# Learning Object-Centric Representations

Adam Roman Kosiorek

Wolfson College
University of Oxford

*A thesis submitted for the degree of*
*Doctor of Philosophy*

Michaelmas 2019

## Abstract

Your abstract text goes here. Check your departmental regulations, but generally this should be less than 300 words. See the beginning of Chapter 2 for more.

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Pellentesque sit amet nibh volutpat, scelerisque nibh a, vehicula neque. Integer placerat nulla massa, et vestibulum velit dignissim id. Ut eget nisi elementum, consectetur nibh in, condimentum velit. Quisque sodales dui ut tempus mattis. Duis malesuada arcu at ligula egestas egestas. Phasellus interdum odio at sapien fringilla scelerisque. Mauris sagittis eleifend sapien, sit amet laoreet felis mollis quis. Pellentesque dui ante, finibus eget blandit sit amet, tincidunt eu neque. Vivamus rutrum dapibus ligula, ut imperdiet lectus tincidunt ac. Pellentesque ac lorem sed diam egestas lobortis.

Suspendisse leo purus, efficitur mattis urna a, maximus molestie nisl. Aenean porta semper tortor a vestibulum. Suspendisse viverra facilisis lorem, non pretium erat lacinia a. Vestibulum tempus, quam vitae placerat porta, magna risus euismod purus, in viverra lorem dui at metus. Sed ac sollicitudin nunc. In maximus ipsum nunc, placerat maximus tortor gravida varius. Suspendisse pretium, lorem at porttitor rhoncus, nulla urna condimentum tortor, sed suscipit nisi metus ac risus.

Aenean sit amet enim quis lorem tristique commodo vitae ut lorem. Duis vel tincidunt lacus. Sed massa velit, lacinia sed posuere vitae, malesuada vel ante. Praesent a rhoncus leo. Etiam sed rutrum enim. Pellentesque lobortis elementum augue, at suscipit justo malesuada at. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Praesent rhoncus convallis ex. Etiam commodo nunc ex, non consequat diam consectetur ut. Pellentesque vitae est nec enim interdum dapibus. Donec dapibus purus ipsum, eget tincidunt ex gravida eget. Donec luctus nisi eu fringilla mollis. Donec eget lobortis diam.

Suspendisse finibus placerat dolor. Etiam ornare elementum ex ut vehicula. Donec accumsan mattis erat. Quisque cursus fringilla diam, eget placerat neque bibendum eu. Ut faucibus dui vitae dolor porta, at elementum ipsum semper. Sed ultrices dui non arcu pellentesque placerat. Etiam posuere malesuada turpis, nec malesuada tellus malesuada.

# Learning Object-Centric Representations


Adam Roman Kosiorek

Wolfson College

University of Oxford

A thesis submitted for the degree of

*Doctor of Philosophy*

Michaelmas 2019

# Acknowledgements


## Personal

This is where you thank your advisor, colleagues, and family and friends.

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Vestibulum feugiat et est at accumsan. Praesent sed elit mattis, congue mi sed, porta ipsum. In non ullamcorper lacus. Quisque volutpat tempus ligula ac ultricies. Nam sed erat feugiat, elementum dolor sed, elementum neque. Aliquam eu iaculis est, a sollicitudin augue. Cras id lorem vel purus posuere tempor. Proin tincidunt, sapien non dictum aliquam, ex odio ornare mauris, ultrices viverra nisi magna in lacus. Fusce aliquet molestie massa, ut fringilla purus rutrum consectetur. Nam non nunc tincidunt, rutrum dui sit amet, ornare nunc. Donec cursus tortor vel odio molestie dignissim. Vivamus id mi erat. Duis porttitor diam tempor rutrum porttitor. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed condimentum venenatis consectetur. Lorem ipsum dolor sit amet, consectetur adipiscing elit.

Aenean sit amet lectus nec tellus viverra ultrices vitae commodo nunc. Mauris at maximus arcu. Aliquam varius congue orci et ultrices. In non ipsum vel est scelerisque efficitur in at augue. Nullam rhoncus orci velit. Duis ultricies accumsan feugiat. Etiam consectetur ornare velit et eleifend.

Suspendisse sed enim lacinia, pharetra neque ac, ultricies urna. Phasellus sit amet cursus purus. Quisque non odio libero. Etiam iaculis odio a ex volutpat, eget pulvinar augue mollis. Mauris nibh lorem, mollis quis semper quis, consequat nec metus. Etiam dolor mi, cursus a ipsum aliquam, eleifend venenatis ipsum. Maecenas tempus, nibh eget scelerisque feugiat, leo nibh lobortis diam, id laoreet purus dolor eu mauris. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Nulla eget tortor eu arcu sagittis euismod fermentum id neque. In sit amet justo ligula. Donec rutrum ex a aliquet egestas.

## Institutional

If you want to separate out your thanks for funding and institutional support, I don't think there's any rule against it. Of course, you could also just remove the subsections and do one big traditional acknowledgement section.

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut luctus tempor ex at pretium. Sed varius, mauris at dapibus lobortis, elit purus tempor neque,

facilisis sollicitudin felis nunc a urna. Morbi mattis ante non augue blandit pulvinar. Quisque nec euismod mauris. Nulla et tellus eu nibh auctor malesuada quis imperdiet quam. Sed eget tincidunt velit. Cras molestie sem ipsum, at faucibus quam mattis vel. Quisque vel placerat orci, id tempor urna. Vivamus mollis, neque in aliquam consequat, dui sem volutpat lorem, sit amet tempor ipsum felis eget ante. Integer lacinia nulla vitae felis vulputate, at tincidunt ligula maximus. Aenean venenatis dolor ante, euismod ultrices nibh mollis ac. Ut malesuada aliquam urna, ac interdum magna malesuada posuere.

# Abstract

Your abstract text goes here. Check your departmental regulations, but generally this should be less than 300 words. See the beginning of Chapter 2 for more.

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Pellentesque sit amet nibh volutpat, scelerisque nibh a, vehicula neque. Integer placerat nulla massa, et vestibulum velit dignissim id. Ut eget nisi elementum, consectetur nibh in, condimentum velit. Quisque sodales dui ut tempus mattis. Duis malesuada arcu at ligula egestas egestas. Phasellus interdum odio at sapien fringilla scelerisque. Mauris sagittis eleifend sapien, sit amet laoreet felis mollis quis. Pellentesque dui ante, finibus eget blandit sit amet, tincidunt eu neque. Vivamus rutrum dapibus ligula, ut imperdiet lectus tincidunt ac. Pellentesque ac lorem sed diam egestas lobortis.

Suspendisse leo purus, efficitur mattis urna a, maximus molestie nisl. Aenean porta semper tortor a vestibulum. Suspendisse viverra facilisis lorem, non pretium erat lacinia a. Vestibulum tempus, quam vitae placerat porta, magna risus euismod purus, in viverra lorem dui at metus. Sed ac sollicitudin nunc. In maximus ipsum nunc, placerat maximus tortor gravida varius. Suspendisse pretium, lorem at porttitor rhoncus, nulla urna condimentum tortor, sed suscipit nisi metus ac risus.

Aenean sit amet enim quis lorem tristique commodo vitae ut lorem. Duis vel tincidunt lacus. Sed massa velit, lacinia sed posuere vitae, malesuada vel ante. Praesent a rhoncus leo. Etiam sed rutrum enim. Pellentesque lobortis elementum augue, at suscipit justo malesuada at. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Praesent rhoncus convallis ex. Etiam commodo nunc ex, non consequat diam consectetur ut. Pellentesque vitae est nec enim interdum dapibus. Donec dapibus purus ipsum, eget tincidunt ex gravida eget. Donec luctus nisi eu fringilla mollis. Donec eget lobortis diam.

Suspendisse finibus placerat dolor. Etiam ornare elementum ex ut vehicula. Donec accumsan mattis erat. Quisque cursus fringilla diam, eget placerat neque bibendum eu. Ut faucibus dui vitae dolor porta, at elementum ipsum semper. Sed ultrices dui non arcu pellentesque placerat. Etiam posuere malesuada turpis, nec malesuada tellus malesuada.

# Contents

# List of Figures

# List of Abbreviations

**1-D, 2-D** . . . One- or two-dimensional, referring in this thesis to spatial dimensions in an image.

**Otter** . . . . . One of the finest of water mammals.

**Hedgehog** . . . Quite a nice prickly friend.

*Neque porro quisquam est qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit...*

*There is no one who loves pain itself, who seeks after it and wants to have it, simply because it is pain...*

— Cicero's *de Finibus Bonorum et Malorum*

# 1 Introduction

## Contents



## 1.1 Motivation

The rapid advance of minimally-invasive cardiac procedures promises improvements in patient safety, procedure efficacy, and access to treatment. While percutaneous coronary intervention (PCI) has become routine and highly effective **bravata_systematic_2007**, catheter procedures in areas such as electrophysiology (EP) and valve replacement are still coming of age. This progress is driven by demographics and the improvement in general cardiac care, as patients surviving initial cardiac events go on to require treatment for sequelae **foot_demographics_2000**. The growing need for advanced treatment is being answered by developments in catheter technology and procedures. These tools are continually advancing to access and manipulate an ever-broader range of anatomy **sousa_new_2005**.

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Maecenas sagittis dolor at nulla feugiat, vitae iaculis est rutrum. Mauris eu sem eros. Sed id faucibus urna. In egestas eros et sapien egestas imperdiet. In hac habitasse platea dictumst.

Phasellus vitae varius tortor. Mauris nec sollicitudin enim. Suspendisse molestie leo nec mauris molestie, nec imperdiet magna vehicula. Phasellus sodales tortor dui, a lacinia turpis congue at. Pellentesque mattis dui non libero commodo, at accumsan ex ultrices. Integer eget ex eget dui cursus euismod et accumsan felis. Nullam laoreet sodales dui, ut finibus elit varius a. Sed elementum orci quis libero ullamcorper, eget egestas enim convallis. Sed nibh libero, tincidunt ultricies nibh quis, lobortis placerat mauris. Maecenas at laoreet risus, nec dictum libero. Donec accumsan, orci eu tempus mattis, nisl arcu auctor turpis, ac sollicitudin justo orci nec nulla.

Nam eget sem sed ligula vehicula iaculis. In non arcu a nisl interdum gravida. Nam egestas erat non turpis sagittis vestibulum. Praesent est metus, facilisis eu commodo sed, sagittis et est. Duis scelerisque luctus erat, elementum pulvinar felis bibendum a. Morbi hendrerit rhoncus consectetur. Vestibulum nec odio finibus, blandit turpis eget, dignissim orci. Curabitur eu ligula auctor, porttitor nulla non, maximus turpis. Nunc sed quam at est varius interdum eu vitae odio. Vestibulum egestas dapibus nulla sit amet fermentum.

Vestibulum ut neque urna. Ut nec odio lobortis, ultricies nulla quis, ultricies tellus. Nam ac iaculis sapien. Vivamus vitae risus id tortor interdum pellentesque. Quisque lorem lectus, sagittis vel metus et, sagittis finibus justo. Curabitur pulvinar odio tellus, eu vehicula est dictum eget. Morbi sed justo justo. Vivamus enim nibh, facilisis pretium luctus quis, ullamcorper quis ipsum. Pellentesque a mi a elit euismod malesuada.

Vestibulum interdum est vel orci tincidunt auctor. Nunc tristique nulla nec blandit fermentum. Maecenas id libero ut justo dictum sodales. Nullam justo sapien, dignissim vel enim at, porta pharetra metus. Integer euismod quam eget ligula gravida euismod. Pellentesque commodo, quam sit amet bibendum tempor, nisi odio varius mauris, et accumsan justo ex sed nunc. Cras bibendum nibh ac dolor volutpat, non elementum orci pulvinar. Maecenas et porttitor nulla. Suspendisse sapien massa, dapibus at blandit et, rhoncus suscipit velit. Fusce molestie, velit eget sagittis suscipit, est libero aliquam libero, in iaculis mi tellus ac nunc.

## 1.2 Contribution

Sed in rhoncus lectus. Mauris vulputate purus non malesuada pulvinar. Curabitur ullamcorper hendrerit elit, id vulputate libero sagittis vel. Pellentesque ac faucibus est. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos himenaeos. Integer venenatis, nisl eleifend pellentesque consequat, sem tortor malesuada ante, ut tincidunt elit tortor sit amet nunc.

Cras vehicula ipsum sit amet dui rutrum ultrices. Integer eu eleifend odio. Praesent tempor, libero id ullamcorper euismod, lectus diam lobortis mauris, id venenatis arcu sem vitae purus. Pellentesque luctus tristique metus quis mollis. Praesent ullamcorper neque velit, sed iaculis est convallis sit amet. Quisque nec massa ut magna lobortis imperdiet. Quisque rhoncus purus eget mollis aliquet. Donec vehicula viverra nisl, sed posuere turpis vulputate non. Donec malesuada, eros id interdum volutpat, ipsum orci luctus quam, non pulvinar urna ipsum eget purus. Nam hendrerit condimentum tristique.

Proin metus velit, tempor at fringilla non, dictum eu felis. Proin volutpat enim ut fermentum aliquam. Nam dictum nisi eu nisl viverra fermentum. Pellentesque tristique arcu non orci congue faucibus. Fusce sit amet nisl fringilla, feugiat turpis vitae, eleifend ante. Suspendisse elementum, lectus non pulvinar bibendum, lectus massa faucibus turpis, vitae porta risus sem quis metus. Maecenas id sapien et dui lobortis imperdiet nec eu mi. Quisque porttitor tincidunt nisi, eget sagittis orci. Nunc mattis erat malesuada facilisis viverra. Maecenas sodales iaculis nisi vel tincidunt. Morbi aliquet nibh ac facilisis consectetur. In ultrices libero quis massa porttitor cursus. Quisque suscipit ac tortor eget aliquet. Ut eget lacus vel orci viverra maximus at at purus.

Nam massa neque, varius nec suscipit id, cursus ac mi. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. In hac habitasse platea dictumst. Vivamus facilisis nunc quis dictum consectetur. Sed congue sed magna non auctor. Vestibulum accumsan sit amet erat non congue. Sed at condimentum mi, sed scelerisque urna. Etiam tristique pulvinar rutrum. Donec

semper nulla vitae rutrum semper. Maecenas ultrices nibh at orci sodales tincidunt sit amet vitae arcu. Curabitur interdum tincidunt ipsum, nec tincidunt nunc dapibus in. Nunc sit amet elementum massa, ut ornare lacus. Vivamus convallis fringilla erat, non suscipit sapien convallis eu. Nunc viverra lectus sit amet turpis viverra, eget iaculis purus rhoncus.

Morbi eu lectus arcu. Sed fringilla dui ut magna commodo, a malesuada ante pellentesque. Donec ornare facilisis pellentesque. Nulla vitae fringilla velit. Nunc id tellus nisl. Maecenas pretium elit lectus, nec consectetur nunc vulputate et. Sed facilisis magna nec gravida hendrerit. Sed a cursus nisl, in rhoncus massa. Curabitur ut nibh interdum, tempor risus vel, scelerisque nibh. Mauris quis ipsum sed risus tempor convallis ut a eros.

*Alles Gescheite ist schon gedacht worden.*
*Man muss nur versuchen, es noch einmal zu denken.*

*All intelligent thoughts have already been thought;*
*what is necessary is only to try to think them again.*

— Johann Wolfgang von Goethe
**von_goethe_wilhelm_1829**

# 2 Background

## Contents



## 2.1 Introduction

This document introduction won't serve as a complete primer on LaTeX. There are plenty of those online, and googling your questions will often get you answers, especially from `http://tex.stackexchange.com`.

Instead, let's talk a little about a few of the features and packages lumped into this template situation. The `savequote` environment at the beginning of chapters can add some wittiness to your thesis. If you don't like the quotes, just remove that block.

For when it comes time to do corrections, there are two useful commands here. First, the `mccorrect` command allows you to highlight a short correction like this one. When the thesis is typeset normally, the correction will just appear as part of the text. However, when you declare `\correctionstrue` in the main `Oxford_Thesis.tex` file, that correction will be highlighted in blue. That might

be useful for submitting a post-viva, corrected copy to your examiners so they can quickly verify you've completed the task.

For larger chunks, like this paragraph or indeed entire figures, you can use the `mccorrection` environment. This environment highlights paragraph-sized and larger blocks with the same blue colour.

Read through the `Oxford_Thesis.tex` file to see the various options for one- and two-sided printing, including or excluding the separate abstract page, and turning corrections and draft footer on or off, and the separate option to centre your text on the page (for PDF submission) or offset it (for binding). There is also a separate option for master's degree submissions, which changes identifying information to candidate number and includes a word count. (Unfortunately, LaTeXhas a hard time doing word counts automatically, so you'll have to enter the count manually if you require this.)

## 2.2 Cardiac Imaging

Within months of Röntgen's discovery of the X-ray in 1895**gagliardi__rontgen__1996**, cardiac pathology was being investigated via non-invasive imaging **gagliardi__cardiac__1996**. Over the intervening years, cardiac imaging modalities and techniques have advanced significantly. Clinically, cardiac imaging is used for two broad purposes: diagnosis of pathophysiology and guidance of interventional procedures. These applications impose different requirements on imaging equipment, image acquisition time, computational complexity, spatial and temporal resolution, and tissue discrimination. The common diagnostic and interventional cardiac imaging techniques in current clinical practice are reviewed below. An accessible introduction to the physics of medical imaging can be found in Webb's *Introduction to Biomedical Imaging* **webb__introduction__2002**. A comprehensive overview of the use of imaging in clinical cardiology is presented in Leeson's *Cardiovascular Imaging* **leeson__cardiovascular__2011**.

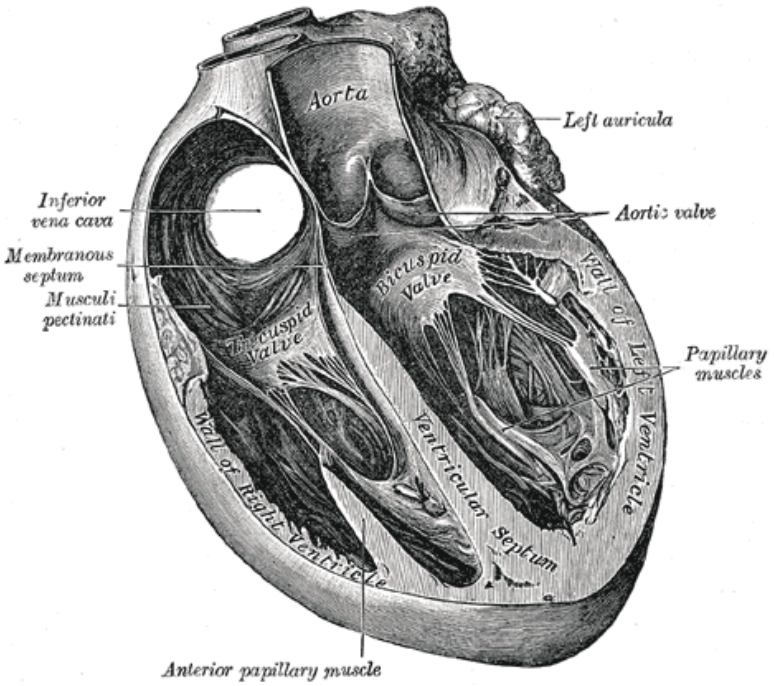



**Figure 2.1:** Four-chamber illustration of the human heart. Clockwise from upper-left: right atrium, left atrium, left ventricle, right ventricle.

### 2.2.1 Diagnostic Imaging

Beyond the chest X-ray ('plain film'), the key non-invasive imaging modalities in diagnostic cardiology are echocardiography, magnetic resonance imaging, and X-ray computed tomography, which are reviewed below. Nuclear medicine, including positron emission tomography (PET) and single-photon emission computed tomography (SPECT), are not discussed here, as they do not play a role in the chapters to follow.

#### Echocardiography

The use of acoustic waves for medical diagnosis, inspired by naval sonar, was initially developed in the 1940s **gagliardi_ultrasonography_1996**. By 1954, the first clinically useful cardiac ultrasound – examining motion of the mitral valve in stenosis – was reported **edler_ultrasonic_1957**. These early scans were one-dimensional images ('A-mode'), sometimes repeated to generate a time axis ('M-mode'). The sector-scanning probe was developed in the 1970s **bom_ultrasonic_1971**; **griffith_sector_1974**,

leading to the ‘B-mode’ that a modern cardiologist would recognise as an echocardiogram.

# 3

# Hierarchical Attentive Recurrent Tracking

# Abstract

Class-agnostic object tracking is particularly difficult in cluttered environments as target specific discriminative models cannot be learned *a priori*. Inspired by how the human visual cortex employs spatial attention and separate "where" and "what" processing pathways to actively suppress irrelevant visual features, this work develops a hierarchical attentive recurrent model for single object tracking in videos. The first layer of attention discards the majority of background by selecting a region containing the object of interest, while the subsequent layers tune in on visual features *particular* to the tracked object. This framework is fully differentiable and can be trained in a purely data driven fashion by gradient methods. To improve training convergence, we augment the loss function with terms for auxiliary tasks relevant for tracking. Evaluation of the proposed model is performed on two datasets: pedestrian tracking on the KTH activity recognition dataset and the more difficult KITTI object tracking dataset.

## 3.1 Introduction

In computer vision, designing an algorithm for model-free tracking of anonymous objects is challenging, since no target-specific information can be gathered *a priori* and yet the algorithm has to handle target appearance changes, varying lighting conditions and occlusion. To make it even more difficult, the tracked object often constitutes but a small fraction of the visual field. The remaining parts may contain *distractors*, which are visually salient objects resembling the target but hold no relevant information. Despite this fact, recent models often process the whole image, which exposes them to noise and increases the associated computational cost or they use heuristic methods to decrease the size of search regions. This in contrast to human visual perception, which does not process the visual field in its entirety, but rather acknowledges it briefly and focuses on processing small fractions thereof, which we dub *visual attention*.

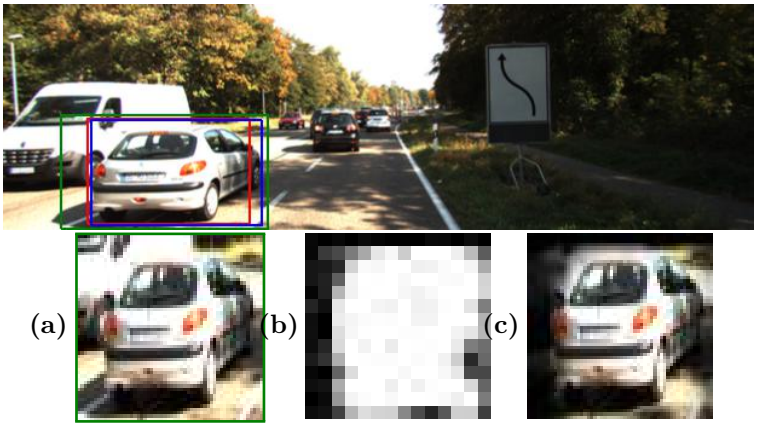



**Figure 3.1:** KITTI image with the ground-truth and predicted bounding boxes and an attention glimpse. The lower row corresponds to the hierarchical attention of our model: $1^{st}$ layer extracts an attention glimpse (a), the $2^{nd}$ layer uses appearance attention to build a location map (b). The $3^{rd}$ layer uses the location map to suppress distractors, visualised in (c).

Attention mechanisms have recently been explored in machine learning in a wide variety of contexts **Vinyals2014**; **Jaderberg2015**, often providing new capabilities to machine learning algorithms **Graves2016**; **Wierstra2015draw**; **Eslami2016**. While they improve efficiency **Graves2014recurrent** and performance on state-of-the-art machine learning benchmarks **Vinyals2014**, their architecture is much simpler than that of the mechanisms found in the human visual cortex **Dayan2001**. Attention has also been long studied by neuroscientists **Ungerleider2000**, who

believe that it is crucial for visual perception and cognition **Olshausen2016foveal**, since it is inherently tied to the architecture of the visual cortex and can affect the information flow inside it. Whenever more than one visual stimulus is present in the receptive field of a neuron, all the stimuli compete for computational resources due to the limited processing capacity. Visual attention can lead to suppression of distractors by reducing the size of the receptive field of a neuron and by increasing sensitivity at a given location in the visual field (*spatial attention*). It can also amplify activity in different parts of the cortex, which are specialised in processing different types of features, leading to response enhancement with respect to those features (*appearance attention*). The functional separation of the visual cortex is most apparent in two distinct processing pathways. After leaving the eye, the sensory inputs enter the primary visual cortex (known as *V1*) and then split into the *dorsal stream*, responsible for estimating spatial relationships (*where*), and the *ventral stream*, which targets appearance-based features (*what*).

Inspired by the general architecture of the human visual cortex and the role of attention mechanisms, this work presents a biologically-inspired recurrent model for single object tracking in videos (*cf.* Section 3.3). Tracking algorithms typically use simple motion models and heuristics to decrease the size of the search region. It is interesting to see whether neuroscientific insights can aid our computational efforts, thereby improving the efficiency and performance of single object tracking. It is worth noting that visual attention can be induced by the stimulus itself (due to, e. g., high contrast) in a *bottom-up* fashion or by back-projections from other brain regions and working memory as *top-down* influence. The proposed approach exploits this property to create a feedback loop that steers the *three* layers of visual attention mechanisms in our hierarchical attentive recurrent tracking (*HART*) framework, see Figure 3.1. The first stage immediately discards spatially irrelevant input, while later stages focus on producing target-specific filters to emphasise visual features *particular* to the object of interest.

The resulting framework is end-to-end trainable and we resort to maximum likelihood estimation (MLE) for parameter learning. This follows from our interest

in estimating the distribution over object locations in a sequence of images, given the initial location from whence our tracking commenced. Formally, given a sequence of images $\mathbf{x}_{1:T} \in \mathbb{R}^{H \times W \times C}$, where the superscript denotes height, width and the number of channels of the image, respectively, and an initial location for the tracked object given by a bounding box $\mathbf{b}_1 \in \mathbb{R}^4$, the conditional probability distribution factorises as

$$p(\mathbf{b}_{2:T} \mid \mathbf{x}_{1:T}, \mathbf{b}_1) = \int p(\boldsymbol{h}_1 \mid \mathbf{x}_1, \mathbf{b}_1) \prod_{t=2}^{T} \int p(\mathbf{b}_t \mid \boldsymbol{h}_t) p(\boldsymbol{h}_t \mid \mathbf{x}_t, \mathbf{b}_{t-1}, \boldsymbol{h}_{t-1}) \, \mathrm{d}\boldsymbol{h}_t \, \mathrm{d}\boldsymbol{h}_1, \tag{3.1}$$

where we assume that motion of an object can be described by a Markovian state $\boldsymbol{h}_t$. Our bounding box estimates are given by $\widehat{\mathbf{b}}_{2:T}$, found by the MLE of the model parameters. In sum, our contributions are threefold: Firstly, a hierarchy of attention mechanisms that leads to suppressing distractors and computational efficiency is introduced. Secondly, a biologically plausible combination of attention mechanisms and recurrent neural networks is presented for object tracking. Finally, our attention-based tracker is demonstrated using real-world sequences in challenging scenarios where previous recurrent attentive trackers have failed.

Next we briefly review related work (Section 3.2) before describing how information flows through the components of our hierarchical attention in Section 3.3. Section 3.4 details the losses applied to guide the attention. Section 3.5 presents experiments on KTH and KITTI datasets with comparison to related attention-based trackers. Section 4.6 discusses the results and intriguing properties of our framework and Section 3.7 concludes the work. Code and results are available online[1].

## 3.2 Related Work

A number of recent studies have demonstrated that visual content can be captured through a sequence of spatial glimpses or foveation **Graves2014recurrent**; **Wierstra2015draw**. Such a paradigm has the intriguing property that the computational complexity is proportional to the number of steps as opposed

[1] `https://github.com/akosiorek/hart`

to the image size. Furthermore, the fovea centralis in the retina of primates is structured with maximum visual acuity in the centre and decaying resolution towards the periphery, **Olshausen2016foveal** show that if spatial attention is capable of zooming, a regular grid sampling is sufficient. **Jaderberg2015** introduced the spatial transformer network (STN) which provides a fully differentiable means of transforming feature maps, conditioned on the input itself. **Eslami2016** use the STN as a form of attention in combination with a recurrent neural network (RNN) to sequentially locate and identify objects in an image. Moreover, **Eslami2016** use a latent variable to estimate the presence of additional objects, allowing the RNN to adapt the number of time-steps based on the input. Our spatial attention mechanism is based on the two dimensional Gaussian grid filters of **Kahou2015ratm** which is both fully differentiable and more biologically plausible than the STN.

Whilst focusing on a specific location has its merits, focusing on particular appearance features might be as important. A policy with feedback connections can learn to adjust filters of a convolutional neural network (CNN), thereby adapting them to features present in the current image and improving accuracy **Stollenga2014**. **Brabandere2016dfn** introduced dynamic filter network (DFN), where filters for a CNN are computed on-the-fly conditioned on input features, which can reduce model size without performance loss. **Karl2017** showed that an input-dependent state transitions can be helpful for learning latent Markovian state-space system. While not the focus of this work, we follow this concept in estimating the expected appearance of the tracked object.

In the context of single object tracking, both attention mechanisms and RNNs appear to be perfectly suited, yet their success has mostly been limited to simple monochromatic sequences with plain backgrounds **Kahou2015ratm**. **Cheung2016gtc** applied STNs **Jaderberg2015** as attention mechanisms for real-world object tracking, but failed due to exploding gradients potentially arising from the difficulty of the data. **Ning2016** achieved competitive performance by using features from an object detector as inputs to a long-short memory network (LSTM), but requires processing of the whole image at each time-step.

Two recent state-of-the-art trackers employ convolutional Siamese networks which can be seen as an RNN unrolled over two time-steps **Held2016**; **Valmadre2017**. Both methods explicitly process small search areas around the previous target position to produce a bounding box offset **Held2016** or a correlation response map with the maximum corresponding to the target position **Valmadre2017**. We acknowledge the recent work[2] of **Gordon2017** which employ an RNN based model and use explicit cropping and warping as a form of non-differentiable spatial attention. The work presented in this paper is closest to **Kahou2015ratm** where we share a similar spatial attention mechanism which is guided through an RNN to effectively learn a motion model that spans multiple time-steps. The next section describes our additional attention mechanisms in relation to their biological counterparts.

## 3.3 Hierarchical Attention

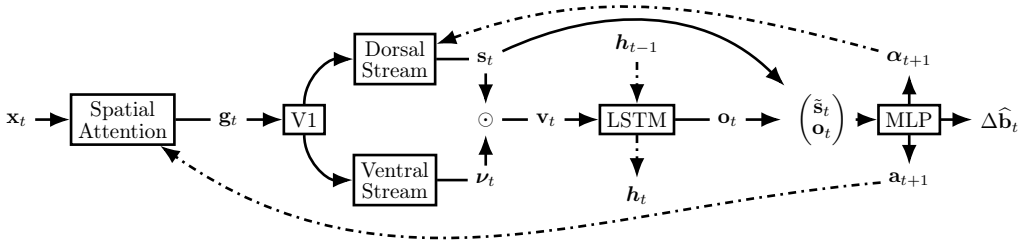



**Figure 3.2:** Hierarchical Attentive Recurrent Tracking. Spatial attention extracts a glimpse $\mathbf{g}_t$ from the input image $\mathbf{x}_t$. V1 and the ventral stream extract appearance-based features $\boldsymbol{\nu}_t$ while the dorsal stream computes a foreground/background segmentation $\mathbf{s}_t$ of the attention glimpse. Masked features $\mathbf{v}_t$ contribute to the working memory $\boldsymbol{h}_t$. The LSTM output $\mathbf{o}_t$ is then used to compute attention $\mathbf{a}_{t+1}$, appearance $\boldsymbol{\alpha}_{t+1}$ and a bounding box correction $\Delta\widehat{\mathbf{b}}_t$. Dashed lines correspond to temporal connections, while solid lines describe information flow within one time-step.

Inspired by the architecture of the human visual cortex, we structure our system around working memory responsible for storing the motion pattern and an

[2] **Gordon2017** only became available at the time of submitting this paper.

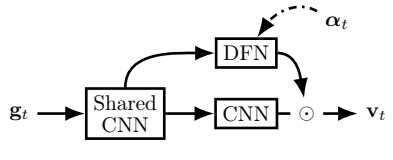



**Figure 3.3:** Architecture of the appearance attention. V1 is implemented as a CNN shared among the dorsal stream (DFN) and the ventral stream (CNN). The $\odot$ symbol represents the Hadamard product and implements masking of visual features by the foreground/background segmentation.

appearance description of the tracked object. If both quantities were known, it would be possible to compute the expected location of the object at the next time step. Given a new frame, however, it is not immediately apparent which visual features correspond to the appearance description. If we were to pass them on to an RNN, it would have to implicitly solve a data association problem. As it is non-trivial, we prefer to model it explicitly by outsourcing the computation to a separate processing stream conditioned on the expected appearance. This results in a location-map, making it possible to neglect features inconsistent with our memory of the tracked object. We now proceed with describing the information flow in our model.

Given attention parameters $\mathbf{a}_t$, the *spatial attention* module extracts a glimpse $\mathbf{g}_t$ from the input image $\mathbf{x}_t$. We then apply *appearance attention*, parametrised by appearance $\boldsymbol{\alpha}_t$ and comprised of V1 and dorsal and ventral streams, to obtain object-specific features $\mathbf{v}_t$, which are used to update the hidden state $\boldsymbol{h}_t$ of an LSTM. The LSTM's output is then decoded to predict both spatial and appearance attention parameters for the next time-step along with a bounding box correction $\Delta\widehat{\mathbf{b}}_t$ for the current time-step. Spatial attention is driven by top-down signal $\mathbf{a}_t$, while appearance attention depends on top-down $\boldsymbol{\alpha}_t$ as well as bottom-up (contents of the glimpse $\mathbf{g}_t$) signals. Bottom-up signals have local influence and depend on stimulus salience at a given location, while top-down signals incorporate global context into local processing. This attention hierarchy, further enhanced by recurrent connections, mimics that of the human visual cortex **Ungerleider2000**. We now describe the individual components of the system.

**Spatial Attention** Our spatial attention mechanism is similar to the one used by **Kahou2015ratm**. Given an input image $\mathbf{x}_t \in \mathbb{R}^{H\times W}$, it creates two matrices $\mathbf{A}_t^x \in \mathbb{R}^{w\times W}$ and $\mathbf{A}_t^y \in \mathbb{R}^{h\times H}$, respectively. Each matrix contains one Gaussian per row; the width and positions of the Gaussians determine which parts of the image are extracted as the attention glimpse. Formally, the glimpse $\mathbf{g}_t \in \mathbb{R}^{h\times w}$ is defined as

$$\mathbf{g}_t = \mathbf{A}_t^y \mathbf{x}_t \left(\mathbf{A}_t^x\right)^\mathsf{T}. \tag{3.2}$$

Attention is described by centres $\mu$ of the Gaussians, their variances $\sigma^2$ and strides $\gamma$ between centers of Gaussians of consecutive rows of the matrix, one for each axis. In contrast to the work by **Kahou2015ratm**, only centres and strides are estimated from the hidden state of the LSTM, while the variance depends solely on the stride. This prevents excessive aliasing during training caused when predicting a small variance (compared to strides) leading to smoother convergence. The relationship between variance and stride is approximated using linear regression with polynomial basis functions (up to $4^{th}$ order) before training the whole system. The glimpse size we use depends on the experiment.

**Appearance Attention** This stage transforms the attention glimpse $\mathbf{g}_t$ into a fixed-dimensional vector $\mathbf{v}_t$ comprising appearance and spatial information about the tracked object. Its architecture depends on the experiment. In general, however, we implement V1 : $\mathbb{R}^{h\times w} \to \mathbb{R}^{h_v \times w_v \times c_v}$ as a number of convolutional and max-pooling layers. They are shared among later processing stages, which corresponds to the primary visual cortex in humans **Dayan2001**. Processing then splits into ventral and dorsal streams. The ventral stream is implemented as a CNN, and handles visual features and outputs feature maps $\boldsymbol{\nu}_t$. The dorsal stream, implemented as a DFN, is responsible for handling spatial relationships. Let MLP $\cdot$ denote a multi-layered perceptron. The dorsal stream uses appearance $\boldsymbol{\alpha}_t$ to dynamically compute convolutional filters $\boldsymbol{\psi}_t^{a\times b\times c\times d}$, where the superscript denotes the size of the filters and the number of input and output feature maps, as

$$\boldsymbol{\Psi}_t = \left\{\boldsymbol{\psi}_t^{a_i\times b_i\times c_i\times d_i}\right\}_{i=1}^{K} = \text{MLP}\,\boldsymbol{\alpha}_t. \tag{3.3}$$

The filters with corresponding nonlinearities form $K$ convolutional layers applied to the output of V1. Finally, a convolutional layer with a $1 \times 1$ kernel and a sigmoid non-linearity is applied to transform the output into a spatial Bernoulli distribution $\mathbf{s}_t$. Each value in $\mathbf{s}_t$ represents the probability of the tracked object occupying the corresponding location.

The location map of the dorsal stream is combined with appearance-based features extracted by the ventral stream, to imitate the distractor-suppressing behaviour of the human brain. It also prevents drift and allows occlusion handling, since object appearance is not overwritten in the hidden state when input does not contain features particular to the tracked object. Outputs of both streams are combined as[3]

$$\mathbf{v}_t = \text{MLP}\,\text{vec}(\boldsymbol{\nu}_t \odot \mathbf{s}_t), \tag{3.4}$$

with $\odot$ being the Hadamard product.

**State Estimation** Our approach relies on being able to predict future object appearance and location, and therefore it heavily depends on state estimation. We use an LSTM, which can learn to trade-off spatio-temporal and appearance information in a data-driven fashion. It acts like a working memory, enabling the system to be robust to occlusions and oscillating object appearance e. g., when an object rotates and comes back to the original orientation.

$$\mathbf{o}_t, \boldsymbol{h}_t = \text{LSTM}(\mathbf{v}_t, \boldsymbol{h}_{t-1}), \tag{3.5}$$

$$\boldsymbol{\alpha}_{t+1}, \Delta\mathbf{a}_{t+1}, \Delta\widehat{\mathbf{b}}_t = \text{MLP}\,\mathbf{o}_t, \text{vec}(\mathbf{s}_t), \tag{3.6}$$

$$\mathbf{a}_{t+1} = \mathbf{a}_t + \tanh(\boldsymbol{c})\Delta\mathbf{a}_{t+1}, \tag{3.7}$$

$$\widehat{\mathbf{b}}_t = \mathbf{a}_t + \Delta\widehat{\mathbf{b}}_t \tag{3.8}$$

Equations (3.5) to (3.8) detail the state updates. Spatial attention at time $t$ is formed as a cumulative sum of attention updates from times $t = 1$ to $t = T$, where $\boldsymbol{c}$ is a learnable parameter initialised to a small value to constrain the size of the updates early in training. Since the spatial-attention mechanism is trained to predict where the object is going to go (Section 3.4), the bounding box $\widehat{\mathbf{b}}_t$ is estimated relative to attention at time $t$.

[3] $\text{vec} : \mathbb{R}^{m\times n} \to \mathbb{R}^{mn}$ is the vectorisation operator, which stacks columns of a matrix into a column vector.

## 3.4 Loss

We train our system by minimising a loss function comprised of: a tracking loss term, a set of terms for auxiliary tasks and regularisation terms. Auxiliary tasks are essential for real-world data, since convergence does not occur without them. They also speed up learning and lead to better performance for simpler datasets. Unlike the auxiliary tasks used by **Jaderberg2016**, ours are relevant for our main objective — object tracking. In order to limit the number of hyperparameters, we automatically learn loss weighting. The loss $\mathcal{L}(\cdot)$ is given by

$$\mathcal{L}_{\text{HART}}(\mathcal{D}, \theta) = \lambda_{\text{t}}\mathcal{L}_{\text{t}}(\mathcal{D}, \theta) + \lambda_{\text{s}}\mathcal{L}_{\text{s}}(\mathcal{D}, \theta) + \lambda_{\text{a}}\mathcal{L}_{\text{a}}(\mathcal{D}, \theta) + R(\boldsymbol{\lambda}) + \beta R(\mathcal{D}, \theta), \quad (3.9)$$

with dataset $\mathcal{D} = \{(\mathbf{x}_{1:T}, \mathbf{b}_{1:T})^i\}_{i=1}^M$, network parameters $\theta$, regularisation terms $R(\cdot)$, adaptive weights $\boldsymbol{\lambda} = \{\lambda_{\text{t}}, \lambda_{\text{s}}, \lambda_{\text{d}}\}$ and a regularisation weight $\beta$. We now present and justify components of our loss, where expectations $\mathbb{E}[\cdot]$ are evaluated as an empirical mean over a minibatch of samples $\{\mathbf{x}_{1:T}^i, \mathbf{b}_{1:T}^i\}_{i=1}^M$, where $M$ is the batch size.

**Tracking** To achieve the main tracking objective (localising the object in the current frame), we base the first loss term on Intersection-over-Union (IoU) of the predicted bounding box w. r. t. the ground truth, where the IoU of two bounding boxes is defined as $\text{IoU}(\boldsymbol{a}, \boldsymbol{b}) = \frac{\boldsymbol{a} \cap \boldsymbol{b}}{\boldsymbol{a} \cup \boldsymbol{b}} = \frac{\text{area of overlap}}{\text{area of union}}$. The IoU is invariant to object and image scale, making it a suitable proxy for measuring the quality of localisation. Even though it (or an exponential thereof) does not correspond to any probability distribution (as it cannot be normalised), it is often used for evaluation **VOT2016**. We follow the work by **yu2016unitbox** and express the loss term as the negative log of IoU:

$$\mathcal{L}_{\text{t}}(\mathcal{D}, \theta) = \mathbb{E}_{p\left(\widehat{\mathbf{b}}_{1:T} | \mathbf{x}_{1:T}, \mathbf{b}_1\right)}\Big[-\log \text{IoU}(\widehat{\mathbf{b}}_t, \mathbf{b}_t)\Big], \quad (3.10)$$

with IoU clipped for numerical stability.

**Spatial Attention** Spatial attention singles out the tracked object from the image. To estimate its parameters, the system has to predict the object's motion. In case of an error, especially when the attention glimpse does not contain the tracked object, it is difficult to recover. As the probability of such an event increases with decreasing size of the glimpse, we employ two loss terms. The first one constrains the predicted attention to cover the bounding box, while the second one prevents it from becoming too large, where the logarithmic arguments are appropriately clipped to avoid numerical instabilities:

$$\mathcal{L}_{\mathrm{s}}(\mathcal{D},\theta) = \mathbb{E}_{p(\mathbf{a}_{1:T}|\mathbf{x}_{1:T},\mathbf{b}_1)}\left[-\log\left(\frac{\mathbf{a}_t \cap \mathbf{b}_t}{\mathrm{area}(\mathbf{b}_t)}\right) - \log(1-\mathrm{IoU}(\mathbf{a}_t,\mathbf{x}_t))\right]. \quad (3.11)$$

**Appearance Attention** The purpose of appearance attention is to suppress distractors while keeping full view of the tracked object e. g., focus on a *particular* pedestrian moving within a group. To guide this behaviour, we put a loss on appearance attention that encourages picking out only the tracked object. Let $\tau(\mathbf{a}_t,\mathbf{b}_t) : \mathbb{R}^4 \times \mathbb{R}^4 \to \{0,1\}^{h_v \times w_v}$ be a target function. Given the bounding box $\mathbf{b}$ and attention $\mathbf{a}$, it outputs a binary mask of the same size as the output of V1. The mask corresponds to the the glimpse $\mathbf{g}$, with the value equal to one at every location where the bounding box overlaps with the glimpse and equal to zero otherwise. If we take $H(p,q) = -\sum_z p(z)\log q(z)$ to be the cross-entropy, the loss reads

$$\mathcal{L}_{\mathrm{a}}(\mathcal{D},\theta) = \mathbb{E}_{p(\mathbf{a}_{1:T},\mathbf{s}_{1:T}|\mathbf{x}_{1:T},\mathbf{b}_1)}[H(\tau(\mathbf{a}_t,\mathbf{b}_t),\mathbf{s}_t)]. \quad (3.12)$$

**Regularisation** We apply the L2 regularisation to the model parameters $\theta$ and to the expected value of dynamic parameters $\boldsymbol{\psi}_t(\boldsymbol{\alpha}_t)$ as $R(\mathcal{D},\theta) = \frac{1}{2}\|\theta\|_2^2 + \frac{1}{2}\left\|\mathbb{E}_{p(\boldsymbol{\alpha}_{1:T}|\mathbf{x}_{1:T},\mathbf{b}_1)}[\boldsymbol{\Psi}_t \mid \boldsymbol{\alpha}_t]\right\|_2^2$.

**Adaptive Loss Weights** To avoid hyper-parameter tuning, we follow the work by **Kendall2017adaptive** and learn the loss weighting $\boldsymbol{\lambda}$. After initialising the weights with a vector of ones, we add the following regularisation term to the loss function: $R(\boldsymbol{\lambda}) = -\sum_i \log(\boldsymbol{\lambda}_i^{-1})$.

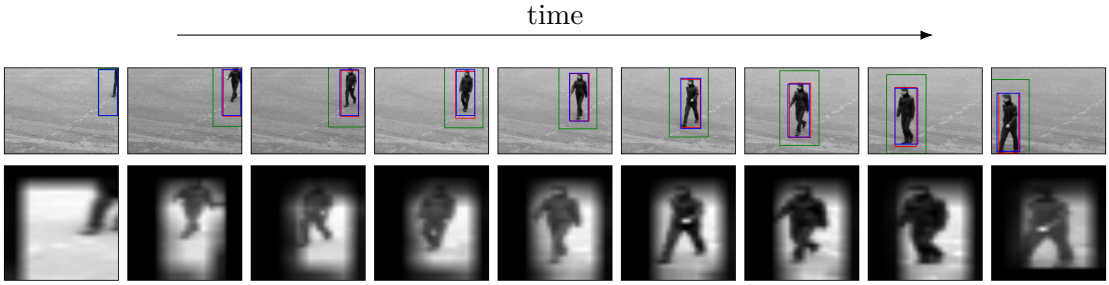



**Figure 3.4:** Tracking results on KTH dataset **KTH_activity_recognition**. Starting with the first initialisation frame where all three boxes overlap exactly, time flows from left to right showing every $16^{th}$ frame of the sequence captured at 25fps. The colour coding follows from Figure 3.1. The second row shows attention glimpses multiplied with appearance attention.

## 3.5 Experiments

### 3.5.1 KTH Pedestrian Tracking

**Kahou2015ratm** performed a pedestrian tracking experiment on the KTH activity recognition dataset **KTH_activity_recognition** as a real-world case-study. We replicate this experiment for comparison. We use code provided by the authors for data preparation and we also use their pre-trained feature extractor. Unlike them, we did not need to upscale ground-truth bounding boxes by a factor of 1.5 and then downscale them again for evaluation. We follow the authors and set the glimpse size $(h, w) = (28, 28)$. We replicate the training procedure exactly, with the exception of using the RMSProp optimiser **Hinton2015RMSProp** with learning rate of $3.33 \times 10^{-5}$ and momentum set to 0.9 instead of the stochastic gradient descent with momentum. The original work reported an IoU of 55.03% on average, on test data, while the presented work achieves an average IoU score of 77.11%, reducing the relative error by almost a factor of two. Figure 3.4 presents qualitative results.

### 3.5.2 Scaling to Real-World Data: KITTI

Since we demonstrated that pedestrian tracking is feasible using the proposed architecture, we proceed to evaluate our model in a more challenging multi-class scenario on the KITTI dataset **Geiger2013**. It consists of 21 high resolution video sequences with multiple instances of the same class posing as potential distractors. We split all sequences into 80/20 sequences for train and test sets, respectively. As

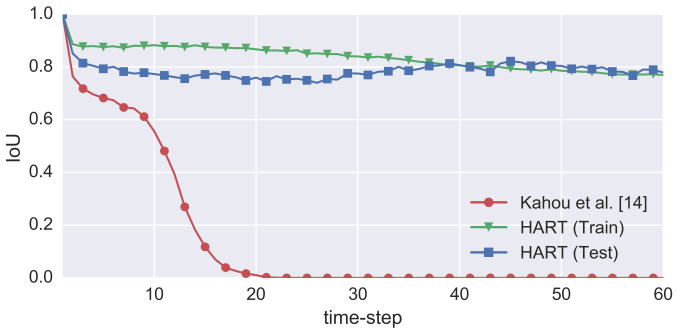



**Figure 3.5:** IoU curves on KITTI over 60 timesteps. HART (train) presents evaluation on the train set (we do not overfit).

| Method | Avg. IoU |
|---|---|
| **Kahou2015ratm** | 0.14 |
| Spatial Att | 0.60 |
| App Att | 0.78 |
| HART | **0.81** |

**Table 3.1:** Average IoU on KITTI over 60 time-steps.

images in this dataset are much more varied, we implement V1 as the first three convolutional layers of a modified AlexNet **Krizhevsky2012**. The original AlexNet takes inputs of size $227 \times 227$ and downsizes them to $14 \times 14$ after *conv3* layer. Since too low resolution would result in low tracking performance, and we did not want to upsample the extracted glimpse, we decided to replace the initial stride of four with one and to skip one of the max-pooling operations to conserve spatial dimensions. This way, our feature map has the size of $14 \times 14 \times 384$ with the input glimpse of size $(h, w) = (56, 56)$. We apply dropout with probability 0.25 at the end of V1. The ventral stream is comprised of a single convolutional layer with a $1 \times 1$ kernel and five output feature maps. The dorsal stream has two dynamic filter layers with kernels of size $1 \times 1$ and $3 \times 3$, respectively and five feature maps each. We used 100 hidden units in the RNN with orthogonal initialisation and Zoneout **Krueger2016** with probability set to 0.05. The system was trained via curriculum learning **Bengio2009**, by starting with sequences of length five and increasing sequence length every 13 epochs, with epoch length decreasing with increasing sequence length. We used the same optimisation settings, with the exception of the learning rate, which we set to $3.33 \times 10^{-6}$.

Table 3.1 and Figure 3.5 contain results of different variants of our model and of the RATM tracker by **Kahou2015ratm** related works. *Spatial Att* does not use appearance attention, nor loss on attention parameters. *App Att* does not apply any loss on appearance attention, while *HART* uses all described modules; it is also our biggest model with 1.8 million parameters. Qualitative results in the form of a video

**(a)** The model with appearance attention loss (top) learns to focus on the tracked object, which prevents an ID swap when a pedestrian is occluded by another one (bottom).

**(b)** Three examples of glimpses and locations maps for a model with and without appearance loss (left to right). Attention loss forces the appearance attention to pick out only the tracked object, thereby suppressing distractors.

**Figure 3.6:** Glimpses and corresponding location maps for models trained with and without appearance loss. The appearance loss encourages the model to learn foreground/background segmentation of the input glimpse.

with bounding boxes and attention are available online [4]. We implemented the RATM tracker of **Kahou2015ratm** and trained with the same hyperparameters as our framework, since both are closely related. It failed to learn even with the initial curriculum of five time-steps, as RATM cannot integrate the frame $\mathbf{x}_t$ into the estimate of $\mathbf{b}_t$ (it predicts location at the next time-step). Furthermore, it uses feature-space distance between ground-truth and predicted attention glimpses as the error measure, which is insufficient on a dataset with rich backgrounds. It did better when we initialised its feature extractor with weights of our trained model but, despite passing a few stags of the curriculum, it achieved very poor final performance.

## 3.6 Discussion

The experiments in the previous section show that it is possible to track real-world objects with a recurrent attentive tracker. While similar to the tracker by **Kahou2015ratm**, our approach uses additional building blocks, specifically: (i) bounding-box regression loss, (ii) loss on spatial attention, (iii) appearance attention with an additional loss term, and (iv) combines all of these in a unified approach. We now discuss properties of these modules.

[4] `https://youtu.be/VvkjmOFRGSs`

**Spatial Attention Loss prevents Vanishing Gradients** Our early experiments suggest that using only the tracking loss causes an instance of the vanishing gradient problem. Early in training, the system is not able to estimate object's motion correctly, leading to cases where the extracted glimpse does not contain the tracked object or contains only a part thereof. In such cases, the supervisory signal is only weakly correlated with the model's input, which prevents learning. Even when the object is contained within the glimpse, the gradient path from the loss function is rather long, since any teaching signal has to pass to the previous timestep through the feature extractor stage. Penalising attention parameters directly seems to solve this issue.

**Is Appearance Attention Loss Necessary?** Given enough data and sufficiently high model capacity, appearance attention should be able to filter out irrelevant input features before updating the working memory. In general, however, this behaviour can be achieved faster if the model is constrained to do so by using an appropriate loss. Figure 3.6 shows examples of glimpses and corresponding location maps for a model with and without loss on the appearance attention. In Fig. 3.6a the model with loss on appearance attention is able to track a pedestrian even after it was occluded by another human. Figure 3.6b shows that, when not penalised, location map might not be very object-specific and can miss the object entirely (right-most figure). By using the appearance attention loss, we not only improve results but also make the model more interpretable.

**Spatial Attention Bias is Always Positive** To condition the system on the object's appearance and make it independent of the starting location, we translate the initial bounding box to attention parameters, to which we add a learnable bias, and create the hidden state of LSTM from corresponding visual features. In our experiments, this bias always converged to positive values favouring attention glimpse slightly larger than the object bounding box. It suggests that, while discarding irrelevant features is desirable for object tracking, the system

as a whole learns to trade off attention responsibility between the spatial and appearance based attention modules.

## 3.7 Conclusion

Inspired by the cascaded attention mechanisms found in the human visual cortex, this work presented a neural attentive recurrent tracking architecture suited for the task of object tracking. Beyond the biological inspiration, the proposed approach has a desirable computational cost and increased interpretability due to location maps, which select features essential for tracking. Furthermore, by introducing a set of auxiliary losses we are able to scale to challenging real world data, outperforming predecessor attempts and approaching state-of-the-art performance. Future research will look into extending the proposed approach to multi-object tracking, as unlike many single object tracking, the recurrent nature of the proposed tracker offers the ability to attend each object in turn.

## Acknowledgements

We would like to thank Oiwi Parker Jones and Martin Engelcke for discussions and valuable insights and Neil Dhir for his help with editing the paper. Additionally, we would like to acknowledge the support of the UK's Engineering and Physical Sciences Research Council (EPSRC) through the Programme Grant EP/M019918/1 and the Doctoral Training Award (DTA). The donation from Nvidia of the Titan Xp GPU used in this work is also gratefully acknowledged.

# 4

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

# Abstract

We present Sequential Attend, Infer, Repeat (SQAIR), an interpretable deep generative model for videos of moving objects. It can reliably discover and track objects throughout the sequence of frames, and can also generate future frames conditioning on the current frame, thereby simulating expected motion of objects. This is achieved by explicitly encoding object presence, locations and appearances in the latent variables of the model. SQAIR retains all strengths of its predecessor, Attend, Infer, Repeat (AIR, **Eslami2016**), including learning in an unsupervised manner, and addresses its shortcomings. We use a moving multi-MNIST dataset to show limitations of AIR in detecting overlapping or partially occluded objects, and show how SQAIR overcomes them by leveraging temporal consistency of objects. Finally, we also apply SQAIR to real-world pedestrian CCTV data, where it learns to reliably detect, track and generate walking pedestrians with no supervision.

## 4.1 Introduction

The ability to identify objects in their environments and to understand relations between them is a cornerstone of human intelligence (**kemp2008discovery**). Arguably, in doing so we rely on a notion of spatial and temporal consistency which gives rise to an expectation that objects do not appear out of thin air, nor do they spontaneously vanish, and that they can be described by properties such as location, appearance and some dynamic behaviour that explains their evolution over time. We argue that this notion of consistency can be seen as an *inductive bias* that improves the efficiency of our learning. Equally, we posit that introducing such a bias towards spatio-temporal consistency into our models should greatly reduce the amount of supervision required for learning.

One way of achieving such inductive biases is through model structure. While recent successes in deep learning demonstrate that progress is possible without explicitly imbuing models with interpretable structure (**lecun2015deep**), recent works show that introducing such structure into deep models can indeed lead to favourable inductive biases improving performance e.g. in convolutional networks (**lecun1989backpropagation**) or in tasks requiring relational reasoning (**Santoro2017**). Structure can also make neural networks useful in new contexts by significantly improving generalization, data efficiency (**jacobsen2016struc**) or extending their capabilities to unstructured inputs (**Graves2016**).

AIR, introduced by **Eslami2016**, is a notable example of such a structured probabilistic model that relies on deep learning and admits efficient amortized inference. Trained without any supervision, AIR is able to decompose a visual scene into its constituent components and to generate a (learned) number of latent variables that explicitly encode the location and appearance of each object. While this approach is inspiring, its focus on modelling individual (and thereby inherently static) scenes leads to a number of limitations. For example, it often merges two objects that are close together into one since no temporal context is available to distinguish between them. Similarly, we demonstrate that AIR struggles to identify

partially occluded objects, e.g. when they extend beyond the boundaries of the scene frame (see Figure 4.7 in Section 4.4.1).

Our contribution is to mitigate the shortcomings of AIR by introducing a sequential version that models sequences of frames, enabling it to discover and track objects over time as well as to generate convincing extrapolations of frames into the future. We achieve this by leveraging temporal information to learn a richer, more capable generative model. Specifically, we extend AIR into a spatio-temporal state-space model and train it on unlabelled image sequences of dynamic objects. We show that the resulting model, which we name Sequential AIR (SQAIR), retains the strengths of the original AIR formulation while outperforming it on moving MNIST digits.

The rest of this work is organised as follows. In Section 4.2, we describe the generative model and inference of AIR. In Section 4.3, we discuss its limitations and how it can be improved, thereby introducing SQAIR, our extension of AIR to image sequences. In Section 4.4, we demonstrate the model on a dataset of multiple moving MNIST digits (Section 4.4.1) and compare it against AIR trained on each frame and VRNN of **Chung2015** with convolutional architectures, and show the superior performance of SQAIR in terms of log marginal likelihood and interpretability of latent variables. We also investigate the utility of inferred latent variables of SQAIR in downstream tasks. In Section 4.4.2 we apply SQAIR on real-world pedestrian CCTV data, where SQAIR learns to reliably detect, track and generate walking pedestrians without any supervision. Code for the implementation on the MNIST dataset[1] and the results video[2] are available online.

## 4.2 Attend, Infer, Repeat (AIR)

AIR, introduced by **Eslami2016**, is a structured variational auto-encoder (VAE) capable of decomposing a static scene $\mathbf{x}$ into its constituent objects, where each object is represented as a separate triplet of continuous latent variables $\mathbf{z} =$

[1] code: github.com/akosiorek/sqair
[2] video: youtu.be/-IUNQgSLE0c

$\{\mathbf{z}^{\text{what},i}, \mathbf{z}^{\text{where},i}, z^{\text{pres},i}\}_{i=1}^{n}$, $n \in \mathbb{N}$ being the (random) number of objects in the scene. Each triplet of latent variables explicitly encodes position, appearance and presence of the respective object, and the model is able to infer the number of objects present in the scene. Hence it is able to count, locate and describe objects in the scene, all learnt in an unsupervised manner, made possible by the inductive bias introduced by the model structure.

**Generative Model** The generative model of AIR is defined as follows

$$
\begin{aligned}
p_\theta(n) &= \text{Geom}(n \mid \theta), \qquad & p_\theta(\mathbf{z}^{\text{w}} \mid n) &= \prod_{i=1}^{n} p_\theta\Big(\mathbf{z}^{w,i}\Big) = \prod_{i=1}^{n} \mathcal{N}\Big(\mathbf{z}^{w,i} | \mathbf{0}, \mathbf{I}\Big), \\
p_\theta(\mathbf{x} \mid \mathbf{z}) &= \mathcal{N}\Big(\mathbf{x} \mid \mathbf{y}_t, \sigma_x^2 \boldsymbol{I}\Big), \qquad & \text{with} \quad \mathbf{y}_t &= \sum_{i=1}^{n} \text{h}_\theta^{\text{dec}}(\mathbf{z}^{\text{what},i}, \mathbf{z}^{\text{where},i}),
\end{aligned} \tag{4.1}
$$

where $\mathbf{z}^{\text{w},i} := (\mathbf{z}^{\text{what},i}, \mathbf{z}^{\text{where},i})$, $z^{\text{pres},i} = 1$ for $i = 1 \dots n$ and $h_\theta^{\text{dec}}$ is the object decoder with parameters $\theta$. It is composed of a *glimpse decoder* $f_\theta^{\text{dec}} : \mathbf{g}_t^i \mapsto \mathbf{y}_t^i$, which constructs an image patch and a spatial transformer (ST, **Jaderberg2015**), which scales and shifts it according to $\mathbf{z}^{\text{where}}$; see Figure 4.1 for details.

**Inference Eslami2016** use a sequential inference algorithm, where latent variables are inferred one at a time; see Figure 4.2. The number of inference steps $n$ is given by $z^{\text{pres},1:n+1}$, a random vector of $n$ ones followed by a zero. The $\mathbf{z}^i$ are sampled sequentially from

$$
q_\phi(\mathbf{z} \mid \mathbf{x}) = q_\phi\Big(z^{\text{pres},n+1} = 0 \mid \mathbf{z}^{\text{w},1:n}, \mathbf{x}\Big) \prod_{i=1}^{n} q_\phi\Big(\mathbf{z}^{\text{w},i}, z^{\text{pres},i} = 1 \mid \mathbf{z}^{1:i-1}, \mathbf{x}\Big), \tag{4.2}
$$

where $q_\phi$ is implemented as a neural network with parameters $\phi$. To implement explaining away, e.g. to avoid encoding the same object twice, it is vital to capture the dependency of $\mathbf{z}^{\text{w},i}$ and $z^{\text{pres},i}$ on $\mathbf{z}^{1:i-1}$ and $\mathbf{x}$. This is done using a RNN $R_\phi$ with hidden state $\boldsymbol{h}^i$, namely: $\boldsymbol{\omega}^i, \boldsymbol{h}^i = R_\phi(\mathbf{x}, \mathbf{z}^{i-1}, \boldsymbol{h}^{i-1})$. The outputs $\boldsymbol{\omega}^i$, which are computed iteratively and depend on the previous latent variables (*cf.* Algorithm 3), parametrise $q_\phi(\mathbf{z}^{\text{w},i}, \mathbf{z}^{\text{pres},i} \mid \mathbf{z}^{1:i-1}, \mathbf{x})$. For simplicity the latter is assumed to factorise such that $q_\phi(\mathbf{z}^{\text{w}}, \mathbf{z}^{\text{pres}} \mid \mathbf{z}^{1:i-1}, \mathbf{x}) = q_\phi(z^{\text{pres},n+1} = 0 \mid \boldsymbol{\omega}^{n+1}) \prod_{i=1}^{n} q_\phi(\mathbf{z}^{\text{w},i} \mid \boldsymbol{\omega}^i) q_\phi(z^{\text{pres},i} = 1 \mid \boldsymbol{\omega}^i)$.

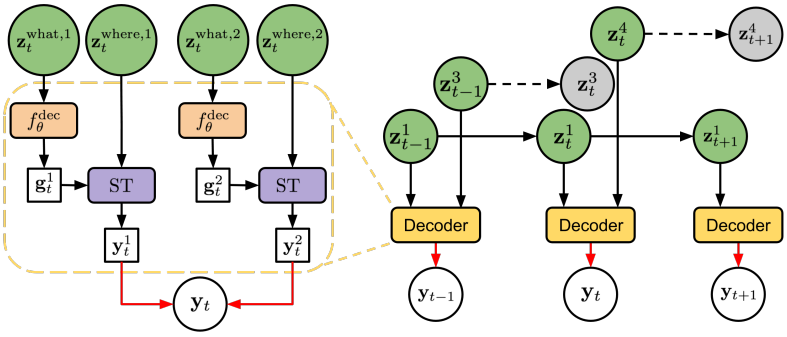



**Figure 4.1:** *Left*: Generation in AIR. The image mean $\mathbf{y}_t$ is generated by first using the *glimpse decoder* $f_\theta^{\text{dec}}$ to map the *what* variables into glimpses $\mathbf{g}_t$, transforming them with the *spatial transformer* ST according to the *where* variables and summing up the results. *Right*: Generation in SQAIR.

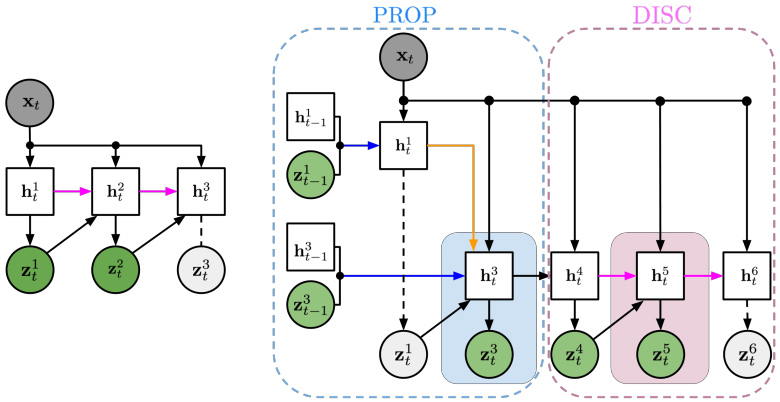



**Figure 4.2:** *Left*: Inference in AIR. The pink RNN attends to the image sequentially and produces one latent variable $\mathbf{z}_t^i$ at a time. Here, it decides that two latent variables are enough to explain the image and $\mathbf{z}_t^3$ is not generated. *Right*: Inference in SQAIR starts with the PROP phase. PROP iterates over latent variables from the previous time-step $t-1$ and updates them based on the new observation $\mathbf{x}_t$. The blue RNN runs forward in time to update the hidden state of each object, to model its change in appearance and location throughout time. The orange RNN runs across all current objects and models the relations between different objects. Here, when attending to $\mathbf{z}_{t-1}^1$, it decides that the corresponding object has disappeared from the frame and *forgets* it. Next, the DISC phase detects new objects as in AIR, but in SQAIR it is also conditioned on the results of PROP, to prevent rediscovering objects. See Figure 4.3 for details of the colored RNNs.

## 4.3 Sequential Attend-Infer-Repeat

While capable of decomposing a scene into objects, AIR only describes single images. Should we want a similar decomposition of an image sequence, it would be desirable to do so in a temporally consistent manner. For example, we might want to detect objects of the scene as well as infer dynamics and track identities of any persistent objects. Thus, we introduce Sequential Attend, Infer, Repeat (SQAIR), whereby AIR is augmented with a state-space model (SSM) to achieve temporal consistency in the generated images of the sequence. The resulting probabilistic model is composed of two parts: Discovery (DISC), which is responsible for detecting (or introducing, in the case of the generation) new objects at every time-step (essentially equivalent to AIR), and Propagation (PROP), responsible for updating (or forgetting) latent variables from the previous time-step given the new observation (image), effectively implementing the temporal SSM. We now formally introduce SQAIR by first describing its generative model and then the inference network.

**Generative Model** The model assumes that at every-time step, objects are first propagated from the previous time-step (PROP). Then, new objects are introduced (DISC). Let $t \in \mathbb{N}$ be the current time-step. Let $\mathcal{P}_t$ be the set of objects propagated from the previous time-step and let $\mathcal{D}_t$ be the set of objects discovered at the current time-step, and let $\mathcal{O}_t = \mathcal{P}_t \cup \mathcal{D}_t$ be the set of all objects present at time-step $t$. Consequently, at every time step, the model retains a set of latent variables $\mathbf{z}_t^{\mathcal{P}_t} = \{\mathbf{z}_t^i\}_{i \in \mathcal{P}_t}$, and generates a set of new latent variables $\mathbf{z}_t^{\mathcal{D}_t} = \{\mathbf{z}_t^i\}_{i \in \mathcal{D}_t}$. Together they form $\mathbf{z}_t := [\mathbf{z}_t^{\mathcal{P}_t}, \mathbf{z}_t^{\mathcal{D}_t}]$, where the representation of the $i^{\text{th}}$ object $\mathbf{z}_t^i := [\mathbf{z}_t^{\text{what},i}, \mathbf{z}_t^{\text{where},i}, z_t^{\text{pres},i}]$ is composed of three components (as in AIR): $\mathbf{z}_t^{\text{what},i}$ and $\mathbf{z}_t^{\text{where},i}$ are real vector-valued variables representing appearance and location of the object, respectively. $z_t^{\text{pres},i}$ is a binary variable representing whether the object is present at the given time-step or not.

At the first time-step ($t = 1$) there are no objects to propagate, so we sample $D_1$, the number of objects at $t = 1$, from the discovery prior $p^D(D_1)$. Then for each object $i \in \mathcal{D}_t$, we sample latent variables $\mathbf{z}_t^{\text{what},i}, \mathbf{z}_t^{\text{where},i}$ from $p^D(z_1^i \mid D_1)$. At

time $t = 2$, the *propagation* step models which objects from $t = 1$ are propagated to $t = 2$, and which objects disappear from the frame, using the binary random variable $(z_t^{\text{pres},i})_{i \in \mathcal{P}_t}$. The *discovery* step at $t = 2$ models new objects that enter the frame, with a similar procedure to $t = 1$: sample $D_2$ (which depends on $\mathbf{z}_2^{\mathcal{P}_2}$) then sample $(\mathbf{z}_2^{\text{what},i}, \mathbf{z}_2^{\text{where},i})_{i \in \mathcal{D}_2}$. This procedure of propagation and discovery recurs for $t = 2, \dots T$. Once the $\mathbf{z}_t$ have been formed, we may generate images $\mathbf{x}_t$ using the exact same generative distribution $p_\theta(\mathbf{x}_t \mid \mathbf{z}_t)$ as in AIR (*cf.* Equation (4.1), Fig. 4.1, and Algorithm 1). In full, the generative model is:

$$p(\mathbf{x}_{1:T}, \mathbf{z}_{1:T}, D_{1:T}) = p^D(D_1, \mathbf{z}_1^{\mathcal{D}_1}) \prod_{t=2}^{T} p^D(D_t, \mathbf{z}_t^{\mathcal{D}_t} | \mathbf{z}_t^{\mathcal{P}_t}) p^P(\mathbf{z}_t^{\mathcal{P}_t} | \mathbf{z}_{t-1}) p_\theta(\mathbf{x}_t | \mathbf{z}_t), \quad (4.3)$$

The *discovery prior* $p^D(D_t, \mathbf{z}_t^{\mathcal{D}_t} | \mathbf{z}_t^{\mathcal{P}_t})$ samples latent variables for new objects that enter the frame. The *propagation prior* $p^P(\mathbf{z}_t^{\mathcal{P}_t} | \mathbf{z}_{t-1})$ samples latent variables for objects that persist in the frame and removes latents of objects that disappear from the frame, thereby modelling dynamics and appearance changes. Both priors are learned during training. The exact forms of the priors are given in Section 4.B.

**Inference** Similarly to AIR, inference in SQAIR can capture the number of objects and the representation describing the location and appearance of each object that is necessary to explain every image in a sequence. As with generation, inference is divided into PROP and DISC. During PROP, the inference network achieves two tasks. Firstly, the latent variables from the previous time step are used to infer the current ones, modelling the change in location and appearance of the corresponding objects, thereby attaining temporal consistency. This is implemented by the *temporal* RNN $\mathrm{R}_\phi^T$, with hidden states $\boldsymbol{h}_t^T$ (recurs in $t$). Crucially, it does not access the current image directly, but uses the output of the *relation* RNN (*cf.* **Santoro2017**). The relation RNN takes relations between objects into account, thereby implementing the *explaining away* phenomenon; it is essential for capturing any interactions between objects as well as occlusion (or overlap, if one object is occluded by another). See Figure 4.7 for an example. These two RNNs together decide whether to retain or to forget objects that have been propagated from the previous time step. During DISC, the network infers further latent variables that

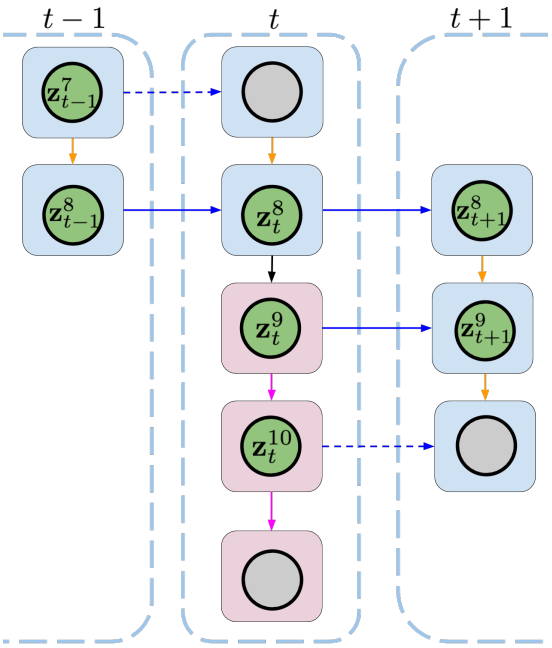
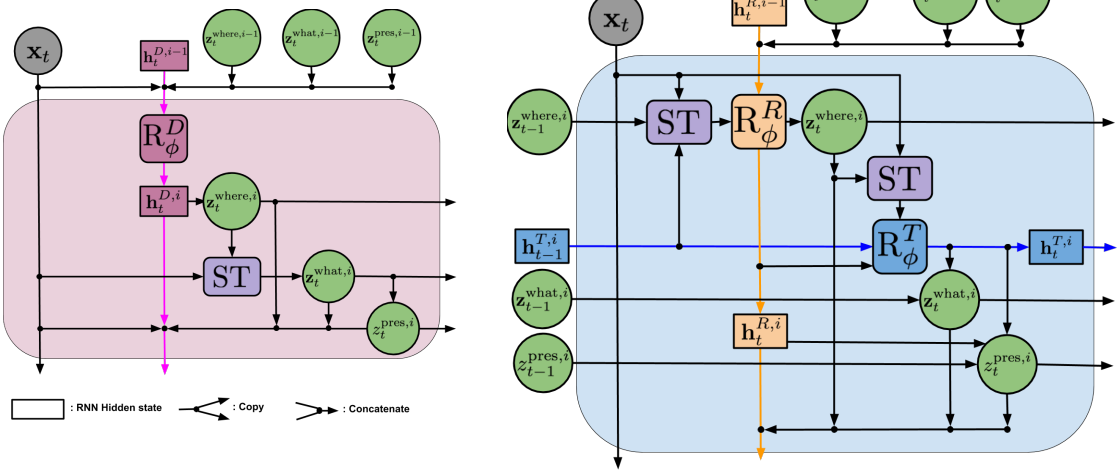



**Figure 4.3:** *Left*: Interaction between PROP and DISC in SQAIR. Firstly, objects are propagated to time $t$, and object $i = 7$ is dropped. Secondly, DISC tries to discover new objects. Here, it manages to find two objects: $i = 9$ and $i = 10$. The process recurs for all remaining time-steps. Blue arrows update the temporal hidden state, orange ones infer relations between objects, pink ones correspond to discovery. *Bottom*: Information flow in a single discovery block (*left*) and propagation block (*right*). In DISC we first predict *where* and extract a glimpse. We then predict *what* and *presence*. PROP starts with extracting a glimpse at a candidate location and updating *where*. Then it follows a procedure similar to DISC, but takes the respective latent variables from the previous time-step into account. It is approximately two times more computationally expensive than DISC. For details, see Algorithms 2 and 3 in Section 4.A.



are needed to describe any new objects that have entered the frame. All latent variables remaining after PROP and DISC are passed on to the next time step.

See Figures 4.2 and 4.3 for the inference network structure . The full variational posterior is defined as

$$q_\phi(D_{1:t}, \mathbf{z}_{1:T} \mid \mathbf{x}_{1:T}) = \prod_{t=1}^{T} q_\phi^D\left(D_t, \mathbf{z}_t^{\mathcal{D}_t} \mid \mathbf{x}_t, \mathbf{z}_t^{\mathcal{P}_t}\right) \prod_{i \in \mathcal{O}_{t-1}} q_\phi^P\left(\mathbf{z}_t^i \mid \mathbf{z}_{t-1}^i, \boldsymbol{h}_t^{T,i}, \boldsymbol{h}_t^{R,i}\right). \quad (4.4)$$

Discovery, described by $q_\phi^D$, is very similar to the full posterior of AIR, *cf.* Equation (4.2). The only difference is the conditioning on $\mathbf{z}_t^{\mathcal{P}_t}$, which allows for a different number of discovered objects at each time-step and also for objects explained by PROP not to be explained again. The second term, or $q_\phi^P$, describes propagation. The detailed structures of $q_\phi^D$ and $q_\phi^P$ are shown in Figure 4.3, while all the pertinent algorithms and equations can be found in Sections 4.A and 4.C, respectively.

**Learning** We train SQAIR as an importance-weighted auto-encoder (IWAE) of **Burda2016**. Specifically, we maximise the importance-weighted evidence lower-bound $\mathcal{L}_{\text{IWAE}}$, namely

$$\mathcal{L}_{\text{IWAE}} = \mathbb{E}_{\mathbf{x}_{1:T} \sim p_{\text{data}}(\mathbf{x}_{1:T})}\left[\mathbb{E}_q\left[\log \frac{1}{K} \sum_{k=1}^{K} \frac{p_\theta(\mathbf{x}_{1:T}, \mathbf{z}_{1:T})}{q_\phi(\mathbf{z}_{1:T} \mid \mathbf{x}_{1:T})}\right]\right]. \quad (4.5)$$

To optimise the above, we use RMSPROP, $K = 5$ and batch size of 32. We use the VIMCO gradient estimator of **Mnih2016** to backpropagate through the discrete latent variables $z^{\text{pres}}$, and use reparameterisation for the continuous ones (**kingma2013auto**). We also tried to use NVIL of **Mnih2014** as in the original work on AIR, but found it very sensitive to hyper-parameters, fragile and generally under-performing.

## 4.4 Experiments

We evaluate SQAIR on two datasets. Firstly, we perform an extensive evaluation on moving MNIST digits, where we show that it can learn to reliably detect, track and generate moving digits (Section 4.4.1). Moreover, we show that SQAIR can simulate moving objects into the future — an outcome it has not been trained for. We also study the utility of learned representations for a downstream task. Secondly, we

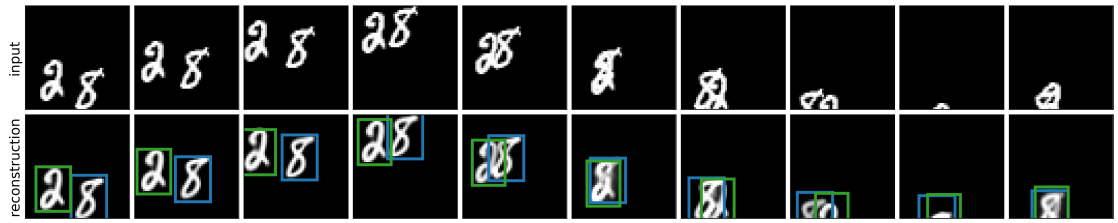



**Figure 4.4:** Input images (top) and SQAIR reconstructions with marked glimpse locations (bottom). For more examples, see Figure 4.H.1 in Section 4.H.

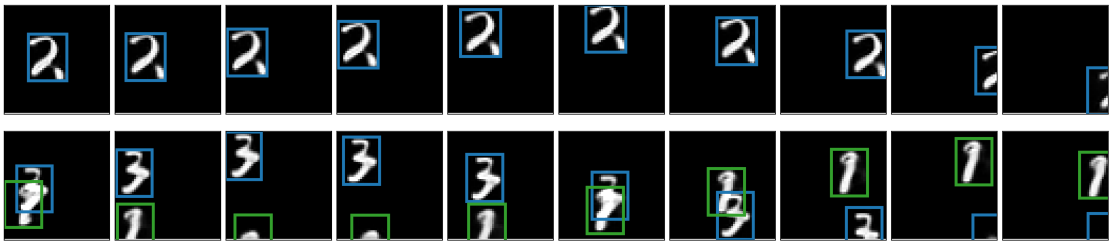


**Figure 4.5:** Samples from SQAIR. Both motion and appearance are consistent through time, thanks to the propagation part of the model. For more examples, see Figure 4.H.3 in Section 4.H.

apply SQAIR to real-world pedestrian CCTV data from static cameras (*DukeMTMC*, **ristani2016performance**), where we perform background subtraction as pre-processing. In this experiment, we show that SQAIR learns to detect, track, predict and generate walking pedestrians without human supervision.

### 4.4.1 Moving multi-mnist

The dataset consists of sequences of length 10 of multiple moving MNIST digits. All images are of size $50 \times 50$ and there are zero, one or two digits in every frame (with equal probability). Sequences are generated such that no objects overlap in the first frame, and all objects are present through the sequence; the digits can move out of the frame, but always come back. See Section 4.F for an experiment on a harder

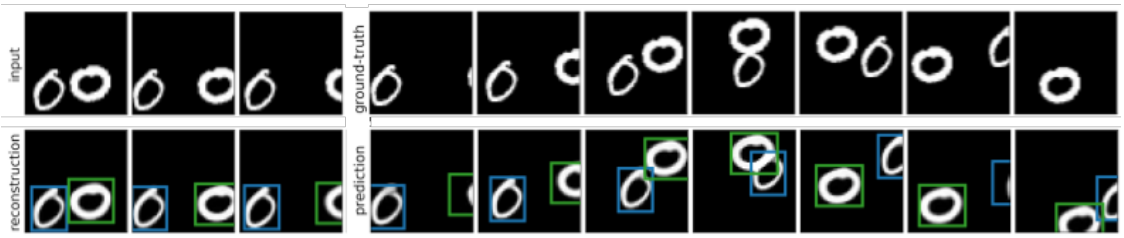



**Figure 4.6:** The first three frames are input to SQAIR, which generated the rest conditional on the first frames.

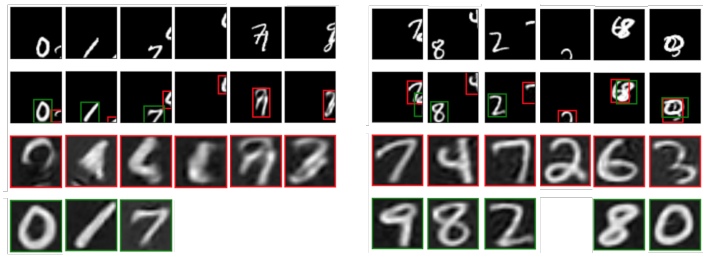


**Figure 4.7:** Inputs, reconstructions with marked glimpse locations and reconstructed glimpses for AIR (left) and SQAIR (right). SQAIR can model partially visible and heavily overlapping objects by aggregating temporal information.

| | $\log p_\theta(\mathbf{x}_{1:T})$ | $\log p_\theta(\mathbf{x}_{1:T} \mid \mathbf{z}_{1:T})$ | $\mathrm{KL}(q_\phi \mid\mid p_\theta)$ | Counting | Addition |
|---|---|---|---|---|---|
| CONV-SQAIR | **6784.8** | **6923.8** | **134.6** | 0.9974 | 0.9990 |
| MLP-SQAIR | 6617.6 | 6786.5 | 164.5 | **0.9986** | **0.9998** |
| MLP-AIR | 6443.6 | 6830.6 | 352.6 | 0.9058 | 0.8644 |
| CONV-VRNN | 6561.9 | 6737.8 | 270.2 | n/a | 0.8536 |
| MLP-VRNN | 5959.3 | 6108.7 | 218.3 | n/a | 0.8059 |

**Table 4.1:** SQAIR achieves higher performance than the baselines across a range of metrics. The third column refers to the Kullback-Leibler (KL) divergence between the approximate posterior and the prior. Counting refers to accuracy of the inferred number of objects present in the scene, while addition stands for the accuracy of a supervised digit addition experiment, where a classifier is trained on the learned latent representations of each frame.

version of this dataset. There are 60,000 training and 10,000 testing sequences created from the respective MNIST datasets. We train two variants of SQAIR: the MLP-SQAIR uses only fully-connected networks, while the CONV-SQAIR replaces the networks used to encode images and glimpses with convolutional ones; it also uses a subpixel-convolution network as the glimpse decoder (**shi2016subpixel**). See Section 4.D for details of the model architectures and the training procedure.

We use AIR and VRNN (**Chung2015**) as baselines for comparison. VRNN can be thought of as a sequential VAE with an RNN as its deterministic backbone. Being similar to a VAE, its latent variables are not structured, nor easily interpretable. For a fair comparison, we control the latent dimensionality of VRNN and the number of learnable parameters. We provide implementation details in Section 4.D.3.

The quantitative analysis consists of comparing all models in terms of the marginal log-likelihood $\log p_\theta(\mathbf{x}_{1:T})$ evaluated as the $\mathcal{L}_{\mathrm{IWAE}}$ bound with $K = 1000$ particles, reconstruction quality evaluated as a single-sample approximation of $\mathbb{E}_{q_\phi}[\log p_\theta(\mathbf{x}_{1:T} \mid \mathbf{z}_{1:T})]$ and the KL-divergence between the approximate posterior

and the prior (Table 4.1). Additionally, we measure the accuracy of the number of objects modelled by SQAIR and AIR. SQAIR achieves superior performance across a range of metrics — its convolutional variant outperforms both AIR and the corresponding VRNN in terms of model evidence and reconstruction performance. The KL divergence for SQAIR is almost twice as low as for VRNN and by a yet larger factor for AIR. We can interpret KL values as an indicator of the ability to compress, and we can treat SQAIR/AIR type of scheme as a version of run-length encoding. While VRNN has to use information to explicitly describe every part of the image, even if some parts are empty, SQAIR can explicitly allocate content information ($\mathbf{z}^{\text{what}}$) to specific parts of the image (indicated by $\mathbf{z}^{\text{where}}$). AIR exhibits the highest values of KL, but this is due to encoding every frame of the sequence independently — its prior cannot take *what* and *where* at the previous time-step into account, hence higher KL. The fifth column of Table 4.1 details the object counting accuracy, that is indicative of the quality of the approximate posterior. It is measured as the sum of $z_t^{\text{pres}}$ for a given frame against the true number of objects in that frame. As there is no $z^{\text{pres}}$ for VRNN no score is provided. Perhaps surprisingly, this metric is much higher for SQAIR than for AIR. This is because AIR mistakenly infers overlapping objects as a single object. Since SQAIR can incorporate temporal information, it does not exhibit this failure mode (*cf.* Figure 4.7). Next, we gauge the utility of the learnt representations by using them to determine the sum of the digits present in the image (Table 4.1, column six). To do so, we train a 19-way classifier (mapping from any combination of up to two digits in the range $[0, 9]$ to their sum) on the extracted representations and use the summed labels of digits present in the frame as the target. Section 4.D contains details of the experiment. SQAIR significantly outperforms AIR and both variants of VRNN on this tasks. VRNN under-performs due to the inability of disentangling overlapping objects, while both VRNN and AIR suffer from low temporal consistency of learned representations, see Section 4.H. Finally, we evaluate SQAIR qualitatively by analyzing reconstructions and samples produced by the model against reconstructions and samples from VRNN. We observe that samples and reconstructions from SQAIR are of better

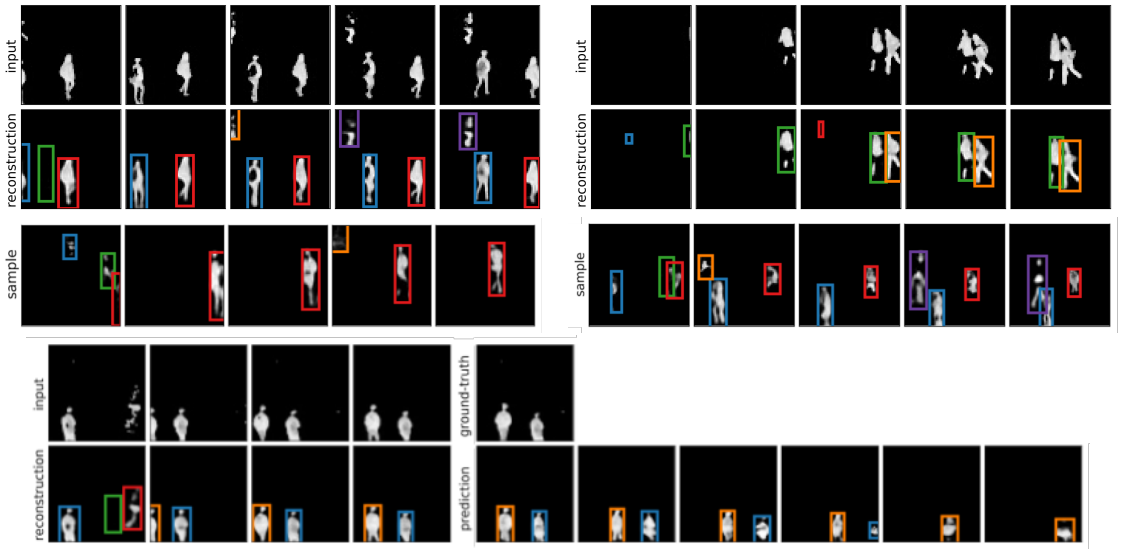



**Figure 4.8:** Inputs on the top, reconstructions in the second row, samples in the third row; rows four and five contain inputs and conditional generation: the first four frames in the last row are reconstructions, while the remaining ones are predicted by sampling from the prior. There is no ground-truth, since we used sequences of length five of training and validation.

quality and, unlike VRNN, preserve motion and appearance consistently through time. See Section 4.H for direct comparison and additional examples. Furthermore, we examine conditional generation, where we look at samples from the generative model of SQAIR conditioned on three images from a real sequence (see Figure 4.6). We see that the model can preserve appearance over time, and that the simulated objects follow similar trajectories, which hints at good learning of the motion model (see Section 4.H for more examples). Figure 4.7 shows reconstructions and corresponding glimpses of AIR and SQAIR. Unlike SQAIR, AIR is unable to recognize objects from partial observations, nor can it distinguish strongly overlapping objects (it treats them as a single object; columns five and six in the figure). We analyze failure cases of SQAIR in Section 4.G.

### 4.4.2 Generative Modelling of Walking Pedestrians

To evaluate the model in a more challenging, real-world setting, we turn to data from static CCTV cameras of the *DukeMTMC* dataset (**ristani2016performance**). As part of pre-precessing, we use standard background subtraction algorithms (**itseez2015opencv**). In this experiment, we use 3150 training and 350 validation

sequences of length 5. For details of model architectures, training and data pre-processing, see Section 4.E. We evaluate the model qualitatively by examining reconstructions, conditional samples (conditioned on the first four frames) and samples from the prior (Figure 4.8 and Section 4.I). We see that the model learns to reliably detect and track walking pedestrians, even when they are close to each other.

There are some spurious detections and re-detections of the same objects, which is mostly caused by imperfections of the background subtraction pipeline — backgrounds are often noisy and there are sudden appearance changes when a part of a person is treated as background in the pre-processing pipeline. The object counting accuracy in this experiment is 0.5712 on the validation dataset, and we noticed that it does increase with the size of the training set. We also had to use early stopping to prevent overfitting, and the model was trained for only 315k iterations ($>$ 1M for MNIST experiments). Hence, we conjecture that accuracy and marginal likelihood can be further improved by using a bigger dataset.

## 4.5 Related Work

**Object Tracking** There have been many approaches to modelling objects in images and videos. Object detection and tracking are typically learned in a supervised manner, where object bounding boxes and often additional labels are part of the training data. Single-object tracking commonly use Siamese networks, which can be seen as an RNN unrolled over two time-steps (**valmadre2017end**). Recently, **kosiorek2017hierch** used an RNN with an attention mechanism in the HART model to predict bounding boxes for single objects, while robustly modelling their motion and appearance. Multi-object tracking is typically attained by detecting objects and performing data association on bounding-boxes (**bewley2016sort**). **schulter2017deepnf** used an end-to-end supervised approach that detects objects and performs data association. In the unsupervised setting, where the training data consists of only images or videos, the dominant approach is to distill the inductive bias of spatial consistency into a discriminative model. **cho2015unsupervised** detect single objects and their parts in images, and

**kwak2015unsupervised**; **xiao2016track** incorporate temporal consistency to better track single objects. SQAIR is unsupervised and hence it does not rely on bounding boxes nor additional labels for training, while being able to learn arbitrary motion and appearance models similarly to HART (**kosiorek2017hierch**). At the same time, is inherently multi-object and performs data association implicitly (*cf.* Section 4.A). Unlike the other unsupervised approaches, temporal consistency is baked into the model structure of SQAIR and further enforced by lower KL divergence when an object is tracked.

**Video Prediction** Many works on video prediction learn a deterministic model conditioned on the current frame to predict the future ones (**ranzato2014video**; **srivastava2015unsupervised**). Since these models do not model uncertainty in the prediction, they can suffer from the multiple futures problem — since perfect prediction is impossible, the model produces blurry predictions which are a mean of possible outcomes. This is addressed in stochastic latent variable models trained using variational inference to generate multiple plausible videos given a sequence of images (**babaeizadeh2017stochastic**; **denton2018stochastic**). Unlike SQAIR, these approaches do not model objects or their positions explicitly, thus the representations they learn are of limited interpretability.

**Learning Decomposed Representations of Images and Videos** Learning decomposed representations of object appearance and position lies at the heart of our model. This problem can be also seen as perceptual grouping, which involves modelling pixels as spatial mixtures of entities. **Greff2016tagger** and **Greff2017neuralem** learn to decompose images into separate entities by iterative refinement of spatial clusters using either learned updates or the Expectation Maximization algorithm; **Ilin2017recurrentln** and **Steenkiste2018relationalnem** extend these approaches to videos, achieving very similar results to SQAIR. Perhaps the most similar work to ours is the concurrently developed model of **Hsieh2018ddpae**. The above approaches rely on iterative inference procedures, but do not exhibit the object-counting behaviour of SQAIR. For this reason, their

computational complexities are proportional to the predefined maximum number of objects, while SQAIR can be more computationally efficient by adapting to the number of objects currently present in an image.

Another interesting line of work is the GAN-based unsupervised video generation that decomposes motion and content (**tulyakov2017mocogan**; **denton2017unsupervised**). These methods learn interpretable features of content and motion, but deal only with single objects and do not explicitly model their locations. Nonetheless, adversarial approaches to learning structured probabilistic models of objects offer a plausible alternative direction of research.

**Bayesian Nonparametric Models** To the best of our knowledge, **neiswanger2012unsupervised** is the only known approach that models pixels belonging to a variable number of objects in a video together with their locations in the generative sense. This work uses a Bayesian nonparametric (BNP) model, which relies on mixtures of Dirichlet processes to cluster pixels belonging to an object. However, the choice of the model necessitates complex inference algorithms involving Gibbs sampling and Sequential Monte Carlo, to the extent that any sensible approximation of the marginal likelihood is infeasible. It also uses a fixed likelihood function, while ours is learnable.

The object appearance-persistence-disappearance model in SQAIR is reminiscent of the Markov Indian buffet process (MIBP) of **Gael2009**, another BNP model. MIBP was used as a model for blind source separation, where multiple sources contribute toward an audio signal, and can appear, persist, disappear and reappear independently. The prior in SQAIR is similar, but the crucial differences are that SQAIR combines the BNP prior with flexible neural network models for the dynamics and likelihood, as well as variational learning via amortized inference. The interface between deep learning and BNP, and graphical models in general, remains a fertile area of research.

## 4.6 Discussion

In this paper we proposed SQAIR, a probabilistic model that extends AIR to image sequences, and thereby achieves temporally consistent reconstructions and samples. In doing so, we enhanced AIR's capability of disentangling overlapping objects and identifying partially observed objects.

This work continues the thread of **Greff2017neuralem**, **Steenkiste2018relationalnem** and, together with **Hsieh2018ddpae**, presents unsupervised object detection & tracking with learnable likelihoods by the means of generative modelling of objects. In particular, our work is the first one to explicitly model object presence, appearance and location through time. Being a generative model, SQAIR can be used for conditional generation, where it can extrapolate sequences into the future. As such, it would be interesting to use it in a reinforcement learning setting in conjunction with Imagination-Augmented Agents (**weber2017imagination**) or more generally as a world model (**ha2018worldm**), especially for settings with simple backgrounds, e. g., games like Montezuma's Revenge or Pacman.

The framework offers various avenues of further research; SQAIR leads to interpretable representations, but the interpretability of *what* variables can be further enhanced by using alternative objectives that disentangle factors of variation in the objects (**kim2018disentangling**). Moreover, in its current state, SQAIR can work only with simple backgrounds and static cameras. In future work, we would like to address this shortcoming, as well as speed up the sequential inference process whose complexity is linear in the number of objects. The generative model, which currently assumes additive image composition, can be further improved by e. g., autoregressive modelling (**oord2016cond**). It can lead to higher fidelity of the model and improved handling of occluded objects. Finally, the SQAIR model is very complex, and it would be useful to perform a series of ablation studies to further investigate the roles of different components.

## Acknowledgements

We would like to thank Ali Eslami for his help in implementing AIR, Alex Bewley and Martin Engelcke for discussions and valuable insights and anonymous reviewers for their constructive feedback. Additionally, we acknowledge that HK and YWT's research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) ERC grant agreement no. 617071.

# Appendix

## 4.A Algorithms

Image generation, described by Algorithm 1, is exactly the same for SQAIR and AIR. Algorithms 2 and 3 describe inference in SQAIR. Note that DISC is equivalent to AIR if no latent variables are present in the inputs.

If a function has multiple inputs and if not stated otherwise, all the inputs are concatenated and linearly projected into some fixed-dimensional space, e. g., Lines 9 and 15 in Algorithm 2. Spatial Transformer (ST, e. g., Line 7 in Algorithm 2) has no learnable parameters: it samples a uniform grid of points from an image $\mathbf{x}$, where the grid is transformed according to parameters $\mathbf{z}^{\text{where}}$. $\text{f}_\phi^1$ is implemented as a perceptron with a single hidden layer. Statistics of $q^P$ and $q^D$ are a result of applying a two-layer multilayer perceptron (MLP) to their respective conditioning sets. Different distributions $q$ do not share parameters of their MLPs. The *glimpse encoder* $\text{h}_\phi^{\text{glimpse}}$ (Lines 8 and 12 in Algorithm 2 and Line 12 in Algorithm 3; they share parameters) and the *image encoder* $\text{h}_\phi^{\text{enc}}$ (Line 3 in Algorithm 3) are implemented as two-layer MLPs or convolutional neural networks (CNNs), depending on the experiment (see Sections 4.D and 4.E for details).

One of the important details of PROP is the proposal glimpse extracted in lines Lines 6 and 7 of Algorithm 2. It has a dual purpose. Firstly, it acts as an information bottleneck in PROP, limiting the flow of information from the current observation $\mathbf{x}_t$ to the updated latent variables $\mathbf{z}_t$. Secondly, even though the information is limited, it can still provide a high-resolution view of the object corresponding to the currently updated latent variable, *given* that the location of the proposal glimpse correctly predicts motion of this object. Initially, our implementation used encoding of the raw observation ($\text{h}_\phi^{\text{enc}}(\mathbf{x}_t)$, similarly to Line 3 in Algorithm 3) as

an input to the relation-RNN (Line 9 in Algorithm 2). We have also experimented with other bottlenecks: (1) low resolution image as an input to the image encoder and (2) a low-dimensional projection of the image encoding before the relation-RNN. Both approaches have led to *ID swaps*, where the order of explaining objects were sometimes swapped for different frames of the sequence (see Figure 4.G.1 in Section 4.G for an example). Using encoded proposal glimpse extracted from a predicted location has solved this issue.

To condition DISC on propagated latent variables (Line 4 in Algorithm 3), we encode the latter by using a two-layer MLP similarly to **zaheer2017deeps**,

$$\mathbf{l}_t = \sum_{i \in \mathcal{P}_t} \mathrm{MLP}\left(\mathbf{z}_t^{\mathrm{what},i}, \mathbf{z}_t^{\mathrm{where},i}\right). \tag{4.6}$$

Note that other encoding schemes are possible, though we have experimented only with this one.

**Algorithm 1:** Image Generation

**Input :** $\mathbf{z}_t^{\mathrm{what}}, \mathbf{z}_t^{\mathrm{where}}$ - latent variables from the current time-step.

$\mathcal{O}_t = \mathrm{indices}\left(\mathbf{z}_t^{\mathrm{what}}\right)$ `// Indices of all present latent variables.`

$\mathbf{y}_t^0 = \mathbf{0}$

**for** $i \in \mathcal{O}_t$ **do**

  $\mathbf{y}_t^{\mathrm{att},i} = \mathrm{f}_\theta^{\mathrm{dec}}\left(\mathbf{z}_t^{\mathrm{what},i}\right)$ `// Decode the glimpse.`

  $\mathbf{y}_t^i = \mathbf{y}_t^{i-1} + \mathrm{ST}^{-1}\left(\mathbf{y}_t^{\mathrm{att},i}, \mathbf{z}_t^{\mathrm{where},i}\right)$

$\hat{\mathbf{x}}_t \sim \mathcal{N}\left(\boldsymbol{x} \mid \boldsymbol{y}_n, \sigma_x^2 \boldsymbol{I}\right)$

**Output :** $\hat{\boldsymbol{x}}$

**Algorithm 2:** Inference for Propagation

**Input :** $\boldsymbol{x}_t$ - image at the current time-step,
$\mathbf{z}^{\text{what}}_{t-1}, \mathbf{z}^{\text{where}}_{t-1}, \mathbf{z}^{\text{pres}}_{t-1}$ - latent variables from the previous time-step
$\boldsymbol{h}^{T}_{t-1}$ - hidden states from the previous time-step.

$\boldsymbol{h}^{R,0}_t, \boldsymbol{z}^{\text{what},0}_t, \boldsymbol{z}^{\text{where},0}_t = \text{initialize}()$
$j = 0$ `// Index of the object processed in the last iteration.`
**for** $i \in \mathcal{O}_{t-1}$ **do**
  **if** $z^{\text{pres},i}_{t-1} == 0$ **then**
    **continue**
  $\hat{\mathbf{z}}^{\text{where},i}_t = \text{f}^1_\phi\left(\mathbf{z}^{\text{where},i}_{t-1}, \boldsymbol{h}^{T,i}_t\right)$ `// Proposal location.`
  $\hat{\mathbf{g}}^i_t = \text{ST}\left(\mathbf{x}_t, \hat{\mathbf{z}}^{\text{where},i}_t\right)$ `// Extract a glimpse from a proposal location.`
  $\hat{\mathbf{e}}^i_t = \text{h}^{\text{glimpse}}_\phi\left(\hat{\mathbf{g}}^i_t\right)$ `// Encode the proposal glimpse.`
  $\mathbf{w}^{R,i}_t, \boldsymbol{h}^{R,i}_t = \text{R}^R_\phi\left(\hat{\mathbf{e}}^i_t, \mathbf{z}^{\text{what},i}_{t-1}, \mathbf{z}^{\text{where},i}_{t-1}, \boldsymbol{h}^{T,i}_{t-1}, \boldsymbol{h}^{R,j}_t, \mathbf{z}^{\text{what},j}_t, \mathbf{z}^{\text{where},j}_t\right)$ `// Relational state, see Equation (4.14).`
  $\mathbf{z}^{\text{where},i}_t \sim q^P_\phi\left(\boldsymbol{z}^{\text{where}} \mid \boldsymbol{z}^{\text{where},k}_{t-1}, \mathbf{w}^{R,i}_t\right)$
  $\mathbf{g}^i_t = \text{ST}\left(\mathbf{x}_t, \mathbf{z}^{\text{where},i}_t\right)$ `// Extract the final glimpse.`
  $\mathbf{e}^i_t = \text{h}^{\text{glimpse}}_\phi\left(\mathbf{g}^i_t\right)$ `// Encode the final glimpse.`
  $\mathbf{w}^{T,i}_t, \boldsymbol{h}^{T,i}_t = \text{R}^T_\phi\left(\mathbf{e}^i_t, \mathbf{z}^{\text{where},i}_t, \boldsymbol{h}^{T,i}_{t-1}, \boldsymbol{h}^{R,i}_t\right)$ `// Temporal state, see Equation (4.15).`
  $\mathbf{z}^{\text{what},i}_t \sim q^P_\phi\left(\mathbf{z}^{\text{what}} \mid \mathbf{e}^i_t, \mathbf{z}^{\text{what},i}_{t-1}, \mathbf{w}^{R,i}_t, \mathbf{w}^{T,i}_t\right)$
  $z^{\text{pres},i}_t \sim q^P_\phi\left(z^{\text{pres}} \mid z^{\text{pres},i}_{t-1}, \mathbf{z}^{\text{what},i}_t, \mathbf{z}^{\text{where},i}_t, \mathbf{w}^{R,i}_t, \mathbf{w}^{T,i}_t\right)$ `// Equation (4.13).`
  $j = i$

**Output :** $\mathbf{z}^{\text{what},\mathcal{P}_t}_t, \mathbf{z}^{\text{where},\mathcal{P}_t}_t, \mathbf{z}^{\text{pres},\mathcal{P}_t}_t$

## 4.B Details for the Generative Model of SQAIR

In implementation, we upper bound the number of objects at any given time by $N$. In detail, the discovery prior is given by

$$p^D\left(D_t, \mathbf{z}^{\mathcal{D}_t}_t \mid \mathbf{z}^{\mathcal{P}_t}_t\right) = p^D(D_t \mid P_t) \prod_{i \in \mathcal{D}_t} p^D(\mathbf{z}^{\text{what},i}_t) p^D(\mathbf{z}^{\text{where},i}_t) \delta_1(z^{\text{pres},i}_t), \tag{4.7}$$

$$p^D(D_t \mid P_t) = \text{Categorical}\left(D_t; N - P_t, p_\theta(P_t)\right), \tag{4.8}$$

where $\delta_x(\cdot)$ is the delta function at $x$, $\text{Categorical}(k; K, p)$ implies $k \in \{0, 1, \ldots, K\}$ with probabilities $p_0, p_1, \ldots, p_K$ and $p^D(\mathbf{z}^{\text{what},i}_t), p^D(\mathbf{z}^{\text{where},i}_t)$ are fixed isotropic Gaus-

**Algorithm 3:** Inference for Discovery

**Input :** $\boldsymbol{x_t}$ - image at the current time-step,
$\mathbf{z}_t^{\mathcal{P}_t}$ - propagated latent variables for the current time-step,
$N$ - maximum number of inference steps for discovery.

$\boldsymbol{h}_t^{D,0}, \boldsymbol{z}_t^{\text{what},0}, \boldsymbol{z}_t^{\text{where},0} = \text{initialize}()$

$j = \text{max\_index}\left(\mathbf{z}_t^{\mathcal{P}_t}\right)$ `// Maximum index among the propagated latent variables.`

$\mathbf{e}_t = \text{h}_\phi^{\text{enc}}\left(\mathbf{x}_t\right)$ `// Encode the image.`

$\mathbf{l}_t = \text{h}_\phi^{\text{enc}}\left(\mathbf{z}_t^{\text{what}}, \mathbf{z}_t^{\text{where}}, \mathbf{z}_t^{\text{pres}}\right)$ `// Encode latent variables.`

**for** $i \in [j+1, \ldots, j+N]$ **do**

  $\mathbf{w}_t^{D,i}, \boldsymbol{h}_t^{D,i} = \text{R}_\phi^D\left(\mathbf{e}_t, \mathbf{l}_t, \mathbf{z}_t^{\text{what},i-1}, \mathbf{z}_t^{\text{where},i-1}, \boldsymbol{h}_t^{D,i-1}\right)$

  $z_t^{\text{pres},i} \sim q_\phi^D\left(z^{\text{pres}} \mid \mathbf{w}_t^{D,i}\right)$

  **if** $z^{\text{pres},i} = 0$ **then**

    break

  $\mathbf{z}_t^{\text{where},i} \sim q_\phi^D\left(\mathbf{z}^{\text{where}} \mid \mathbf{w}_t^{D,i}\right)$

  $\mathbf{g}_t^i = \text{ST}\left(\mathbf{x}_t, \mathbf{z}_t^{\text{where},i}\right)$

  $\mathbf{e}_t^i = \text{h}_\phi^{\text{glimpse}}\left(\mathbf{g}_t^i\right)$ `// Encode the glimpse.`

  $\mathbf{z}_t^{\text{what},i} \sim q_\phi^D\left(\mathbf{z}_t^{\text{what}} \mid \mathbf{e}_t^i\right)$

**Output :** $\mathbf{z}_t^{\text{what},\mathcal{D}_t}, \mathbf{z}_t^{\text{where},\mathcal{D}_t}, \mathbf{z}_t^{\text{pres},\mathcal{D}_t}$

sians. The propagation prior is given by

$$p^P\left(\mathbf{z}_t^{\mathcal{P}_t} \mid \mathbf{z}_{t-1}\right) = \prod_{i \in \mathcal{P}_t} p^P\left(\mathbf{z}_t^{\text{pres},i} \mid \mathbf{z}_{t-1}^{\text{pres},i}, \boldsymbol{h}_{t-1}\right) p^P\left(\mathbf{z}_t^{\text{what},i} \mid \boldsymbol{h}_{t-1}\right) p^P\left(\mathbf{z}_t^{\text{where},i} \mid \boldsymbol{h}_{t-1}\right), \tag{4.9}$$

$$p^P\left(\mathbf{z}_t^{\text{pres},i} \mid \mathbf{z}_{t-1}^{\text{pres},i}, \boldsymbol{h}_{t-1}\right) = \text{Bernoulli}(z_t^{\text{pres},i}; f_\theta(\boldsymbol{h}_{t-1}))\delta_1(z_{t-1}^{\text{pres},i}), \tag{4.10}$$

with $f_\theta$ a scalar-valued function with range $[0, 1]$ and $p^P(\mathbf{z}_t^{\text{what},i}|\boldsymbol{h}_{t-1})$, $p^P(\mathbf{z}_t^{\text{where},i}|\boldsymbol{h}_{t-1})$ both factorised Gaussians parameterised by some function of $\boldsymbol{h}_{t-1}$.

## 4.C Details for the Inference of SQAIR

The propagation inference network $q_\phi^P$ is given as below,

$$q_\phi^P\left(\mathbf{z}_t^{\mathcal{P}_t} \mid \mathbf{x}_t, \mathbf{z}_{t-1}, \boldsymbol{h}_t^{T,\mathcal{P}_t}\right) = \prod_{i \in \mathcal{O}_{t-1}} q_\phi^P\left(\mathbf{z}_t^i \mid \mathbf{x}_t, \mathbf{z}_{t-1}^i, \boldsymbol{h}_t^{T,i}, \boldsymbol{h}_t^{R,i}\right), \tag{4.11}$$

with $\boldsymbol{h}_t^{R,i}$ the hidden state of the relation RNN (see Equation (4.14)). Its role is to capture information from the observation $\mathbf{x}_t$ as well as to model dependencies

between different objects. The propagation posterior for a single object can be expanded as follows,

$$\begin{aligned} q_\phi^P\left(\mathbf{z}_t^i \mid \mathbf{x}_t, \mathbf{z}_{t-1}^i, \boldsymbol{h}_t^{T,i}, \boldsymbol{h}_t^{R,i}\right) = \\ q_\phi^P\left(\mathbf{z}_t^{\text{where},i} \mid \mathbf{z}_{t-1}^{\text{what},i}, \mathbf{z}_{t-1}^{\text{where},i}, \boldsymbol{h}_{t-1}^{T,i}, \boldsymbol{h}_t^{R,i}\right) \\ q_\phi^P\left(\mathbf{z}_t^{\text{what},i} \mid \mathbf{x}_t, \mathbf{z}_t^{\text{where},i}, \mathbf{z}_{t-1}^{\text{what},i}, \boldsymbol{h}_t^{T,i}, \boldsymbol{h}_t^{R,i}\right) \\ q_\phi^P\left(z_t^{\text{pres},i} \mid \mathbf{z}_t^{\text{what},i}, \mathbf{z}_t^{\text{where},i}, z_{t-1}^{\text{pres},i}, \boldsymbol{h}_t^{T,i}, \boldsymbol{h}_t^{R,i}\right). \end{aligned} \tag{4.12}$$

In the second line, we condition the object location $\mathbf{z}_t^{\text{where},i}$ on its previous appearance and location as well as its dynamics and relation with other objects. In the third line, current appearance $\mathbf{z}_t^{\text{what},i}$ is conditioned on the new location. Both $\mathbf{z}_t^{\text{where},i}$ and $\mathbf{z}_t^{\text{what},i}$ are modelled as factorised Gaussians. Finally, presence depends on the new appearance and location as well as the presence of the same object at the previous time-step. More specifically,

$$\begin{aligned} &q_\phi^P\left(z_t^{\text{pres},i} \mid \mathbf{z}_t^{\text{what},i}, \mathbf{z}_t^{\text{where},i}, z_{t-1}^{\text{pres},i}, \boldsymbol{h}_t^{T,i}, \boldsymbol{h}_t^{R,i}\right) \\ &\quad = \text{Bernoulli}\left(z_t^{\text{pres},i} \mid f_\phi\left(\mathbf{z}_t^{\text{what},i}, \mathbf{z}_t^{\text{where},i}, \boldsymbol{h}_t^{T,i}, \boldsymbol{h}_t^{R,i}\right)\right) \delta_1(z_{t-1}^{\text{pres},i}), \end{aligned} \tag{4.13}$$

where the second term is the delta distribution centered on the presence of this object at the previous time-step. If it was not there, it cannot be propagated. Let $j \in \{0, \ldots, i-1\}$ be the index of the most recent present object before object $i$. Hidden states are updated as follows,

$$\boldsymbol{h}_t^{R,i} = \text{R}_\phi^R\left(\mathbf{x}_t, \mathbf{z}_{t-1}^{\text{what},i}, \mathbf{z}_{t-1}^{\text{where},i}, \boldsymbol{h}_{t-1}^{T,i}, \boldsymbol{h}_t^{R,i-1}, \mathbf{z}_t^{\text{what},j}, \mathbf{z}_t^{\text{where},j}\right), \tag{4.14}$$

$$\boldsymbol{h}_t^{T,i} = \text{R}_\phi^T\left(\mathbf{x}_t, \mathbf{z}_t^{\text{where},i}, \boldsymbol{h}_{t-1}^{T,i}, \boldsymbol{h}_t^{R,i}\right), \tag{4.15}$$

where $\text{R}_\phi^T$ and $\text{R}_\phi^R$ are temporal and propagation RNNs, respectively. Note that in Eq. (4.14) the RNN does not have direct access to the image $\mathbf{x}_t$, but rather accesses it by extracting an attention glimpse at a proposal location, predicted from $\boldsymbol{h}_{t-1}^{T,i}$ and $\mathbf{z}_{t-1}^{\text{where},i}$. This might seem like a minor detail, but in practice structuring computation this way prevents ID swaps from occurring, *cf.* Section 4.G. For computational details, please see Algorithms 2 and 3 in Section 4.A.

## 4.D Details of the moving-mnist Experiments

### 4.D.1 Sqair and air Training Details

All models are trained by maximising the evidence lower bound (ELBO) $\mathcal{L}_{IWAE}$ (Equation (4.5)) with the RMSPROP optimizer (**tieleman2012rms**) with momentum equal to 0.9. We use the learning rate of $10^{-5}$ and decrease it to $\frac{1}{3} \cdot 10^{-5}$ after 400k and to $10^{-6}$ after 1000k training iterations. Models are trained for the maximum of $2 \cdot 10^6$ training iterations; we apply early stopping in case of overfitting. SQAIR models are trained with a curriculum of sequences of increasing length: we start with three time-steps, and increase by one time-step every $10^5$ training steps until reaching the maximum length of 10. When training AIR, we treated all time-steps of a sequence as independent, and we trained it on all data (sequences of length ten, split into ten independent sequences of length one).

### 4.D.2 Sqair and air Model Architectures

All models use glimpse size of 20×20 and exponential linear unit (ELU) (**Clevert2015elu**) non-linearities for all layers except RNNs and output layers. MLP-SQAIR uses fully-connected layers for all networks. In both variants of SQAIR, the $\mathrm{R}^D_\phi$ and $\mathrm{R}^R_\phi$ RNNs are the vanilla RNNs. The propagation prior RNN and the temporal RNN $\mathrm{R}^T_\phi$ use gated recurrent unit (GRU). AIR follows the same architecture as MLP-SQAIR. All fully-connected layers and RNNs in MLP-SQAIR and AIR have 256 units; they have 2.9M and 1.7M trainable parameters, respectively.

CONV-SQAIR differs from the MLP version in that it uses CNNs for the glimpse and image encoders and a subpixel-CNN (**shi2016subpixel**) for the glimpse decoder. All fully connected layers and RNNs have 128 units. The encoders share the CNN, which is followed by a single fully-connected layer (different for each encoder). The CNN has four convolutional layers with $[16, 32, 32, 64]$ features maps and strides of $[2, 2, 1, 1]$. The glimpse decoder is composed of two fully-connected layers with $[256, 800]$ hidden units, whose outputs are reshaped into 32 features maps of size 5×5, followed by a subpixel-CNN with three layers of $[32, 64, 64]$ feature maps and strides of $[1, 2, 2]$. All filters are of size $3 \times 3$. CONV-SQAIR has 2.6M trainable parameters.

We have experimented with different sizes of fully-connected layers and RNNs; we kept the size of all layers the same and altered it in increments of 32 units. Values greater than 256 for MLP-SQAIR and 128 for CONV-SQAIR resulted in overfitting. Models with as few as 32 units per layer ($< 0.9$M trainable parameters for MLP-SQAIR) displayed the same qualitative behaviour as reported models, but showed lower quantitative performance.

The output likelihood used in both SQAIR and AIR is Gaussian with a fixed standard deviation set to 0.3, as used by **Eslami2016**. We tried using a learnable scalar standard deviation, but decided not to report it due to unsable behaviour in the early stages of training. Typically, standard deviation would converge to a low value early in training, which leads to high penalties for reconstruction mistakes. In this regime, it is beneficial for the model to perform no inference steps ($z^{\text{pres}}$ is always equal to zero), and the model never learns. Fixing standard deviation for the first 10k iterations and then learning it solves this issue, but it introduces unnecessary complexity into the training procedure.

### 4.D.3 Vrnn Implementation and Training Details

Our VRNN implementation is based on the implementation[3] of Filtering Variational Objectives (FIVO) by **maddison2017filtering**. We use an LSTM with hidden size $J$ for the deterministic backbone of the VRNN. At time $t$, the LSTM receives $\psi^x(\mathbf{x}_{t-1})$ and $\psi^z(\mathbf{z}_{t-1})$ as input and outputs $o_t$, where $\psi^x$ is a data feature extractor and $\psi^z$ is a latent feature extractor. The output is mapped to the mean and standard deviation of the Gaussian prior $p_\theta(\mathbf{z}_t \mid \mathbf{x}_{t-1})$ by an MLP. The likelihood $p_\theta(\mathbf{x}_t \mid \mathbf{z}_t, \mathbf{x}_{t-1})$ is a Gaussian, with mean given by $\psi^{\text{dec}}(\psi^z(\mathbf{z}_t), o_t)$ and standard deviation fixed to be 0.3 as for SQAIR and AIR. The inference network $q_\phi(\mathbf{z}_t \mid \mathbf{z}_{t-1}, \mathbf{x}_t)$ is a Gaussian with mean and standard deviation given by the output of separate MLPs with inputs $[o_t, \psi^x(\mathbf{x}_t)]$.

All aforementioned MLPs use the same number of hidden units $H$ and the same number of hidden layers $L$. The CONV-VRNN uses a CNN for $\psi^x$ and a transposed CNN for $\psi^{\text{dec}}$. The MLP-VRNN uses an MLP with $H'$ hidden units and $L'$ hidden

[3] `https://github.com/tensorflow/models/tree/master/research/fivo`

**Table 4.D.1:** Number of trainable parameters for the reported models.

| | CONV-SQAIR | MLP-SQAIR | MLP-AIR | CONV-VRNN | MLP-VRNN |
|---|---|---|---|---|---|
| number of parameters | 2.6M | 2.9M | 1.7M | 2.6M | 2.1M |

layers for both. ELU were used throughout as activations. The latent dimensionality was fixed to 165, which is the upper bound of the number of latent dimensions that can be used per time-step in SQAIR or AIR. Training was done by optimising the FIVO bound, which is known to be tighter than the IWAE bound for sequential latent variable models (**maddison2017filtering**). We also verified that this was the case with our models on the moving-MNIST data. We train with the RMSPROP optimizer with a learning rate of $10^{-5}$, momentum equal to 0.9, and training until convergence of test FIVO bound.

For each of MLP-VRNN and CONV-VRNN, we experimented with three architectures: small/medium/large. We used $H{=}H'{=}J{=}128/256/512$ and $L{=}L'{=}2/3/4$ for MLP-VRNN, giving number of parameters of 1.2M/2.1M/9.8M. For CONV-VRNN, the number of features maps we used was $[32, 32, 64, 64]$, $[32, 32, 32, 64, 64, 64]$ and $[32, 32, 32, 64, 64, 64, 64, 64, 64]$, with strides of $[2, 2, 2, 2]$, $[1, 2, 1, 2, 1, 2]$ and $[1, 2, 1, 2, 1, 2, 1, 1, 1]$, all with $3 \times 3$ filters, $H{=}J{=}128/256/512$ and $L{=}1$, giving number of parameters of 0.8M/2.6M/6.1M. The largest convolutional encoder architecture is very similar to that in **gulrajani2016pixelvae** applied to MNIST.

We have chosen the medium-sized models for comparison with SQAIR due to overfitting encountered in larger models.

### 4.D.4 Addition Experiment

We perform the addition experiment by feeding latent representations extracted from the considered models into a 19-way classifier, as there are 19 possible outputs (addition of two digits between 0 and 9). The classifier is implemented as an MLP with two hidden layers with 256 ELU units each and a softmax output. For AIR and SQAIR, we use concatenated $\mathbf{z}^{\text{what}}$ variables multiplied by the corresponding $z^{\text{pres}}$ variables, while for VRNN we use the whole 165-dimensional latent vector. We train the

model over $10^7$ training iterations with the ADAM optimizer (**kingma2015adam**) with default parameters (in tensorflow).

## 4.E Details of the *DukeMTMC* Experiments

We take videos from cameras one, two, five, six and eight from the *DukeMTMC* dataset (**ristani2016performance**). As pre-processing, we invert colors and subtract backgrounds using standard OpenCV tools (**itseez2015opencv**), downsample to the resolution of $240 \times 175$, convert to gray-scale and randomly crop fragments of size $64 \times 64$. Finally, we generate 3500 sequences of length five such that the maximum number of objects present in any single frame is three and we split them into training and validation sets with the ratio of $9 : 1$.

We use the same training procedure as for the MNIST experiments. The only exception is the learning curriculum, which goes from three to five time-steps, since this is the maximum length of the sequences.

The reported model is similar to CONV-SQAIR. We set the glimpse size to $28 \times 12$ to account for the expected aspect ratio of pedestrians. Glimpse and image encoders share a CNN with $[16, 32, 64, 64]$ feature maps and strides of $[2, 2, 2, 1]$ followed by a fully-connected layer (different for each encoder). The glimpse decoder is implemented as a two-layer fully-connected network with 128 and 1344 units, whose outputs are reshaped into 64 feature maps of size $7 \times 3$, followed by a subpixel-CNN with two layers of $[64, 64]$ feature maps and strides of $[2, 2]$. All remaining fully-connected layers in the model have 128 units. The total number of trainable parameters is 3.5M.

## 4.F Harder multi-mnist Experiment

We created a version of the multi-MNIST dataset, where objects can appear or disappear at an arbitrary point in time. It differs from the dataset described in Section 4.4.1, where all digits are present throughout the sequence. All other dataset parameters are the same as in Section 4.4.1. Figure 4.F.1 shows an example sequence and MLP-SQAIR reconstructions with marked glimpse locations. The model

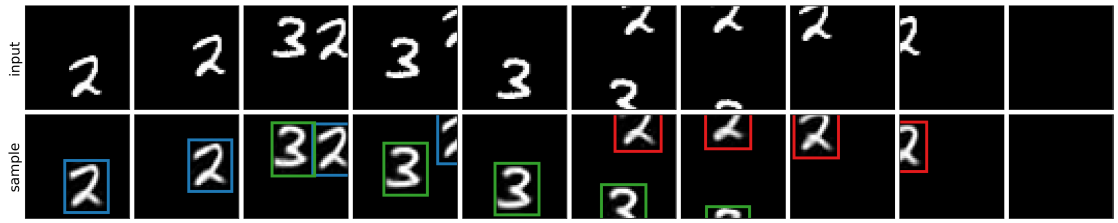



**Figure 4.F.1:** SQAIR trained on a harder version of moving-MNIST. Input images (top) and SQAIR reconstructions with marked glimpse locations (bottom)

has no trouble detecting new digits in the middle of the sequence and rediscovering a digit that was previously present.

## 4.G Failure cases of sqair

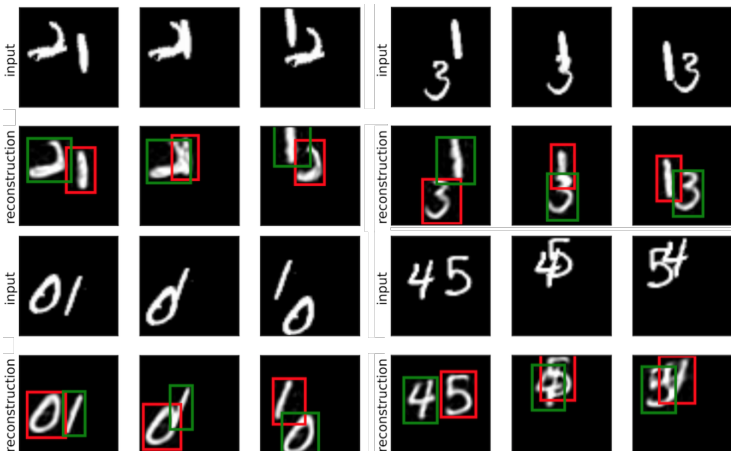



**Figure 4.G.1:** Examples of ID swaps in a version of SQAIR *without* proposal glimpse extraction in PROP (see Section 4.A for details). Bounding box colours correspond to object index (or its identity). When PROP is allowed the same access to the image as DISC, then it often prefers to ignore latent variables, which leads to swapped inference order.

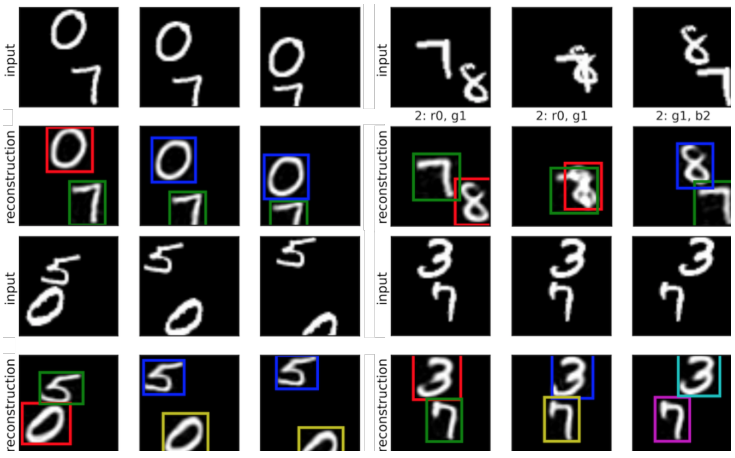



**Figure 4.G.2:** Examples of re-detections in MLP-SQAIR. Bounding box colours correspond to object identity, assigned to it upon discovery. In some training runs, SQAIR converges to a solution, where objects are re-detected in the second frame, and PROP starts tracking only in the third frame (left). Occasionally, an object can be re-detected after it has severely overlapped with another one (top right). Sometimes the model decides to use only DISC and repeatedly discovers all objects (bottom right). These failure mode seem to be mutually exclusive – they come from different training runs.

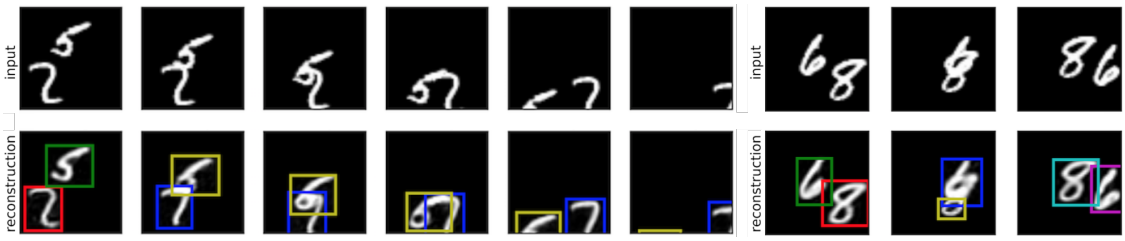
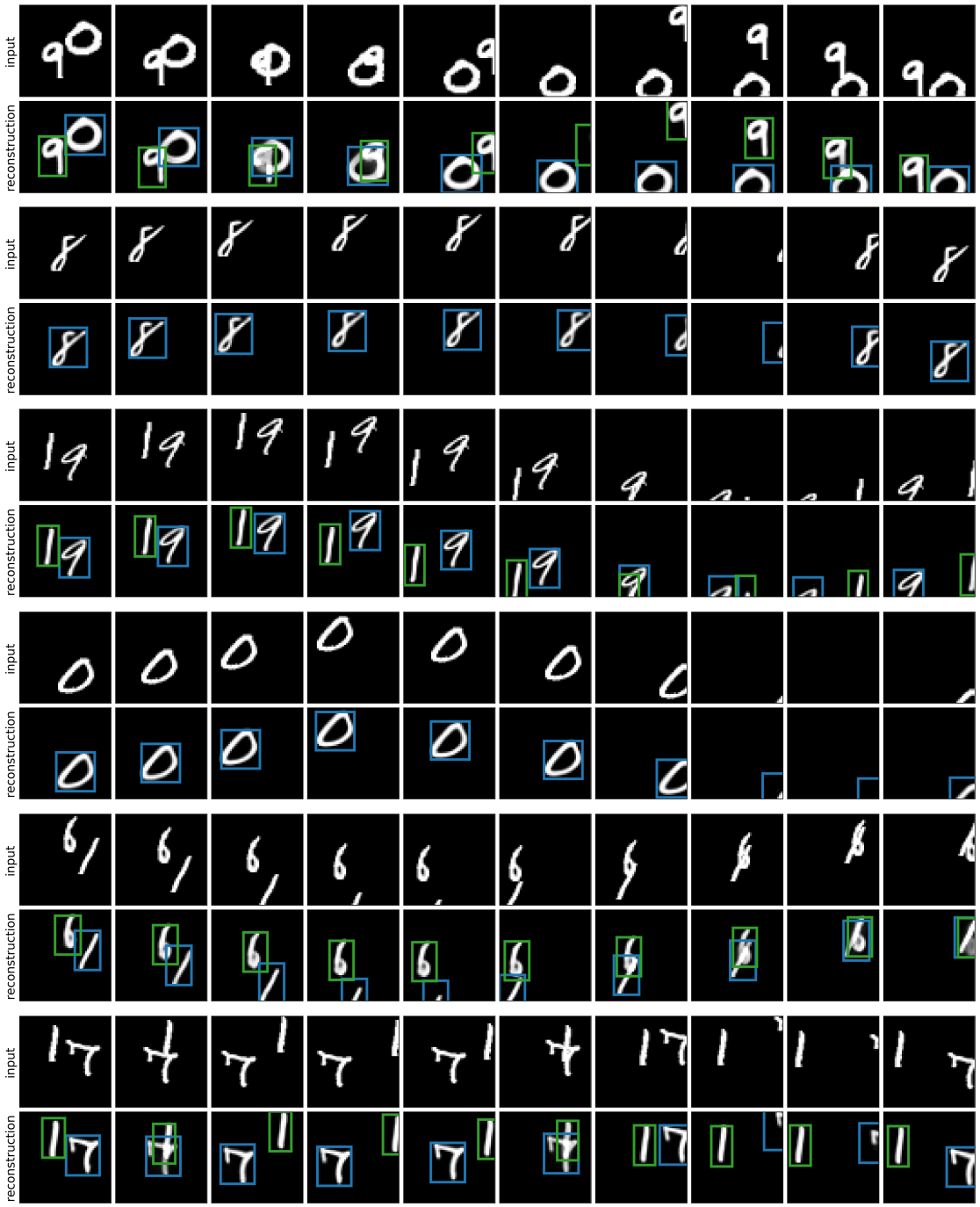


**Figure 4.G.3:** Two failed reconstructions of SQAIR. *Left*: SQAIR re-detects objects in the second time-step. Instead of 5 and 2, however, it reconstructs them as 6 and 7. Interestingly, reconstructions are consistent through the rest of the sequence. *Right:* At the second time-step, overlapping 6 and 8 are explained as 6 and a small 0. The model realizes its mistake in the third time-step, re-detects both digits and reconstructs them properly.

## 4.H Reconstruction and Samples from the Moving-MNIST Dataset

### 4.H.1 Reconstructions



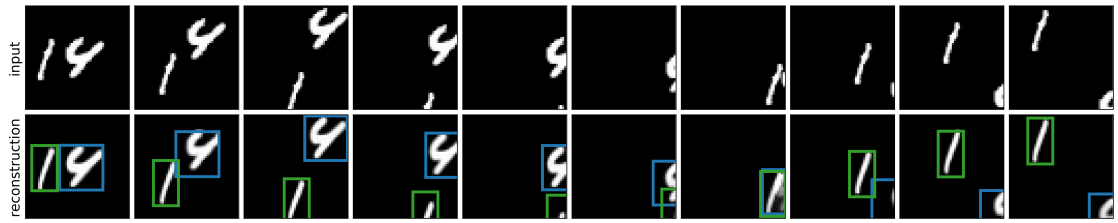



**Figure 4.H.1:** Sequences of input (first row) and SQAIR reconstructions with marked glimpse locations. Reconstructions are all temporally consistent.

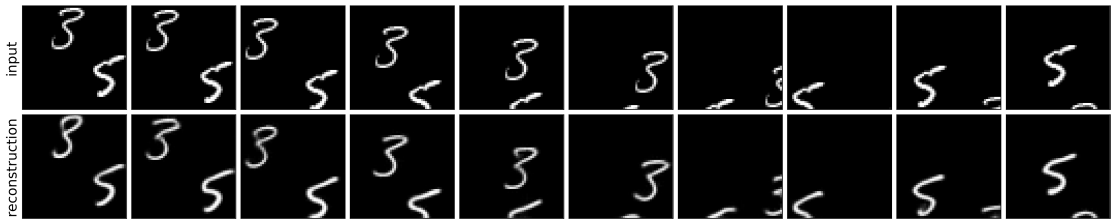



**Figure 4.H.2:** Sequences of input (first row) and CONV-VRNN reconstructions. They are not temporally consistent. The reconstruction at time $t = 1$ is typically of lower quality and often different than the rest of the sequence.

## 4.H.2 Samples

**Figure 4.H.3:** Samples from SQAIR. Both motion and appearance are temporally consistent. In the last sample, the model introduces the third object despite the fact that it has seen only up to two objects in training.

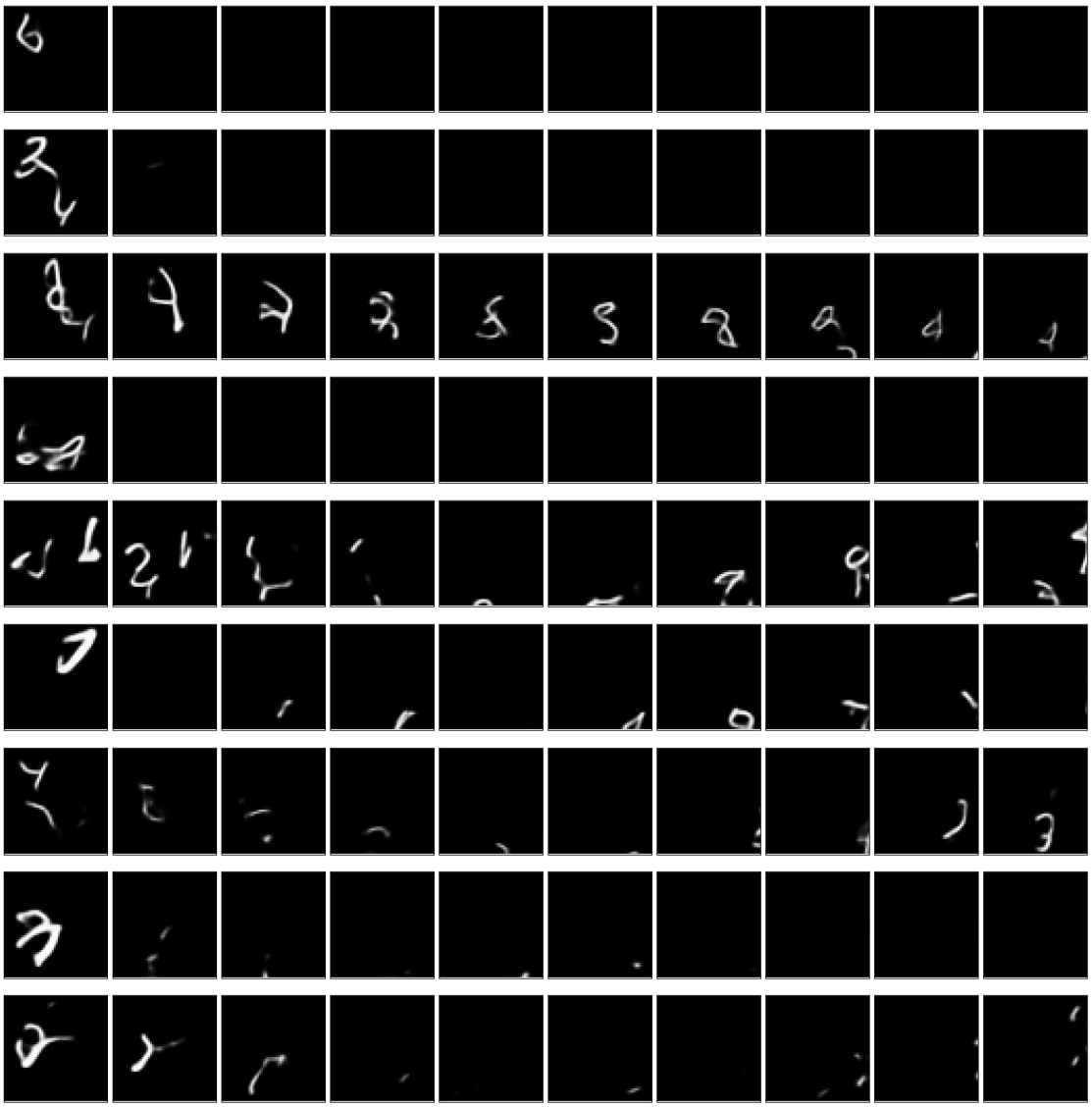


**Figure 4.H.4:** Samples from CONV-VRNN. They show lack of temporal consistency. Objects in the generated frames change between consecutive time-steps and they do not resamble digits from the training set.

### 4.H.3 Conditional Generation


**Figure 4.H.5:** Conditional generation from SQAIR, which sees only the first three frames in every case. Top is the input sequence (and the remaining ground-truth), while bottom is reconstruction (first three time-steps) and then generation.

# 4.I Reconstruction and Samples from the DukeMTMC Dataset

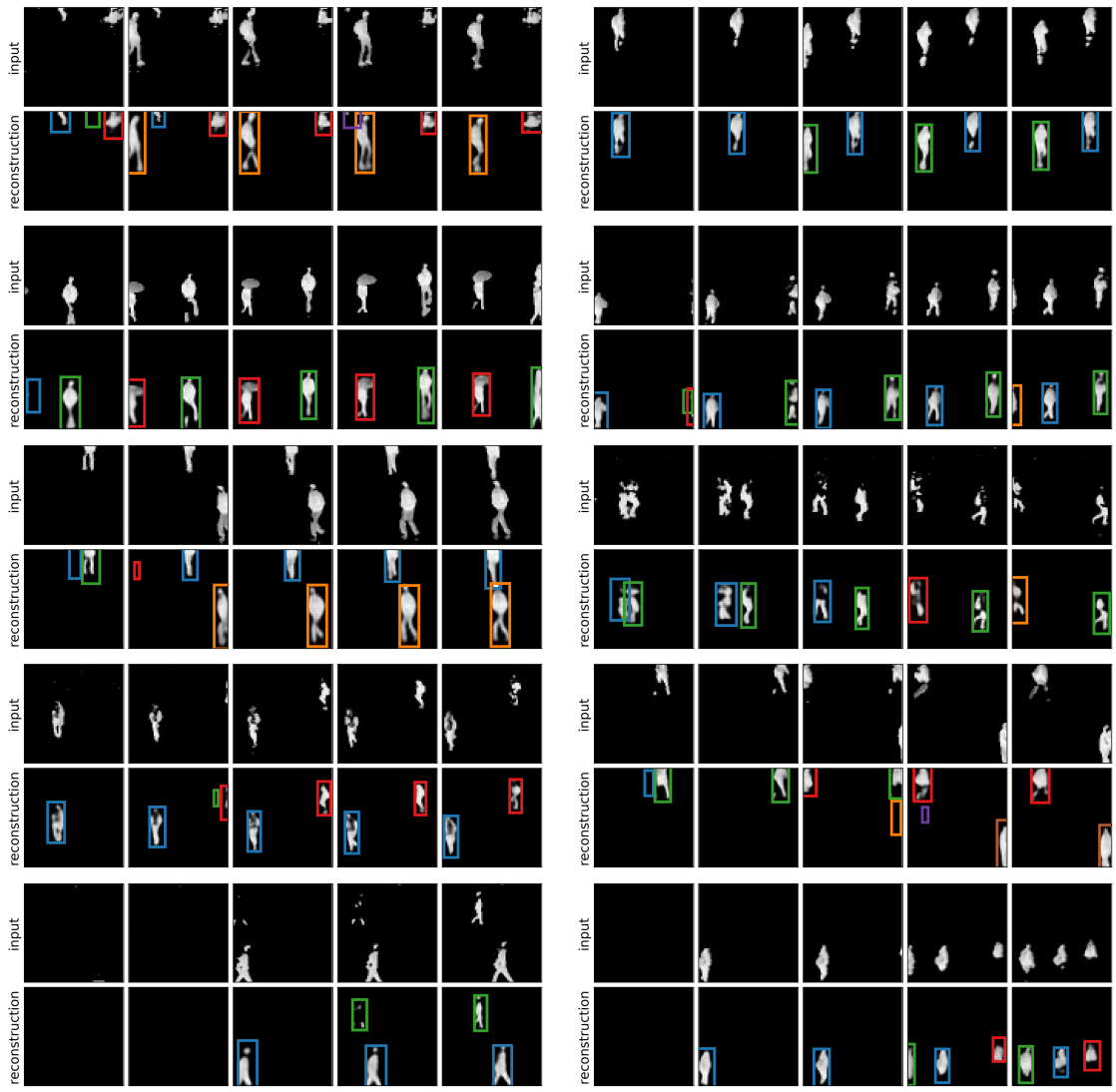



**Figure 4.I.1:** Sequences of input (first row) and SQAIR reconstructions with marked glimpse locations. While not perfect (spurious detections, missed objects), they are temporally consistent and similar in appearance to the inputs.

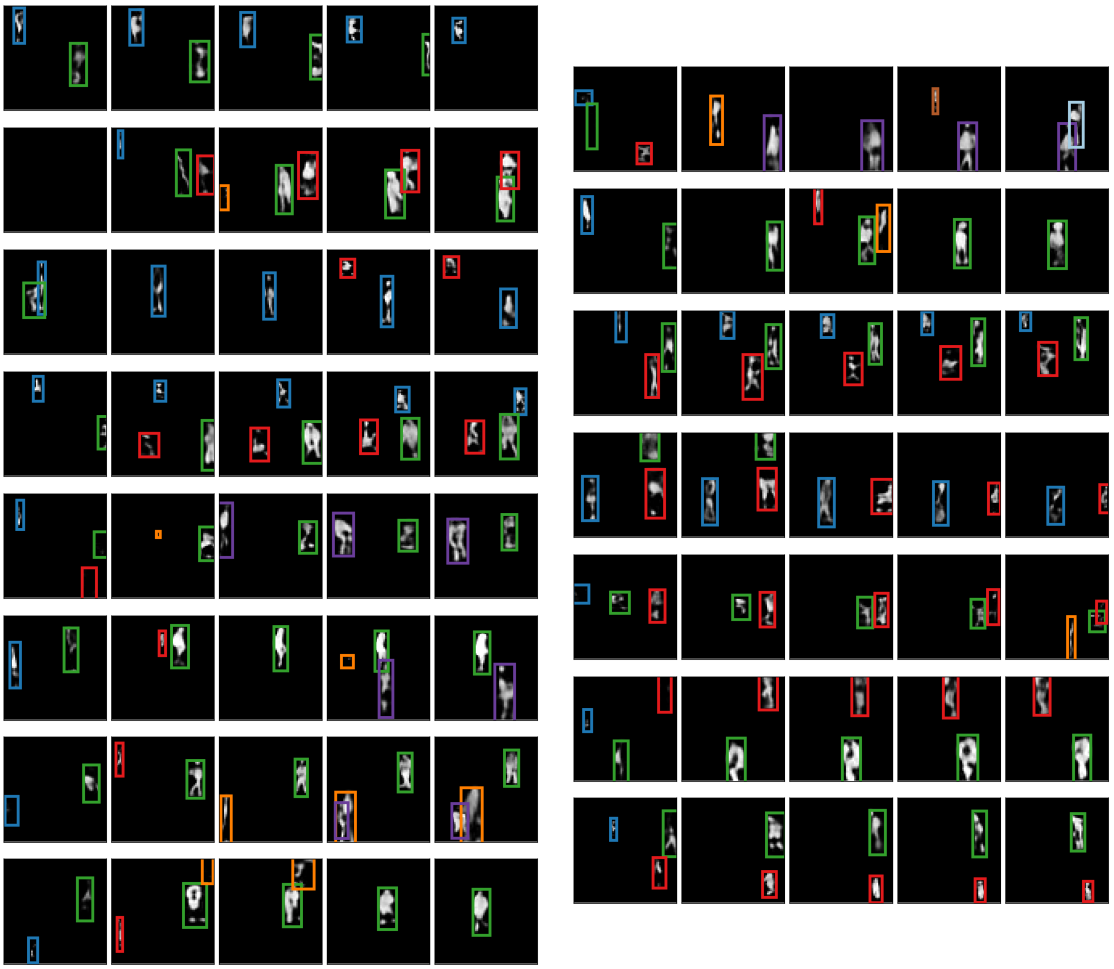


**Figure 4.I.2:** Samples with marked glimpse locations from SQAIR trained on the DukeMTMC dataset. Both appearance and motion is spatially consistent. Generated objects are similar in appearance to pedestrians in the training data. Samples are noisy, but so is the dataset.

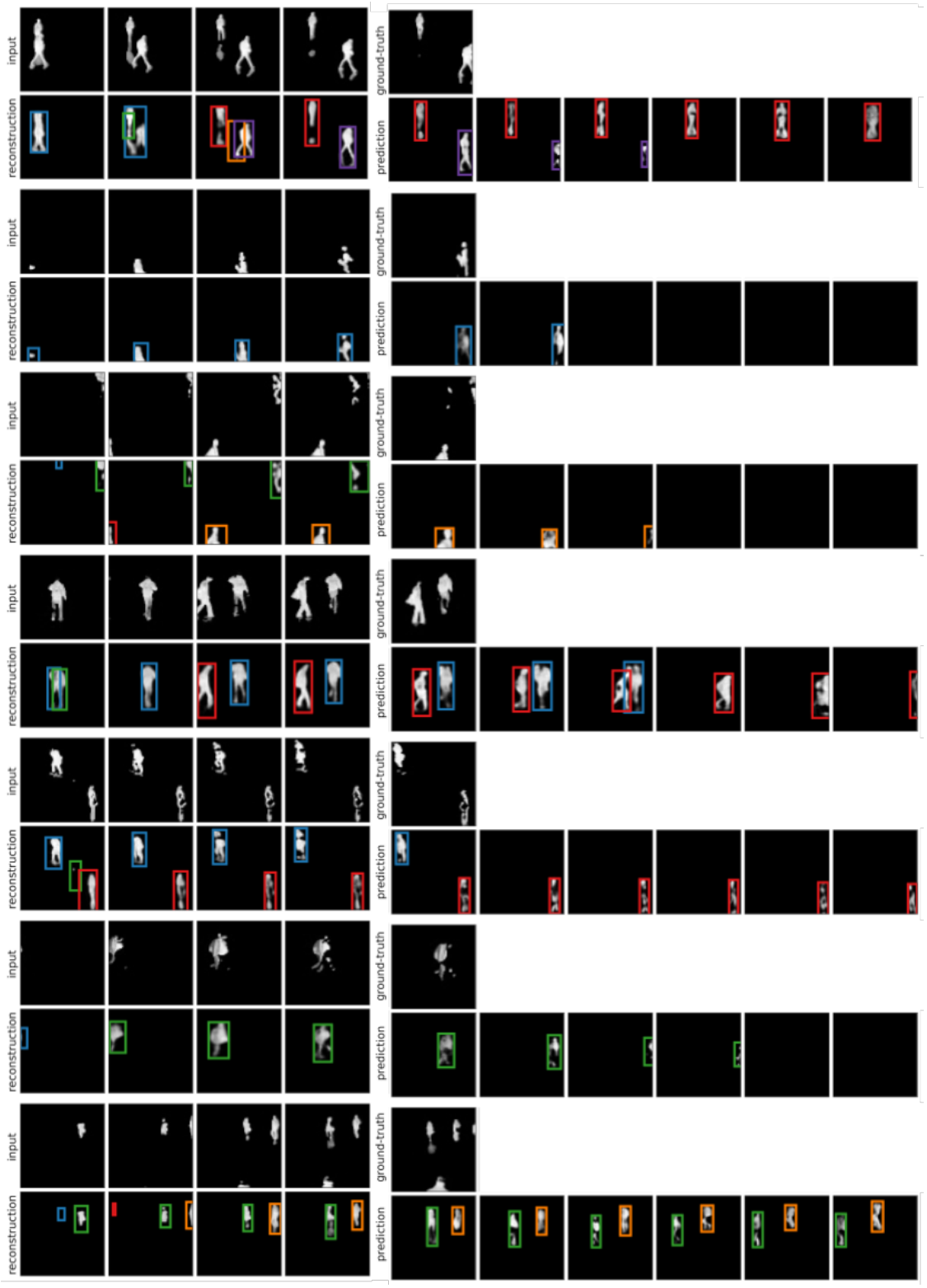



**Figure 4.I.3:** Conditional generation from SQAIR, which sees only the first four frames in every case. Top is the input sequence (and the remaining ground-truth), while bottom is reconstruction (first four time-steps) and then generation.

# Appendices

*Cor animalium, fundamentum eſt vitæ, princeps omnium, Microcoſmi Sol, a quo omnis vegetatio dependet, vigor omnis & robur emanat.*

*The heart of animals is the foundation of their life, the sovereign of everything within them, the sun of their microcosm, that upon which all growth depends, from which all power proceeds.*

— William Harvey **harvey_exercitatio_1628**

# A

# Review of Cardiac Physiology and Electrophysiology

Appendices are just like chapters. Their sections and subsections get numbered and included in the table of contents; figures and equations and tables added up, etc. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed et dui sem. Aliquam dictum et ante ut semper. Donec sollicitudin sed quam at aliquet. Sed maximus diam elementum justo auctor, eget volutpat elit eleifend. Curabitur hendrerit ligula in erat feugiat, at rutrum risus suscipit. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Integer risus nulla, facilisis eget lacinia a, pretium mattis metus. Vestibulum aliquam varius ligula nec consectetur. Maecenas ac ipsum odio. Cras ac elit consequat, eleifend ipsum sodales, euismod nunc. Nam vitae tempor enim, sit amet eleifend nisi. Etiam at erat vel neque consequat.

## A.1 Anatomy

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Donec accumsan cursus neque. Pellentesque eget tempor turpis, quis malesuada dui. Proin egestas, sapien sit amet feugiat vulputate, nunc nibh mollis nunc, nec auctor turpis purus sed metus. Aenean consequat leo congue volutpat euismod. Vestibulum et vulputate

nisl, at ultrices ligula. Cras pulvinar lacinia ipsum at bibendum. In ac augue ut ante mollis molestie in a arcu.

Etiam vitae quam sollicitudin, luctus tortor eu, efficitur nunc. Vestibulum maximus, ante quis consequat sagittis, augue velit luctus odio, in scelerisque arcu magna id diam. Proin et mauris congue magna auctor pretium id sit amet felis. Maecenas sit amet lorem ipsum. Proin a risus diam. Integer tempus eget est condimentum faucibus. Suspendisse sem metus, consequat vel ante eget, porttitor maximus dui. Nunc dapibus tincidunt enim, non aliquam diam vehicula sed. Proin vel felis ut quam porta tempor. Vestibulum elit mi, dictum eget augue non, volutpat imperdiet eros. Praesent ac egestas neque, et vehicula felis.

Pellentesque malesuada volutpat justo, id eleifend leo pharetra at. Pellentesque feugiat rutrum lobortis. Curabitur hendrerit erat porta massa tincidunt rutrum. Donec tincidunt facilisis luctus. Aliquam dapibus sodales consectetur. Suspendisse lacinia, ipsum sit amet elementum fermentum, nulla urna mattis erat, eu porta metus ipsum vel purus. Fusce eget sem nisl. Pellentesque dapibus, urna vitae tristique aliquam, purus leo gravida nunc, id faucibus ipsum magna aliquet ligula. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Proin sem lacus, rutrum eget efficitur sed, aliquam vel augue. Aliquam ut eros vitae sem cursus ultrices ut ornare urna. Nullam tempor porta enim, in pellentesque arcu commodo quis. Interdum et malesuada fames ac ante ipsum primis in faucibus. Curabitur maximus orci purus, ut molestie turpis pellentesque ut.

Donec lacinia tristique ultricies. Proin dignissim risus ut dolor pulvinar mollis. Proin ac turpis vitae nibh finibus ullamcorper viverra quis felis. Mauris pellentesque neque diam, id feugiat diam vestibulum vitae. In suscipit dui eu libero ultrices, et sagittis nunc blandit. Aliquam at aliquet ex. Nullam molestie pulvinar ex vitae interdum. Praesent purus nunc, gravida id est consectetur, convallis elementum nulla. Praesent ex dolor, maximus eu facilisis at, viverra eget nulla. Donec ullamcorper ante nisi. Sed volutpat diam eros. Nullam egestas neque non tortor aliquet, sed pretium velit tincidunt. Aenean condimentum, est ac vestibulum mattis, quam

augue congue augue, mattis ultrices nibh libero non ante. Lorem ipsum dolor sit amet, consectetur adipiscing elit.

Aenean volutpat eros tortor, non convallis sapien blandit et. Maecenas faucibus nulla a magna posuere commodo. Nullam laoreet ante a turpis laoreet malesuada. Phasellus in varius sem. Vestibulum sagittis nibh sed tincidunt blandit. Donec aliquam accumsan odio sit amet lacinia. Integer in tellus diam. Vivamus varius massa leo, vitae ullamcorper metus pulvinar sed. Maecenas nec lorem ornare, elementum est quis, gravida massa. Suspendisse volutpat odio ex, ac ultrices leo ultrices vel. Sed sed convallis ipsum. Pellentesque euismod a nulla sed rhoncus. Sed vehicula urna vitae mi aliquet, non sodales lacus ullamcorper. Duis mattis justo turpis, id tempus est tempus eu. Curabitur vitae hendrerit ligula.

Curabitur non pretium enim, in commodo ligula. Etiam commodo eget ligula a lacinia. Vestibulum laoreet ante tellus, vel congue sapien ornare in. Donec venenatis cursus velit vitae pulvinar. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Suspendisse in metus lectus. Pellentesque gravida dolor eget finibus imperdiet. Duis id molestie tortor. Mauris laoreet faucibus facilisis. Aliquam vitae dictum massa, sit amet dignissim lacus.

Fusce eleifend tellus id ex consequat maximus. Donec ultrices ex ut turpis ornare, non molestie mi placerat. Nulla sit amet auctor nunc, sit amet euismod elit. Phasellus risus tellus, condimentum a metus et, venenatis tristique urna. Cras mattis felis eget ipsum fermentum egestas. Ut augue odio, venenatis id convallis vel, congue quis augue. Maecenas sed maximus est, posuere aliquet tortor. Ut condimentum egestas nisi eu porttitor. Ut mi turpis, posuere id lorem vel, elementum tempor arcu.

Morbi nisl arcu, venenatis non metus ac, ullamcorper scelerisque justo. Nulla et accumsan lorem. Mauris aliquet dui sit amet libero aliquet, in ornare metus porttitor. Integer ultricies urna eu consequat ultrices. Maecenas a justo id purus ultricies posuere sed et quam. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Sed eleifend risus quis aliquet gravida. Nullam ac erat porta est bibendum dictum in a dolor. Nam eget turpis viverra, vulputate

lectus eget, mattis ligula. Nam at tellus eget dui lobortis sodales et ut augue. In vestibulum diam eget mi cursus, ut tincidunt nulla pellentesque.

Aliquam erat volutpat. Sed ultrices massa id ex mattis bibendum. Nunc augue magna, ornare at aliquet gravida, vehicula sed lorem. Quisque lobortis ipsum eu posuere eleifend. Duis bibendum cursus viverra. Nam venenatis elit leo, vitae feugiat quam aliquet sed. Cras velit est, tempus ac lorem sed, pharetra lobortis ipsum. Donec suscipit gravida interdum. Nunc non finibus est. Nullam turpis elit, tempus non ante.

## A.2 Mechanical Cycle

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aenean tellus est, suscipit sed facilisis quis, malesuada at ipsum. Nam tristique urna quis quam iaculis, et mattis orci pretium. Praesent euismod elit vel metus commodo ultrices. Vestibulum et tincidunt ex, in molestie ex. Donec ullamcorper sollicitudin accumsan. Etiam ac leo turpis. Duis a tortor felis. Nullam sollicitudin eu purus ac hendrerit. Nam hendrerit ligula libero, eget finibus orci bibendum a. Aenean ut ipsum magna.

Ut viverra, sapien sed accumsan blandit, nisi sem tempus tellus, at suscipit magna erat ornare nunc. Proin lacinia, nisi ut rutrum malesuada, nibh quam pellentesque nunc, sit amet consectetur purus felis ac tortor. Suspendisse lacinia ipsum eu sapien pellentesque mattis. Mauris ipsum nunc, placerat non diam vel, efficitur laoreet nunc. Sed lobortis, ipsum eget gravida facilisis, sem nulla viverra mi, in placerat eros sem viverra lacus. Aliquam porta aliquet diam vel commodo. Nulla facilisi. Duis erat libero, lobortis vel hendrerit vitae, sagittis id dui. Nulla pretium eros nec quam tincidunt, vel luctus mi aliquam. Integer imperdiet purus in est tristique venenatis. Ut pellentesque, nunc vitae iaculis ultricies, urna turpis dignissim risus, a laoreet felis magna nec erat.

Quisque sollicitudin faucibus ligula, et egestas nibh dictum sit amet. Proin eu mi a lectus congue pretium eu quis arcu. Suspendisse vehicula libero eu ipsum aliquam, vel elementum nibh mattis. Sed sed sapien vitae turpis tristique pulvinar a ut metus. Etiam semper gravida est, mollis gravida est porta ac. Proin eget tincidunt

erat. Maecenas ultrices erat eget purus ultricies, ut lacinia arcu dictum. Nam et nisi sit amet ex congue mattis vel eget lorem. Aliquam erat volutpat. Pellentesque porttitor nibh vitae elementum consectetur. Aenean et est lobortis, congue sapien non, ullamcorper sapien. Ut facilisis sem non dapibus vehicula.

Mauris euismod odio dolor, sit amet gravida mauris placerat et. Curabitur nec dolor non nibh molestie lobortis dignissim non ante. Nullam rutrum lobortis ultrices. Aenean ex erat, elementum sed maximus id, posuere id quam. Proin rutrum ex elit, pretium aliquam risus finibus at. Aenean egestas orci velit, sed aliquet sapien condimentum a. Duis consequat, arcu eu viverra venenatis, dolor lorem gravida lectus, non aliquet nisi sem at augue. Donec laoreet blandit luctus. Aenean vehicula nisl vel faucibus luctus. Sed ut semper velit, vitae laoreet magna. Sed at interdum magna.

Sed iaculis faucibus odio, eu aliquam purus efficitur vel. Cras at nulla ac enim congue varius ut et nulla. Integer blandit mattis augue.

## A.3 Electrical Cycle

Lorem ipsum dolor sit amet, consectetur adipiscing elit. In faucibus condimentum rhoncus. Ut dictum nisl id risus gravida lobortis. Sed vehicula mollis tellus ut varius. Fusce eget egestas dui, et commodo dui. Proin sollicitudin interdum tempus. Nullam in elit a enim fringilla bibendum. Vestibulum sodales pellentesque condimentum. Nulla facilisi. Nunc et dolor in nulla eleifend dictum at vel ligula. Aliquam ut velit non elit ullamcorper porta ac et ex. Fusce ornare magna non nunc vestibulum, eget molestie quam dictum. In interdum aliquam odio, in posuere tellus convallis quis. Curabitur non diam elit. Proin vulputate orci diam, a tincidunt ante luctus eu. Ut a viverra ligula. Curabitur pulvinar tempus tellus eget suscipit.

Aliquam posuere massa at ante dapibus congue. Curabitur ullamcorper tortor eget consectetur aliquet. Mauris tempor magna id mauris fringilla, a varius erat blandit. Nam eleifend ullamcorper placerat. Phasellus augue tortor, volutpat bibendum lorem nec, fringilla volutpat nisl. Mauris cursus urna metus, vel eleifend orci iaculis ut. Sed sit amet scelerisque massa, quis consequat dui. Donec

semper sem dui, ac placerat velit egestas vel. Nulla facilisi. Quisque tellus eros, sagittis malesuada augue ut, faucibus dictum nulla. Vestibulum non dapibus erat, ut consequat libero. Ut turpis mi, dapibus commodo libero lobortis, maximus vestibulum lectus. Vestibulum sit amet sapien dapibus, tincidunt leo in, suscipit arcu. Sed in erat bibendum, laoreet eros eu, pellentesque justo. Nulla sodales purus neque, eget maximus ipsum consequat at. Maecenas a nisl sagittis, tempus ipsum sed, dictum mauris.

Suspendisse posuere odio lacus, at auctor tortor vehicula sed. Phasellus suscipit ornare enim vitae placerat. Sed viverra purus vel sapien tempor, quis iaculis erat laoreet. Aenean vel nunc vestibulum, ornare nunc ac, mollis urna. Aenean ultrices felis ipsum, ac semper est ullamcorper in. Donec in justo varius, egestas tortor ut, venenatis augue. Duis mattis, ligula quis lacinia fringilla, tellus neque accumsan ipsum, vitae tempor metus elit vel nibh. Curabitur porttitor urna nec sapien tempor, et porttitor velit malesuada.

Suspendisse aliquam nisl quis placerat vulputate. Proin dapibus ipsum ac ante sagittis, volutpat auctor sem dapibus. Nam in facilisis odio. Integer ante mauris, eleifend et pulvinar in, venenatis quis ligula. Phasellus posuere sollicitudin tortor eget euismod. Maecenas mollis tortor eget justo vulputate sagittis. Etiam hendrerit massa quis ex molestie sodales. Quisque facilisis erat lacus, id convallis sem suscipit bibendum. Integer dui urna, pharetra sed porta sed, bibendum ut odio. Donec placerat at lectus egestas consequat. Sed id rhoncus est, vitae vulputate sapien. Fusce tempus quam lorem, id ornare turpis sodales sed. Integer aliquet urna eget condimentum consequat. Vestibulum quis dui vel ligula posuere luctus id nec turpis.

Nam vitae placerat lacus. Mauris scelerisque interdum volutpat. Nunc aliquet tristique enim, sit amet molestie felis ullamcorper vitae. Nullam sollicitudin orci orci, in condimentum tellus consectetur in. Nam id justo justo. Fusce eget finibus est. Proin id tortor nec quam cursus vehicula. Aliquam ultrices eros eros, a tincidunt elit eleifend auctor.

Nullam consectetur dapibus ligula sit amet efficitur. Nunc non posuere sapien. Vivamus dui nisl, aliquam id ipsum non, pulvinar ornare neque. Nunc rhoncus

pretium congue. Fusce id laoreet enim. Cras sed massa in eros bibendum auctor in nec sem. Nam commodo, velit id porta consequat, mi arcu gravida lorem, ut aliquam elit ante quis dui. Quisque in massa sed nibh blandit dictum.

Vestibulum molestie consectetur porttitor. Donec tincidunt vel orci at pharetra. Nullam id felis sit amet nulla tempus lacinia. Integer egestas ullamcorper massa, ut ultricies diam congue sit amet. Cras sit amet velit at nibh vehicula finibus a et lorem. Cras odio metus, venenatis ut ultrices non, ornare ac orci. Morbi et nulla dui. Mauris dictum molestie nibh, eu efficitur lorem accumsan quis.

## A.4 Cellular Electromechanical Coupling

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nullam vitae consectetur metus, ac maximus ex. Quisque vitae ex eu lectus ultricies consequat vel non lorem. Etiam odio ipsum, tempus ut lobortis in, molestie ac leo. Vivamus mollis feugiat bibendum. Vestibulum eget venenatis quam. Aenean faucibus, massa sed ullamcorper porta, arcu nunc iaculis velit, quis consectetur purus neque placerat nibh. Vestibulum elit nunc, dignissim vulputate venenatis et, sodales non massa. Proin leo ligula, vehicula eu aliquam varius, posuere a dolor. Donec iaculis auctor neque, sit amet gravida libero porta vel. Vivamus consequat elementum lacus, at bibendum mauris egestas nec. Fusce fermentum diam eu dolor ornare, vitae vestibulum leo interdum. Morbi luctus libero quis dictum laoreet. Etiam semper porta ante, vel ullamcorper enim sodales quis.

Nullam eu nisi faucibus, fermentum ex auctor, tempor arcu. Phasellus condimentum erat mi, condimentum malesuada ligula congue venenatis. Nullam gravida imperdiet urna quis cursus. Ut tempus nec purus eget posuere. Cras non nulla sit amet justo aliquet pellentesque nec sed eros. Nam aliquam nisl urna, in placerat magna gravida venenatis. Donec interdum vel magna ullamcorper molestie. Nunc felis neque, rhoncus fringilla faucibus sit amet, ultrices sed magna. Maecenas malesuada hendrerit diam in ultrices. Nam libero urna, volutpat ut auctor eget, interdum sed odio. Vestibulum suscipit mauris nec augue ornare, ut eleifend nulla gravida. Proin imperdiet, mauris quis consectetur porta, leo dui convallis leo,

id lobortis massa diam eu libero. Aenean hendrerit vel ante aliquam venenatis. Pellentesque bibendum pretium odio, ut sagittis lectus feugiat a. Donec porttitor vulputate lacus.

Nunc volutpat efficitur lacus in aliquet. Nullam non iaculis diam, at ultrices diam. Proin vehicula vulputate cursus. Morbi tempus sapien id urna lobortis interdum. Maecenas elementum sagittis elementum. Donec at sodales velit, a posuere tortor. Nulla id hendrerit tortor. Sed semper velit in magna sagittis pulvinar. Nulla nec arcu molestie, ultricies sapien sit amet, sollicitudin nisi. Donec nisi massa, suscipit ut dignissim quis, lacinia id leo.

Suspendisse ut mi metus. Morbi tincidunt ligula in porttitor consectetur. Integer eu urna urna. Suspendisse potenti. Mauris sit amet felis eu diam auctor ullamcorper. Morbi in porta nisi. Nam ante tortor, venenatis vitae tempor sed, sagittis vitae velit. In semper orci sit amet nisi ullamcorper varius. Aenean dignissim ultrices imperdiet. Maecenas lacinia enim id neque porttitor iaculis. Curabitur laoreet ante ut urna dignissim, id sollicitudin metus consectetur. Aenean massa ipsum, auctor vel ante vel, blandit dignissim libero. Fusce interdum ac magna et interdum.