Sequential Attend, Infer, Repeat:
Generative Modelling of Moving Objects

Adam R. Kosiorek[1,2], Hyunjik Kim[2],
Ingmar Posner[1], Yee Whye Teh[2]

Poster #24

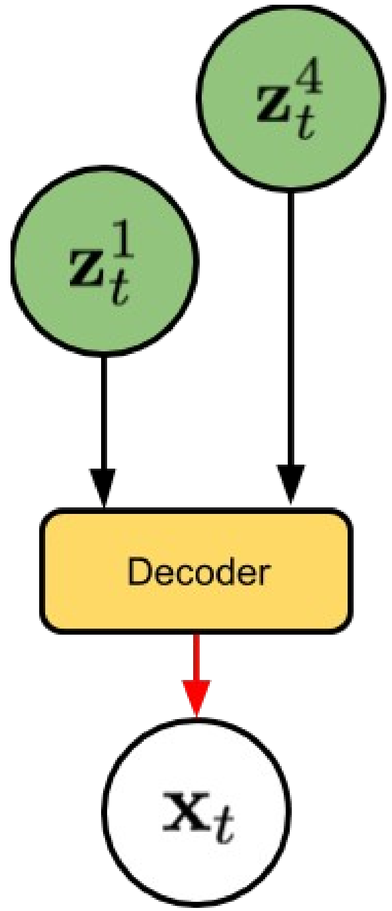
[1] Applied AI Lab, Oxford Robotics Institute
[2] Department of Statistics, University of Oxford

NeurIPS 2018

Attend, Infer, Repeat[1]

[1] Eslami et. al., “Attend, Infer, Repeat”, *NIPS* 2016.

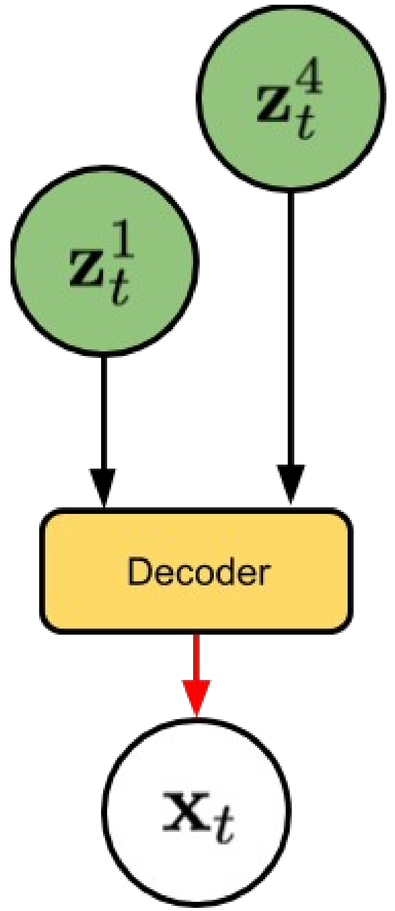
Attend, Infer, Repeat[1] (AIR):



[1] Eslami et. al., “Attend, Infer, Repeat”, *NIPS* 2016.

Attend, Infer, Repeat[1] (AIR):

- Variational Autoencoder (VAE)

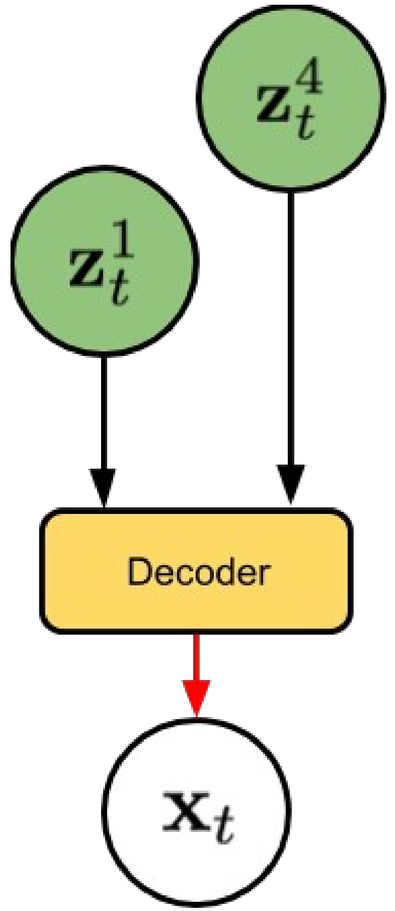


[1] Eslami et. al., “Attend, Infer, Repeat”, *NIPS* 2016.

Attend, Infer, Repeat

Attend, Infer, Repeat[1] (AIR):

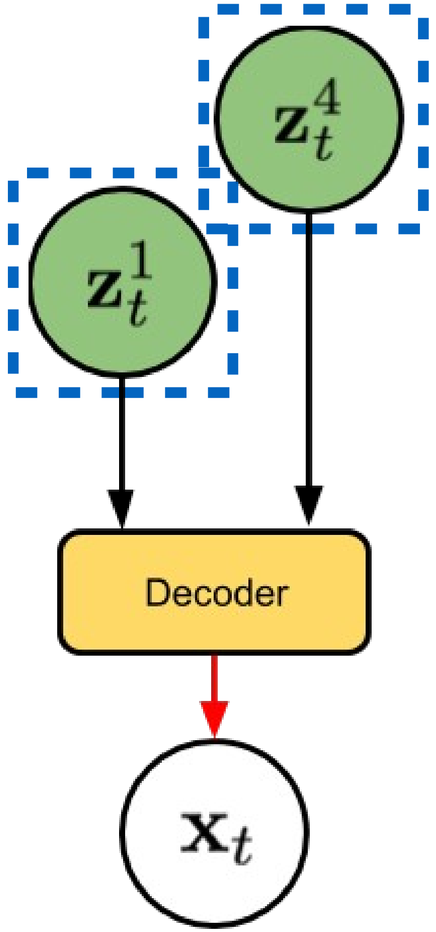
- Variational Autoencoder (VAE)
- Decomposes an image into objects



[1] Eslami et. al., “Attend, Infer, Repeat”, *NIPS* 2016.

Attend, Infer, Repeat[1] (AIR):

- Variational Autoencoder (VAE)
- Decomposes an image into objects
- Explains each object with a separate latent variable

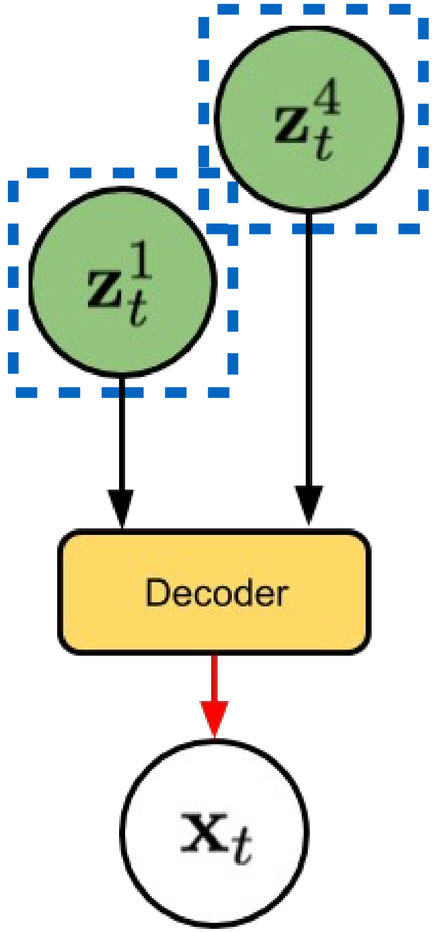


[1] Eslami et. al., “Attend, Infer, Repeat”, *NIPS* 2016.

Attend, Infer, Repeat[1] (AIR):

- Variational Autoencoder (VAE)
- Decomposes an image into objects
- Explains each object with a separate latent variable

Here, we have two objects with superscripts 1 and 4



[1] Eslami et. al., “Attend, Infer, Repeat”, *NIPS* 2016.

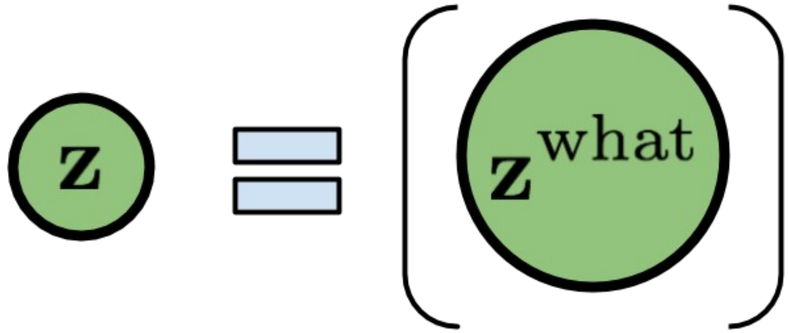
AIR: Latent Variables

Objects are explained by separate latent variables

AIR: Latent Variables

Objects are explained by separate latent variables

what: *Gaussian, how does it look like?*

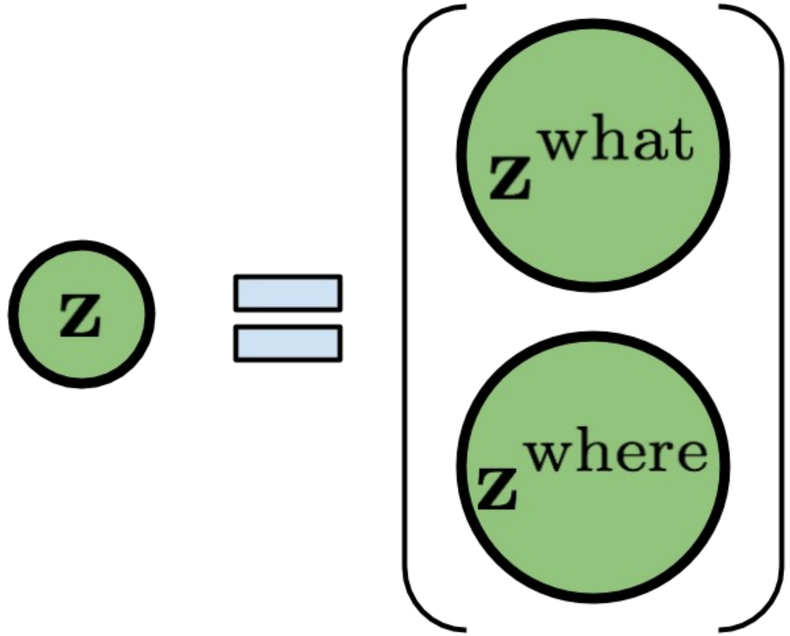


AIR: Latent Variables

Objects are explained by separate latent variables

what: *Gaussian, how does it look like?*

where: *Gaussian, where and how big is it?*



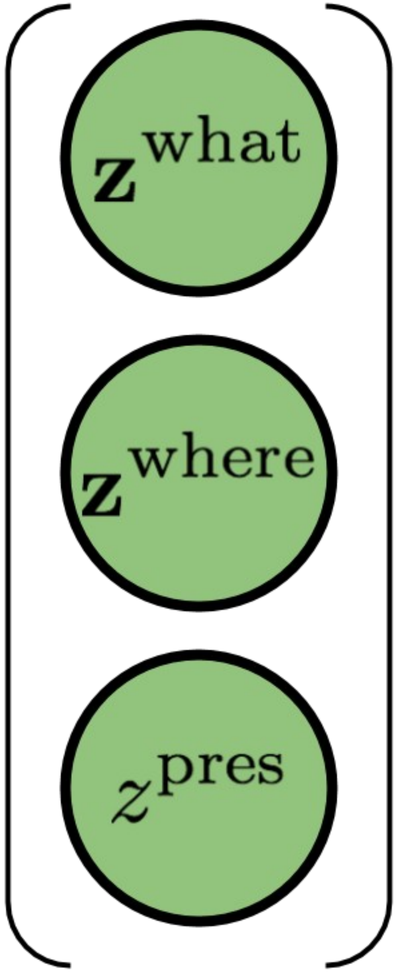
AIR: Latent Variables

Objects are explained by separate latent variables

what: *Gaussian, how does it look like?*

where: *Gaussian, where and how big is it?*

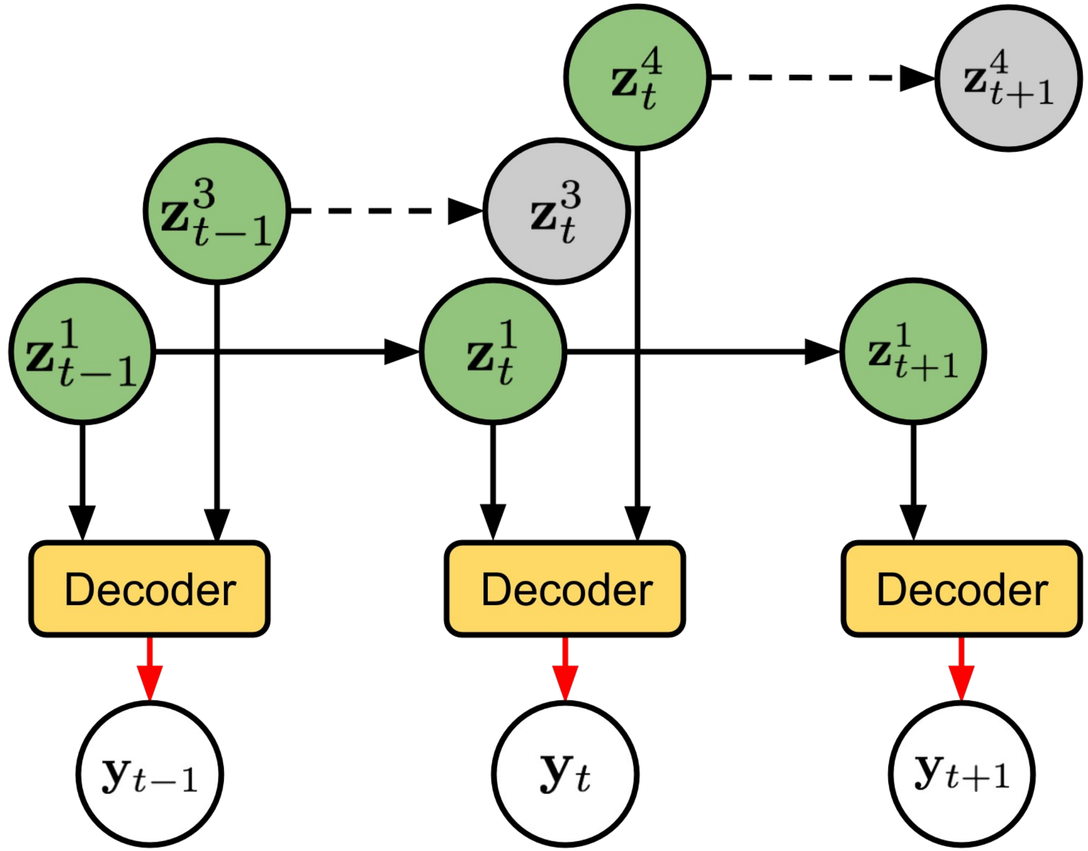
presence: *Bernoulli, does it exist?*



Sequential Attend, Infer, Repeat

SQAIR: Generative Model

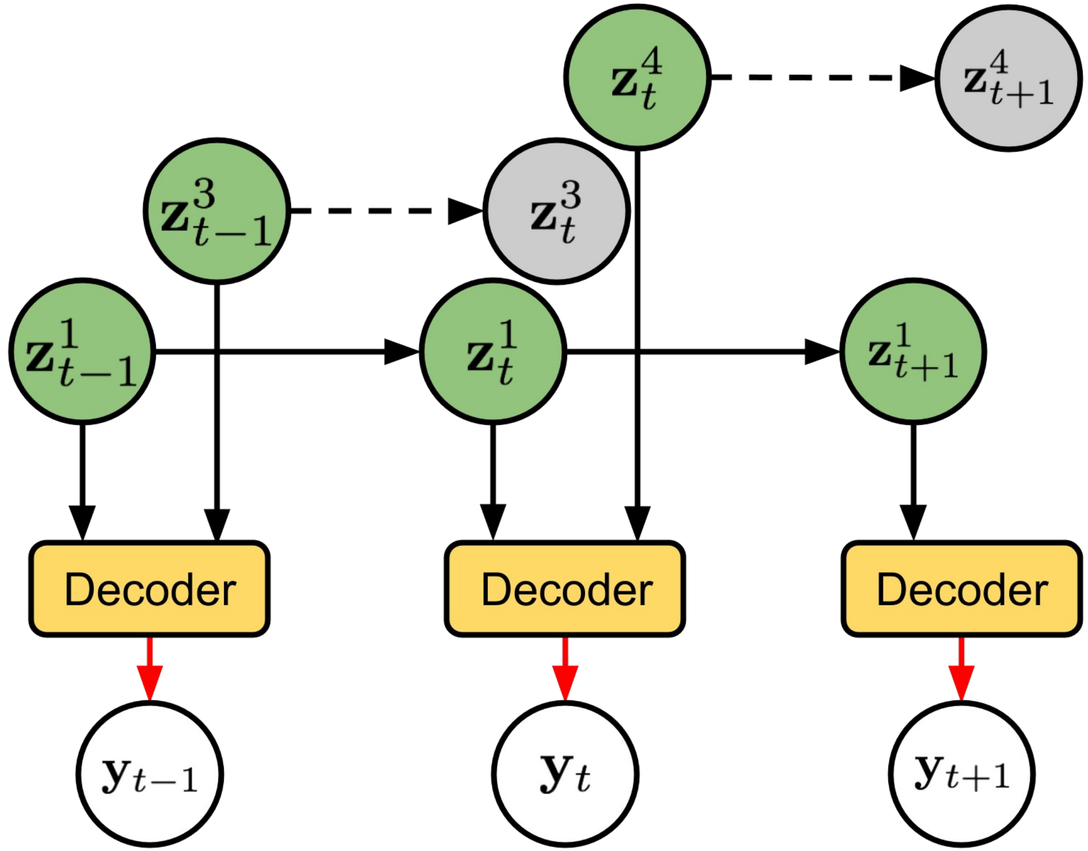
Sequential Attend, Infer Repeat (SQAIR) extends AIR to image sequences



SQAIR: Generative Model

Sequential Attend, Infer Repeat (SQAIR)
extends AIR to image sequences

Like AIR: model objects
with separate latent variables

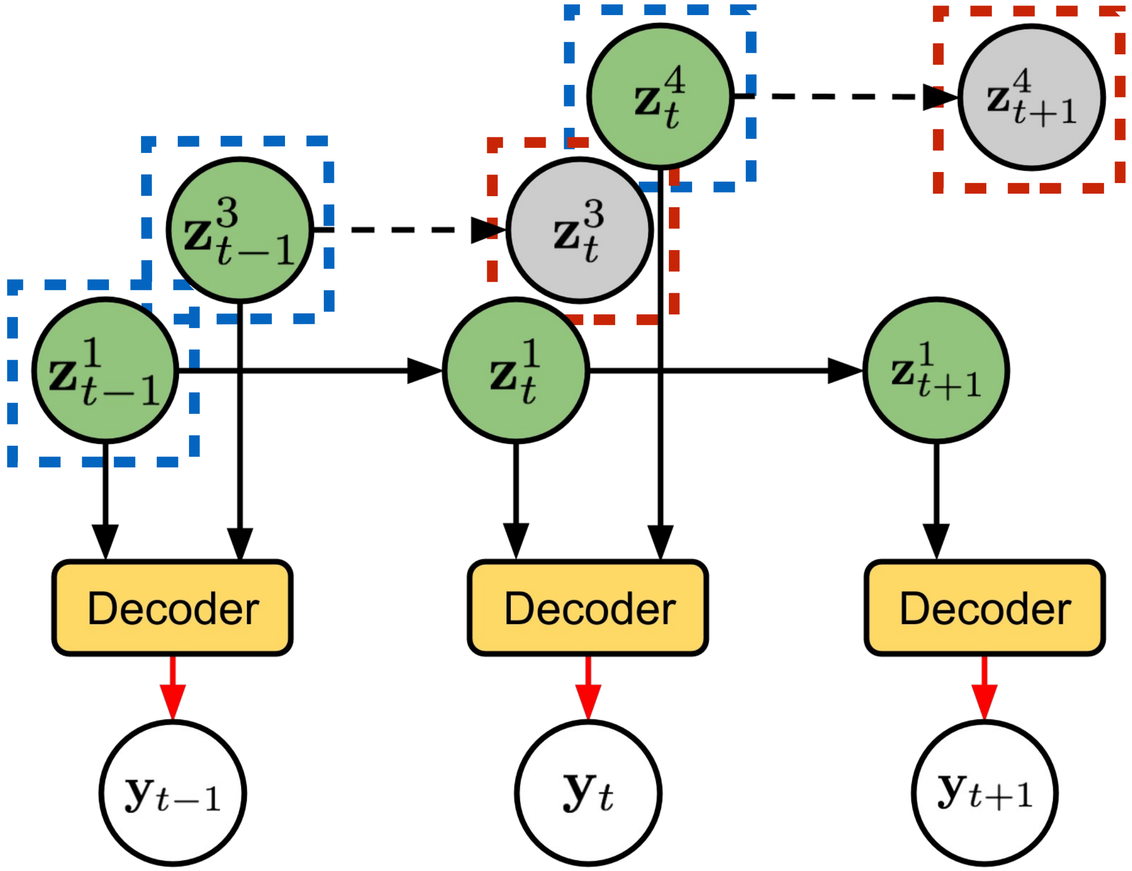


SQAIR: Generative Model

Sequential Attend, Infer Repeat (SQAIR)
extends AIR to image sequences

Like AIR: model objects
with separate latent variables

Objects can appear and
disappear in every frame



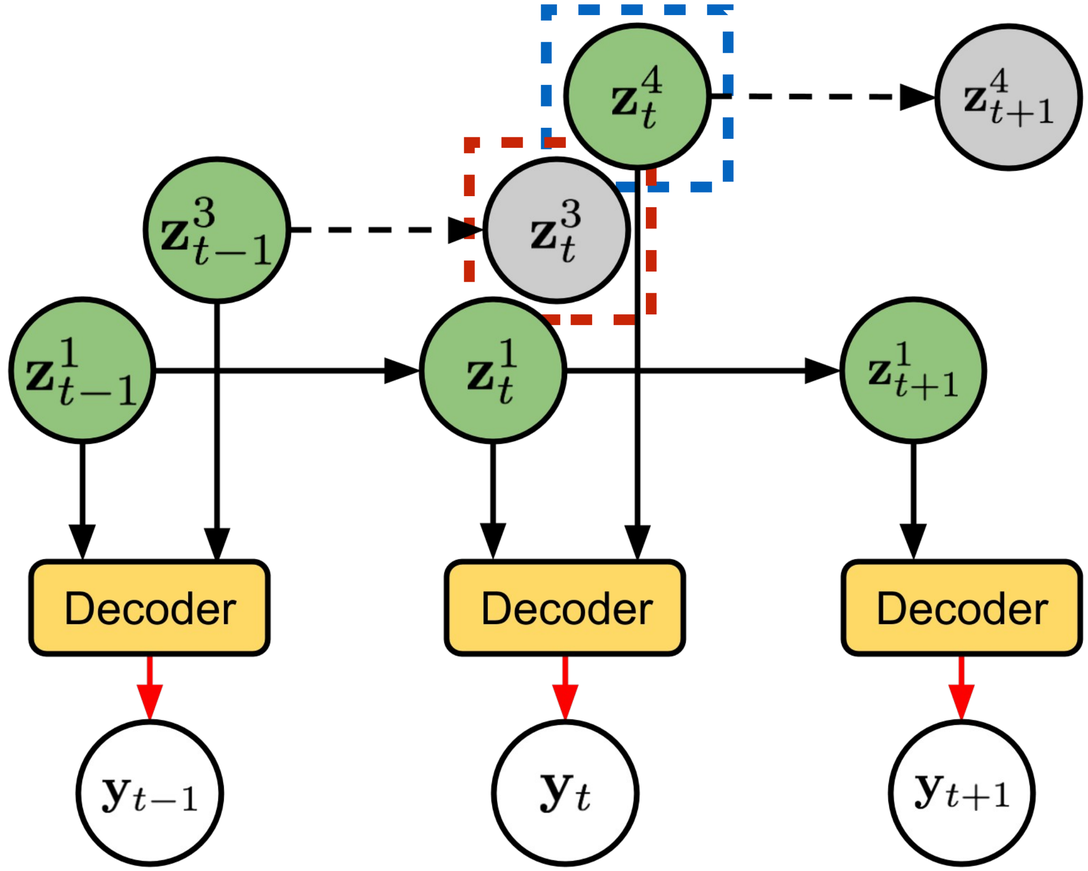
SQAIR: Generative Model

Sequential Attend, Infer Repeat (SQAIR)
extends AIR to image sequences

Like AIR: model objects
with separate latent variables

Objects can appear and
disappear in every frame

Here, object 4 appeared and
object 3 disappeared in frame t



MNIST: Reconstructions

SQAIR can model sequences of moving objects

MNIST: Reconstructions

SQAIR can model sequences of moving objects

like this one

MNIST: Reconstructions

SQAIR can model sequences of moving objects

like this one

any VAE could reconstruct it

MNIST: Reconstructions

SQAIR can model sequences of moving objects

like this one

any VAE could reconstruct it

SQAIR:

one latent variable per object

knows their location

maintains identity (unlike AIR)

MNIST: Samples

Once trained, we can sample from SQAIR

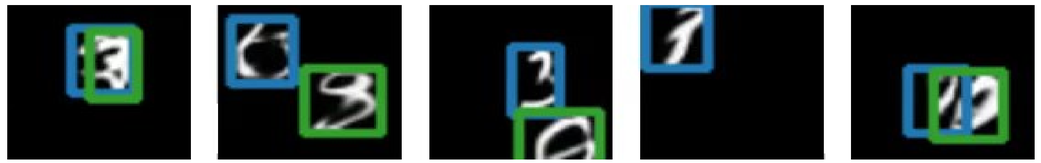
Check what the model learned

MNIST: Samples

Once trained, we can sample from SQAIR

Check what the model learned

Object appearance does not change between frames

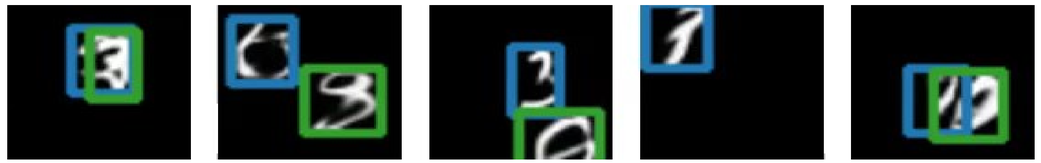


MNIST: Samples

Once trained, we can sample from SQAIR

Check what the model learned

Object appearance does not change between frames

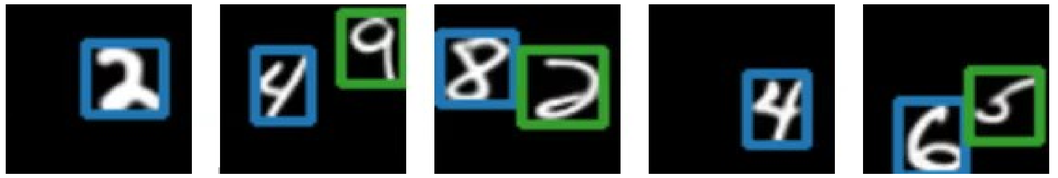


Motion is consistent with motion patterns in the training set

MNIST: Conditional Generation

Condition the model on three frames

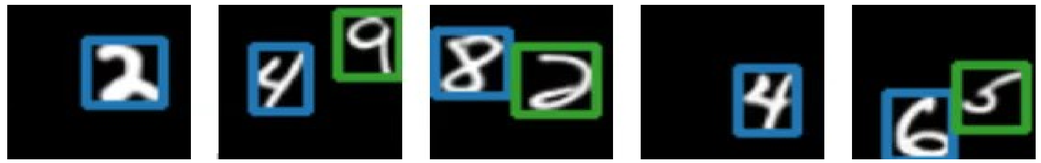
Predict the next 97 frames
by sampling from the prior



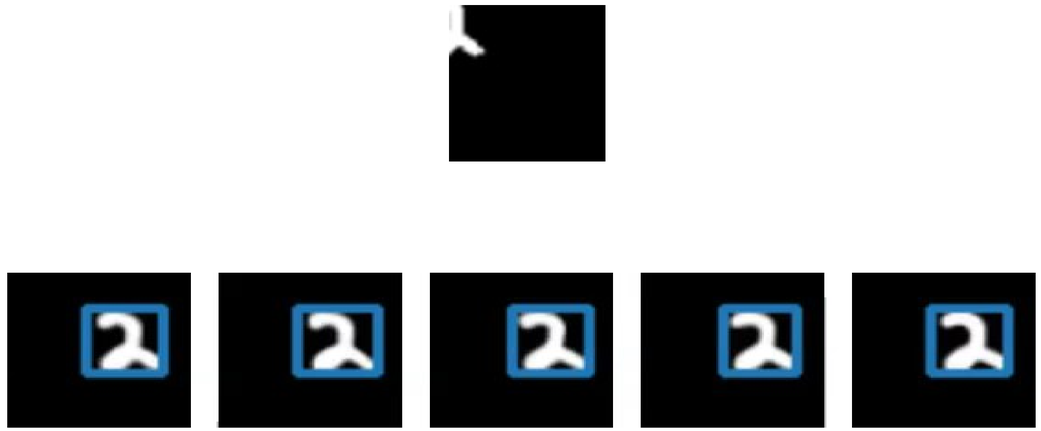
MNIST: Conditional Generation

Condition the model on three frames

Predict the next 97 frames
by sampling from the prior



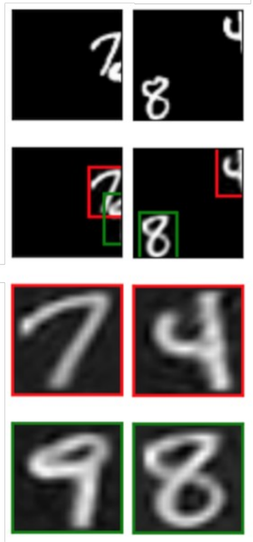
For every conditioning sequence,
we can imagine different rollouts



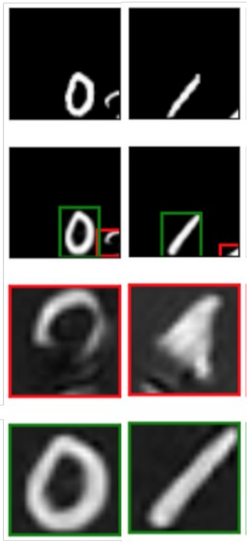
SQAIR vs AIR

Reconstruction from partial observations

SQAIR



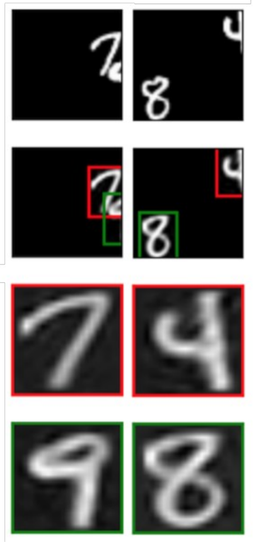
AIR



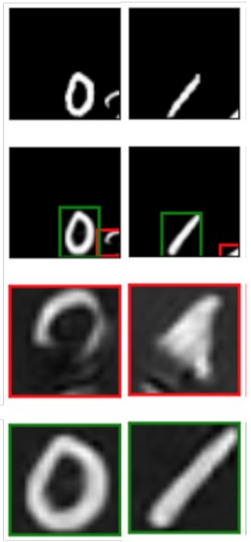
SQAIR vs AIR

Reconstruction from partial observations

SQAIR



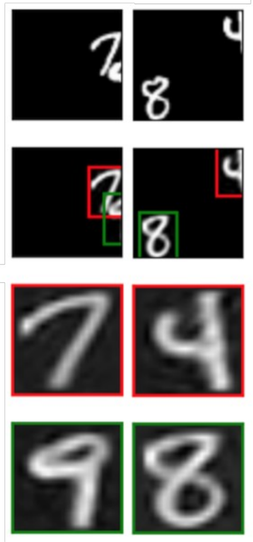
AIR



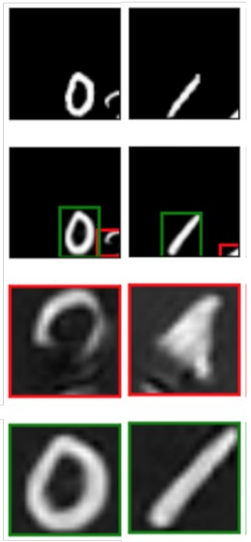
SQAIR vs AIR

Reconstruction from partial observations

SQAIR



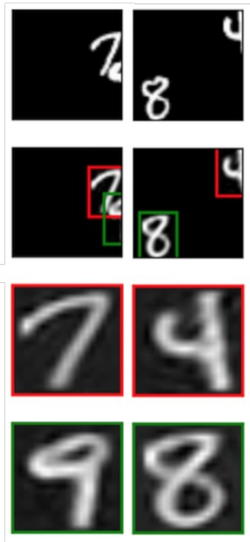
AIR



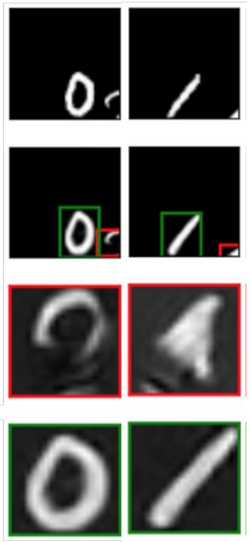
SQAIR vs AIR

Reconstruction from partial observations

SQAIR

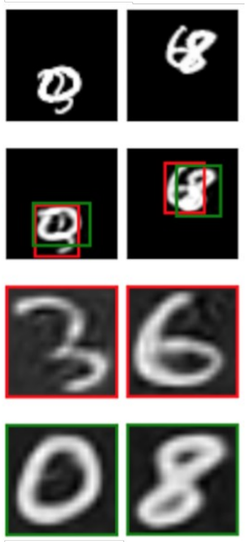


AIR

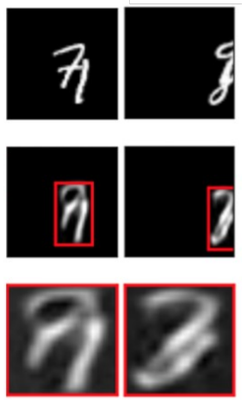


Disentangling overlapping objects

SQAIR



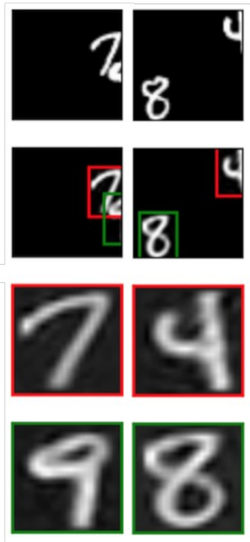
AIR



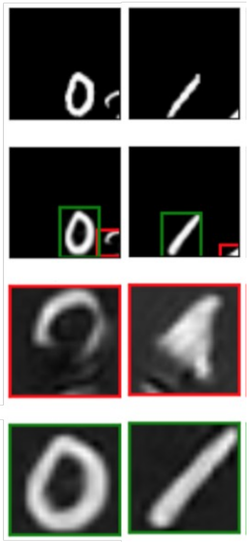
SQAIR vs AIR

Reconstruction from partial observations

SQAIR

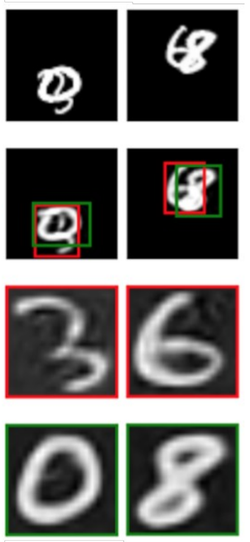


AIR

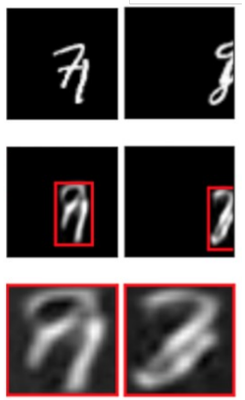


Disentangling overlapping objects

SQAIR



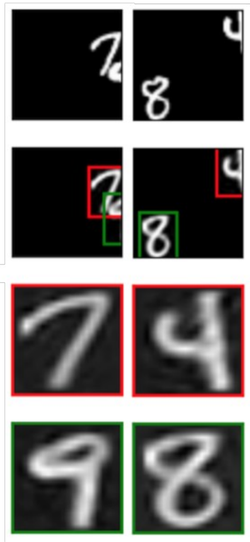
AIR



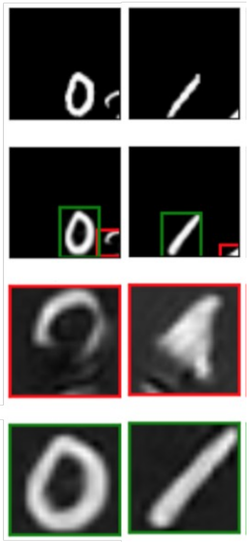
SQAIR vs AIR

Reconstruction from partial observations

SQAIR

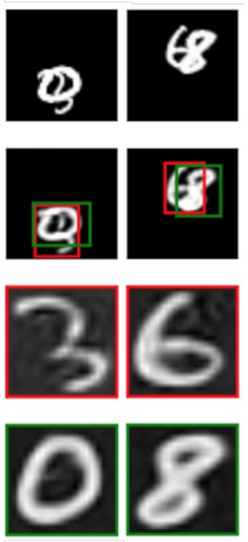


AIR

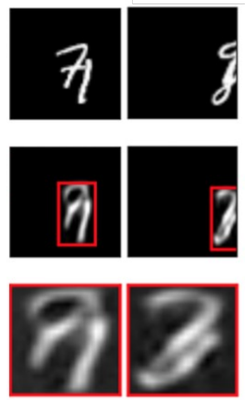


Disentangling overlapping objects

SQAIR



AIR

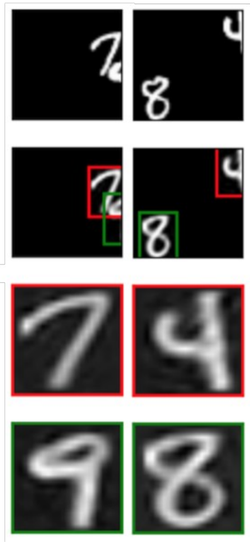


missing objects!

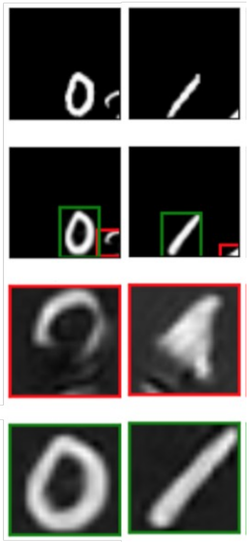
SQAIR vs AIR

Reconstruction from partial observations

SQAIR

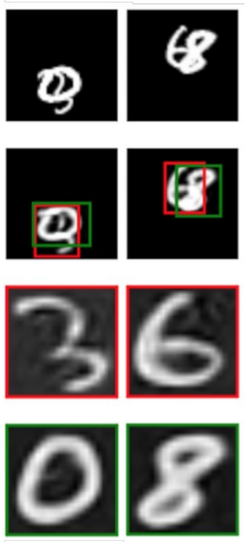


AIR

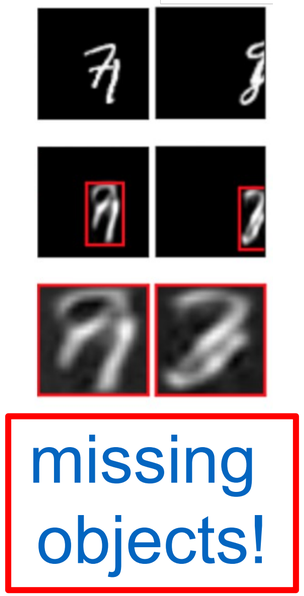


Disentangling overlapping objects

SQAIR



AIR

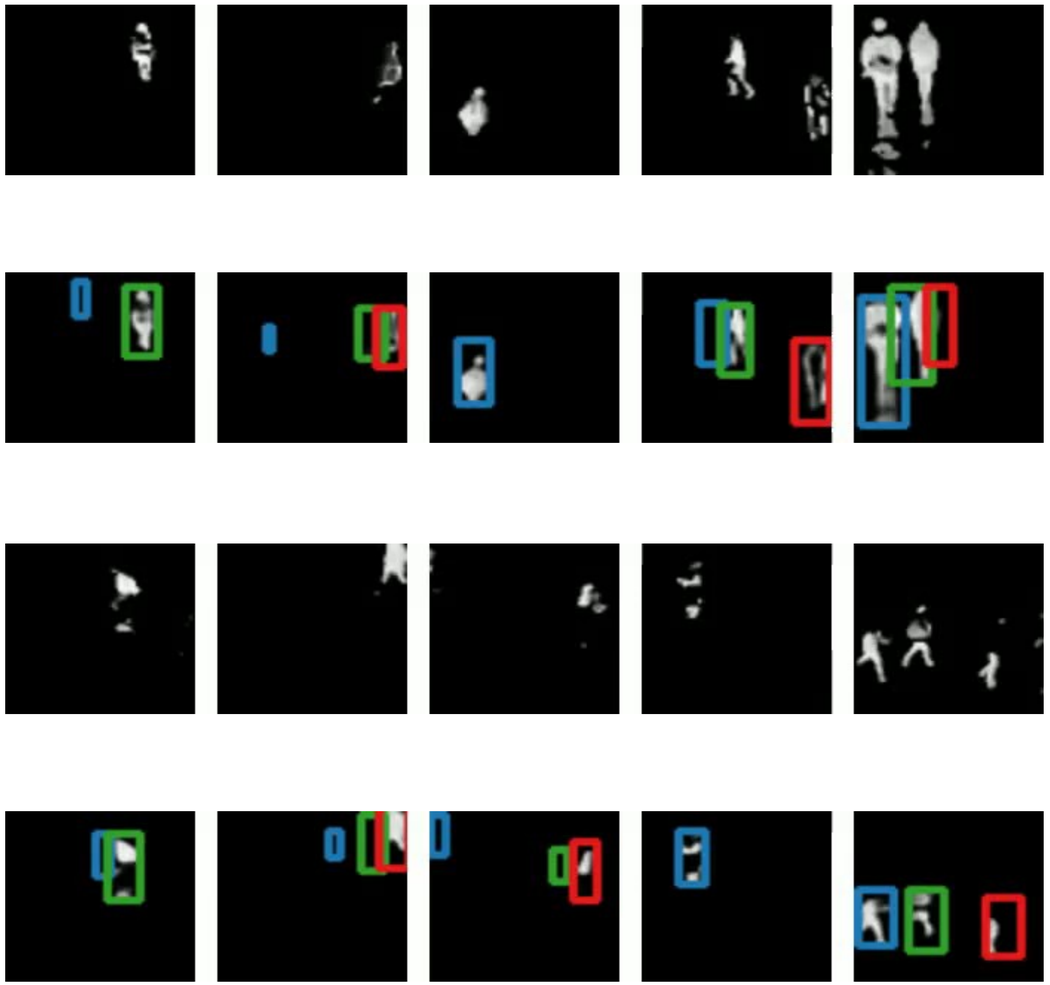


missing objects!

Real World Data: Unsupervised Detection & Tracking of Pedestrians

DukeMTMC: Reconstructions

DukeMTMC dataset[2] contains videos from static CCTV cameras

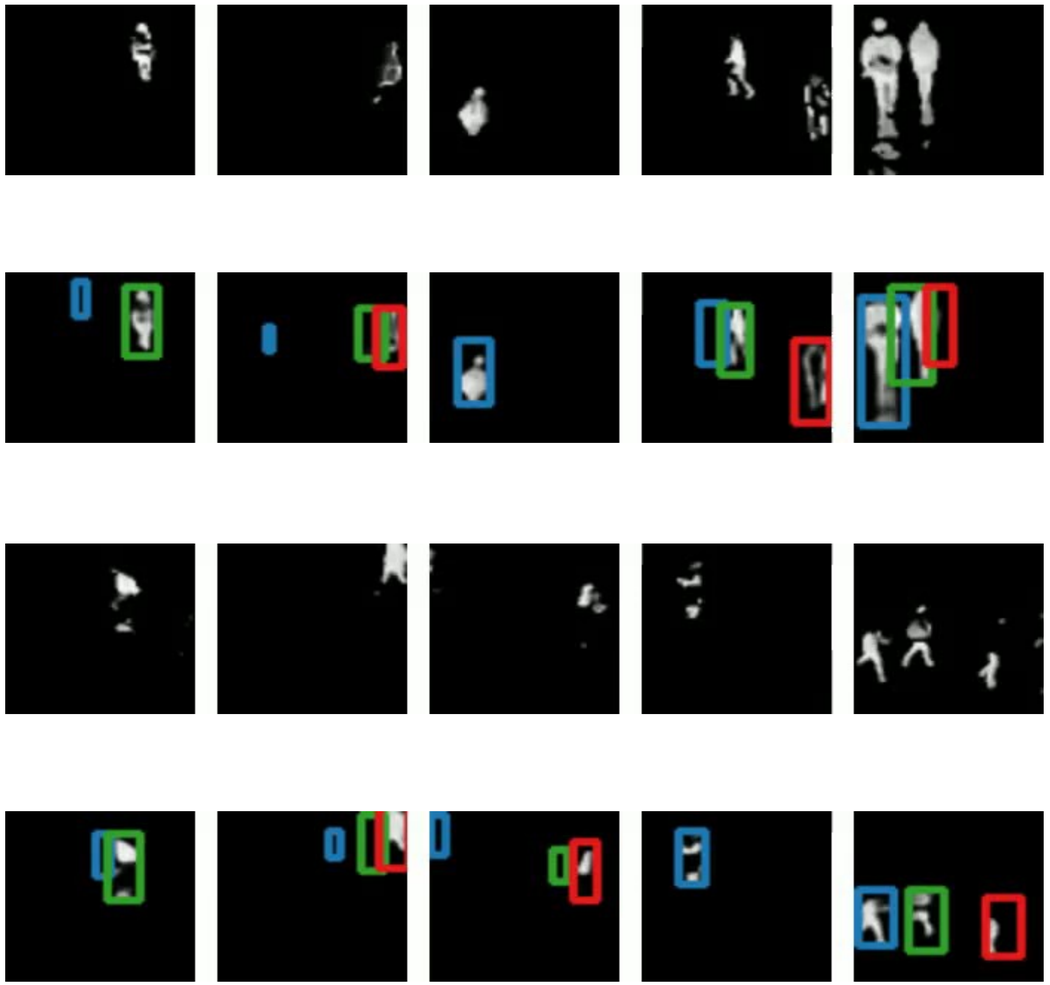


[2] Ristani et. al., “Performance Measures and a Data Set for Multi-Target, Multi-Camera Tracking”, *ECCV workshop*, 2016.

DukeMTMC: Reconstructions

DukeMTMC dataset[2] contains videos from static CCTV cameras

Pre-process by removing backgrounds and inverting colours



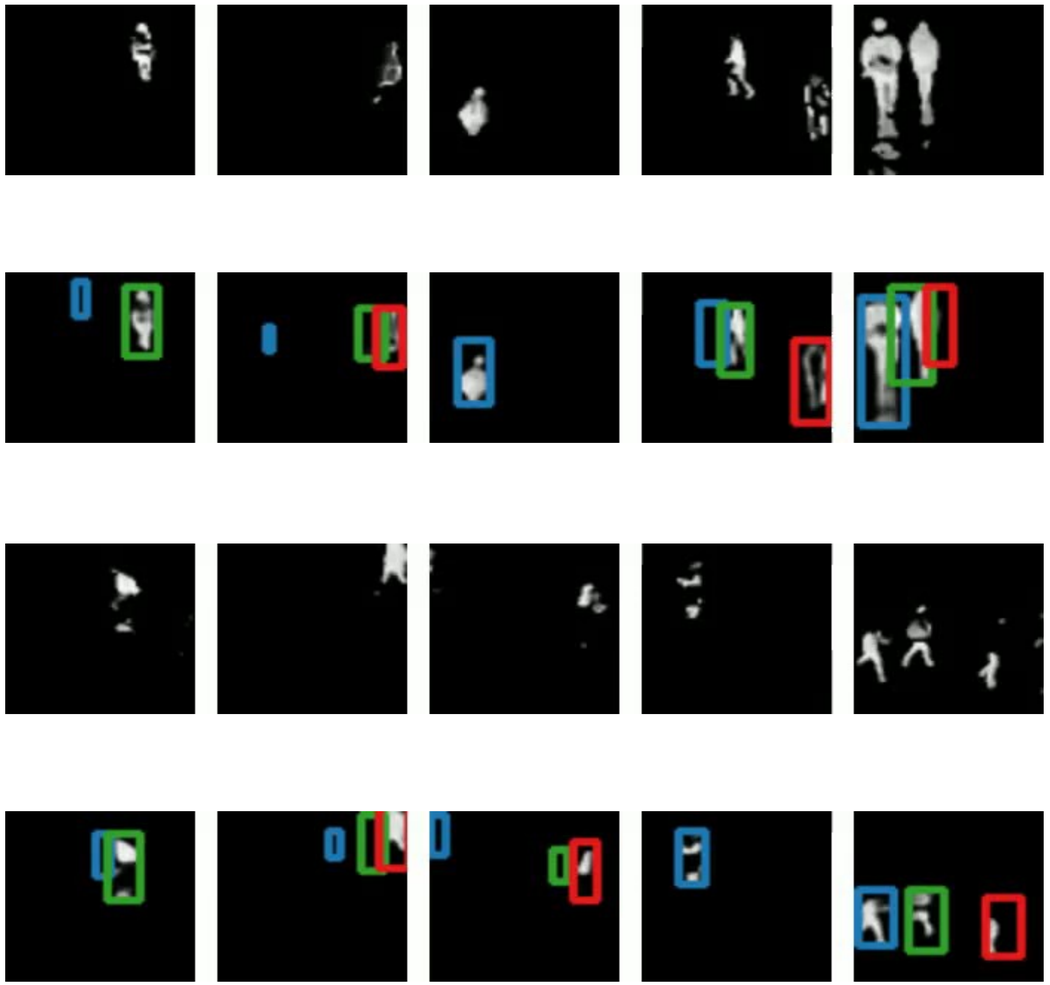
[2] Ristani et. al., “Performance Measures and a Data Set for Multi-Target, Multi-Camera Tracking”, *ECCV workshop*, 2016.

DukeMTMC: Reconstructions

DukeMTMC dataset[2] contains videos from static CCTV cameras

Pre-process by removing backgrounds and inverting colours

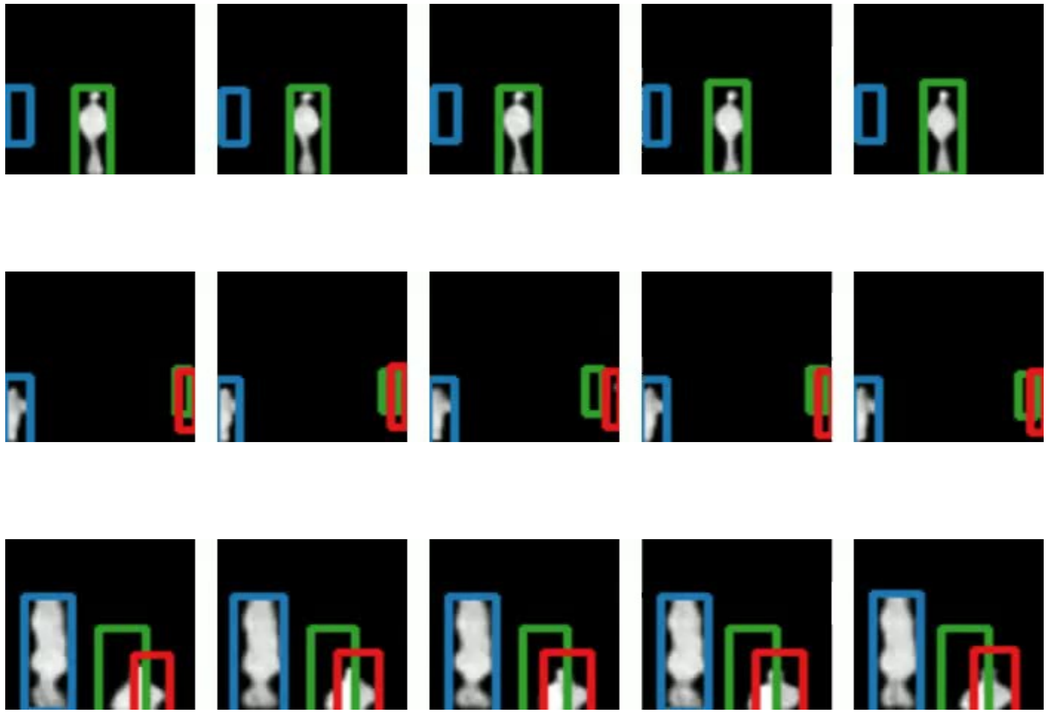
SQAIR learns to detect & track pedestrians without human supervision!



[2] Ristani et. al., “Performance Measures and a Data Set for Multi-Target, Multi-Camera Tracking”, *ECCV workshop*, 2016.

DukeMTMC: Conditional Generation

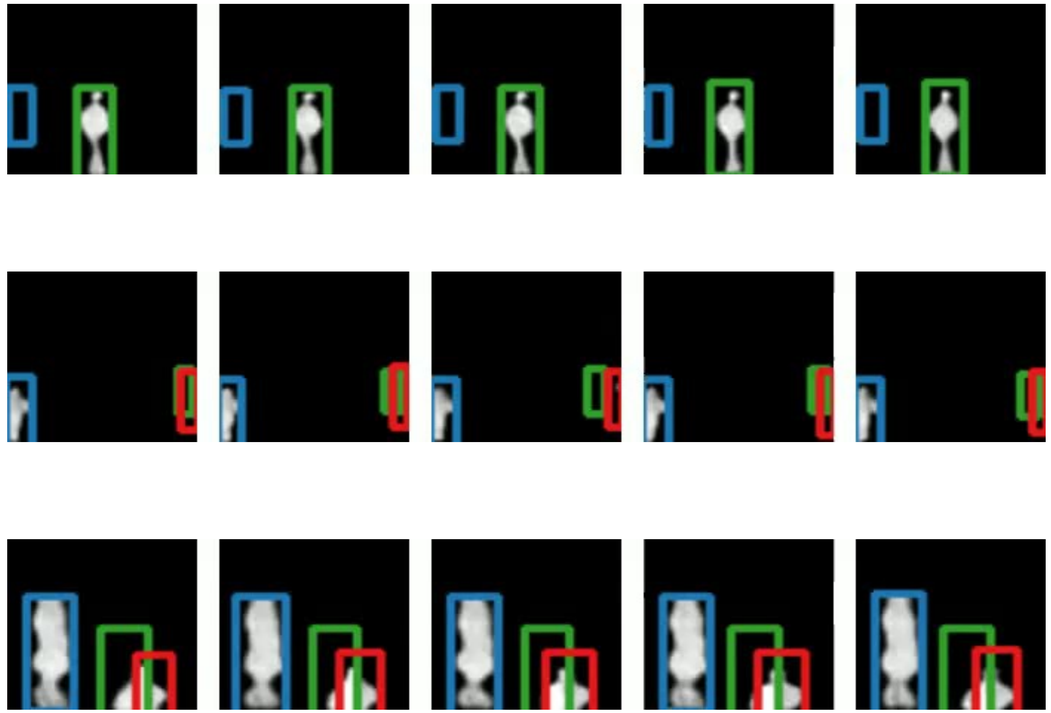
SQAIR trained on sequences of five frames



DukeMTMC: Conditional Generation

SQAIR trained on sequences of five frames

- Condition the model on five frames
- Predict the next 15 frames by sampling from the prior

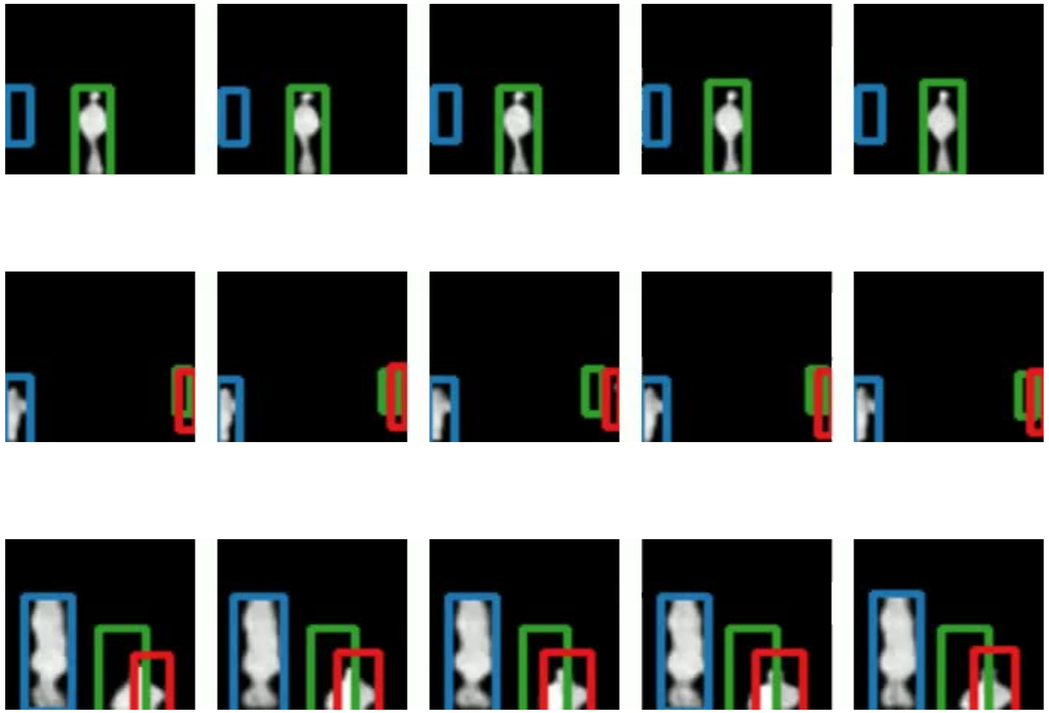


DukeMTMC: Conditional Generation

SQAIR trained on sequences of five frames

- Condition the model on five frames
- Predict the next 15 frames by sampling from the prior

Each row contains five different predictions for the same sequence



Code:

/akosiorek/SQAIR

Poster #24