Relational Neural Expectation Maximization:

Unsupervised Discovery of Objects and their Interactions

Sjoerd van Steenkiste
sjoerd@idsia.ch

Michael Chang
mbchang@berkeley.edu

Klaus Greff
klaus@idsia.ch

Jürgen Schmidhuber
juergen@idsia.ch

Scuola universitaria professionale della Svizzera italiana SUPSI
IDSIA
Università della Svizzera italiana

Summary

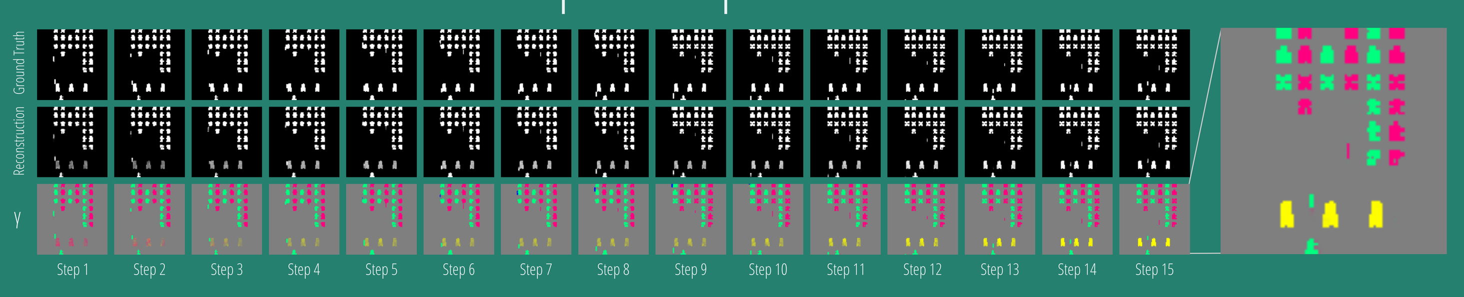
We introduce Relational Neural Expectation Maximization (R-NEM), a novel approach to common-sense physical reasoning that learns to discover objects and model their physical interactions from raw visual images

- R-NEM adds relational structure to Neural Expectation Maximization (N-EM), an unsupervised method that learns compositional object representations
- This enables it to learn interactions between objects and build a predictive model of a visual scene
- Using prior knowledge about the compositional nature of human perception, R-NEM factors interactions between object-pairs and learns efficiently
- On videos of bouncing balls we find that R-NEM learns an accurate world model that can be used for simulation
- R-NEM can extrapolate learned knowledge to scenes with additional objects and demonstrates a sense of object permanence when faced with occlusion

Motivation

- We humans rely on common-sense physical reasoning for many everyday tasks
- It is facilitated by the discovery and representation of objects, which serve as primitives of a compositional system
- This allows us to decompose a visual scene into distinct parts, describe relations between them and reason about their dynamics as well as the consequences of their interactions
- Successful previous approaches to physical reasoning incorporate such prior knowledge in their design, but require supervised information about objects
- Neural approaches that operate in pixel space offer an alternative but thus far have failed due to their lack of compositionality at the representational level of objects
- We address these problems by combining recent advances in symbol-like representation learning with insights from successful previous approaches to physical reasoning

Example: Space Invaders



R-NEM

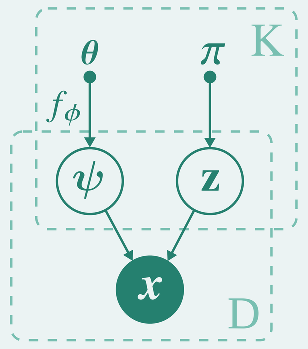
R-NEM is an unsupervised approach to common-sense physical reasoning that adds relational structure to Neural Expectation Maximization (N-EM):

N-EM

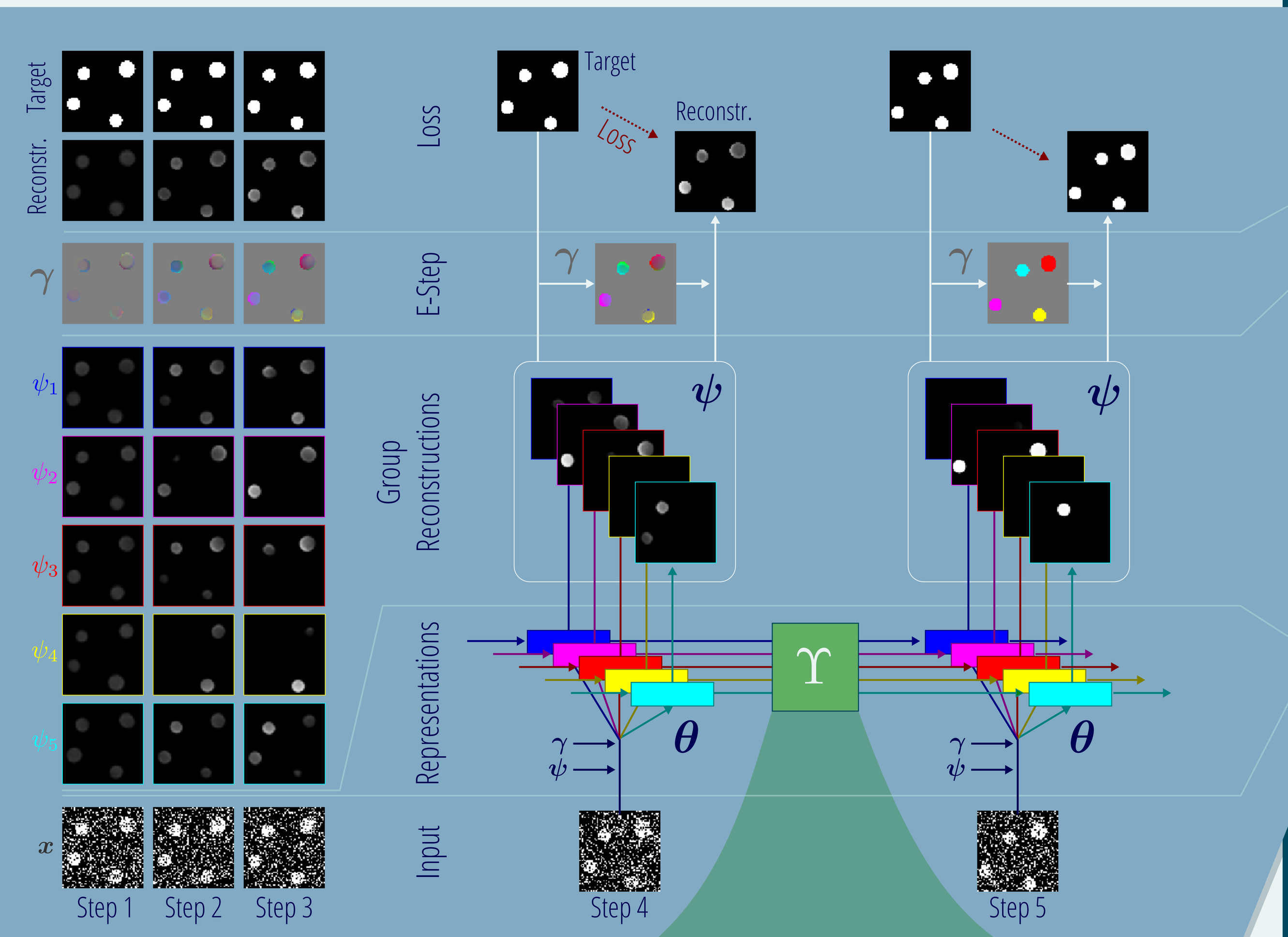
Derived from generalized EM inference in a spatial mixture model with a generative network f_ϕ in each component distribution $P(\boldsymbol{x}|\psi = f_\phi(\boldsymbol{\theta}))$

f_ϕ is trained to implement a distribution over images of objects given their representational form θ

For a given image its loss is computed throughout the unrolled generalized EM inference steps



- K copies of an RNN with an encoder-decoder architecture
- Each copy receives as input $\gamma_k(\psi_k(t-1) - \boldsymbol{x}(t))$ and outputs $\psi_k(t)$ corresponding to the future state of the world
- At each step γ is updated based on how well each RNN is able to model $\boldsymbol{x}(t+1)$ (E-step)
- Each RNN is encouraged to model a single object (or background)
- Once this has been achieved each $\boldsymbol{\theta}_k$ will correspond to an object-representation



Adding relational structure in the recurrence allows interactions between objects to be modelled:

$$\boldsymbol{\theta}_k^{(t)} = \mathrm{RNN}(\tilde{\boldsymbol{x}}^{(t)}, \Upsilon_k(\boldsymbol{\theta}^{(t-1)})) := \sigma(\boldsymbol{W} \cdot \tilde{\boldsymbol{x}}^{(t)} + \boldsymbol{R} \cdot \Upsilon_k(\boldsymbol{\theta}^{(t-1)}))$$

The inductive bias incorporated in Υ reflects our modelling assumptions about the interactions between objects in the environment

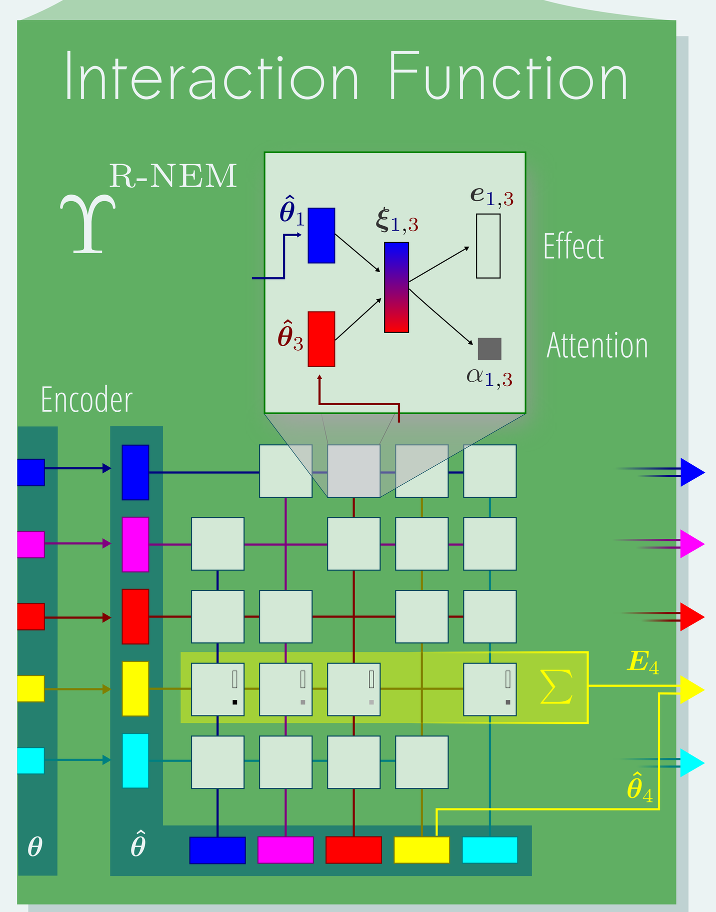
Here we adopt general but guiding constraints on how objects interact with one another

- The same physics apply to all objects
- Interactions among objects are factored into pairs
- The effect of one object on another object is fully determined by their states

Imposing these constraints preserves compositionality and allows interactions among objects to be learned efficiency

Using the identify function for Υ we obtain N-EM and are unable to model interactions between objects

Interaction Function



$$\hat{\boldsymbol{\theta}}_k = \mathrm{MLP}^{enc}(\boldsymbol{\theta}_k) \qquad \alpha_{k,i} = \mathrm{MLP}^{att}(\boldsymbol{\xi}_{k,i})$$

$$\boldsymbol{\xi}_{k,i} = \mathrm{MLP}^{emb}([\hat{\boldsymbol{\theta}}_k; \hat{\boldsymbol{\theta}}_i]) \qquad \boldsymbol{e}_{k,i} = \mathrm{MLP}^{eff}(\boldsymbol{\xi}_{k,i})$$

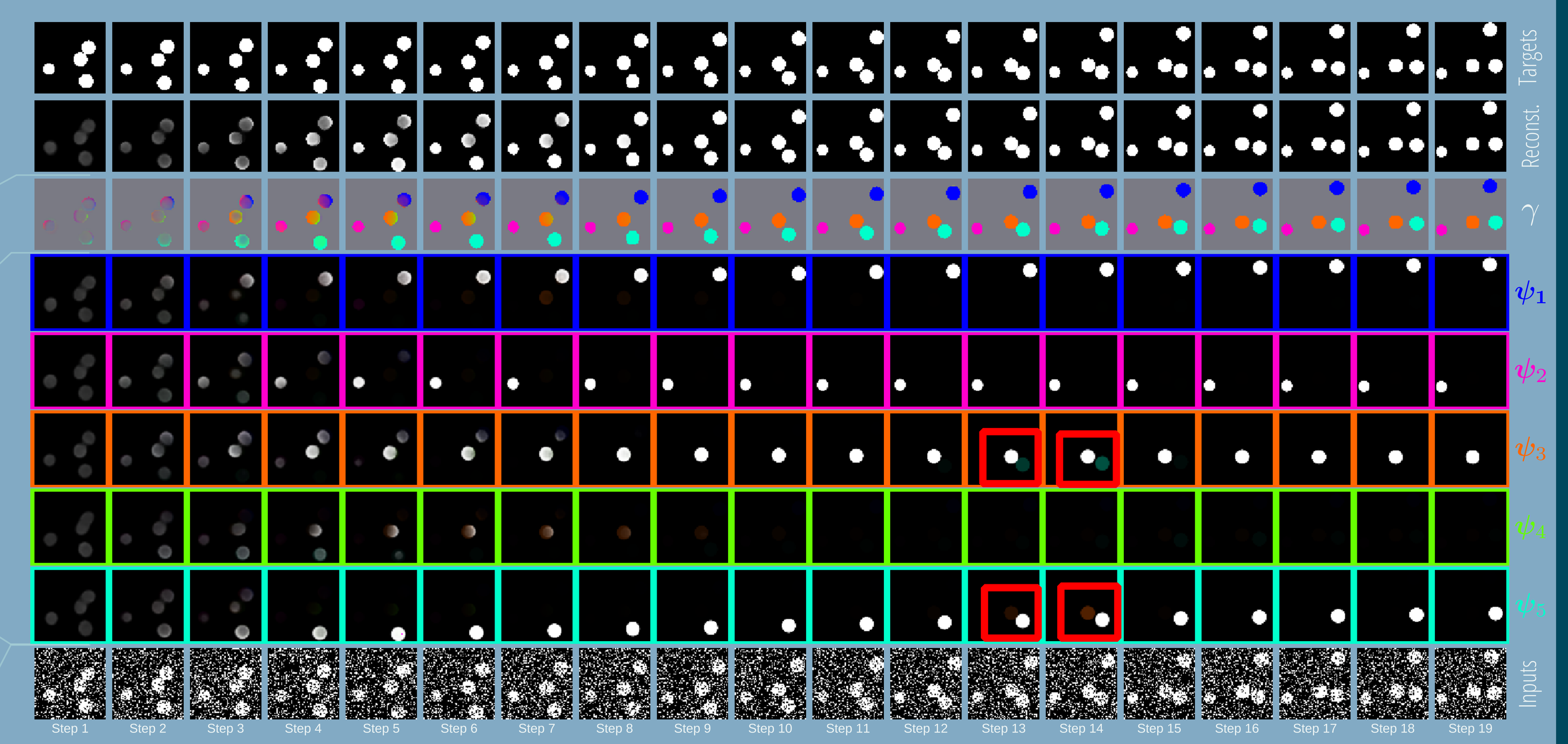
$$\Upsilon_k^{\text{R-NEM}}(\boldsymbol{\theta}) = [\hat{\boldsymbol{\theta}}_k; \boldsymbol{E}_k] \qquad \boldsymbol{E}_k = \sum_{i \neq k} \alpha_{k,i} \cdot \boldsymbol{e}_{k,i}$$

Results

R-NEM accurately models sequences of bouncing balls, unlike RNN & LSTM

Compared to N-EM it greatly reduces the relational loss

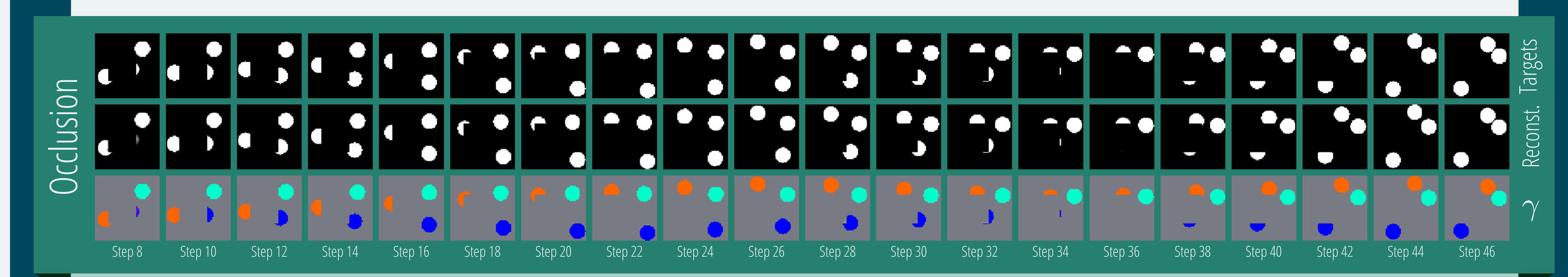
Increasing K is beneficial for grouping and lowers the loss



R-NEM can extrapolate learned physical dynamics

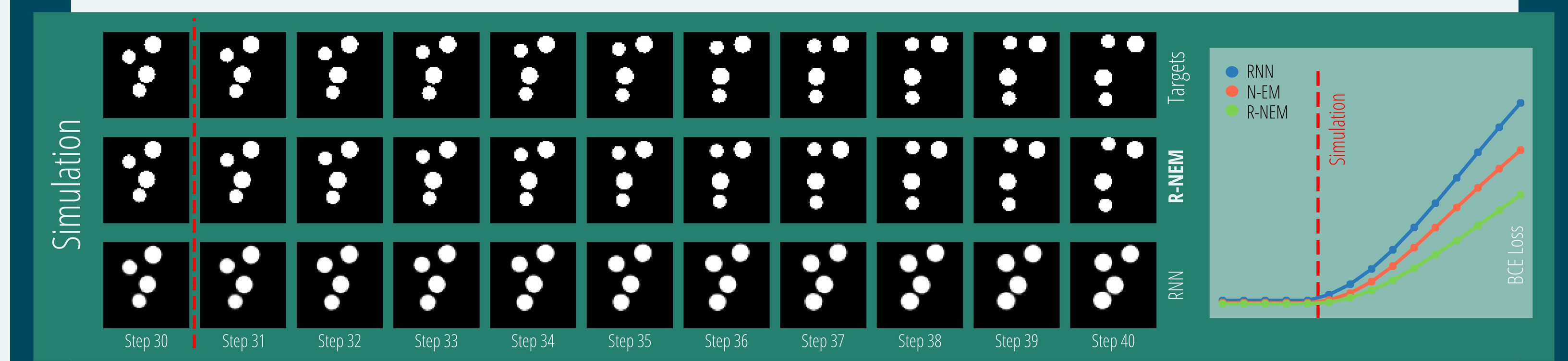
Trained on 5 balls, R-NEM is able to accurately model sequences with 6-8 balls

Unable to represent individual objects LSTM & RNN drop in relative performance



R-NEM is able to build a predictive model of the world in face of occlusion and demonstrates a sense of object permanence

By assigning persistence and identity to objects their interactions can be modelled without actually being observed



R-NEM builds a model of the environment that can be used for simulation

GitHub/SjoerdvanSteenkiste/Relational-NEM