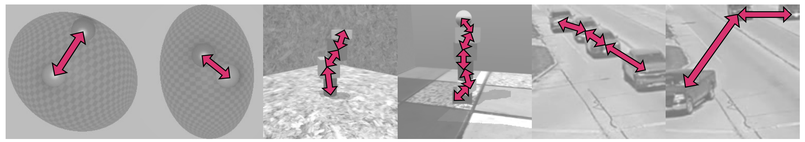
RELATE: Physically Plausible Multi-Object Scene Synthesis Using Structured Latent Spaces

Sébastien Ehrhardt, Oliver Groth, Áron Monszpart, Martin Engelcke,
Ingmar Posner, Niloy J. Mitra, Andrea Vedaldi

NIANTIC · Adobe Research · UCL · UNIVERSITY OF OXFORD

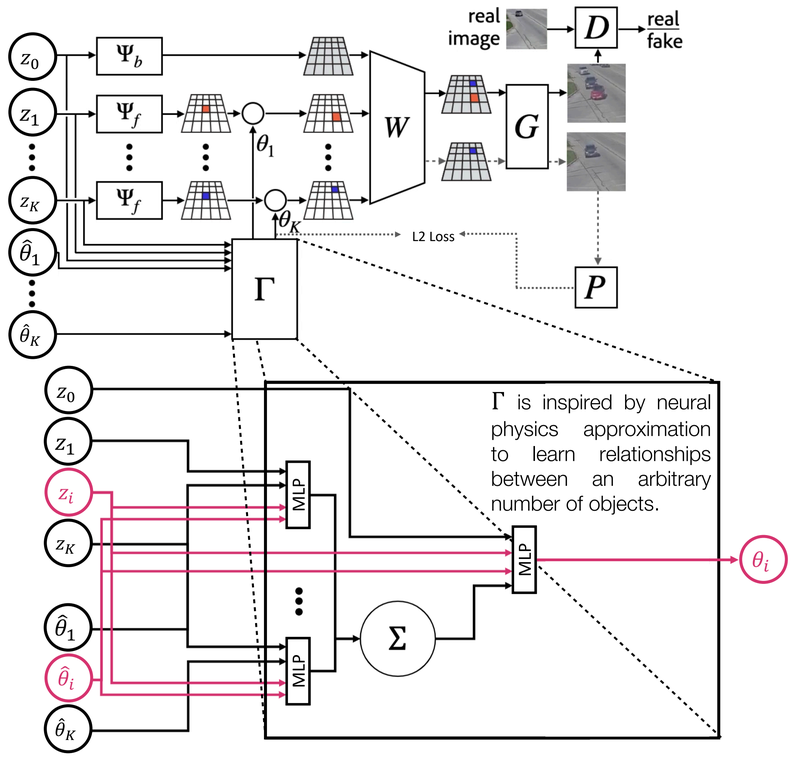
General Problem

Most component-wise image generation techniques assume **independence** between individual objects. We overcome this limitation by proposing a model which **learns interactions between objects in images**.



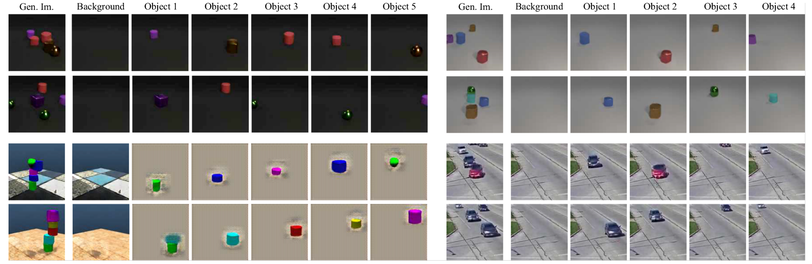
Our Method

We start from a 2D version of BlockGAN [1] where individual scene components – **background and foreground objects – are represented by appearance z_0 and pairs of appearance and pose vectors $(z_i, \hat{\theta}_i)$, respectively**. Each appearance vector is converted to a tensor by a module Ψ. We augment the model with a relationship module Γ that adjusts the independently sampled $\hat{\theta}_i$ to enhance physical plausibility of the scene. The structured scene tensor W is finally transformed by the generator network G to produce an image.



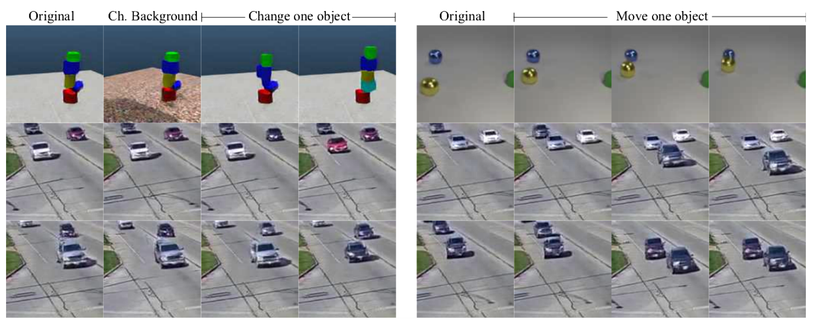
Compositional Generation

RELATE represents a scene **component-wise**. Below are renderings of the latent vectors of generated images from all experimentation datasets.

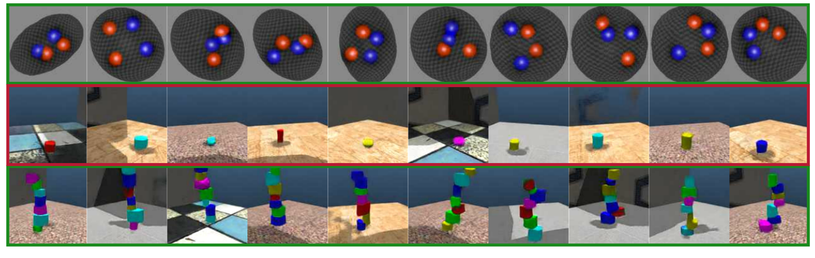


Scene Editing

RELATE is also amenable to scene editing. The model allows to edit a scene's **background** or an object's **appearance or position**. The car examples also exemplify how the change of the background also affects the rendering of the shadows.



RELATE allows for immediate generalization to **fewer or more objects** by simply *adding* **or** *removing* **components in latent space**. Γ ensures global spatial consistency.



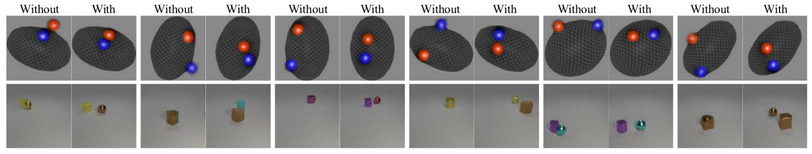
Generation Results

As measured by **FID (lower is better)**, RELATE **outperforms SOTA object-centric models** in image generation. Its performance is also **on par with monolithic GAN baselines** such as DRAGAN while generating images with higher (128x128) resolution.

		CLEVR5-vbg	CLEVR	ShapeStacks	RealTraffic
monolithic latent spaces	DCGAN [2]	361.8	247.8	197.6	47.6
	DRAGAN [3]	84.4	108.0	**57.2**	**38.8**
object-centric latent spaces	GENESIS [4]	169.4	151.3	233.0	167.1
	OCF [5]	83.1	N/A	N/A	N/A
	BlockGAN2D [1]	53.3	78.1	99.3	57.9
	RELATE (ours)	**36.4**	**62.9**	**95.7**	**42.0**

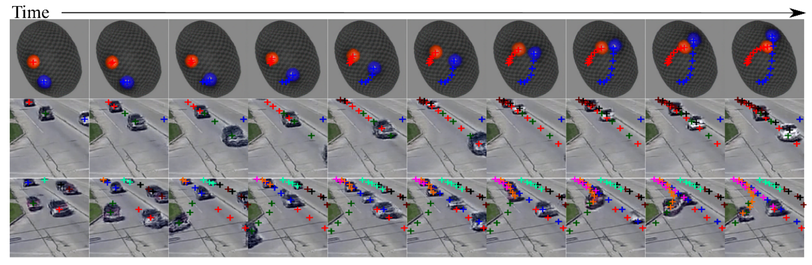
Ablation

Our module Γ captures the relation **between objects and background** (top) and **other objects** (bottom), e.g. by constraining positions or resolving object intersections.



Temporal Extension

RELATE can be extended to generate video sequences. In this case, Γ models **relations over space and time** and the **discriminator** *D* operates on the **sequence level.**



[1] Thu et al, BlockGAN: Learning 3D Object-aware Scene Representations from Unlabelled Images, NeurIPS 2020.
[2] Radford et. al, Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. ICLR 2016.
[3] Kodali et. al, On Convergence and Stability of Gans. arXiv:1705.07215 2017.
[4] Engelcke et. al, Genesis: Generative Scene Inference and Sampling with Object-Centric Latent Representations. ICLR 2016.
[5] Anciukevicius et. al, Object-Centric Image Generation with Factored Depths, Locations, and Appearances. arXiv:2004.00642 2020.

Please visit our project page to find our datasets, source code, pre-trained models and video results.