations. In *Proc. of the International Conf. on Learning Representations (ICLR)*, 2020.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proc. of the International Conf. on Learning Representations (ICLR)*, 2015.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *Proc. of the International Conf. on Learning Representations (ICLR)*, 2014.

Adam Kosiorek, Hyunjik Kim, Yee Whye Teh, and Ingmar Posner. Sequential attend, infer, repeat: Generative modelling of moving objects. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Zhixuan Lin, Yi-Fu Wu, Skand Vishwanath Peri, Weihao Sun, Gautam Singh, Fei Deng, Jindong Jiang, and Sungjin Ahn. Space: Unsupervised object-oriented scene representation via spatial attention and decomposition. *arXiv.org*, 2001.02407, 2020.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. *arXiv.org*, 2006.15055, 2020.

Joe Marino, Yisong Yue, and Stephan Mandt. Iterative amortized inference. In *Proc. of the International Conf. on Machine learning (ICML)*, 2018.

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

A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler. MOT16: A benchmark for multi-object tracking. *arXiv.org*, 1603.00831, 2016.

Volodymyr Mnih, Nicolas Heess, Alex Graves, and Koray Kavukcuoglu. Recurrent models of visual attention. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015.

Sjoerd van Steenkiste, Michael Chang, Klaus Greff, and Jürgen Schmidhuber. Relational neural expectation maximization: Unsupervised discovery of objects and their interactions. In *Proc. of the International Conf. on Learning Representations (ICLR)*, 2018. URL https://openreview.net/forum?id=ryH20GbRW.

Rishi Veerapaneni, John D. Co-Reyes, Michael Chang, Michael Janner, Chelsea Finn, Jiajun Wu, Joshua B. Tenenbaum, and Sergey Levine. Entity abstraction in visual model-based reinforcement learning. *arXiv.org*, 1910.12827, 2019.

Paul Voigtlaender, Michael Krause, Aljosa Osep, Jonathon Luiten, Berin Balachandar Gnana Sekar, Andreas Geiger, and Bastian Leibe. Mots: Multi-object tracking and segmentation. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Julius von Kügelgen, Ivan Ustyuzhaninov, Peter Gehler, Matthias Bethge, and Bernhard Schölkopf. Towards causal generative scene models via competition of experts. *arXiv.org*, 2004.12906, 2020.

Nicholas Watters, Loic Matthey, Matko Bosnjak, Christopher P. Burgess, and Alexander Lerchner. COBRA: Data-efficient model-based RL through unsupervised object discovery and curiosity-driven exploration. *arXiv.org*, 1905.09275, 2019a.

Nicholas Watters, Loic Matthey, Christopher P. Burgess, and Alexander Lerchner. Spatial broadcast decoder: A simple architecture for learning disentangled representations in VAEs. *arXiv.org*, 1901.07017, 2019b.

Jinyang Yuan, Bin Li, and Xiangyang Xue. Generative modeling of infinite occluded objects for compositional scene representation. In *Proc. of the International Conf. on Machine learning (ICML)*, 2019.