IEEE International Conference on Image Processing (ICIP), Kuala Lumpur, Malaysia, October 2023.

OBJECT-CENTRIC VIDEO PREDICTION VIA DECOUPLING OF OBJECT DYNAMICS AND INTERACTIONS

Angel Villar-Corrales[†] *Ismail Wahdan*[†] *Sven Behnke*

Autonomous Intelligent Systems, University of Bonn, Germany

ABSTRACT

We present a framework for object-centric video prediction, i.e., parsing a video sequence into objects, and modeling their dynamics and interactions in order to predict the future object states from which video frames are rendered. To facilitate the learning of meaningful spatio-temporal object representations and forecasting of their states, we propose two novel object-centric video prediction (OCVP) transformer modules, which decouple the processing of temporal dynamics and object interactions. We show how OCVP predictors outperform object-agnostic video prediction models on two different datasets. Furthermore, we observe that OCVP modules learn consistent and interpretable object representations. Animations and code to reproduce our results can be found in our project website[1].

Index Terms— Object-centric video prediction, scene parsing, object-centric learning, future frame prediction, transformers

1. INTRODUCTION

Humans perceive the world by parsing scenes into background and multiple foreground objects that can interact with each other [1]. Scene modeling approaches that are equipped with inductive biases for such decomposition have the ability to obtain modular and structured representations with desirable properties such as sample efficiency, improved generalization, and more robust representations [2, 3]. In recent years, unsupervised approaches to decompose images or video sequences into their object-centric components achieved impressive results on downstream tasks, such as object discovery [4, 5, 6] or object tracking [7, 8]. Despite these recent advances in object-centric decomposition, modeling temporal dynamics and object interactions from visual observations alone remains challenging.

To model spatio-temporal object dynamics, we present a framework for object-centric video prediction. Our approach, depicted in Fig. 1, uses a scene parsing module to extract object representations from video frames, learns a sequence prediction module to model temporal dynamics and object interactions using these representations, and renders future video frames from predicted object states.

A key component in our framework is the sequence predictor, which models multi-object dynamics to predict future object states. Different predictor designs embody distinct inductive biases, which affect their suitability for object prediction. We investigate multiple predictors for object-centric video prediction and propose two novel object-centric video prediction (OCVP) transformers, which decouple the processing of temporal dynamics and object interactions.

In summary, our contributions are: **(1)** We present a framework for object-centric video prediction, which decomposes video frames into object representations and models object dynamics and interactions. **(2)** We propose two novel object-centric predictor modules, which decouple the processing of temporal dynamics and object interactions. **(3)** Our prediction framework using our OCVP modules outperforms object-agnostic models for the task of video prediction while learning consistent interpretable object representations.

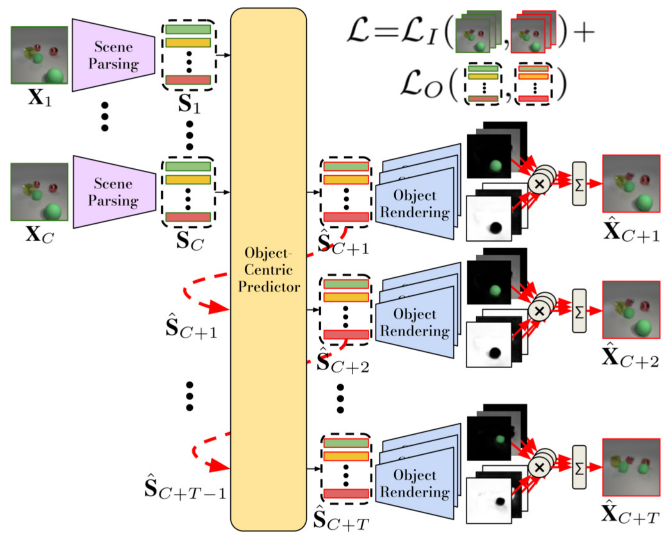


Fig. 1: Overview of our object-centric prediction framework. We decompose the seed video frames into object slot representations and learn an autoregressive object-centric predictor to model the object dynamics and interactions so as to predict future object states. Predicted slot representations are independently rendered into object images and masks, which are combined to generate the subsequent video frames. Our approach is trained by simultaneously minimizing the slot prediction and video frame prediction mean squared errors.

2. RELATED WORK

Object-Centric Learning: Our approach is inspired by recent works on unsupervised image and video object-centric decomposition. Image decomposition approaches [4, 6, 9, 10] design models with suitable inductive biases to enforce object-centric decomposition; whereas sequential models [11, 5] leverage the object movement in video sequences to identify objects. Our work extends the SAVi [5] video decomposition model to perform object-centric video prediction. However, our proposed OCVP modules are general and could be integrated with a variety of decomposition models.

Video Prediction: Future frame video prediction is the task of forecasting future video frames conditioned on past frames. Many different approaches have been proposed for this task, including 3D

This work was funded by grant BE 2556/16-2 (Research Unit FOR 2535 Anticipating Human Behavior) of the German Research Foundation (DFG).

† denotes equal contribution.

[1] `https://sites.google.com/view/ocvp-vp`

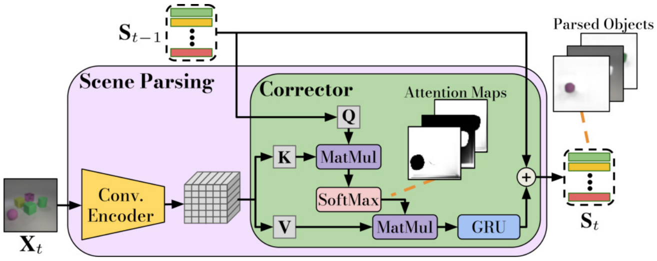


Fig. 2: Scene parsing module. We parse seed frames into object slots using a convolutional encoder and a Slot Attention corrector.

convolutions [12, 13], RNNs [14, 15, 16], and transformers [17, 18]. Despite recent advances, most existing methods model temporal changes using image-level features, without explicitly modeling the composition of the video frames or object dynamics. A few works follow more structured object-centric approaches for video prediction. Farazi *et al.* [19] employ local phase differences in order to separate foreground objects from a static background while modeling the foreground motion. The approaches presented in [20, 21, 22, 23] employ structured or object-centric representations to perform video prediction, at the cost of requiring explicit human supervision, error-prone Hungarian alignment operations, or being only applicable to very simple 2D datasets. The most similar approach to ours, developed concurrently with this work, is SlotFormer [24], which also combines SAVi [5] with an autoregressive transformer to perform object-centric video prediction. In this work, we further investigate the role of the predictor and propose two novel object-centric transformers that decouple object dynamics and interactions.

3. METHOD

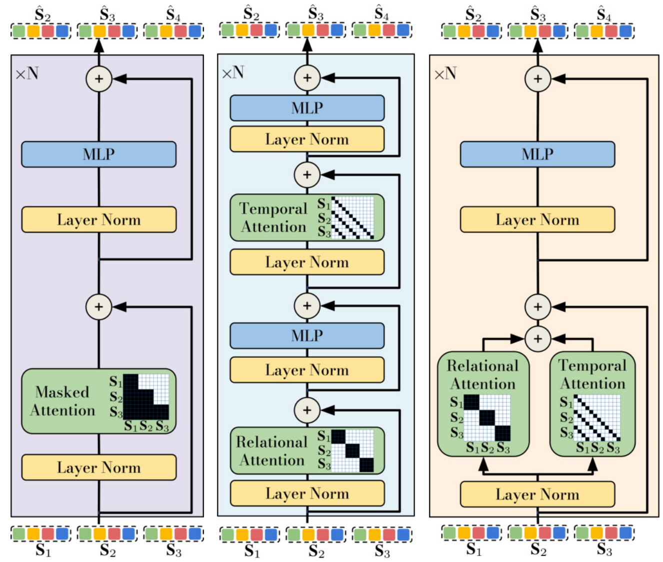
Video prediction is defined as the task of, given C seed video frames $\boldsymbol{\mathcal{X}}_{1:C}$, predicting the subsequent T frames $\hat{\boldsymbol{\mathcal{X}}}_{C+1:C+T}$. In this work, we address this task in an object-centric manner, i.e., decomposing the C seed frames into object representations $\boldsymbol{\mathcal{S}}_{1:C} = (\mathbf{S}_1, ..., \mathbf{S}_C)$, where $\mathbf{S}_t = (\mathbf{s}_t^1, ..., \mathbf{s}_t^N) \in \mathbb{R}^{N \times D}$ is the set of D-dimensional object embeddings parsed from image t, and modeling the object dynamics and interactions to predict future object states and video frames.

3.1. Scene Parsing

We employ SAVi [5], a recursive encoder-decoder model with a state composed of N permutation-invariant object slots, to parse the seed video frames $\boldsymbol{\mathcal{X}}_{1:C}$ into their object components $\boldsymbol{\mathcal{S}}_{1:C}$, as depicted in Fig. 2. The slots $\mathbf{S}_0$ are randomly initialized and recursively refined to bind to the objects in the video frames. At time step t, we encode the input frame into multiple feature maps that represent semantic high-level information of the input. These feature maps are fed to a Slot Attention [4] corrector, which updates the previous slots based on visual features from the current frame. Slot Attention performs cross-attention with the attention coefficients normalized over the slot dimension, thus encouraging the slots to compete to represent parts of the input, and then updates the slot representations using a Gated Recurrent Unit [25] (GRU). Namely, Slot Attention updates the previous slots $\mathbf{S}_{t-1}$ by:

$$\mathbf{U}_t = \underset{\mathbf{Q}}{\mathrm{softmax}}(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{D}})\mathbf{V}, \qquad \mathbf{S}_t = \mathbf{S}_{t-1} + \mathrm{GRU}(\mathbf{U}_t, \mathbf{S}_{t-1}), \quad (1)$$

where $\mathbf{K}$ and $\mathbf{V}$ are linear projections of the features maps, and $\mathbf{Q}$ is a linear projection of the previous slots. The output of this module is a set of object slots $\mathbf{S}_t$, representing the objects of the input frame.



(a) Transformer [26] (b) OCVP-Seq (c) OCVP-Par

Fig. 3: Transformer predictors employed in our object-centric video prediction framework. The transformer predictor (3a) jointly processes all object slots using masked self-attention, whereas our OCVP modules decouple the processing of temporal dynamics and object interactions using specialized attention blocks, i.e., temporal and relational attention, sequentially (3b) or in parallel (3c).

3.2. Object-Centric Video Predictor Modules

Given the object slots $(\mathbf{S}_1, ..., \mathbf{S}_C)$, the object-centric predictor module models the temporal dynamics and object interactions to forecast the subsequent object states $\hat{\mathbf{S}}_{C+1}$ while maintaining temporally consistent and interpretable object representations.

To this end, we propose two novel object-centric video prediction (OCVP) transformer modules, depicted in Fig. 3, which are explicitly designed to decouple the modeling of the temporal and object dimensions by employing two specialized multi-head self-attention mechanisms, i.e., *temporal attention* and *relational attention*.

Temporal Attention: In the temporal attention block, each object slot is updated by aggregating information from that same object up to the current time step, without modeling interactions between distinct objects. Given object slots $(\mathbf{S}_1, ..., \mathbf{S}_C)$, these are rearranged into temporal histories $(\mathbf{S}^1, ..., \mathbf{S}^N)$, where $\mathbf{S}^n = (\mathbf{s}_1^n, ..., \mathbf{s}_C^n) \in \mathbb{R}^{C \times D}$ denotes the temporal history of the n-th object slot up to the current time step C. Following the transformer design [26], slots in the history are first linearly mapped into key $\mathbf{K}$, query $\mathbf{Q}$ and value $\mathbf{V}$ embeddings via learnable projections and then jointly processed via dot-product attention:

$$\mathrm{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \underset{\mathbf{K}}{\mathrm{softmax}}(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{D}})\mathbf{V}, \quad (2)$$

whose outcome yields slot representations that have been updated according to their temporal history.

Relational Attention: In contrast, relational attention models the multi-object interactions and relationships by jointly processing all slots from the same time step. The computation of this module is analogous to the one of temporal attention, with the notable difference that, instead of processing temporal histories from each independent object, it jointly processes all object slots from the same time step. The outcome of this block yields object slots that have been updated according to their relations with other objects.

Table 1: Video prediction quantitative evaluation. Best two results are highlighted in boldface and underlined. We rank the methods according to their performance, averaged across all datasets and metrics.

		$\textbf{Obj3D}_{5\rightarrow5}$						$\textbf{MOVi-A}_{6\rightarrow8}$						
		NumPreds = 15			**NumPreds = 25**			**NumPreds = 8**			**NumPreds = 18**			**Mean**
		PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	**Rank↓**
Object Agnostic	CopyLast	25.40	0.911	0.060	23.98	0.876	0.093	23.48	0.790	0.128	22.83	0.778	0.172	6.33
	ConvLSTM [27]	**34.32**	0.929	0.046	28.95	0.849	0.112	27.12	0.877	0.248	23.85	<u>0.831</u>	0.337	4.83
	PhyDNet [28]	30.08	0.921	0.034	28.21	**0.892**	0.053	**28.36**	**0.893**	0.168	**26.49**	**0.870**	0.199	3.5
Object Centric	LSTM [29]	31.13	0.900	0.039	28.83	0.849	0.071	27.82	0.827	0.089	26.14	0.784	0.133	5.17
	Transformer [24]	32.89	<u>0.931</u>	**0.025**	<u>30.87</u>	**0.892**	**0.041**	27.97	0.832	0.085	26.18	0.785	0.134	2.92
	OCVP-Seq	<u>33.10</u>	**0.932**	**0.025**	**30.93**	0.891	**0.041**	<u>27.99</u>	<u>0.834</u>	**0.082**	26.24	0.789	**0.127**	**1.83**
	OCVP-Par	32.99	<u>0.931</u>	**0.025**	30.85	0.890	0.043	<u>27.99</u>	0.832	<u>0.083</u>	**26.31**	0.788	**0.127**	<u>2.67</u>

OCVP Modules: Employing our structured attention blocks, we propose two distinct OCVP modules. *OCVP-Seq* (Fig. 3b) cascades both attention blocks, thus iteratively refining the slots with relational and temporal information, respectively; whereas *OCVP-Par* (Fig. 3c) applies both attention blocks in parallel and sums their outcomes to aggregate the temporal and object information.

3.3. Video Rendering

We use the SAVi decoder to independently render an object image and a mask from each individual predicted slot. The object masks are normalized across slots using a softmax, and then combined with the rendered objects via a weighted sum to generate video frames:

$$\mathbf{o}_t^n, \mathbf{m}_t^n = f_{\text{dec}}(\hat{\mathbf{s}}_t^n), \;\; \tilde{\mathbf{m}}_t^n = \underset{N}{\text{softmax}}(\mathbf{m}_t^n), \; \hat{\mathbf{X}}_t = \sum_{n=1}^{N} \mathbf{o}_t^n \odot \tilde{\mathbf{m}}_t^n. \quad (3)$$

3.4. Model Training and Inference

Our object-centric video prediction approach is trained in two separate stages. First, SAVi is trained to parse video frames into their object components. Then, given the pretrained SAVi model, we train the predictor to forecast future object states.

More precisely, we parse the seed frames $\boldsymbol{\mathcal{X}}_{1:C}$ into their object components $\boldsymbol{\mathcal{S}}_{1:C}$ using the pretrained SAVi encoder. Then, the extracted object slots are fed to the predictor module in order to forecast the subsequent object states $\hat{\mathbf{S}}_{C+1}$, which are rendered into object images and masks using the pretrained SAVi decoder and then combined to generate the subsequent video frame $\hat{\mathbf{X}}_{C+1}$. We apply this process in an autoregressive way to predict, conditioned on all previous parsed and predicted object slots, the subsequent T object slots $\hat{\boldsymbol{\mathcal{S}}}_{C+1:C+T}$ and video frames $\hat{\boldsymbol{\mathcal{X}}}_{C+1:C+T}$.

We train the predictor by minimizing the combined loss:

$$\mathcal{L} = \mathcal{L}_I(\boldsymbol{\mathcal{X}}, \hat{\boldsymbol{\mathcal{X}}}) + \mathcal{L}_O(\boldsymbol{\mathcal{S}}, \hat{\boldsymbol{\mathcal{S}}}), \quad (4)$$

$$\mathcal{L}_I = \frac{1}{T} \sum_{t=1}^{T} ||\hat{\mathbf{X}}_{C+t} - \mathbf{X}_{C+t}||_2^2, \quad (5)$$

$$\mathcal{L}_O = \frac{1}{T \cdot N} \sum_{t=1}^{T} \sum_{n=1}^{N} ||\hat{\mathbf{s}}_{C+t}^n - \mathbf{s}_{C+t}^n||_2^2, \quad (6)$$

where $\mathcal{L}_I$ measures the future frame prediction error, and $\mathcal{L}_O$ measures the error when predicting future object slots.

4. EXPERIMENTS

4.1. Datasets and Experimental Details

We evaluate our object-centric video prediction framework on two distinct object-centric datasets, namely Obj3D and MOVi-A.

Table 2: Object-centric prediction evaluation on MOVi-A.

	NumPreds = 8		**NumPreds = 18**		**Mean**
	ARI↑	mIoU↑	ARI↑	mIoU↑	**Rank↓**
LSTM [29]	0.619	0.578	0.431	0.500	4
Transformer [24]	**0.640**	0.585	**0.452**	0.508	2
OCVP-Seq	0.632	**0.588**	<u>0.443</u>	**0.511**	**1.75**
OCVP-Par	<u>0.636</u>	<u>0.586</u>	0.435	**0.511**	2

Obj3D [23] contains 2920 train and 200 test synthetic sequences in which a moving ball collides with static 3D geometric objects (e.g. spheres or cubes), setting them in motion. We train our predictors using $C = 5$ seed frames to predict the subsequent $T = 5$ frames.

MOVi-A [5] contains 9703 train and 250 test sequences with up to ten objects similar to those in Obj3D, but with more complex dynamics, occlusions and collisions. We train our predictors using $C = 6$ seed frames to predict the subsequent $T = 8$ video frames.

Following previous works [24, 4], we evaluate the visual quality of predicted video frames using video prediction metrics (PSNR, SSIM and LPIPS), and evaluate the ability to model object dynamics by measuring the ARI and mIoU between ground-truth instance segmentation and the forecasted object masks.

4.2. Evaluation

We evaluate our object-centric prediction framework using different predictor modules, namely LSTM [29], Transformer [24], and our two proposed OCVP modules, for different prediction horizons (*NumPreds*). We compare our object-centric approach with two existing object-agnostic prediction models: ConvLSTM [27] and PhyDNet [28]. Additionally, we include a CopyLast baseline that naively copies the last seed frame.

Video Prediction: In Table 1, we quantitatively report the visual quality of the predicted frames. We aggregate scores by measuring how each model ranks compared to the others, and average ranks across metrics and prediction horizons. Our OCVP modules achieve the overall best performance, with OCVP-Seq widely outranking all other models, and outperform the object-agnostic baselines and other predictors. This is most notable on the perceptual LPIPS metric, where our OCVP modules achieve perceptually realistic predictions, whereas object-agnostic baselines are even outperformed by the simple CopyLast baseline for longer prediction horizons. Fig. 4 depicts qualitative comparisons on Obj3D (Fig. 4a) and MOVi-A (Fig. 4b) between two object-agnostic baselines (ConvLSTM and PhyDNet), and our object-centric prediction framework using two distinct predictors. Object-agnostic baselines, which lack explicit knowledge about objects, lead to predictions of very low quality after a few time steps; whereas object-centric models, specially our OCVP-Seq, achieve accurate and temporally consistent predictions.

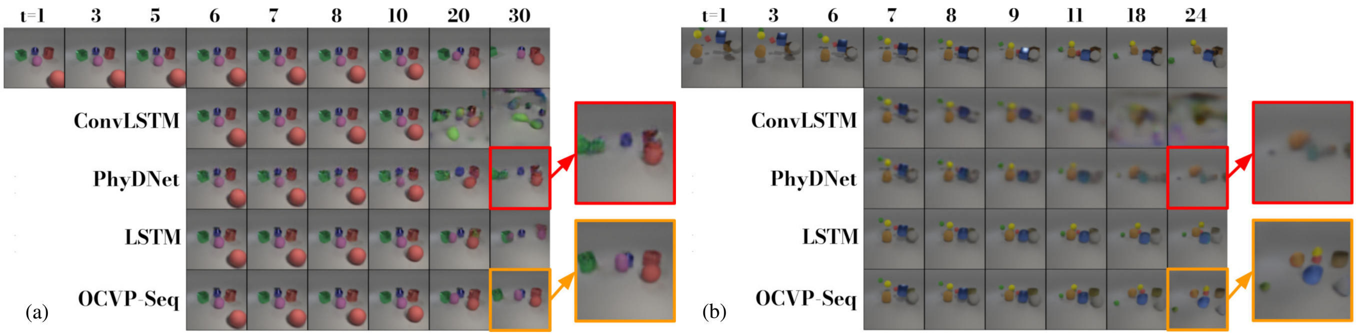


Fig. 4: Qualitative comparison on Obj3D (4a) and MOVi-A (4b). We display three seed and six target frames (top row), as well as predictions from four different models. OCVP-Seq achieves the most accurate predictions among the compared methods.

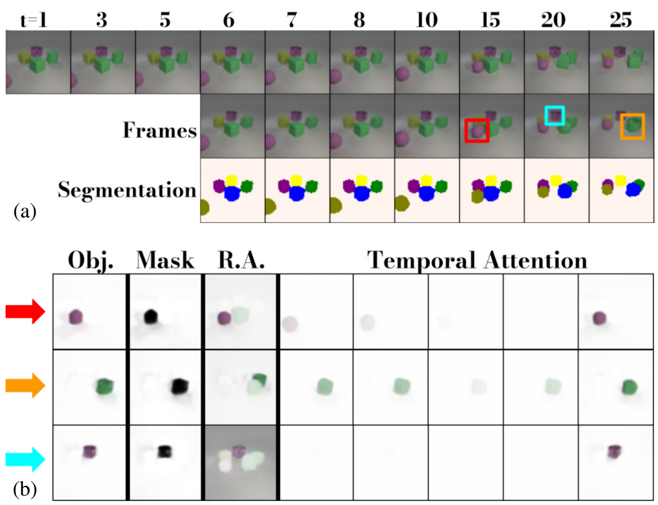


Fig. 5: a) Qualitative segmentation forecasting results on Obj3D using OCVP-Seq. b) Interpretable representations for three object predictions. We visualize the predicted object images and masks, and their relational (R.A.) and temporal attention masks.

Object-Centric Evaluation: We evaluate the ability of different predictors to model object dynamics. We convert the predicted object masks into segmentation maps by computing the `argmax` over all objects, and compare them in Table 2 with ground-truth segmentation masks on MOVi-A. Our OCVP-Seq module achieves the best overall performance, while all transformer-based predictors outperform the LSTM module. Fig. 5 depicts frames and segmentations predicted by OCVP-Seq for an Obj3D sequence, as well as interpretable intermediate representations for three object predictions. We display the predicted object images and masks, and illustrate the corresponding relational (R.A.) and temporal attention masks. For dynamic and interacting objects (purple ball or green cube), relational attention focuses on neighboring and colliding objects, whereas temporal attention focuses mostly on the current input, but also attends to multiple previous time steps. In contrast, for static objects (purple cube), relational attention divides the attention between the corresponding object and the background, whereas temporal attention focuses only on the last time step.

4.3. Ablation Study

We train several modified versions of our OCVP-Seq predictor on Obj3D in order to understand how different design choices affect the video prediction performance. The results are reported in Table 3.

Table 3: OCVP-Seq Ablation Study on Obj3D. We investigate the effect of residual connections, teacher forcing (TF), skipping the slots from the first frame, and the number of predictor layers.

	Num Preds = 15			Num Preds = 25		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
w/o Residual	31.13	0.906	0.032	29.35	0.862	0.056
w/ Residual	**33.10**	**0.932**	**0.025**	**30.93**	**0.891**	**0.041**
w/o TF	**33.10**	**0.932**	**0.025**	**30.93**	**0.891**	**0.041**
w/ TF	32.62	0.923	**0.025**	27.67	0.876	0.047
w/o Skip First	33.10	**0.932**	**0.025**	30.93	**0.891**	**0.041**
w/ Skip First	**33.14**	**0.932**	**0.025**	**30.97**	**0.891**	**0.041**
Layers = 2	33.04	0.931	**0.025**	30.78	0.889	0.043
Layers = 4	**33.10**	**0.932**	**0.025**	**30.93**	**0.891**	**0.041**
Layers = 6	32.79	0.929	0.026	30.73	0.889	0.043

Residual Connection: Using a residual predictor ($\hat{\mathbf{S}}_t = \hat{\mathbf{S}}_{t-1} + f_{\text{pred}}(\hat{\mathbf{S}}_{t-1})$) widely outperforms its counterpart. We conclude that a residual predictor refines the input slot representations, allowing for more temporally consistent predictions.

Teacher Forcing: When training with teacher forcing, the predictor module does not learn to handle its own imperfect predictions, thus leading to poor performance for longer prediction horizons.

Skip First: Skipping the often imperfect slots from the first time step does not to help the model to predict better frames, often leading to the same SSIM and LPIPS scores.

Num. Layers: Using four predictor modules leads to the best video prediction performance, especially for longer prediction horizons.

5. CONCLUSION

We presented a framework for object-centric video prediction, which decomposes video frames into object components and models the object dynamics and their interactions to generate future video frames. To achieve an improved prediction performance, we proposed two OCVP transformer predictor modules, which decouple the processing of temporal dynamics and object interactions. In our experiments, we showed how our prediction framework using an OCVP module outperforms object-agnostic baselines, while also learning interpretable and temporally consistent object representations. Our models also accurately forecast the segmentation of the scene. In future work, we will use our rich spatio-temporal object representations for action and anticipative behavior planning. We release code to reproduce our results in our project website[2].

[2] `https://sites.google.com/view/ocvp-vp`

6. REFERENCES

[1] Scott P Johnson, "Object perception," in *Oxford Research Encyclopedia of Psychology*. 2018.

[2] Klaus Greff, Sjoerd Van Steenkiste, and Jürgen Schmidhuber, "On the binding problem in artificial neural networks," *arXiv preprint arXiv:2012.05208*, 2020.

[3] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio, "Toward causal representation learning," *Proceedings of the IEEE*, vol. 109, no. 5, pp. 612–634, 2021.

[4] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf, "Object-centric learning with slot attention," in *International Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

[5] Thomas Kipf, Gamaleldin F. Elsayed, Aravindh Mahendran, Austin Stone, Sara Sabour, Georg Heigold, Rico Jonschkowski, Alexey Dosovitskiy, and Klaus Greff, "Conditional Object-Centric Learning from Video," in *International Conference on Learning Representations (ICLR)*, 2022.

[6] 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.

[7] Zhen He, Jian Li, Daxue Liu, Hangen He, and David Barber, "Tracking by animation: Unsupervised learning of multi-object attentive trackers," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[8] Jindong Jiang, Sepehr Janghorbani, Gerard De Melo, and Sungjin Ahn, "SCALOR: Generative world models with scalable object representations," in *International Conference on Learning Representations (ICLR)*, 2019.

[9] Angel Villar-Corrales and Sven Behnke, "Unsupervised image decomposition with phase-correlation networks," in *International Conference on Computer Vision Theory and Applications (VISAPP)*, 2022.

[10] 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," in *International Conference on Learning Representations (ICLR)*, 2020.

[11] Adam R Kosiorek, Hyunjik Kim, Ingmar Posner, and Yee Whye Teh, "Sequential attend, infer, repeat: Generative modelling of moving objects," in *International Conference on Neural Information Processing Systems (NeurIPS)*, 2018.

[12] Zhangyang Gao, Cheng Tan, Lirong Wu, and Stan Z Li, "SimVP: Simpler yet better video prediction," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[13] Hsu-kuang Chiu, Ehsan Adeli, and Juan Carlos Niebles, "Segmenting the future," *IEEE Robotics and Automation Letters*, vol. 5, no. 3, pp. 4202–4209, 2020.

[14] Emily Denton and Rob Fergus, "Stochastic video generation with a learned prior," in *International Conference on Machine Learning (ICML)*, 2018.

[15] Angel Villar-Corrales, Ani Karapetyan, Andreas Boltres, and Sven Behnke, "MSPred: Video prediction at multiple spatio-temporal scales with hierarchical recurrent networks," in *British Machine Vision Conference (BMVC)*, 2022.

[16] Yunbo Wang, Haixu Wu, Jianjin Zhang, Zhifeng Gao, Jianmin Wang, Philip Yu, and Mingsheng Long, "PredRNN: A recurrent neural network for spatiotemporal predictive learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, pp. 1–1, 2022.

[17] Dirk Weissenborn, Oscar Täckström, and Jakob Uszkoreit, "Scaling autoregressive video models," in *International Conference on Learning Representations (ICLR)*, 2020.

[18] Ruslan Rakhimov, Denis Volkhonskiy, Alexey Artemov, Denis Zorin, and Evgeny Burnaev, "Latent video transformer," in *International Conference on Computer Vision Theory and Applications (VISAPP)*, 2021.

[19] Hafez Farazi, Jan Nogga, and Sven Behnke, "Local frequency domain transformer networks for video prediction," in *International Joint Conference on Neural Networks (IJCNN)*, 2021.

[20] Yi-Fu Wu, Jaesik Yoon, and Sungjin Ahn, "Generative video transformer: Can objects be the words?," in *International Conference on Machine Learning (ICML)*, 2021.

[21] Yufei Ye, Maneesh Singh, Abhinav Gupta, and Shubham Tulsiani, "Compositional video prediction," in *IEEE/CVF International Conference on Computer Vision (CVPR)*, 2019.

[22] Antonia Creswell, Rishabh Kabra, Chris Burgess, and Murray Shanahan, "Unsupervised object-based transition models for 3D partially observable environments," *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[23] Zhixuan Lin, Yi-Fu Wu, Skand Peri, Bofeng Fu, Jindong Jiang, and Sungjin Ahn, "Improving generative imagination in object-centric world models," in *International Conference on Machine Learning (ICML)*, 2020.

[24] Ziyi Wu, Nikita Dvornik, Klaus Greff, Thomas Kipf, and Animesh Garg, "SlotFormer: Unsupervised visual dynamics simulation with object-centric models," *arXiv preprint arXiv:2210.05861*, 2022.

[25] 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," in *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

[26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[27] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

[28] Vincent Le Guen and Nicolas Thome, "Disentangling physical dynamics from unknown factors for unsupervised video prediction," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[29] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.