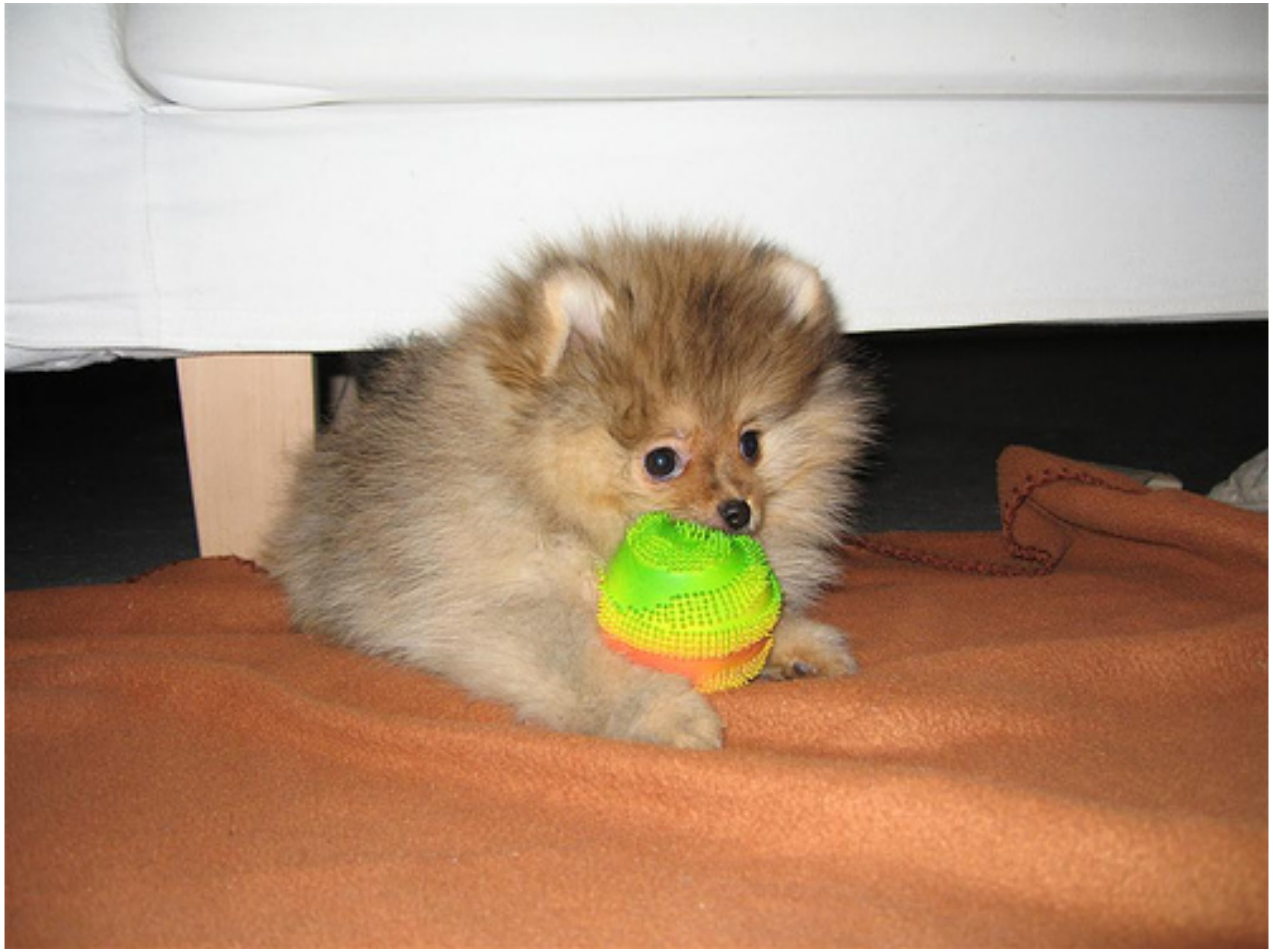
Learning visual representations for active perception

Bruno Olshausen
Helen Wills Neuroscience Institute, School of Optometry
Redwood Center for Theoretical Neuroscience, UC Berkeley

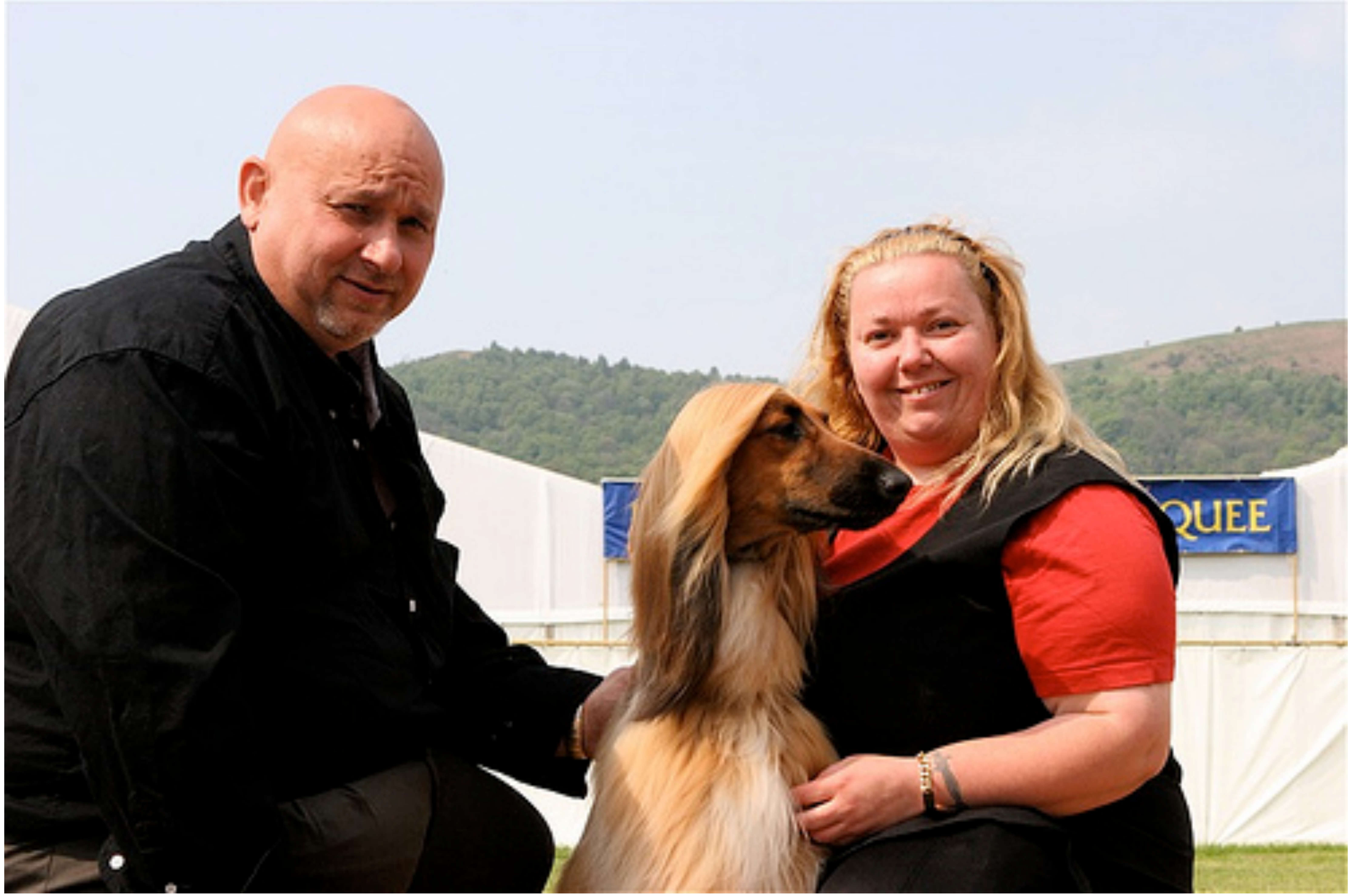
Brian Cheung
Vision Science

Eric Weiss
Neuroscience

REDWOOD CENTER
for Theoretical Neuroscience



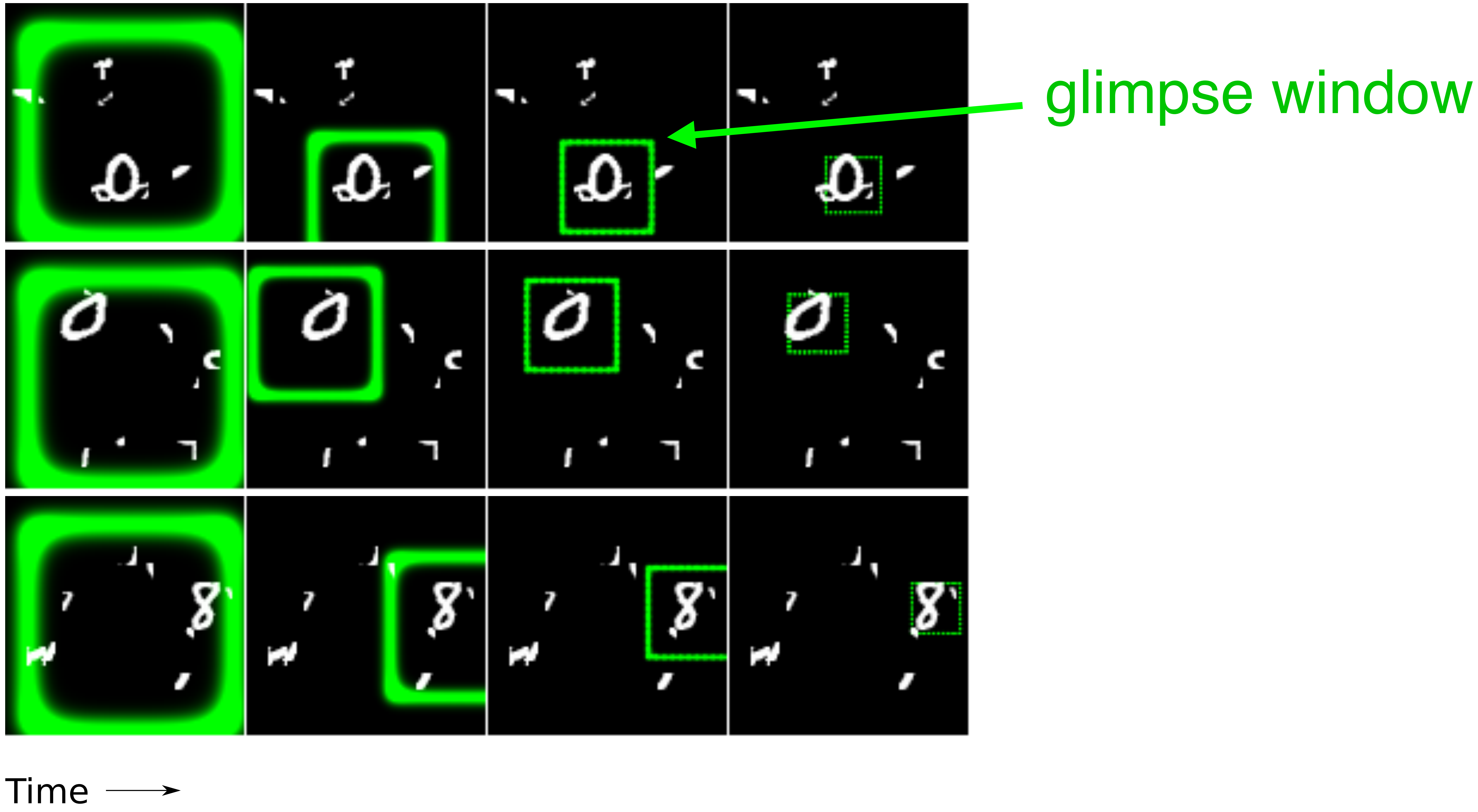
Correct label: Pomeranian



Correct label: Afghan hound

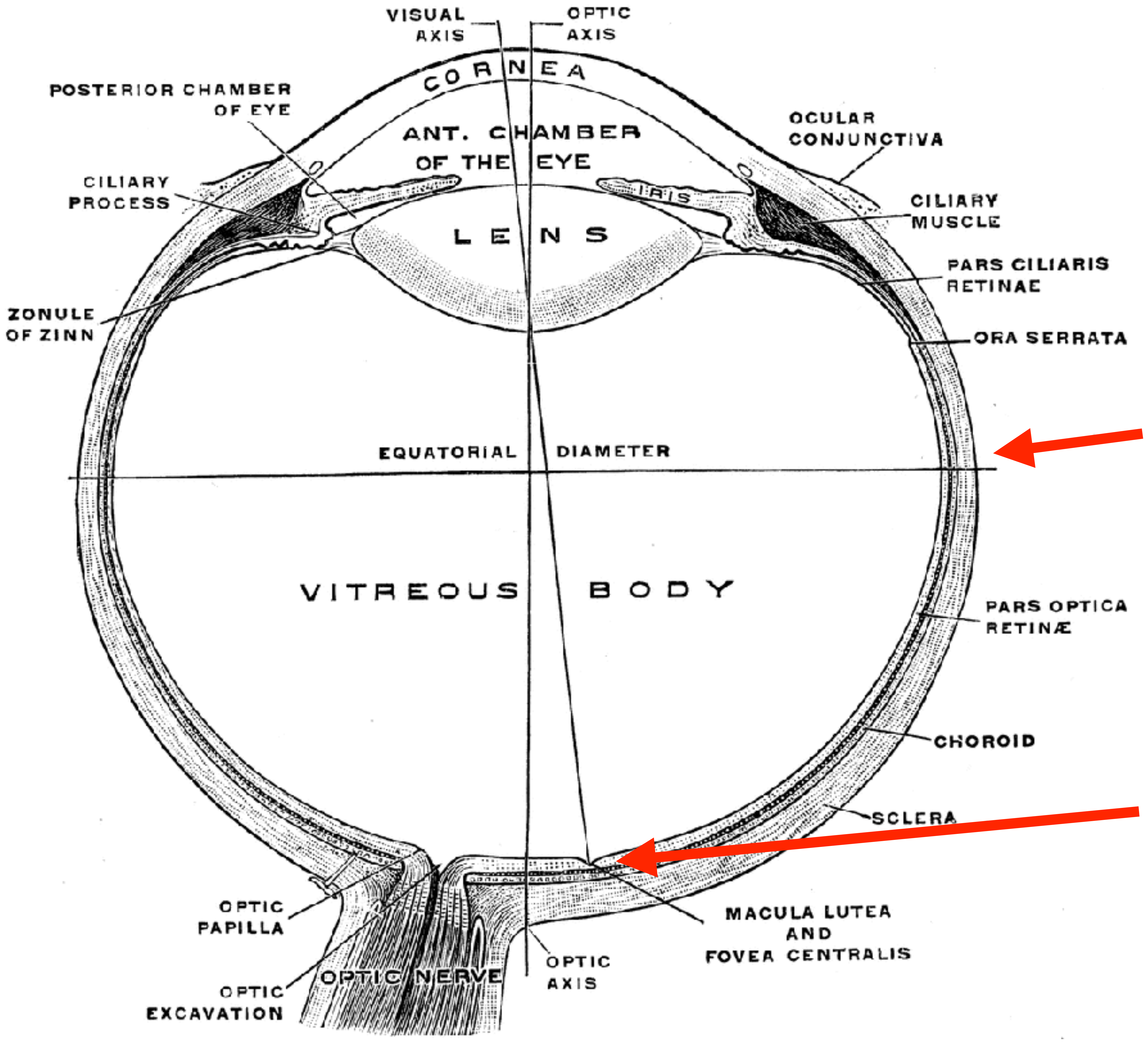
DRAW: A Recurrent Neural Network For Image Generation

Karol Gregor KAROLG@GOOGLE.COM
Ivo Danihelka DANIHELKA@GOOGLE.COM
Alex Graves GRAVESA@GOOGLE.COM
Danilo Jimenez Rezende DANILOR@GOOGLE.COM
Daan Wierstra WIERSTRA@GOOGLE.COM
Google DeepMind



Two questions

- What is the optimal sampling lattice for the glimpse window?
- How is information combined across glimpses?

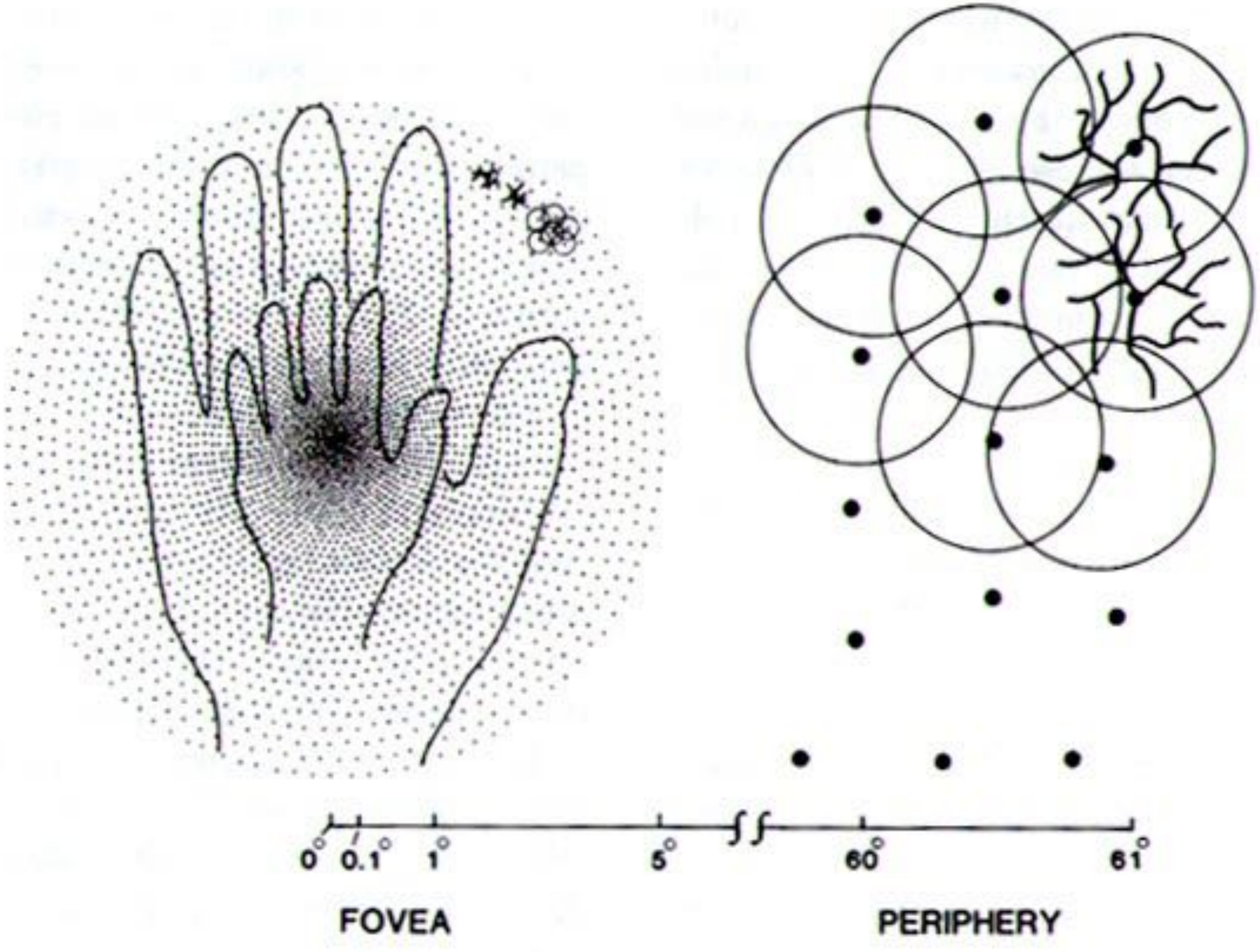


Low resolution

High resolution

Retinal ganglion cell sampling lattice

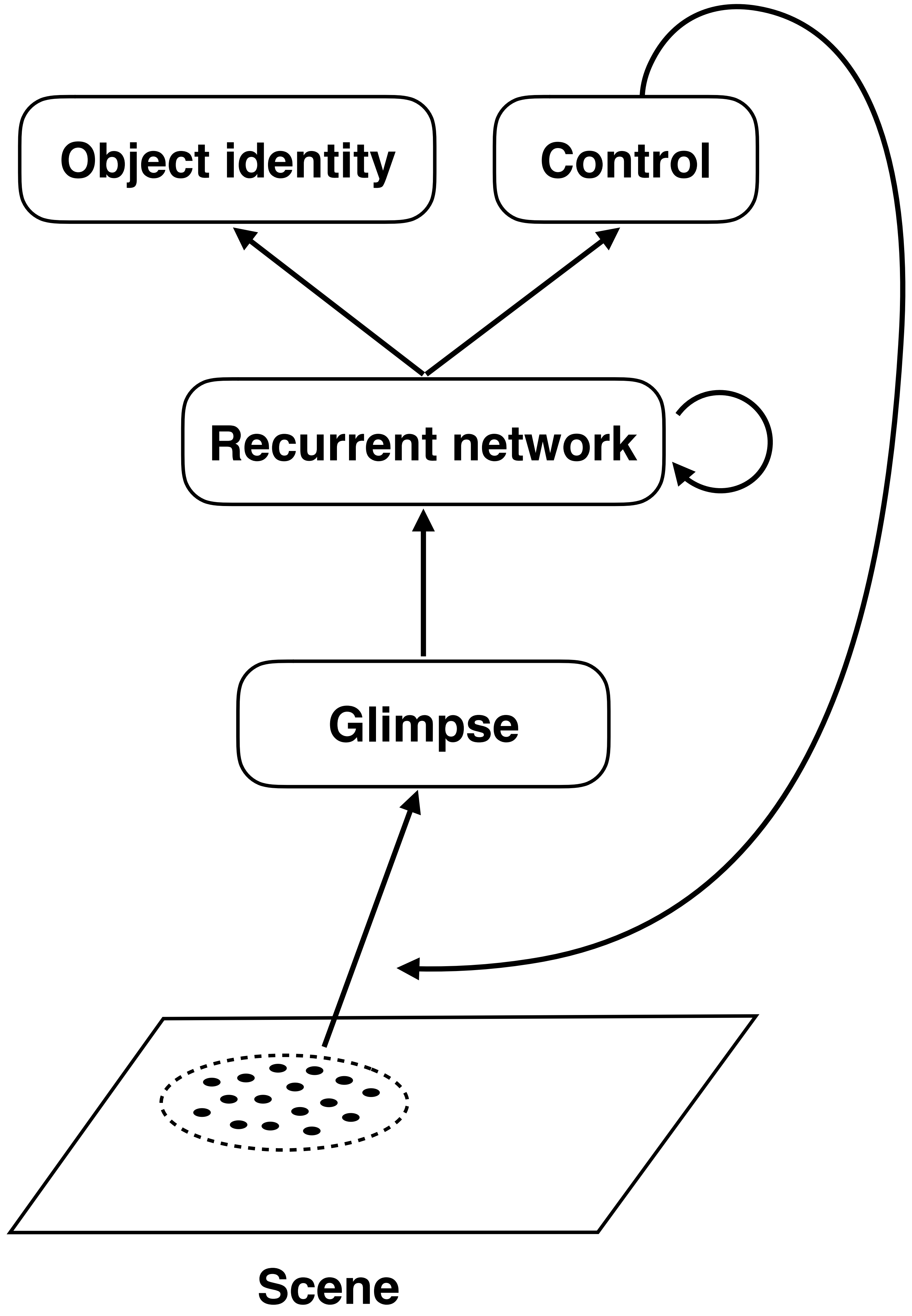
(shown at one dot for every 20 ganglion cells)



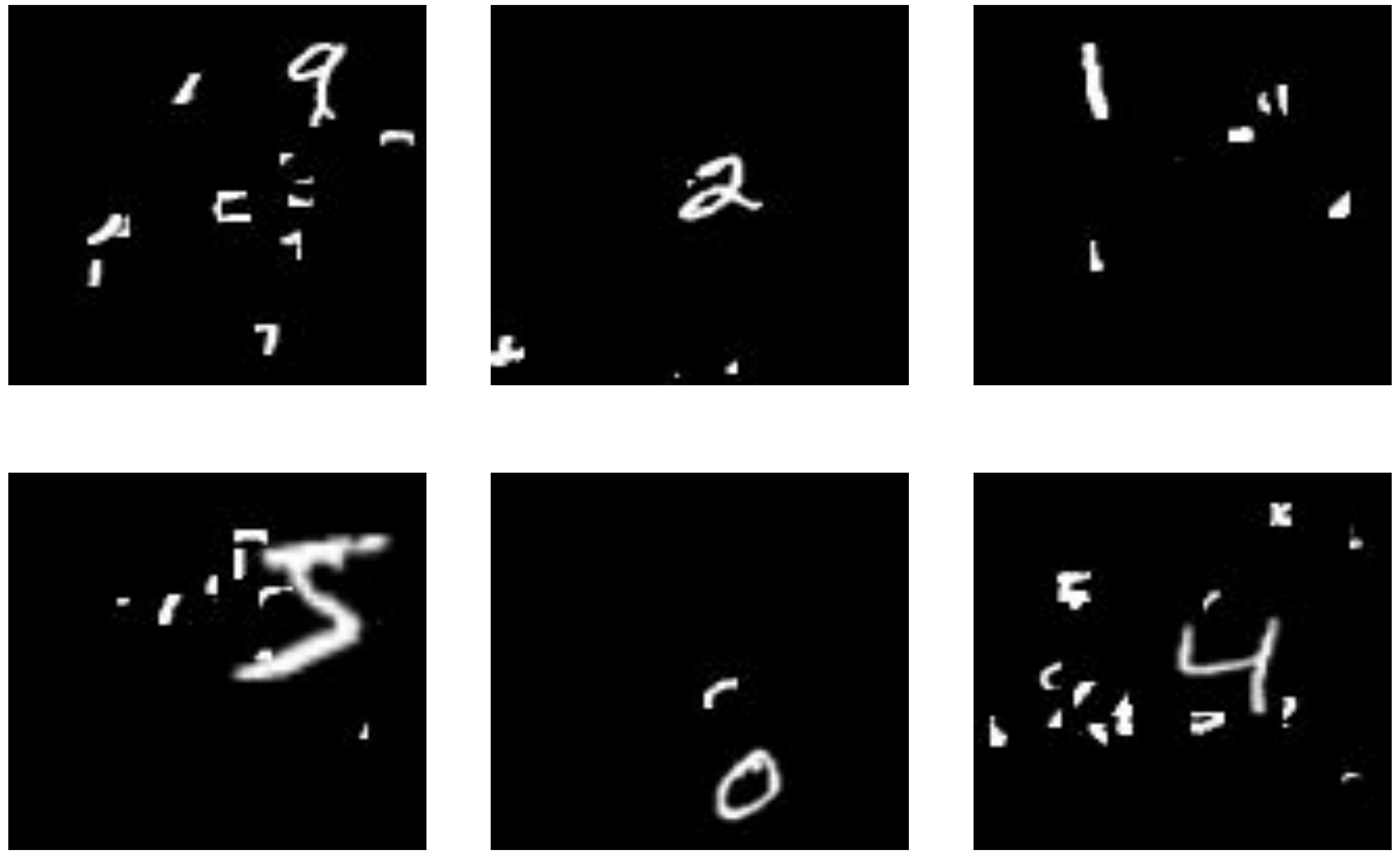
$$\Delta \approx .01(|E| + 1)$$

Anderson & Van Essen (1995)

Learning the glimpse window sampling lattice



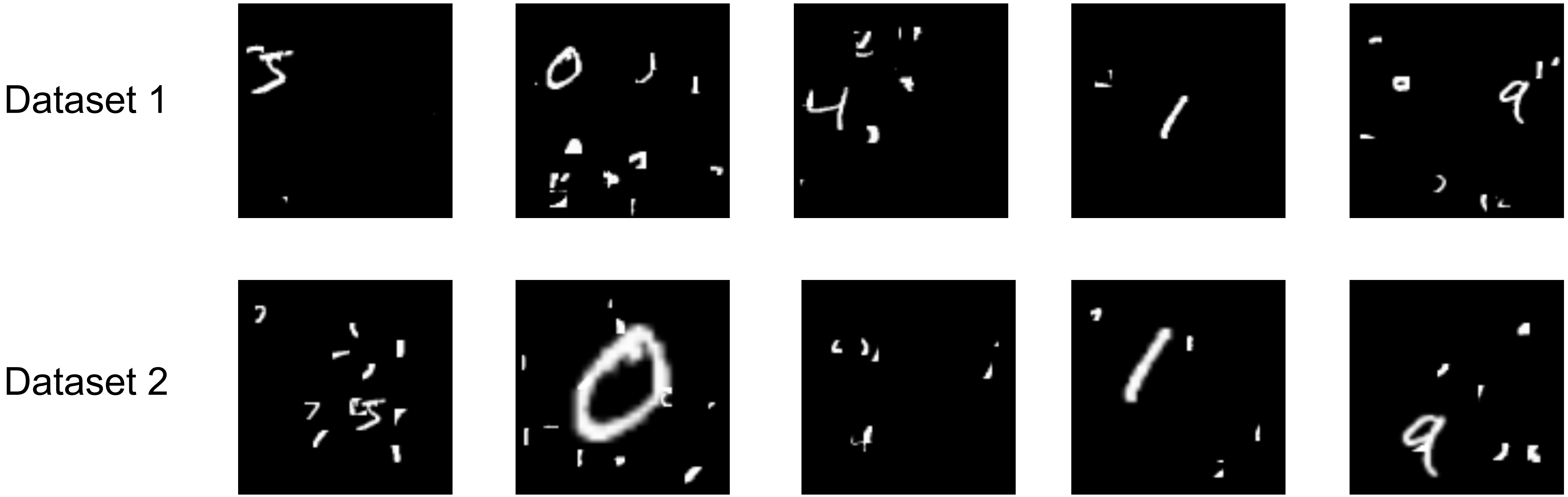
- Network is trained to correctly classify the digit in the scene.
- To do this it must find a digit and move its glimpse window to that location.



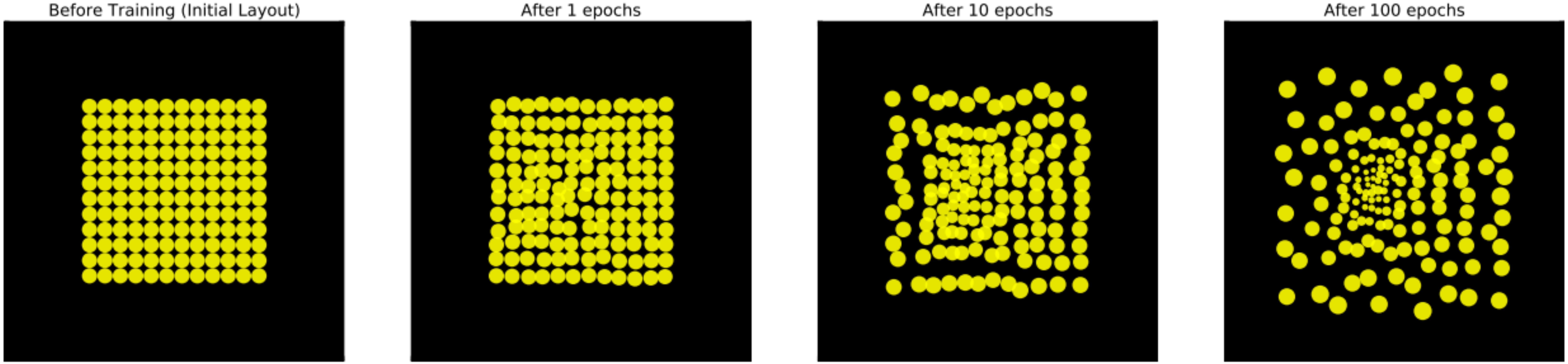
Example MNIST scenes

Visual Search Task

Find and Classify the MNIST digit



Evolution of the sampling lattice during training



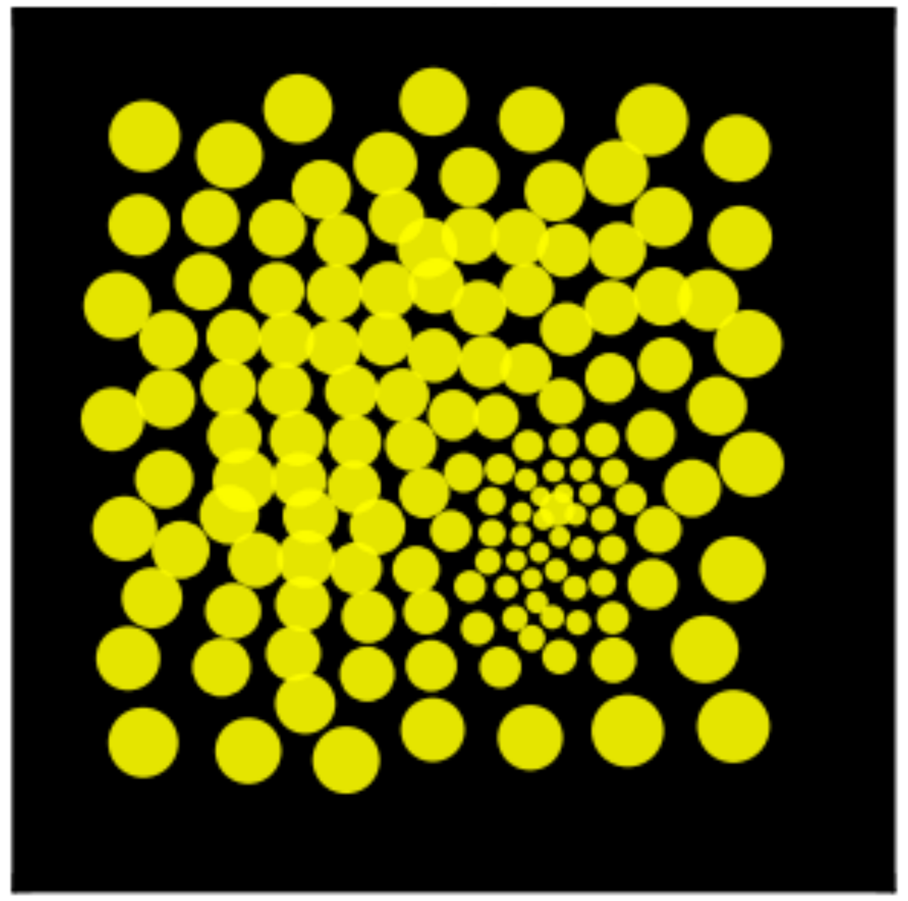
Before Training (Initial Layout)

After 1 epochs

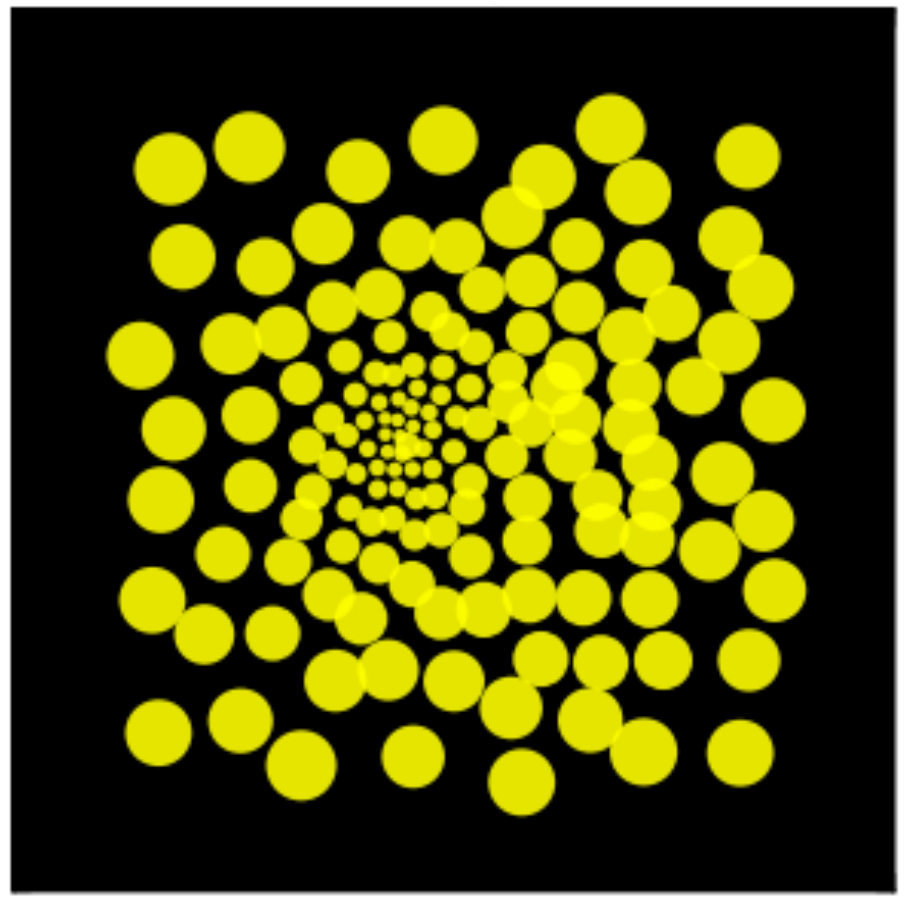
After 10 epochs

After 100 epochs

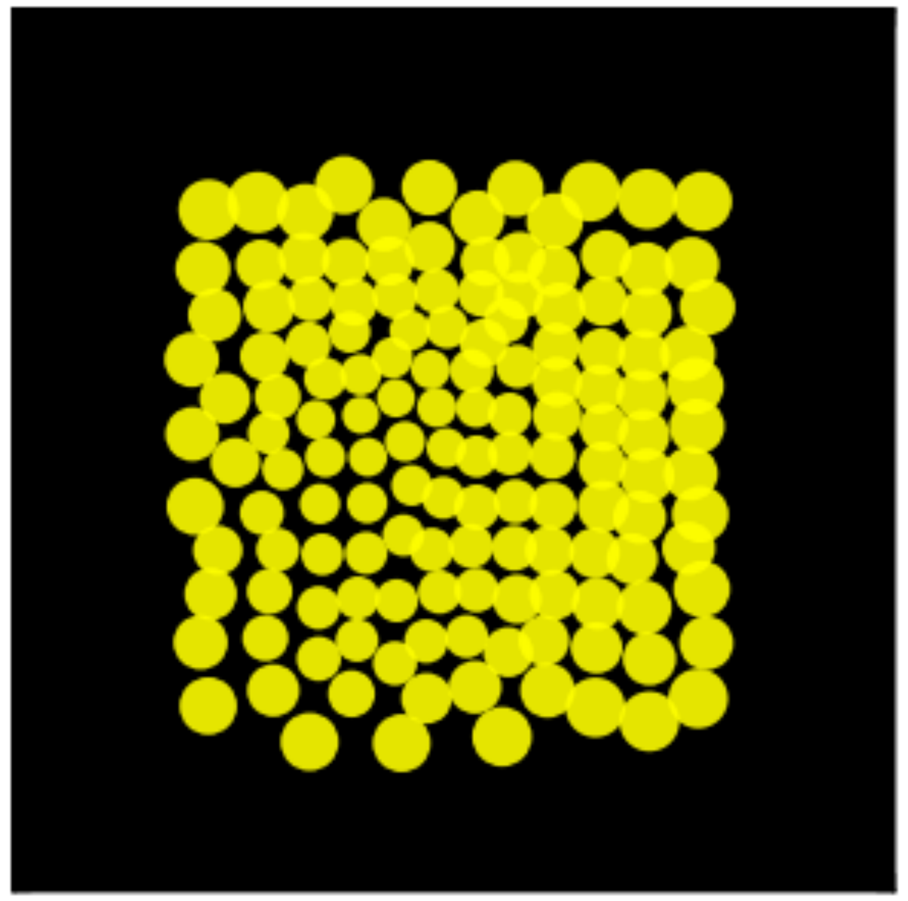
Learned sampling lattices for different conditions



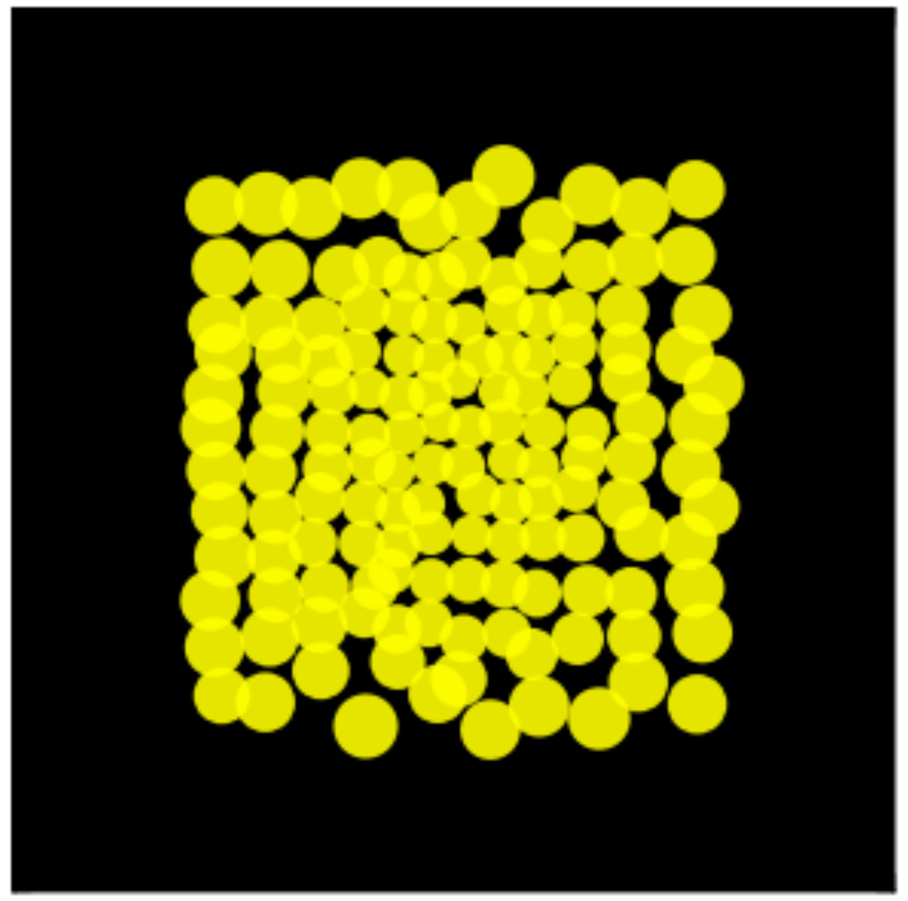
Translation only
(Dataset 1)



Translation only
(Dataset 2)

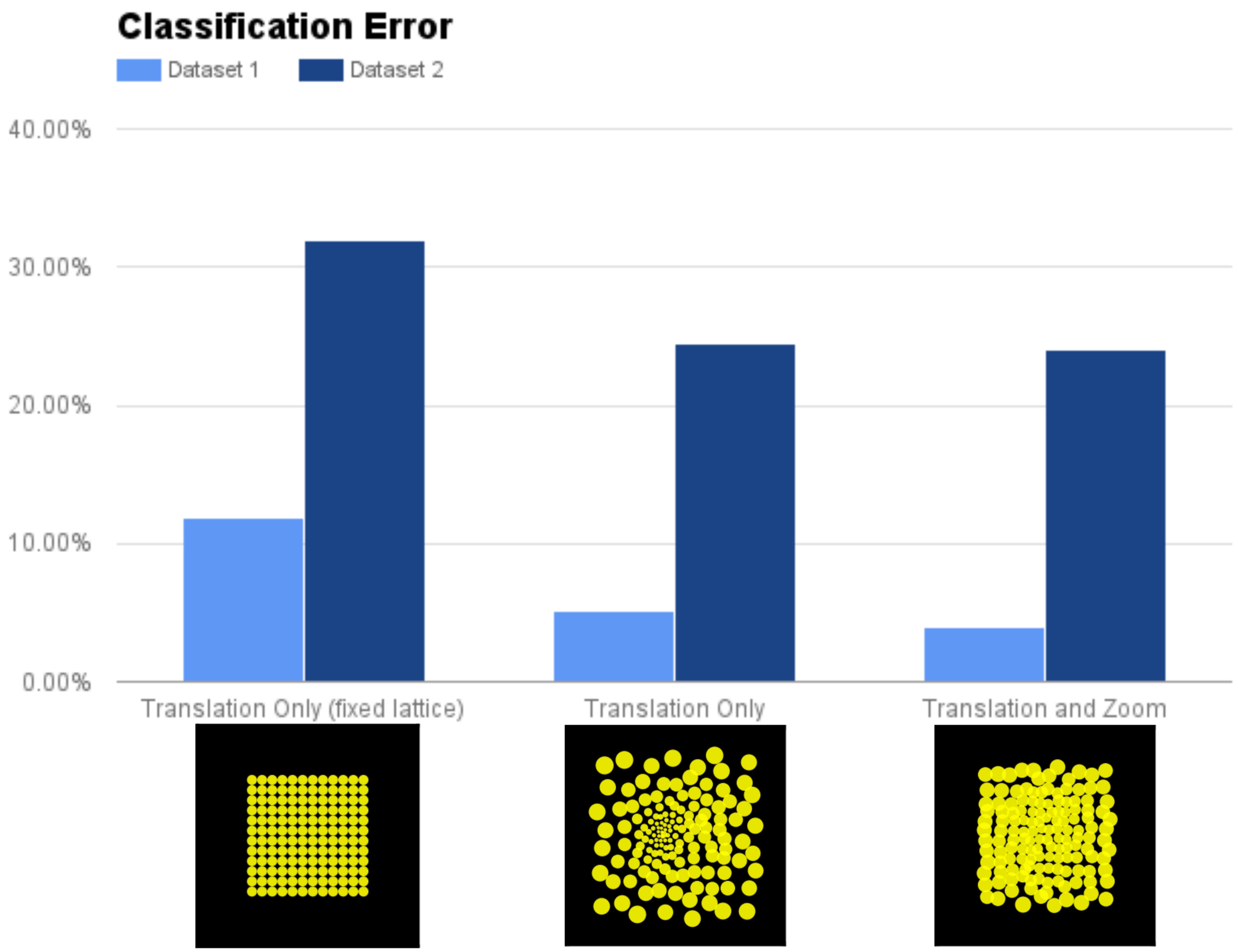


Translation & zoom
(Dataset 1)



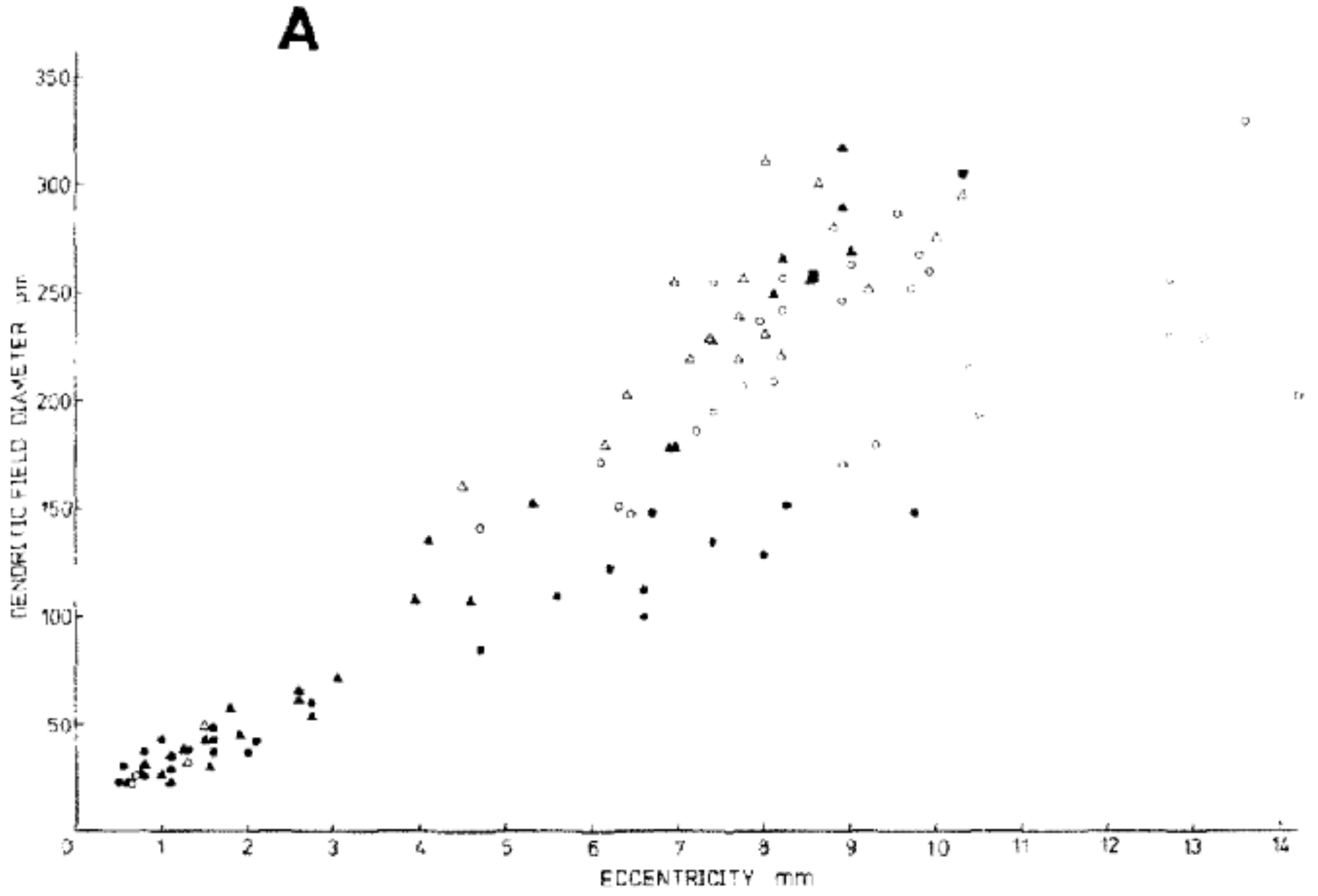
Translation & zoom
(Dataset 2)

Visual Search Performance



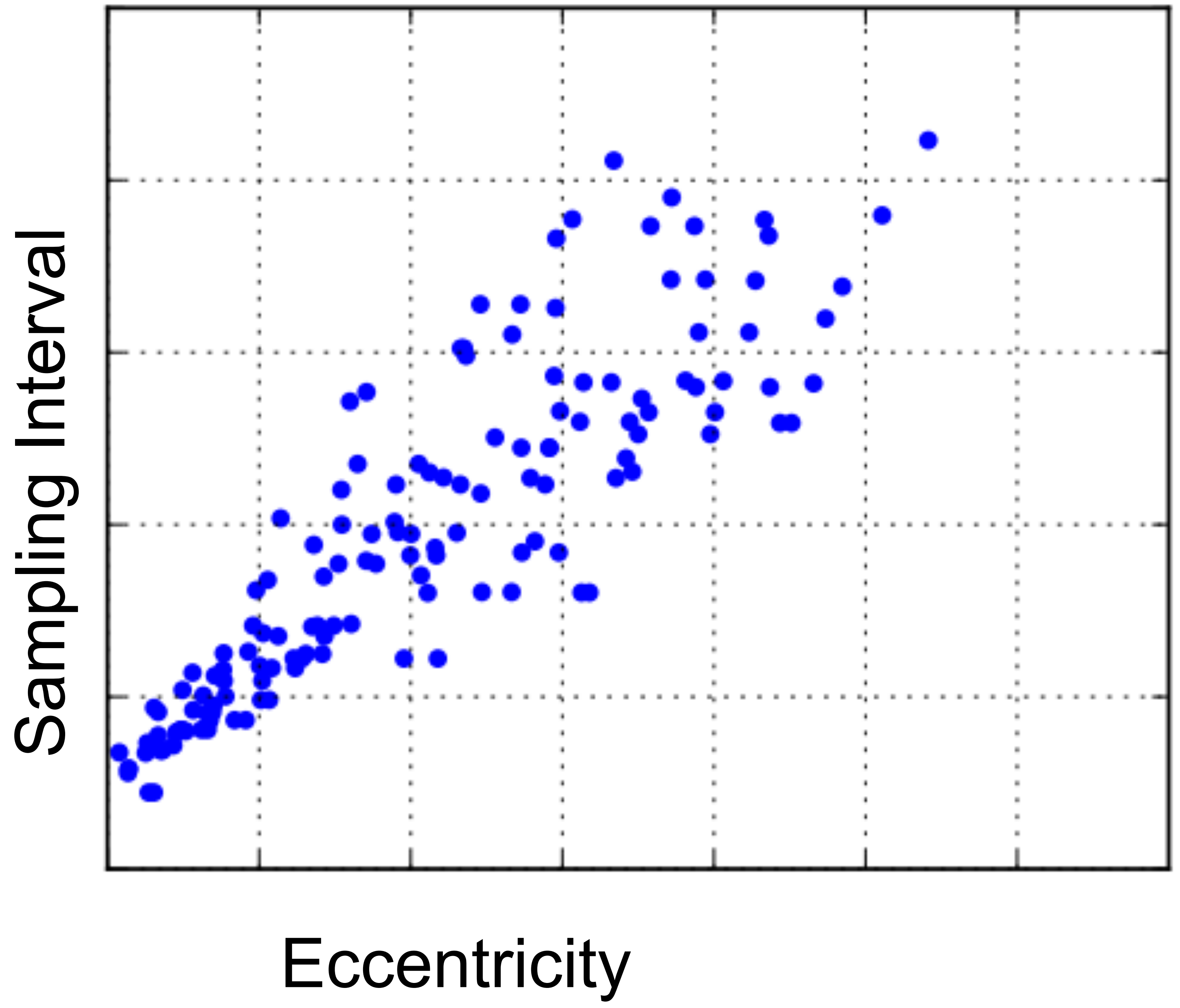
Comparison to primate retina

Macaque Retina



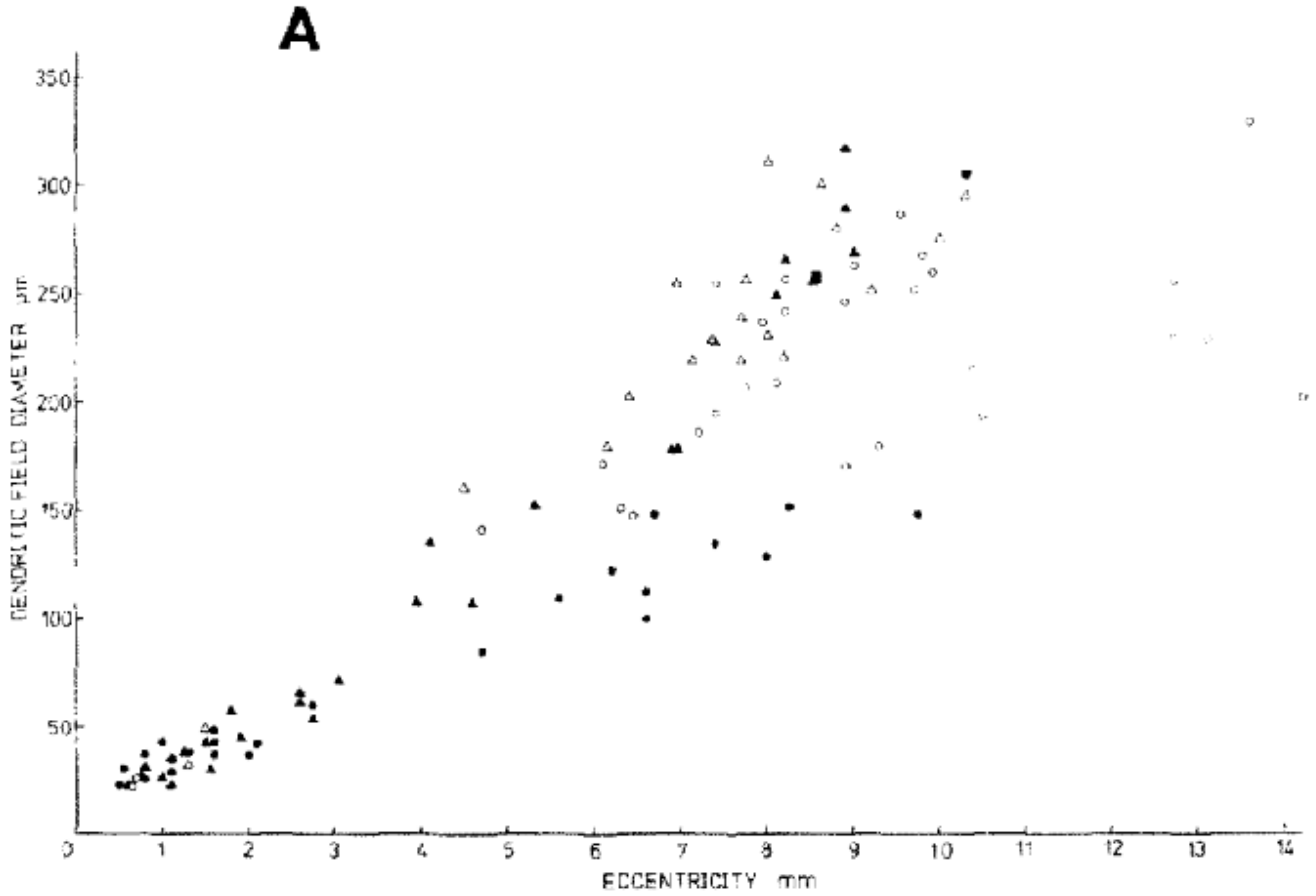
Perry, Oehler, Cowey 1984

Model



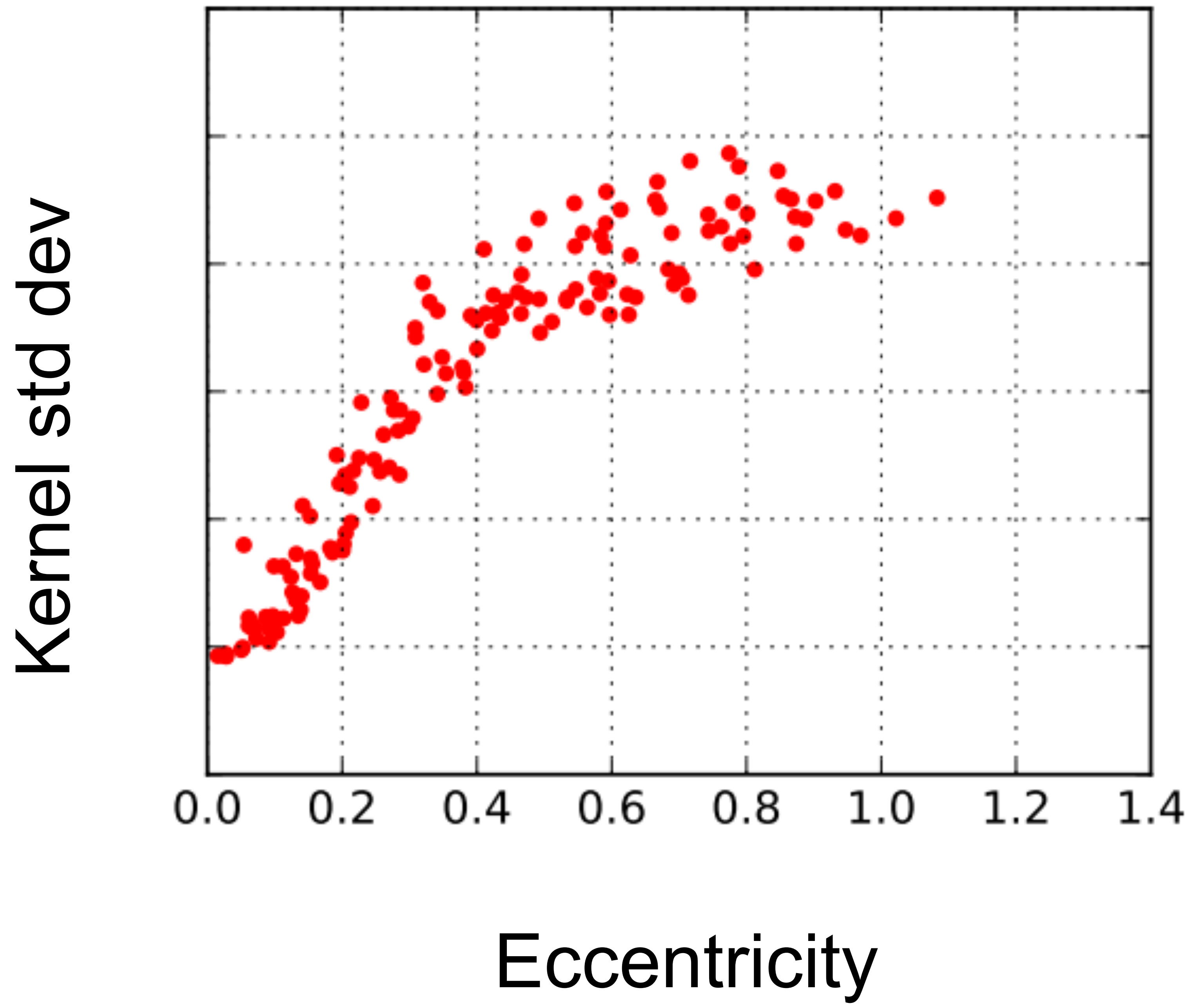
Comparison to primate retina

Macaque Retina



Perry, Oehler, Cowey 1984

Model



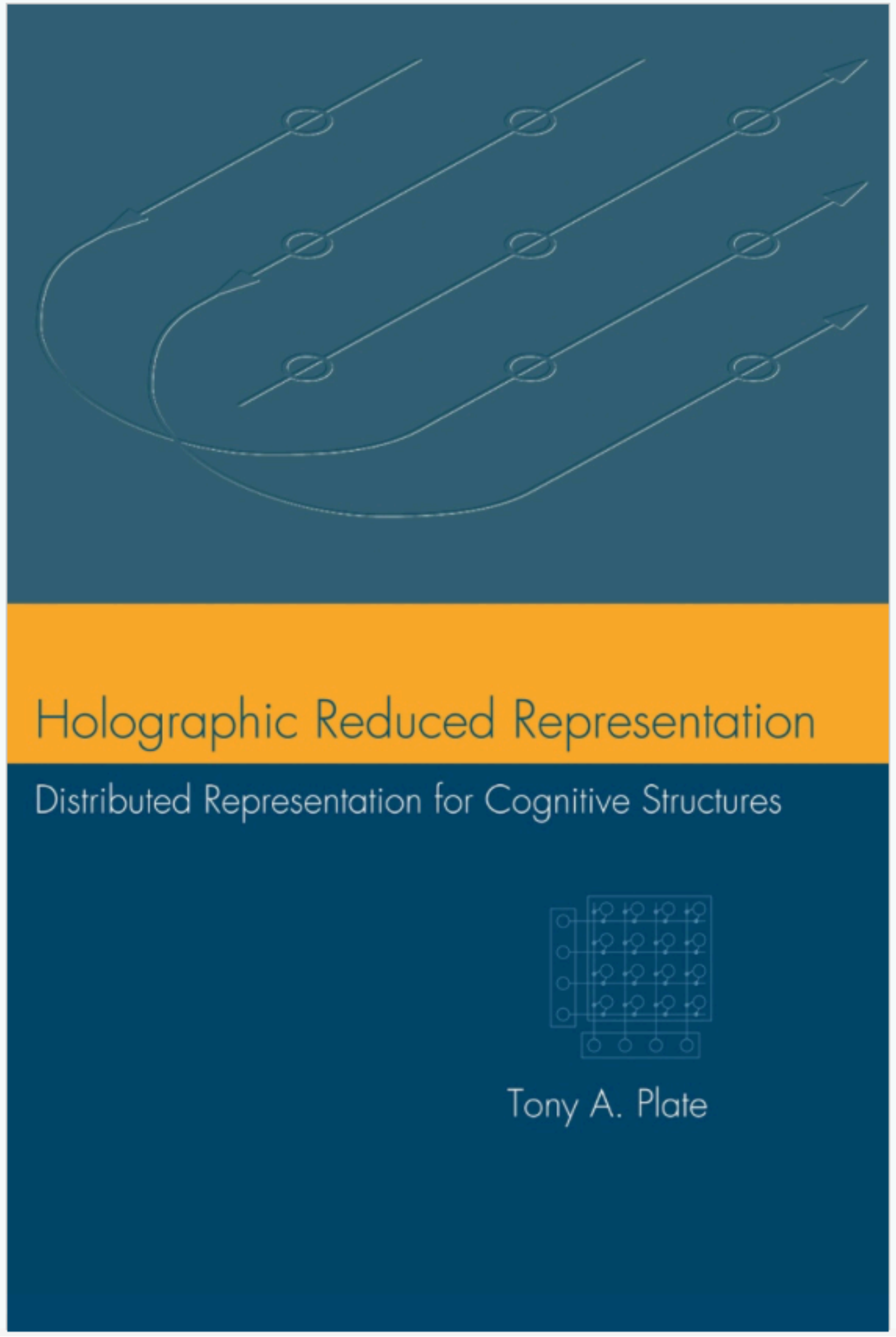
How is information combined across glimpses?

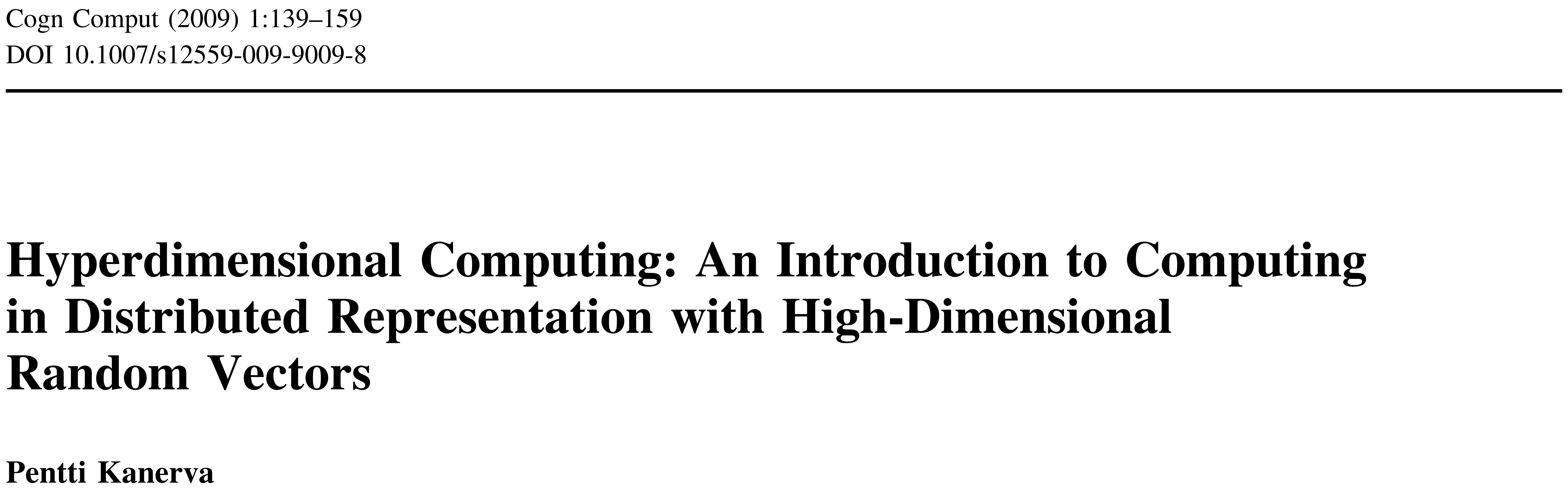
Two things must be encoded and combined at each fixation:

1) *position* of the glimpse window
2) *contents* of the glimpse window

What is required is to *bind* these two things together!

A scene may then be represented as a superposition of such bindings.


Holographic Reduced Representation
Distributed Representation for Cognitive Structures
Tony A. Plate



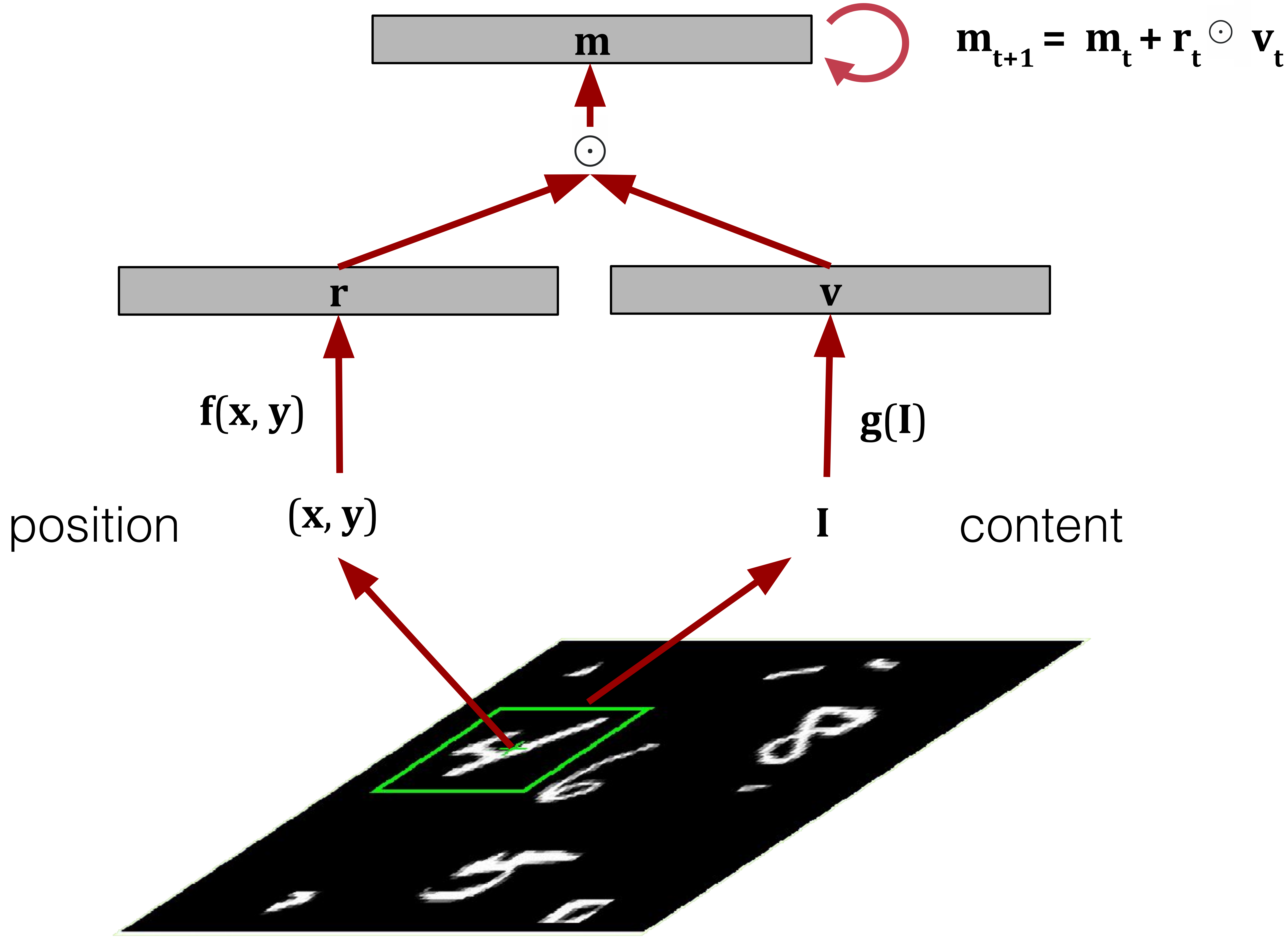
Cogn Comput (2009) 1:139–159
DOI 10.1007/s12559-009-9009-8

Hyperdimensional Computing: An Introduction to Computing in Distributed Representation with High-Dimensional Random Vectors

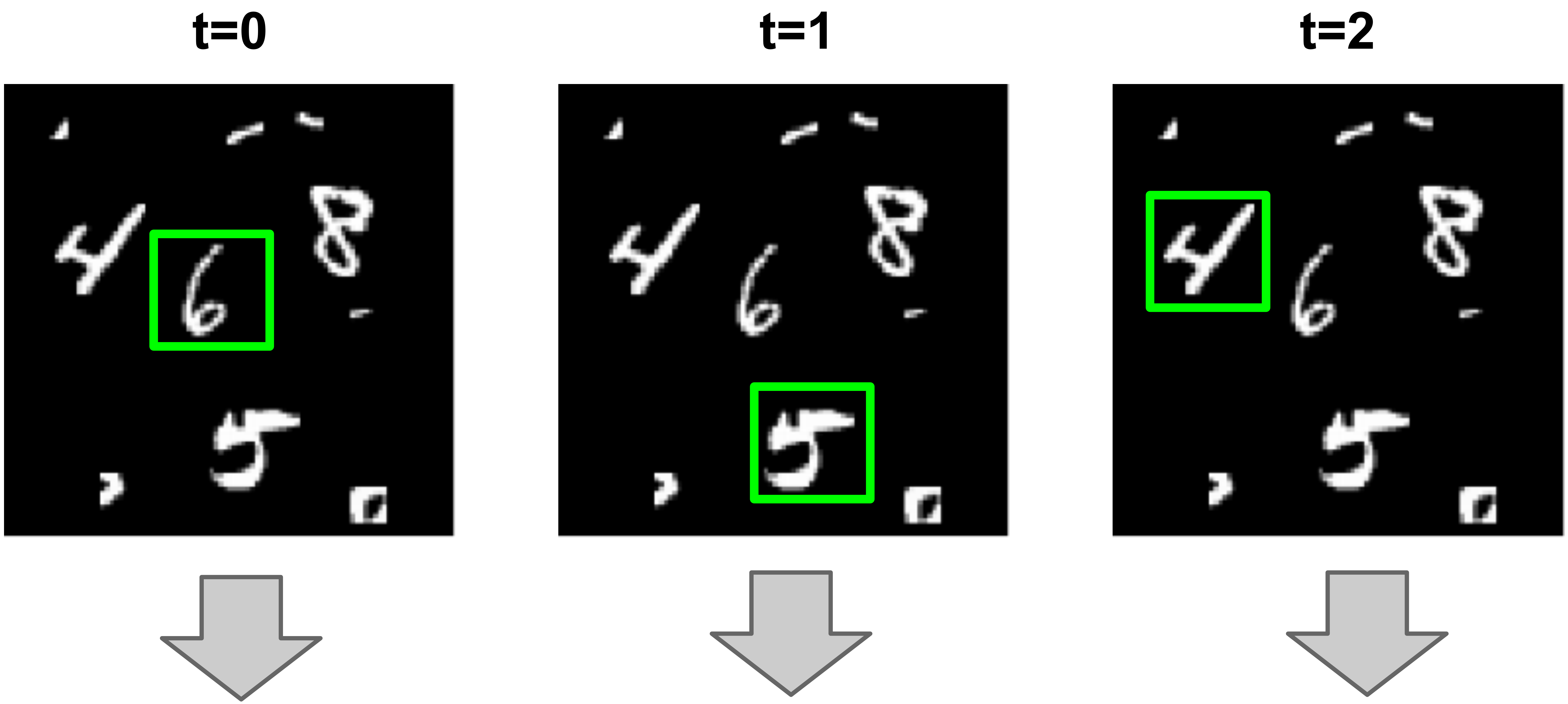
Pentti Kanerva

- binding without growing dimensionality
- fully distributed representation
- mathematical framework for storing and recovering information:
  - multiplication for binding
  - addition for combining
  - operators and inverses

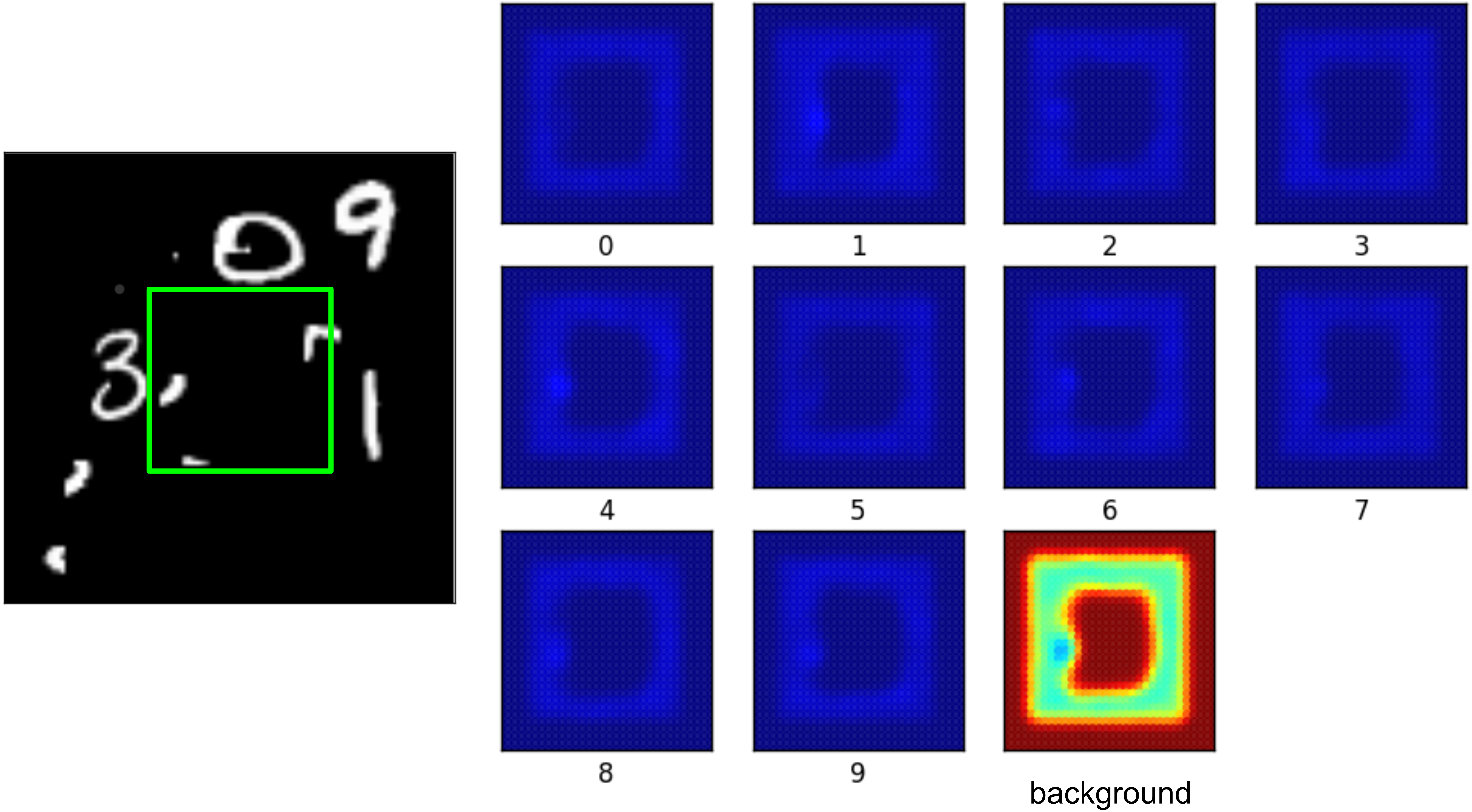
Network for binding and combining

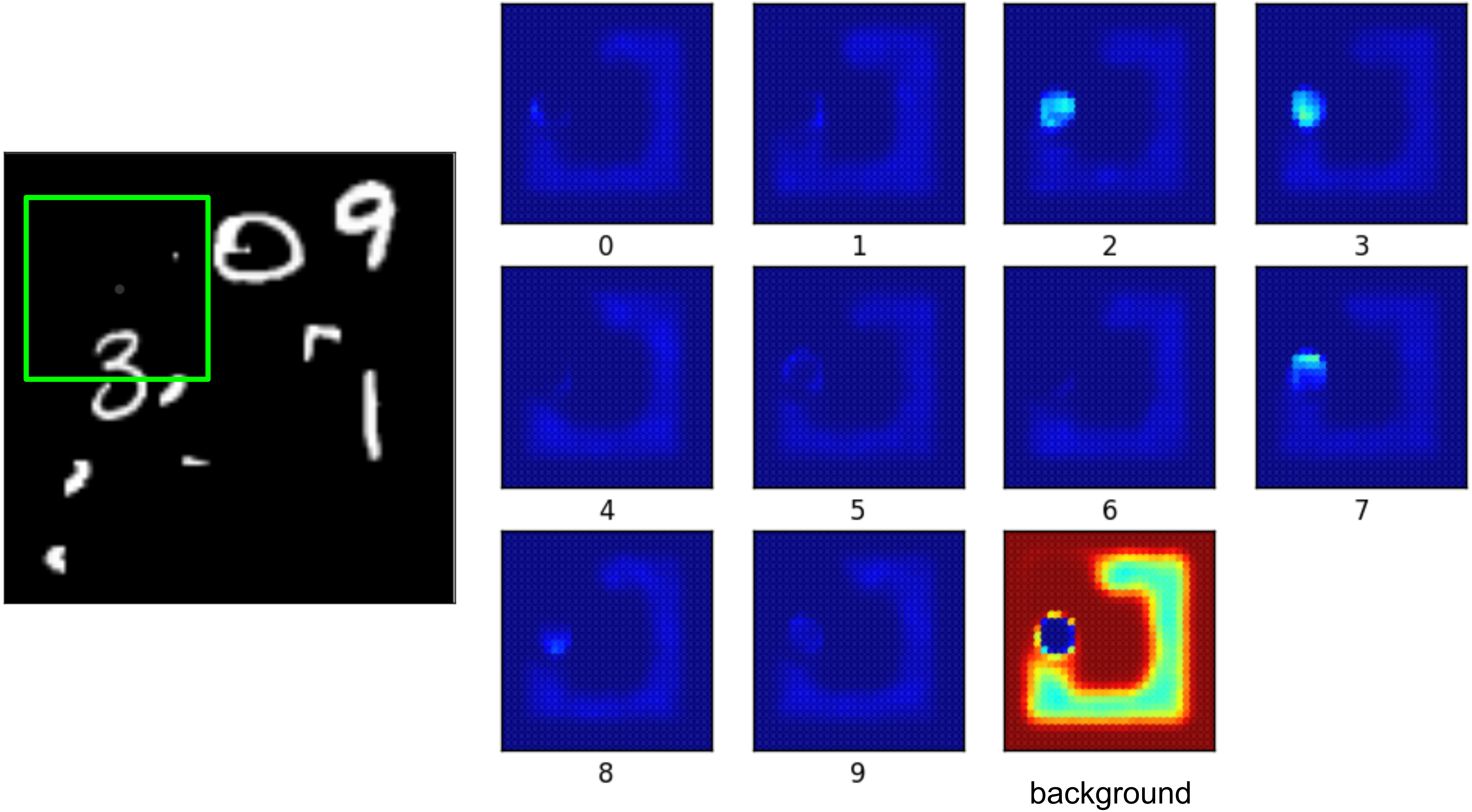


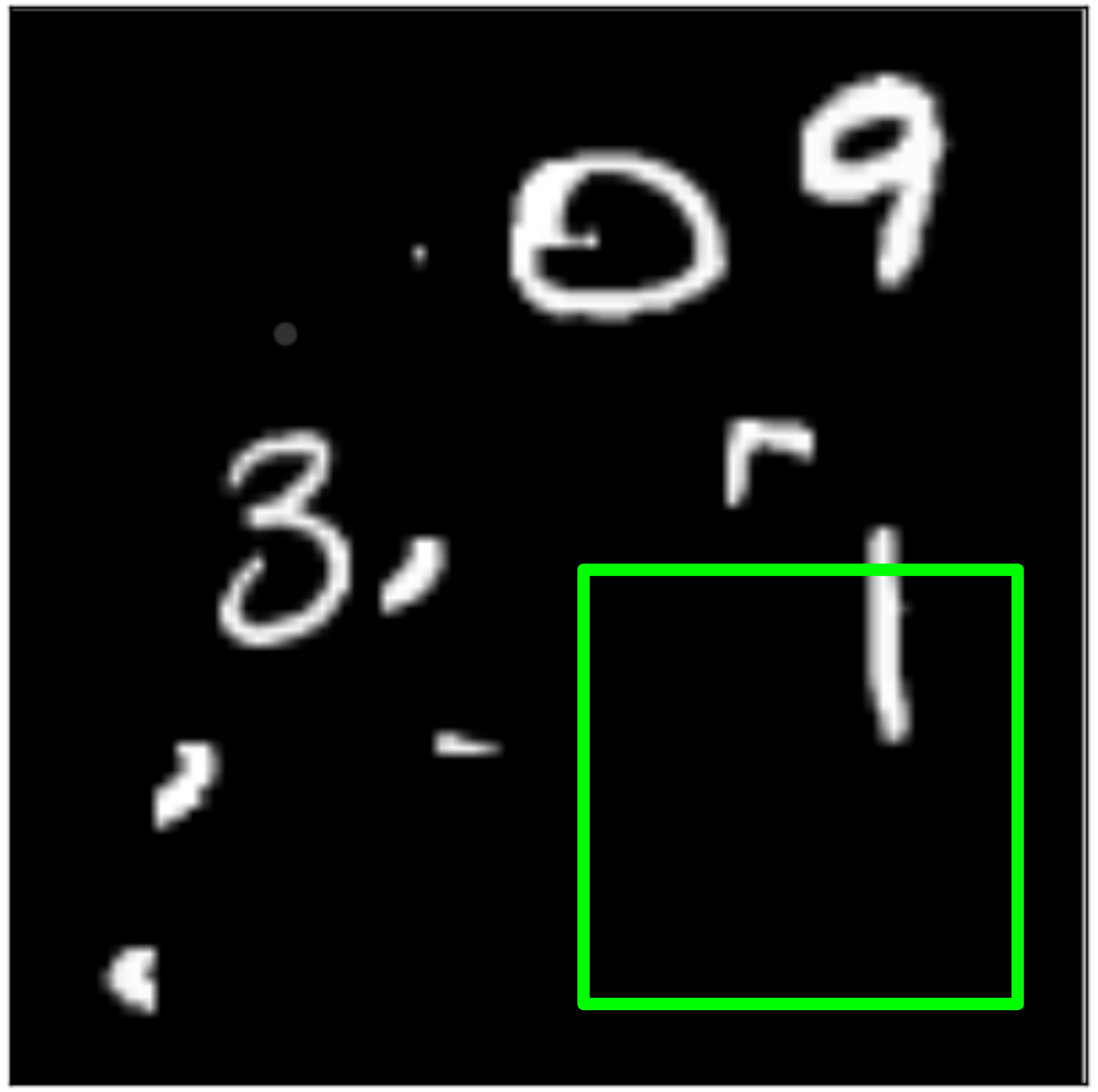
Example

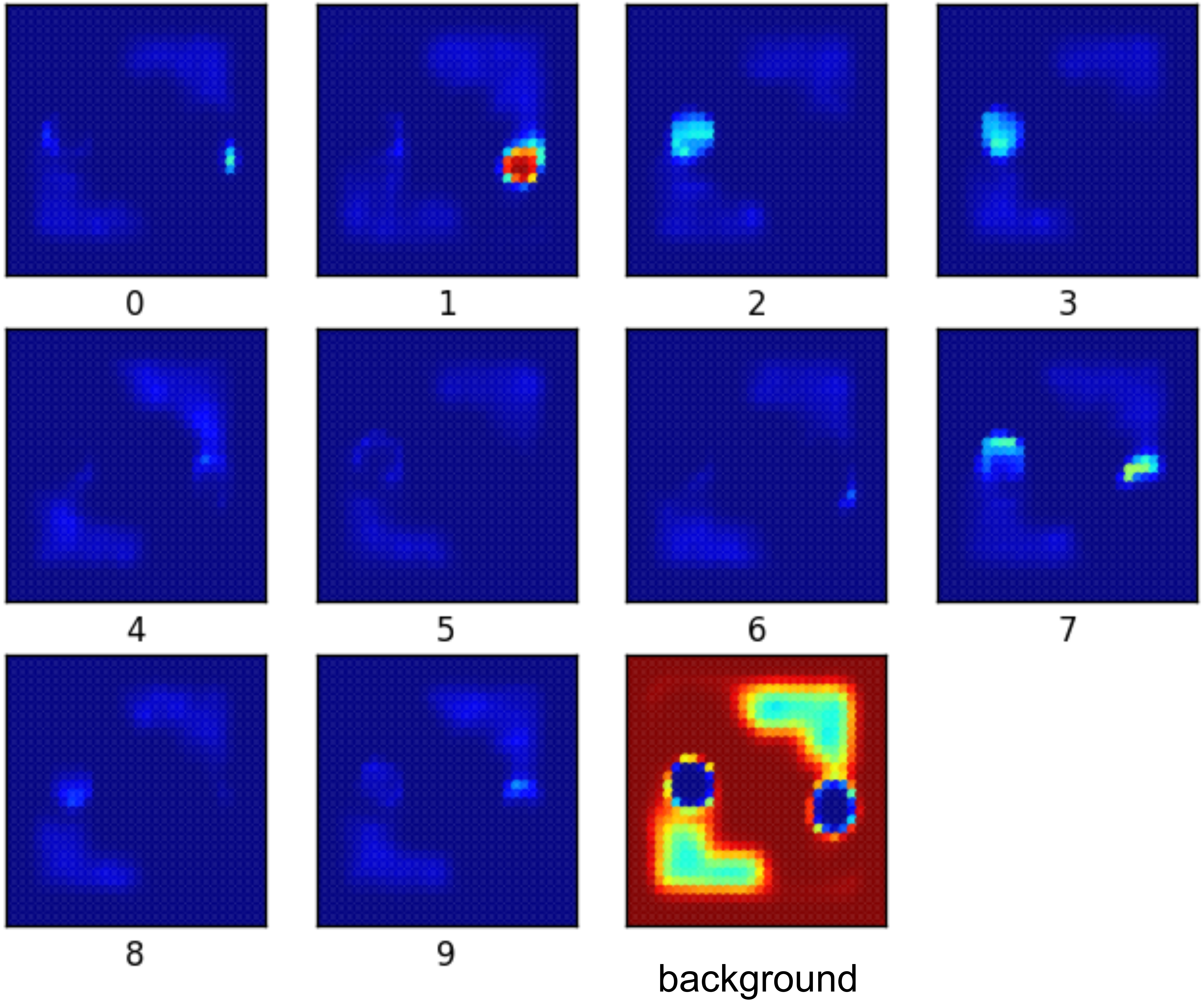


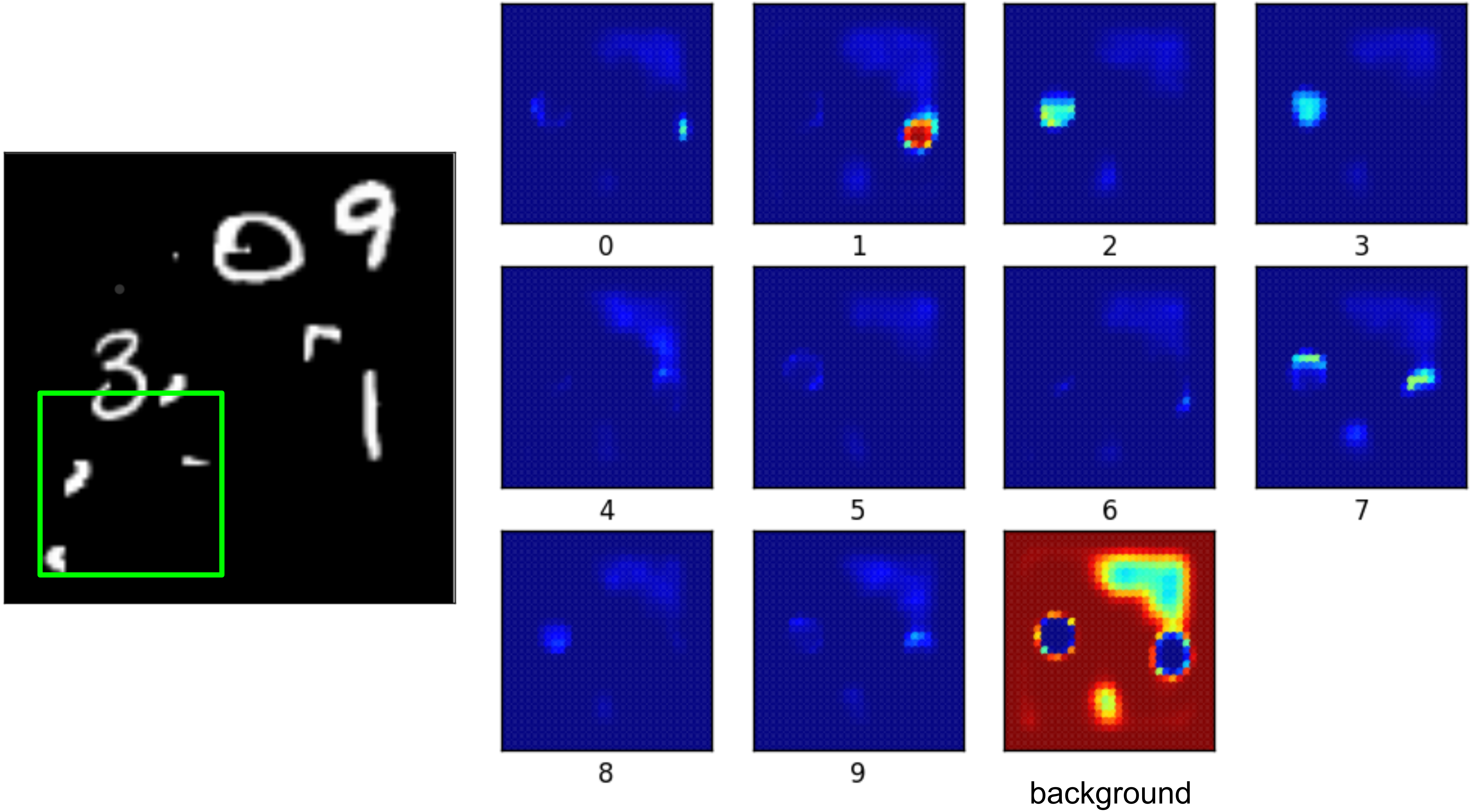
$$\mathbf{v}_6 \odot \mathbf{r}_{t=0} \quad + \quad \mathbf{v}_5 \odot \mathbf{r}_{t=1} \quad + \quad \mathbf{v}_4 \odot \mathbf{r}_{t=2} \quad = \quad \mathbf{m}$$

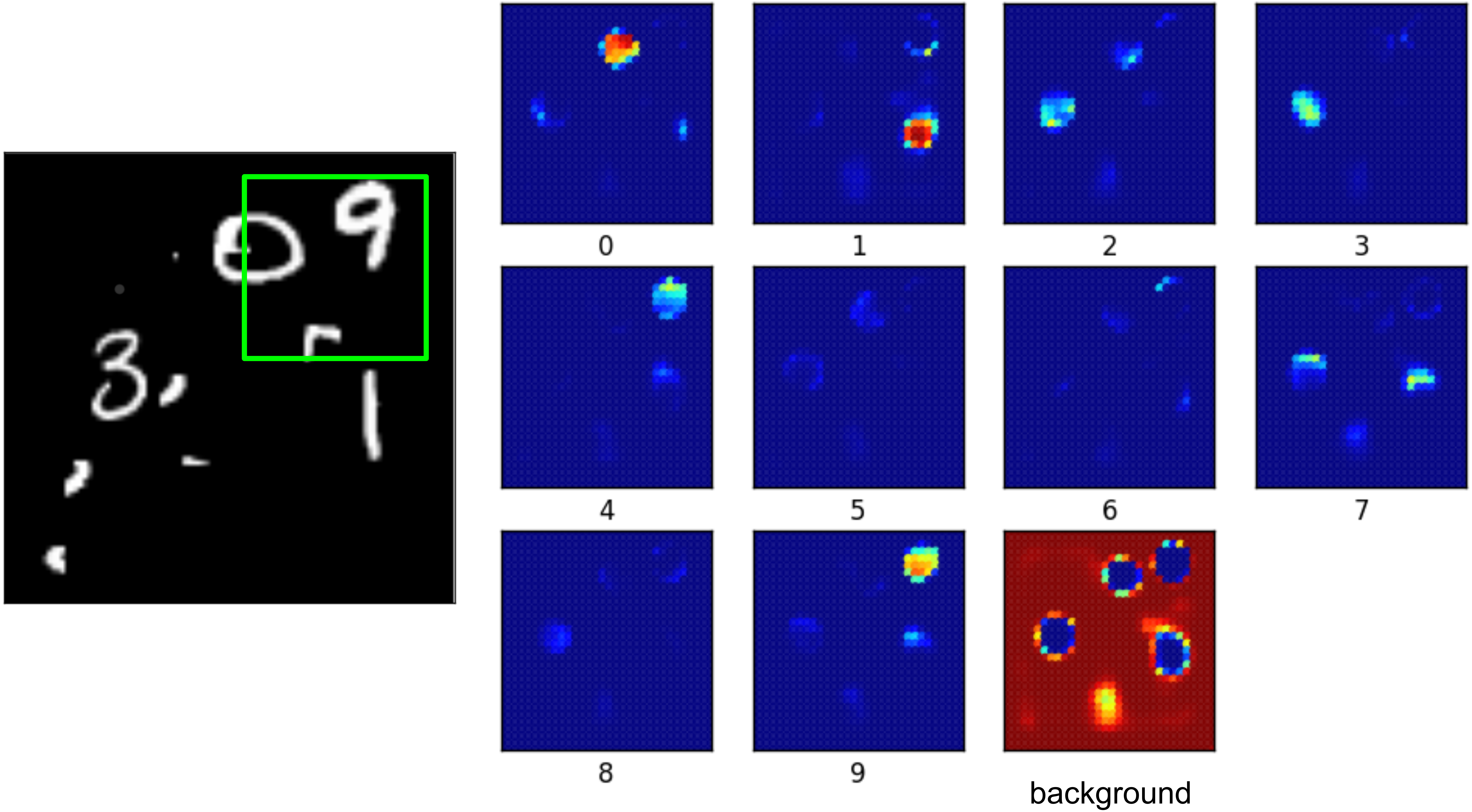

0
1
2
3
4
5
6
7
8
9
background

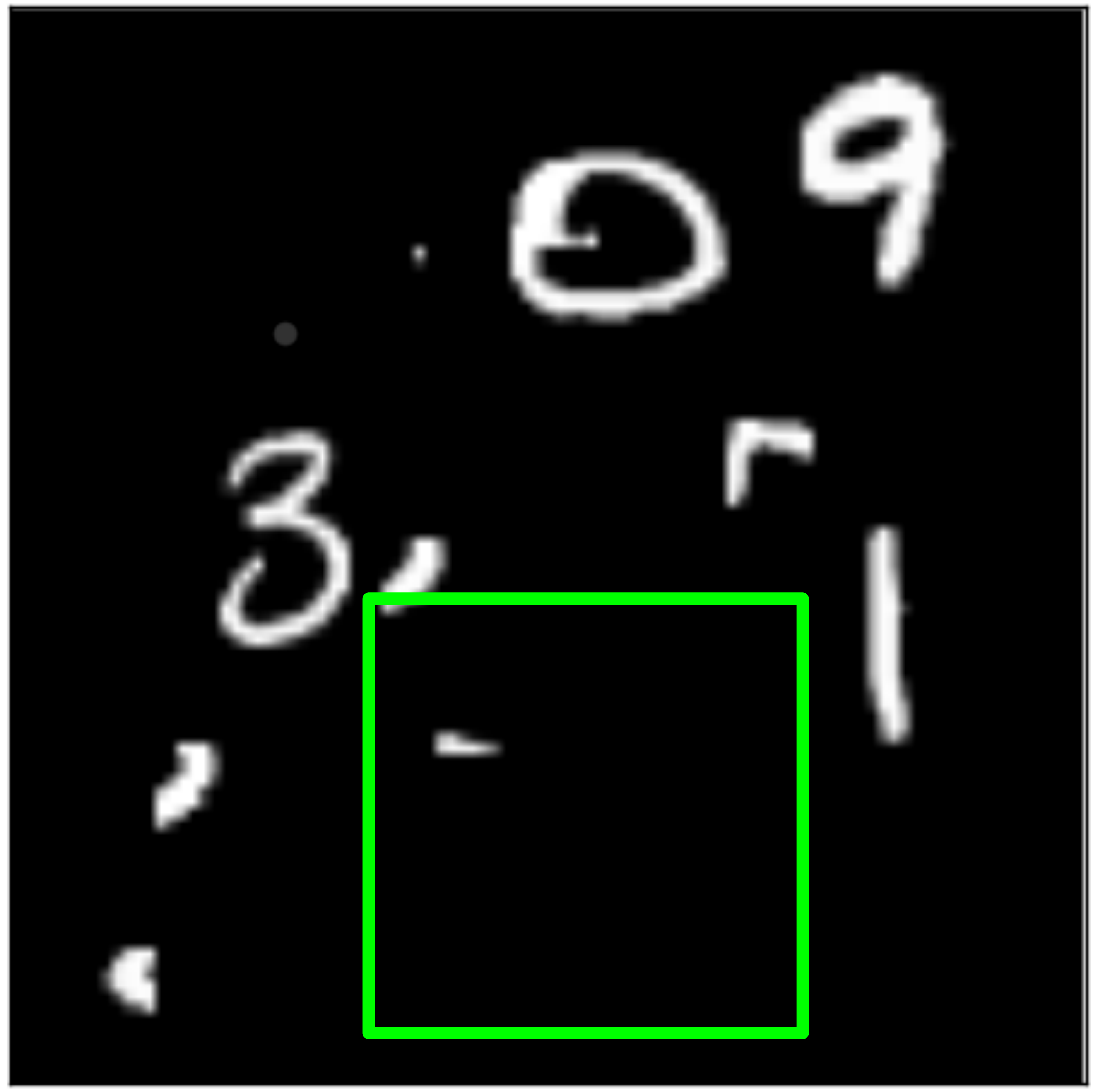

0
1
2
3
4
5
6
7
8
9
background

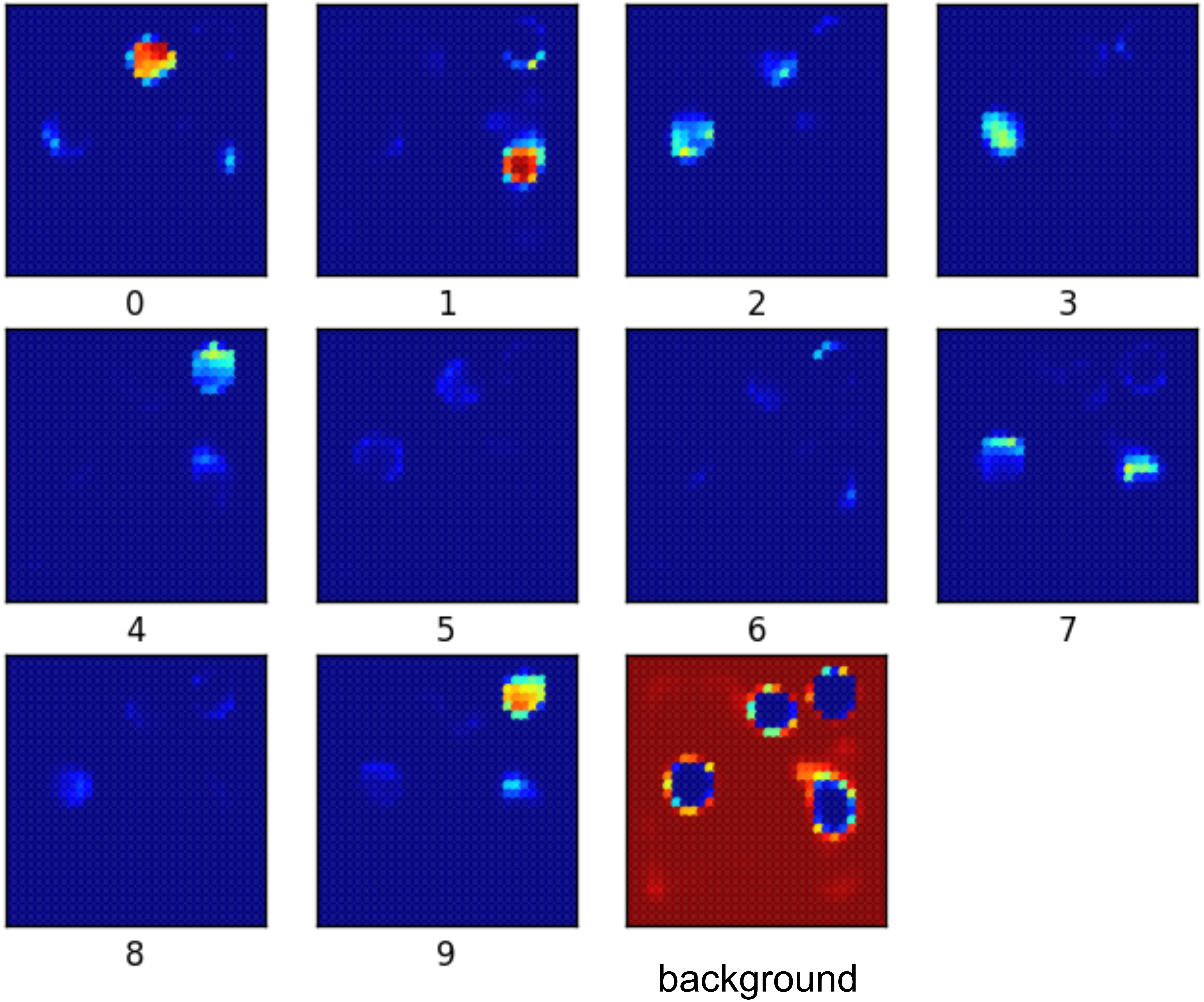


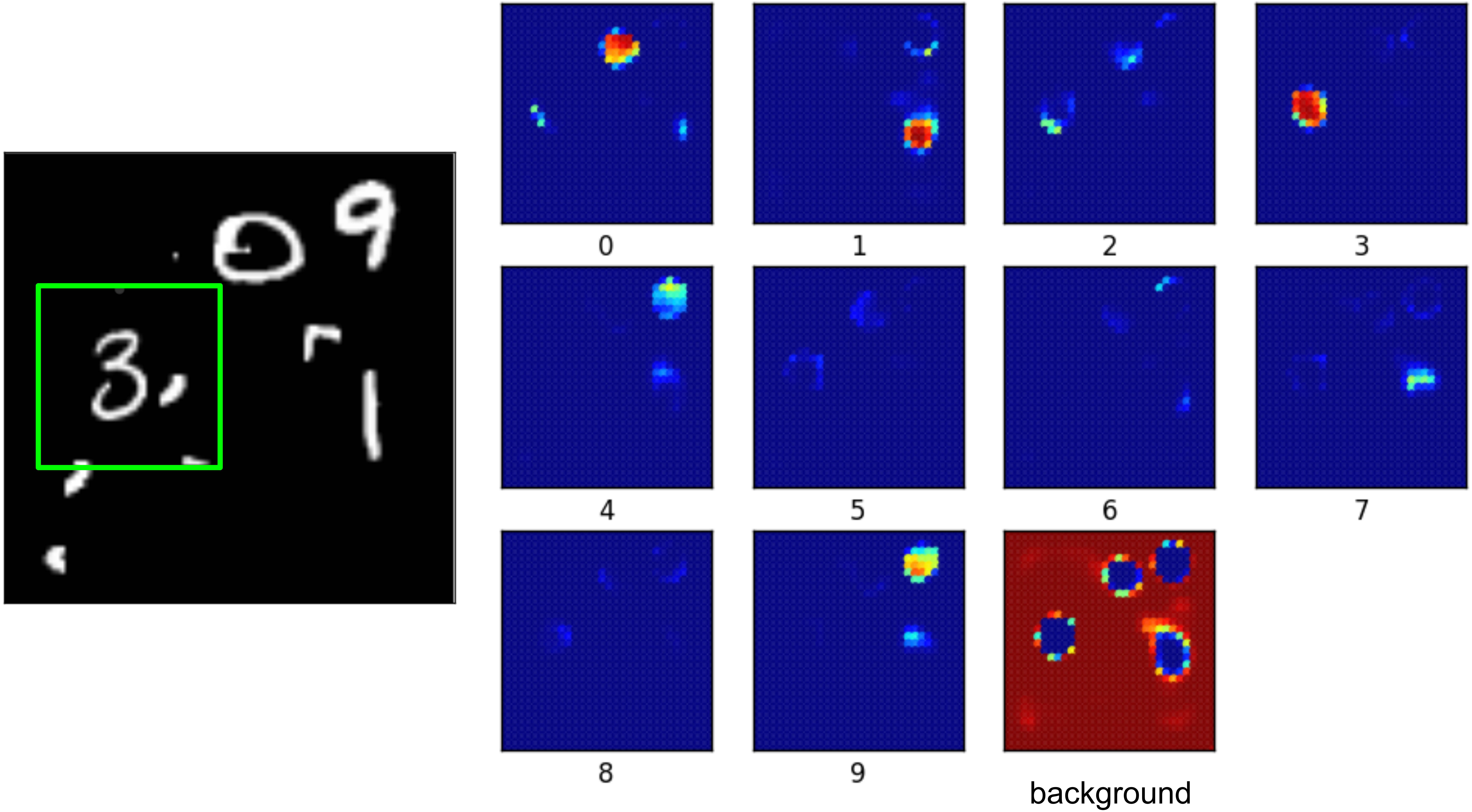

0
1
2
3
4
5
6
7
8
9
background

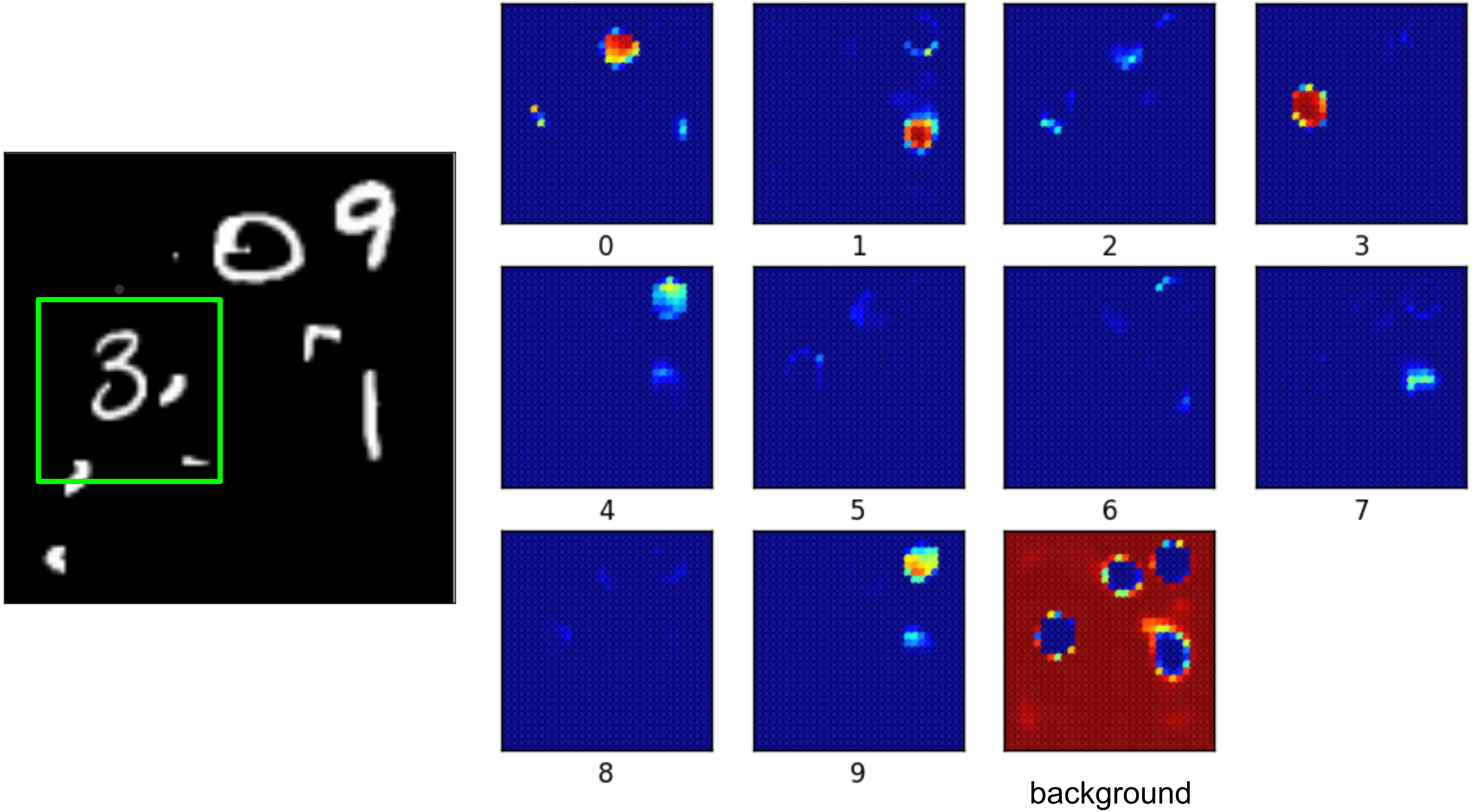

0
1
2
3
4
5
6
7
8
9
background

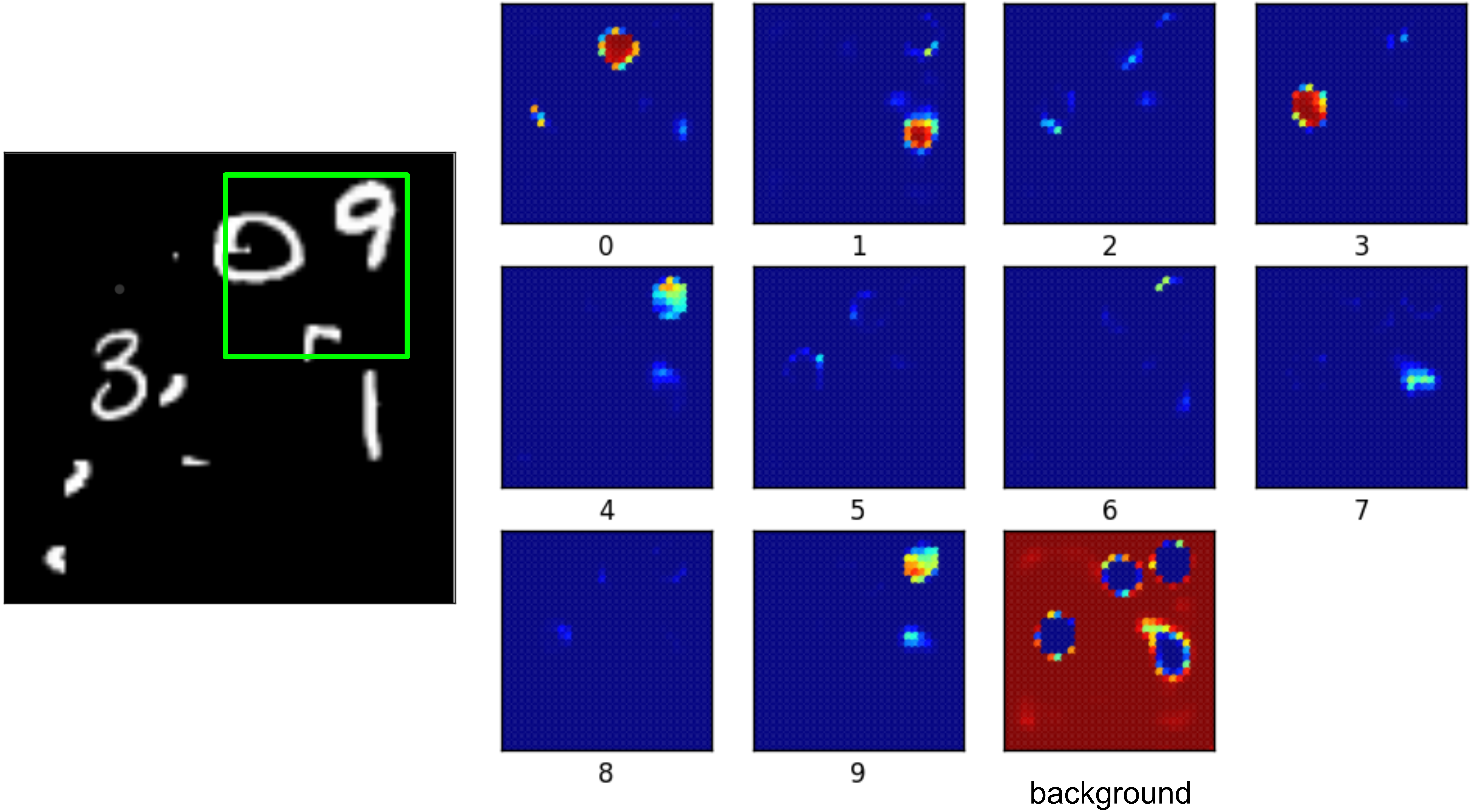

0
1
2
3
4
5
6
7
8
9
background

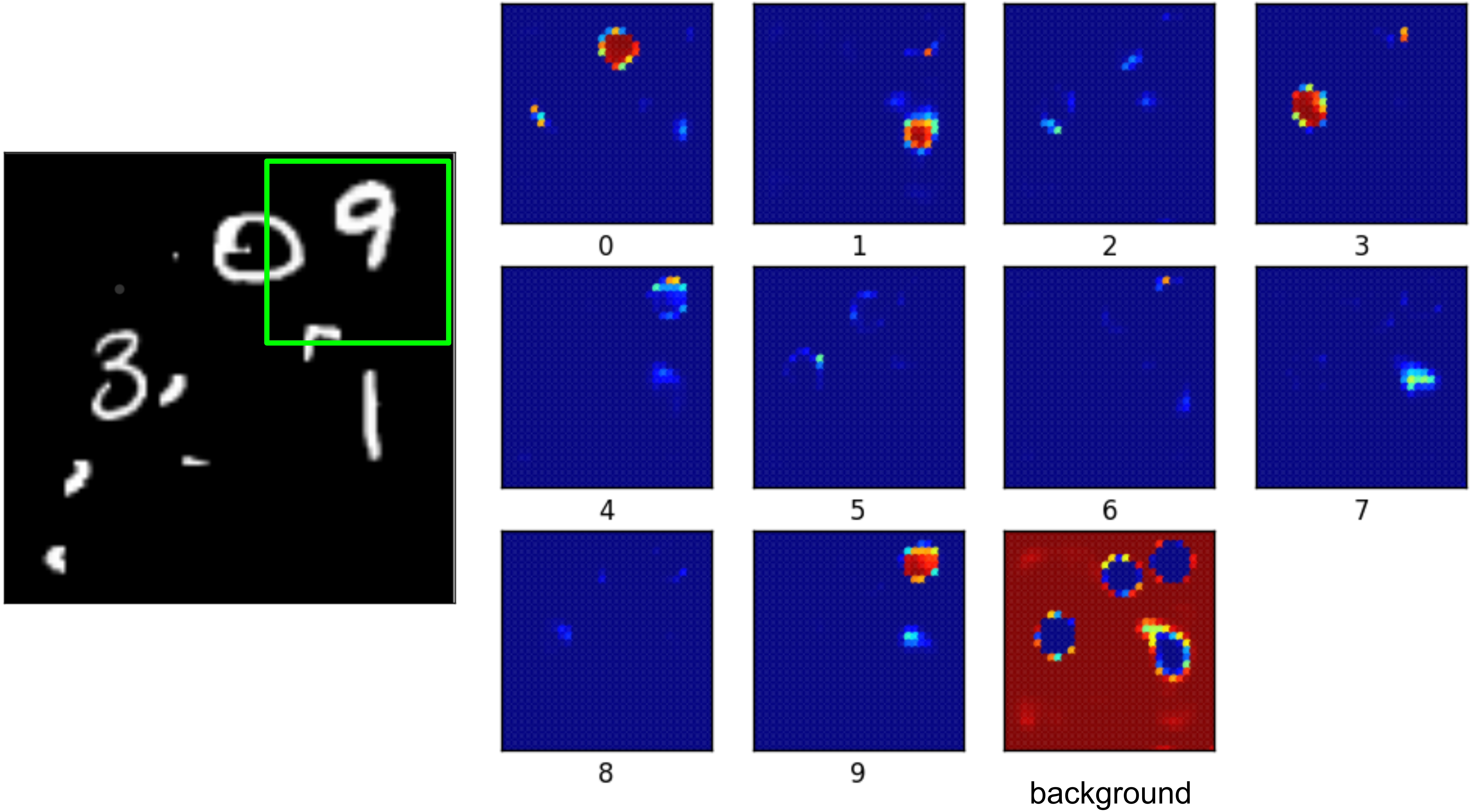



0
1
2
3
4
5
6
7
8
9
background


0
1
2
3
4
5
6
7
8
9
background

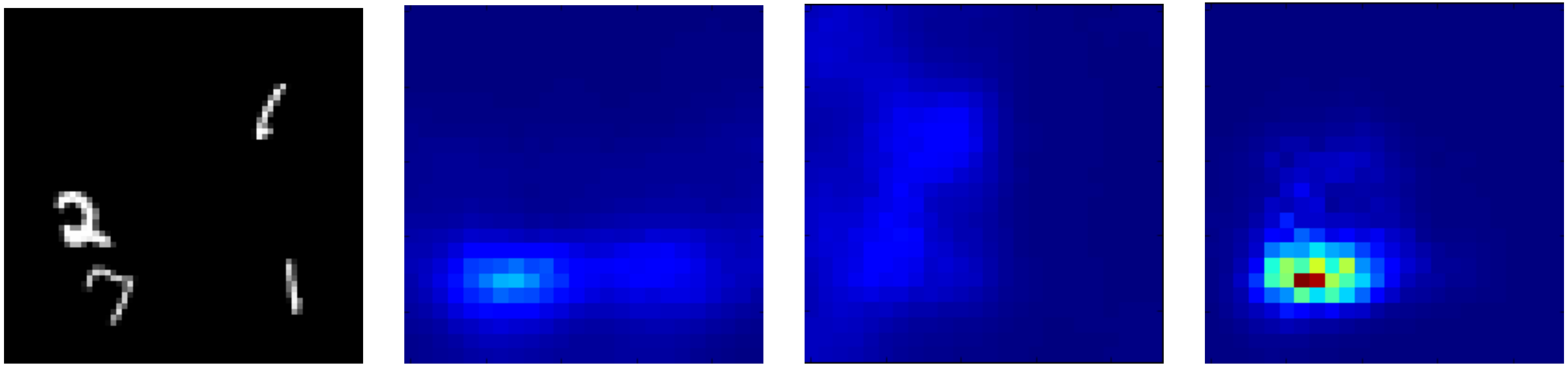

0
1
2
3
4
5
6
7
8
9
background


0
1
2
3
4
5
6
7
8
9
background


0
1
2
3
4
5
6
7
8
9
background

Spatial reasoning

What is below a ‘2’ and to the left of a ‘1’?



(a) Example image (b) ”below a 2” (c) ”to the left of a 1” (d) Combined

Main points

- Visual scenes require the ability to represent *compositional* structure.

- *Active* sensing strategies, such as eye movements, allows us to acquire information and build a scene representation with limited neural resources.

- A *foveated* image sampling lattice similar to the primate retina emerges as the optimal solution for visual search, but only for an eye without the ability to zoom.

- Neural networks with the ability to *bind* and *combine* information across saccades are capable of building up a scene representation that supports spatial reasoning.