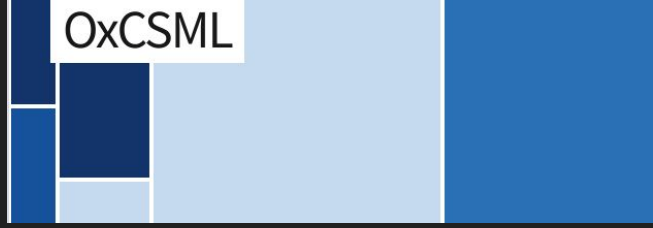


Attention Mechanisms

Knowing Where to Look Improves Visual Reasoning

Adam R. Kosiorek

Contributors:
Alex Bewley
Hyunjik Kim
Ingmar Posner
Yee Whye Teh

#reworkDL London, 20.09.2018

Visual Distractions

Why more is not always better?

- Irrelevant information
- Computational cost
- Noise
- Harder credit assignment

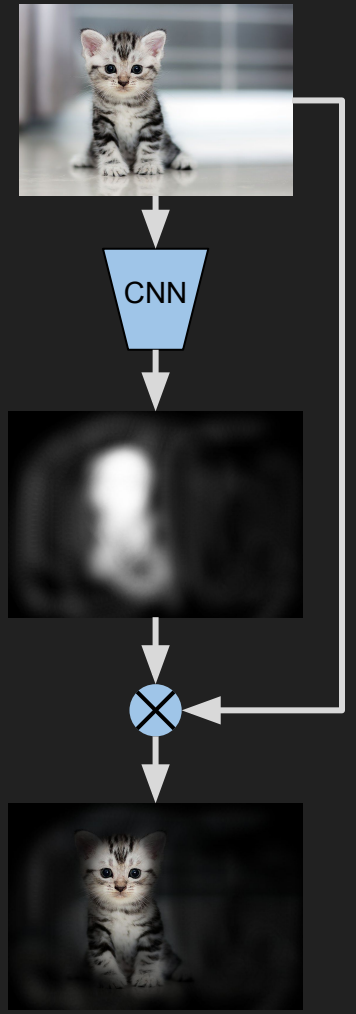
Visual Attention

Soft-attention

Blacking out irrelevant information

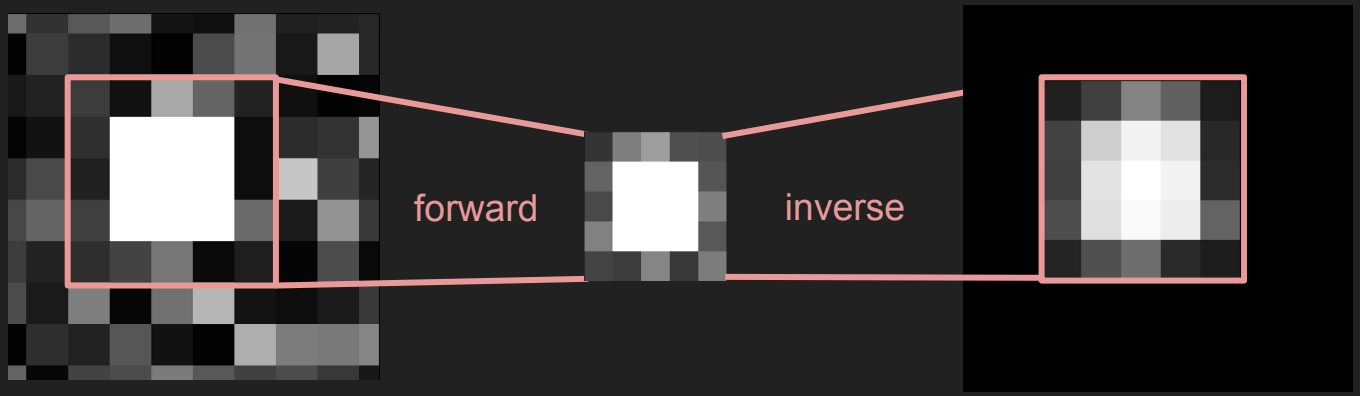
- Fully differentiable
- Structure depends on parametrisation
  - Typically given by a conv-net (for images)

    => spatial correlation
  - No structure in case of overfitting or MLPs
- Computationally wasteful
  - We need to process the whole image!



Spatial Transformer

Parametric Affine Transform



Original image

Extracted glimpse

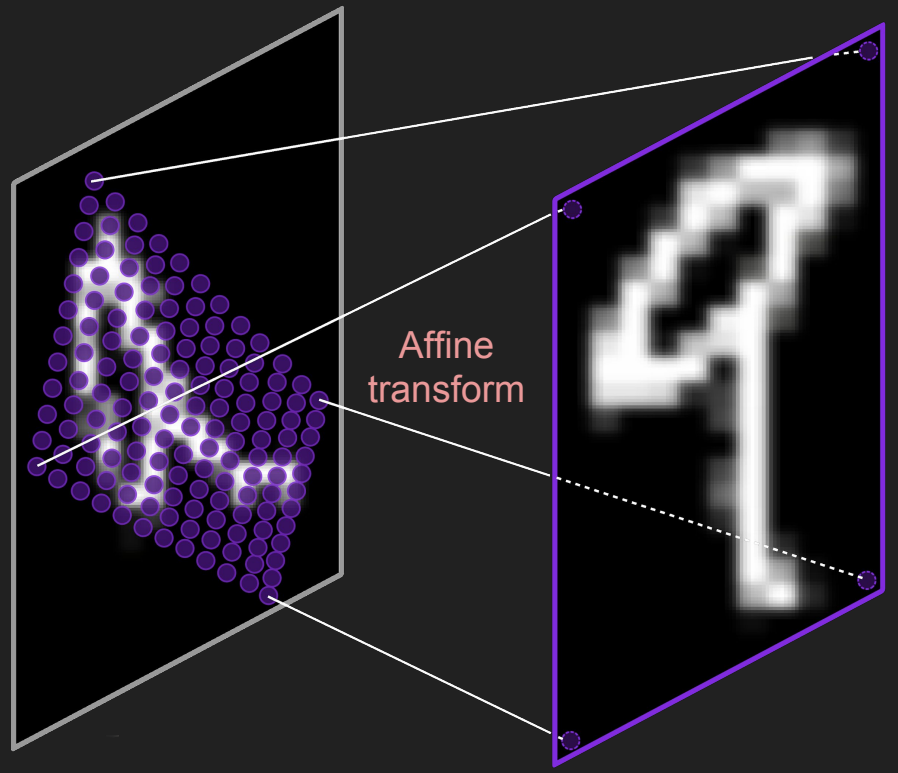
Inverted transformation

- Differentiable
- Computationally efficient

Jaderberg *et. al.*, “Spatial Transformer Networks”, NIPS 2015

Spatial Transformer

Parametric Affine Transform

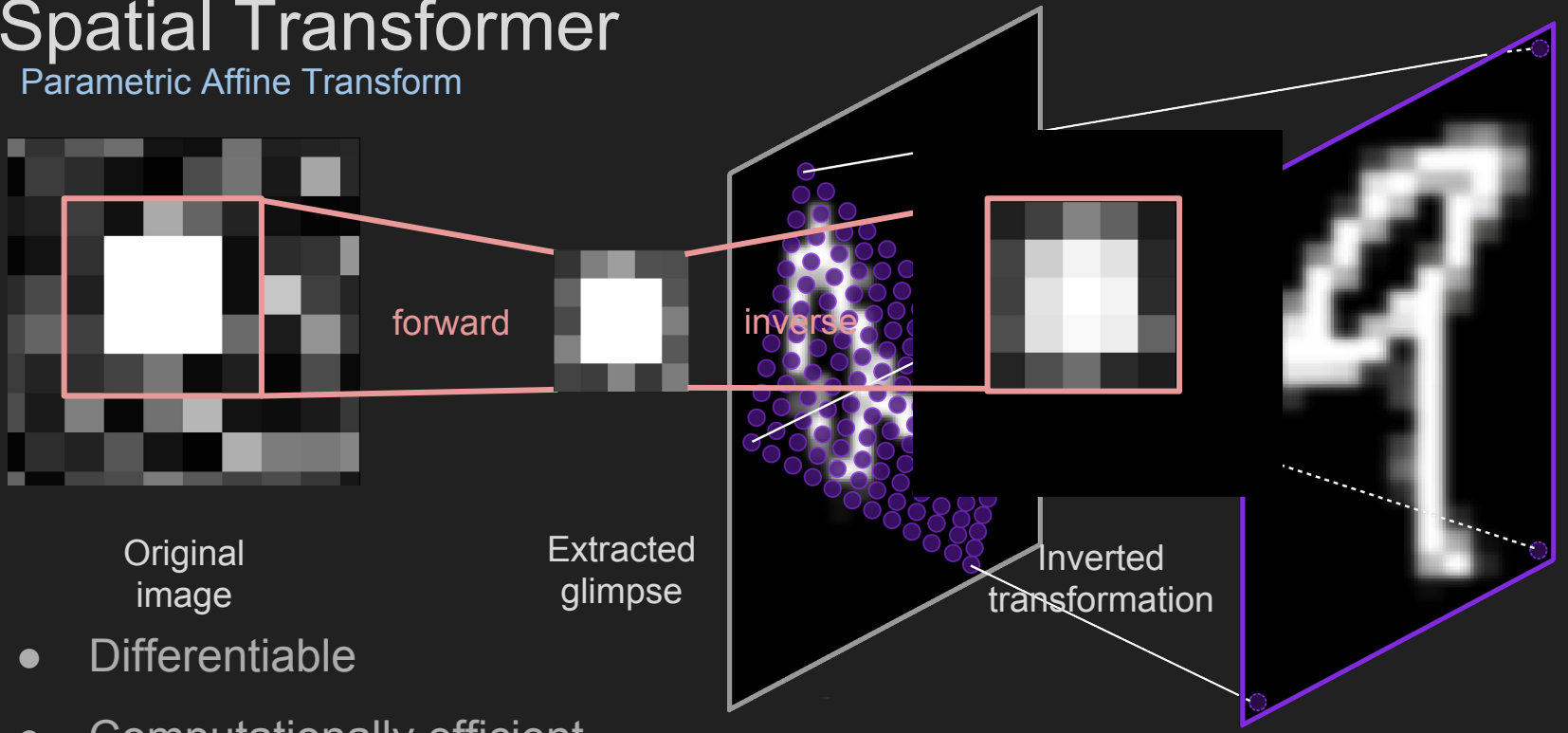


- Differentiable
- Computationally efficient
- Extract a parametric structure from an image

Jaderberg *et. al.*, “Spatial Transformer Networks”, NIPS 2015

Spatial Transformer

Parametric Affine Transform



- Differentiable
- Computationally efficient
- Extract a parametric structure from an image

Jaderberg *et. al.*, “Spatial Transformer Networks”, NIPS 2015

Object Tracking in Videos

Problem Setup

- One object at a time
- Class-agnostic
  - No pre-trained object detectors or classifiers
- Initialised with a known location

Adrian Rosebrock, “OpenCV Object Tracking Demo #3”, pyimagesearch 2018.

Here is an idea

Context: No need to look at the whole image

Here is an idea

Context: No need to look at the whole image

Motion: Can we estimate how the object moves?

Here is an idea

Context: No need to look at the whole image

Motion: Can we estimate how the object moves?

Appearance: Can we estimate how its appearance changes?

Here is an idea

Context: No need to look at the whole image

Motion: Can we estimate how the object moves?

Appearance: Can we estimate how its appearance changes?

Solution: Attention + Recurrent Neural Nets

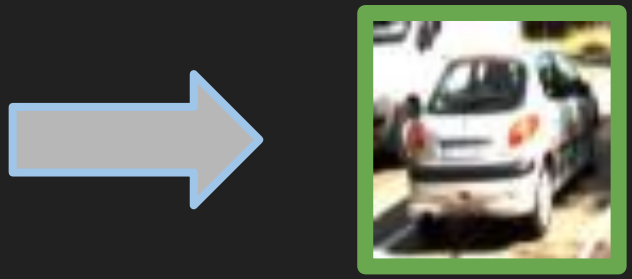
Hierarchical Attentive Recurrent Tracking (HART)

Kosiorek *et. al.*, “Hierarchical Attentive Recurrent Tracking”, NIPS 2017

Hierarchical Attentive Recurrent Tracking (HART)

Learn to look before learning to track

- Idea: To track an object, look at the object - not the image

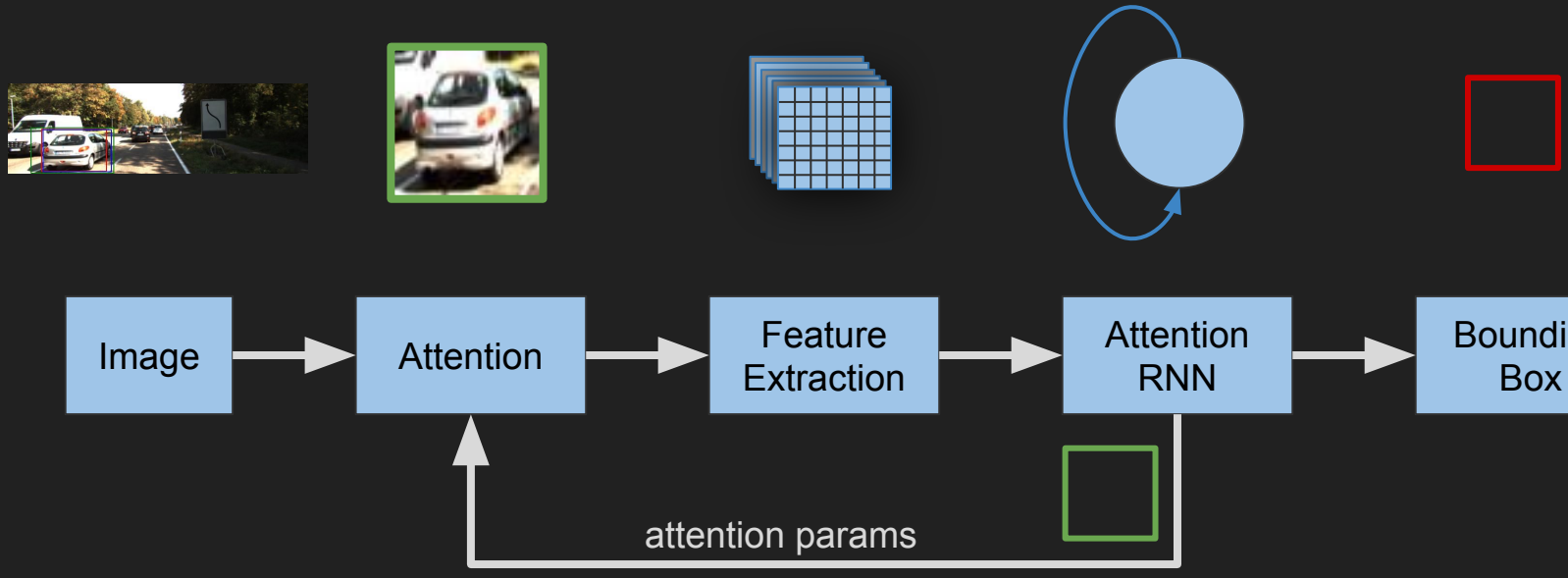


  - Higher computational efficiency
  - Easier credit assignment for learning
  - Bonus: learned motion and appearance model

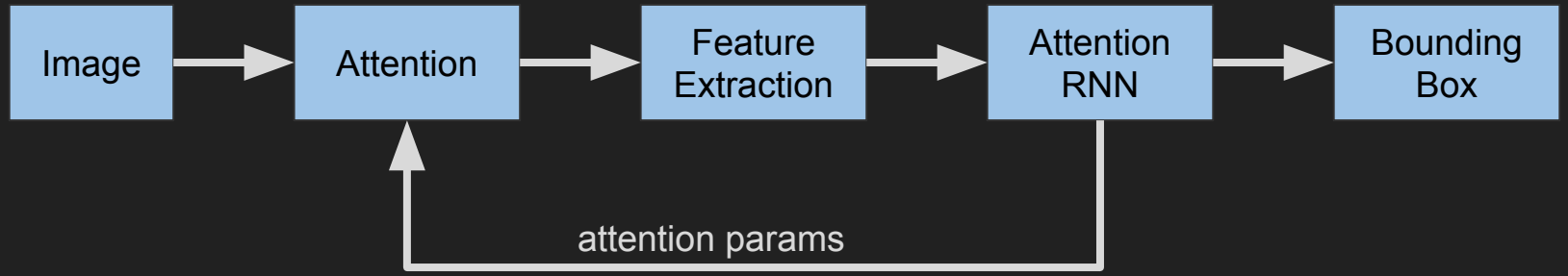
Kosiorek *et. al.*, “Hierarchical Attentive Recurrent Tracking”, NIPS 2017

Hierarchical Attentive Recurrent Tracking

How does it work?



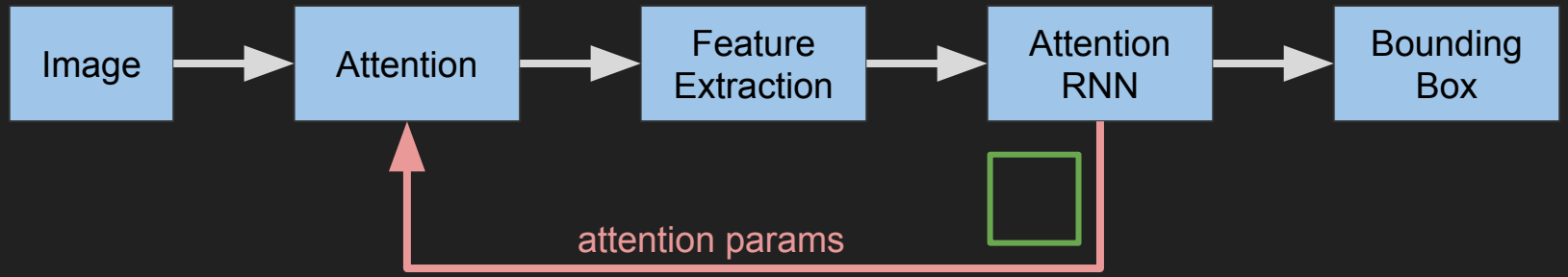
Hierarchical Attentive Recurrent Tracking



- We start with an image

Kosiorek *et. al.*, “Hierarchical Attentive Recurrent Tracking”, NIPS 2017

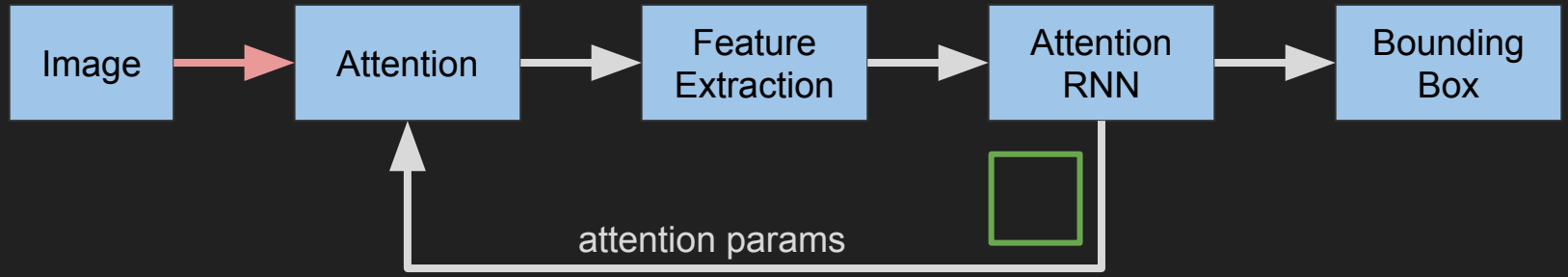
Hierarchical Attentive Recurrent Tracking



- Predict object location based on its past motion

Kosiorek *et. al.*, “Hierarchical Attentive Recurrent Tracking”, NIPS 2017

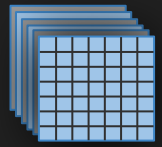
Hierarchical Attentive Recurrent Tracking

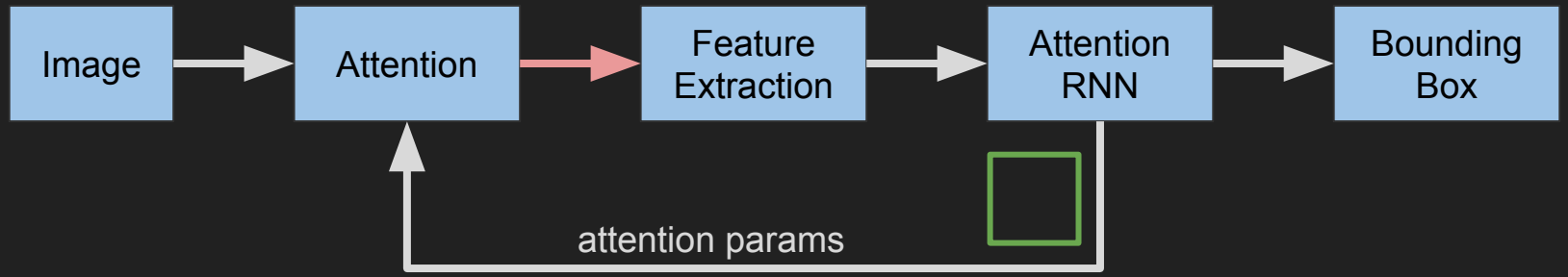


- Extract a glimpse from this location (Spatial Transformer)

Kosiorek *et. al.*, “Hierarchical Attentive Recurrent Tracking”, NIPS 2017

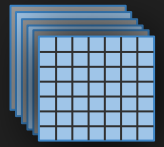
Hierarchical Attentive Recurrent Tracking

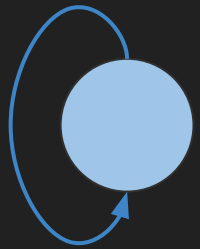


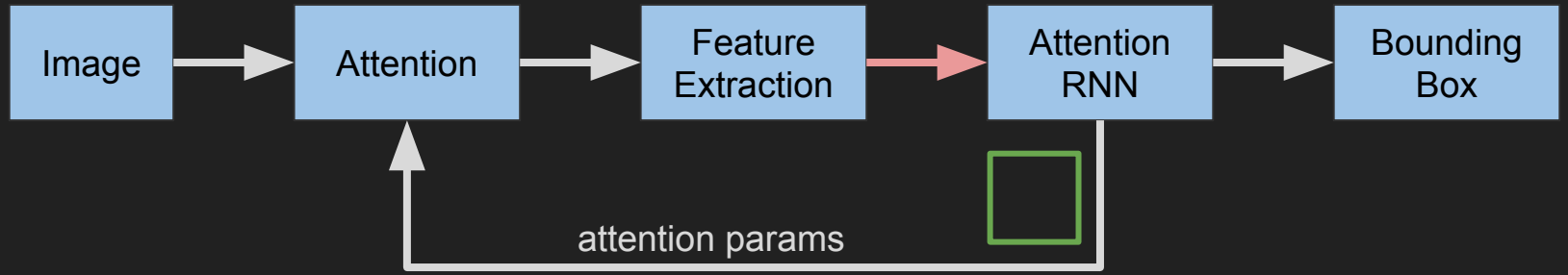


- Extract features from the glimpse

Hierarchical Attentive Recurrent Tracking

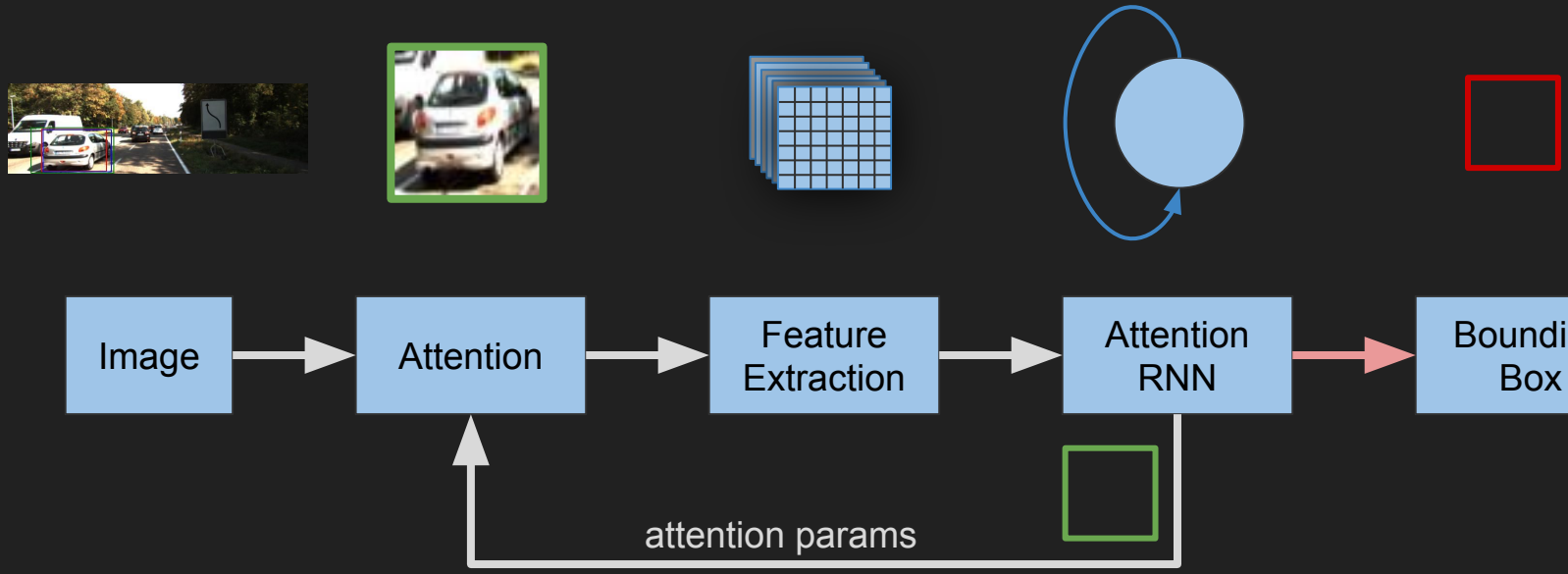






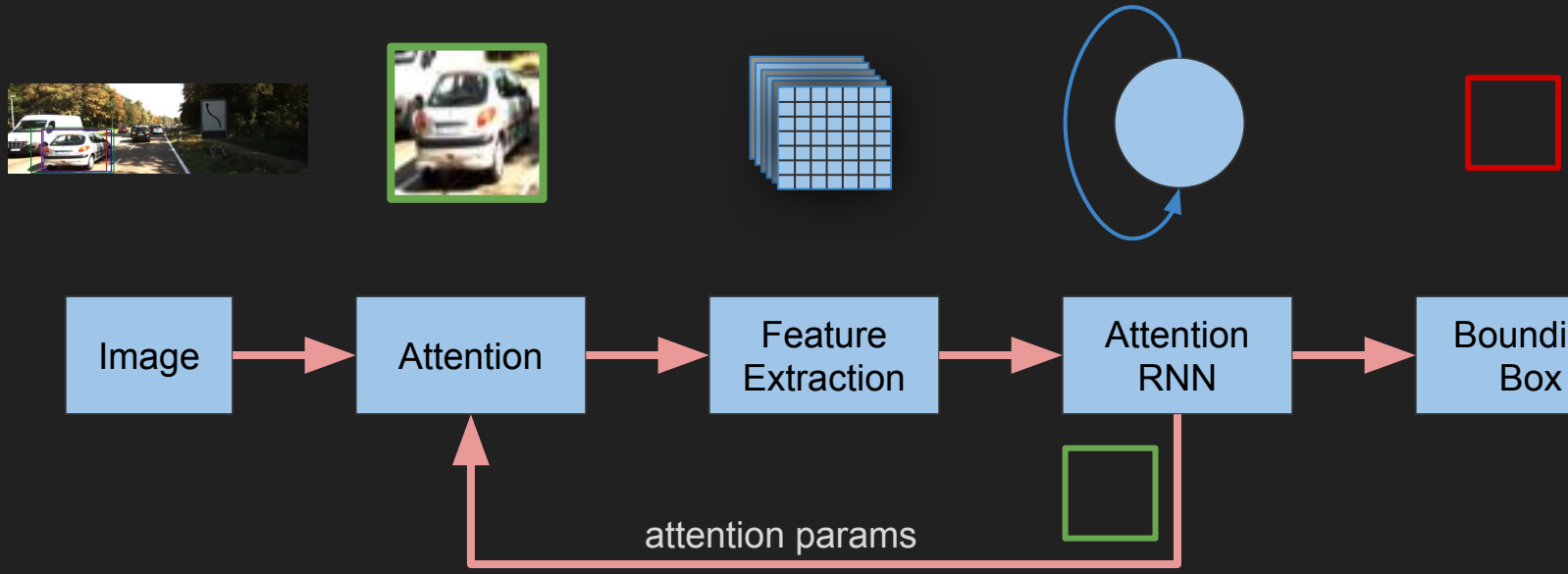
- Update hidden state

Hierarchical Attentive Recurrent Tracking



- Predict a bounding box

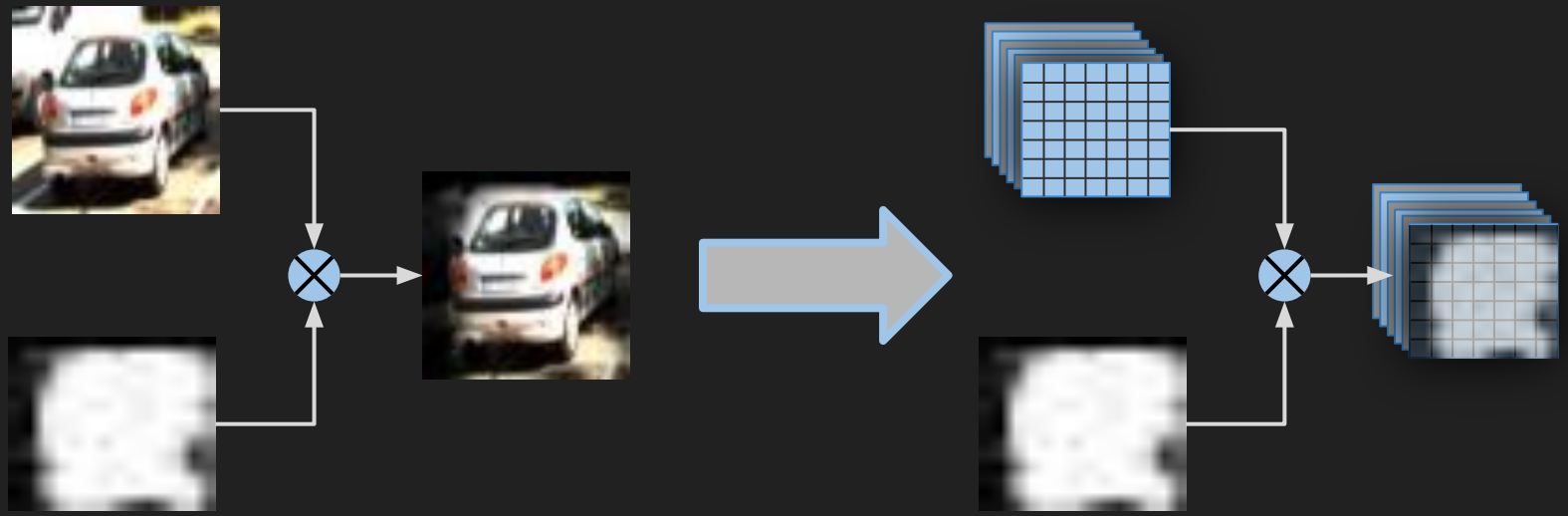
Hierarchical Attentive Recurrent Tracking



- Repeat

Hierarchical Attentive Recurrent Tracking

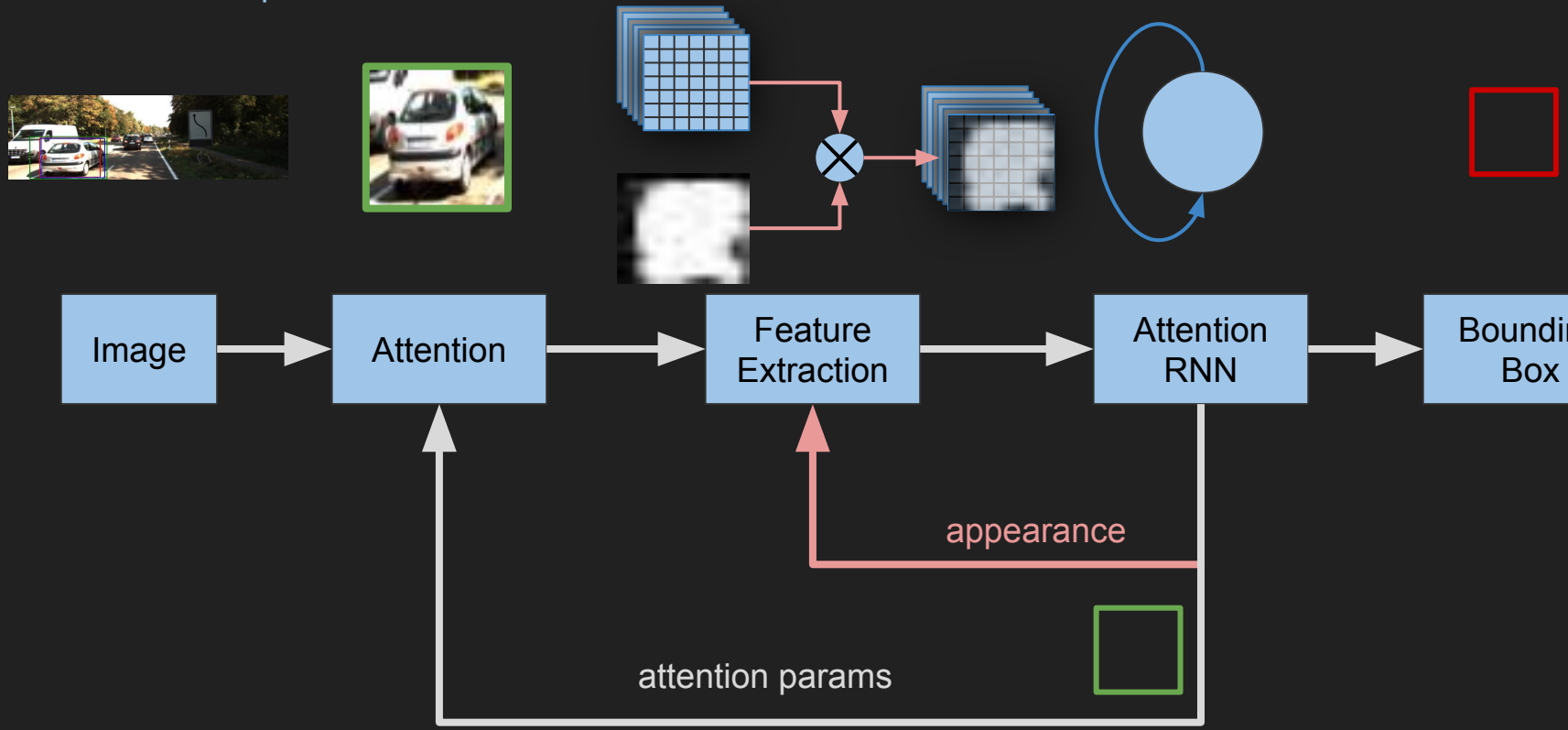
Feature Attention



- Detect the object within an attention glimpse
  - Reduces noise
  - Makes it easier to remember how the object looks like (no appearance drift over time)
- Mask in the feature space

Hierarchical Attentive Recurrent Tracking

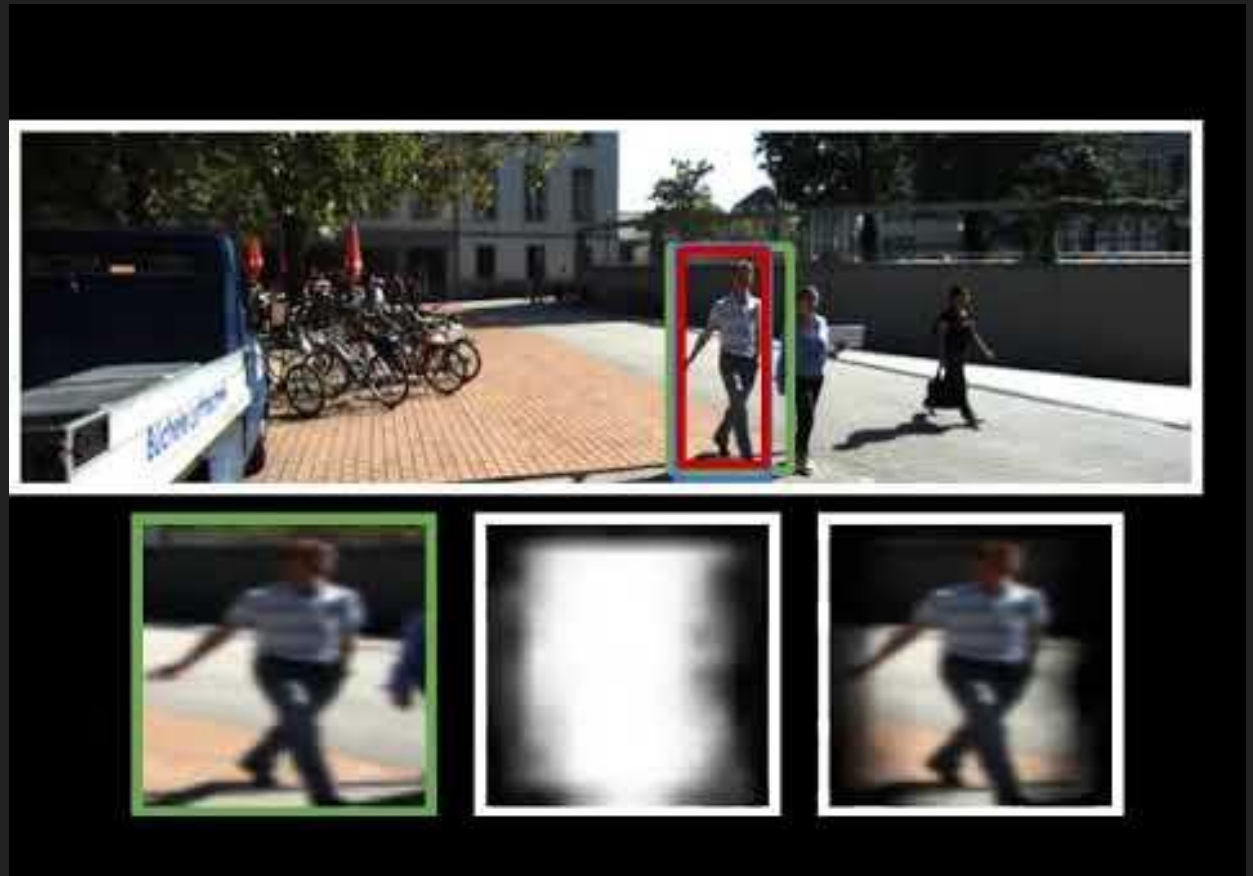
Full Model: Spatial & Feature Attention



Kosiorek *et. al.*, “Hierarchical Attentive Recurrent Tracking”, NIPS 2017

Hierarchical Attentive Recurrent Tracking

Tracking pedestrians, cyclists and cars

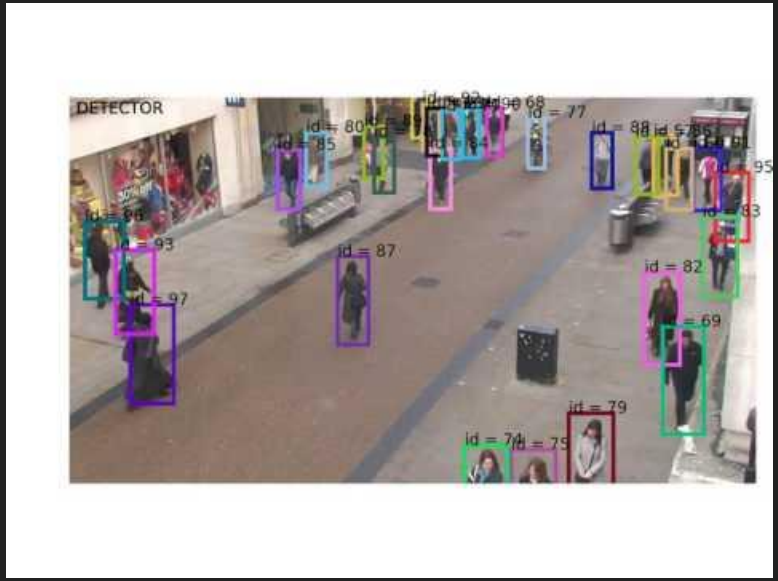


Ground-truth

Predicted

Attention

Object Tracking: can we do better?



- Can we track multiple objects at once?
- Can we do without supervision?
- Can we generate moving objects?

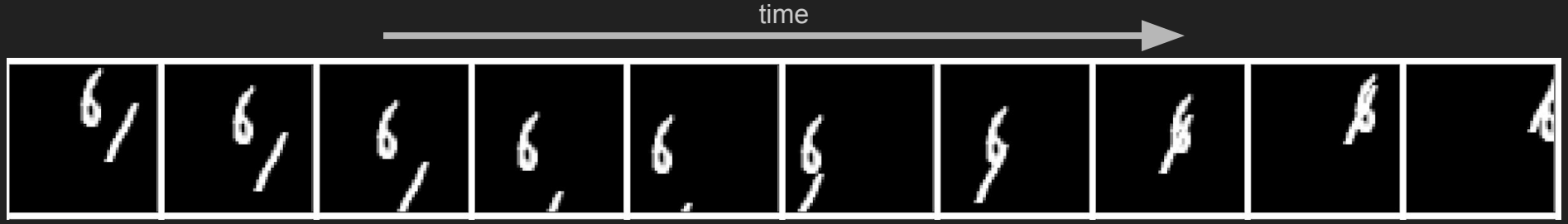
Bewley *et. al.*, “Simple Online and Realtime Tracking”, ICIP 2016.

Sequential Attend, Infer, Repeat (SQAIR)

Kosiorek *et. al.*, “Sequential Attend, Infer, Repeat”, NIPS 2018

Sequential Attend, Infer, Repeat (SQAIR)

Generative & Unsupervised with strong Object Priors



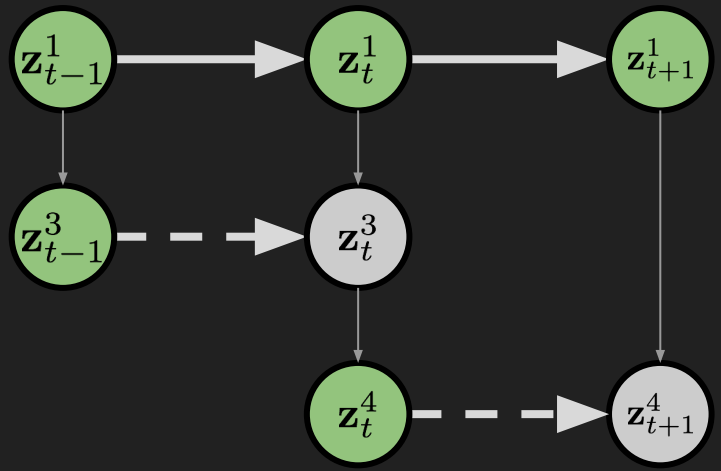
- Objects tend to be spatially and temporally consistent
  - They don’t appear out of nothing
  - They don’t disappear suddenly
  - No teleportation
- Common intuition, but how do we encode it into an ML model?

Kosiorek *et. al.*, “Sequential Attend, Infer, Repeat”, NIPS 2018

Sequential Attend, Infer, Repeat (SQAIR)

Spatial & Temporal Consistency of Objects

- If it moves together, it belongs together
- Model every object by a separate glimpse
  - Spatial consistency
- An objects depends on:
  - *A previous version of itself (strongly)*
  - *Other objects (weakly)*
  - Temporal consistency



Kosiorek *et. al.*, “Sequential Attend, Infer, Repeat”, NIPS 2018

Sequential Attend, Infer, Repeat (SQAIR)

Unsupervised Object Detection & Tracking

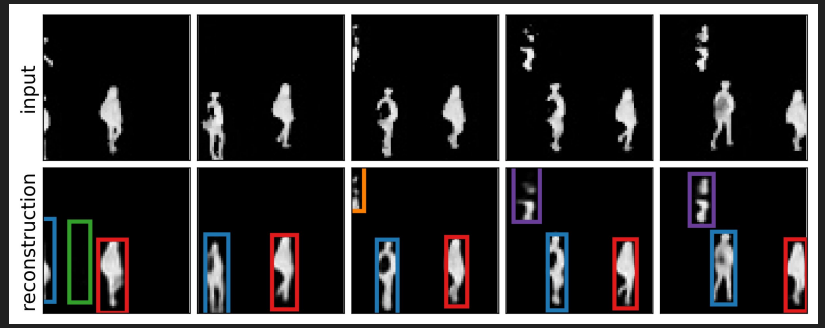
input output input output

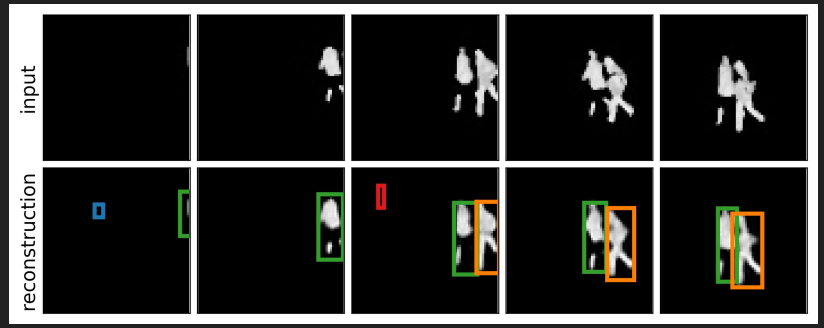
- Trained on 10 time-steps
- Here:
  - 100 time-steps
  - More noise in motion

Kosiorek *et. al.*, “Sequential Attend, Infer, Repeat”, NIPS 2018

Sequential Attend, Infer, Repeat (SQAIR)

Unsupervised Detection & Tracking of Pedestrians



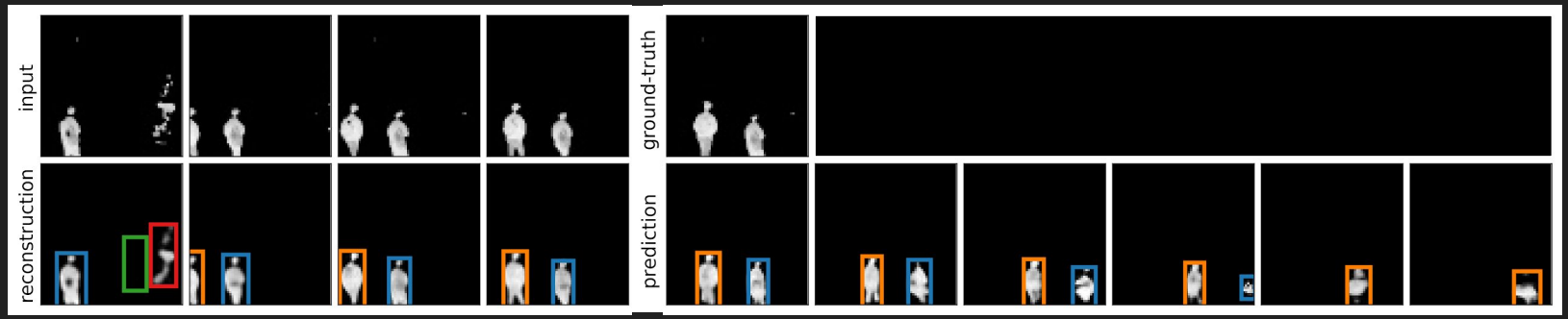


- We applied SQAIR to videos from static CCTV cameras
- It learns to reliably detect & track pedestrians
  - We can also predict future motions by sampling from the prior, as in MNIST

Kosiorek *et. al.*, “Sequential Attend, Infer, Repeat”, NIPS 2018

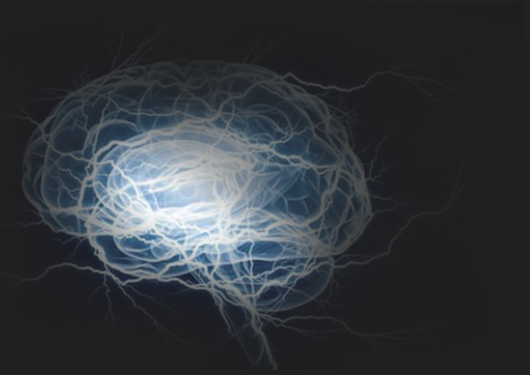
Sequential Attend, Infer, Repeat (SQAIR)

Conditional Generation



- Model trained only on 5 time-steps
- Learns to generate future
- Can be conditioned on initial frames

Kosiorek *et. al.*, “Sequential Attend, Infer, Repeat”, NIPS 2018



Questions?

More at my blog:

akosiorek.github.io

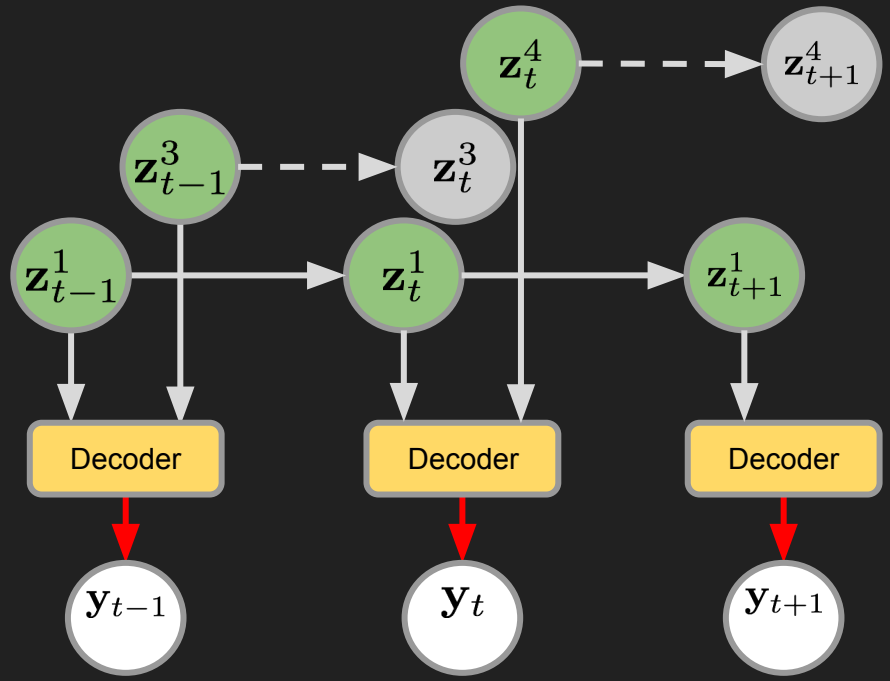
Code:

akosiorek.github.com

Sequential Attend, Infer, Repeat (SQAIR)

From Objects to Images

- Assume that every frame might have some new objects and some old ones
- Use separate latent variables for every object
- New objects should be discovered
- Old objects should be propagated or forgotten



Kosiorek *et. al.*, “Sequential Attend, Infer, Repeat”, NIPS 2018