DeepMind

SIMONe: View-Invariant, Temporally-Abstracted Object Representations via Unsupervised Video Decomposition

Rishabh Kabra, Daniel Zoran, Goker Erdogan, Loic Matthey,
Antonia Creswell, Matthew Botvinick, Alexander Lerchner, Christopher P. Burgess

Desiderata

- Decomposition of multi-object scenes from RGB videos without supervision.
- Handles changing camera pose; simultaneous inference of *scene contents* and *viewpoint* from correlated views (i.e. sequential observations of a moving agent).
- Learning of structure across diverse scene instances (i.e. procedurally sampled contents);
- Object representations which summarize static object attributes like color or shape, view-dissociated properties like position or size, as well as time-abstracted trajectory features like direction of motion;
- No explicit assumptions of 3D geometry, no explicit dynamics model, no specialized renderer, and few a priori modeling assumptions about the objects being studied; and
- Simple, scalable modules (for inference and rendering) to enable large-scale use.

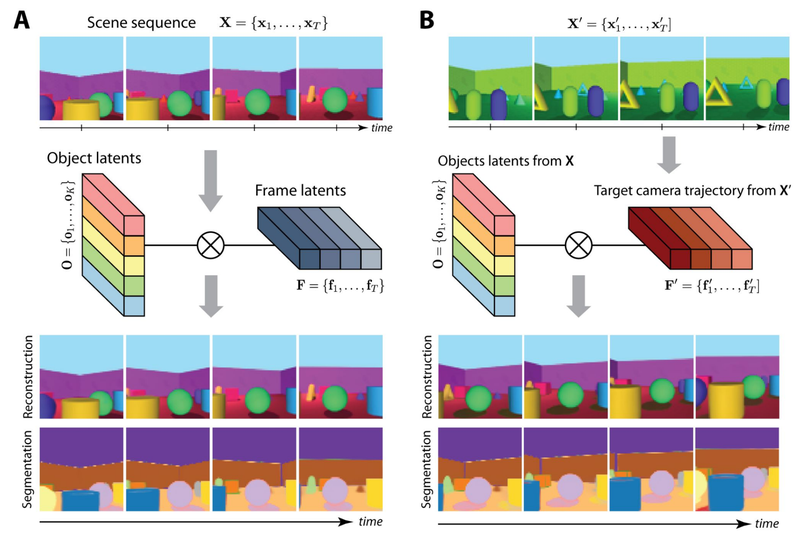
Related Work

1) **Scene decomposition models** (e.g. MONet, IODINE, GENESIS, SPACE, AIR): Decompose images into component objects and object features.
2) **Multi-view scene rendering models** (e.g. GQN, SRNs, NeRF): Decompose allocentric scene structure and a variable viewpoint.
3) **Simultaneous localization and mapping:** Infer scene representations by exploring environment. Object-centric and learning based solutions are active research areas.

Overview

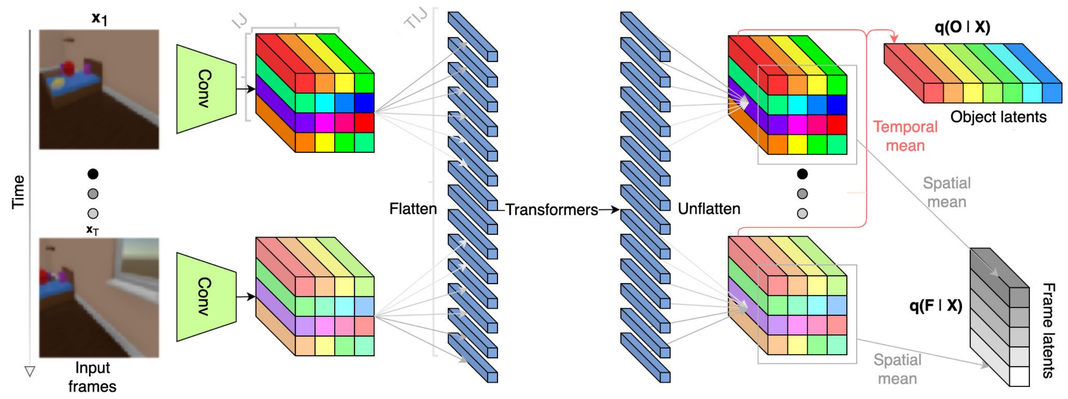
A. SIMONe decomposes and factorizes a sequence **X** into *scene content* ("**object latents**", constant across a sequence) and *view/global content* ("**frame latents**", one per frame) without supervision. Spatio-temporal inference naturally allows stable object tracking.

B. Object latents inferred from a sequence **X** can be re-composed with the frame latents of a different (i.i.d.) sequence **X'** to generate a consistent rendering of the same scene (i.e. objects and their properties, relative arrangements, and segmentation assignments) from entirely different viewpoints.



Model

- **Latent structure**: K object latents are invariant across all frames in the sequence, and expected to summarize information over time. T frame latents are invariant across objects but vary with time.
- **Generative process**: Each pixel is modeled as a GMM with K components. The decoder is queried for a specific time-step t and pixel location $\mathbf{l}_i$ and is a simple MLP.
- **Inference**: Spatio-temporal data is jointly processed via transformers. Feature maps attend to one another across space and time.
- **Loss**: The model is trained using a β-weighted ELBO with two KL terms and a reconstruction loss.

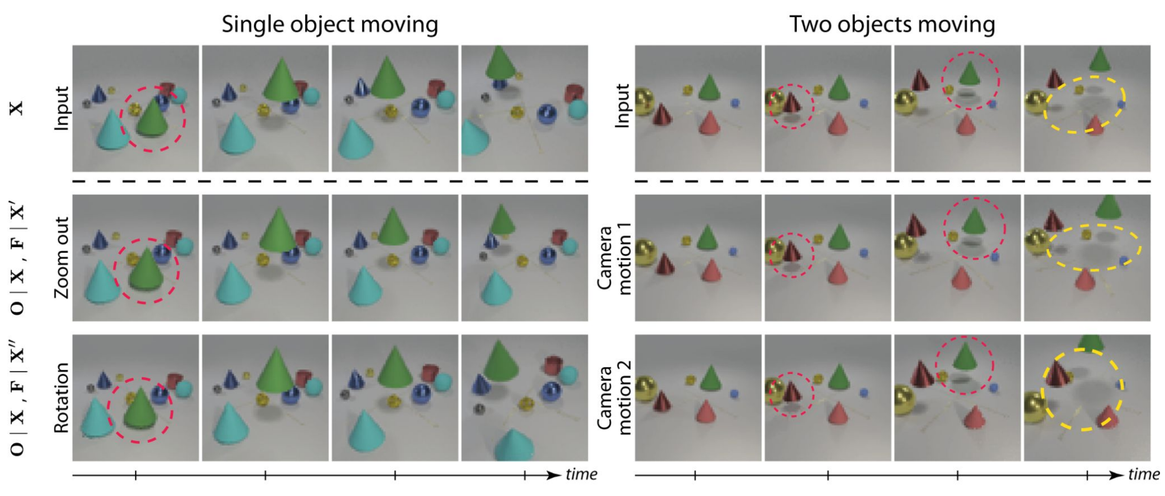


Segmentation performance

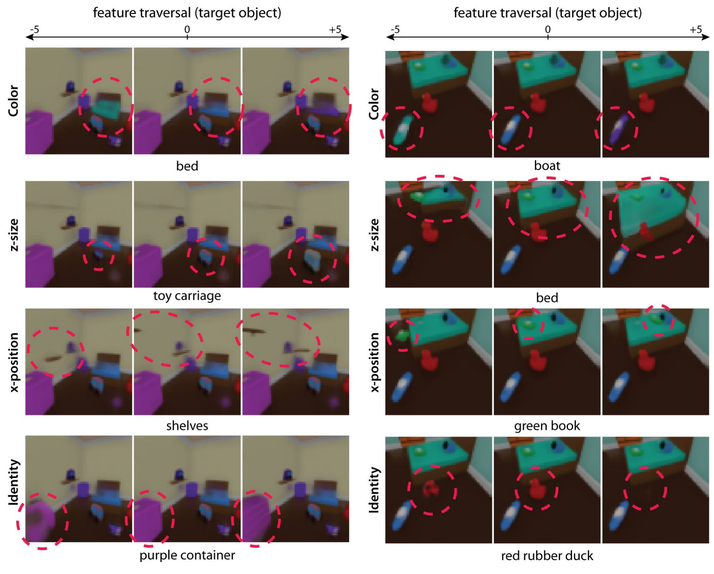
We evaluate models on still images via Static ARI-F. We also evaluate them across space and time via Video ARI-F, penalizing models which fail to track objects stably.

	Static ARI-F			Video ARI-F			
	MONet	SA	S-IODINE	MONet	SA	S-IODINE	SIMONe
Objects Room 9	0.886 (±0.061)	0.784 (±0.138)	0.695 (±0.007)	0.865 (±0.007)	0.066 (±0.014)	0.673 (±0.0.002)	0.936 (±0.010)
CATER	0.937 (±0.004)	0.923 (±0.076)	0.728 (±0.032)	0.412 (±0.012)	0.073 (±0.006)	0.668 (±0.033)	0.918 (±0.036)
Playroom	0.647 (±0.012)	0.653 (±0.024)	0.439 (±0.009)	0.442 (±0.010)	0.059 (±0.002)	0.356 (±0.006)	0.800 (±0.043)

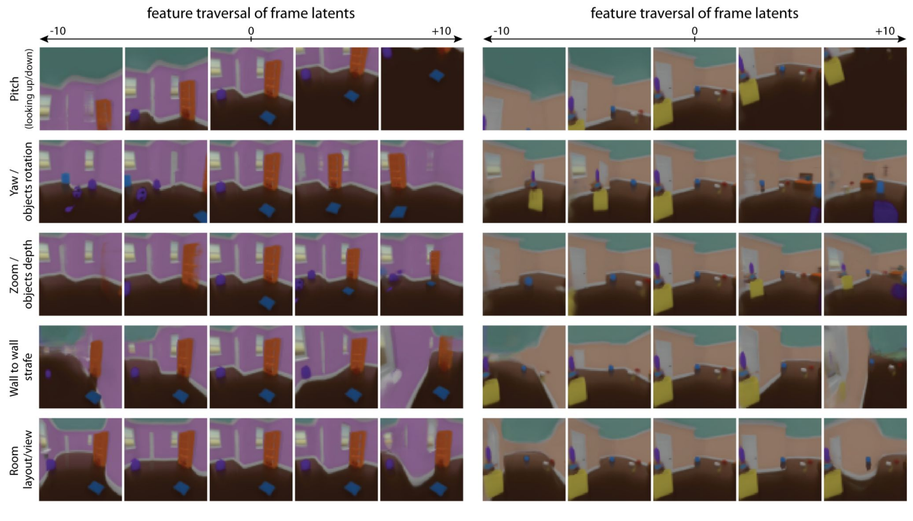
Temporal abstraction



Object representations



View representations



Quantitative analysis

Predicting camera pose from frame latents:

	Linear($\mathbf{f}_t$)	MLP($\mathbf{f}_t$)	MLP($\mathbf{o}_1$, ..., $\mathbf{o}_K$)
Camera location	0.832	0.944	-0.017
Camera orientation (Rodrigues)	0.800	0.954	0.272

Predicting object dynamics from object latents:

	MLP($\mathbf{o}_k$)	MLP($\mathbf{o}_k, t$)	MLP($\mathbf{o}_k, \mathbf{f}_t, t$)	MLP($\{\mathbf{o}_j : j \neq k\}$)
Trained on all objects	0.710 ± 0.006	0.871 ± 0.006	0.876 ± 0.003	-0.062 ± 0.006
Trained on moving objects	0.724 ± 0.007	0.894 ± 0.004	0.898 ± 0.005	-0.022 ± 0.025