Multi-Object Network (MONet):

From scene decomposition to a transferable model of attention

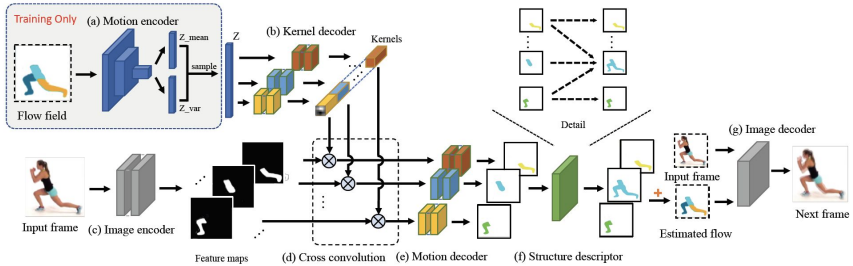
Gongqi Li & Benjamin Midler

Not this Monet

www.claude-monet.com

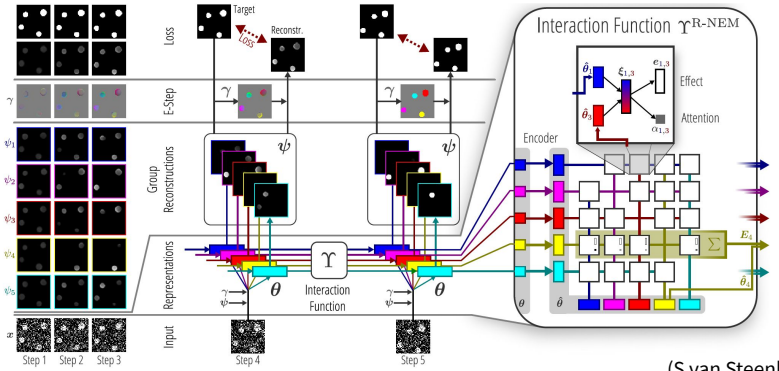
Architectures for Scene Decomposition

PSD



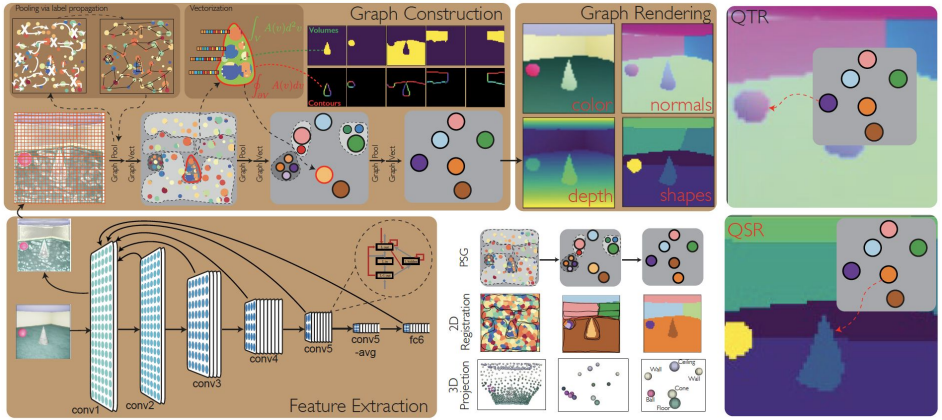
(Zhenjia Xu, 2019)

Relational-NEM



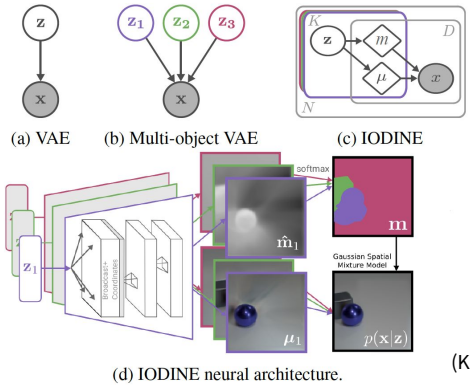
(S van Steenkiste, 2018)

Physical-Scene Graph



(DM Bear, 2020)

IODINE

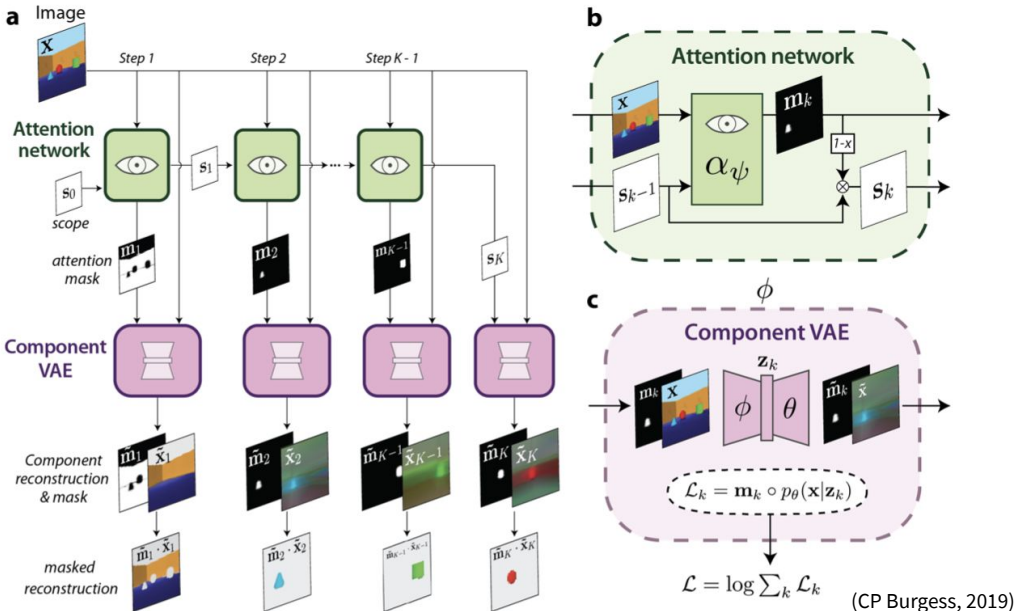


(K Geff, 2019)

MONet Architecture

- Original Motivation:
  - “Where those basic building blocks share meaningful properties, interactions and other regularities across scenes, such decompositions can simplify reasoning and facilitate imagination of novel scenarios.”
  - Unsupervised decomposition of an image into constituent components.
    - Most current approaches to object decomposition involve supervision, namely explicitly labeled segmentations in the dataset [Ronneberger et al., 2015, Jégou et al., 2017, He et al., 2017].
  - Representing image components as distinct entities promises to improve efficiency and transfer performance.

MONet Architecture

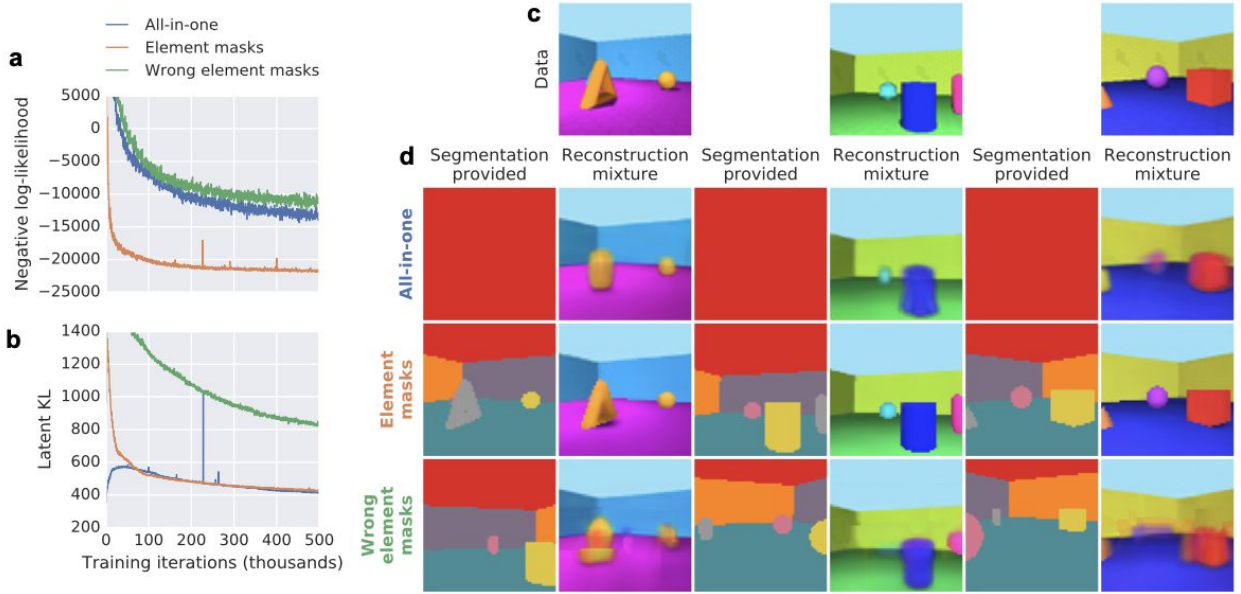


(CP Burgess, 2019)

$$\mathcal{L}(\phi;\theta;\psi;\mathbf{x}) = -\log \sum_{k=1}^{K} \mathbf{m}_k p_\theta(\mathbf{x}|\mathbf{z}_k) + \beta D_{KL}\Big(\prod_{k=1}^{K} q_\phi(\mathbf{z}_k|\mathbf{x},\mathbf{m}_k) \parallel p(\mathbf{z})\Big) + \gamma D_{KL}\big(q_\psi(\mathbf{c}|\mathbf{x}) \parallel p_\theta(\mathbf{c}|\{\mathbf{z}_k\})\big)$$

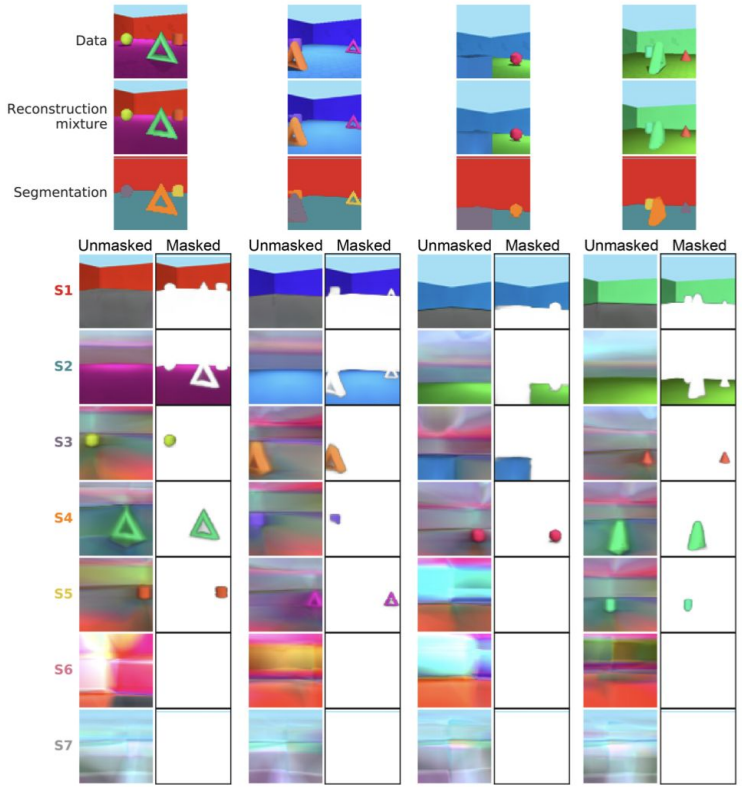
MONet Architecture

What are the results:



MONet Architecture

Scene decomposition:

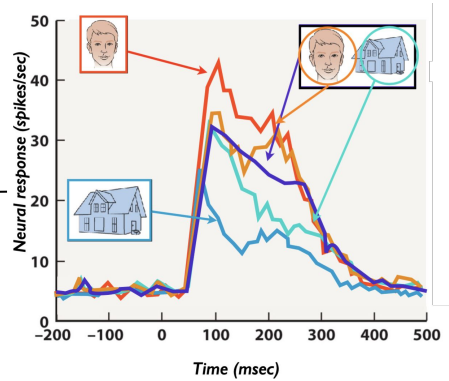


Question

How valid is MONet as a model of attention?

Biological Attention (Top-Down)

- Requires widespread neural activation (Beck et al., 2001).
  - Input from the parietal attention networks (dorsal visual stream), frontal eye field, etc.
- Particularly implicates the vertical occipital fasciculus.
- “The new VOF measurements provide insight into the communication between ventral stream regions involved in form perception and dorsal stream regions involved in eye movements and attention” (Yeatman et al., 2014).



Gain enhancement

From Justin Gardner, Psych 50

A brain-imaging discovery by Stanford scientists resolves a century-old argument

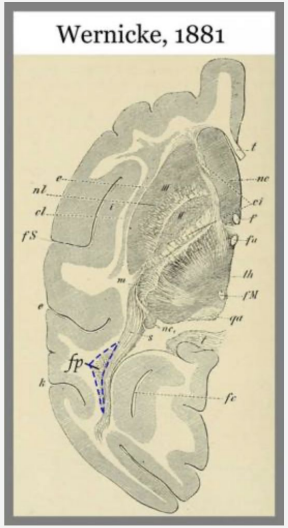
Nov 20 2014 | Stanford Report
BY AMY ADAMS

What started a few years ago as a brain-imaging study turned into a scientific mystery that eventually ended in the basement of Stanford's Lane Medical Library, within the pages of a book first published in 1881 and last checked out in 1912.

That journey, published this week in the *Proceedings of the National Academy of Sciences* , revealed the long and contentious history of an otherwise innocuous tract of nerve fibers of the visual system, running from just below to just above the ear. It also revealed the many ways scientific knowledge has been gained and lost over the centuries, and in some cases written out of history through a combination of scientific in-fighting and, at times, poor record-keeping.

The journey began when then-graduate student Jason Yeatman, co-first author of the recent paper, was carrying out brain-imaging studies to better understand how kids learn to read, in the lab of Brian Wandell, a professor of psychology. Yeatman noticed that all the brain images in his study contained a structure that didn't appear in any texts.

Either he'd discovered a new brain pathway or someone else had discovered it first, but the discovery and researcher had been lost to



The original reference to the vertical occipital fasciculus was published by Carl Wernicke in 1881. The dashed blue lines outline where Wernicke located the region. Jason Yeatman and Kevin Weiner found the illustration in Lane Medical Library.

Link

MONet Attention

- Recurrently generates an attention mask, outputting the portion of the current and subsequent scope that remains to be explained by the VAE.
- Segmentation into attention slots.
- Constrains the VAE to reconstruct the image portion masked by the attention network.

Biological Attention

- Enhances the sensitivity of visual neurons (does not impact selectivity).
- Involves broad range of neural structures, particularly inputs from the dorsal visual stream.
- Can attention slots be analogized to saccades?

Project Premise

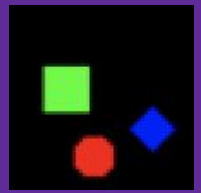
What are we specifically trying to figure out?

- If MONet's approach to attention modeling is valid, it should not only be capable of image decomposition, but also of other tasks associated with occipito-parietal attentional interactions.
  - "The anatomy suggests that the VOF carries signals between ventral regions that encode object properties including form, identify, and color information" (Takemura et al., 2014).
- What other abilities are exaptations of the original function of scene decomposition?

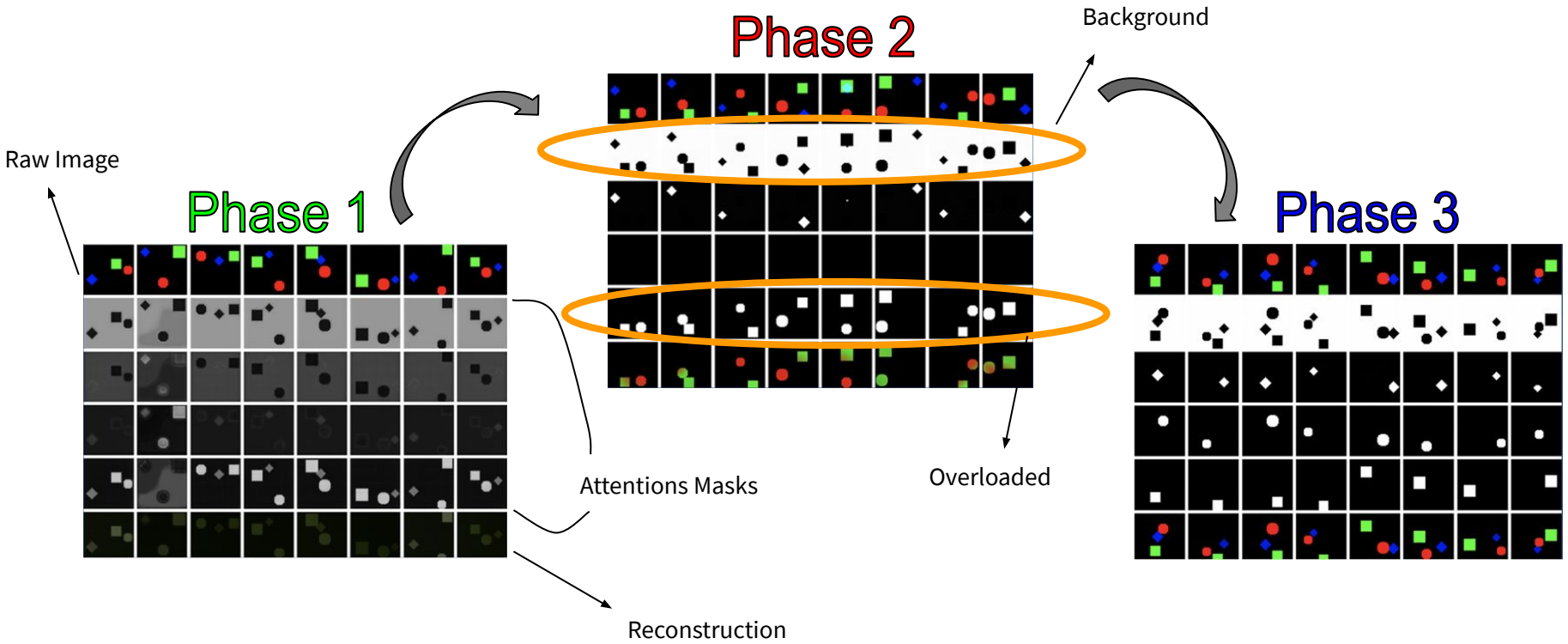
Methods

A tale of two empirically testable hypotheses

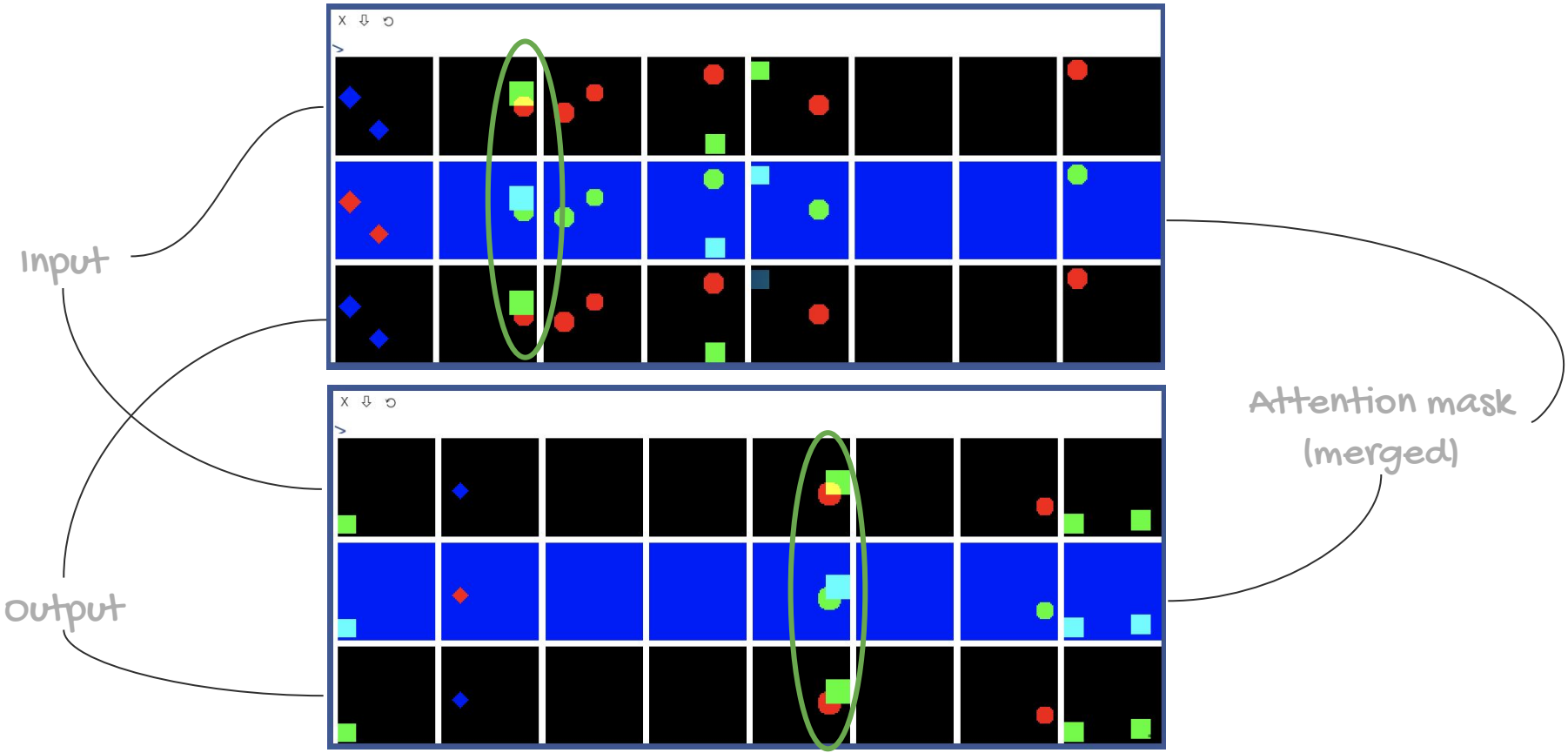
- Creating 2D images
- Training/dataset parameters



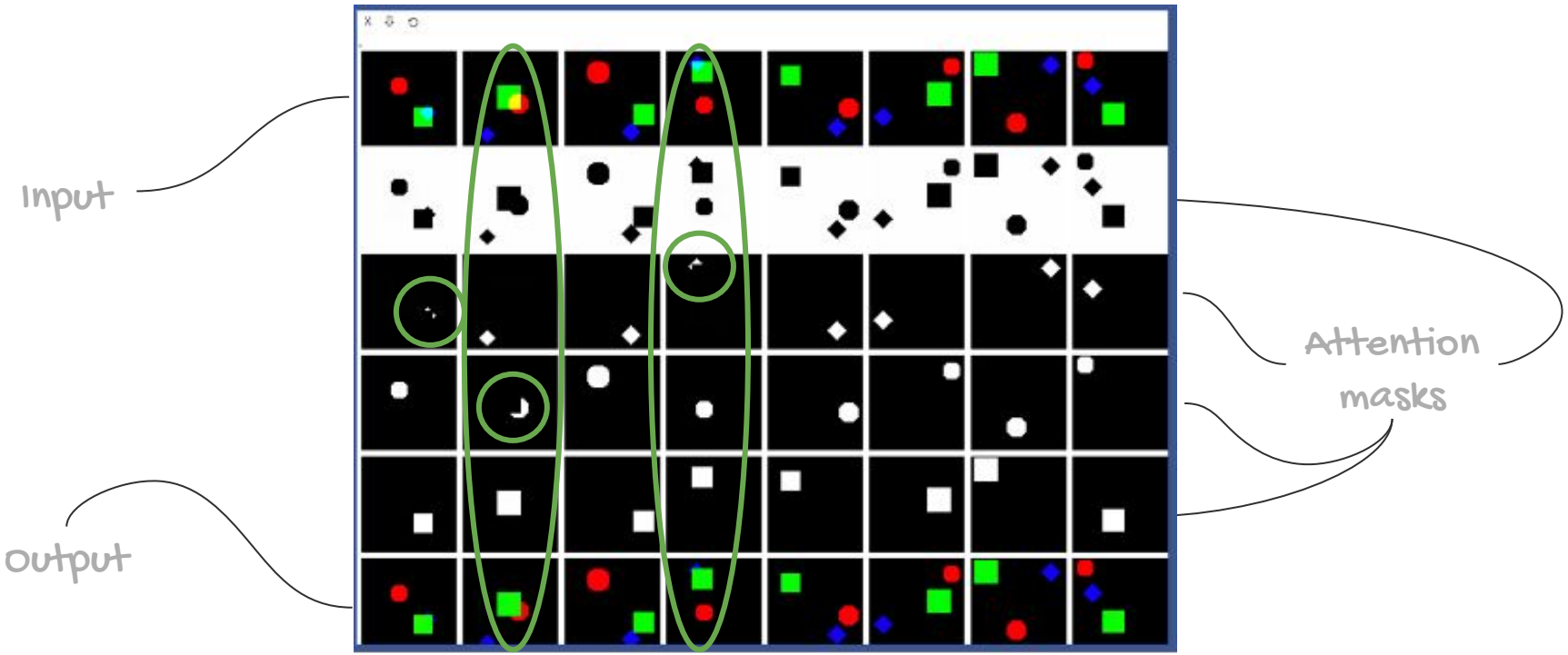
Learning Trajectory



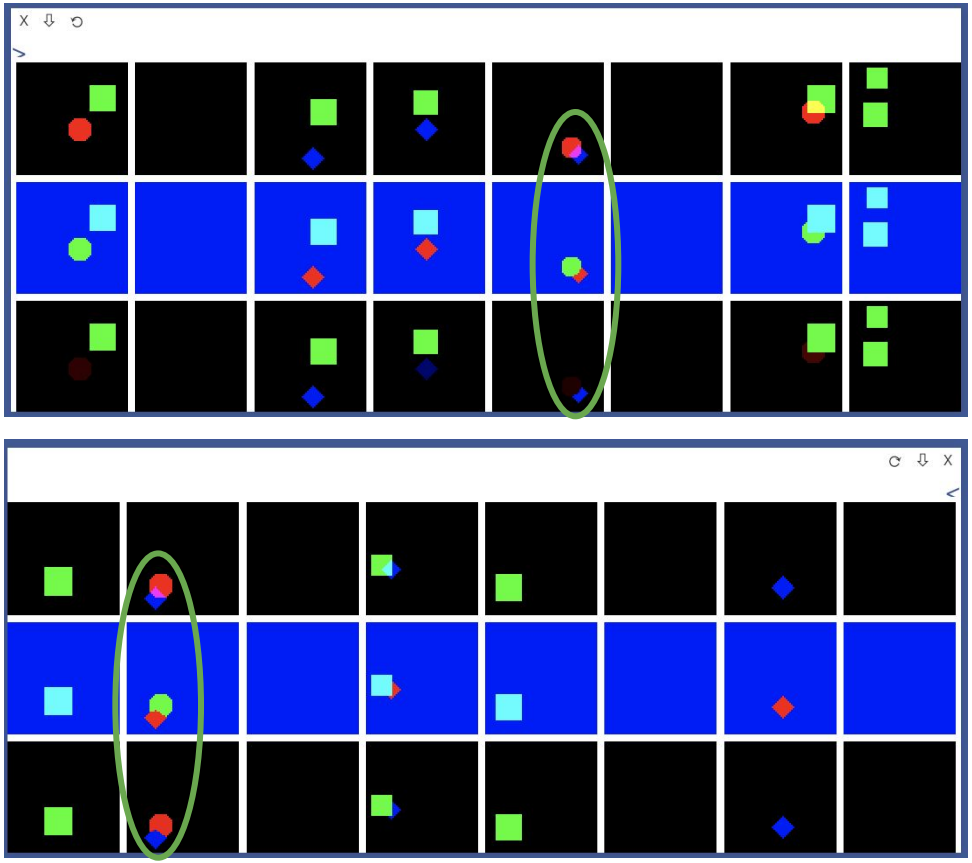
Occlusion (Successes)



Occlusion (Successes)

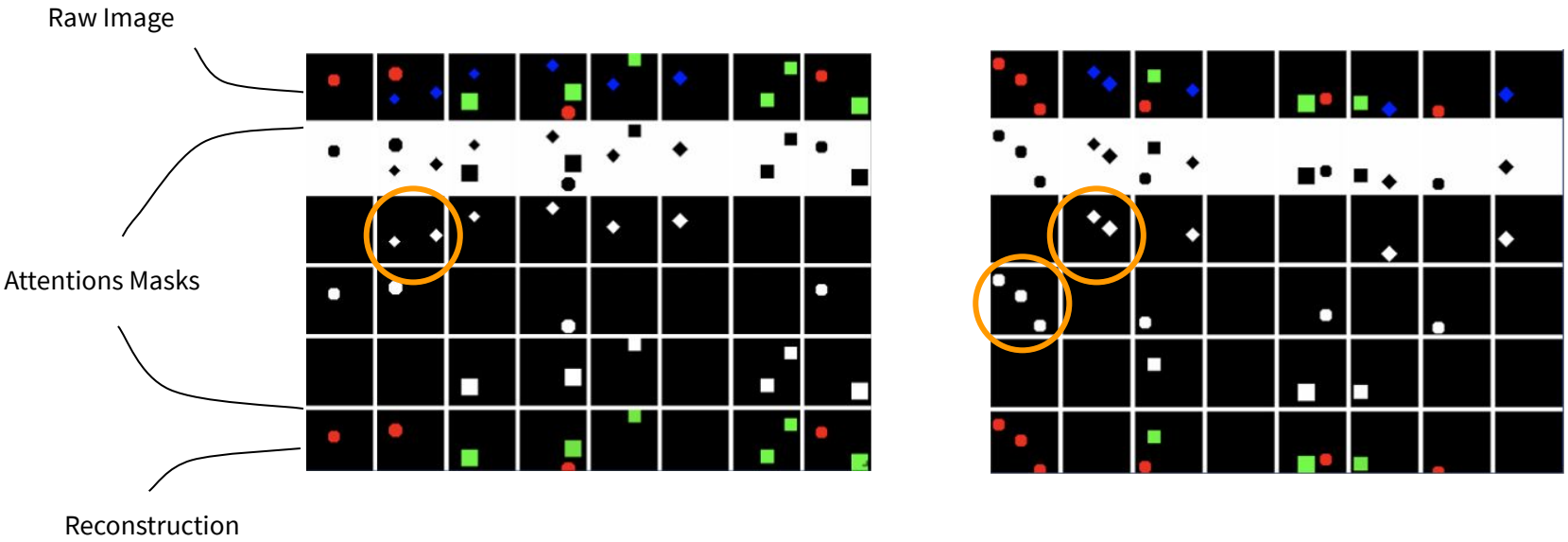


Occlusion (Not so Successful)



Visual Number Sense

- Number of activated attention masks indicates the number of objects
- Broken down by color



Conclusions

What have we learned?

- While the MONet attention mask is well-suited for scene decomposition, it does not generalize to other attention-related tasks.
  - Specifically, while the model does display color invariance, it does not have a sense of form or object constancy independent to color.