Comparing deep neural networks against humans: Object recognition with weak signals

Felix Wichmann

Neural Information Processing Group and
Bernstein Center for Computational Neuroscience,
Eberhard Karls Universität Tübingen

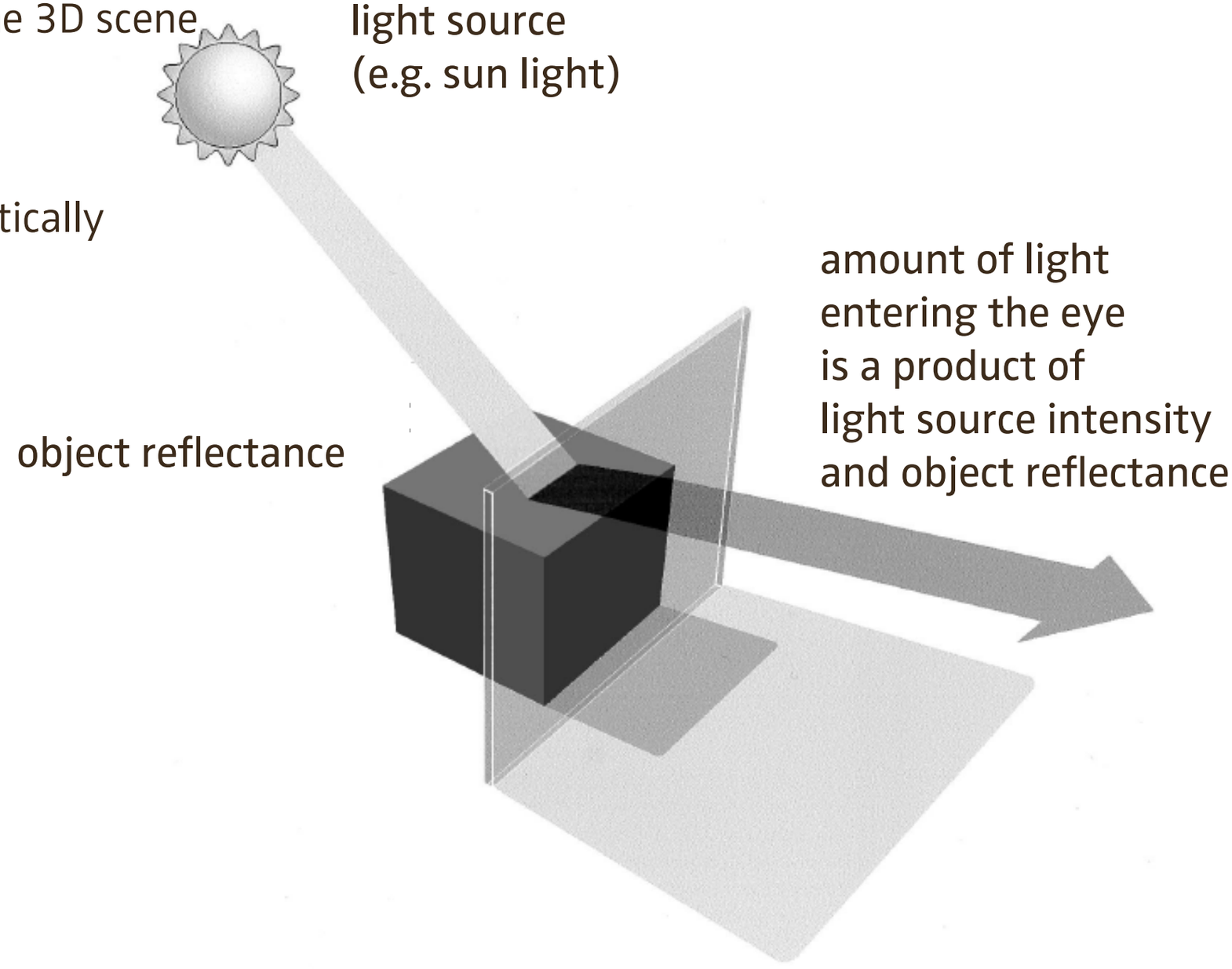
Max Planck Institute for Intelligent Systems, Tübingen

One way to think about vision: inverse optics

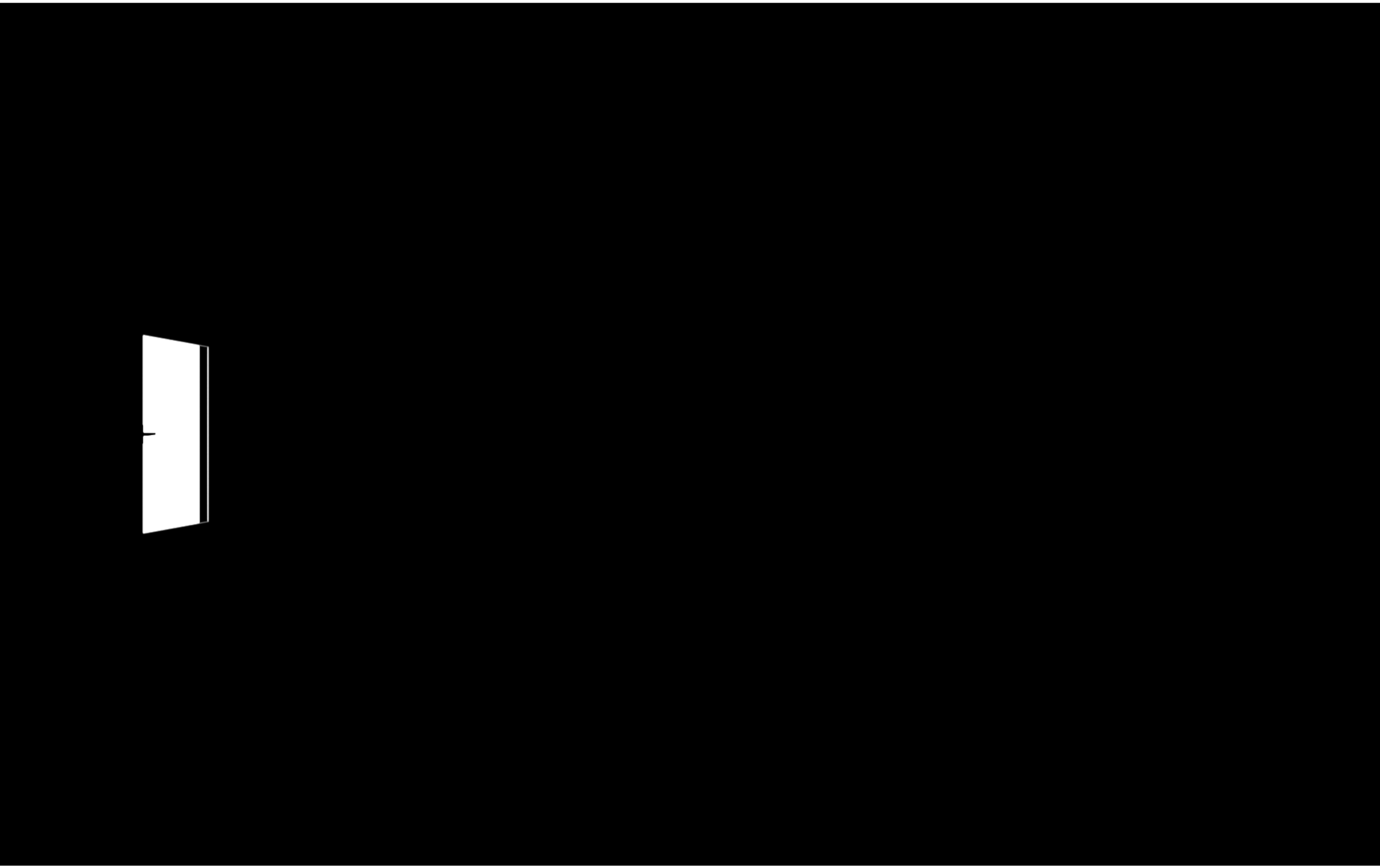
Laws of physics "generate" 2D images on our retinae from 3D scenes (forward optics / rendering)

Starting point to think about visual perception: we want to infer the 3D scene from the 2D retinal images: inverse optics!

But: Inverse optics is mathematically impossible.



N = o



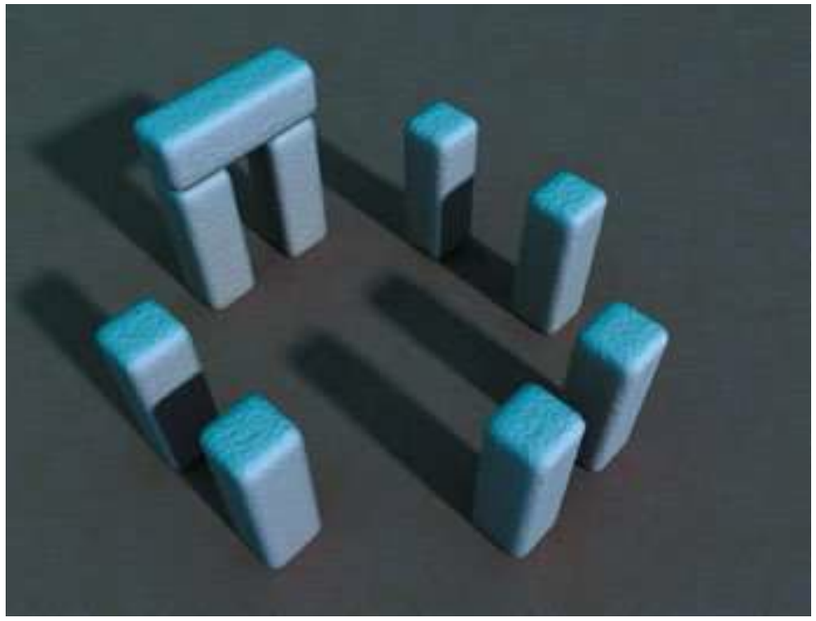
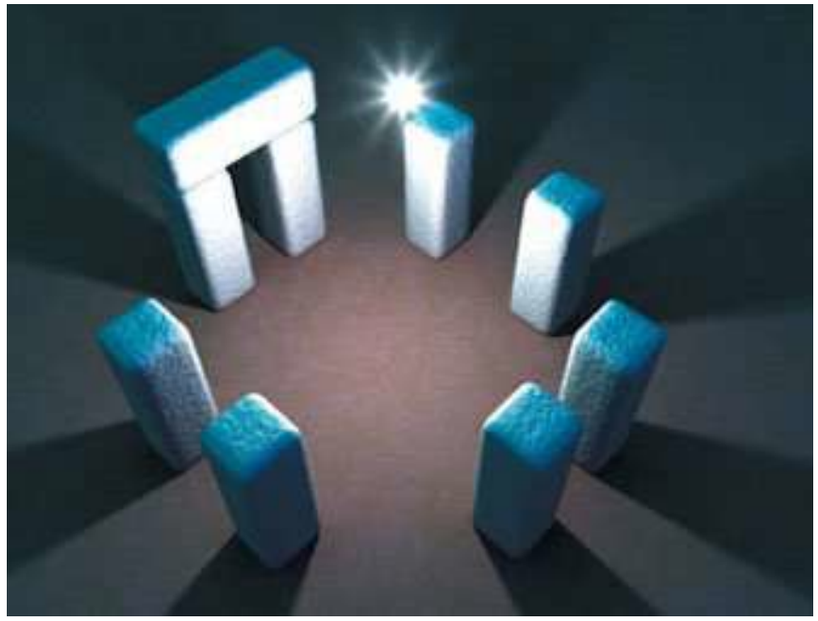
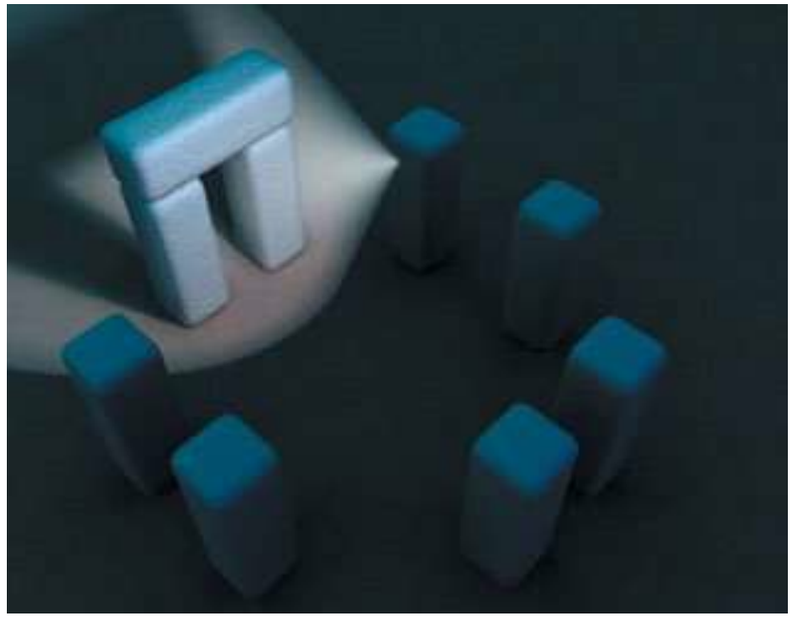
N = 1

N = 2

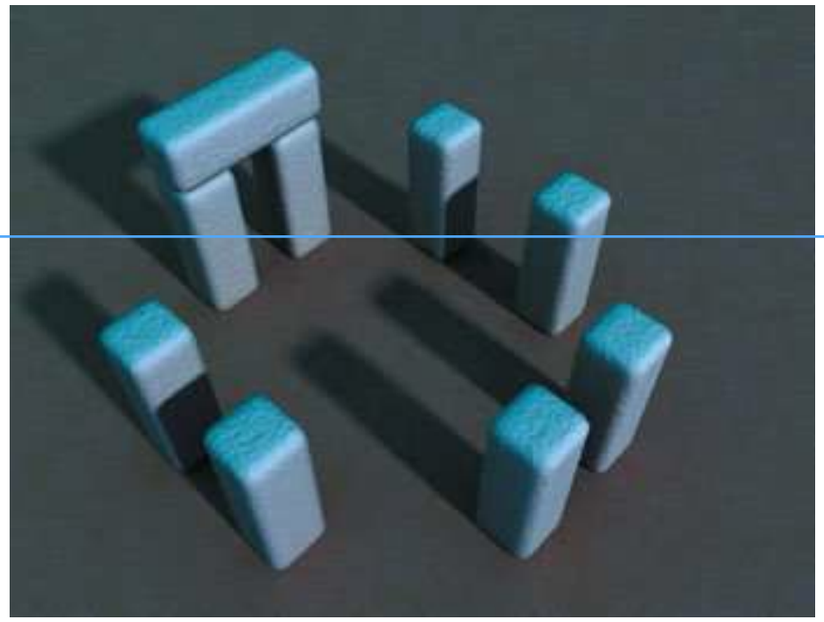
N = 5

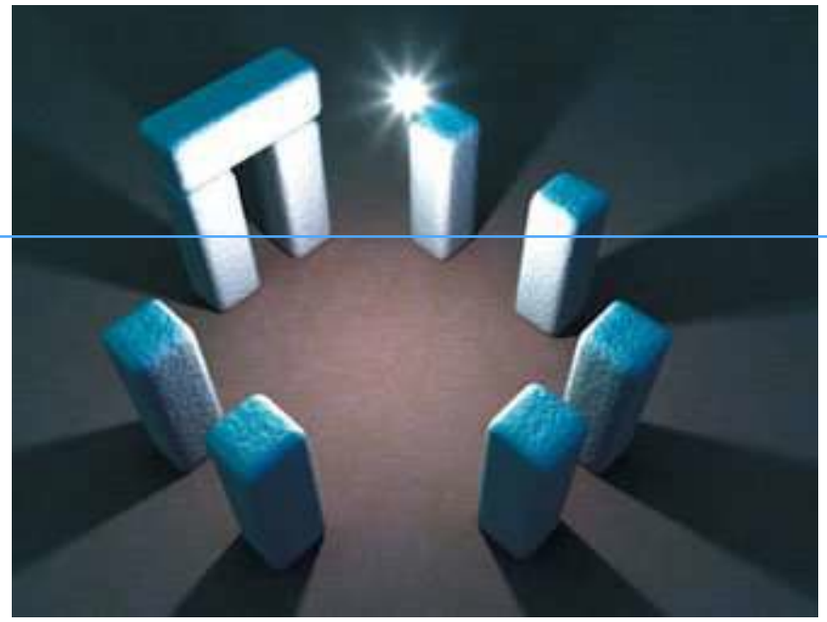
N = 9

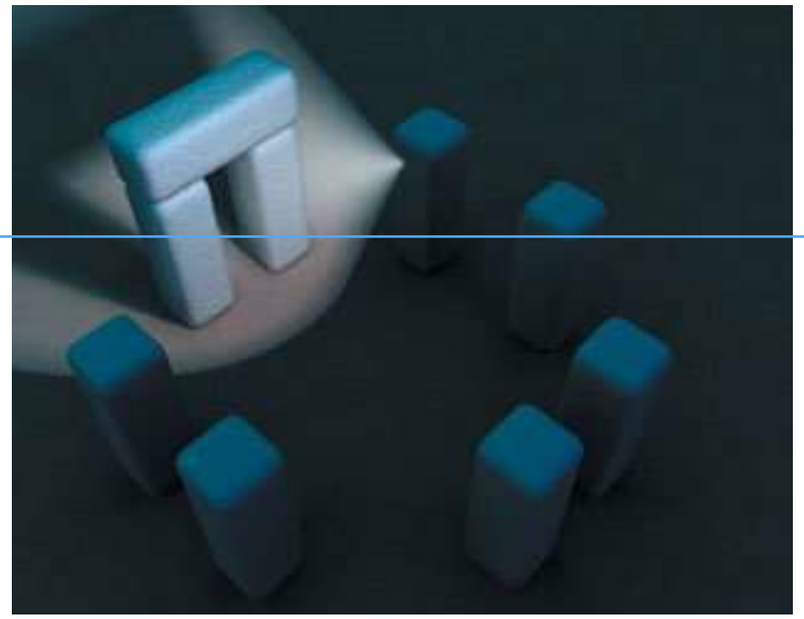
N = 24 (considered fully rendered)

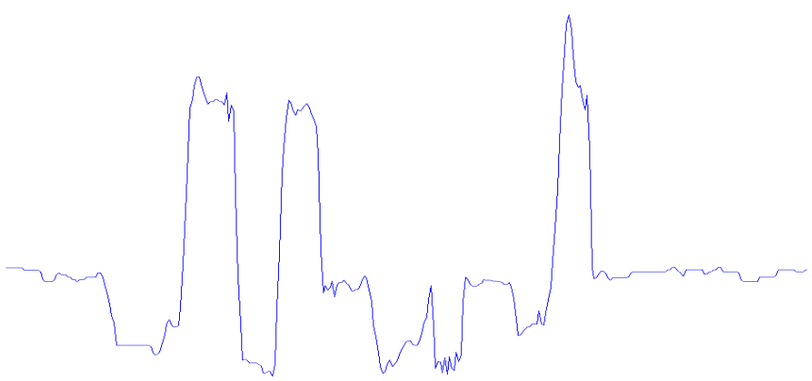




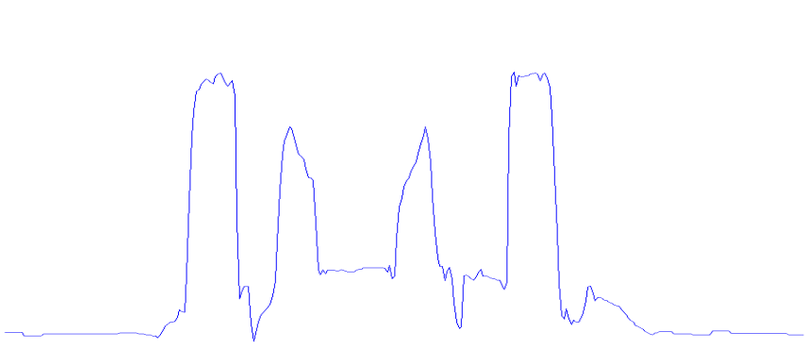
modified from Matthias Bethge

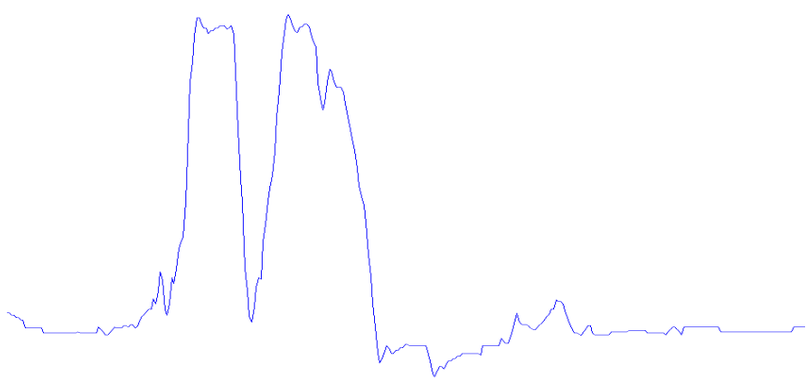




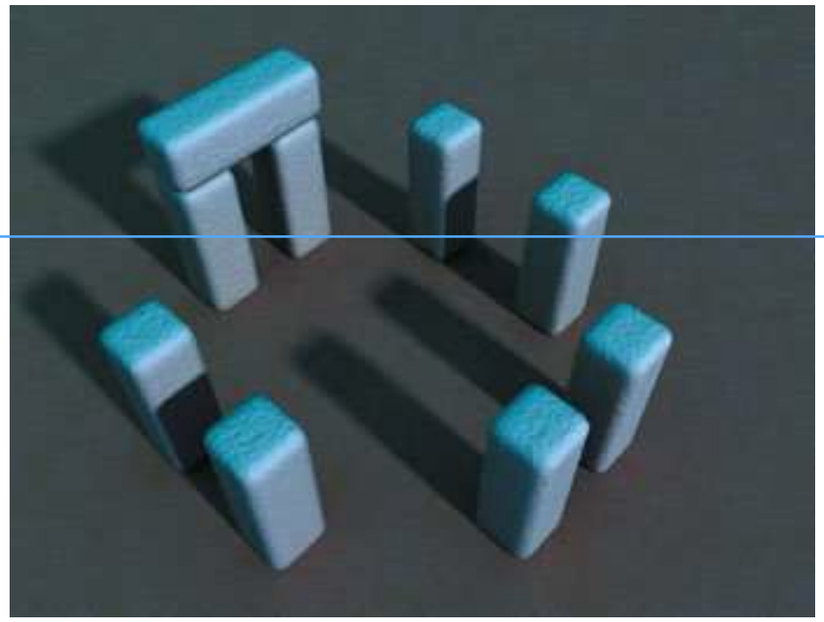


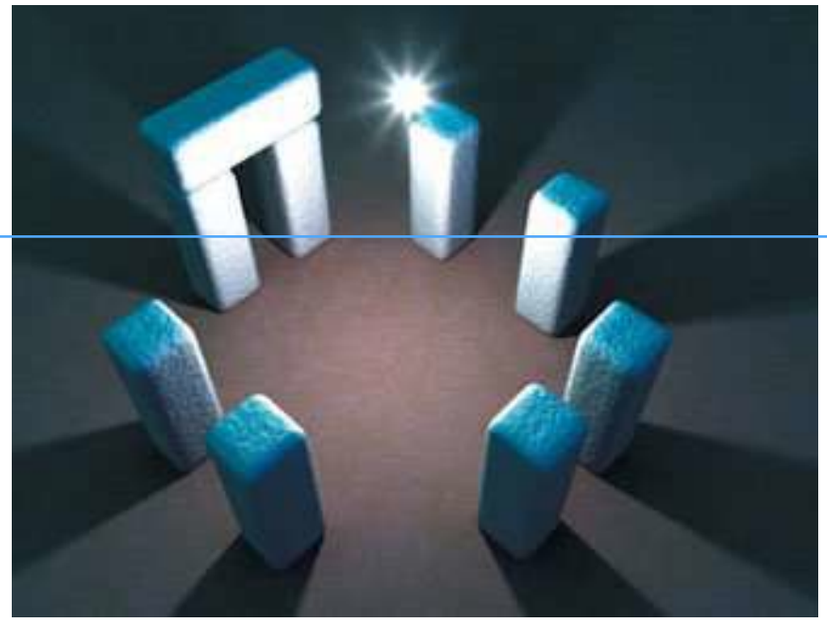


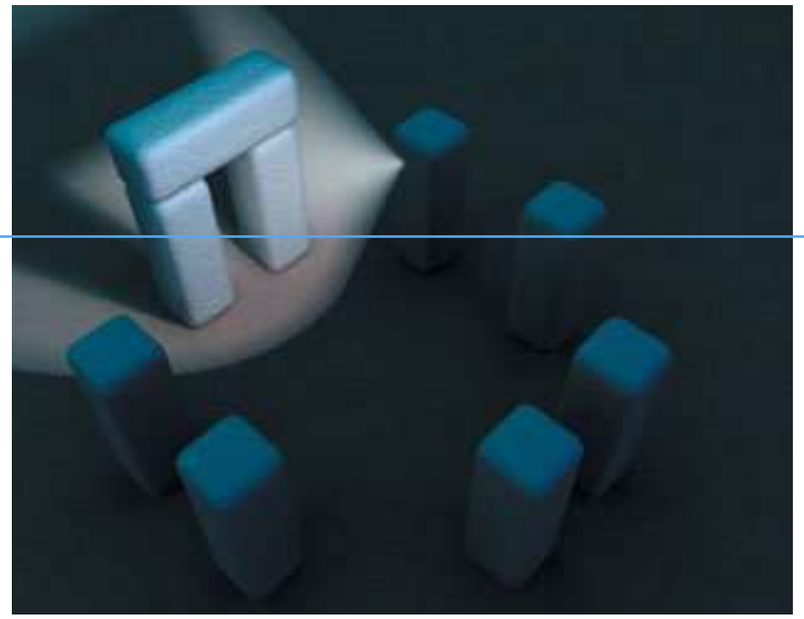


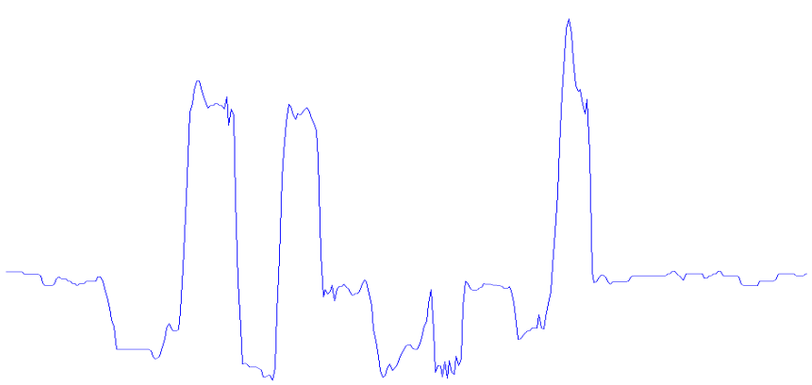


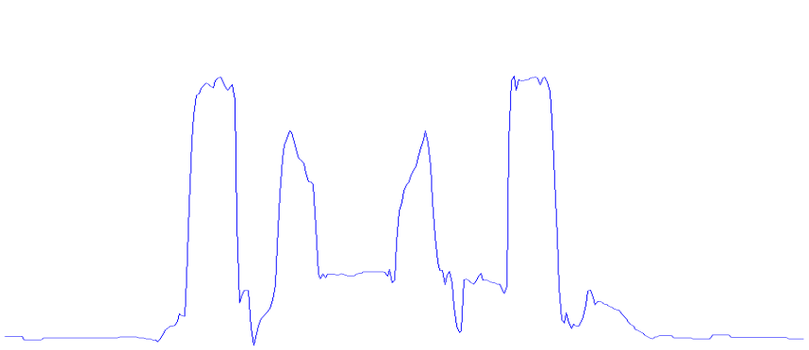
modified from Matthias Bethge

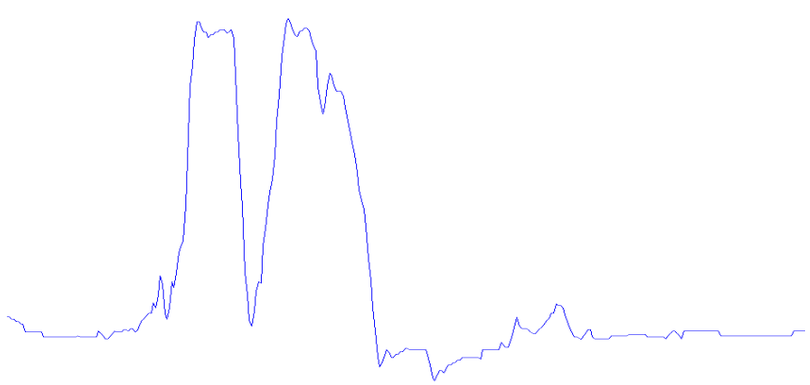






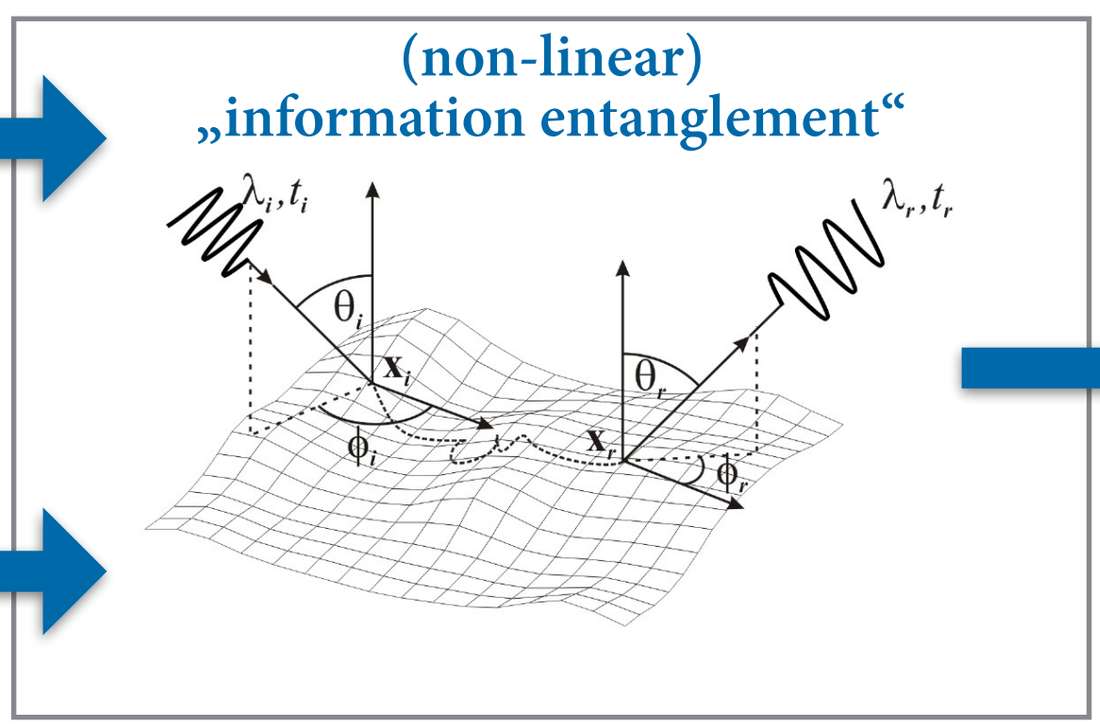






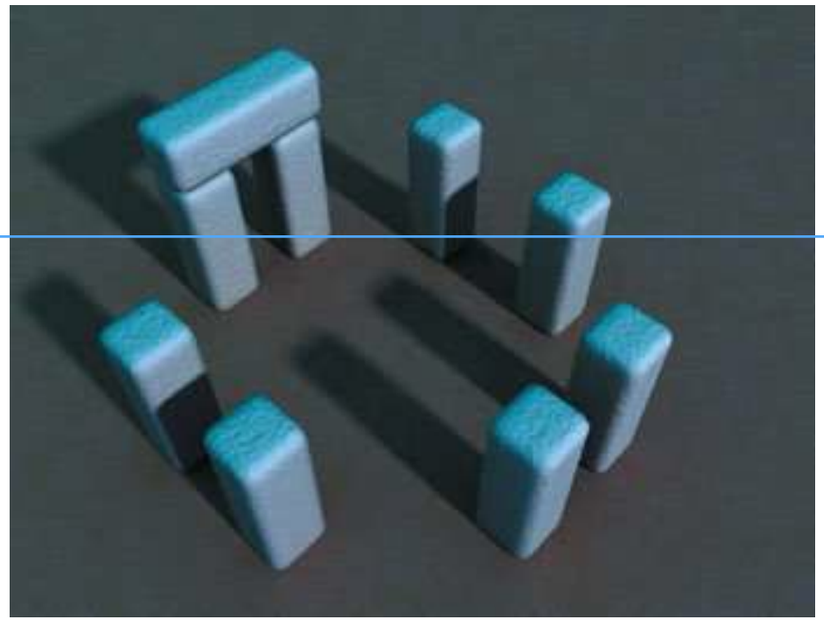
illumination
(„light field“)

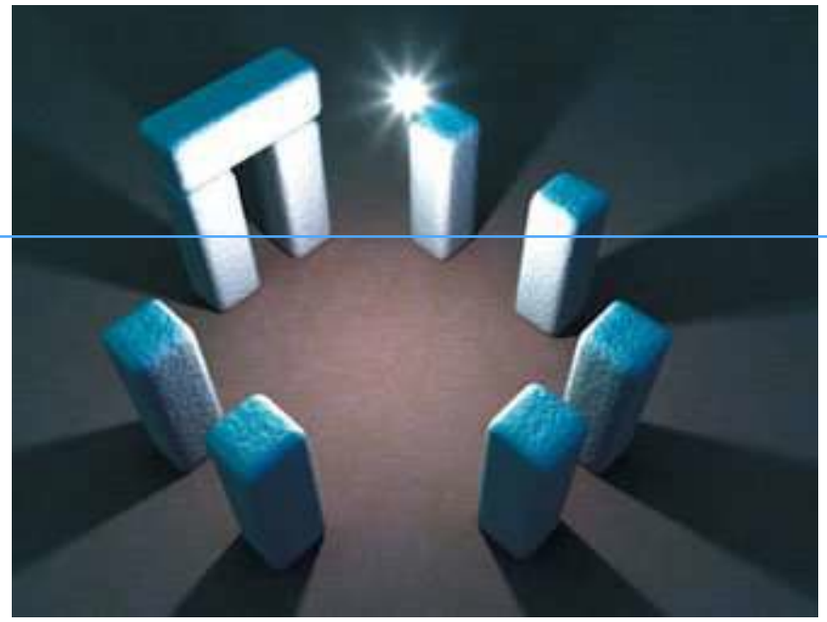
objects & surfaces
(geometry, materials)

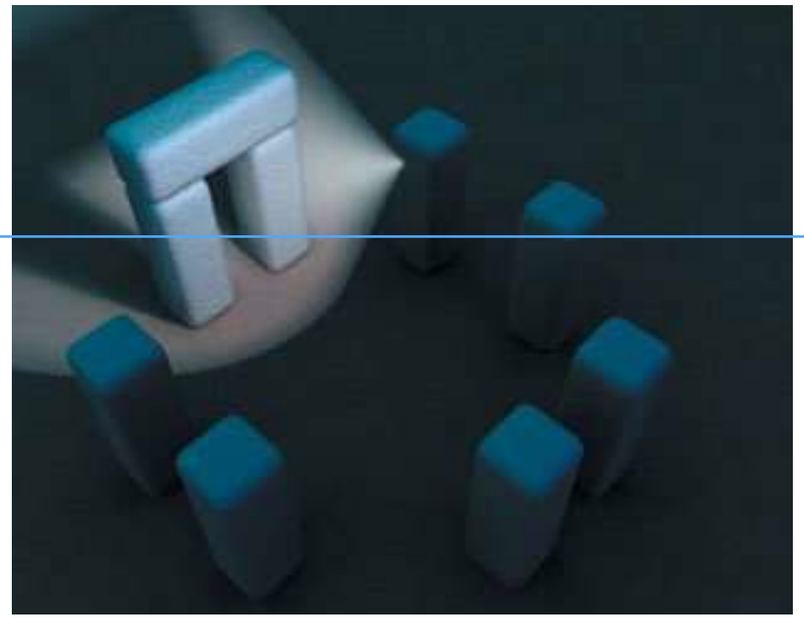


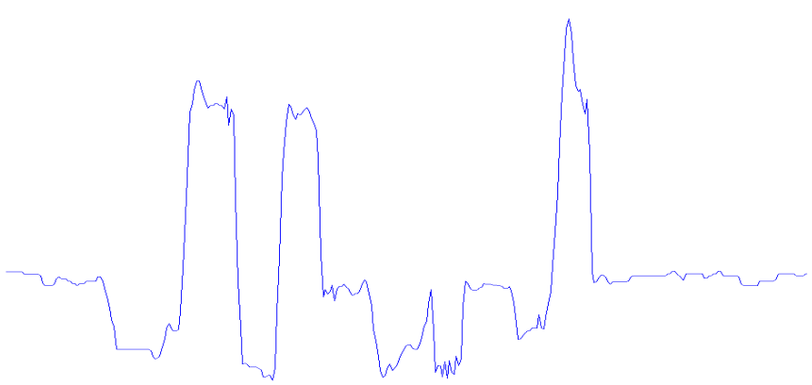
resulting
image

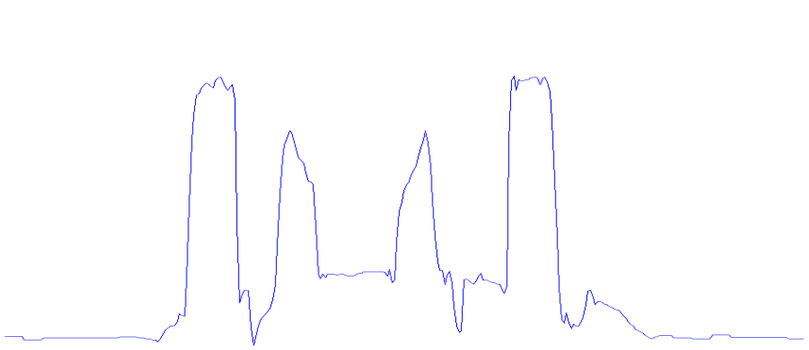
modified from Matthias Bethge

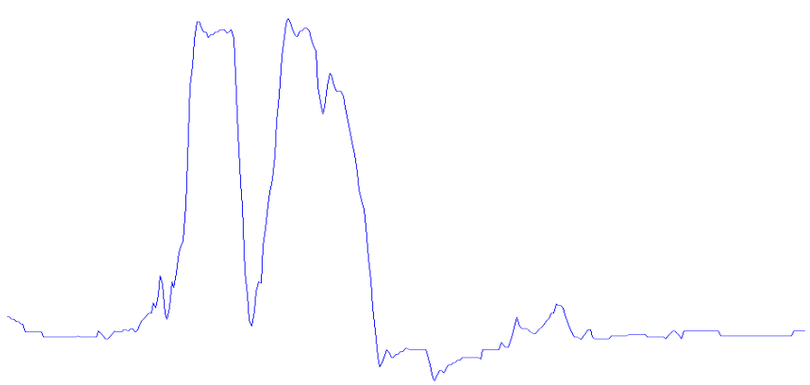








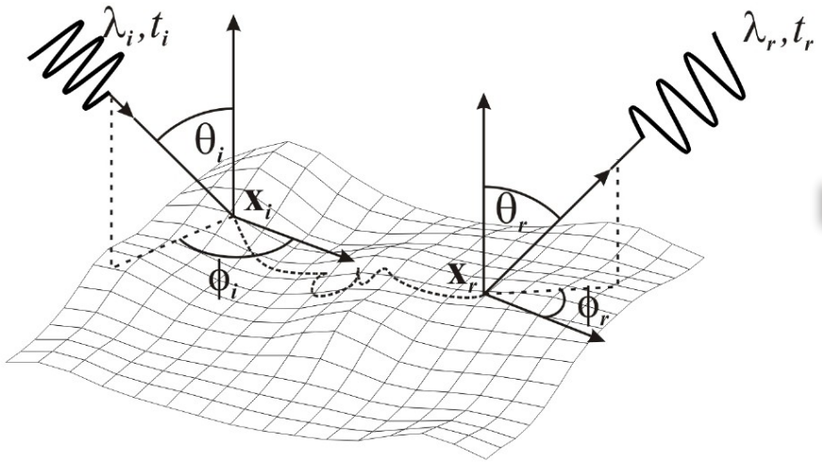




illumination
(„light field“)

objects & surfaces
(geometry, materials)

(non-linear)
„information entanglement“



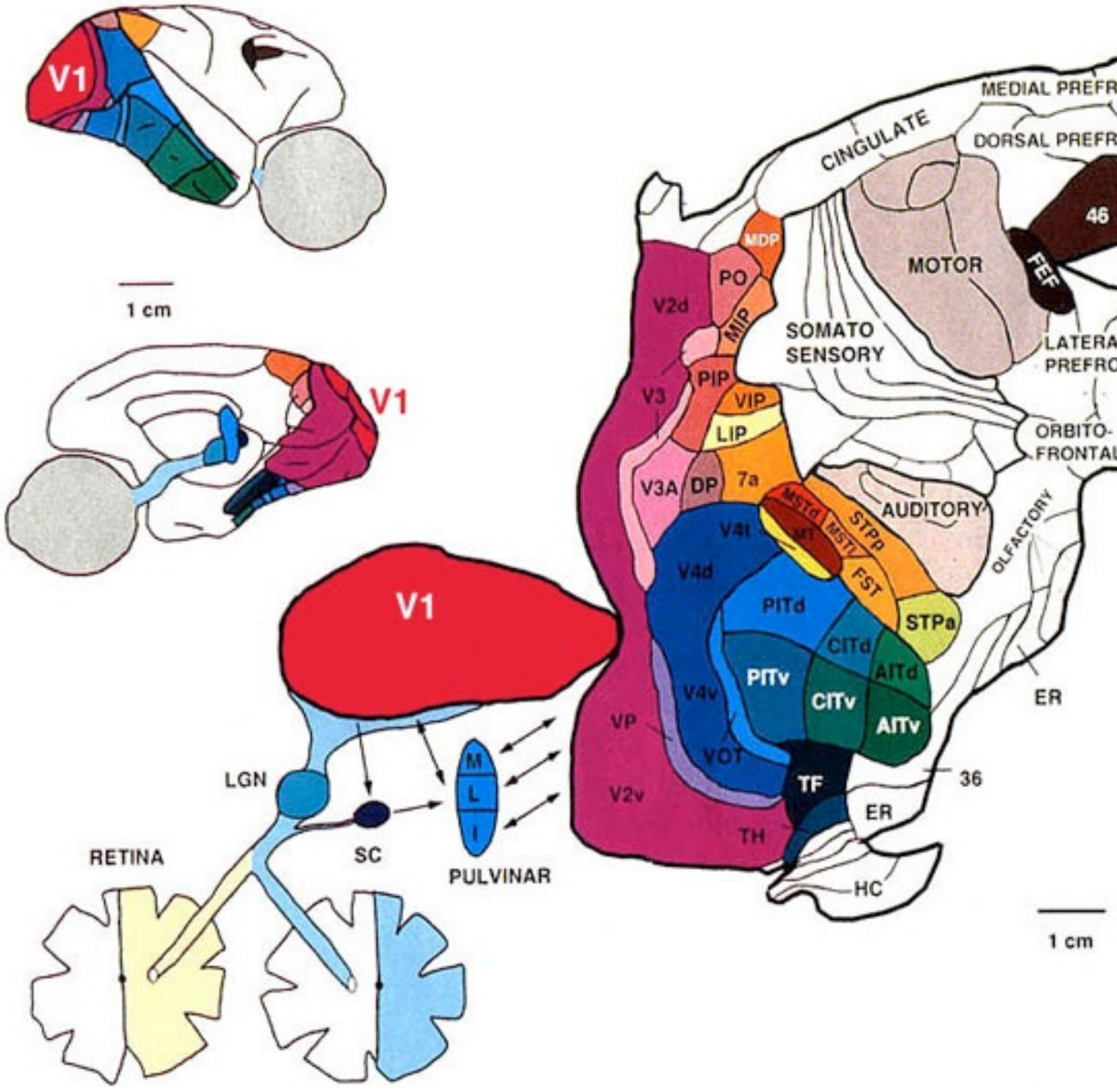
resulting
image

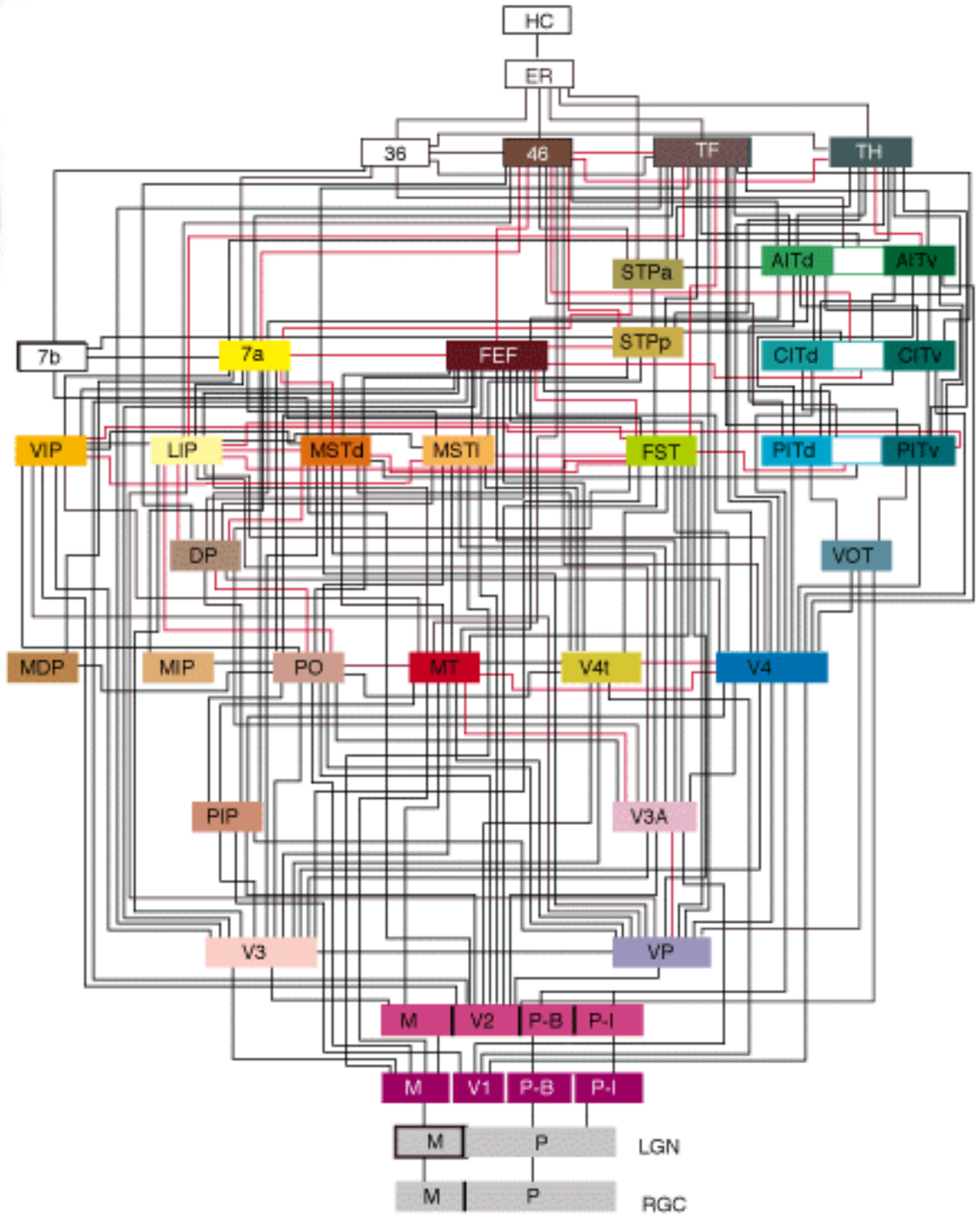
visual inference („untangling“)

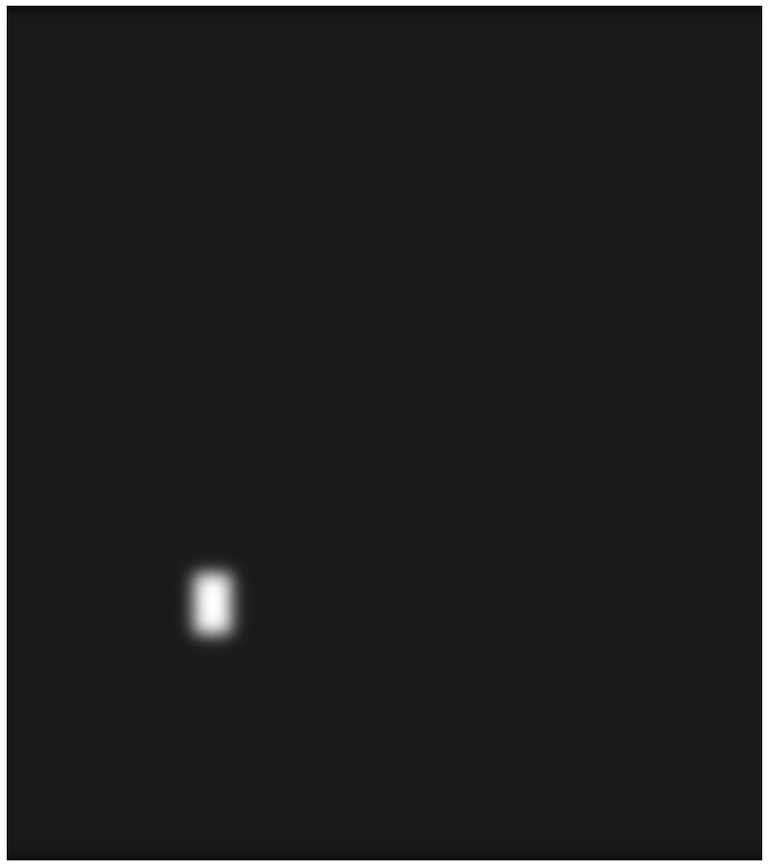
modified from Matthias Bethge

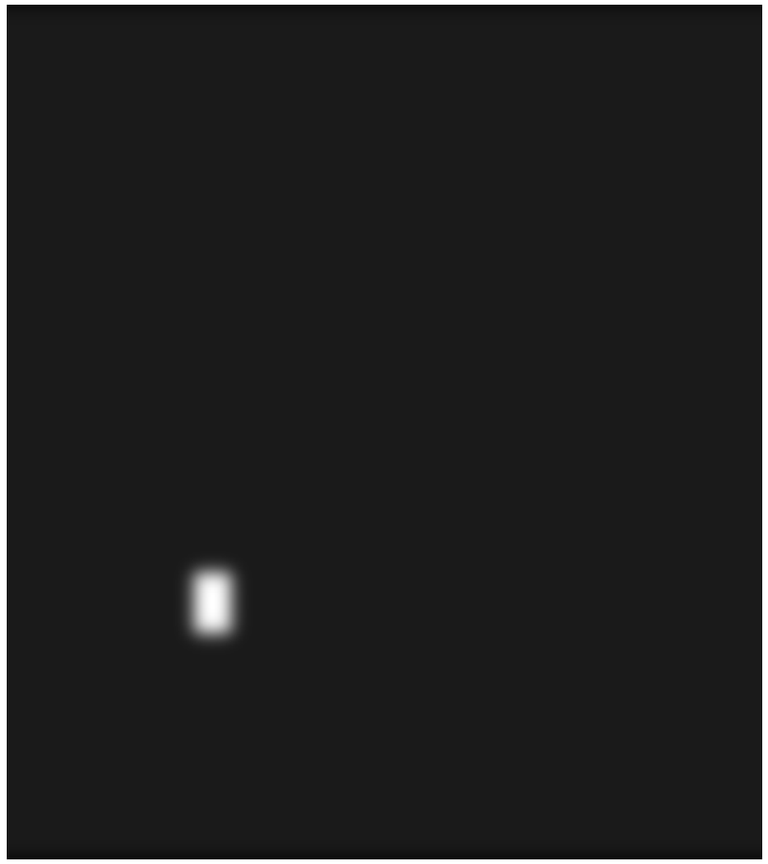
Why do things look as they do?

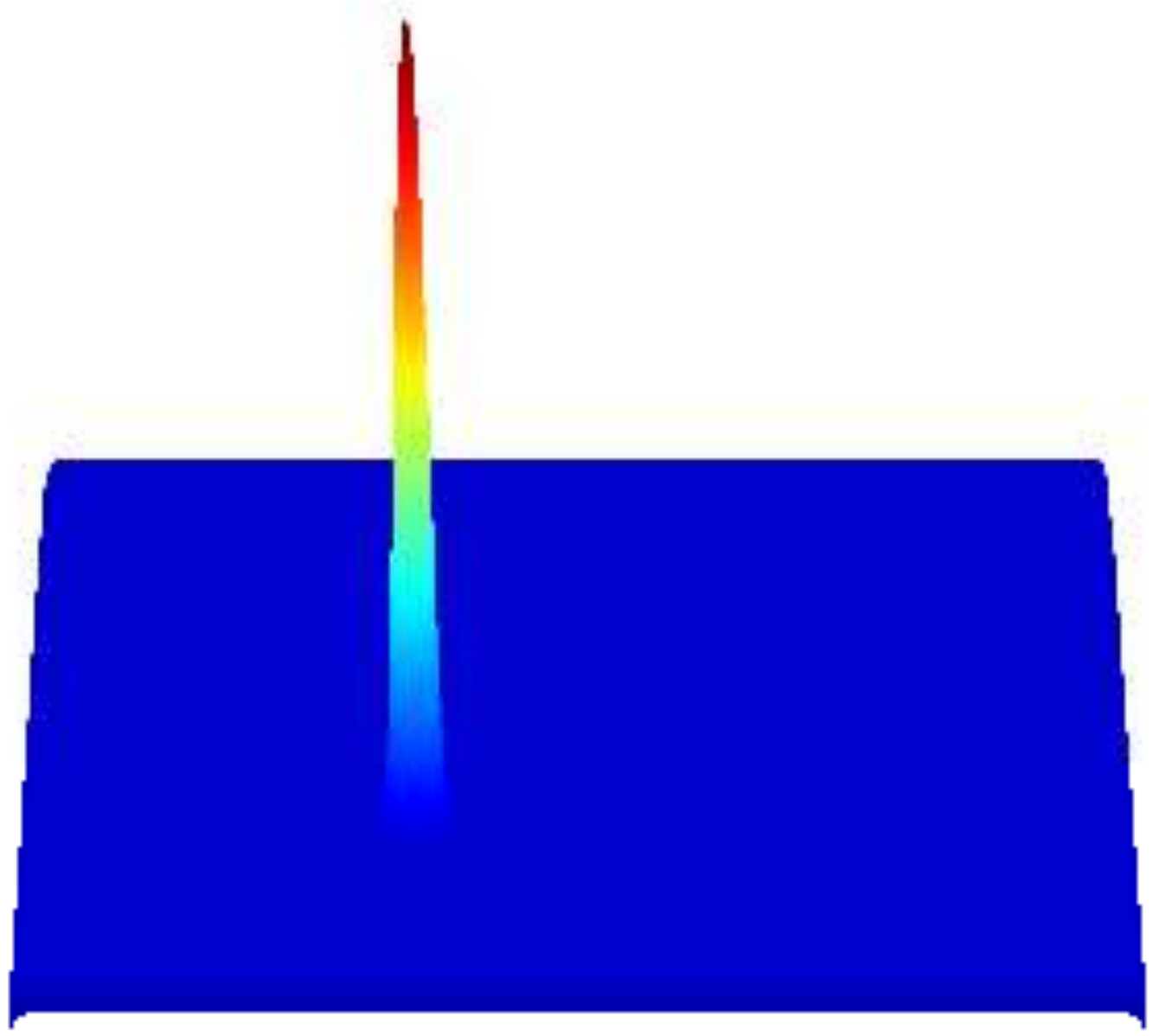
Kurt Koffka, *Principles of Gestalt Psychology*, 1935

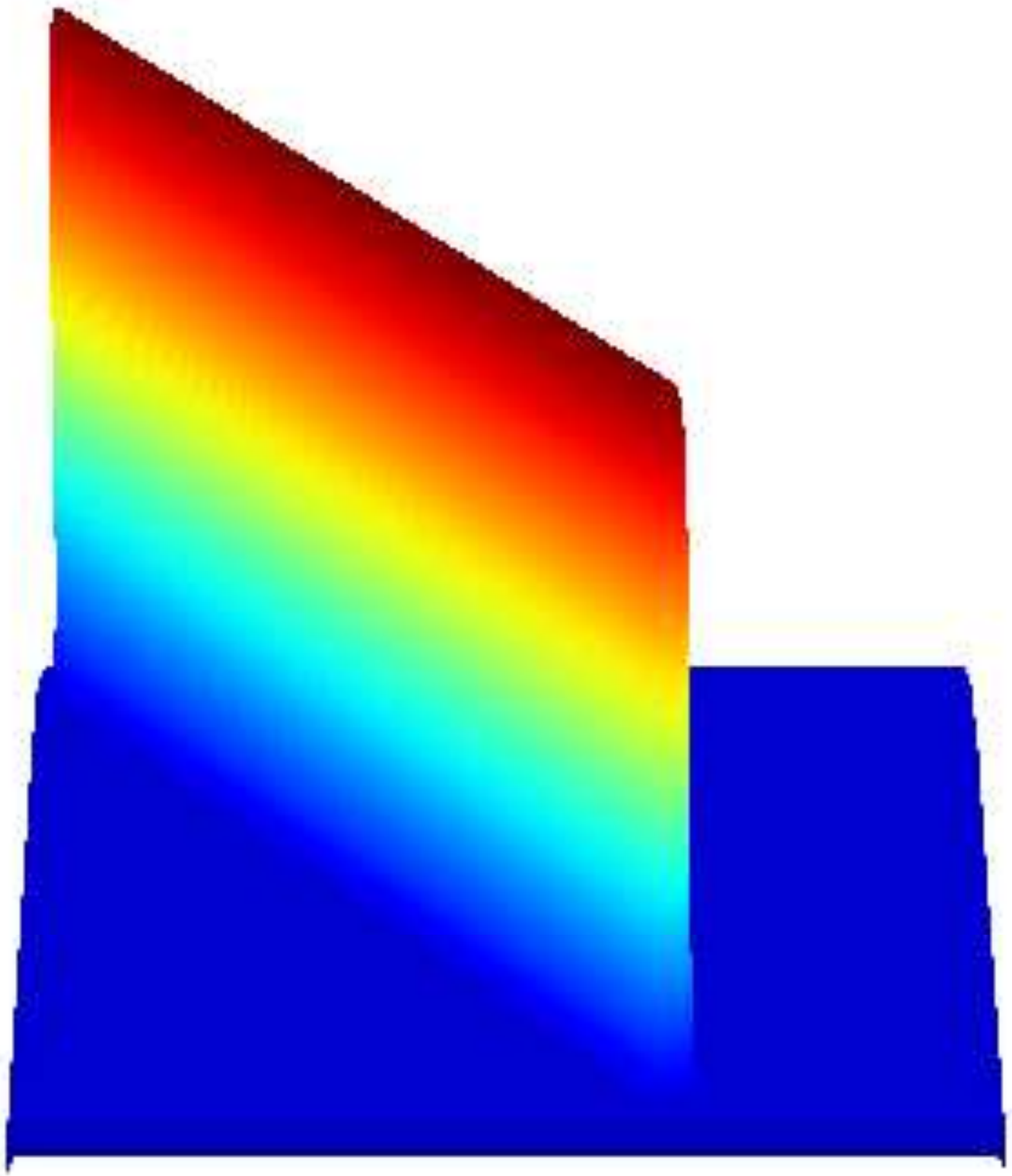

V1
1 cm
V1
V1
LGN
RETINA
SC
M
L
I
PULVINAR
MEDIAL PREFRONTAL
CINGULATE
DORSAL PREFRONTAL
46
MOTOR
FEF
SOMATO SENSORY
LATERAL PREFRONTAL
ORBITO-FRONTAL
AUDITORY
OLFACTORY
ER
36
MDP
PO
V2d
MIP
PIP
V3
VIP
LIP
V3A
DP
7a
MSTd
MT
MSTl
STPp
FST
STPa
V4t
V4d
PITd
CITd
AITd
PITv
CITv
AITv
V4v
VP
VOT
V2v
TF
TH
ER
HC
1 cm

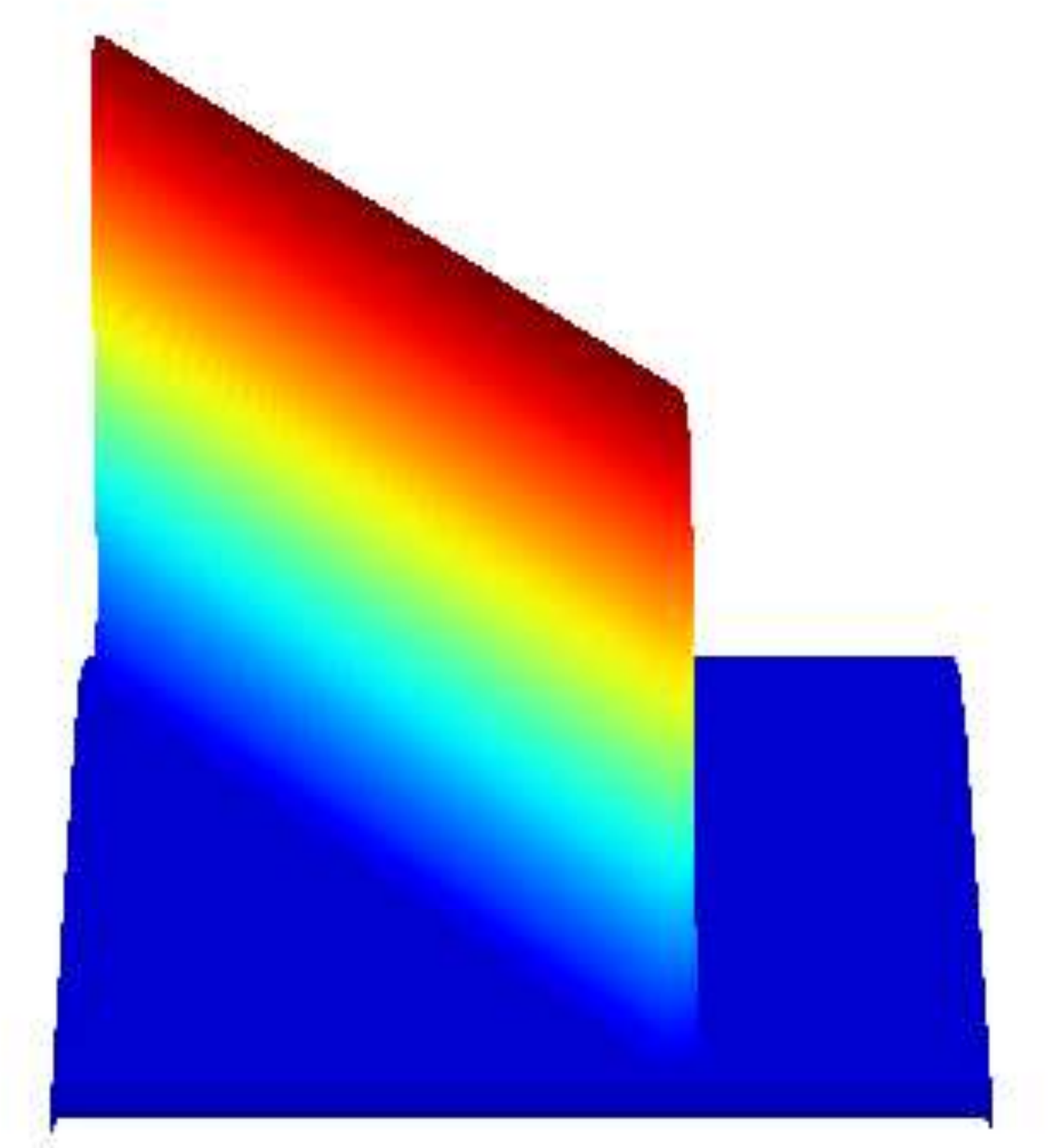

HC
ER
36
46
TF
TH
STPa
AITd
AITv
7b
7a
FEF
STPp
CITd
CITv
VIP
LIP
MSTd
MSTl
FST
PITd
PITv
DP
VOT
MDP
MIP
PO
MT
V4t
V4
PIP
V3A
V3
VP
M
V2
P-B
P-I
M
V1
P-B
P-I
M
P
LGN
M
P
RGC

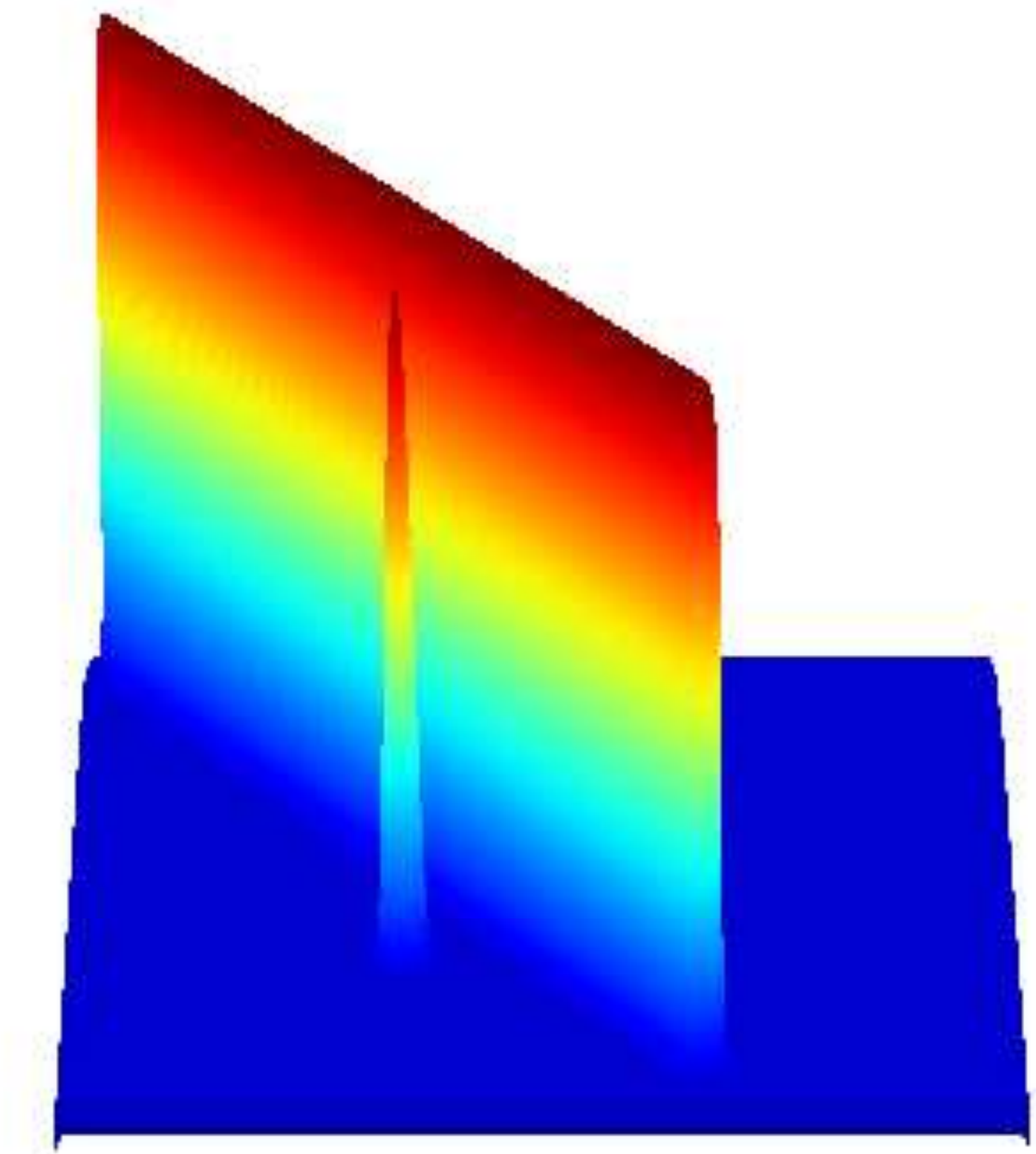


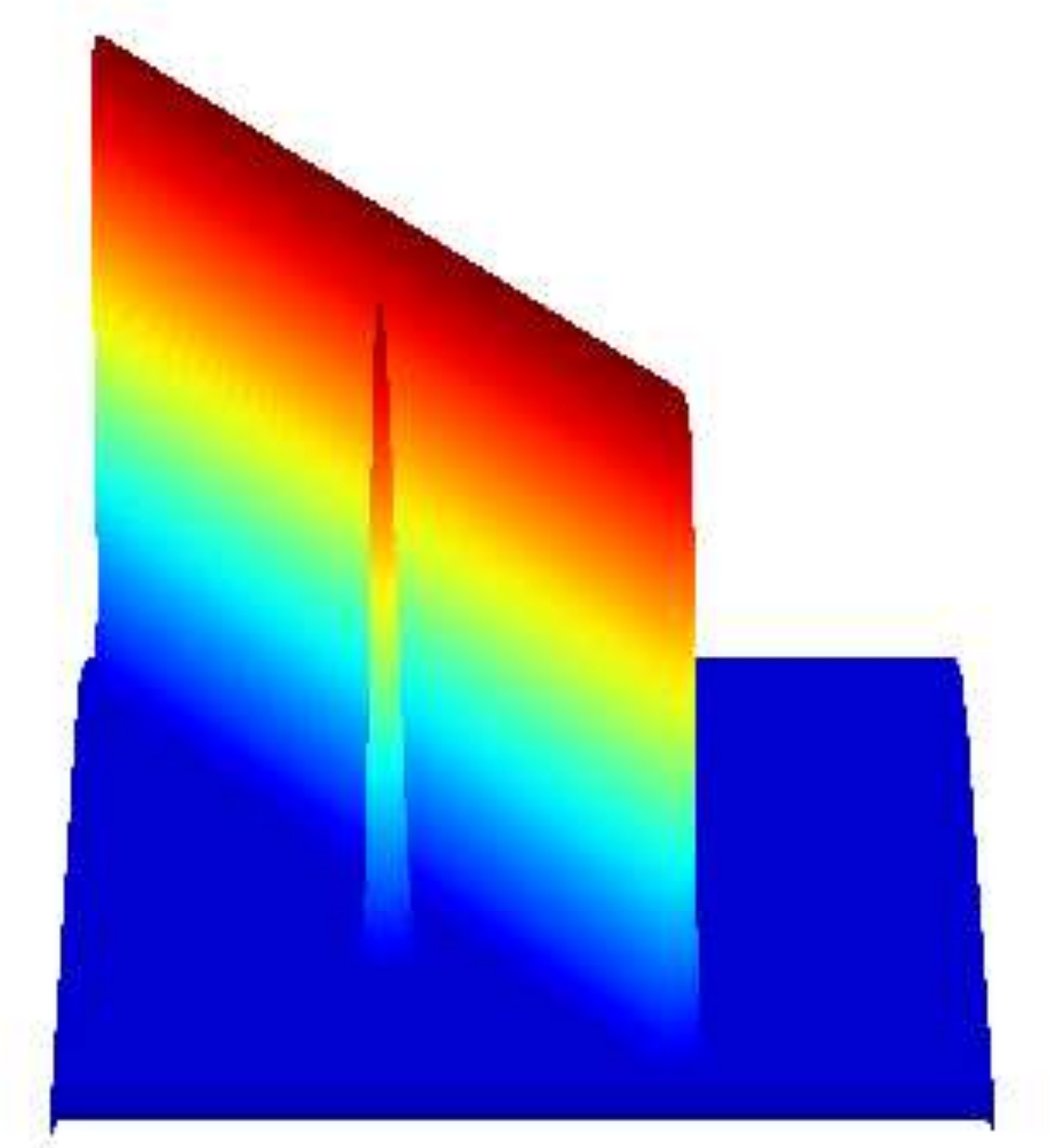


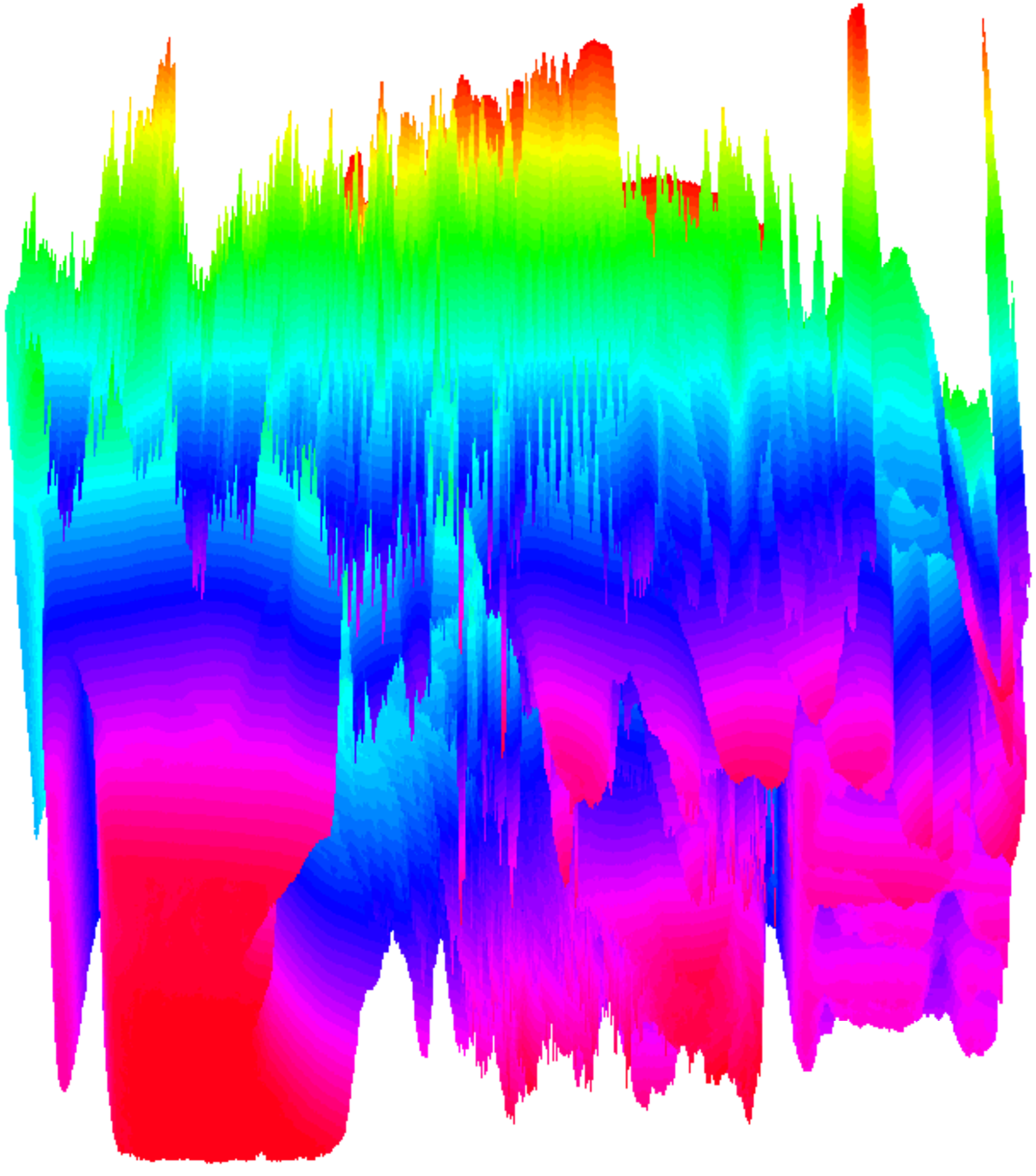


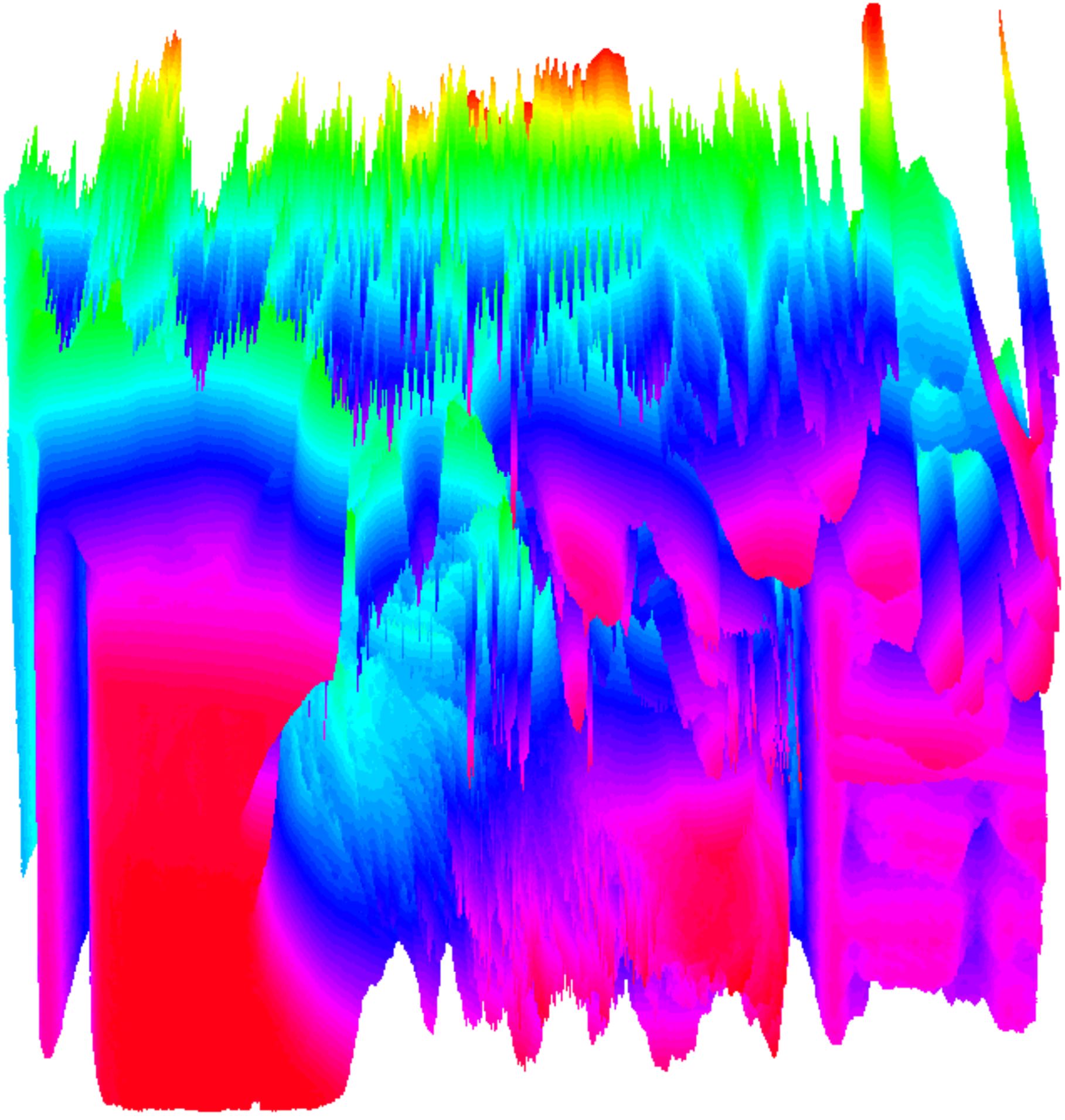


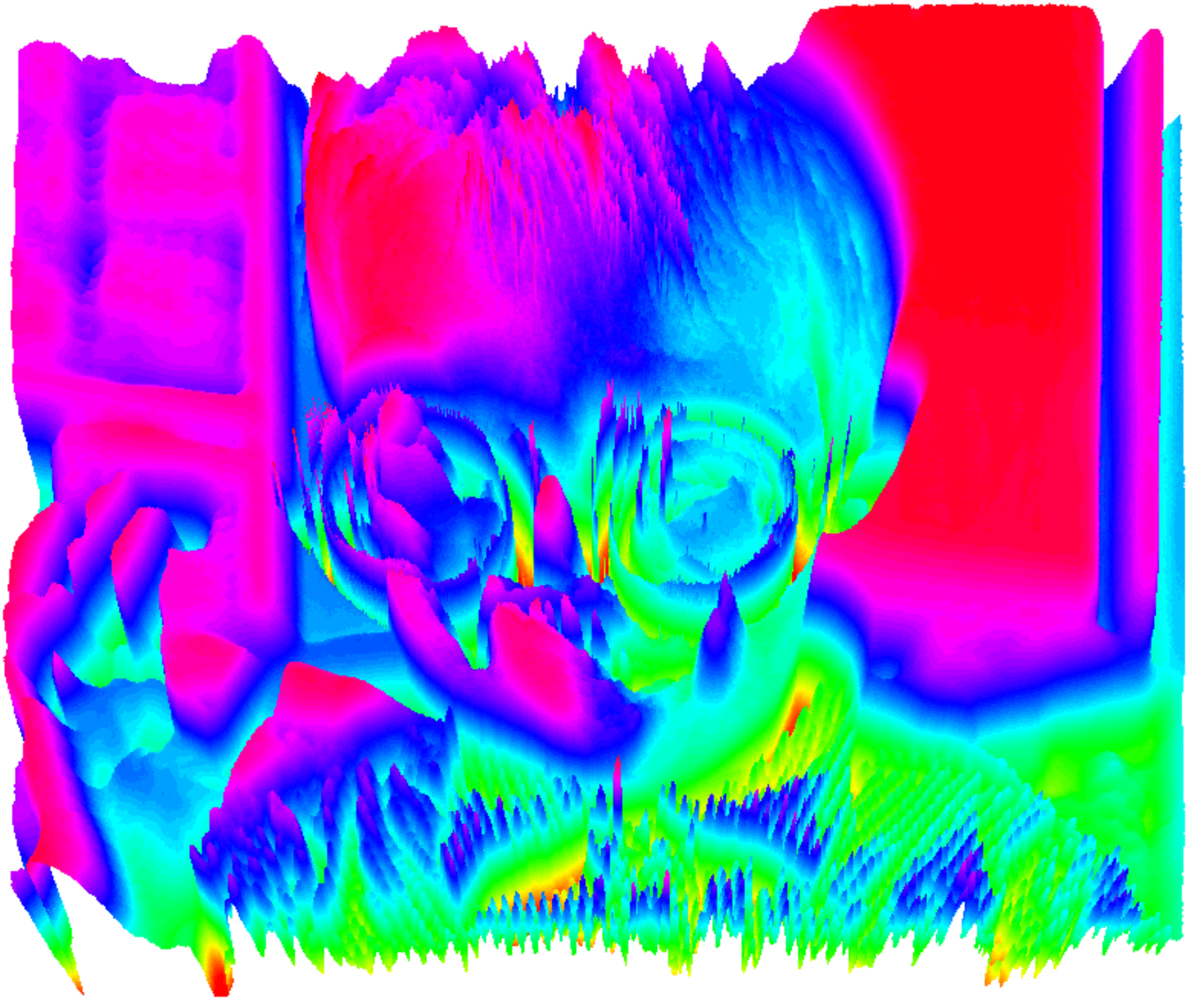






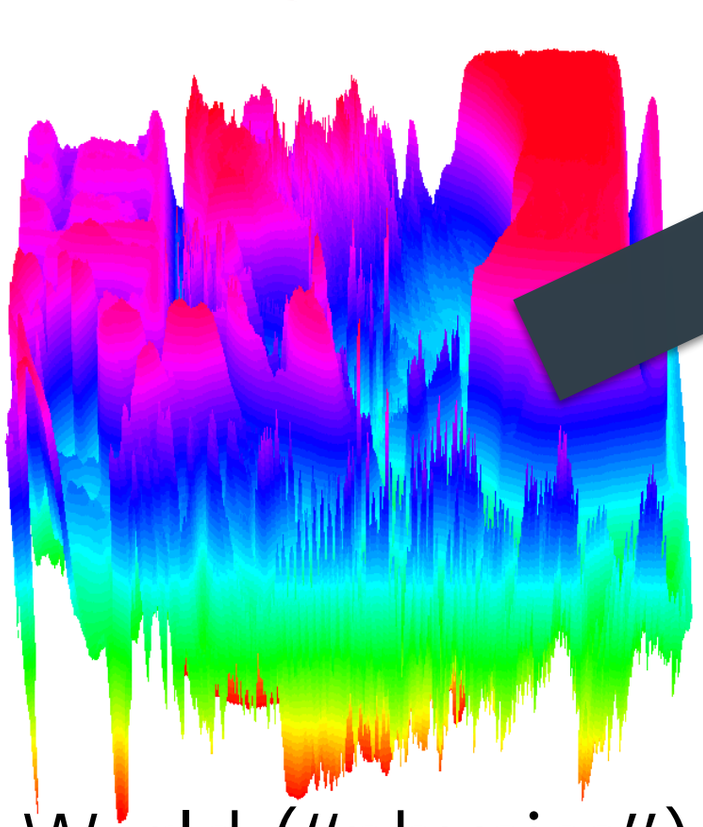






„Visual Percept“
Subjective experience

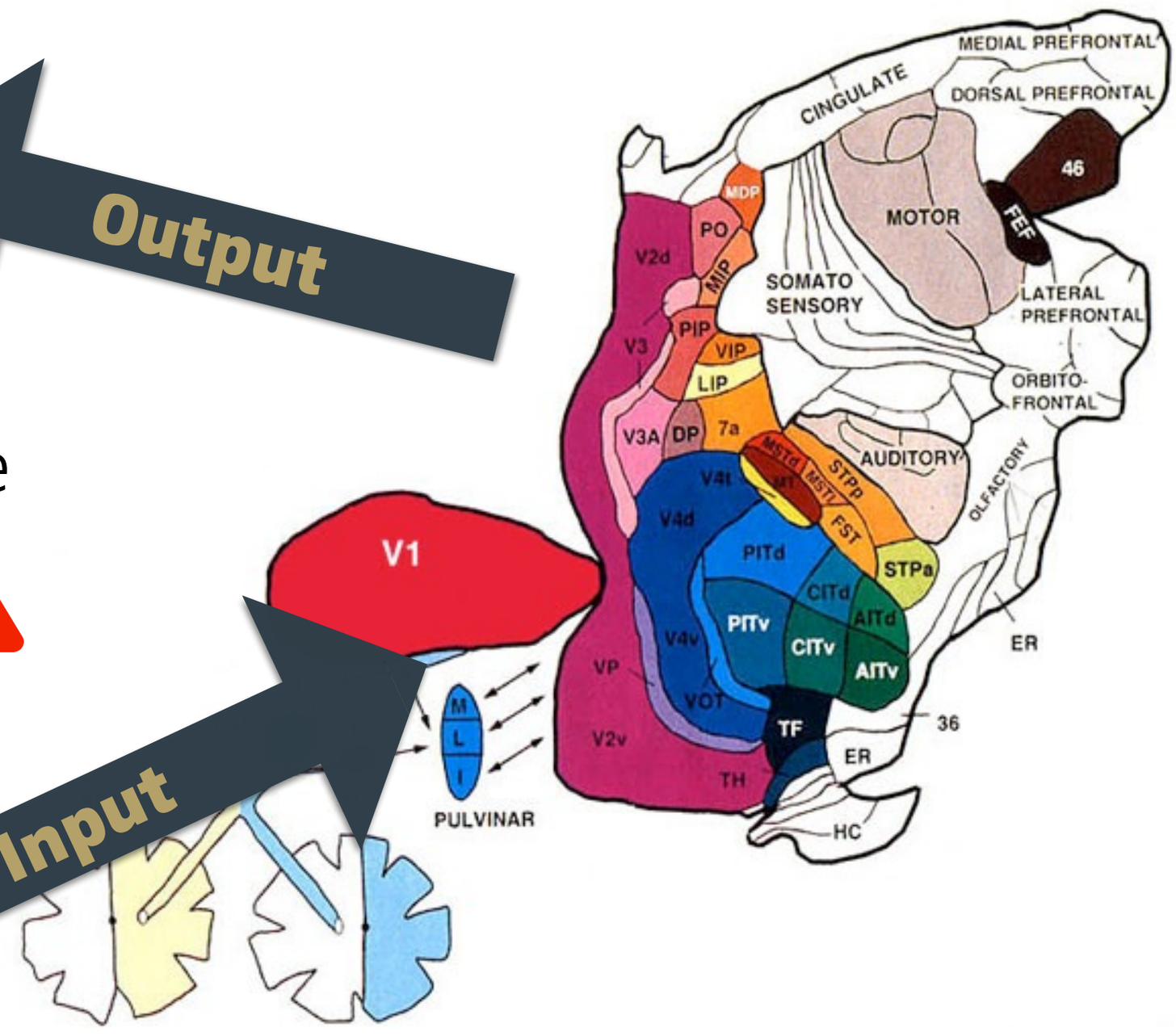
≠



World (“physics”)

Output

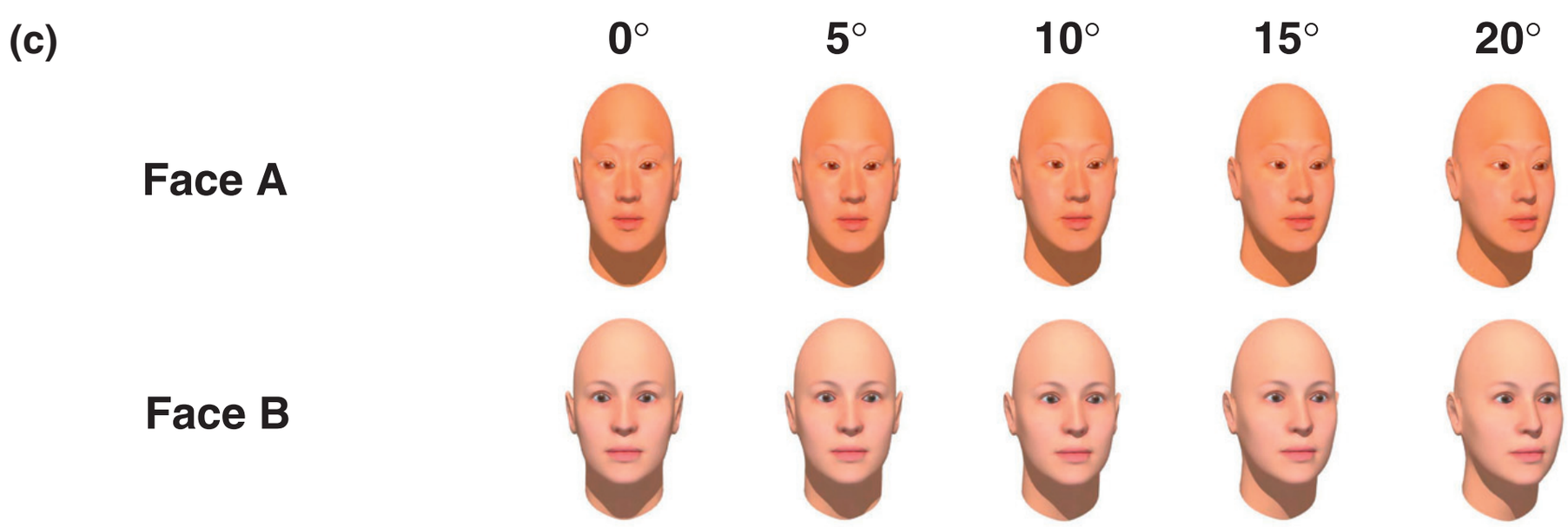
Input



Why do things look as they do?

"At a functional level, visual object recognition is at the center of understanding how we think about what we see. Object identification is a primary end state of visual processing and a critical precursor to interacting with and reasoning about the world."

(Peissig & Tarr, 2007, p. 76)



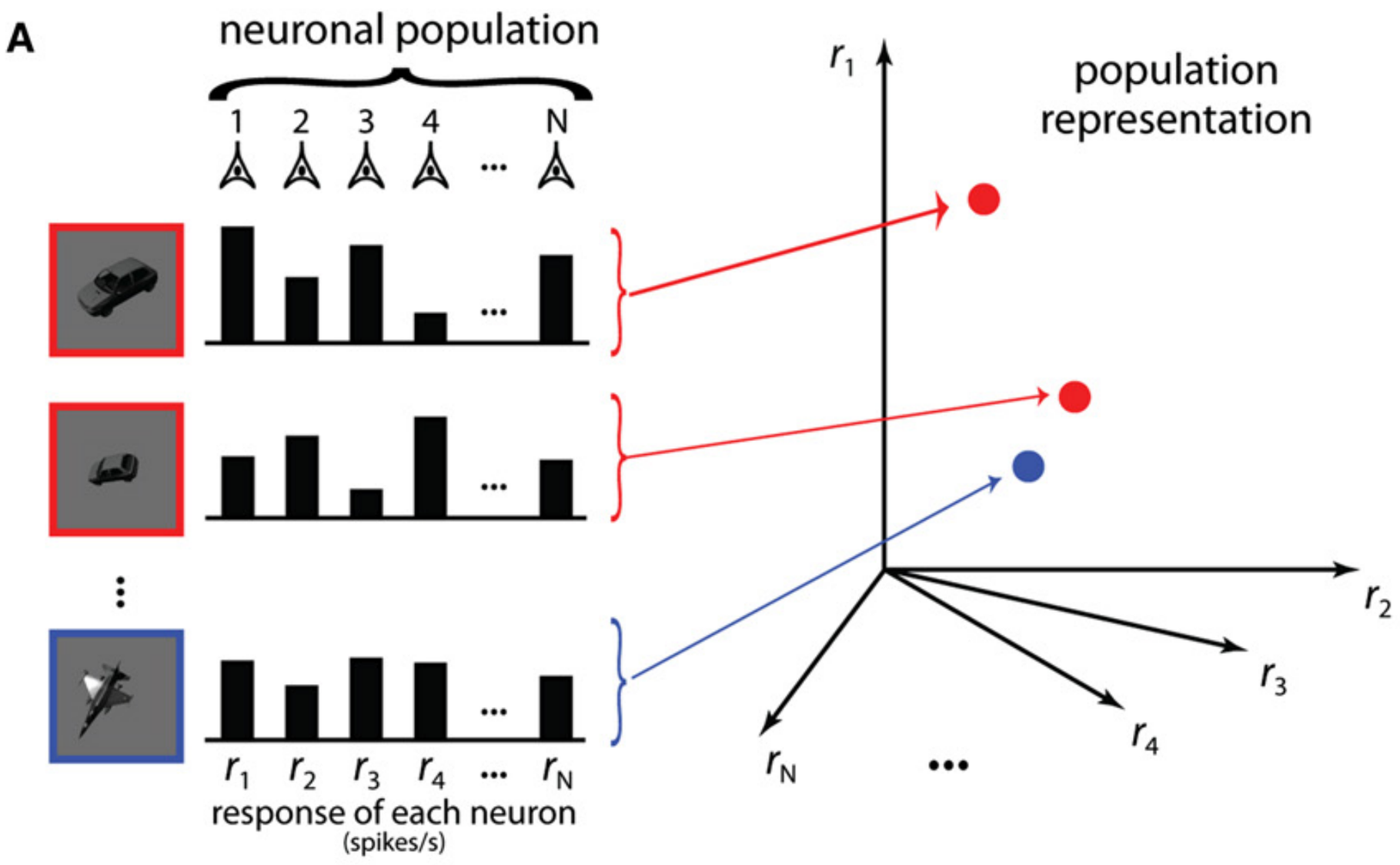
Cox (2014)

Neuron
Perspective

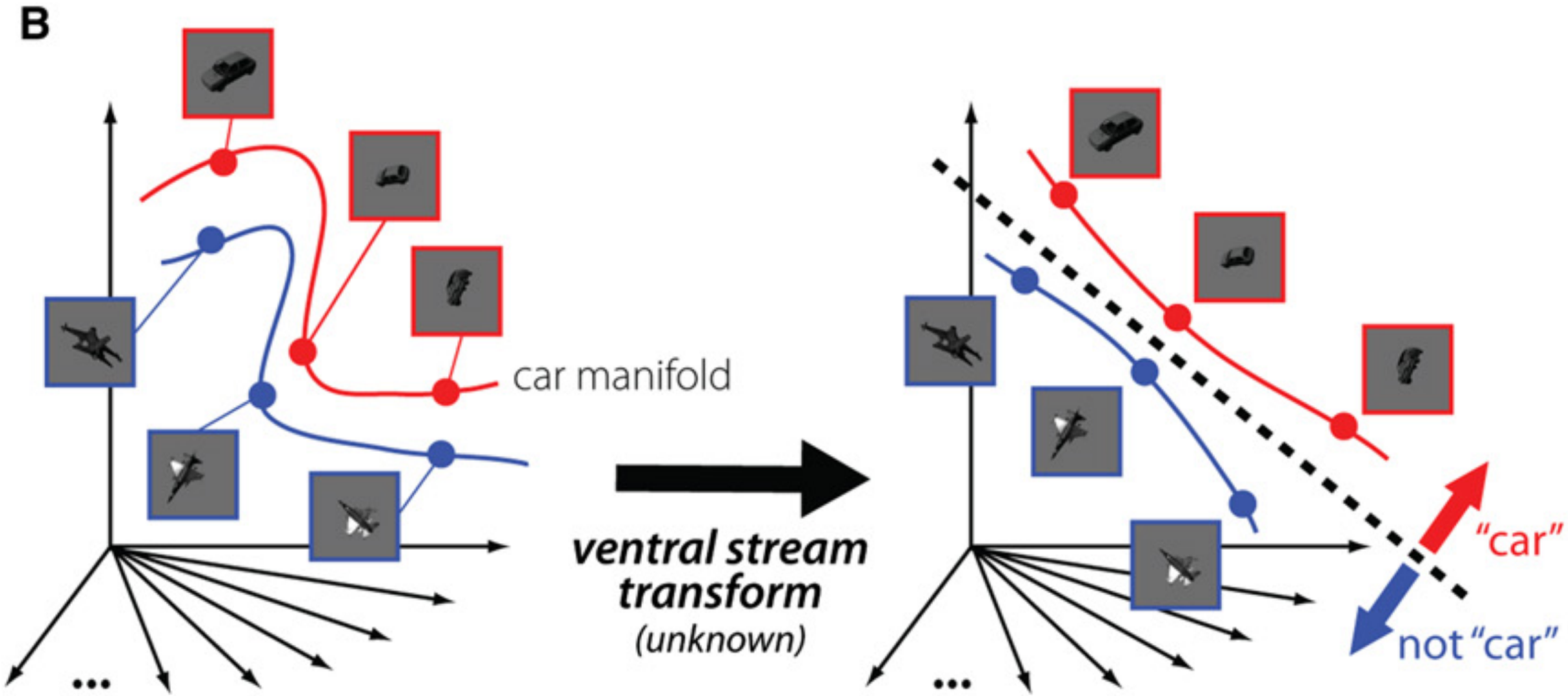
How Does the Brain Solve Visual Object Recognition?

James J. DiCarlo,[1,*] Davide Zoccolan,[2] and Nicole C. Rust[3]
[1]Department of Brain and Cognitive Sciences and McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[2]Cognitive Neuroscience and Neurobiology Sectors, International School for Advanced Studies (SISSA), Trieste, 34136, Italy
[3]Department of Psychology, University of Pennsylvania, Philadelphia, PA 19104, USA
*Correspondence: dicarlo@mit.edu
DOI 10.1016/j.neuron.2012.01.010

Mounting evidence suggests that 'core object recognition,' the ability to rapidly recognize objects despite substantial appearance variation, is solved in the brain via a cascade of reflexive, largely feedforward computations that culminate in a powerful neuronal representation in the inferior temporal cortex. However, the algorithm that produces this solution remains poorly understood. Here we review evidence ranging from individual neurons and neuronal populations to behavior and computational models. We propose that understanding this algorithm will require using neuronal and psychophysical data to sift through many computational models, each based on building blocks of small, canonical subnetworks with a common functional goal.



DiCarlo et al. (2012)



DiCarlo et al. (2012)

Fundamentals of Neural Networks

Interest in shallow, 2-layer artificial neural networks (ANN)—so-called **perceptrons**—began in the late 1950s and early 60s (Frank Rosenblatt), based on Warren McCulloch and Walter Pitts's as well Donald Hebb's ideas of computation by neurons from the 1940s.

Second wave of ANN research and interest in psychology—often termed **connectionism**—after the publication of the **parallel distributed processing** (PDP) books by David Rumelhart and James McClelland (1986), using the backpropagation algorithm as a learning rule for multi-layer networks.

Three-layer network with (potentially infinitely many) hidden units in the intermediate layer is a universal function approximator (Kurt Hornik, 1991).

Non-convex optimization problems during backpropagation training, and lack of data and computing power limited the usefulness of the ANNs:

Universal function approximator in theory, but in practice three-layer ANNs could often not successfully solve complex problems.

Fundamentals of Neural Networks (cont'd)

Breakthrough again with so-called **deep neural networks** or **DNN**s, widely known since the 2012 NIPS-paper by ALEX KRIZHEVSKY ET AL.

DNN: loose terminology to refer to networks with at least two hidden or intermediate layers, typically at least five to ten (or up to dozens):

1. Massive increase in labelled training data (''the internet''),
2. computing power (GPUs),
3. simple non-linearities (ReLU) instead of sigmoids,
4. convolutional rather than fully connected layers,

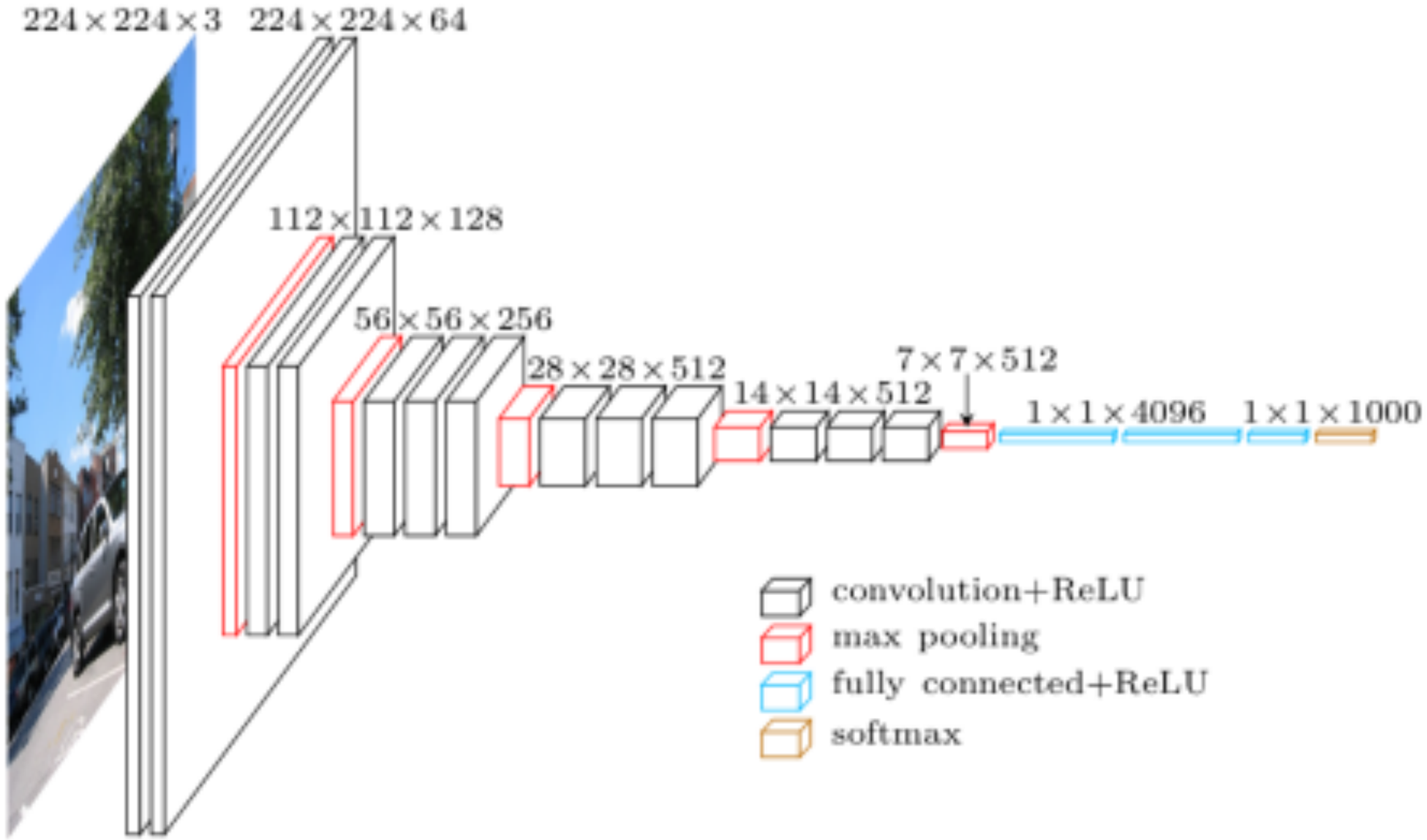
and

5. *weight sharing* across deep layers

appear to be the critical ingredients for the current success of DNNs, and makes them the current method of choice in ML, particular in application.

At least superficially DNNs appear to be similar to the human object recognition system: convolutions (''filters'', ''receptive fields'') followed by non-linearities and pooling is thought to be the canonical computation of cortex, at least within sensory areas.

Example: VGG-16



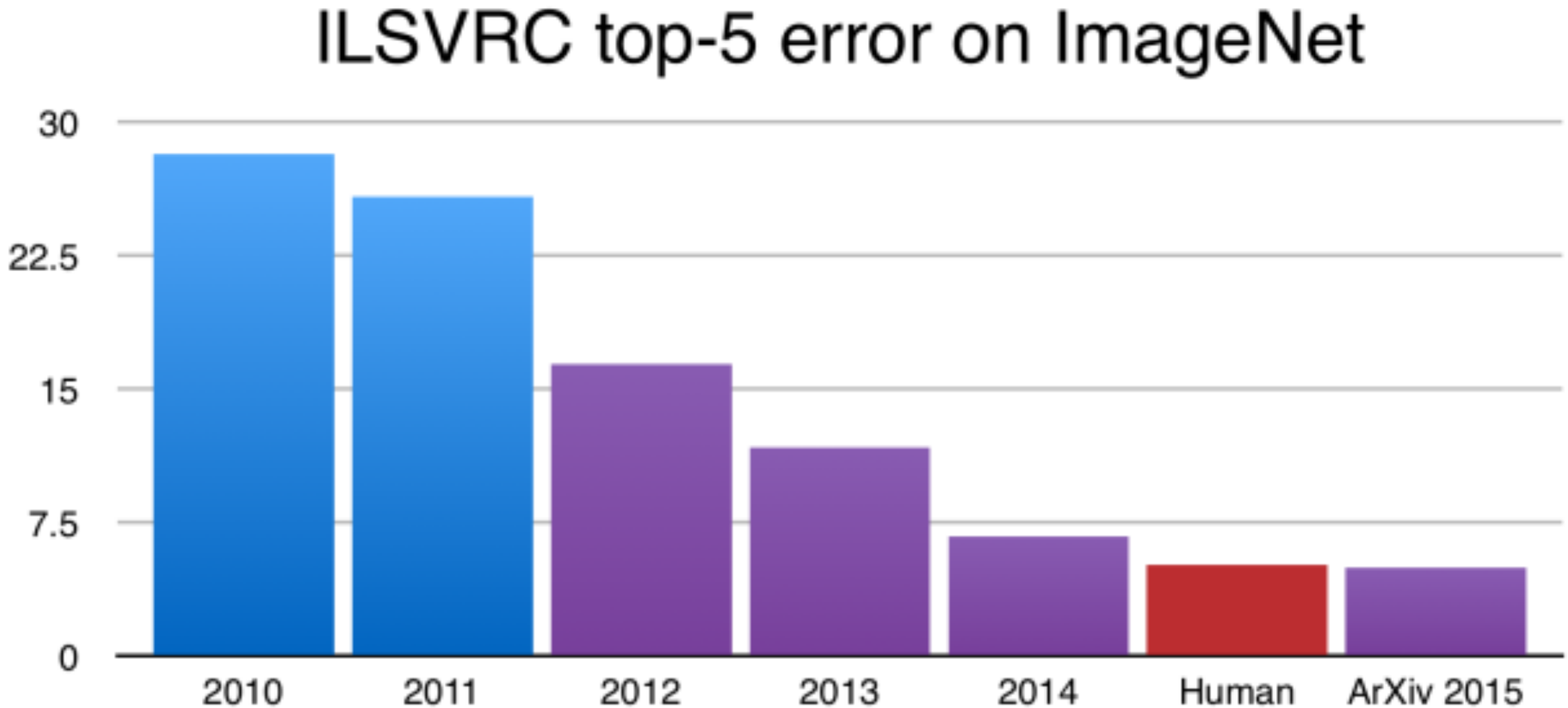
https://www.cs.toronto.edu/~frossard/post/vgg16/#architecture

VGG16 by Simonyan & Zisserman (2014); 92.7% top-5 test accuracy on ImageNet

What changed vision research in 2012?

ImageNet challenge: 1000 categories, 1.2 million training images.

AlexNet by Krizhevsky, Sutskever & Hinton (2012) appears on the stage, and basically reduces the prediction error by nearly 50%:



Recent studies suggest that state-of-the-art convolutional 'deep' neural networks (DNNs) capture important aspects of human object perception. We hypothesized that these successes might be partially related to a human-like representation of object shape.

Kubilius et al. (2016), PLoS Comp. Biol., p. 1

… deep neural networks can match or even exceed human-level performance in pattern recognition … , and they develop representations that are remarkably similar to those found in the mammalian neocortex. These observations suggest that something akin to deep learning may, in fact, be occurring in the real brain.

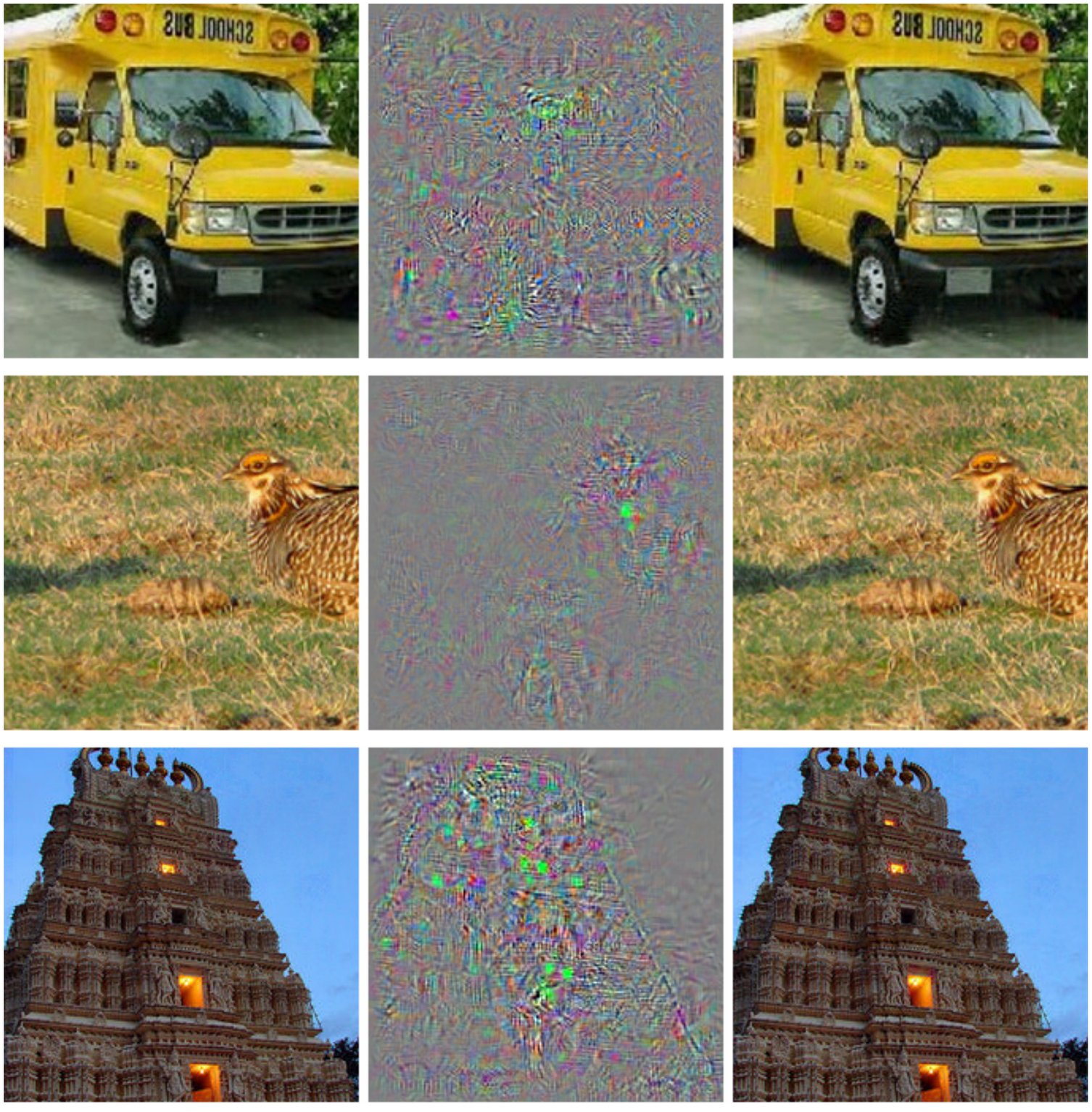
Cosyne, 2017, Workshop on ‘‘Deep learning’’ and the brain (Day 2)

It is a capital mistake to theorize
before one has data.

(Sherlock Holmes)

Arthur Conan Doyle (1891). *A Scandal in Bohemia.*
The Strand Magazine, July issue.

Adversarial attacks?



Szegedy et al. (2014)

Adversarial examples? (cont'd)

Reese
Witherspoon

Russel
Crowe

Sharif et al. (2016)

Adversarial examples? (cont'd)






Sharif et al. (2016)

Adversarial attacks, random perturbations and generalisation in DNNs

Adversarial attacks show generalisation errors of DNNs—however, only to carefully designed stimuli, exploiting the knowledge of the weights and gradients in the DNN.

Data augmentation (re-training) often leads to robustness against a specific adversarial attack, but it does not guarantee robustness against adversarial perturbations in general.

Strong argument against DNNs using similar computations as human vision?

The susceptibility of deep neural networks to adversarial examples exposes one of the most striking differences in the sensory decision making of humans and machines. (from https://robust.vision/benchmark)

Human vision suffers from so-called visual illusions, carefully designed stimuli, leading the visual system astray—illusions as adversarial stimuli?

What about generalisation abilities—robustness— of DNNs and humans to weak signals and to randomly degraded stimuli rather than carefully engineered "freak" stimuli?

Images and categories

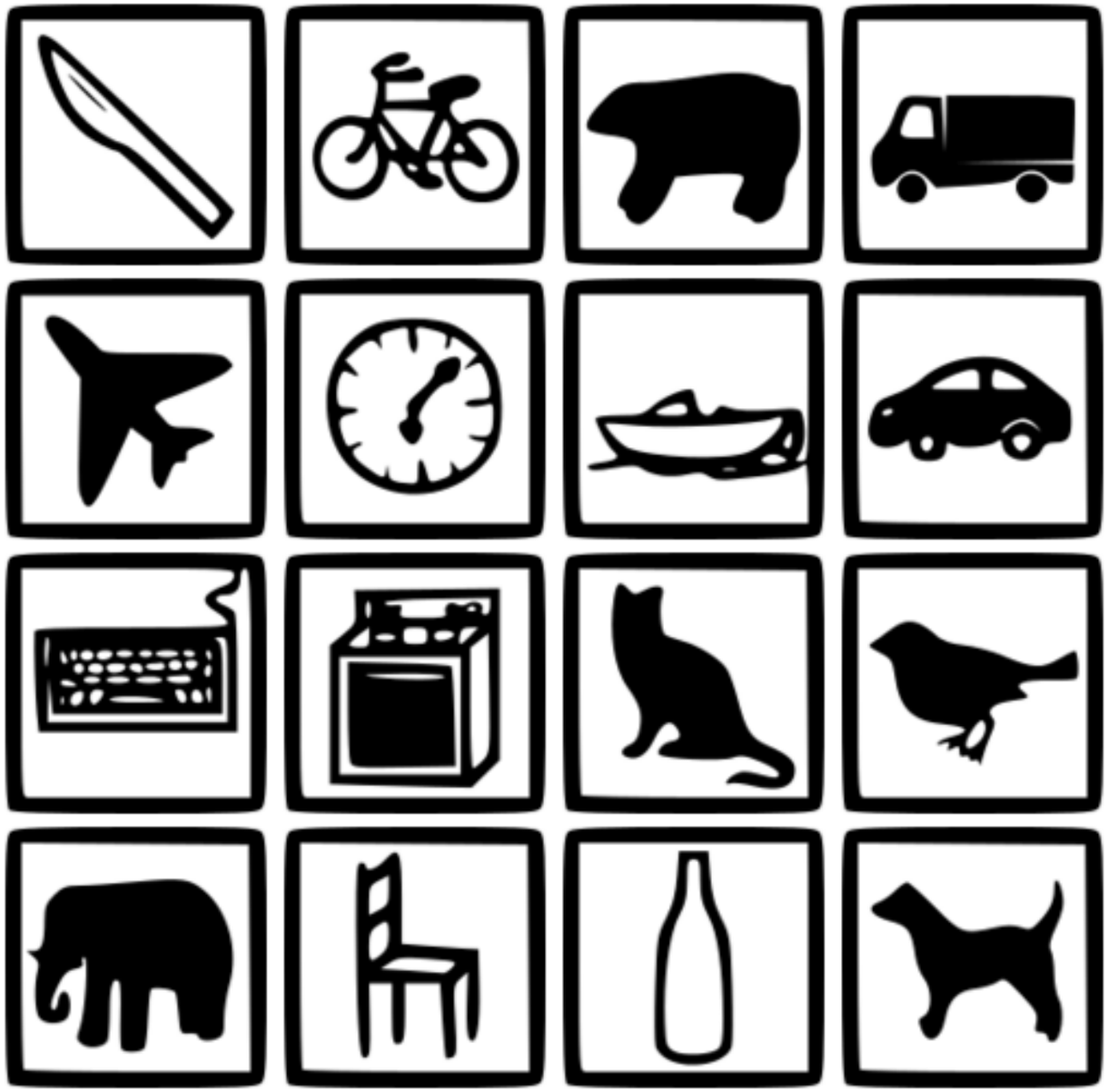
All images from the training set of ImageNet 2012 database.

To compare human observers to DNNs, a categorisation in 1000+ classes at different psychological levels is not optimal.

MS COCO database is structured according to 91 basic or entry-level categories, making it an excellent source for an object recognition task using human observers.

We used MS COCO categories with images from ImageNet, mapping, if possible, the ImageNet label to a MS COCO entry-level category.

We retained 16 non-ambigious categories with 213,555 images.

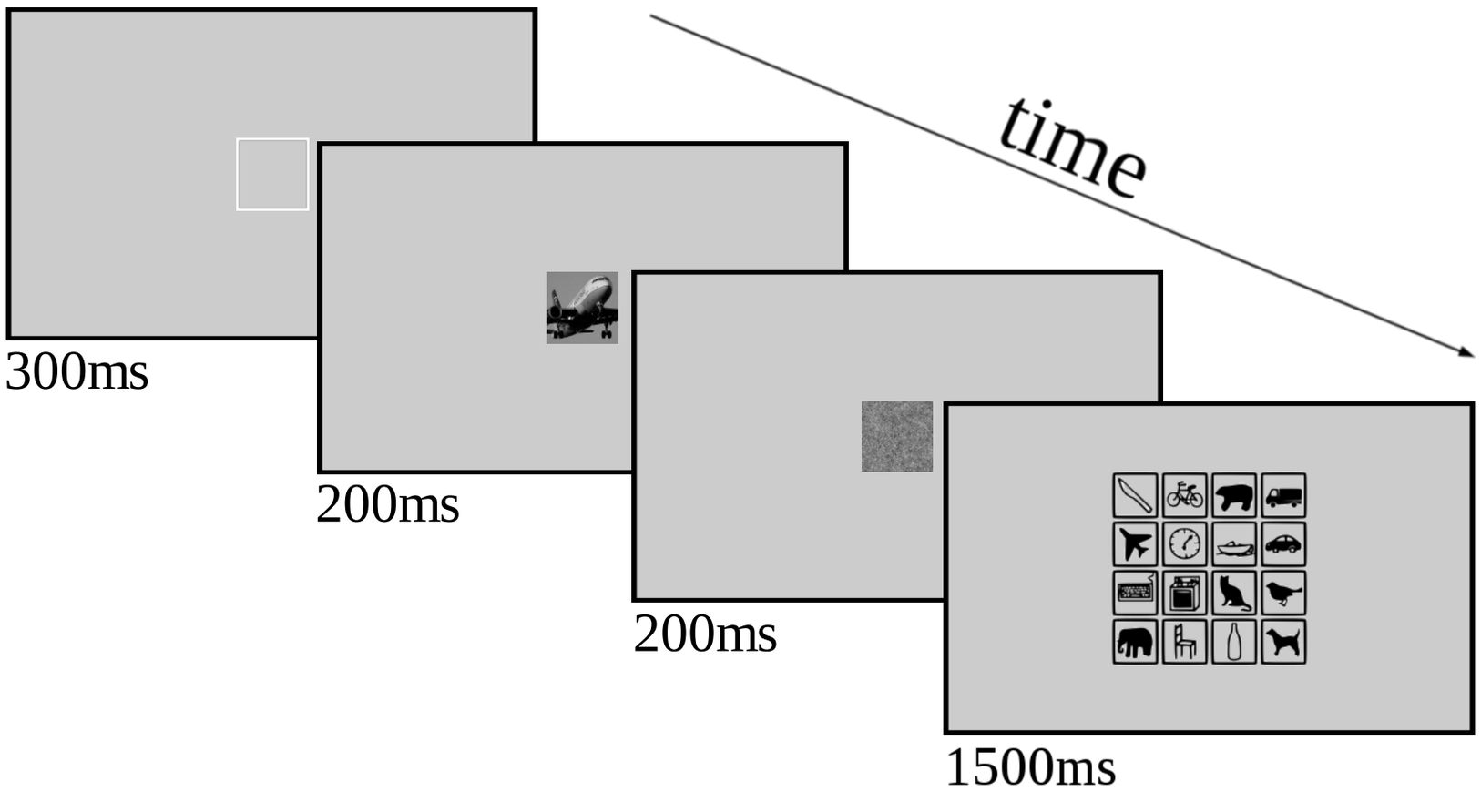


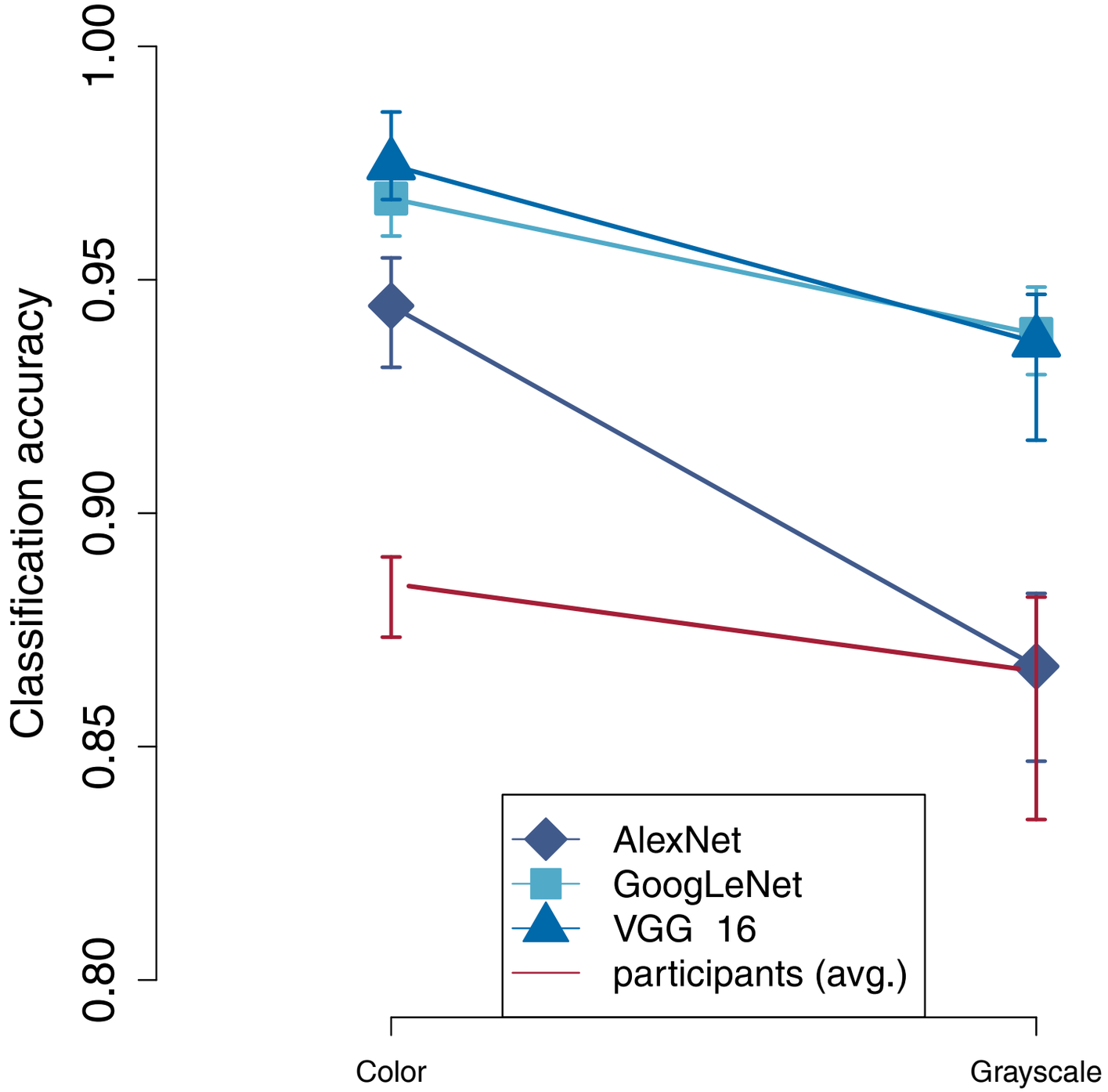
DNNs and methods

Three well-known, successful and architecturally different DNNs:
AlexNet, VGG-16, GoogleLeNet.

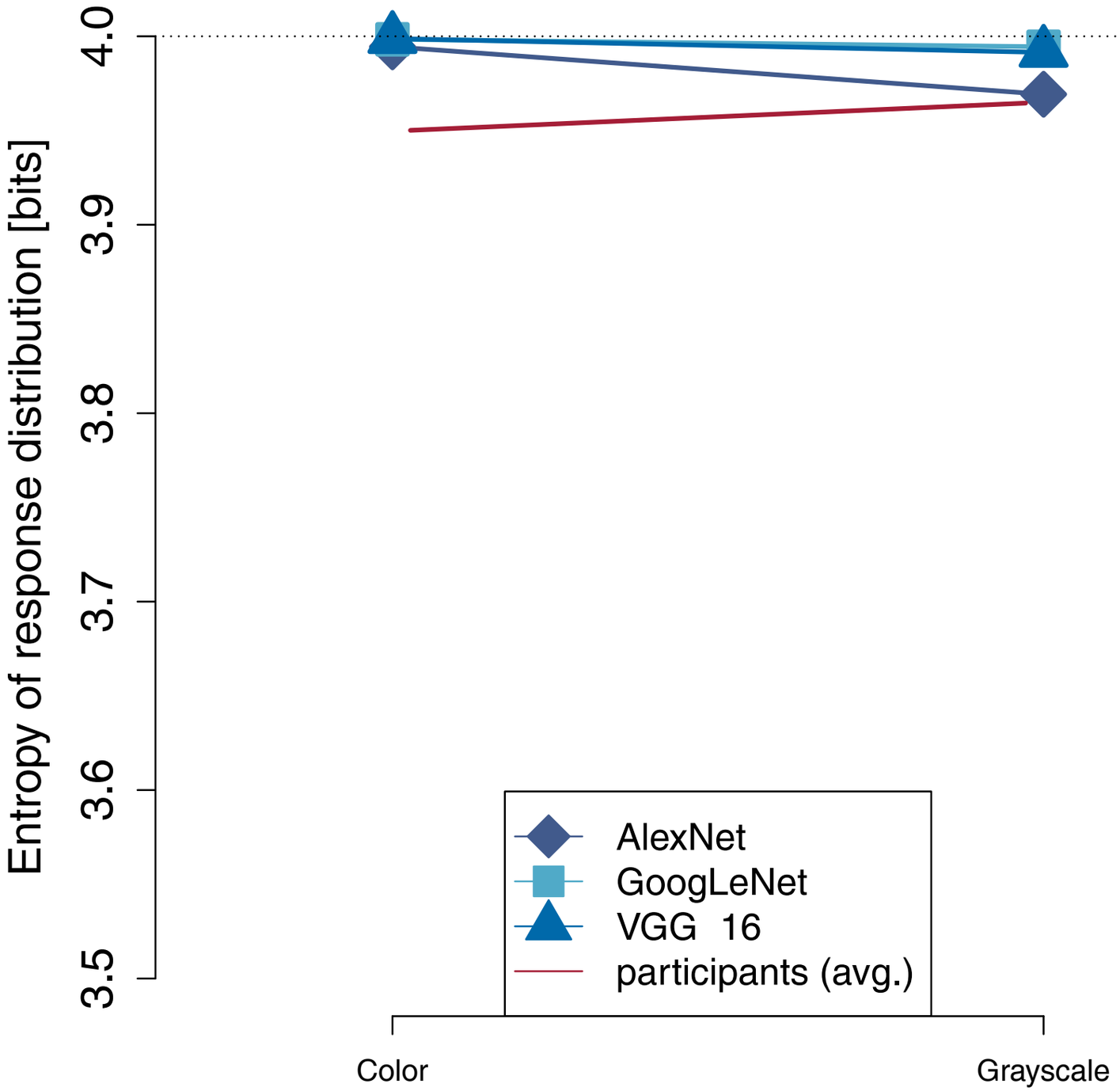
Experimental protocol chosen to allow fair comparison between humans and DNNs as models of the human visual system for core object recognition:

- short presentation time (200 ms)
- followed by a high contrast 1/f noise mask (200 ms)
- fast-paced responding (1500 ms, mouse to select one of 16 icons)



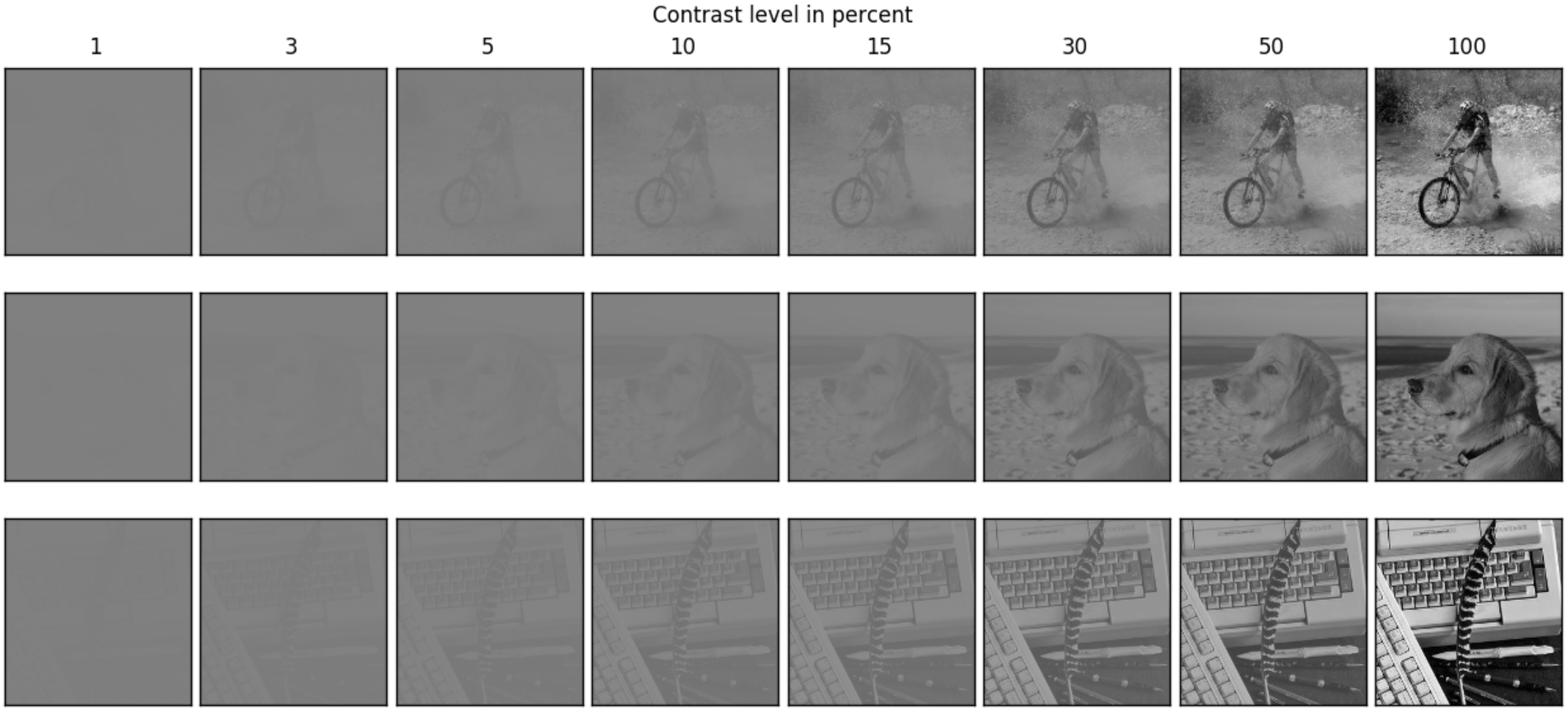


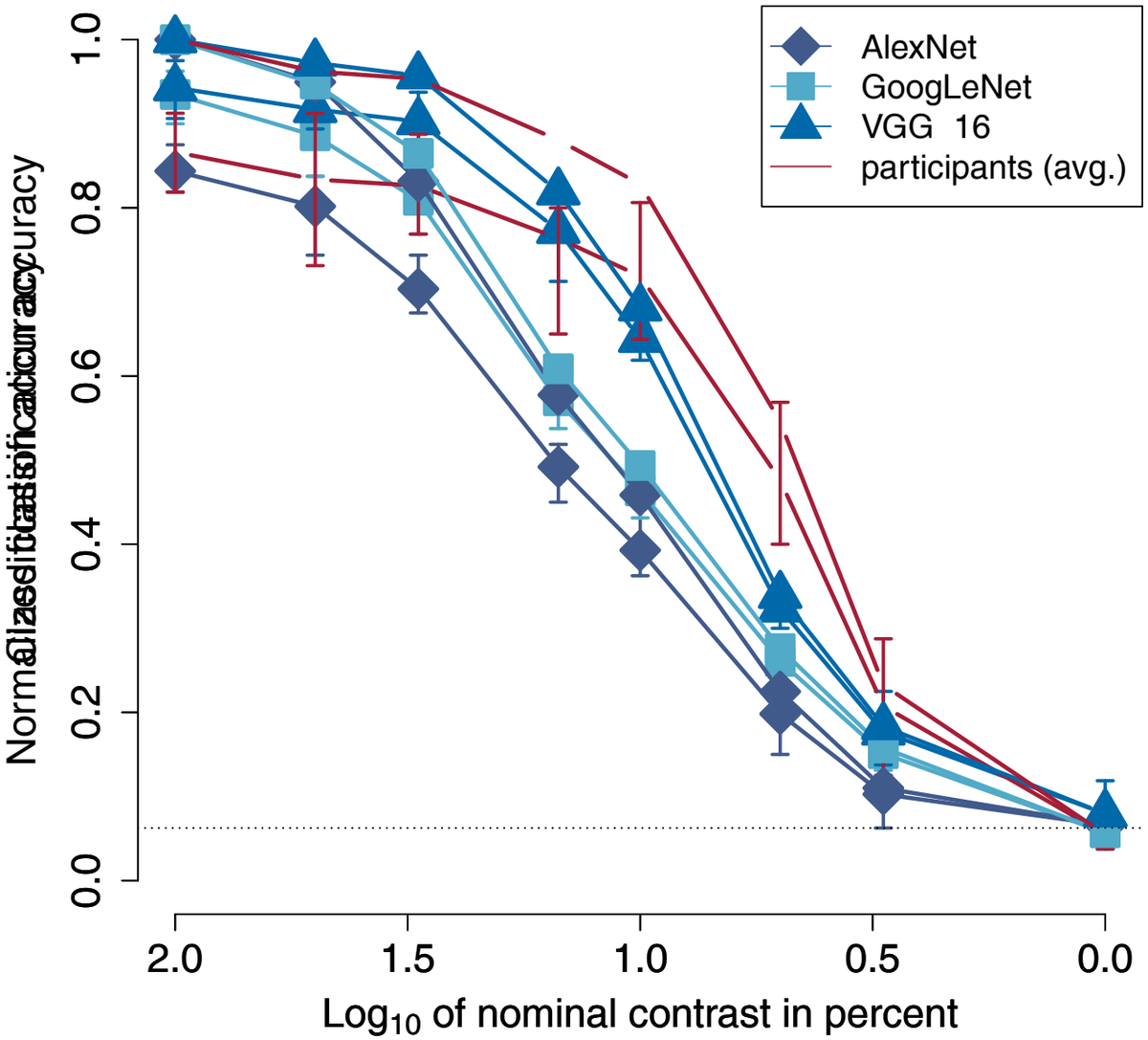
Geirhos, Janssen, Schütt, Rauber, Bethge and Wichmann. (2017). Comparing deep neural networks against humans: object recognition when the signal gets weaker. *arXiv,* 1706.06969v1, 1-31.

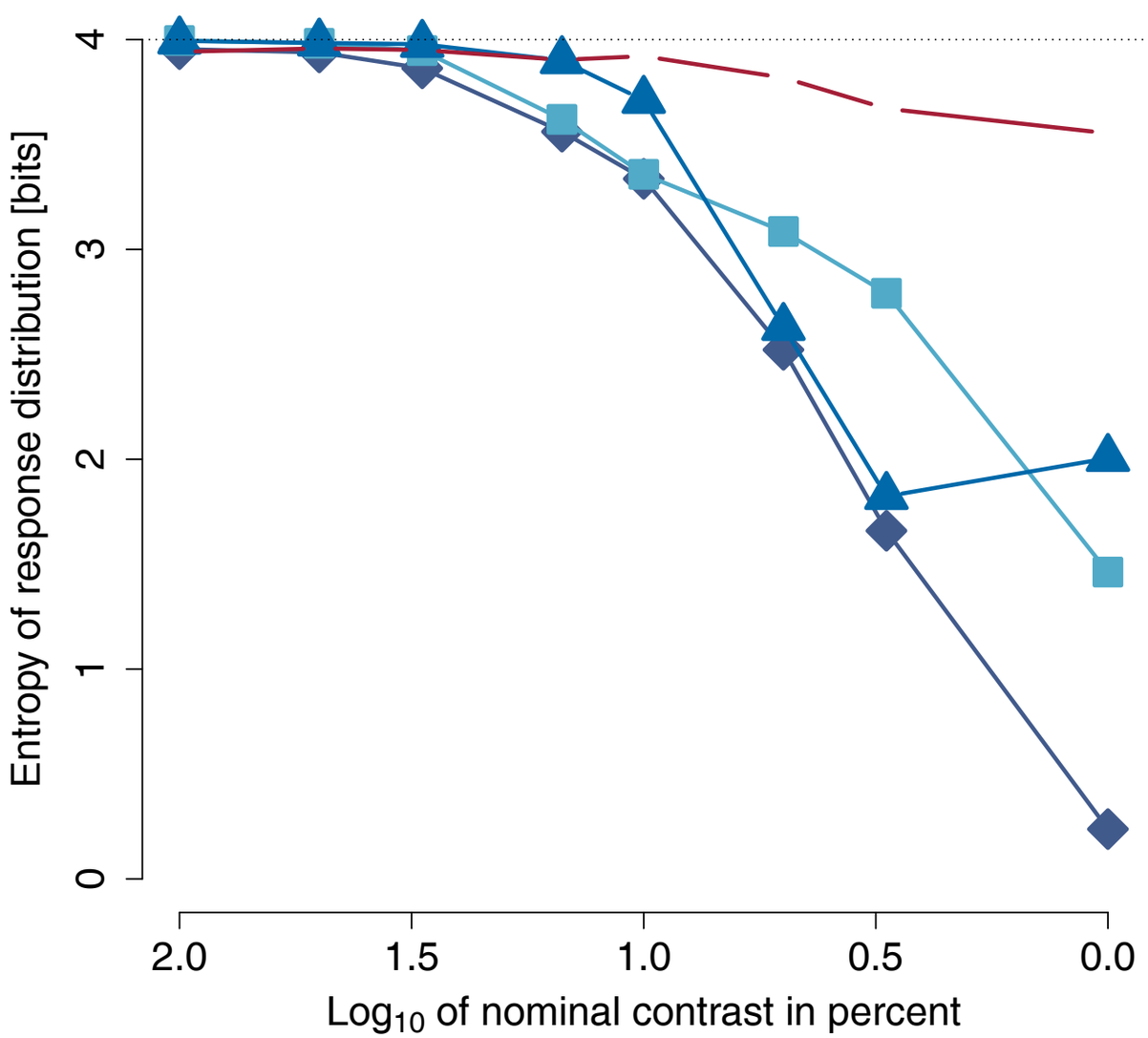


Geirhos et al. (2017)

Contrast reduction







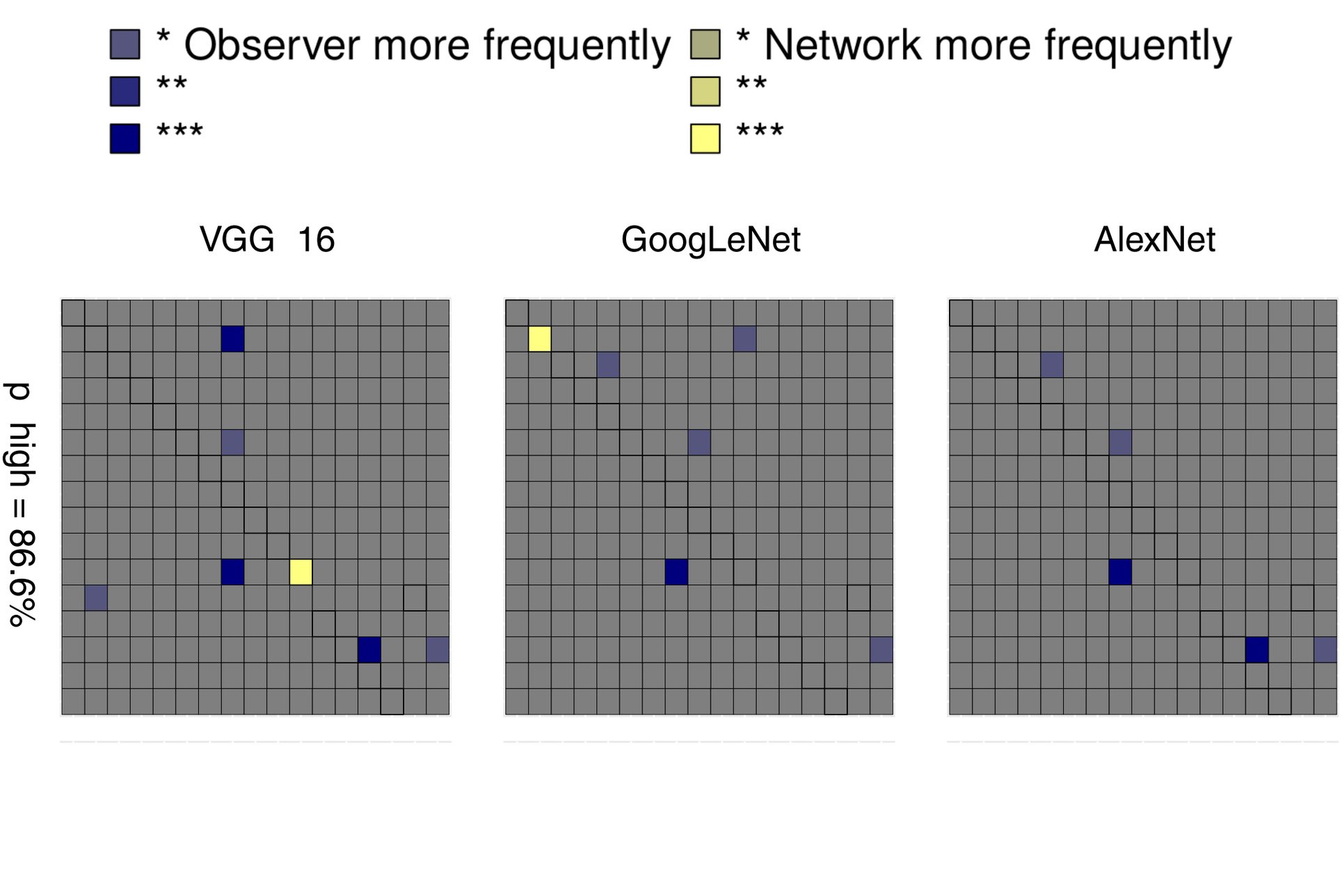
Geirhos et al. (2017)

4.2	7.5	0.0	0.8	0.0	0.8	1.7	1.7	0.8	0.0	2.5	1.7	3.3	0.0	1.7	1.7
0.0	0.0	0.0	0.0	0.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	90.0
0.0	0.0	0.0	1.7	0.0	0.0	0.0	0.8	0.8	1.7	0.0	0.0	1.7	4.2	85.0	0.0
0.8	0.0	0.0	0.0	0.0	1.7	0.0	0.0	0.0	0.8	0.0	0.0	1.7	79.2	2.5	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.7	1.7	0.0	0.0	86.7	0.0	1.7	0.0
0.0	1.7	0.0	0.0	0.0	0.0	0.0	0.8	0.0	0.0	0.0	95.0	0.0	0.8	0.0	0.0
0.0	6.7	0.0	1.7	0.0	0.0	0.0	11.7	0.0	0.0	87.5	0.8	0.0	2.5	0.0	0.0
0.8	0.0	0.8	0.0	0.0	1.7	0.0	0.0	1.7	93.3	0.0	0.0	0.0	1.7	0.0	0.8
0.0	0.0	0.0	0.0	0.0	1.7	0.8	0.0	90.8	0.0	0.0	0.0	0.0	3.3	1.7	0.0
0.0	9.2	0.8	3.3	0.0	0.8	0.0	77.5	0.0	0.8	4.2	0.0	2.5	2.5	1.7	0.0
0.8	0.8	0.0	0.0	0.8	0.0	96.7	0.8	1.7	0.0	0.8	0.0	0.0	0.0	0.0	4.2
0.0	0.0	0.0	0.0	0.0	89.2	0.0	0.0	2.5	0.0	1.7	0.0	4.2	1.7	4.2	0.8
3.3	1.7	0.0	0.8	97.5	2.5	0.8	0.0	0.0	0.8	0.0	0.8	0.0	0.0	0.8	1.7
0.0	2.5	0.0	90.8	0.0	0.8	0.0	1.7	0.0	0.0	1.7	0.8	0.0	1.7	0.0	0.0
0.0	0.0	97.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.7	0.8	0.0
0.0	69.2	0.8	0.8	0.8	0.8	0.0	5.0	0.0	0.8	1.7	0.8	0.0	0.0	0.0	0.0
90.0	0.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.8	0.0	0.8

Response

Percent
75
50
25
0

Presented category


* Observer more frequently
**
***
* Network more frequently
**
***
VGG 16
GoogLeNet
AlexNet
p high = 86.6%

* Observer more frequently
**
***
* Network more frequently
**
***

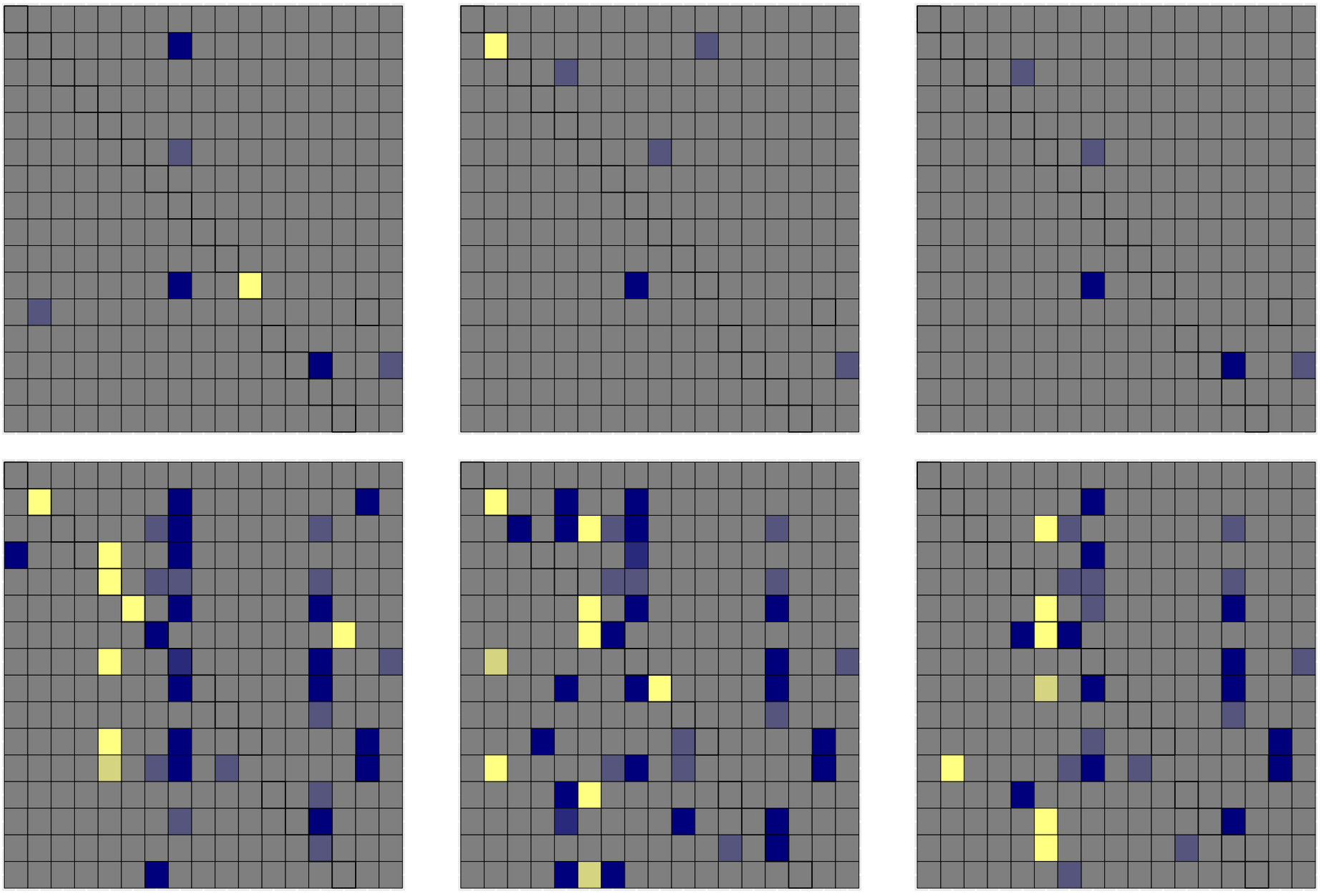
VGG 16

GoogLeNet

AlexNet

p high = 86.6%

p threshold = 47.6%



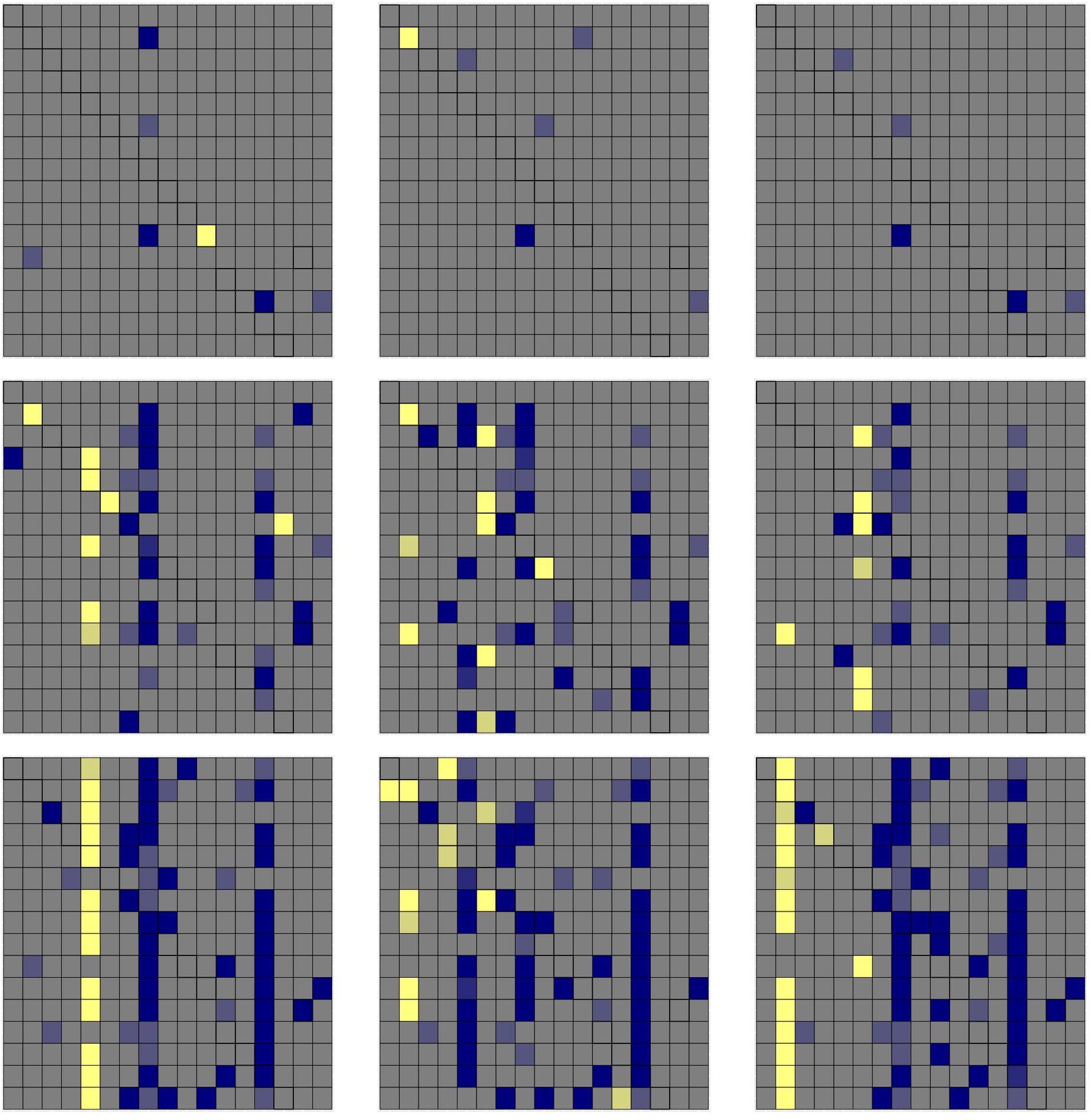
* Observer more frequently
**
***
* Network more frequently
**
***

VGG 16 | GoogLeNet | AlexNet

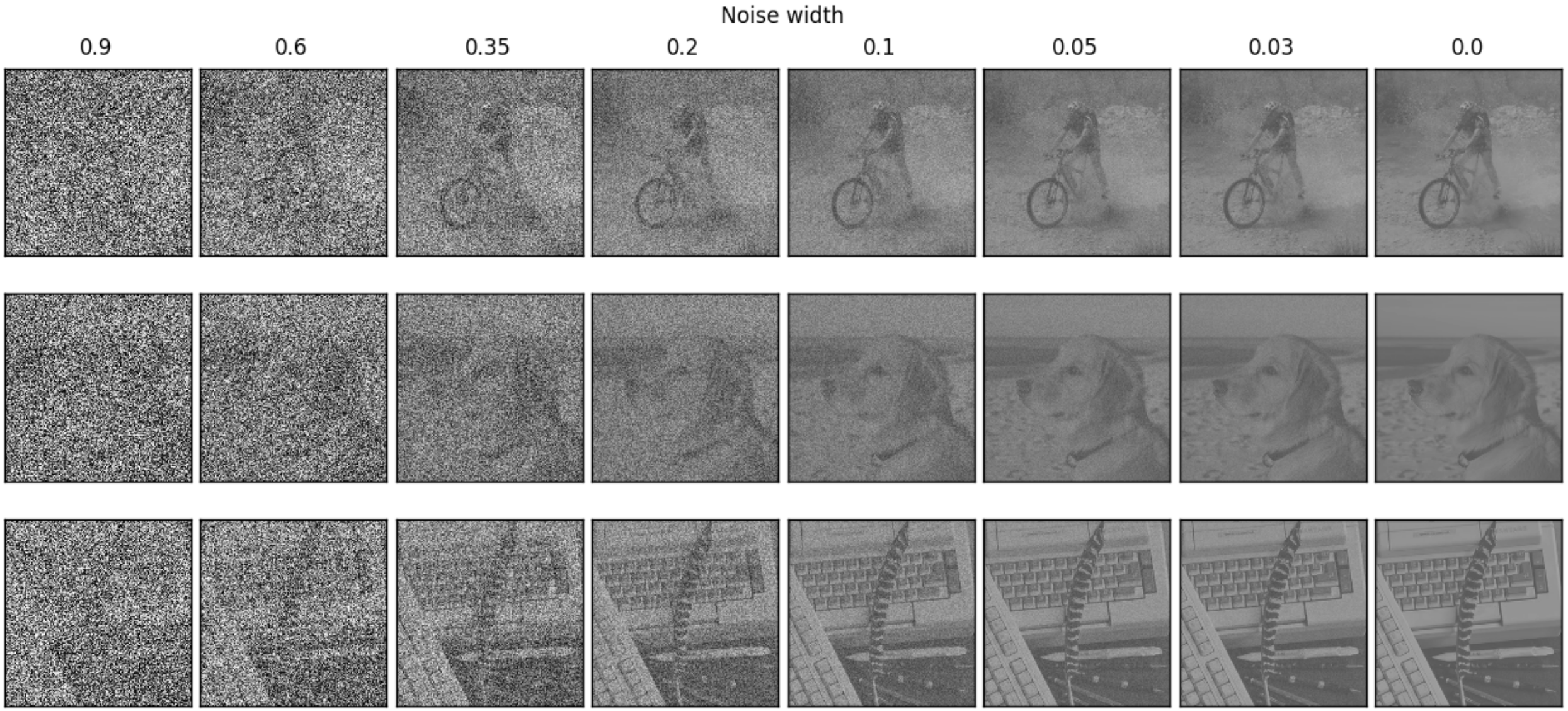
p high = 86.6%

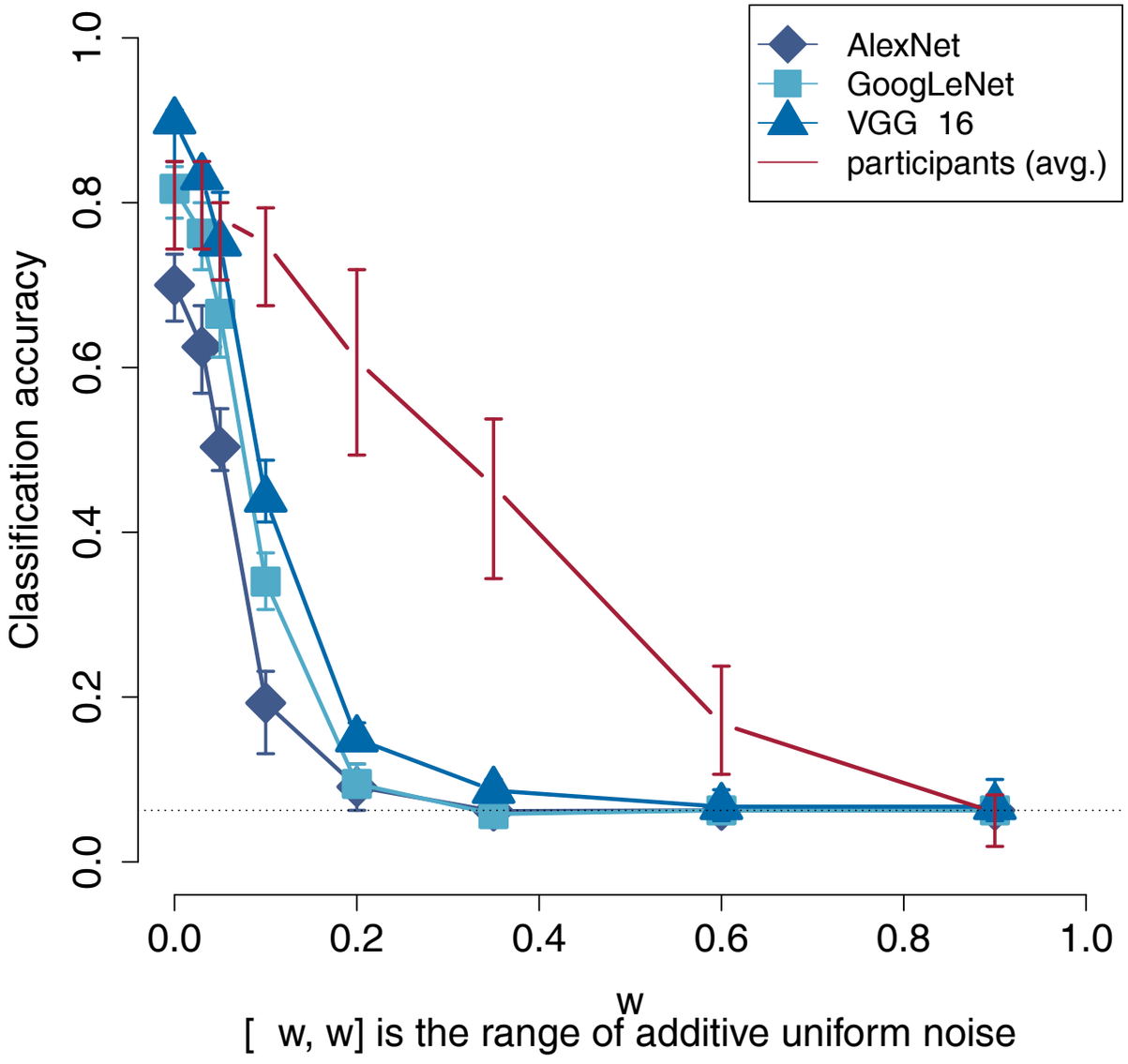
p threshold = 47.6%

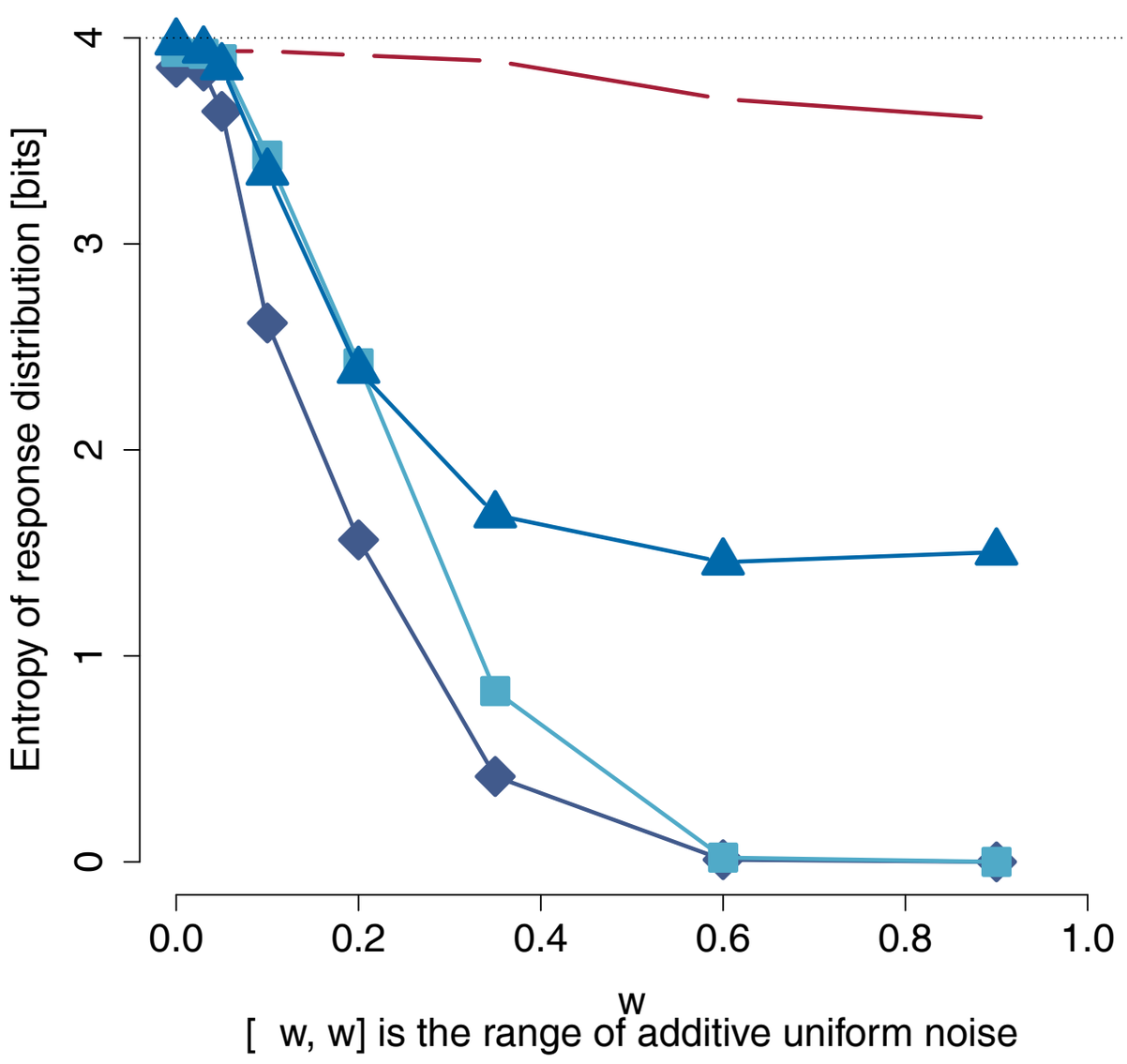
p low = 20.8%



Additive uniform noise

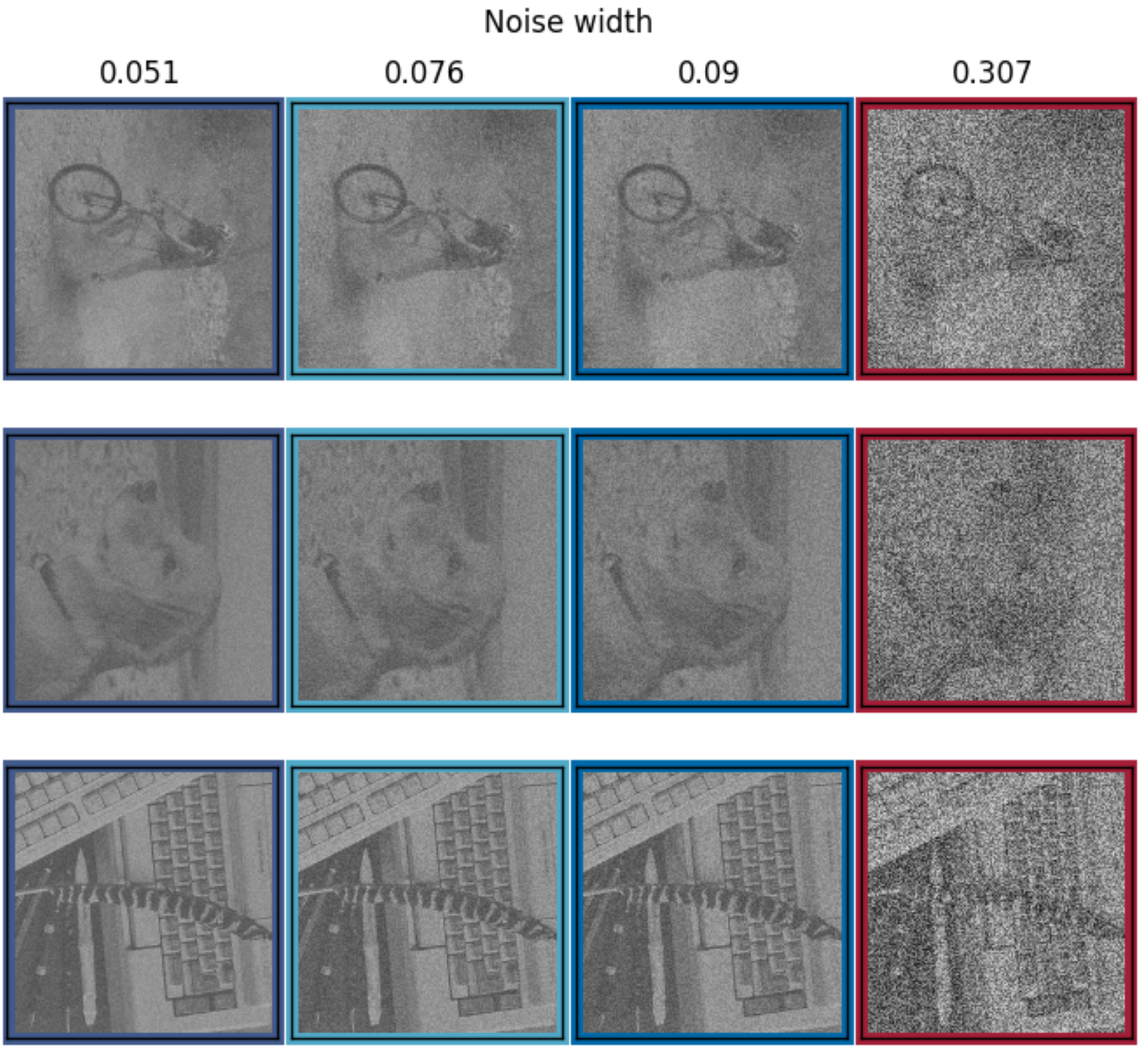


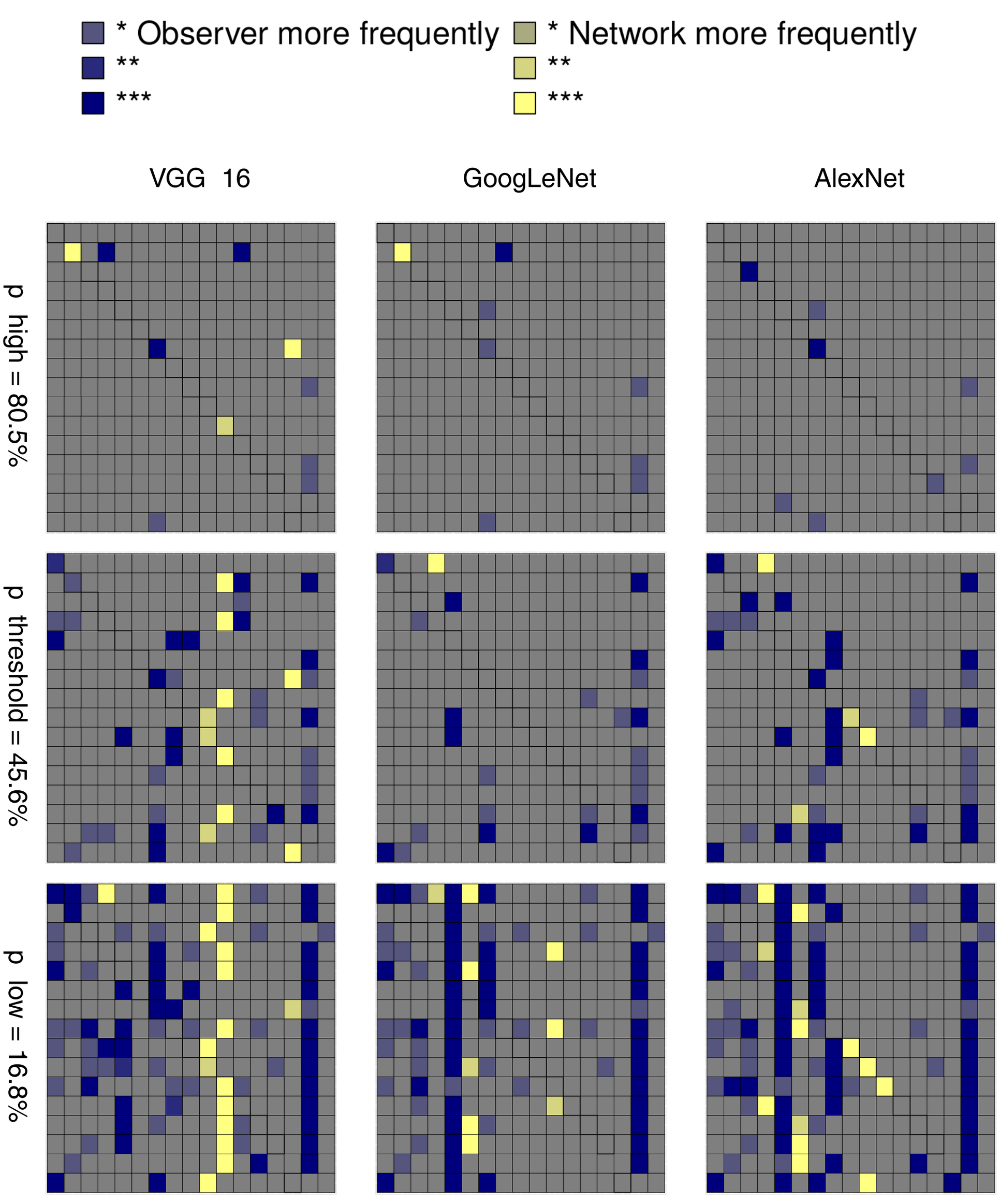




Geirhos et al. (2017)

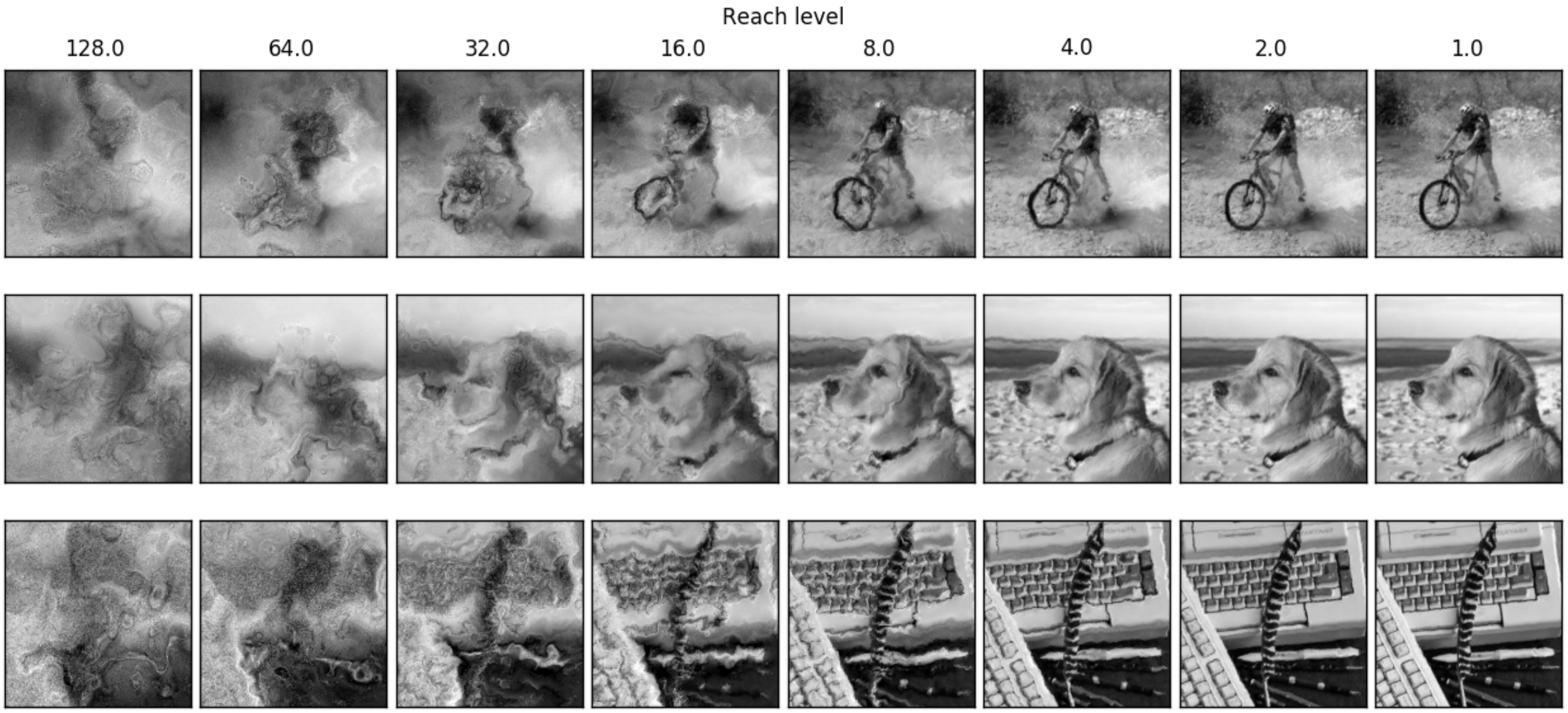
Geirhos et al. (2017)

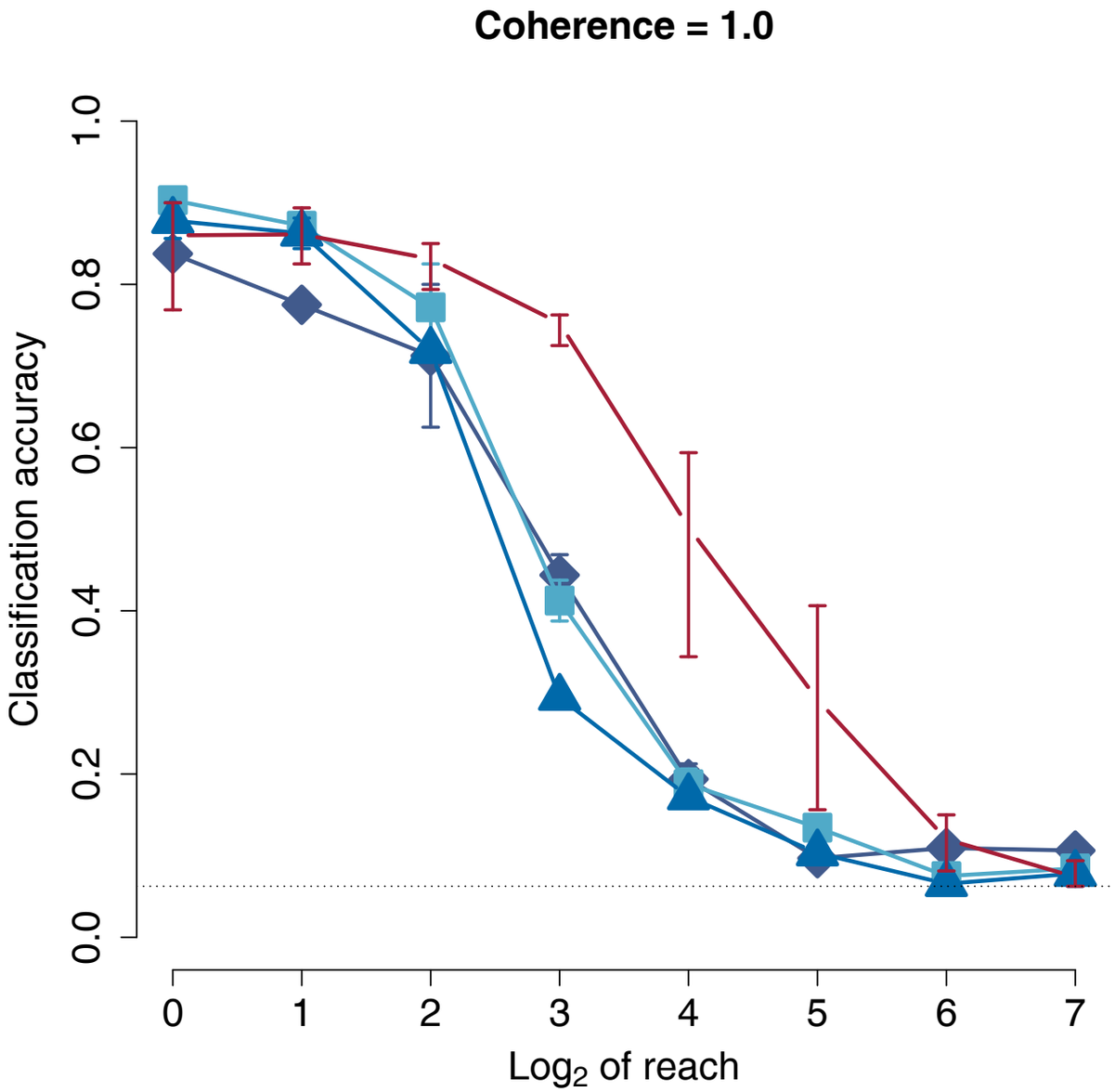


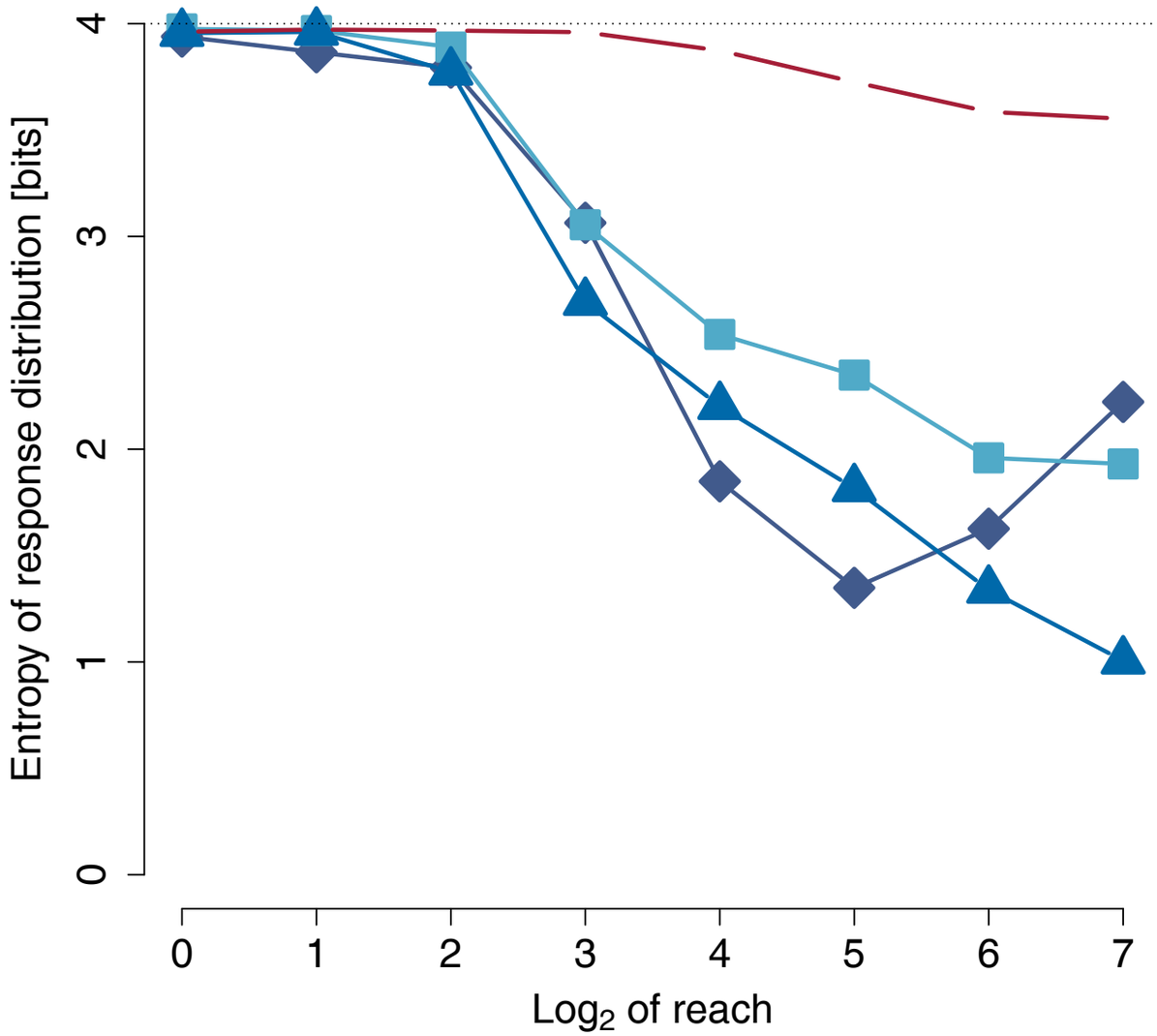

* Observer more frequently
**
***
* Network more frequently
**
***
VGG 16
GoogLeNet
AlexNet
p high = 80.5%
p threshold = 45.6%
p low = 16.8%

Simple fix through data augmentation?

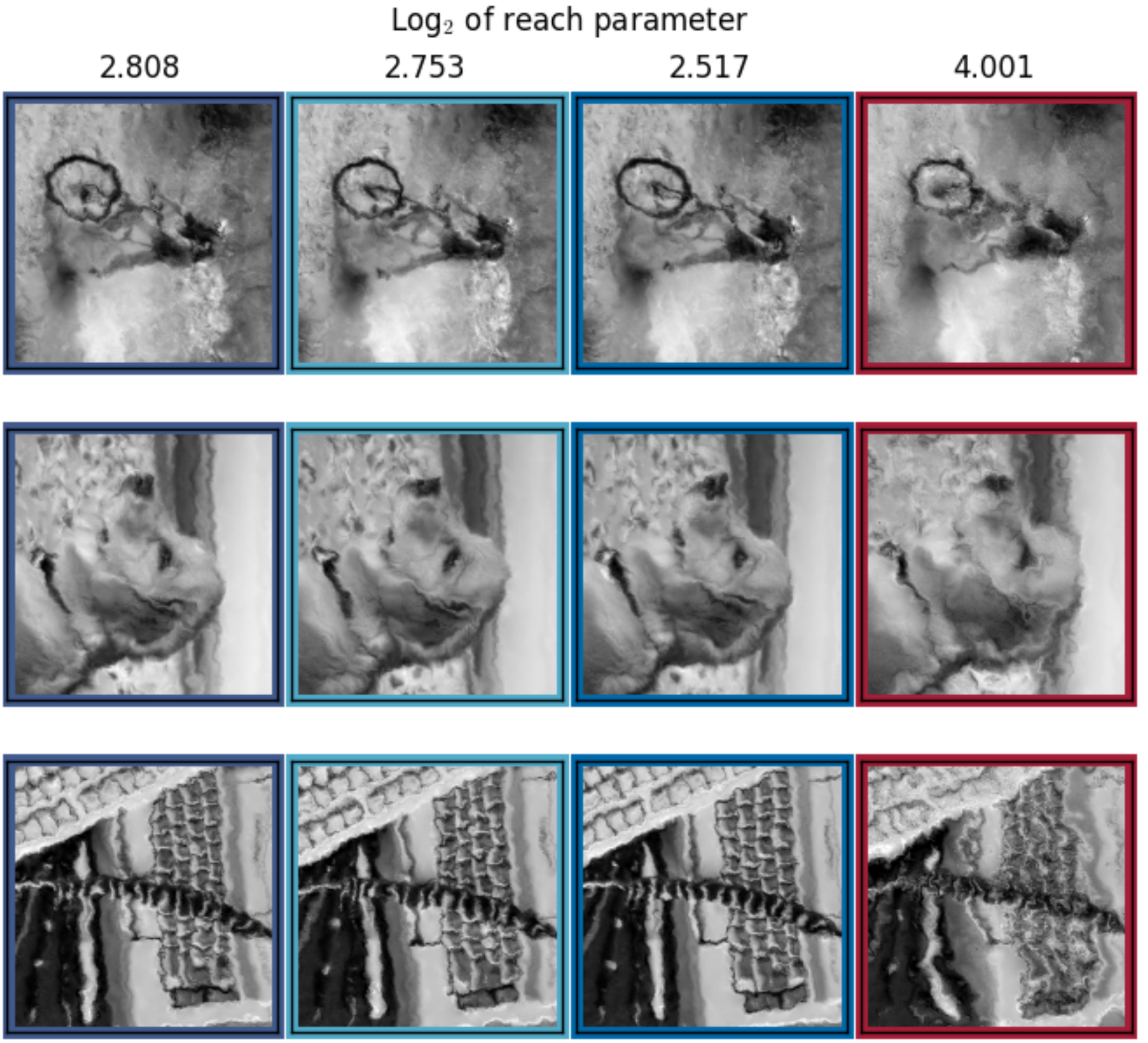
Eidolon distortions (maximal coherence)







Geirhos et al. (2017)


Log2 of reach parameter
2.808
2.753
2.517
4.001

Comparison of DNNs against human observers: object recognition when signals get weaker

For colour images AlexNet, GoogleLeNet and VGG-16 are better than human observers at a **16-category core object recognition task** under "feedforward-only" psychophysical conditions (~96% versus 88% correct).

Human observers are, however, more robust to:

1. **Contrast reduction**
2. **Visual noise (both additive uniform noise as well as random pixel flips)**
3. **Eidolon distortions (from maximal to zero coherence)**

Furthermore, confusion difference matrices show that all tested DNNs and human observers diverge in their recognition behaviour with weaker signals.

Comparison of DNNs against human observers: object recognition when signals get weaker (cont'd)

True for additional experiments exploring image rotations, false colours, power spectrum equalisation, phase noise, low- and high-pass filtering: most often human observers more robust, always diverging response entropy, i.e. differing error patterns when the task gets more difficult (low performance, weak signals; Medina-Temme et al., in preparation).

Claims about strong behavioural—and implied algorithmic—similarities between current DNNs and human observers appear somewhat overstated: in vision science (current/standard) DNNs are perhaps **powerful tools to study—rather than models of—**the human visual system.

We show that using non-linearities that include rectification and local contrast normalization is the single most important ingredient for good accuracy on object recognition benchmarks.

Jarrett, K., Kavukcuoglu, K., Ranzato, M. A., and LeCun, Y. (2009). What is the best multi-stage architecture for object recognition? *IEEE International Conference on Computer Vision*, p. 2146.

Thoughts and speculations

i. Local gain control: ubiquitous in all (?) biological sensory system: **Normalization as a canonical neural computation** (Carandini & Heeger, 2012)

ii. "Hand-crafted" early vision model with only seven free parameters and divisive contrast-gain control predicts a lot of vision data (Schütt & Wichmann, 2017).

iii. Local normalization known to be useful in the context of DNNs (Jarret et al., 2009; c.f. Ren, Liao, Urtasun, Sinz, Zemel, 2016, arXiv): necessary ingredient to achieve more similarity between DNNs and biology?

iv. Our results show we must go beyond prediction performance when evaluating computational models as models of human vision: e.g. response entropy and confusion difference matrices.

v. Striking similar performance of AlexNet, VGG-16 and GoogLeNet when probed with weaker signals, despite very large architectural differences: why?

Thanks

Felix Wichmann

Neural Information Processing Group and
Bernstein Center for Computational Neuroscience,
Eberhard Karls Universität Tübingen

Max Planck Institute for Intelligent Systems, Tübingen