OBJECT DYNAMICS DISTILLATION FOR SCENE DECOMPOSITION AND REPRESENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

The ability to perceive scenes in terms of abstract entities is crucial for us to achieve higher-level intelligence. Recently, several methods have been proposed to learn object-centric representations of scenes with multiple objects, yet most of which focus on static scenes. In this paper, we work on object dynamics and propose Object Dynamics Distillation Network (ODDN), a framework that distillates explicit object dynamics (e.g., velocity) from sequential static representations. ODDN also builds a relation module to model object interactions. We verify our approach on tasks of video reasoning and video prediction, which are two important evaluations for video understanding. The results show that the reasoning model with visual representations of ODDN performs better in answering reasoning questions around physical events in a video compared to the previous state-of-the-art methods. The distilled object dynamics also could be used to predict future video frames given two input frames, involving occlusion and objects collision. In addition, our architecture brings better segmentation quality and higher reconstruction accuracy.

1 INTRODUCTION

Humans learn to decompose scenes into independent objects to understand environment at a very early age. This pattern of scene understanding promotes the generation of cognition and common sense of physics, for example, the velocity of an object after a collision will change. However, the neural networks seem to lack the ability to acquire a compositional understanding of the world in terms of structured objects, which is crucial for generalizing and higher-level tasks such as planning and reasoning(Greff et al., 2020). To tackle this problem, many scene decomposition approaches have been proposed to segment an image with ground-truth segmentation labels as supervision(Ronneberger et al., 2015; Long et al., 2015; Badrinarayanan et al., 2017; Chen et al., 2017), which are difficult to obtain in most cases. Besides, these methods do not provide structured representations of segmented areas. Recently, several unsupervised methods have become popular in learning decomposition and object-centric representations for scenes with multiple objects(Hsieh et al., 2018; Van Steenkiste et al., 2018; Burgess et al., 2019; Locatello et al., 2020; Greff et al., 2017; Eslami et al., 2016; Kosiorek et al., 2018), and some of those learn disentangled features which represent crucial properties (e.g., 'color', 'size') of objects in the same format. The work of van Steenkiste et al. (2019) has proved that disentangled representations do in fact lead to better downstream performance.

Beyond static scenes (images), people understand dynamic physical events and predict how the future might unfold(Battaglia et al., 2013). Even very young infants have the intuition that temporarily occluded objects remain coherent wholes and follow spatially contiguous paths which is based on object dynamics(Baillargeon, 1987). However, existing scene decomposition and representation methods focus on static scenes, which means that when interpreting dynamic scenes (videos), those approaches process one frame at a time, and do not encode objects' dynamic properties. We argue that object dynamics is significant for both understanding a video and predicting the future.

In this work, we aim to distillate and disentangle object dynamics in an unsupervised manner and present ODDN (Object Dynamics Distillation Network). ODDN is built on a spatial mixtures framework as IODINE(Greff et al., 2019), a prime example of VAE(Kingma & Welling, 2013) based architecture. Basically, the model cluster pixels with an iterative amortized variational inference

procedure and segment and represent complex static scenes into disentangled latents. Our key insight is that the object dynamics hide in discrepancies between objects' static latents of different frames, thus, we design our model to learn object dynamics from their states of the current and previous frames. Afterward, we build a relation module to model object interactions. We evaluate ODDN on tasks of video understanding and future frames prediction, which are two important aspects of video understanding. Results show that representations with object dynamics perform better in reasoning task than representations with static properties only, and the relation module endows ODDN with the capability of predicting the future, taking object interactions (e.g., collision) into consideration. In addition, we notice that incorporating object dynamics and relations into the scene decomposition framework benefits the segmentation and reconstruction quality.

We highlight contributions of our approach: (1) Distillation of disentangled object dynamics from raw video inputs in an unsupervised manner; (2) State-of-the-art performance on CLEVRER(Yi et al., 2019) which focus on video understanding and causal reasoning; (3) Capability of future frames prediction in complex 3D scenes including physical events (occlusion, collision); (4) Better scene decomposition for spatial mixture architecture in terms of segmentation and reconstruction.

2 RELATED WORK

Object representation for static scenes. Representation learning is an important topic in deep learning. Several recent lines of work making breakthroughs in this area learns representations without explicit supervision. NEM(Greff et al., 2017) builds a spatial mixture model based on EM algorithm to discover a compositional object representation but fails to cope with colored or 3D data. AIR(Eslami et al., 2016) achieves scene interpretation by using a recurrent neural network that attends to one element at a time and encodes positions and appearances explicitly in latent variables. AIR is able to decompose 3D scenes, however, it generates partial images separately and adds them together to form the final image, which leads to its restriction: they can't be applied to scenes with occluded objects. MoNet(Burgess et al., 2019) uses multiple attention steps to tackle complex 3D scene, which is the first being capable of discovering object masks and completing partially occluded objects. IODINE(Greff et al., 2019) uses iterative variational inference to refine the inferred latents in each encoding step.

Object representation for dynamic scenes. Methods mentioned above are designed to decompose static scenes, hence they do not encode object dynamics in the latent space. Relying on NEM, R-NEM(Van Steenkiste et al., 2018) models relations between objects to learn physical interaction which enables the model to predict future frames for video input. SQAIR(Kosiorek et al., 2018) further extends AIR to image sequence to discover and track objects throughout the video frames, yet both extensions bear drawbacks of their predecessor. ViMON(Weis et al., 2020) is a video-extension of MoNet by building temporal connections between latents with GRU(Cho et al., 2014), ViMON learns a linear transformation to predict next time-step latents, encouraging dynamics encoded. Similarly, PROVIDE(Zablotskaia et al., 2020) applies 2D-LSTM(Graves et al., 2007) to IODINE to video data, and is able to simulate future frames. Although ViMON and PROVIDE encode object dynamics from sequential data, they do not model interactions between objects, thus they can't handle collision. Besides, object dynamics of both methods are spread in latent space implicitly, which lacks interpretability. It has been proved that disentangled representations allow one to learn models with better performance for downstream reasoning tasks with fewer samples, therefore, in this work, we aim to represent object dynamics explicitly. Instead of using RNNs, we match objects' latents between the current and previous frames and distillate their dynamics by applying a single-layer transformer encoder. On this basis, object relations could also be built explicitly because a collision between rigid objects changes their dynamics only.

Self-attention and reasoning. Transformers(Vaswani et al., 2017) have been widely studied recently. The essential self-attention mechanism aggregates weighted sequence information and updates them to obtain better representation for each input element. In our approach, the transformer is used to align object latents in different frames and distillate dynamics from matched stationary representations. Previous approaches like MAC (V+)(Yi et al., 2019), NS-DR(Yi et al., 2019), and DCL(Chen et al., 2021) build neuro-symbolic architectures which normally combine logic engine and language parser with a perception network, and achieve better performance than fully neural network on CLEVRER. However, ALOE(Ding et al., 2021) proves that neural networks can learn

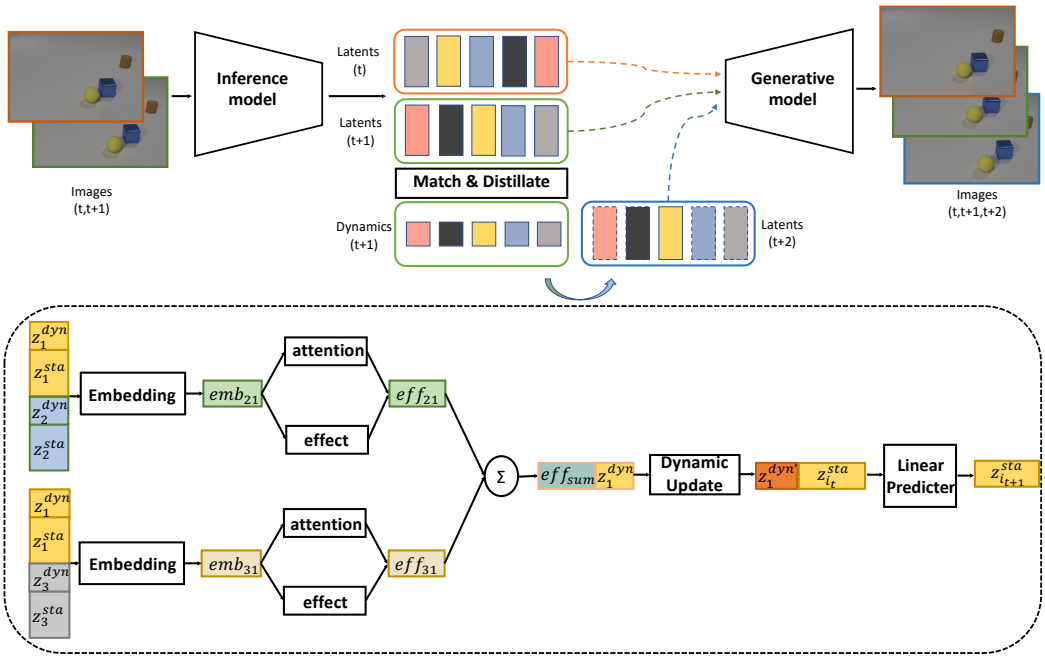

Figure 1: Overview of ODDN. The inference model assigns inputs into K slots per image, each slot encodes object or background into latent space, shown in different colors. We take use of the attention mechanism of transformer to align objects' latents of two input images and encode the aggregated representation into a low-dimensional vector to obtain disentangled object dynamics. Object latents (sta) and dynamics (dyn) are used to model interactions between object-pairs, we show the relation module with three slots as an example, which compute the effects z_2 and z_3 have on z_1. We concat the updated object dynamics and object latents to predict the next frame object latents. The Generative model outputs both reconstructions of input images and prediction of future frame images.

to reason about physical events and outperforms recent neuro-symbolic methods by a large margin. ALOE makes use of cross-modality self-attention and object-centric representations, as well as self-supervised predictive learning like BERT(Devlin et al., 2018). We take ALOE as our reasoning model and omit the self-supervision part for simplicity.

3 METHOD

Our goal is to distillate object dynamics explicitly purely based on visual observations. Recent works decompose static scenes into independent latents by randomly assigning different objects to a fixed number of slots, which share weights to obtain latents with a common format. However, representations of these works do not encode object dynamics for sequential input, thus, our key idea is that building upon these successful multi-object scene decomposing and representation approaches, object dynamics could be derived from the corresponding latents of two consecutive frames.

3.1 OBJECT DYNAMICS DISTILLATION NETWORK

We illustrate the framework of ODDN in Figure 1. Formally, ODDN takes as input a video, then the inference model processes all frames $x_{1:N}$ independently to obtain K object latents which encode object static properties. For frames $x_{2:N-1}$, we distillate objects' dynamics from their latents of current and previous frames. Our relation module takes as inputs object latents and dynamics of frames $x_{2:N-1}$, then calculates object interactions and predicts object latents of frames $x_{3:N}$.

Generative model. The input image $x \in \mathbb{R}^D$ could be reconstructed by K latents $z_k \in \mathbb{R}^M$, each of which corresponds to one object or pure background in the scene. Firstly, latent z_k is decoded into

pixel-wise mean μ_{ik} and mask logit $\hat{m}_{ik}$. Then, the K mask logits are normalized with a softmax function to ensure that the sum of K masks for each pixel is 1. At last, μ and m parameterize the final spatial Gaussian mixture distribution:

$$p(x|z) = \prod_{i=1}^{D} \sum_{k=1}^{K} m_{ik} \mathcal{N}(x_i; \mu_{ik}, \sigma^2), \tag{1}$$

and the input image can be reconstructed as $\hat{x} = \sum_{k=1}^{K} m_k \mu_k$.

Inference model. The inference model of ODDN follows Greff et al. (2019), by using amortized variational inference to refine posterior $q_\lambda(z|x)$ iteratively where $\lambda = \{\mu_z, \sigma_z\}$. The main reason to use such an iterative process is that the standard variational inference framework leads to the multimodal problem which means that any pixel could be captured by any slot. IODINE updates posterior λ as follows:

$$z^t \sim q_\lambda(z^t|x) \tag{2}$$

$$\lambda^{t+1} \leftarrow \lambda^t + f_\phi(z^t, x, a), \tag{3}$$

where the posterior parameters λ start with an arbitrary guess, and the first latent representations z^1 are sampled from q_λ. In each refinement step, input x together with a set of variables computed from z^t (means μ, masks m, mean gradient $\nabla_\mu \mathcal{L}$, mask gradient $\nabla_m \mathcal{L}$) a are fed into a refinement network $f\phi$, which outputs additive updates to the posterior.

ODDN predict future objects state in latent space given the previous two frames:

$$\hat{z}_i = W_{pred}([z_{i-1}^{dyn}; z_{i-1}^{sta}]); \quad i > 2, \tag{4}$$

where z^{dyn} encodes object dynamics and z^{dyn} refers to the output of the inference model. The next-frame latents are predicted with a learnable transformation W_{pred}, details will be introduced in section 3.2 and section 3.3.

The whole network is trained end-to-end and our loss consists of $\mathcal{L}(x)$=$\mathcal{L}_{prediction}$+$\mathcal{L}_{reconstruction}$, in detail:

$$\mathcal{L}_{reconstruction}(x) = \sum_{i=1}^{N} \sum_{t=1}^{T} \left[-log(p(x_i|z_i^t)) + \beta KL(q_\lambda(z_i^t|x_i, z_i^{<t})||p(z) \right] \tag{5}$$

$$\mathcal{L}_{prediction}(x) = \sum_{i=3}^{N} \left[-log(p(x_i|\hat{z}_i)) \right] \tag{6}$$

where T is the maximum refinement step number during the inference stage, we measure and optimize the KL divergence between the variational posterior and the prior in $\mathcal{L}_{reconstruction}$.

3.2 DYNAMIC DISTILLATION

We aim to derive object dynamics from sequential object latents. However, for existing scene decomposition and representation approaches, object-to-slot assignments usually switch unpredictably in a video. To match object latents in different frames correctly and further distillate object dynamics, we draw support from the Transformer.

Revisit attention mechanism of Transformer. The key component of the Transformer is the attention module which could be mathematically described as:

$$Attention(q, K, V) = \sum_{i=1}^{|V|} \alpha_i W_v v_i; \quad \alpha_i \propto exp[(W_q q)^T (W_k k_i)], \tag{7}$$

where $q \in \mathbb{R}_d$; K consists of $|K|$ elements $(k_1, k_2, \cdots, k_{|K|})$ with $k_i \in \mathbb{R}_d$; V consists of $|V|$ elements $(v_1, v_2, \cdots, v_{|V|})$ with $v_i \in \mathbb{R}_d$.

This attention function map input query q and a set of key(k)-value(v) pairs to an updated representation, W_q, W_k, W_v are learnable parameters. Besides, $\alpha_i \geq 0, \sum_{i=1}^{|v|} \alpha_i = 1$, and usually $|K| = |V|$.

Model architecture. The inference model decomposes the input video into $N * K$ static object latents, where K is the slot number for each frame. We put these sequence consisting latent vectors $\{z_{11}, \cdots, z_{1K}, \cdots, z_{N1}, \cdots, z_{NK}\}$ into one single Transformer encoder layer to match corresponding slots which usually switch their assignments with objects several times in a video. After finding object correspondence in different frames, the Transformer aggregates latents information and distillates object dynamics which are encoded in low-dimensional representation space. Mathematically, the dynamic distillation process for the k-th slot in the n-th frame could be described as:

$$z'_{nk} = Attention(z_{nk}, \{z_{n-1}, z_n\}, \{z_{n-1}, z_n\}); \quad n > 1, \tag{8}$$

$$z^{dyn}_{nk} = FFN(Norm(z'_{nk} + z_{nk})), \tag{9}$$

where z^{dyn} is a 4-d vector transformed by the fully connected feed-forward network. This learning process is performed in parallel with a pre-defined attention mask to ensure that latents of the current frame only attend to ones one frame before.

3.3 RELATION MODULE.

To model interactions between object pairs, we compute all object-pair effects and let it act on object dynamics, and learn an attention coefficient for each effect. The object dynamics are updated as:

$$emb_{k,i} = W_{emb}([z^{dyn}_k; z^{sta}_k; z^{dyn}_i; z^{sta}_i]), \tag{10}$$

$$att_{k,i} = Sigmoid(W_{att}(emb_{k,i})), \tag{11}$$

$$eff_{k,i} = W_{eff}(emb_{k,i}), E_k = \sum_{i \neq k} att_{k,i} * eff_{k,i}, \tag{12}$$

$$z^{dyn'}_k = z^{dyn}_k + W_{dyn}(E_k), \tag{13}$$

where W_{emb} first embeds representations of object-pair, then W_{att} learns attention coefficient based the embedding and W_{emb} learns the effect z_i has on z_k. At last, the summed effects output an update with W_{dyn}.

We draw inspiration from NPE(Chang et al., 2016) and R-NEM(Van Steenkiste et al., 2018), both architectures model object relations. NPE has access to ground truth representations of scenes and computes effect for each neighborhood object-pair with a neighborhood mask based on ground-truth object position, and finally, the summed effects act on object velocity. R-NEM represents objects in an unsupervised fashion and implicitly encodes sequential information with an RNN(Hochreiter & Schmidhuber, 1997). Instead of using a neighborhood mask, R-NEM learns soft attention for all object-pairs. Our approach learns to represent object dynamics explicitly. Thus, we compute all object-pair effects and let it act on object dynamics as NPE, and learn an attention coefficient for each effect as R-NEM.

Object's motion consists with the basic physics concept that object trajectory has to be continuous through space. After distilling and updating object dynamics, we predict next-frame object latents by simply learning a transformation W_{pred}, and $W_{pred} \in \mathbb{R}^{(L+4)\times L}$, $\hat{z}^{sta}_{t+1} = W_{pred}([z^{dyn'}_t; z^{sta}_t])$.

To alleviate error accumulating when implementing multi-step predicting in the inference stage, we use student forcing strategy during training: instead of only using z_t to predict $\hat{z}_{t+1}$, $\hat{z}_t$ is also used to iteratively roll out a predicted image sequence.

4 EXPERIMENTS

In this section, we design experiments in the perspective of representation, prediction, and scene decomposition. In particular, we study how ODDN performs on tasks of video understanding and reasoning, video prediction, reconstruction, and segmentation. To quantify segmentation quality, we measure the similarity between ground-truth (instance) segmentations and our predicted object masks using the Adjusted Rand Index(Rand, 1971) (ARI) and The Foreground Adjusted Rand Index (F-ARI) which is a modification of the ARI score ignoring background pixels.

Table 1: Comparison with the state-of-the-art methods MAC (V+), NS-DR, DCL, and ALOE on CLEVRER, ALOE takes static scene representaions from Monet as visual input while ODDN-ALOE use representations of ODDN.

Model	descriptive	explanatory	predictive	counterfactual
MAC (V+)	86.4	22.3	42.9	25.1
NS-DR	88.1	79.6	68.7	42.2
DCL	90.7	82.8	82.0	46.5
ALOE	94.0	96.0	87.5	75.6
ODDN-ALOE	**95.8**	**97.0**	**91.8**	**80.1**

Table 2: Ablations of how different object representations perform on CLEVRER, all experiments are conducted based on ALOE without self-supervision, "Reported" is the result from original paper with Monet features.

Model	dimension	dyn	descriptive	explanatory	predictive	counterfactual
Reported	16		91.0	92.8	82.8	68.7
PROVIDE	16	$\surd$	75.2	77.4	65.4	39.4
IODINE	16		93.5	94.9	84.5	73.4
ODDN w/o dyn	16		95.1	96.1	89.1	76.8
ODDN	20	$\surd$	**95.8**	**97.0**	**91.8**	**80.1**

Dataset. CLEVRER is a synthetic video dataset of moving and colliding objects. Each video contains 128 frames at resolution 480×320. We pre-train ODDN on the entire CLEVRER training set, in order to promote convergence of our Relation Module, we extract images every 4 frames and ensure that at least one collision event is included forming CLEVRER-collision and we finetune ODDN on CLEVRER-collision. For testing, we use the validation set which has ground truth masks, we sample 1k sub-clips containing 6 objects, each sub-clip consists of 10 frames.

Baselines. We compare our method to IODINE and PROVIDE. IODINE is designed for static images, thus it neither encodes object dynamics nor considers object relations. PROVIDE is an extension of IODINE to fit sequential inputs. The major difference between PROVIDE and our approach is that PROVIDE encodes object dynamics implicitly while we distillate object dynamics explicitly. In detail, PROVIDE uses 2D-LSTM to achieve temporally conditioned and iterative amortized inference for posterior refinement, as a result, the final sampled latents in every time-step contain dynamic information which could be used to predict the future. Besides, PROVIDE does not model object relations, which means when predicting the future, object latents are processed separately.

4.1 VIDEO UNDERSTANDING AND REASONING

We study the representation power of ODDN on CLEVRER, which poses different types of questions for each video in CLEVRER dataset. Those questions could test the model's understanding of a video comprehensively, including descriptive questions ("how many objects are moving"), explanatory questions("the reason why certain collision happens"), predictive questions("what will happen next"), and counterfactual questions("what would happen if").

Table 1 shows that ODDN-ALOE outperforms ALOE on all question types even without self-supervision, especially on predictive and counterfactual questions, and achieves state-of-the-art performance. Both predictive and counterfactual types of questions are related to predicting how objects will move. This result proves that representations with object dynamics could help the neural model to do better understanding and reasoning around physical events. We also do ablation experiments to study how different representations affect the model's performance, results are shown in Table 2. We can tell that representations of PROVIDE lead to a significant decline compared to IODINE, and the reason will be discussed in Section. 4.3. On the contrary, representations of ODDN which encode object dynamics (dim=20) outperform IODINE (dim=16) by a large margin, especially on predictive (91.8 vs 84.5) and counterfactual (80.1 vs 73.4) questions. In addition, we obtain ODDN features without dynamics by reducing 4 dimension dynamics. Results show that ODDN without

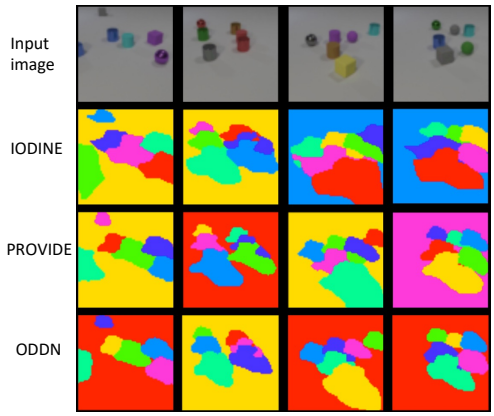

Figure 2: Unsupervised Segmentation results of IODINE, PROVIDE, and ODDN. Each color corresponding to one slot index whose assignment is unpredictable. Both PROVIDE and ODDN generate better segmentation masks than IODINE.

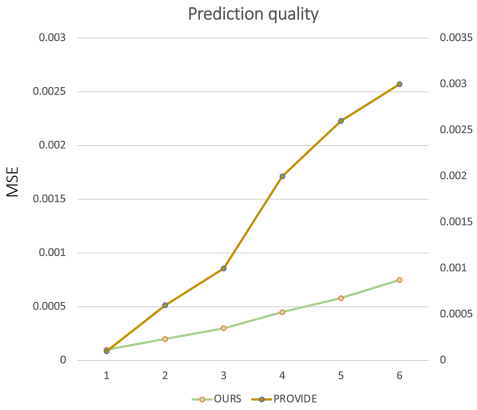

Figure 3: Comparison of MSE for 6 predicted frames between PROVIDE and ODDN.

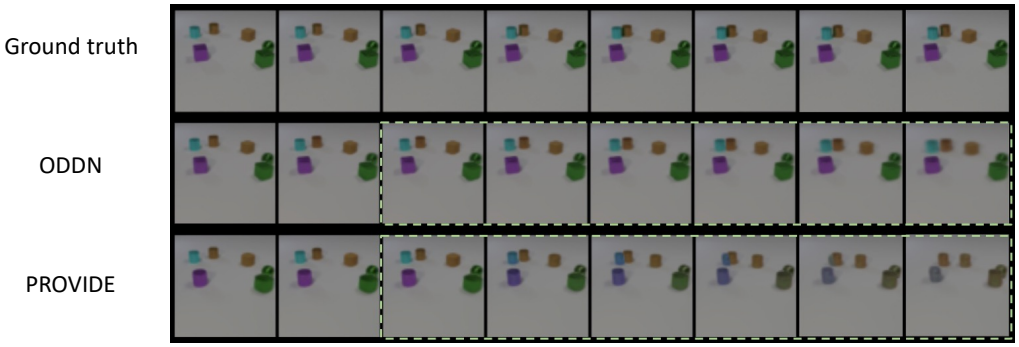

Figure 4: Illustration of video prediction results. PROVIDE and our model both take as input the first two frames, the first two rows of the second and third lines are reconstructed results from inferenced latents, base on which, the next 6 frames are simulated iteratively.

dynamics still performs better than IODINE, which means that ODDN not only distillates useful object dynamics but also generates higher quality object static representations.

4.2 PREDICTION

ODDN achieves predicting future frames in an iterative fashion. Unlike PROVIDE, ODDN models object dynamics and interactions explicitly. Results could be observed in Figure 4, which shows that ODDN could output reasonable predictions and handle objects collision. PROVIDE could only predict object's future state independently, and the reason is that PROVIDE does not cope with objects interaction, thus, objects blend when they suppose to collide and change their dynamics.

ODDN also generates more stable predictions comparing to PROVIDE which fails to capture the object's correct properties (shape, color) after a few frames, as shown in Figure 3, the MSE of predicted frames of our model increase much more slowly than PROVIDE.

4.3 DECOMPOSITION

Reconstruction. We show how different architectures perform in terms of reconstruction quality in Table 3. Comparing to IODINE, PROVIDE brings some negative effects, as shown in Figure

Table 3: We compare ODDN with baseline models on scene decomposition quality, ODDN* indicates training with the long-range prediction (6 frames).

Model	temporal-consistency	MSE	ARI	F-ARI
IODINE		0.000100	0.096	95.55
PROVIDE	√	0.000148	0.228	95.17
ODDN	√	**0.000075**	0.156	**95.63**
ODDN*	√	0.000103	**0.233**	95.36

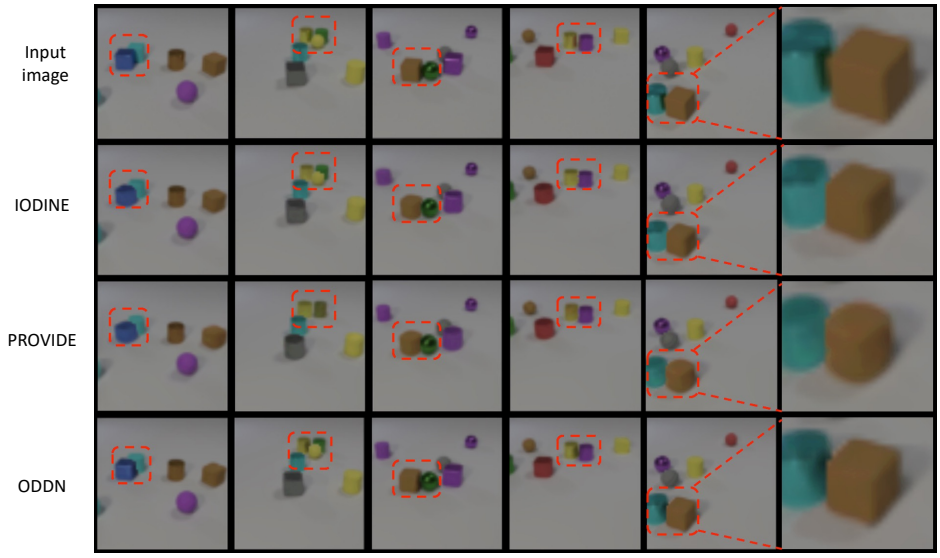

Figure 5: Comparison of reconstruction results of our approach with IODINE and PROVIDE, we highlight dense areas with dashed boxes.

5. PROVIDE occasionally fails to capture the right object shape. Instead, we notice that ODDN performs better than IODINE. Specifically, the reconstruction of close objects is more precise. We believe that this benefits from the relation module which relies on object surface details to process objects interaction (weather collision occurs and how object dynamics update). The reconstruction performance is positively correlated to object representation quality which has been verified in Section. 4.1.

Segmentation. ODDN and baseline models decode object latents into pixel values and segmentation masks. The major contribution of PROVIDE is that it models temporal dependencies between latents across frames to obtain better segmentation results. However, Table 3 indicates that PROVIDE hurts the model's reconstruction capability. Instead, we show that the future frame prediction task in our architecture also improves the segmentation performance and achieves good reconstruction quality at the same time.

The ARI further improves when increasing the steps of future object states to predict. We believe that the long-range prediction supervision encourages ODDN to attend more on foreground objects which may change its position and view over time. This temporal consistency is similar to PROVIDE, which could be the main reason for both models to achieve better segmentation performance than IODINE. Results are shown in Figure 2 and Table 3.

4.4 Component visualization

Attention. We visualize the attention coefficients of our relation module, results are shown in Figure 6. The cyan cylinder mostly attends on the brown cylinder and the attention score achieves the maximum when the collision happens, while other objects which do not get involved in the

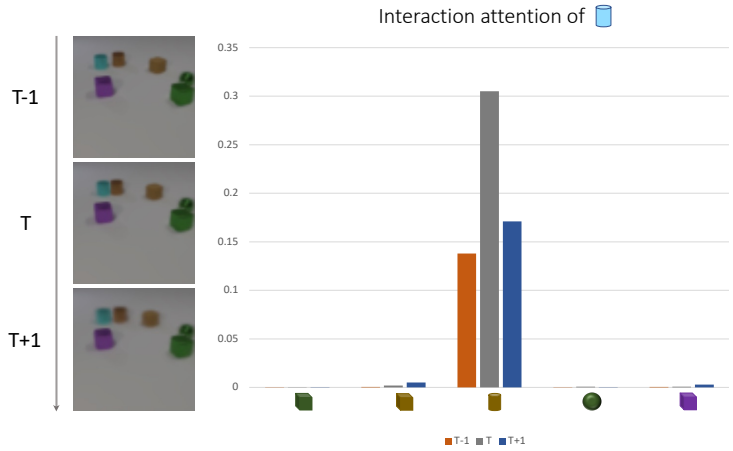

Figure 6: Attention coefficients visualization for relation module. We plot attention scores the cyan cylinder have on other objects for three frames, and collision happens on the second frame.

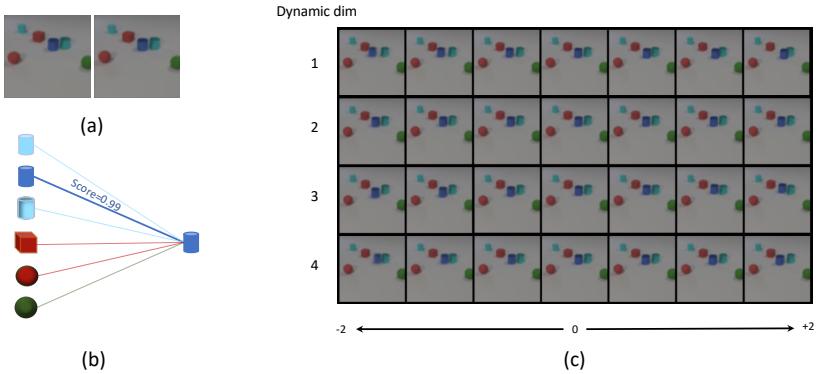

Figure 7: (a) Visualization of attention map of the blue cylinder in the matching phase of dynamics distillation, latents of the second frame will update it's representations by attending and merge information from latents of the first frame. line intensity indicates the magnitude of attention probability. (b) Disentanglement of learned object dynamics, rows indicate single dynamic dimensions. We adjust the factor value on the corresponding dimension of the blue cylinder, and visualize the effects on the predicted next frame.

interaction process barely gain attention. We also draw an attention map for the Transformer in Figure 7(a), which shows that the attention model could match latents in different frames precisely.

Disentanglement. Disentanglement is an important property that describes how well representations separate features into distinct factors. In Figure 7(b), we show that different dimensions of the learned object dynamics corresponding to different moving directions, and the factor value indicates the object's velocity magnitude.

5 CONCLUSION

We propose a unsupervised learning framework ODDN to decompose temporal scene of multiple objects with complex appearance and dynamic. ODDN distills explicit object dynamics which are proved to be significant in down-stream physical reasoning task. Our approach also incorporates object interactions which endows the model with the capability of predicting future frames, the prediction supervision also benefits the segmentation mask and reconstruction quality.

REFERENCES

Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.

Renee Baillargeon. Object permanence in $31/2$-and $41/2$-month-old infants. *Developmental psychology*, 23(5):655, 1987.

Peter W Battaglia, Jessica B Hamrick, and Joshua B Tenenbaum. Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, 110(45):18327–18332, 2013.

Christopher P Burgess, Loic Matthey, Nicholas Watters, Rishabh Kabra, Irina Higgins, Matt Botvinick, and Alexander Lerchner. Monet: Unsupervised scene decomposition and representation. *arXiv preprint arXiv:1901.11390*, 2019.

Michael B Chang, Tomer Ullman, Antonio Torralba, and Joshua B Tenenbaum. A compositional object-based approach to learning physical dynamics. *arXiv preprint arXiv:1612.00341*, 2016.

Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4): 834–848, 2017.

Zhenfang Chen, Jiayuan Mao, Jiajun Wu, Kwan-Yee Kenneth Wong, Joshua B Tenenbaum, and Chuang Gan. Grounding physical concepts of objects and events through dynamic visual reasoning. *arXiv preprint arXiv:2103.16564*, 2021.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

David Ding, Felix Hill, Adam Santoro, Malcolm Reynolds, and Matt Botvinick. Attention over learned object embeddings enables complex visual reasoning, 2021.

SM Eslami, Nicolas Heess, Theophane Weber, Yuval Tassa, David Szepesvari, Geoffrey E Hinton, et al. Attend, infer, repeat: Fast scene understanding with generative models. *Advances in Neural Information Processing Systems*, 29:3225–3233, 2016.

Alex Graves, Santiago Fernández, and Jürgen Schmidhuber. Multi-dimensional recurrent neural networks. In *International conference on artificial neural networks*, pp. 549–558. Springer, 2007.

Klaus Greff, Sjoerd Van Steenkiste, and Jürgen Schmidhuber. Neural expectation maximization. *arXiv preprint arXiv:1708.03498*, 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 *International Conference on Machine Learning*, pp. 2424–2433. PMLR, 2019.

Klaus Greff, Sjoerd van Steenkiste, and Jürgen Schmidhuber. On the binding problem in artificial neural networks. *arXiv preprint arXiv:2012.05208*, 2020.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

Jun-Ting Hsieh, Bingbin Liu, De-An Huang, Li Fei-Fei, and Juan Carlos Niebles. Learning to decompose and disentangle representations for video prediction. *arXiv preprint arXiv:1806.04166*, 2018.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Adam R Kosiorek, Hyunjik Kim, Ingmar Posner, and Yee Whye Teh. Sequential attend, infer, repeat: G@inproceedingsgraves2007multi, title=Multi-dimensional recurrent neural networks, author=Graves, Alex and Fernández, Santiago and Schmidhuber, Jürgen, booktitle=International conference on artificial neural networks, pages=549–558, year=2007, organization=Springer enerative modelling of moving objects. *arXiv preprint arXiv:1806.01794*, 2018.

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

Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, 2015.

William M Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850, 1971.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Sjoerd Van Steenkiste, Michael Chang, Klaus Greff, and Jürgen Schmidhuber. Relational neural expectation maximization: Unsupervised discovery of objects and their interactions. *arXiv preprint arXiv:1802.10353*, 2018.

Sjoerd van Steenkiste, Francesco Locatello, Jürgen Schmidhuber, and Olivier Bachem. Are disentangled representations helpful for abstract visual reasoning? *arXiv preprint arXiv:1905.12506*, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

Marissa A Weis, Kashyap Chitta, Yash Sharma, Wieland Brendel, Matthias Bethge, Andreas Geiger, and Alexander S Ecker. Unmasking the inductive biases of unsupervised object representations for video sequences. *arXiv preprint arXiv:2006.07034*, 2020.

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. Clevrer: Collision events for video representation and reasoning. *arXiv preprint arXiv:1910.01442*, 2019.

Polina Zablotskaia, Edoardo A Dominici, Leonid Sigal, and Andreas M Lehrmann. Unsupervised video decomposition using spatio-temporal iterative inference. *arXiv preprint arXiv:2006.14727*, 2020.