An anatomically-inspired model of visual cortex suggests that the inverted face effect starts at the level of V1

Gary Cottrell

Computer Science and Engineering
University of California, San Diego
Gary's Unbelievable Research Unit (GURU)

UCSDCSE
Computer Science and Engineering

NSF

Euclidean coordinates are the wrong prior for models of primate vision

Gary Cottrell

Computer Science and Engineering
University of California, San Diego
Gary's Unbelievable Research Unit (GURU)

UCSDCSE
Computer Science and Engineering

NSF

CNN Priors: Properties of the visual world

Four important properties:

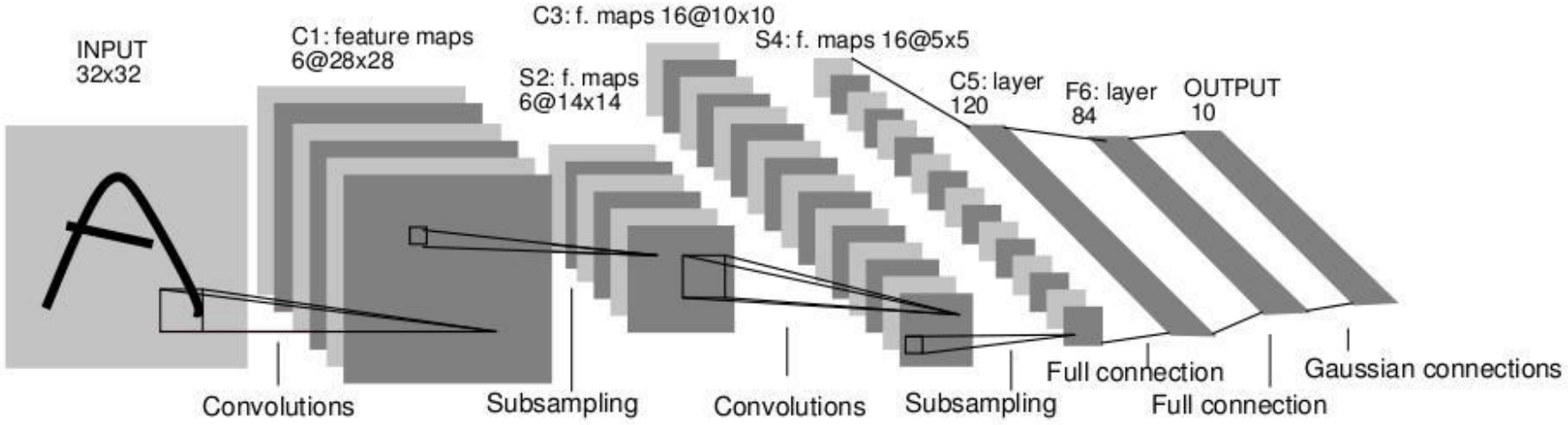
1. Nearby pixels correlate the most with nearby pixels – not pixels far away: *locality*
2. If a feature is useful in one place, it's useful in others (*stationary statistics of filters*)
3. Identity of an object (usually) doesn't depend on its location in the image: *translation invariance*
4. Objects are made of parts: *Compositionality*

Convolutional Neural Networks

Began with LeCun et al. 1989

(actually, for backprop nets, it began with Chapter 8 of the PDP books (1986)
"The T-C problem" pp. 348-352, Rumelhart, McClelland, and the PDP group.
And they reference Fukushima, 1980)

1. Small, local receptive fields and learned features (kernels): locality
2. These are replicated across the image: If a feature is useful in one place, it's useful in others
3. Convolution + Spatial pooling: *translation invariance*
4. Objects are made of parts*:* receptive fields get larger deeper in the net



Convolutional Neural Nets

- Are currently the best model we have for the primate visual system (at least the ventral-temporal lobe part of it).

- But there's something fundamentally wrong with them as models of primate vision!

What is the right prior for vision models?

- One thing not considered: Resolution of the image
  - CNN's have high resolution everywhere
  - With translation invariance, this is good for recognizing objects at any location in the image, but…
  - Fixate on the cross, then tell me what word appears on the right:

+ word

  - So much for high resolution + translation invariance!
  - We only have high resolution in the fovea (about the width of your thumbnail at arm's length) – while CNNs have high resolution everywhere.

Outline

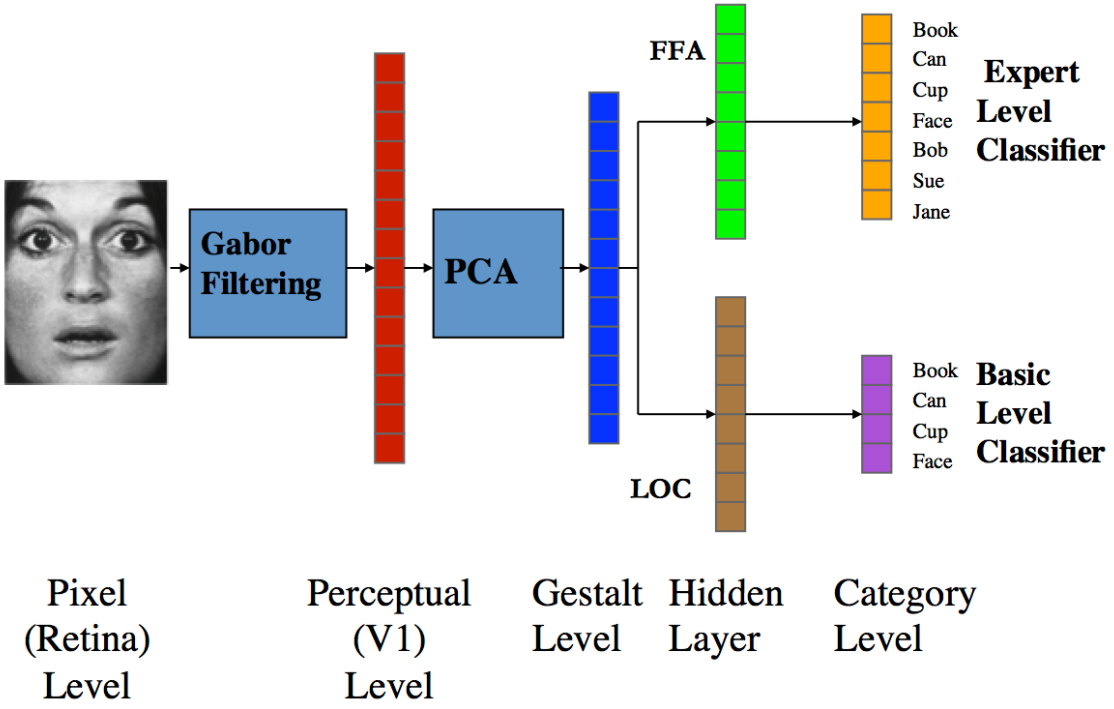
- **The Model (briefly)**
- A (more) anatomically-inspired version:

  The Model 2.0

- How The Model 2.0 accounts for the Face Inversion Effect

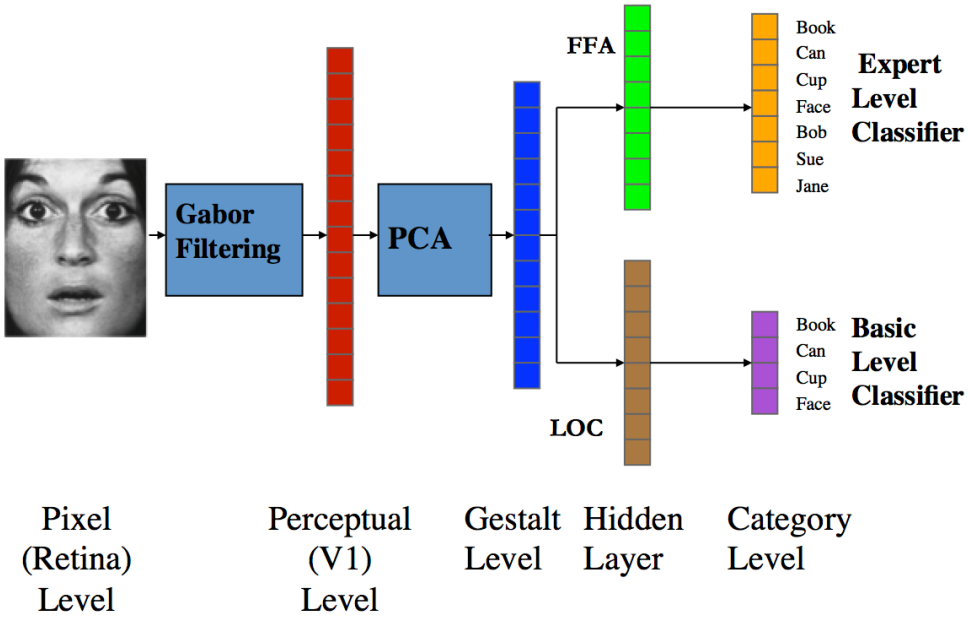
The Model ™

Developed over the last 33 years or so…



- Starts with an actual image as input (now called "image computable")
- Applies a convolution by multi-scale Gabor filters
  (hey! It's convolutional!)
- Does a form of statistical pooling: PCA (can be learned by a simple autoencoder)
  (hey! It does pooling!)
- Hidden layers learn features in the service of the task
  (hey! It's deep!)

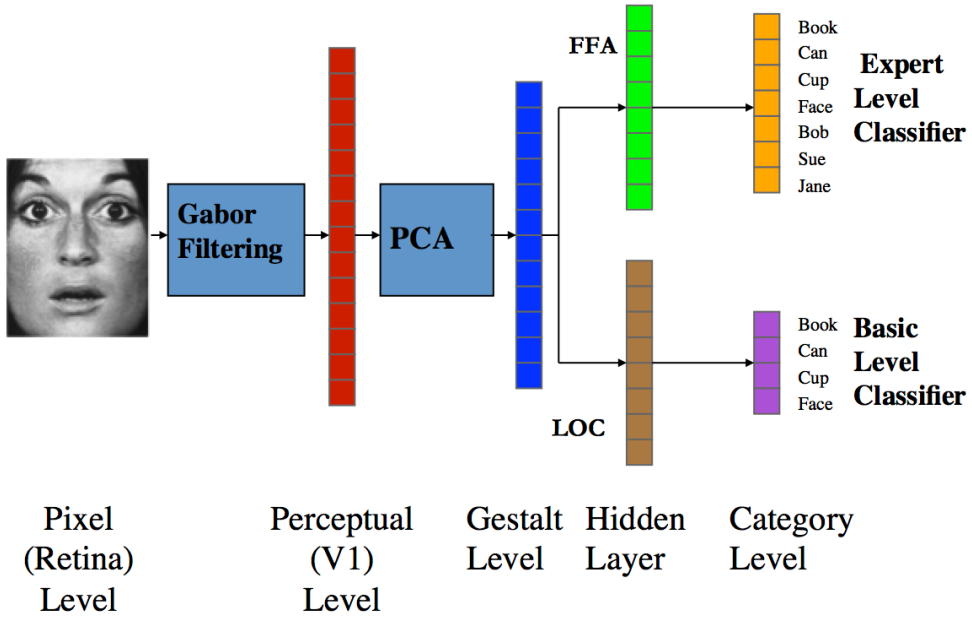
The Model ™



Approach:

- Train The Model to do the same tasks people do (recognize faces, emotions, objects and letters) (now called a "task-driven" vision model)
- Use the model to explain behavioral, developmental, and imaging results (without fitting parameters!)
- Example: Why the Fusiform Face Area is recruited for other tasks (Tong, M.H., Joyce, C.A., and Cottrell, G.W. (2008) *Brain Research*).

The Model ™



Lots of data accounted for...

Outline

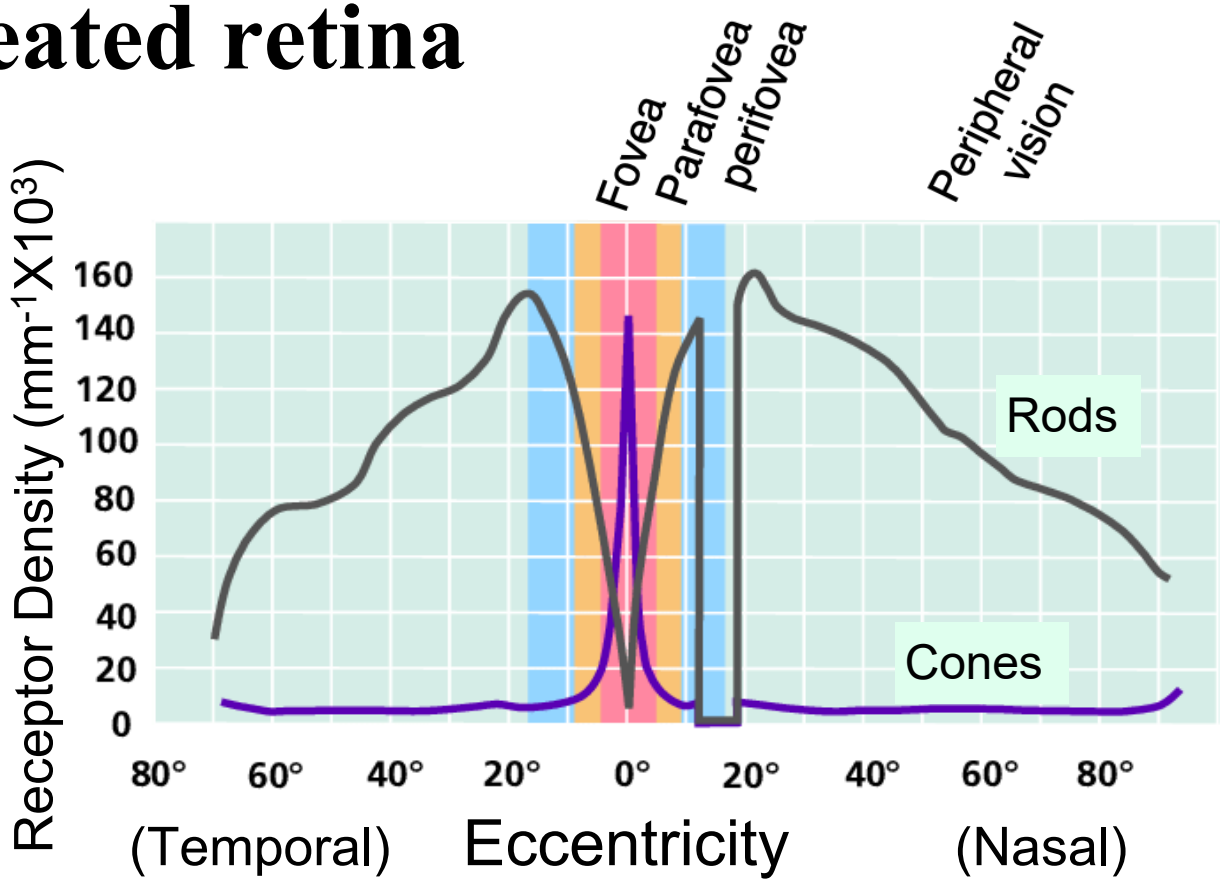
- The Model (briefly)
- **A (more) anatomically-inspired version:**

  The Model 2.0
- How The Model 2.0 accounts for the Face Inversion Effect

The Model 2.0

- Deeper
- More anatomical constraints
  - Foveated retina
  - Log-polar mapping from the visual field to V1
  - Central and Peripheral pathways
  - Sampling of the image using salience

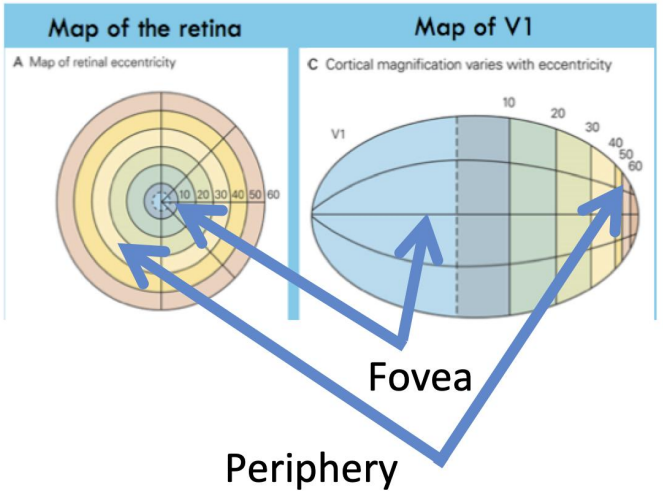
Anatomical Constraints: The foveated retina



The fovea has a high density of photoreceptors, which drop off in the periphery

As a result, we move our eyes about 172,000 times a day

Anatomical Constraints: The log-polar transform



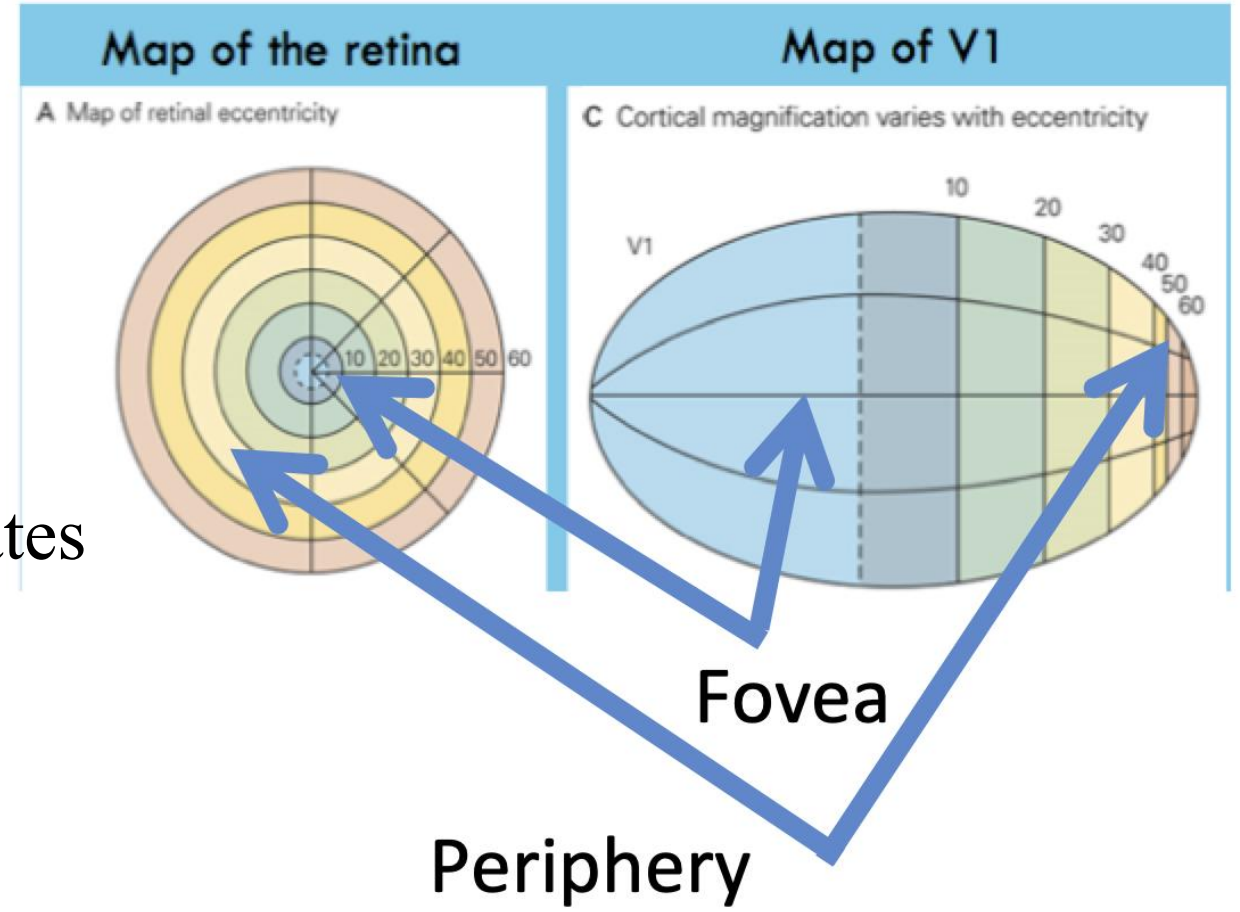
- There is a log-polar transform from the visual field to V1
- This separates central from peripheral representations
- The result is a huge "cortical magnification" of the central visual field representation - and a corresponding shrinkage of peripheral vision
- (This is not *really* magnification: It just results from there being more receptors in the fovea, fewer in the periphery)

Anatomical Constraints: The log-polar transform

- The mapping from the visual field to V1 is approximately log-polar

  $(\log r, \theta)$ coordinates

- This separates central from peripheral representations

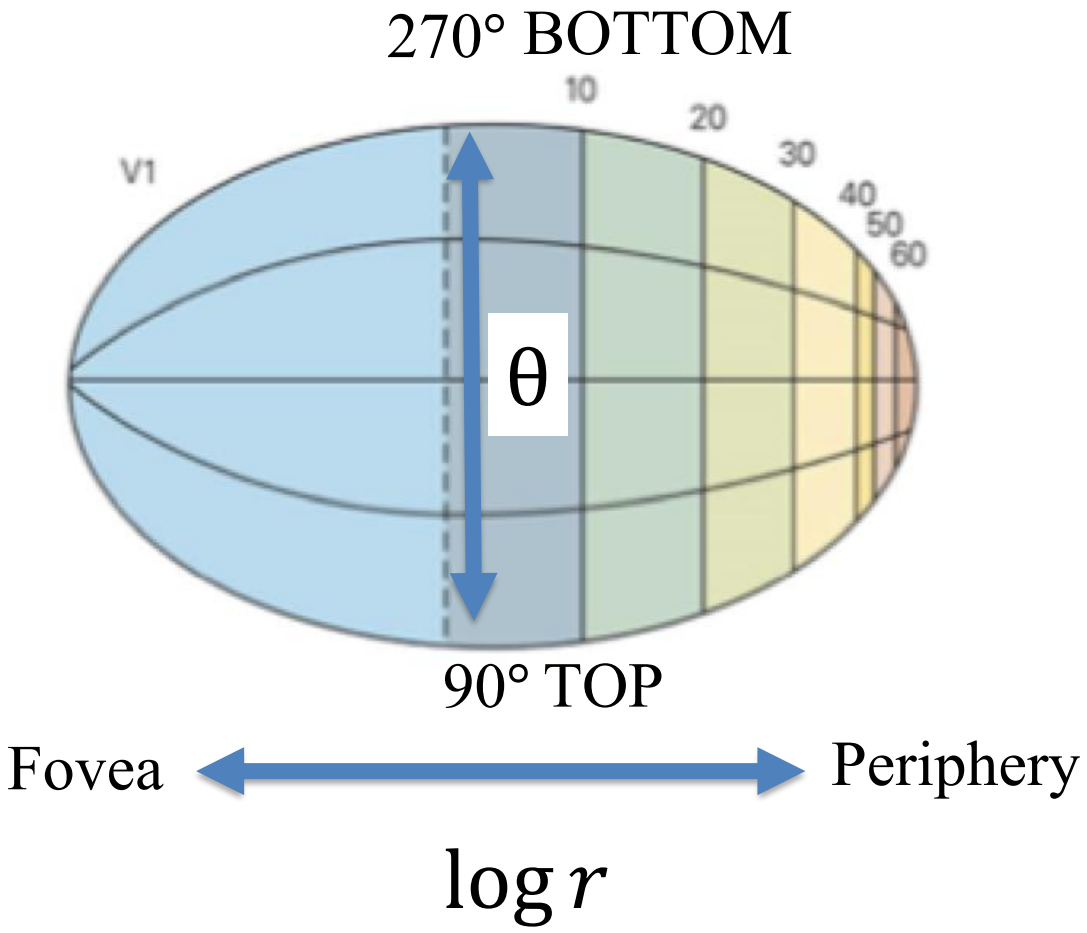


Anatomical Constraints: The log-polar transform

- The mapping from the visual field to V1 is approximately log-polar

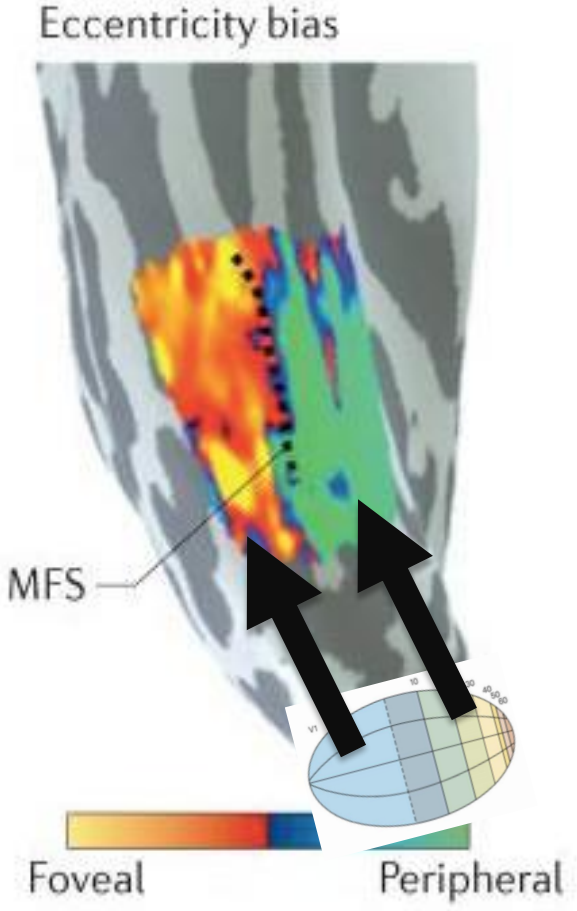
  $(\log r, \theta)$ coordinates

- This separates central from peripheral representations



Anatomical Constraints: The central and peripheral pathways

- Now the map is “formatted” for two pathways...
- The central pathway innervates the ventral stream (Fusiform Face Area, Lateral Occipital Cortex, IT)
- The peripheral pathway innervates the Parahippocampal Place Area



Picture credit:
Grill-Spector & Weiner, 2014

Advantages of Log-Polar Representation

- Biologically plausible
- Still retinotopic – standard convnets can process it
- Scale and Rotation equivariant – just a translation

Scale Equivariance

Geoff at multiple scales

Geoff log-polarized at multiple scales

I find the hoops CV people jump through to achieve this (image pyramids, three networks at different scales) rather baroque..

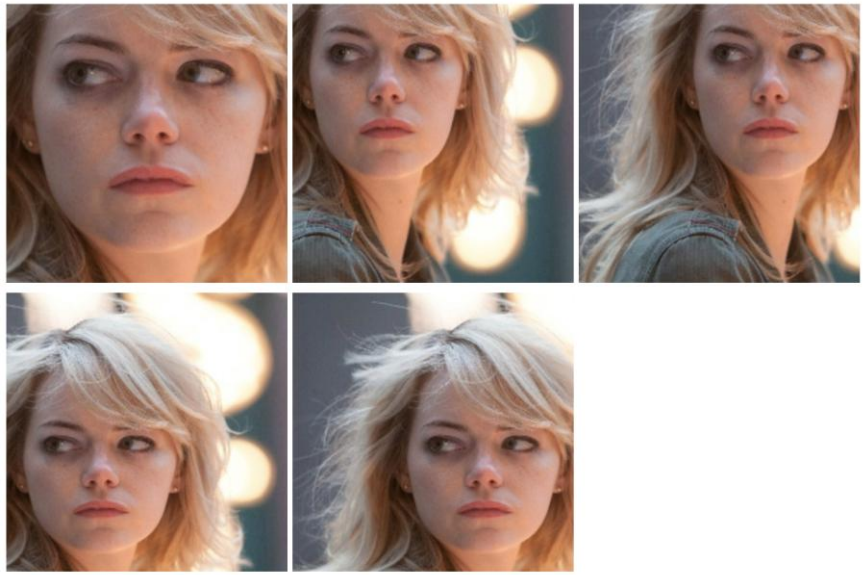
Rotation Equivariance

- Rotation is just a shift up and down

- When these are fed to a standard convnet – which is translation invariant (mostly) – you get rotation and scale *invariance*

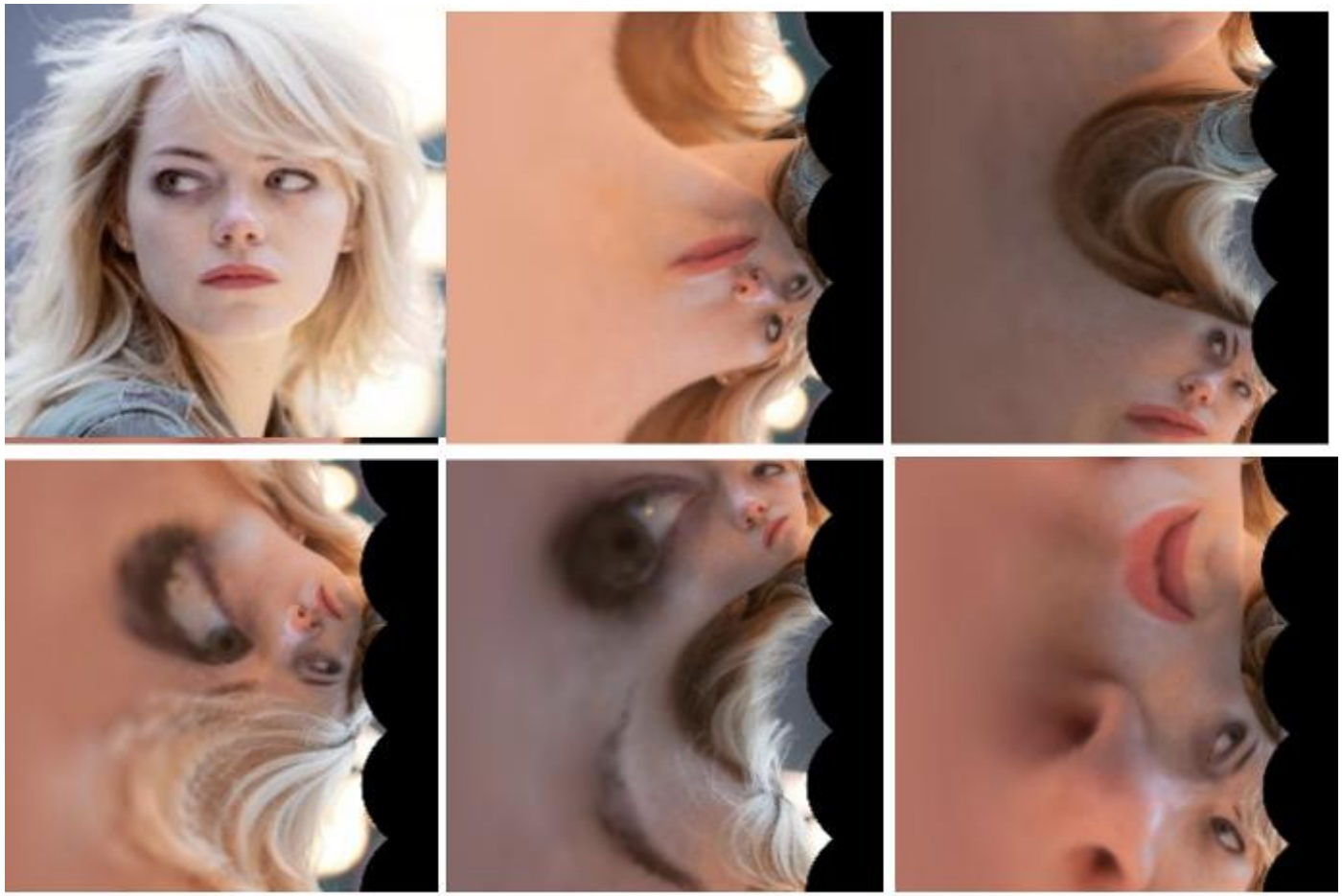
But *translation* invariance is lost

- We make up for this by "fixating" the image at different points (learning appearance at different "translations")
- Top: standard crops for training a convnet
- Bottom: Sampling the image with different fixations - eyes, mouth, nose



But *translation* invariance is lost

Note that this amounts to *self-augmentation* of the data!



Outline

- The Model (briefly)
- The Model 2.0
- **How The Model 2.0 accounts for the Face Inversion Effect**

The Face Inversion Effect

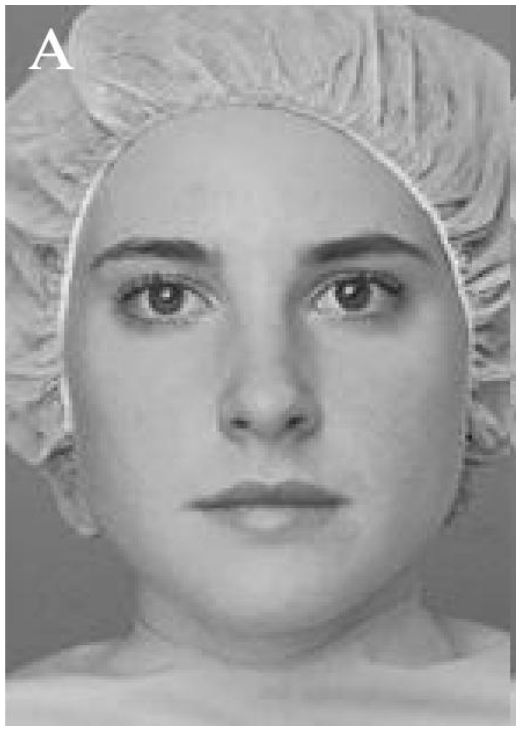
- Faces appear "strange" when inverted (Köhler, 1940)
- Köhler devised a clever experiment to test whether it was the *relationship to the environment* or *the projection on the retina*:
- He had an assistant hold a picture of a face right-side up while he looked at it through his legs.
- Result: It still looked "strange"!
- Inverted, it looked normal
- Conclusion: It's the projection on the retina, not the relation to the surround.

The Face Inversion Effect

- There's been some other work since…;-)
- Yin (1969): Faces are more difficult inverted compared to inverted mono-oriented objects (houses), which are more difficult than objects
- Diamond and Carey: It's an expertise effect:
  - Dog show judges are similarly disrupted by inverted pictures of dogs
  - Non-experts actually performed *better* in some experiments than dog show judges on inverted pictures of dogs
- Farah & Tanaka (1995): It's due to holistic (configural) processing
- Yovel & Kanwisher (2005): It happens in the FFA, not earlier

The Face Inversion Effect
Could it start in V1???

- **Face recognition is an example of visual expertise:**
  - **Fine-level discrimination of homogeneous categories**
  - **The arrangement of features (configural information) is critical.**



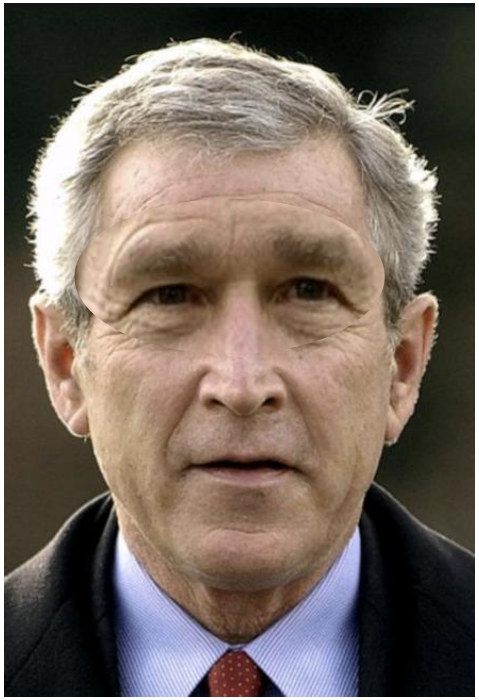
Mondloch, Le Grand, & Maurier, 2002

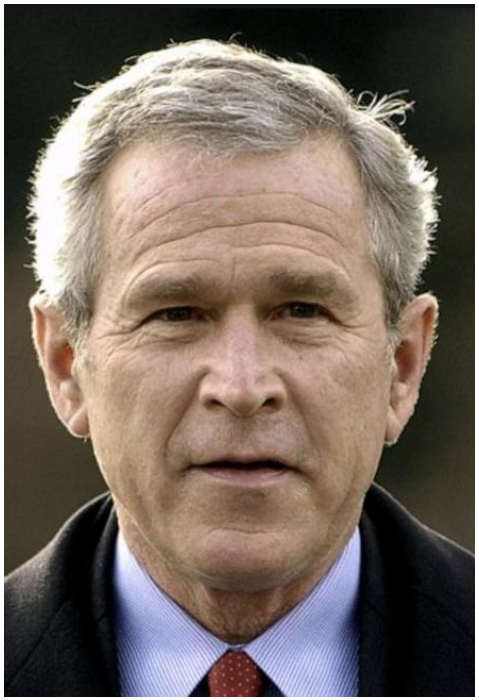
A test - which is the real George Dubya?

A test - which is the real George Dubya (right side up this time)?

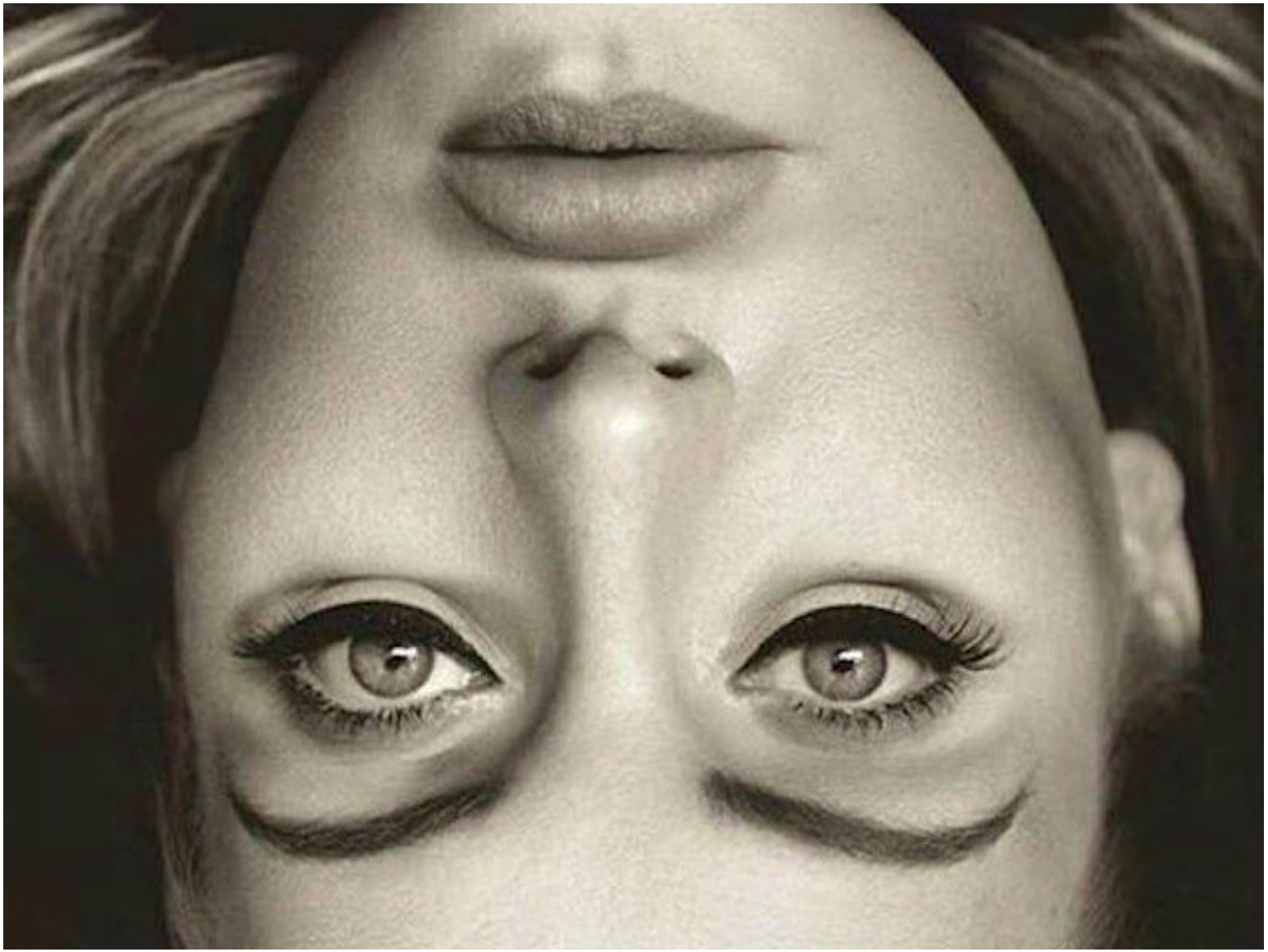
Anatomy Matters: Face Inversion

- **Face recognition is an example of visual expertise:**
  - **Fine-level discrimination of homogeneous categories**
  - **The arrangement of features (configural information) is critical.**

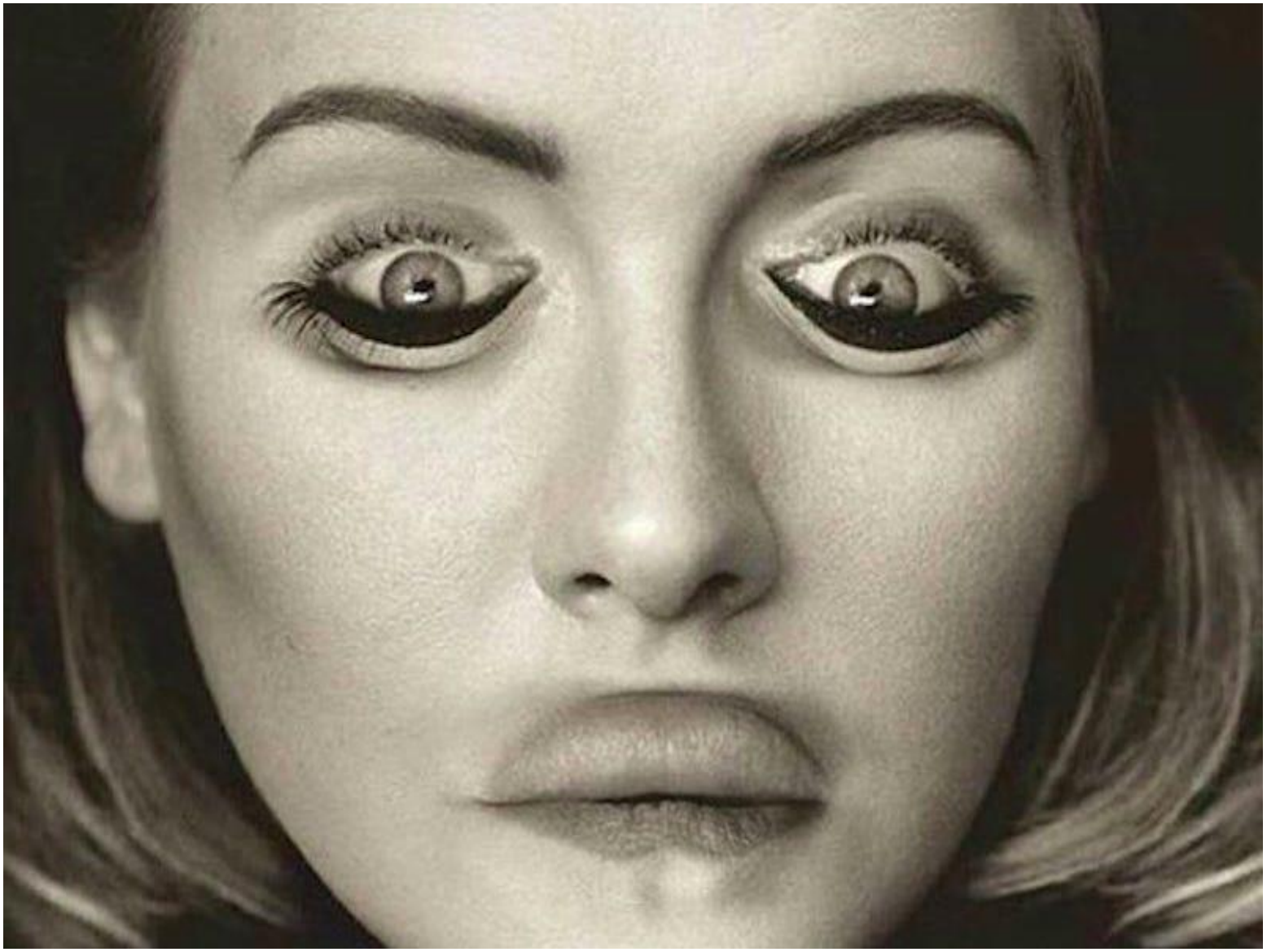




The Thatcher Effect



The Thatcher Effect

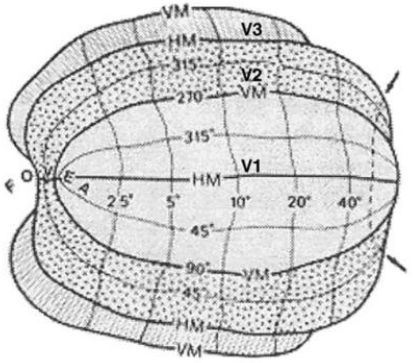


But wait!

- *How could an inversion effect happen when the model is rotation invariant???*
- Answer: There's a topological difference between rotation and scale in the log-polar plane

Rotation vs. Scale Invariance in V1: topologically different!

- In the log-polar representation:
  - Scale is just a shift left or right.
  - Rotation is just a shift up and down
- But there's a fundamental difference!



(Far left) The map from the visual field to V1 starts at the vertical meridian and proceeds clockwise. (Left) The mapping of the visual field starts at 90°, at the bottom of V1, and continues to 270°, at the top of V1 (J.R. Polimeni et al., 2006).

- Your brain is not a torus – so when an input rotates, it "falls off" the top and appears again at the other side, re-arranging the feature configuration – and configuration matters in face recognition!

Rotation Invariance: topologically different!

- In the log-polar representation:
  - Scale is just a shift left or right.
  - Rotation is just a shift up and down
- But there's a fundamental difference!
- When an input rotates, it "falls off" the top and appears again at the other side, re-arranging the features – and configuration matters in face recognition!

Rotation Invariance: topologically different!

- In the log-polar representation:
  - Scale is just a shift left or right.
  - Rotation is just a shift up and down
- But there's a fundamental difference!
- When an input rotates, it "falls off" the top and appears again at the other side, re-arranging the features – and configuration matters in face recognition!

Nose next to left eye

Rotation Invariance: topologically different!

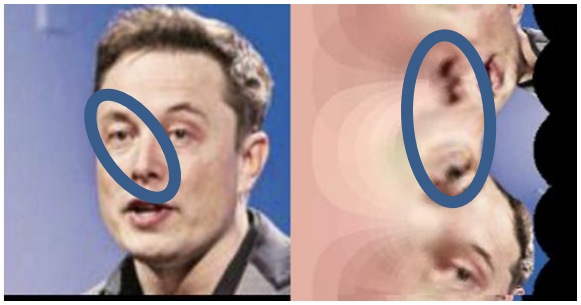
- In the log-polar representation:
  - Scale is just a shift left or right.
  - Rotation is just a shift up and down
- But there's a fundamental difference!
- When an input rotates, it "falls off" the top and appears again at the other side, re-arranging the features – and configuration matters in face recognition!

Nose next to left eye

Small rotation – same configuration, just shifted up.

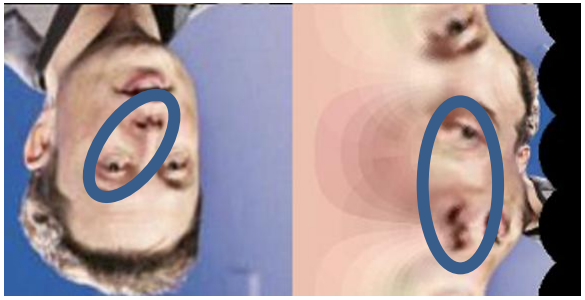
Rotation Invariance: topologically different!

- In the log-polar representation:
  - Scale is just a shift left or right.
  - Rotation is just a shift up and down
- But there's a fundamental difference!
- When an input rotates, it "falls off" the top and appears again at the other side, re-arranging the features – and configuration matters in face recognition!



Nose next to left eye

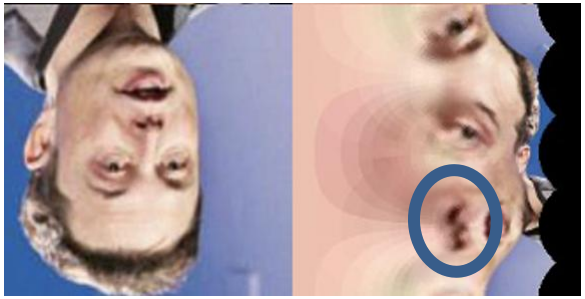
Small rotation – same configuration, just shifted up.



Inversion - Log-polar representation puts the nose next to *right* eye

Rotation Invariance: topologically different!

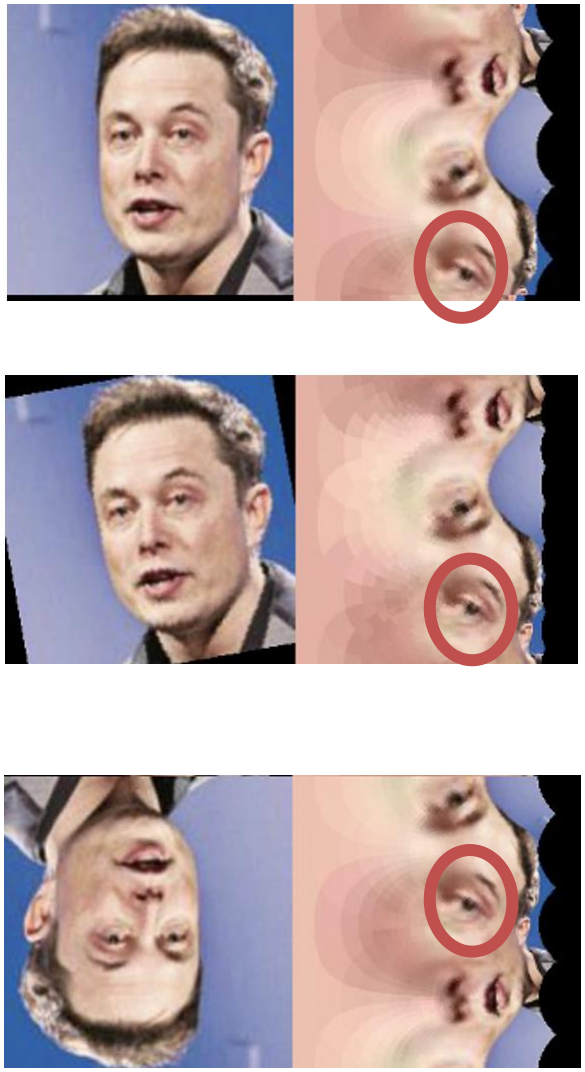
- In the log-polar representation:
  - Scale is just a shift left or right.
  - Rotation is just a shift up and down
- But there's a fundamental difference!
- When an input rotates, it "falls off" the top and appears again at the other side, re-arranging the features – and configuration matters in face recognition!



NOTE here: *The features themselves (eyes, nose, mouth) are identical, just shifted*

Rotation Invariance: topologically different!

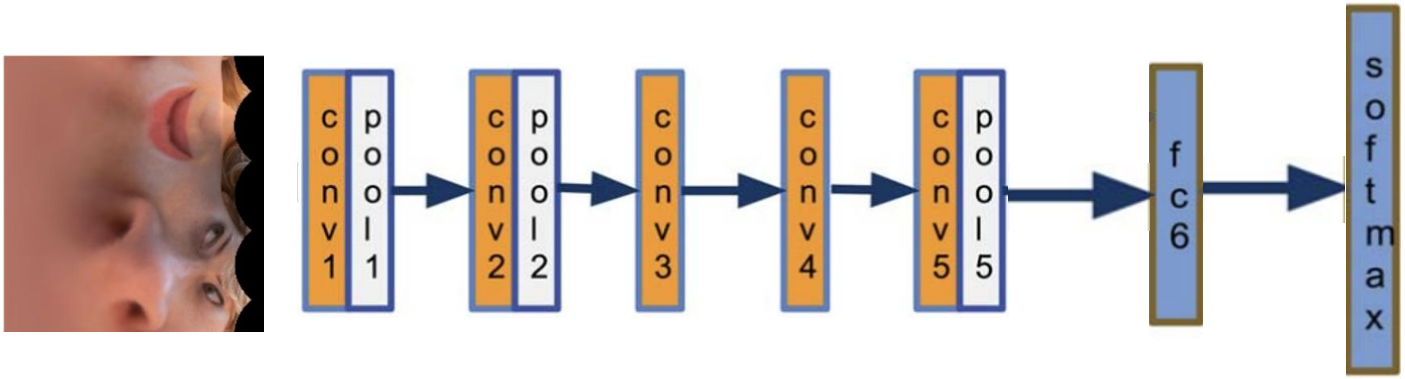
- In the log-polar representation:
  - Scale is just a shift left or right.
  - Rotation is just a shift up and down
- But there's a fundamental difference!
- When an input rotates, it "falls off" the top and appears again at the other side, re-arranging the features – and configuration matters in face recognition!



NOTE here: *The features themselves (eyes, nose, mouth) are identical, just shifted*

Experiments

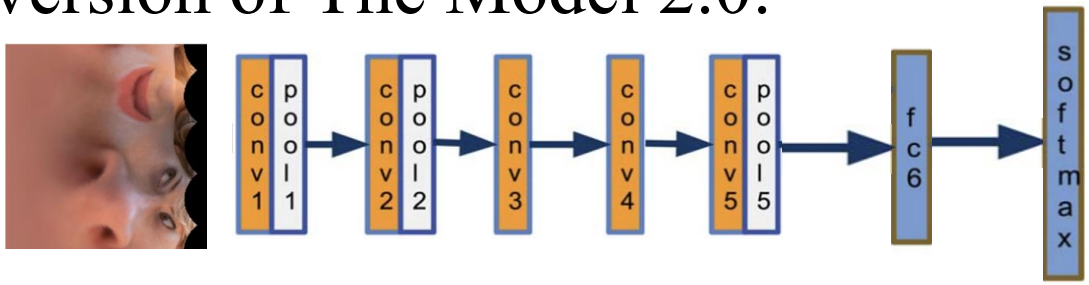
- We used a simplified version of The Model 2.0:
  - Foveated retina
  - Log-polar mapping
  - Dubbed “LPNet”



This is an artist’s conception – we actually used ResNet50

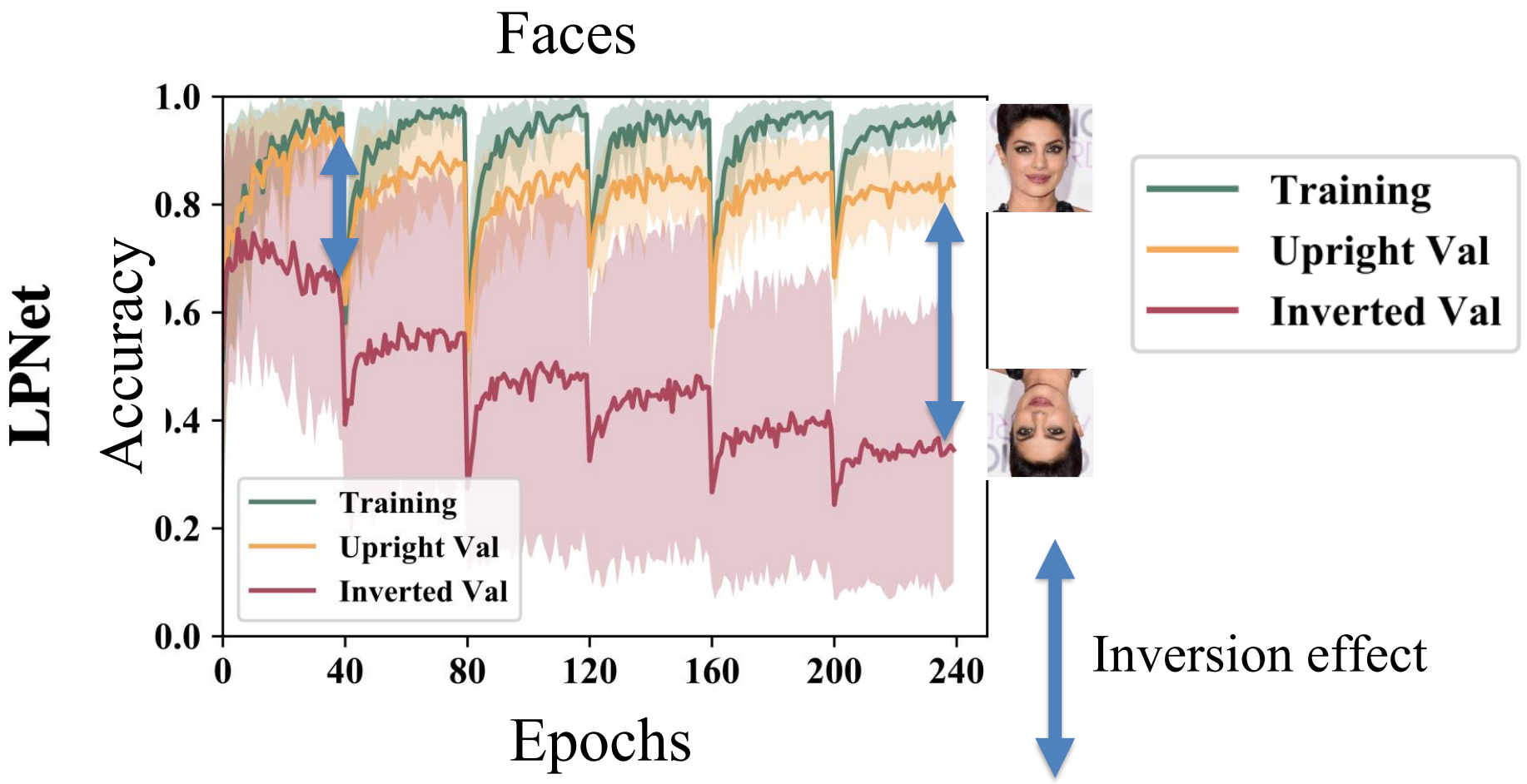
Experiments

- We used a simplified version of The Model 2.0:
  - Foveated retina
  - Log-polar mapping
  - Dubbed “LPNet”



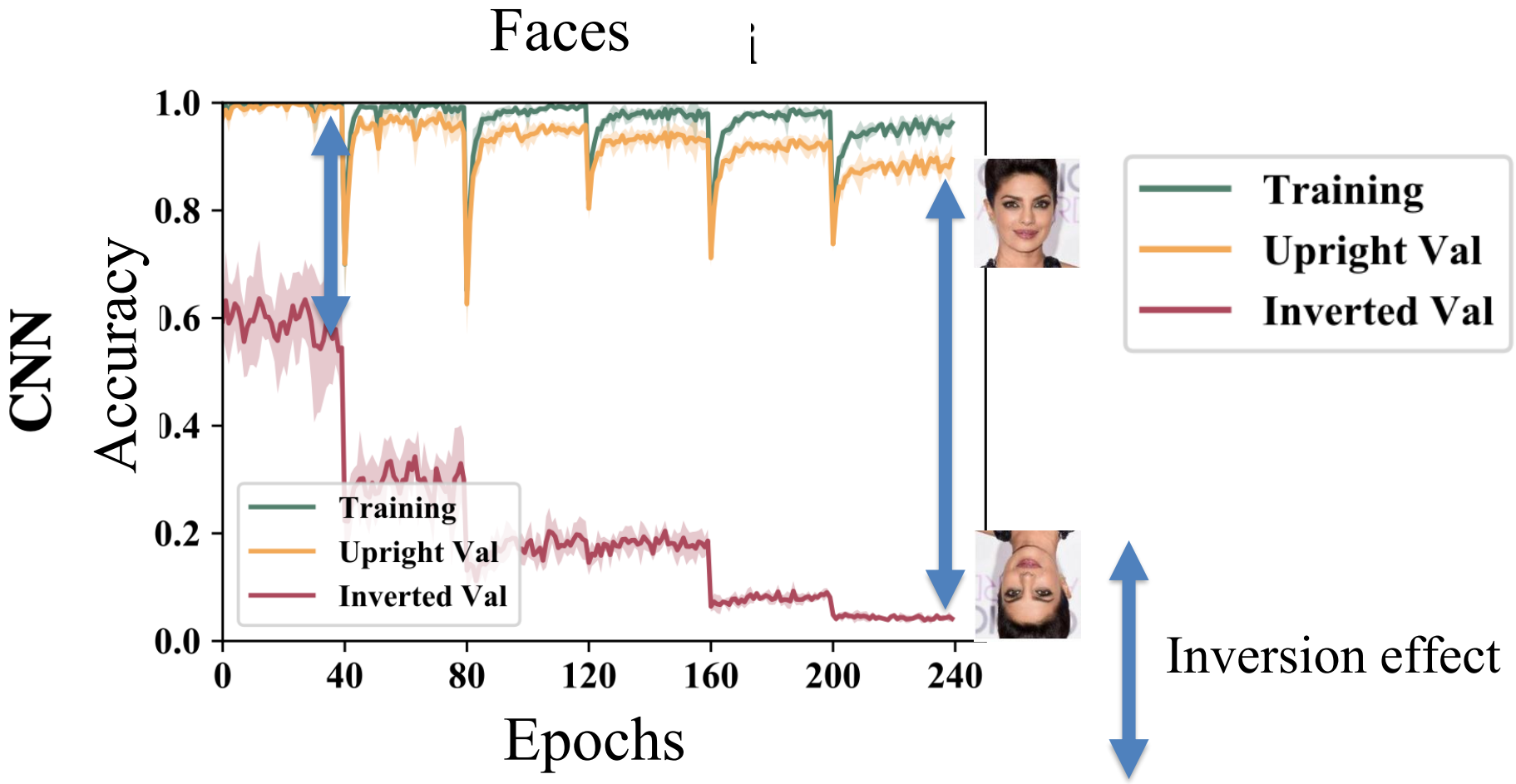
- Then we trained it on face recognition (identity), doubling the identities every 40 epochs in order to simulate increasing exposure to different people over development
- Tested it on upright and inverted held-out faces
- Similarly, trained a “vanilla” CNN (Euclidean coordinates) on the same data

LPNet Results: inversion effect grows with expertise



Inflections represent doubling of the number of identities: 4, 8, 16, 32, 64, 128, simulating experience

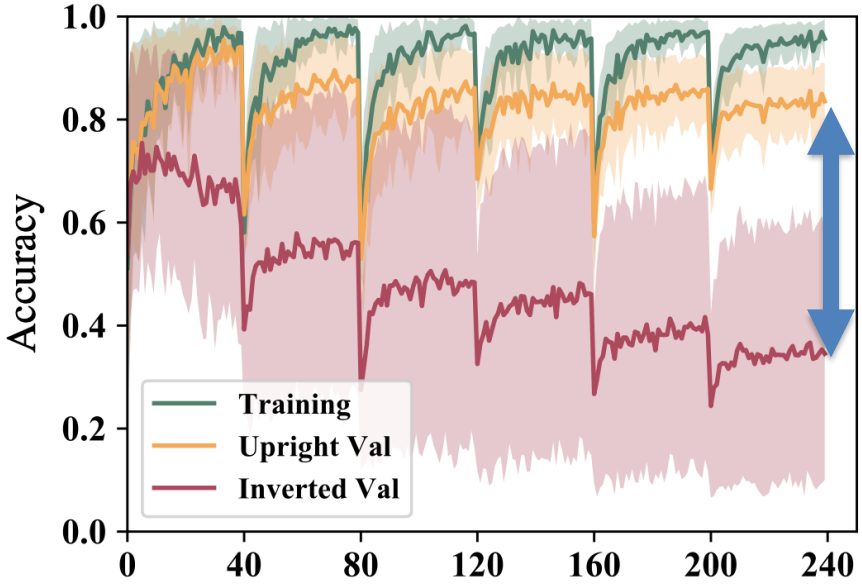
CNN Results: inversion effect grows with expertise



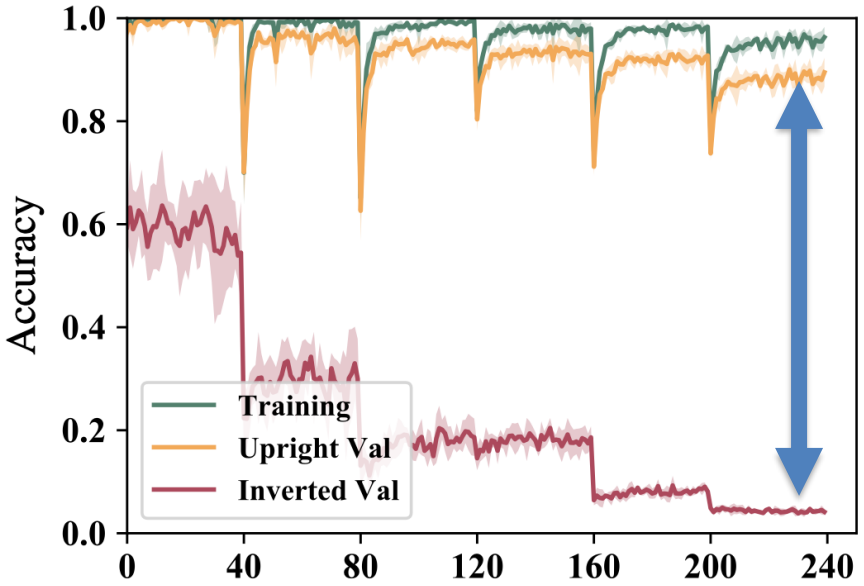
Inflections represent doubling of the number of identities: 4, 8, 16, 32, 64, 128, simulating experience

But note the difference!

LPNet



CNN



The CNN drops to nearly chance!

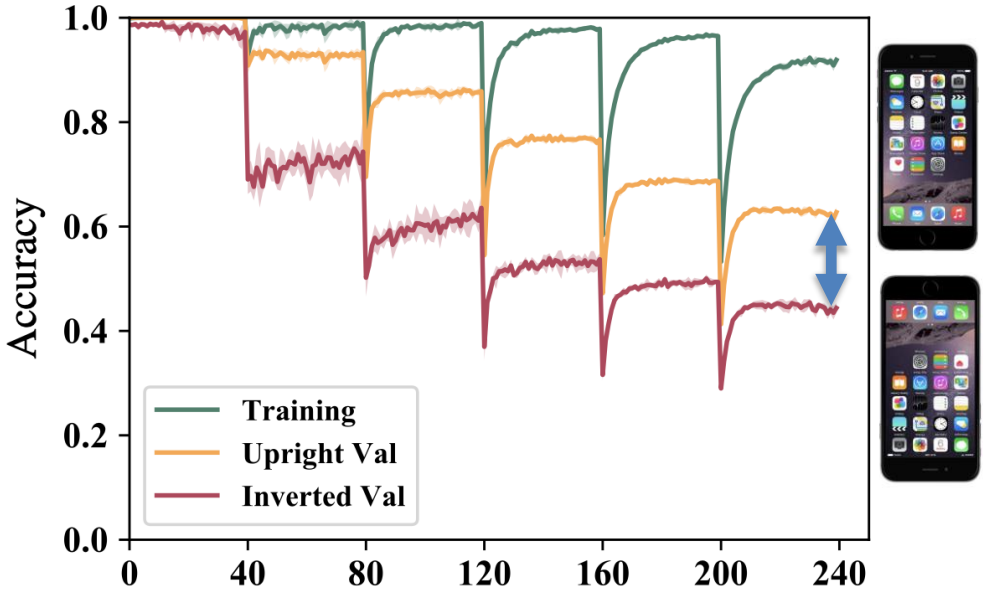
=> CNNs can't handle inversion!

So…LPNet is more like humans…

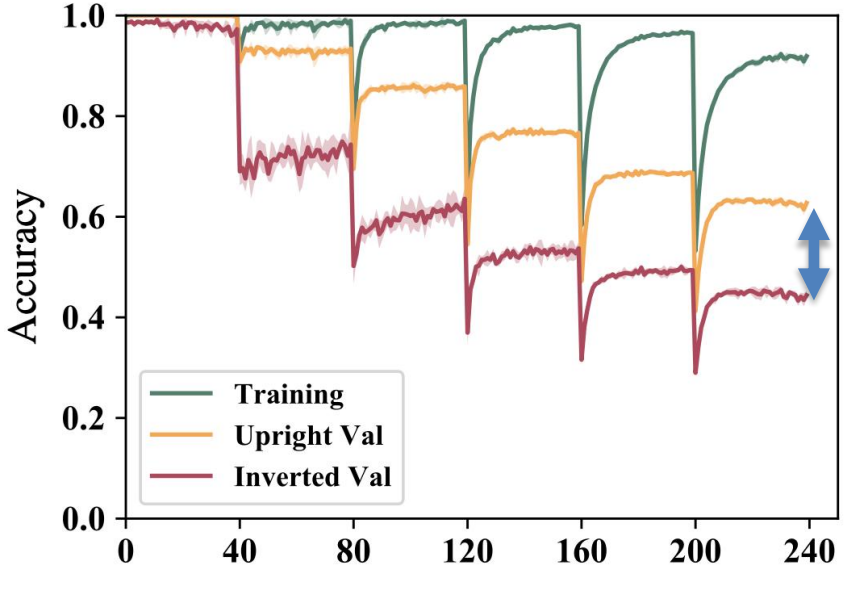
Inversion effect

This inversion effect is *much* less severe for objects

LPNet



CNN



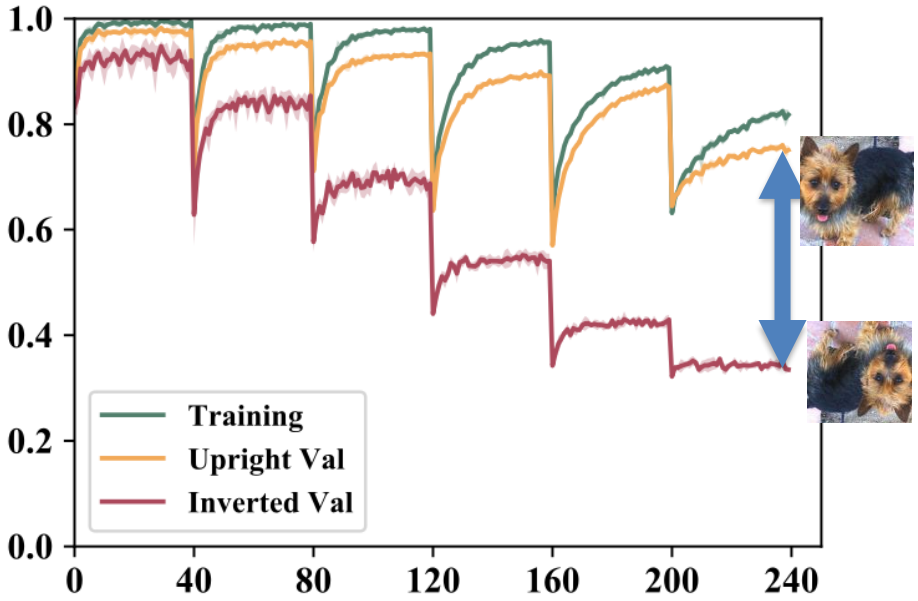
This is consistent with Yin's (1969) results

Objects are a *basic level* category:
Suggests that the network is just using features,
not *configuration* of features

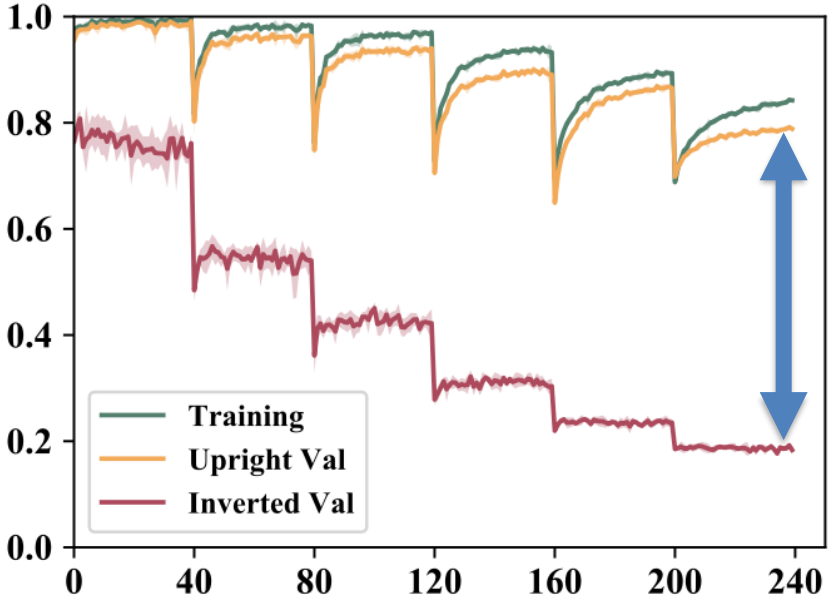
This inversion effect returns for objects of expertise:

Here, we trained the networks to be dog experts

LPNet



CNN



The networks were trained on

4, 8, 16, 32, 64, and 117 dog breeds from ImageNet

Note: Diamond and Carey showed inversion effects for Dog Show Judges

Overall, our networks match Yin's (1969) results:

Size of inversion effect:

Faces > Houses > Objects

(data not shown for houses)

Here, houses are a mono-oriented object – we only experience them upright, like faces, but we are not house experts.

Similarly, we match Diamond & Carey's result that dog show judges show an inversion effect – so inversion effects are a result of expertise.

There's nothing special about faces *per se*

Conclusions

- Anatomy matters!
- The log-polar network has nice properties – inversion effects fall out of the topology of the cortex and the nature of the problem:
  - Fine-level discrimination of homogeneous categories (faces and dogs) vs.
  - Basic-level categorization (much smaller effect)

We conclude that **using Euclidean coordinates and high resolution everywhere are the wrong prior for models of human vision.**

THANKS! To the following Gurons and NSF

Kira Fleischer

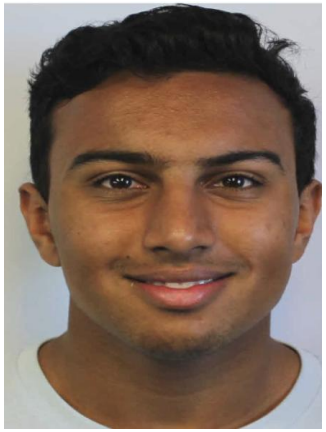
Nikita Kachappilly

Xavier Chen

Alexander Tahan

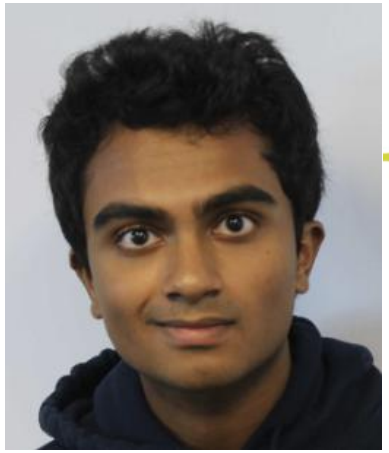
Panqu Wang

Martha Gahl



Shubham Kulkarni

Meilu Yuan



Arun Sugumar

And thank YOU for listening!

Questions?