Unsupervised object-centric video generation and decomposition in 3D

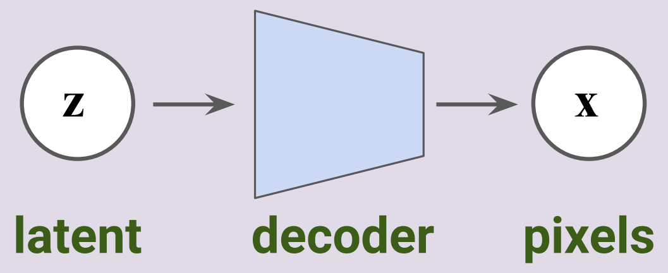
Paul Henderson
Christoph H. Lampert
IST AUSTRIA

NEURAL INFORMATION PROCESSING SYSTEMS

Generative models

Classic

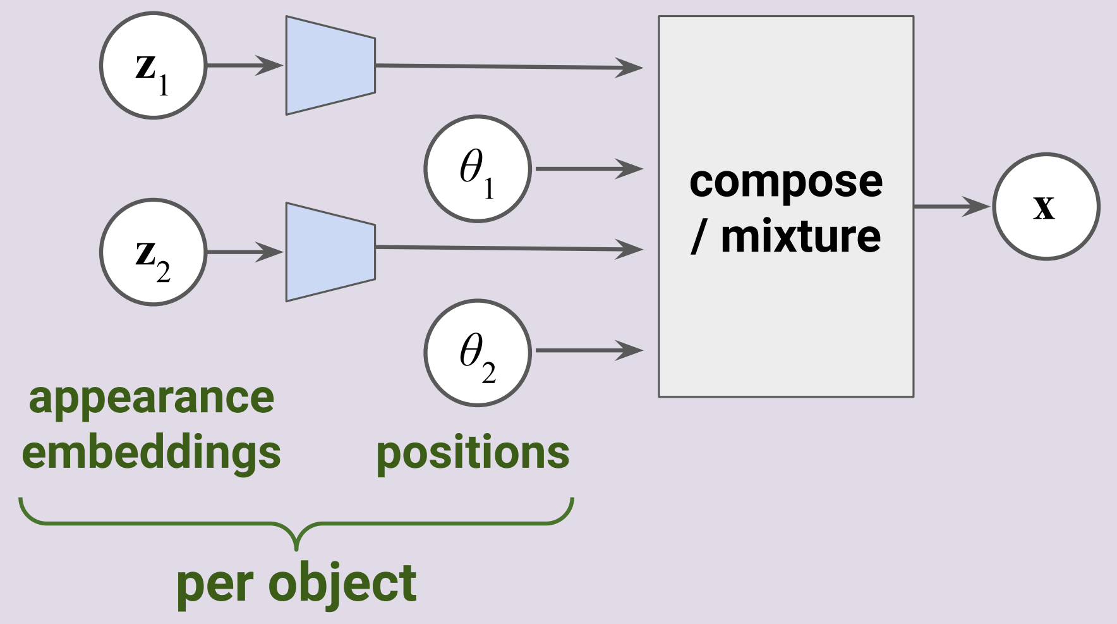
VAE [Kingma, ICLR 2014]
GAN [Goodfellow, NIPS 2014]



- single opaque latent – not interpretable
- only support generation – no inference

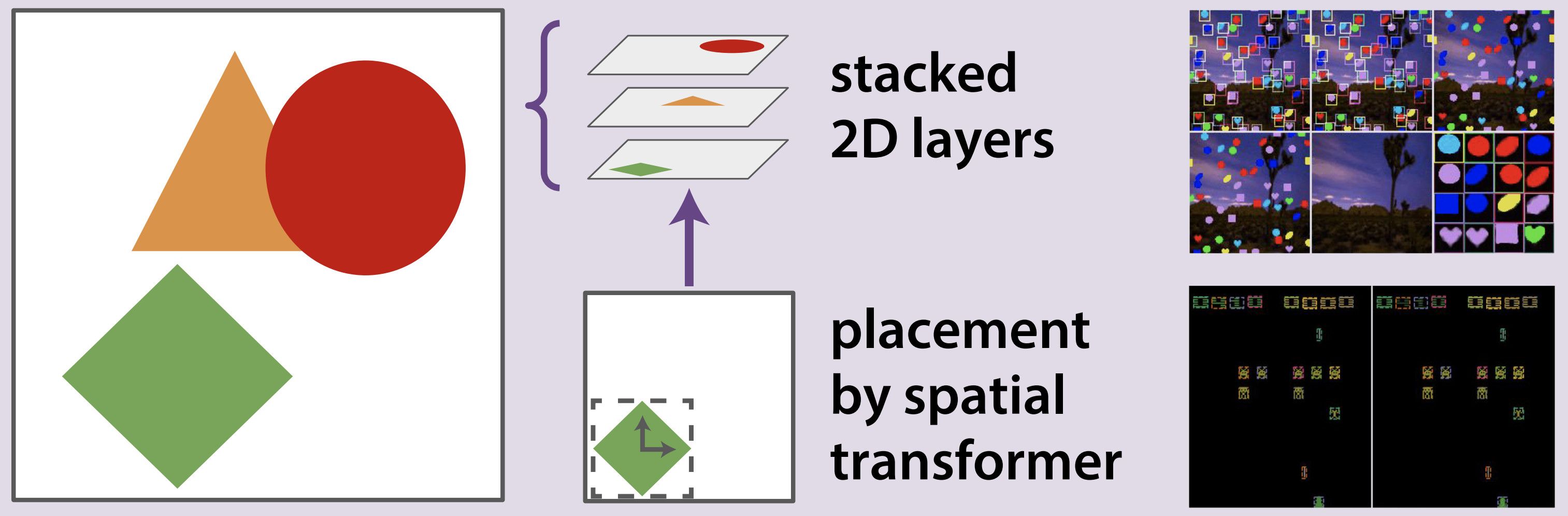
Object-centric

AIR [Eslami, NIPS 2016]
SCALOR [Jiang, ICLR 2020]
SQAIR [Kosiorek, NeurIPS 2018]
SPACE [Lin, ICLR 2020]



- structured latents – **interpretable** and **compositional**
- if learn an object appearance once, can model at any location ...i.e. appearance and (2D) location are **disentangled**
- support **inference of scene structure**: segmentation, etc. ...and this is learnt **without supervision**, just maximising the pixel likelihood

Existing 2D object-centric models



- **2D sprites,** with xy positions, scales and depth ordering
- rendering by spatial transformer + alpha blending
- do not learn a scene-level prior (e.g. collision avoidance)
- work well on videos that consist of independently-moving 2D sprites with slowly-changing appearance

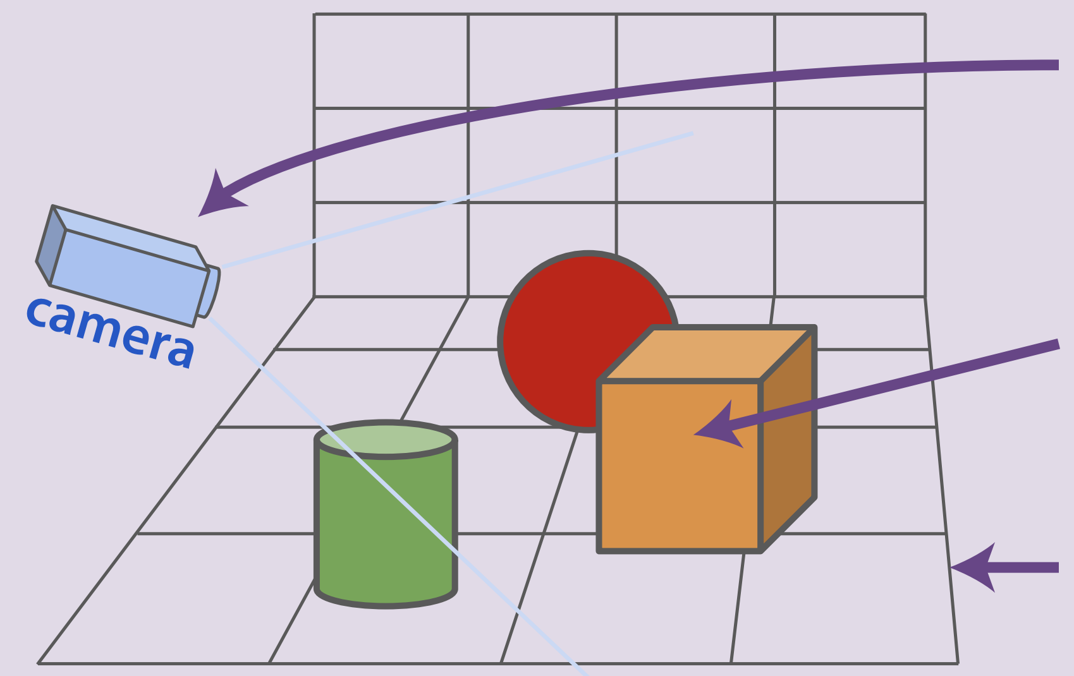
SCALOR [Jiang, ICLR 2020] • SILOT [Crawford, AAAI 2020]

Our model

Key idea

- the world is built out of **3D objects** (not 2D sprites!)

...so: **model video as view observed by a camera moving through a scene consisting of multiple 3D objects, and a 3D background**

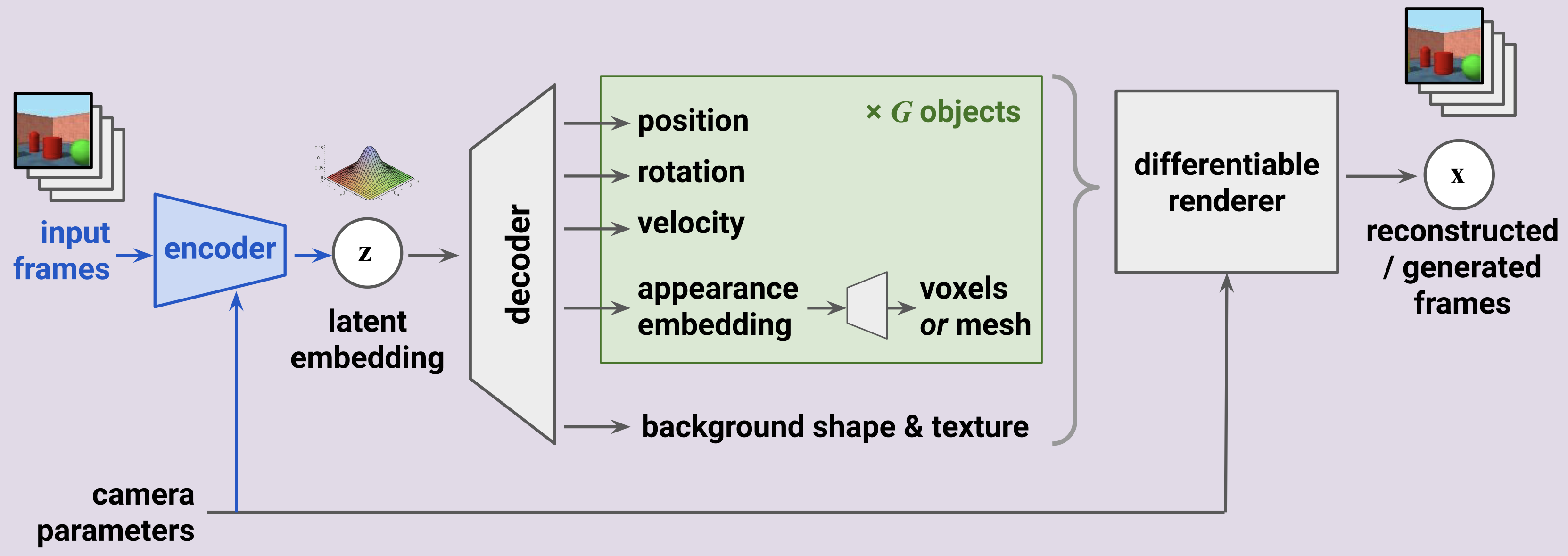


camera moves through scene capturing frames

objects = **voxels,** or **textured mesh** (we evaluate both); **static or moving**

background = textured mesh

Probabilistic model



- have a 3D **grid** of G candidate objects; each may be present or not
- **single Gaussian latent** z embeds all information about the scene
  - includes object/background appearances and motion
  - allows learning **inter-object dependencies,** e.g. avoid collisions
- decoders map z to per-object...
  - **appearance codes,** which are decoded independently to explicit 3D appearances (voxel RGBAs / mesh vertex offsets & texture)
  - **3D locations, rotations,** and **velocities**
  - binary **presence indicator**
- **differentiably render** each object, then **composite** together
  - camera parameters (extrinsic + intrinsic) treated as known
- trained like a VAE
  - add an **encoder** that maps a video to its latent z
  - maximise **ELBO** (variational bound on likelihood)

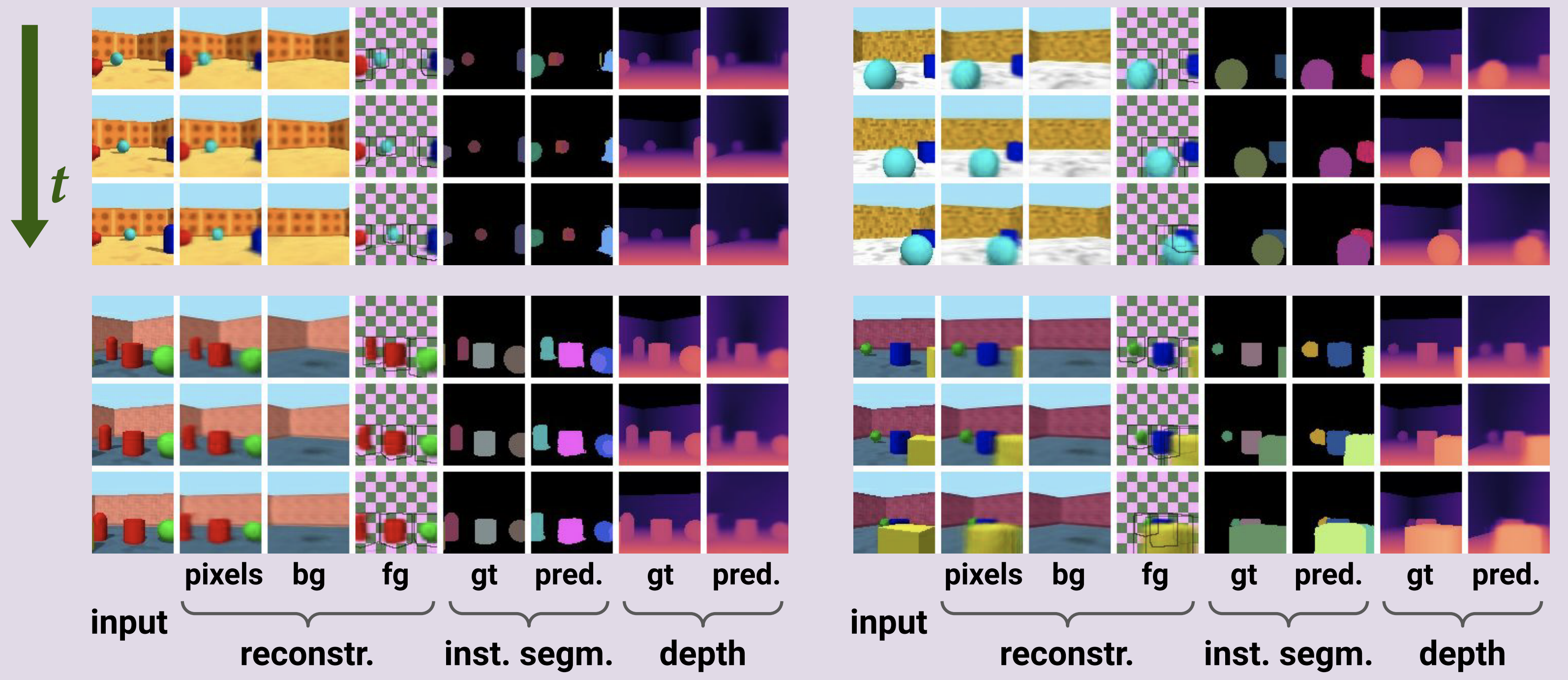
Results

more at **https://www.pmh47.net/o3v/**

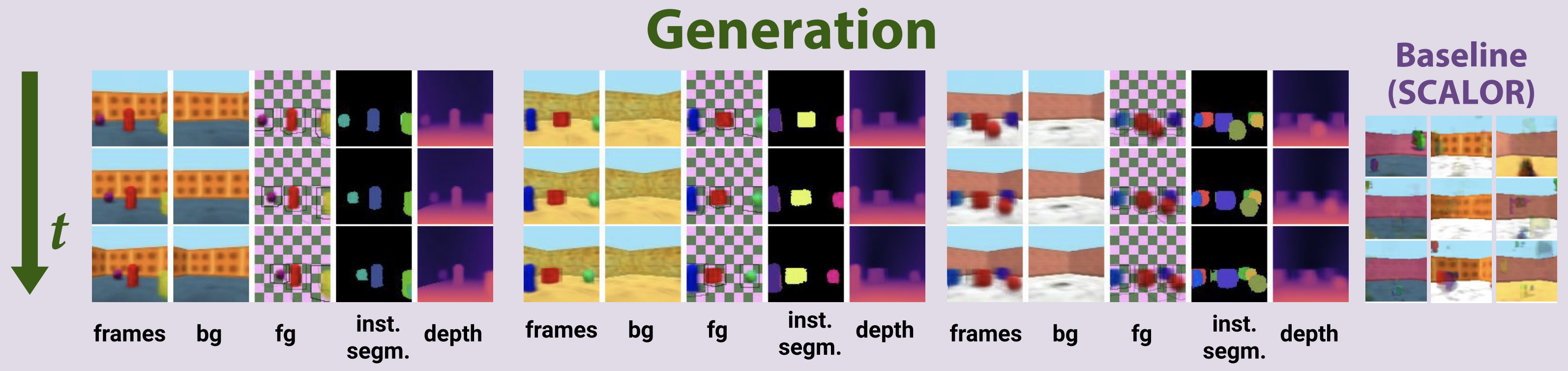
Rooms

- inspired by GQN [Eslami, Science 2018]
- 3-5 static objects, random colours

Inference (unsupervised scene decomposition)



Generation

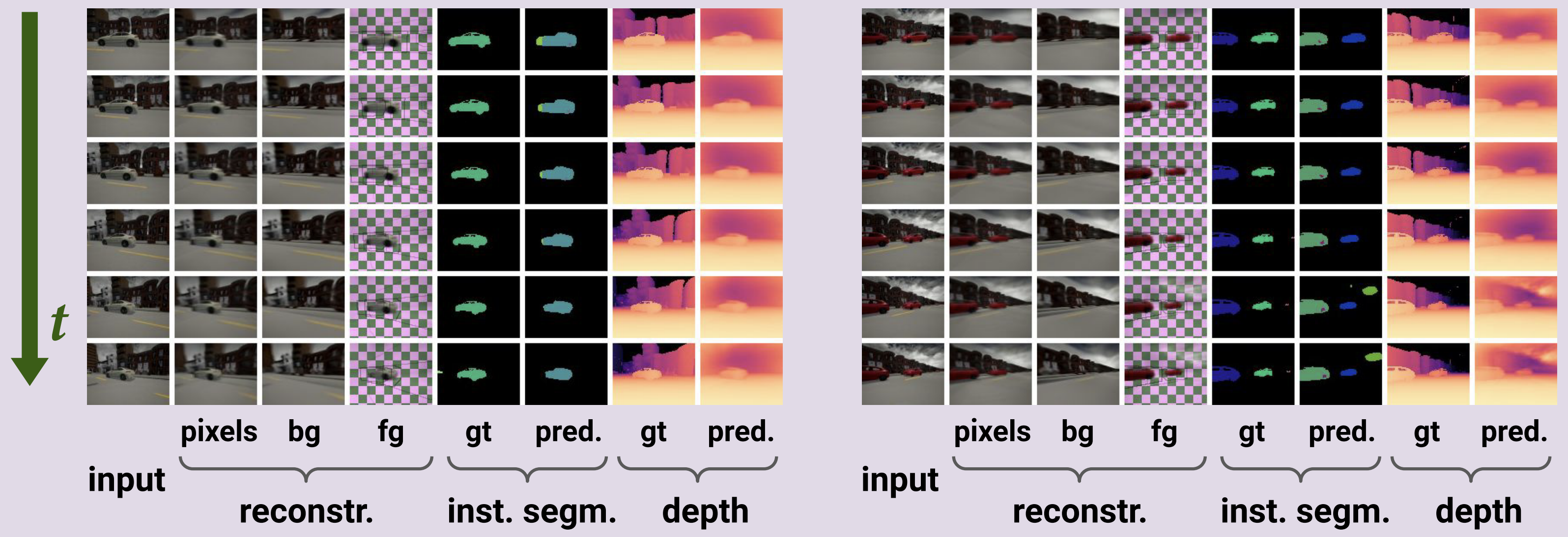


Ours (voxel objects)	FVD = 341.8	KID = 0.108
Ours (mesh objects)	FVD = 383.1	KID = 0.106
Baseline (GENESIS)	-	KID = 0.083 (single frames only!)
Baseline (SCALOR)	FVD = 1421.6	KID = 0.148

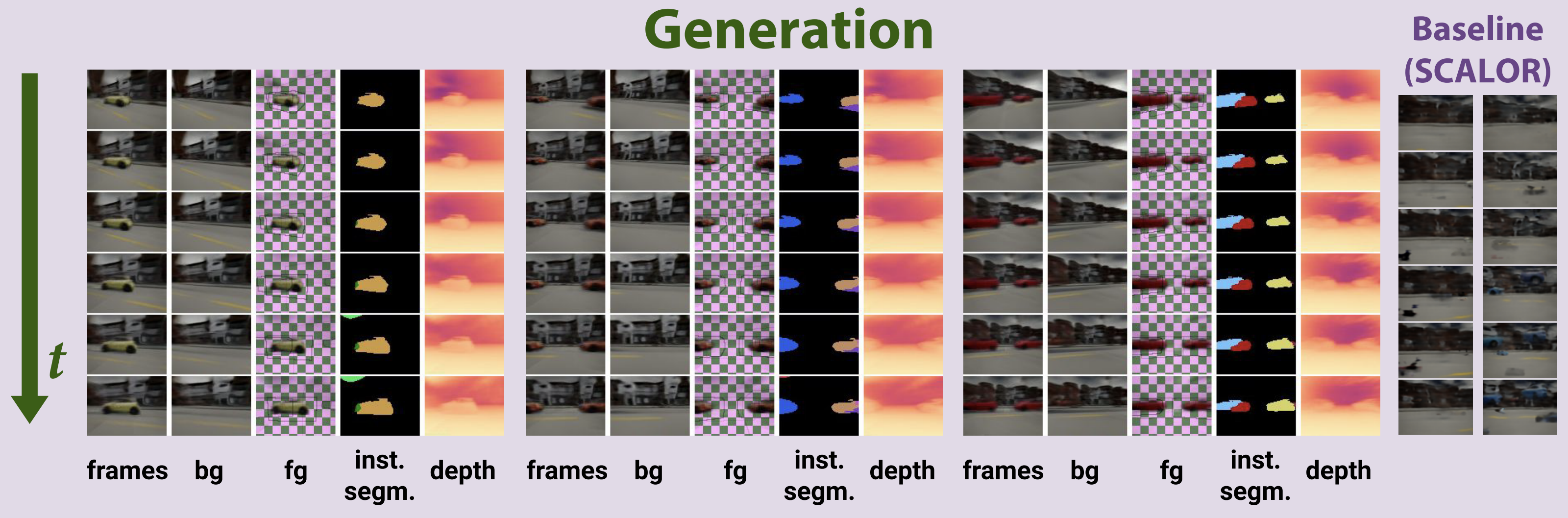
Traffic

- created using CARLA [Dosovitskiy, CoRL 2017]
- 1-3 cars driving along a straight road

Inference (unsupervised scene decomposition)



Generation


Baseline (SCALOR)
t
frames
bg
fg
inst. segm.
depth