# Face Processing in Humans and Deep Convolutional Neural Networks

20.08.2024

Benedikt Wirth

# The role of facial identification in the 21st century

- Face recognition most prevalent method of person identification in the pre-digital world
- In the digital world, facial recognition still common method of person identification, for example in:
  - User identification
  - Migration control
  - Criminal prosecution


Daniela Klette

# The role of facial identification in the 21st century

- Artificial face recognition systems as driver of efficiency and increased reliabilty in many application scenarios
- But potentially dire consequences of errors

***Eight Months Pregnant and Arrested After False Facial Recognition Match***

Porcha Woodruff thought the police who showed up at her door to arrest her for carjacking were joking. She is the first woman known to be wrongfully accused as a result of facial recognition technology.

Share full article 1473

**What went wrong with the electronic passport gates at UK airports?**

Passengers waited
checks on Tuesday
malfunctioned

The Register

**Cybercriminals are stealing iOS users' face scans to break into mobile banking accounts**

Deepfake-enabled attacks against Android and iPhone users are netting criminals serious cash

**A false facial recognition match sent this innocent Black man to jail**

# Motivation and Outline

Understand characteristics of human face recognition

Find (dis)similarities between human and artificial face recognition

Discern when/why human face recognition fails

Discern when/why artificial face recognition fails

- Improve understanding of human face recognition
- Improve performance of artificial face recognition
- Infer guidelines for human-AI collaboration

# Characteristics of human face processing

# Human face recognition in everyday life

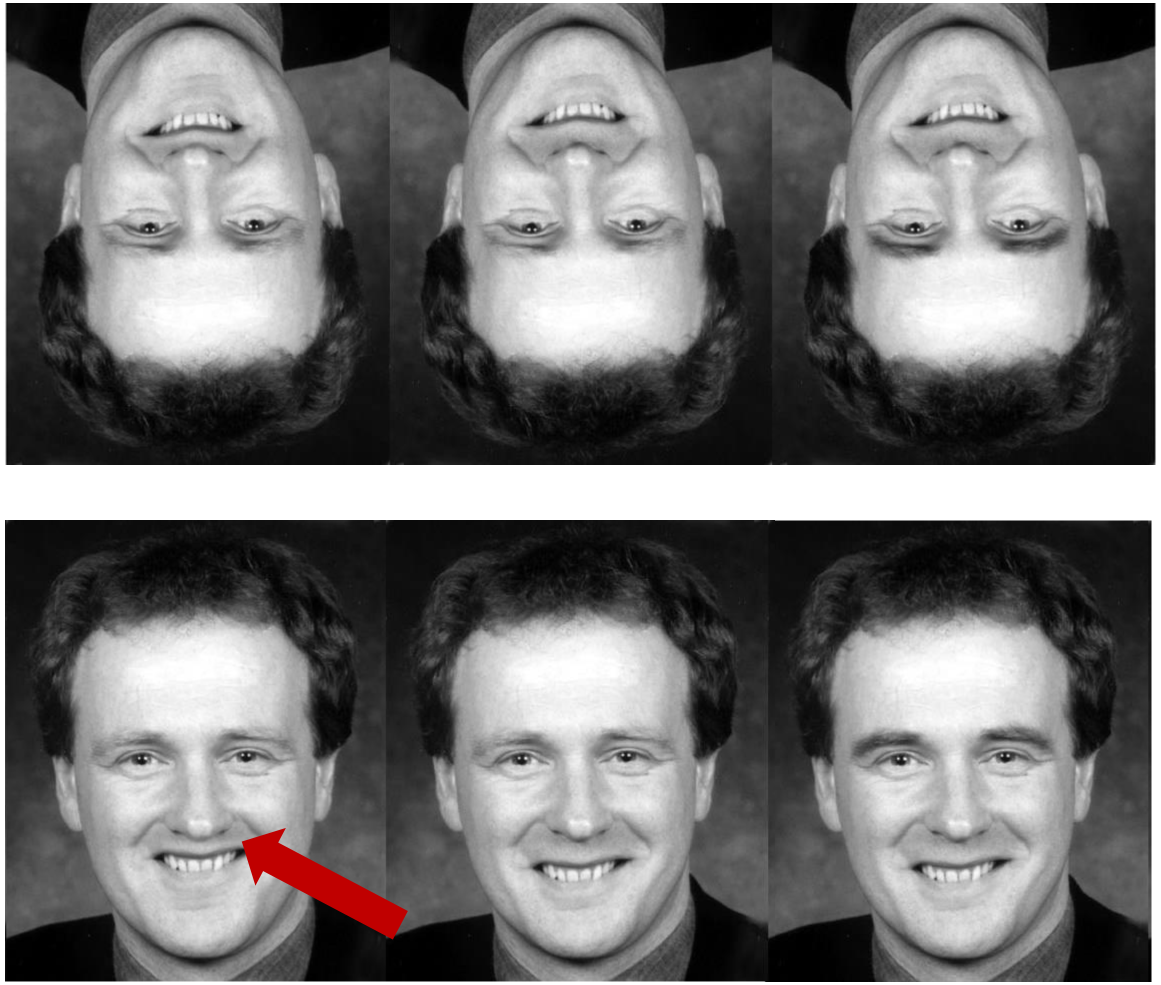
- The average person knows about 5000 faces (Jenkins et al., 2018)
  - Range between 1000 and 10000
- Robust recognition of familiar faces despite differences in:
  - Perspective
  - Illumination
  - Physical distance
  - Emotional expression
  - Age
- But: Face recognition is actually a hard task!

# Human face recognition in everyday life

- The average person knows about 5000 faces (Jenkins et al., 2018)
  - Range between 1000 and 10000
- Robust recognition of familiar faces despite differences in:
  - Perspective
  - Illumination
  - Physical distance
  - Emotional expression
  - Age
- But: Face recognition is actually a hard task!
- Why are we nevertheless able to robustly recognize (familiar) faces?

# Holistic face processing in humans

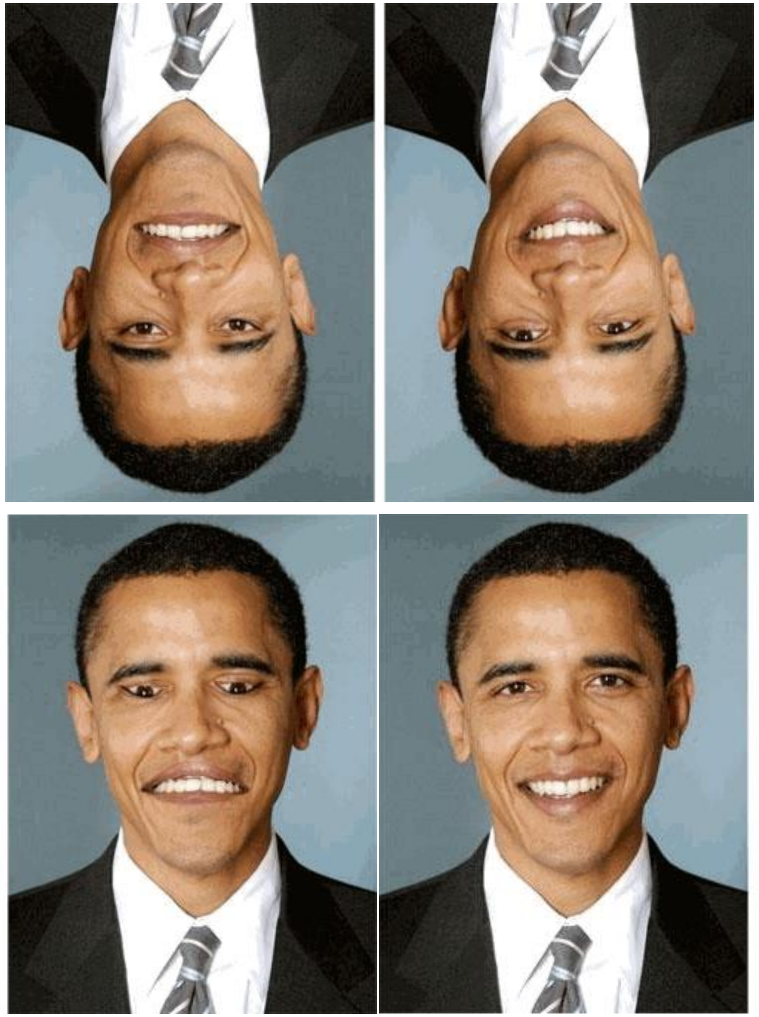
- In contrast to other object categories, faces are assumed to be processed holistically
  - Not only individual features
  - But also spatial relationships between aspects of the face (second order relations, configural processing)
- Empirical evidence for holistic processing:
  - Inversion effect



Hancock al. (2000)

# Holistic face processing in humans

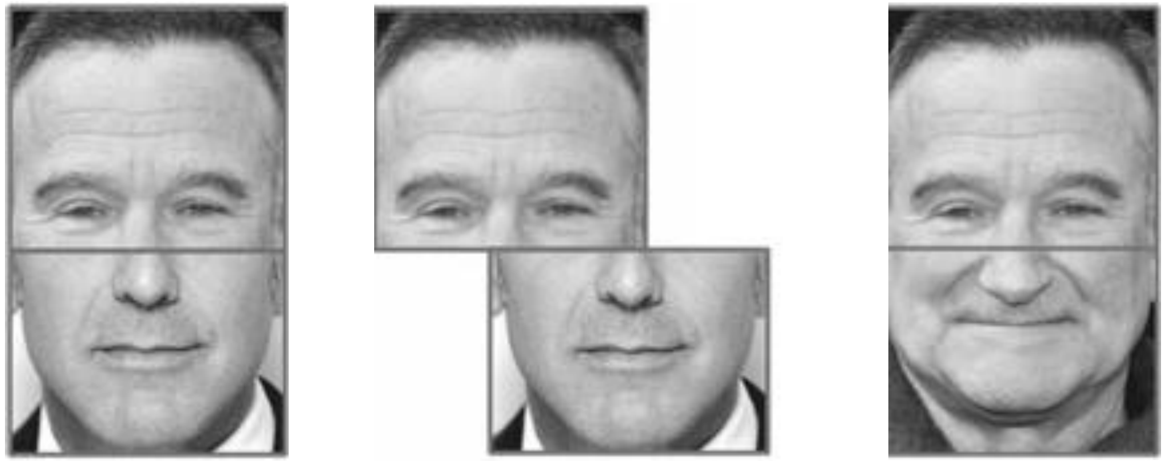
- In contrast to other object categories, faces are assumed to be processed holistically
  - Not individual features
  - But spatial relationships between aspects of the face (second order relations, configural processing)
- Empirical evidence for holistic processing:
  - Inversion effect
  - Thatcher illusion



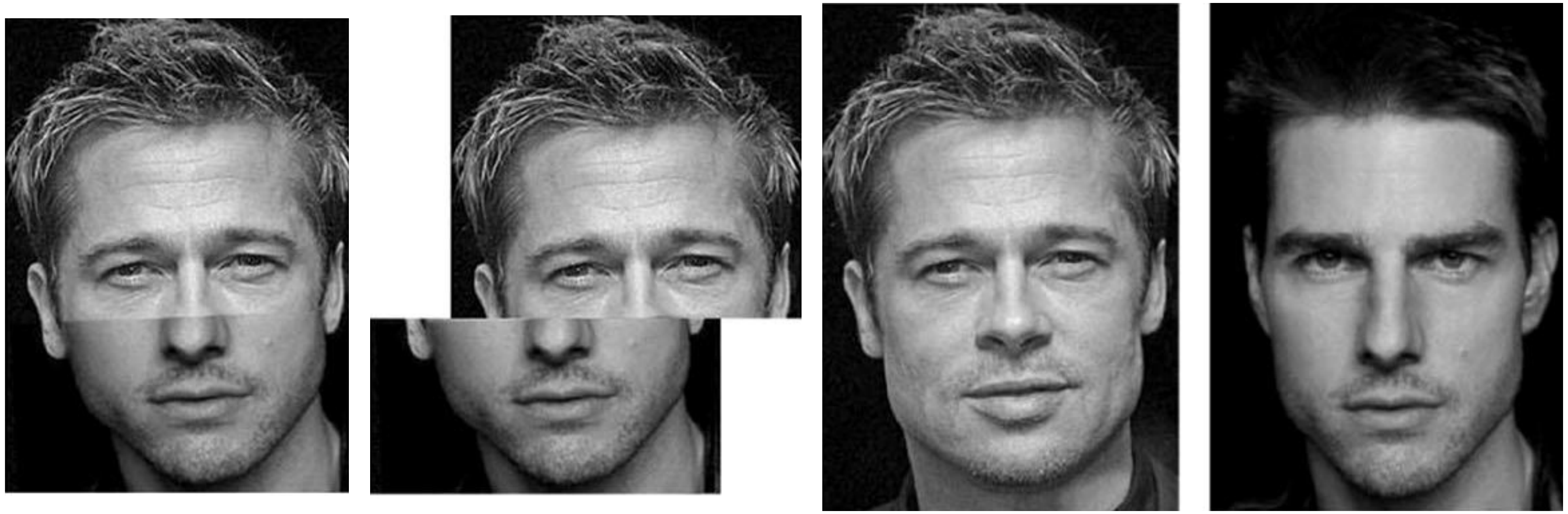
Thompson (1980)

# Holistic face processing in humans

- In contrast to other object categories, faces are assumed to be processed holistically
  - Not individual features
  - But spatial relationships between aspects of the face (second order relations, configural processing)
- Empirical evidence for holistic processing:
  - Inversion effect
  - Thatcher illusion
  - Composite-face effect



Murphy et al. (2017)

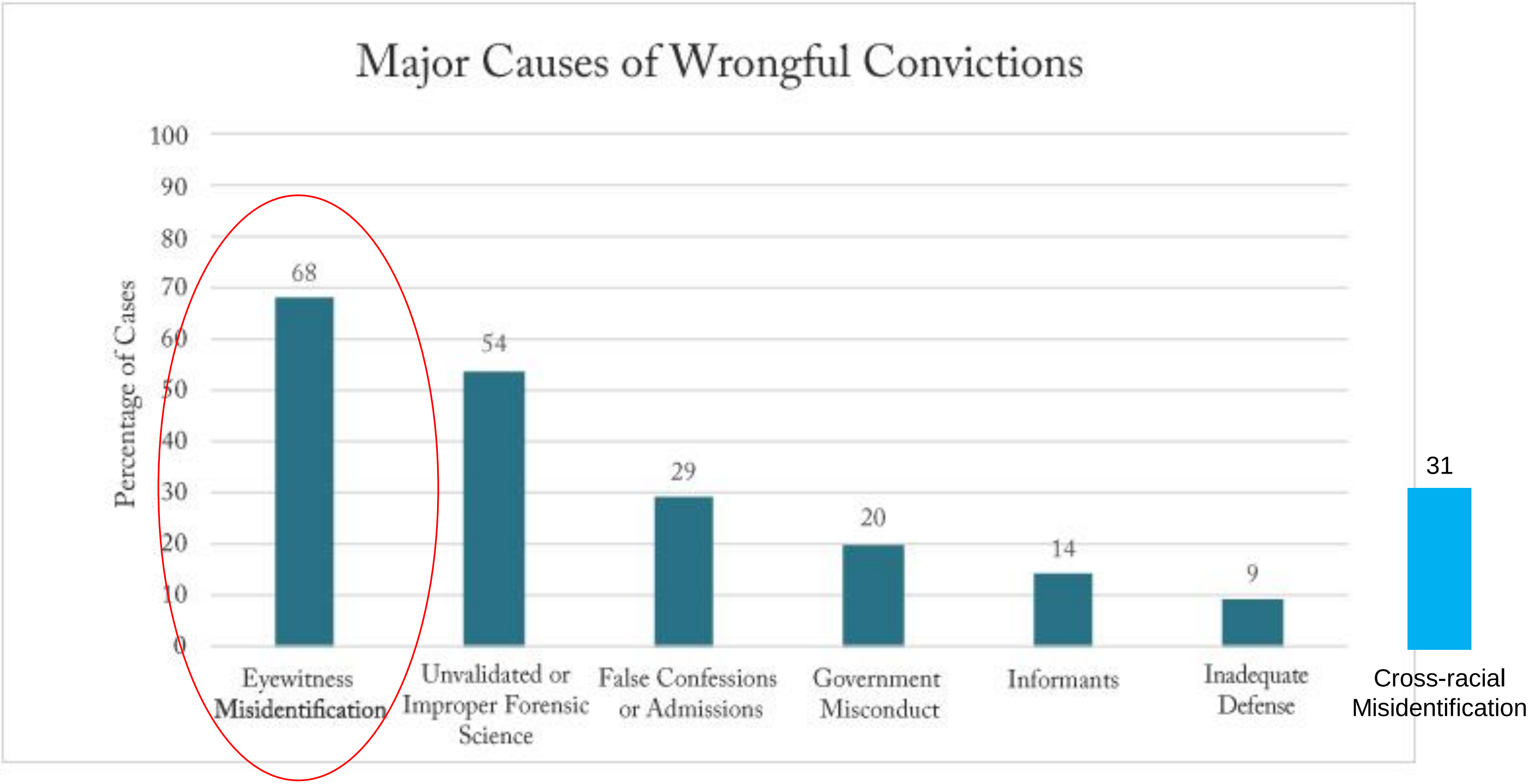


Tanaka & Simonyi (2016)

# When face recognition fails…

# … The consequences can be dire

- Innocence project
  - NGO with the aim to overturn wrongful convictions
  - 300 succesful suspensions of wrongful convictions
  - Mostly due to new DNA evidence



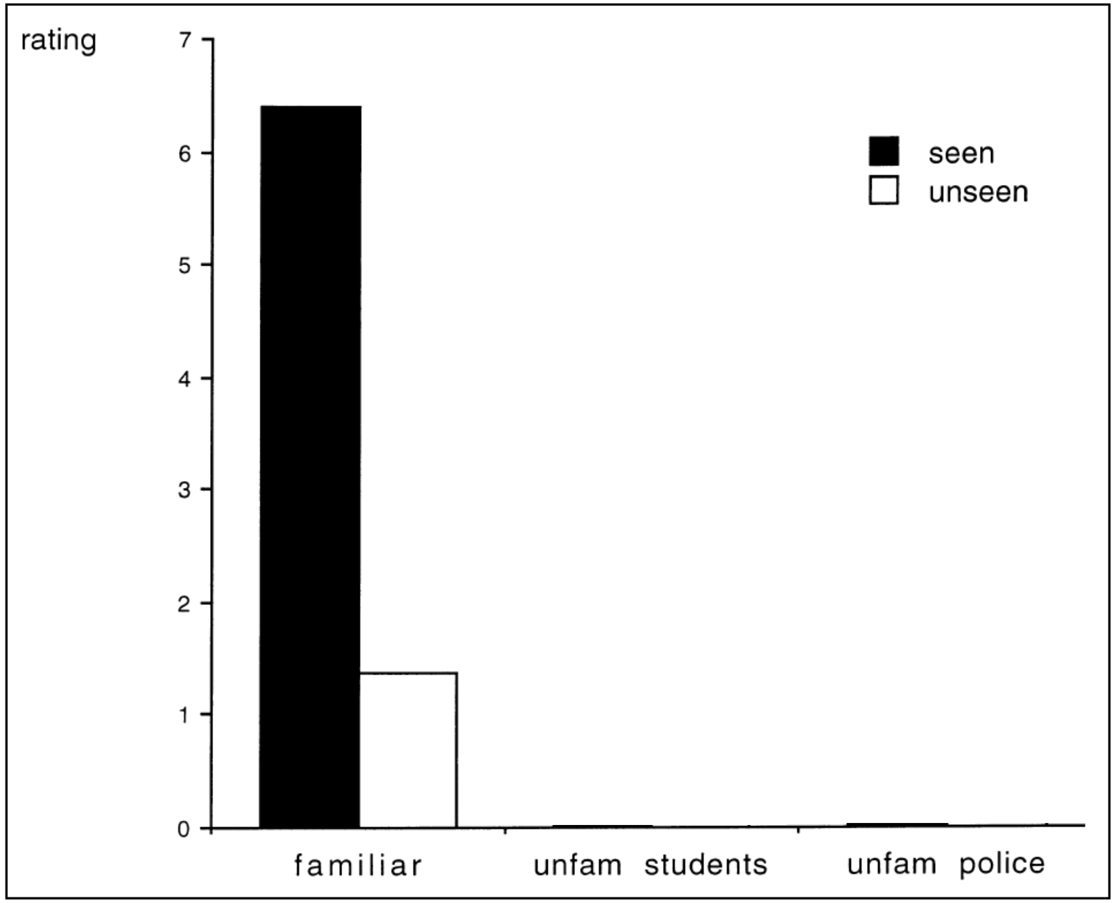
# Recognition of familiar versus unfamiliar faces

- Identification of people from CCTV (Burton et al. 1999)
  - Presentation of 10 CCTV images showing 10 different lecturers entering a university building
  - Participants:
    - 20 students familiar with the lecturers (same department)
    - 20 students unfamiliar with the lecturers (different department)
    - 20 police officers (average 13.5 years of service)
  - Task:
    - Presentation of 20 high quality images (10 previously seen, 10 not seen)
    - Indicate on a scale from 1 (definitely not seen) to 7 (definitely seen) whether you have seen the depicted person in the videos

# Recognition of familiar versus unfamiliar faces

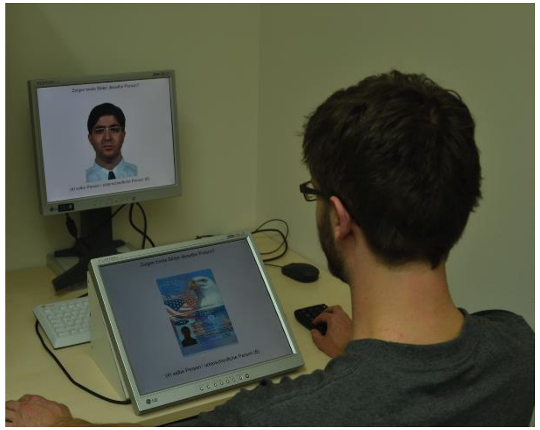
- Results:
  - Near-perfect performance of students familiar with the shown lecturers
  - Poor performance of students unfamiliar with the shown lecturers
  - Poor performance of police officers unfamiliar with the shown lecturers

- Conclusion: Familiarity critical factor of face recognition performance



# Unfamiliar face matching in novices and professional experts

- Simulation of everyday passport controls (Wirth & Carbon, 2017)
  - Presentation of 192 pairs of faces consisting of
    - Passport photograph
    - Large-sized  photograph
  - Task: Indicate whether both photographs depict the same person

  - Participants:
    - 48 novices (students without  specific passport-matching experience)
    - 96 officers of the German Federal Police:
      - 48 officers with short job experience ($M$ = 5.7 years)
      - 48 officers with long job experience ($M$ = 22.7 years)

BUNDESPOLIZEI

# Unfamiliar face matching in novices and professional experts

## Match trials

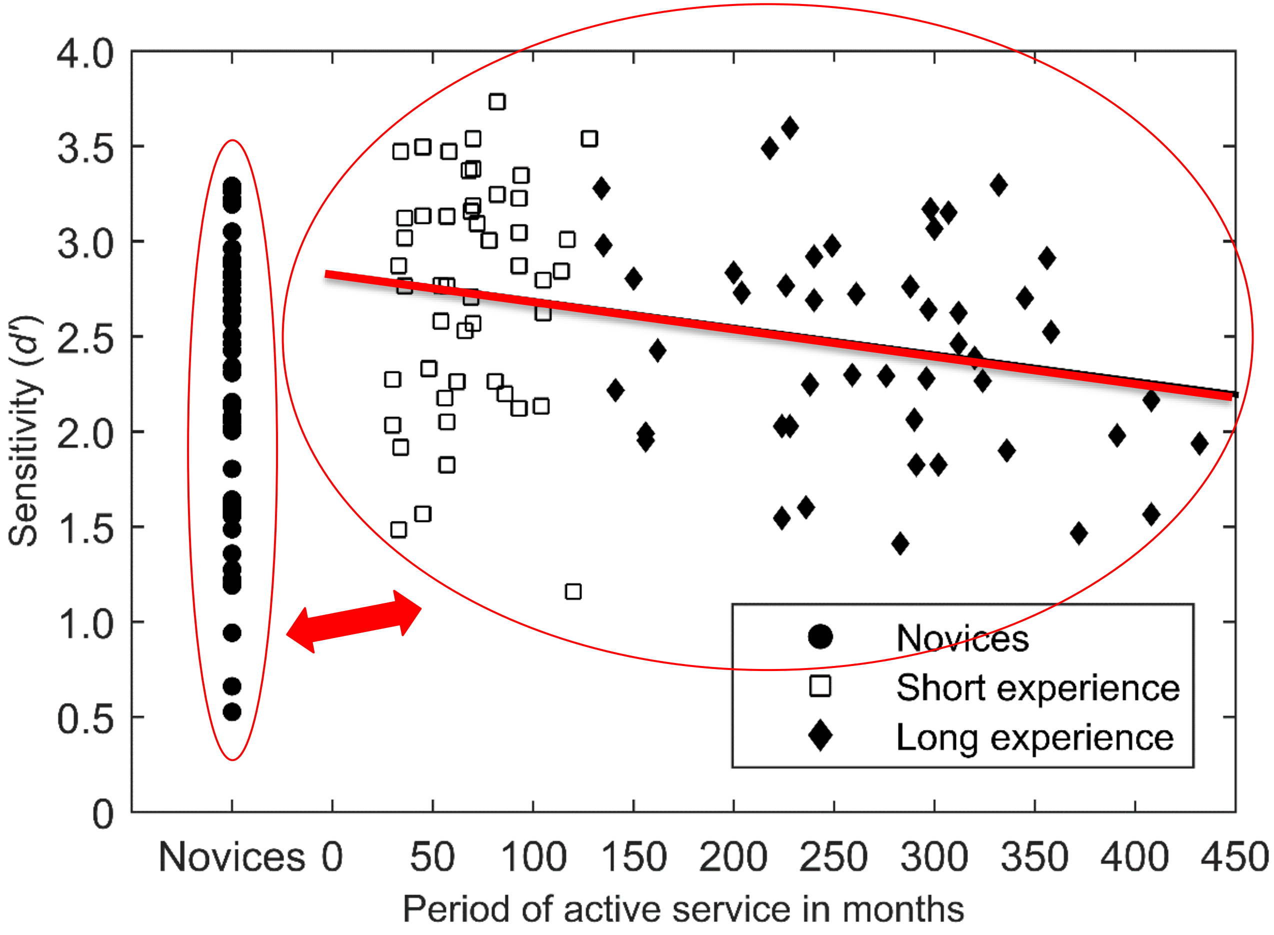
| Unmanipulated | Paraphernalia | Distinctive features | Hairstyle |
|---|---|---|---|
|  |  |  |  |

## Mismatch trials

| Unmanipulated | Paraphernalia | Distinctive features | Hairstyle |
|---|---|---|---|
|  |  |  |  |

# Unfamiliar face matching in novices and professional experts

- Results
  - Police officers‘ matching performance significantly higher than novice performance
  - But within group of police officers, decreasing performance with increasing professional experience



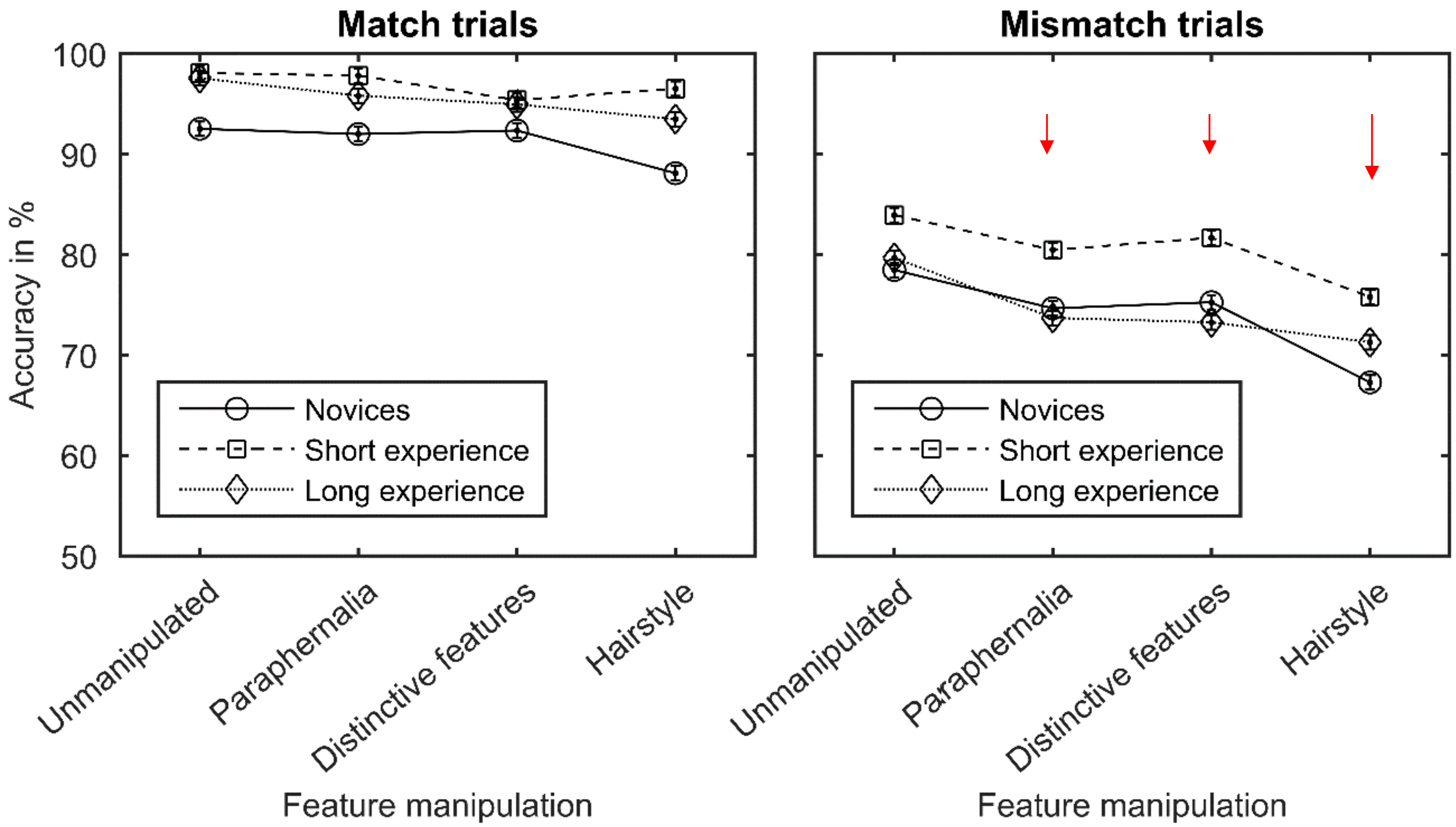
# Unfamiliar face matching in novices and professional experts

Match trials:

- High accuracy close to ceiling
- Hardly affected by feature manipulations

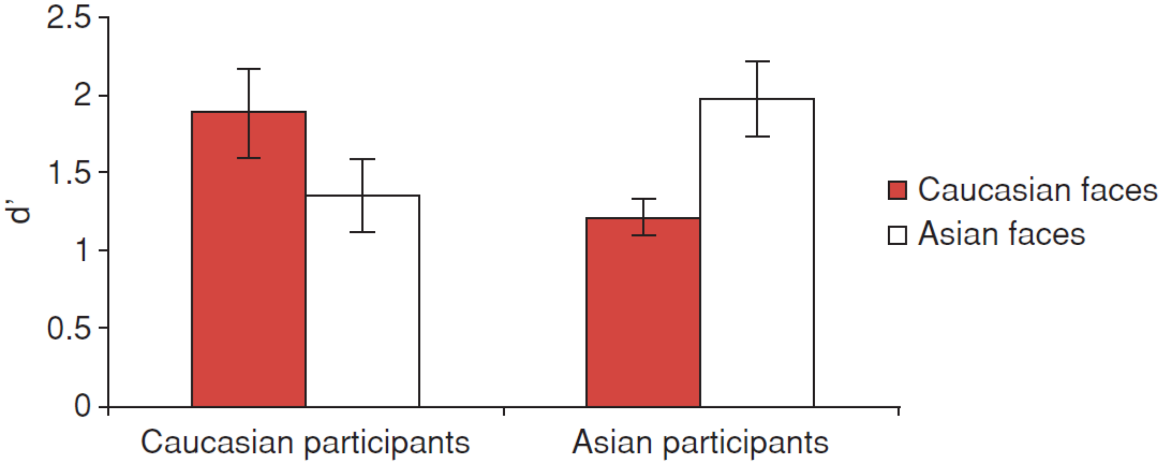
Mismatch trials:

- Lower accuracy rates
- Significant impairments by all feature manipulations



# The own-race bias

- Investigation of face recognition in White and Asian participants (Michel et al., 2006)
  - Sequential presentation of 20 White and 20 Asian faces
  - Task
    - Memorize the presented faces
    - Subsequently presentation of 40 faces (20 previously presented, 20 new)
    - Indicate whether a given face was seen before or not
  - Resuts:
    - Better performance for White faces in White participants
    - Better performance for Asian faces in Asian participants



# The own-race bias

Two groups of explanatory theories for the own-race bias (ORB)

**Cognitive theories**

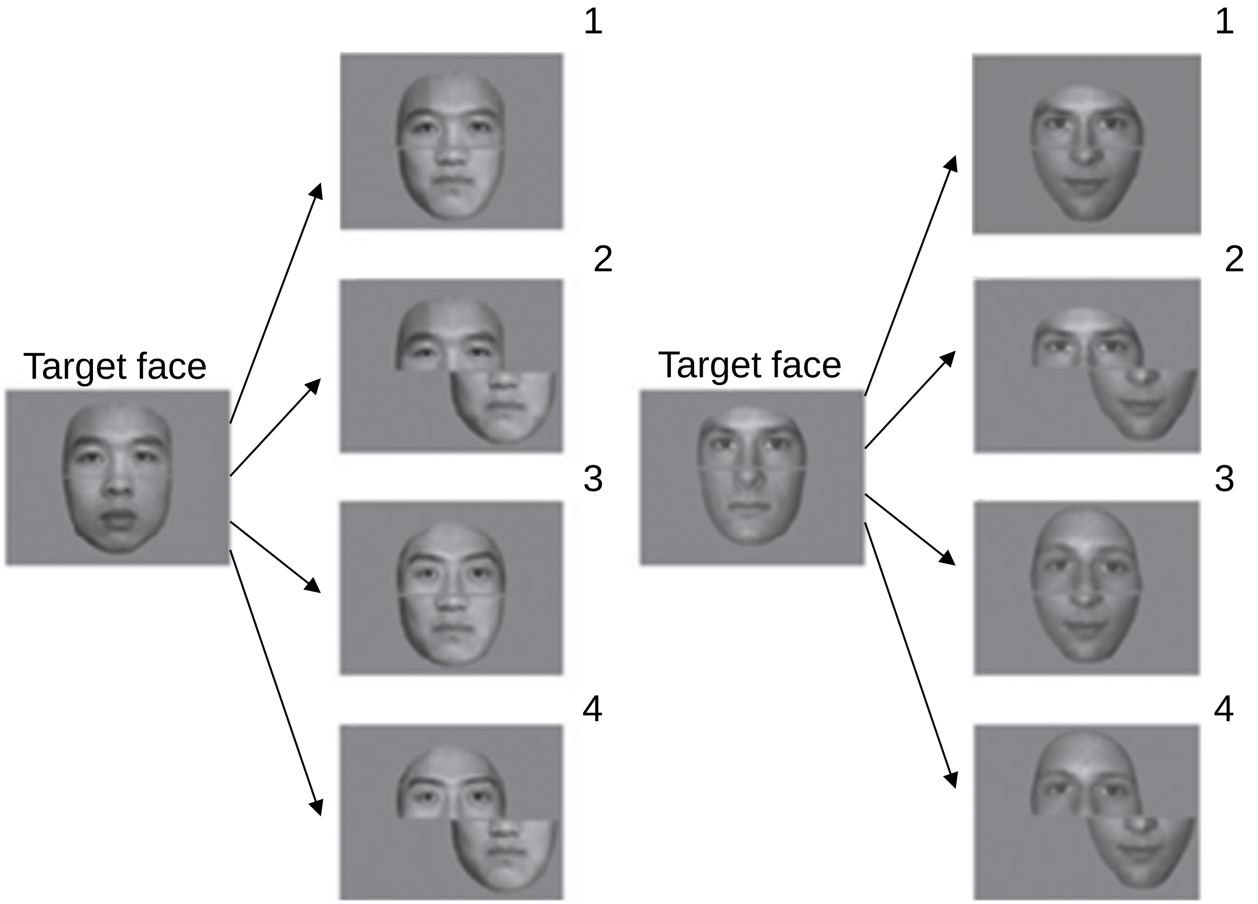
- ORB consequence of reduced experience with other-race faces
- Cognitive system not trained to be sensitive to diagnostic features of other-race faces
- Reduced use of holistic processing
- Correlation between strength of ORB and frequency of contact with people from different ethnicities (during childhood)

**Social-psychological theories**

- ORB consequence of social ingroup-outgroup processes
- Categorization of other-race faces on superordinate level as outgroup
- Less motivation to remember outgroup faces
- Correlation between strength of ORB and racial prejudice

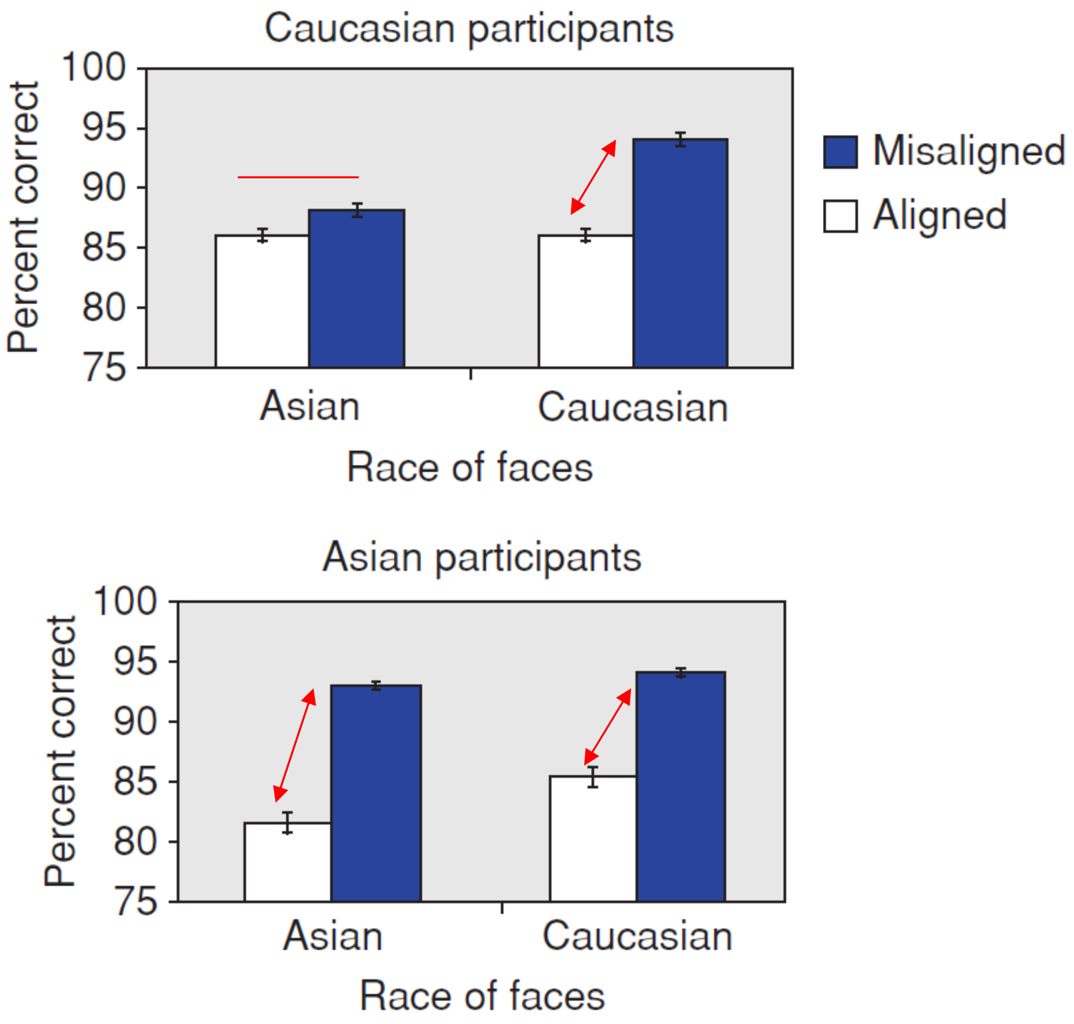
# The own-race bias

- Study by Michel et al. (2006)
  - Presentation of a target face and subsequently a comparison face
  - Task: Is the top half of the target face and the comparison face the same?
  - Four different types of comparison faces:
    1. Same/Aligned
    2. Same/Misaligned
    3. Different Aligned
    4. Different Misaligned



# The own-race bias

- Results:
  - White participants:
    - Significant composite-face effect for White faces
    - No composite-face effect for Asian faces
  - Asian participants:
    - Significant composite-face effect for Asian faces
    - Significant, but reduced composite-face effect for White faces
- Conclusion: Holistic processing reduced or even completely elimited for other-race faces



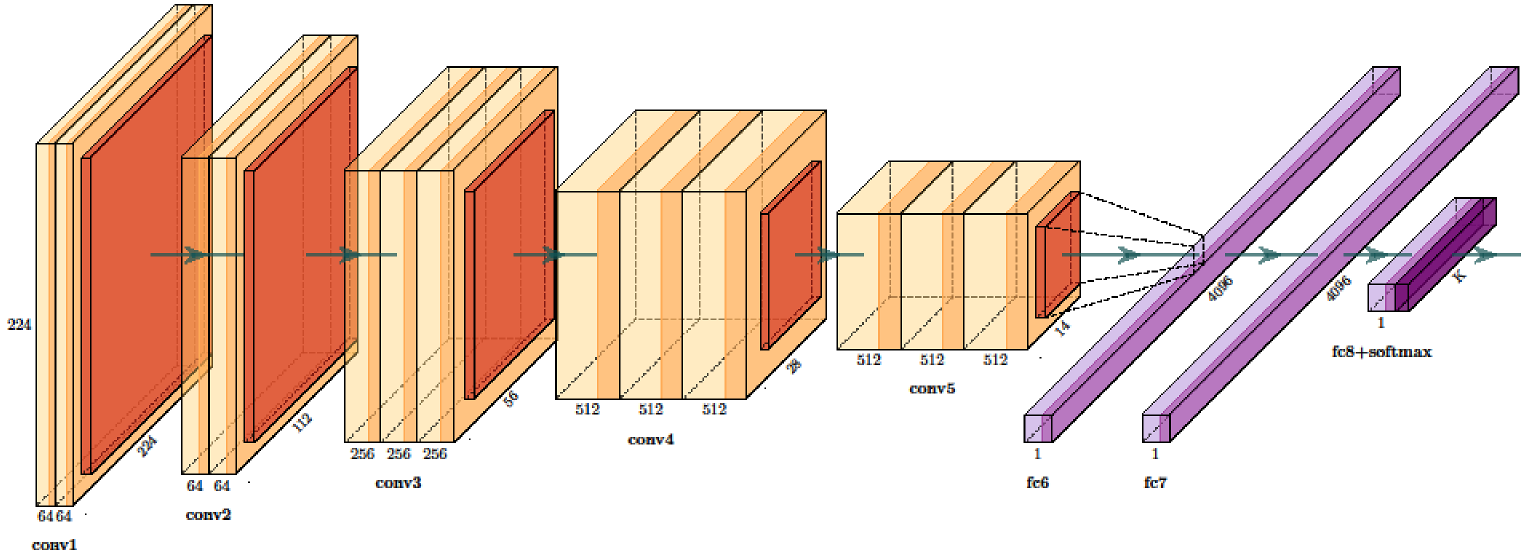
# Interim summary

- Substantial differences between processing of faces versus processing of other objects
  - Holistic processing (i.e., processing of the „gestalt“ of a face)
  - Processing of spatial distances between individual face parts
- Advantage of holistic processing: Robust recognition of familiar faces across different illumination conditions, perspectives, emotional expressions, etc.
- Processing of unfamiliar (and especially of other-race) faces:
  - Less use of holistic processing
  - Less reliable and robust

# Deep Convolutional Neural Networks (DCNNs)

# Face processing in AI: Deep Convolutional Neural Networks

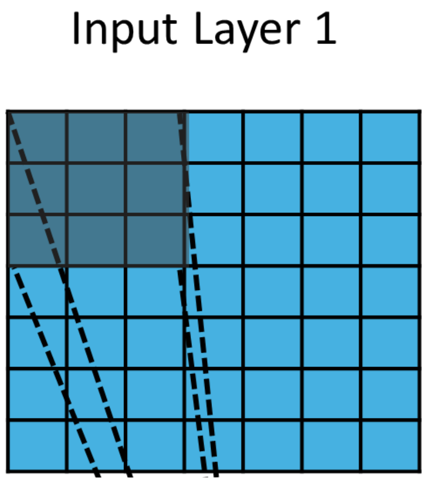
- Network architecture developed for image classification:
- Characterized by
  - Convolution operations
  - High number of layers
  - Purely Feed-forward operations



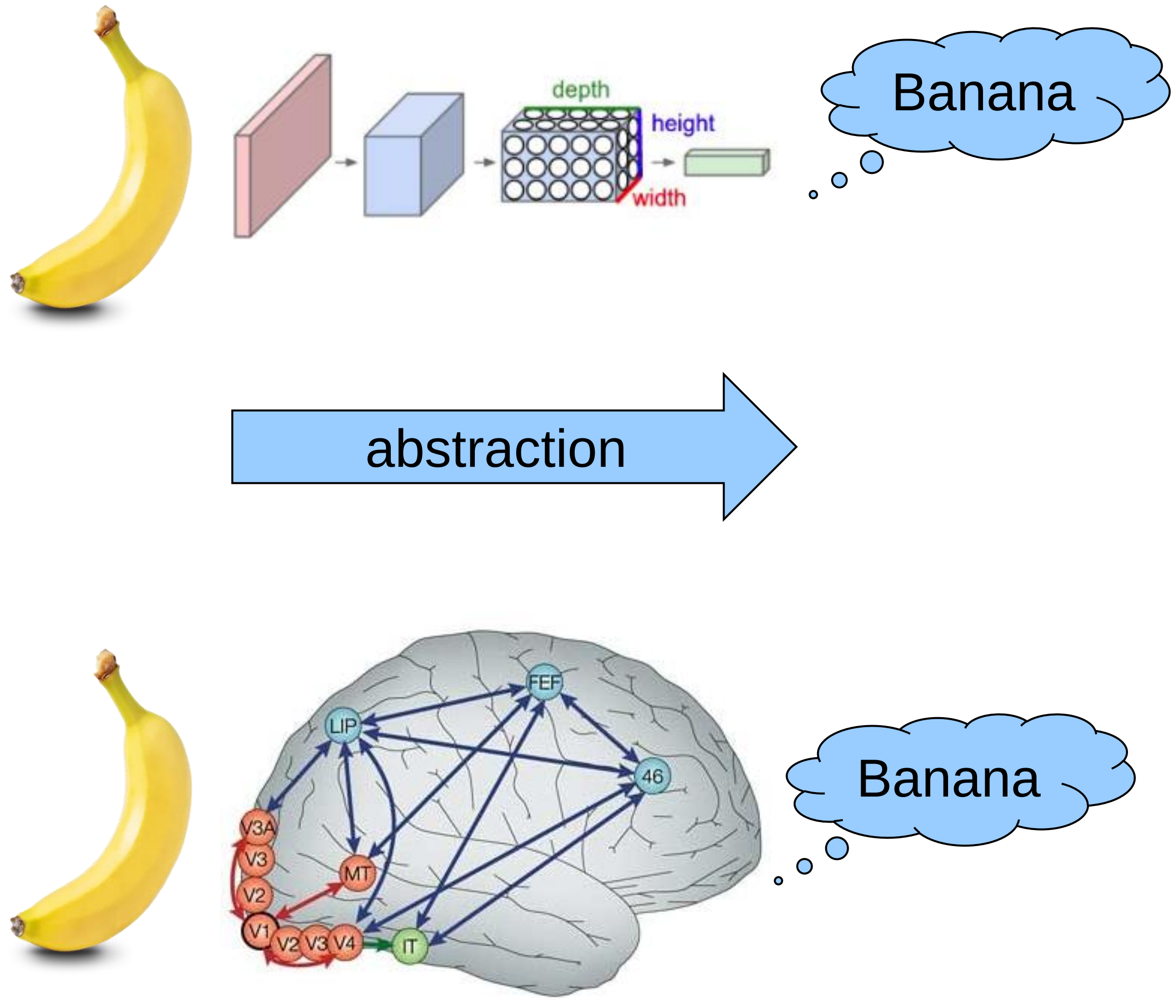

- Idea: Do not process individual pixels but small image patches to decrease the number of neurons required
- Popular architectures:
  - AlexNet (Krizhevsky et al., 2012)
  - VGG-16 (Simonyan & Zisserman, 2015)
  - ResNet-50 (He et al., 2016)

# The principles of convolution and receptive field size

- Input image size: 7×7 pixels
- Conv layer 1:
  - Kernel size: 3×3
  - Stride: 2
  - Output size: 3×3
  - Receptive field size: 3×3
- Conv Layer 2:
  - Kernel size: 3×3
  - Output size: 1×1
  - Receptive field size: 7×7



# Structural (dis)similarities between humans and DCNNs




Despite generally similar architecture, differences in operations:

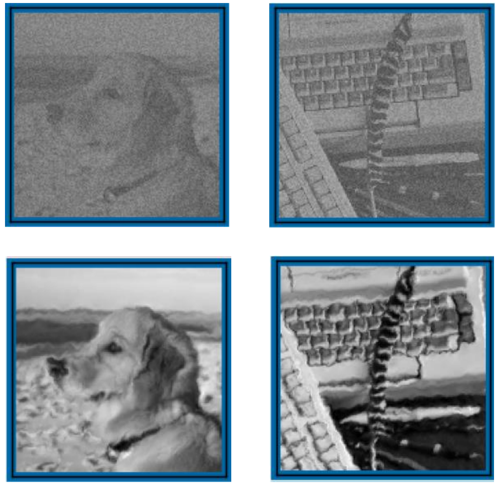
DCNNs:
- Convolution
- Pooling
- Normalization

Human Brain:
- Center-surround antagonism
- Cortical magnification
- Local receptive fields without weight sharing

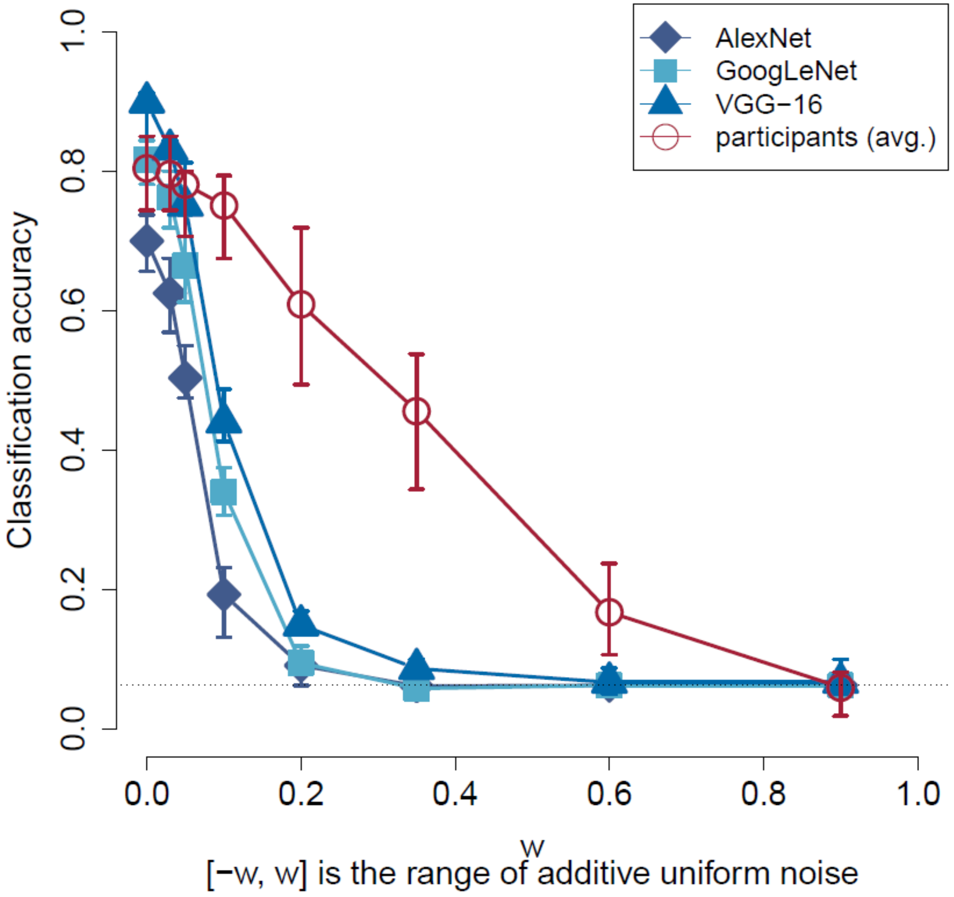
# Object Recognition in Humans and DCNNs

# Differences between humans and DCNNs in object recognition

- Comparison of recognition accuracy for 16 ImageNet classes (Geirhos et al., 2018)
- Restricted viewing conditions for humans:
  - Image presented for 200 ms
  - Masked with pink noise for 200 ms
- Four different image manipulations
  - Greyscale
  - Reduced Contrast
  - Added noise
  - Eidolon



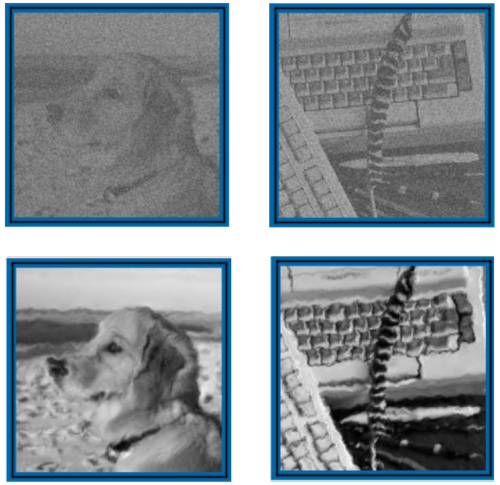
➔ Human object recognition more robust than DCNN recognition



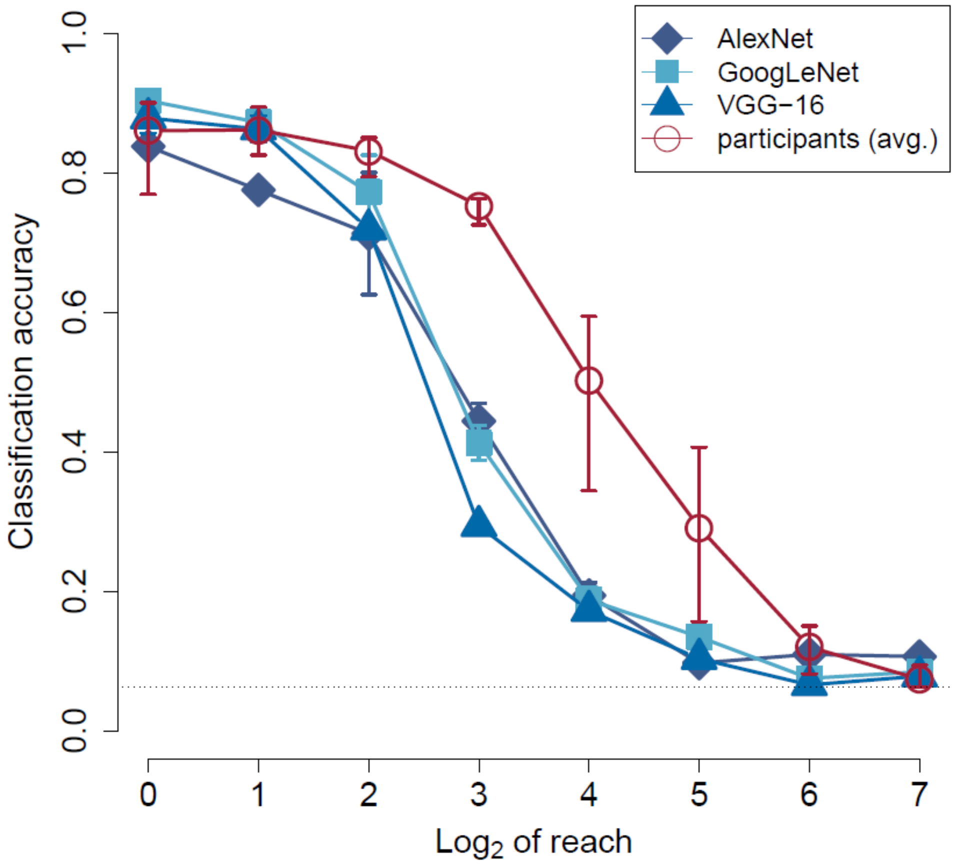

50%-Accuracy examples

# Differences between humans and DCNNs in object recognition

- Comparison of recognition accuracy for 16 ImageNet classes (Geirhos et al., 2018)
- Restricted viewing conditions for humans:
  - Image presented for 200 ms
  - Masked with pink noise for 200 ms
- Four different image manipulations
  - Greyscale
  - Reduced Contrast
  - Added noise
  - Eidolon



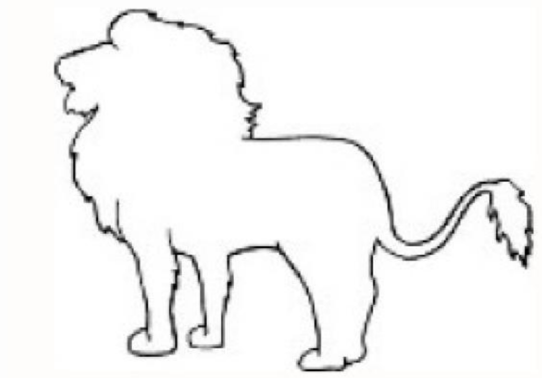
➔ Human object recognition more robust than DCNN recognition




50%-Accuracy examples

# Texture versus global shape in DCNN object recognition

- DCNNs have problems classifying objects with unusual texture (Baker et al., 2018)
  - Silhouettes
  - Outlines
  - Glass figurines



- In cue-conflict images, humans show a bias towards shape, DCNNs towards texture (Baker et al., 2018; Geirhos et al., 2019)

„Lobster!“

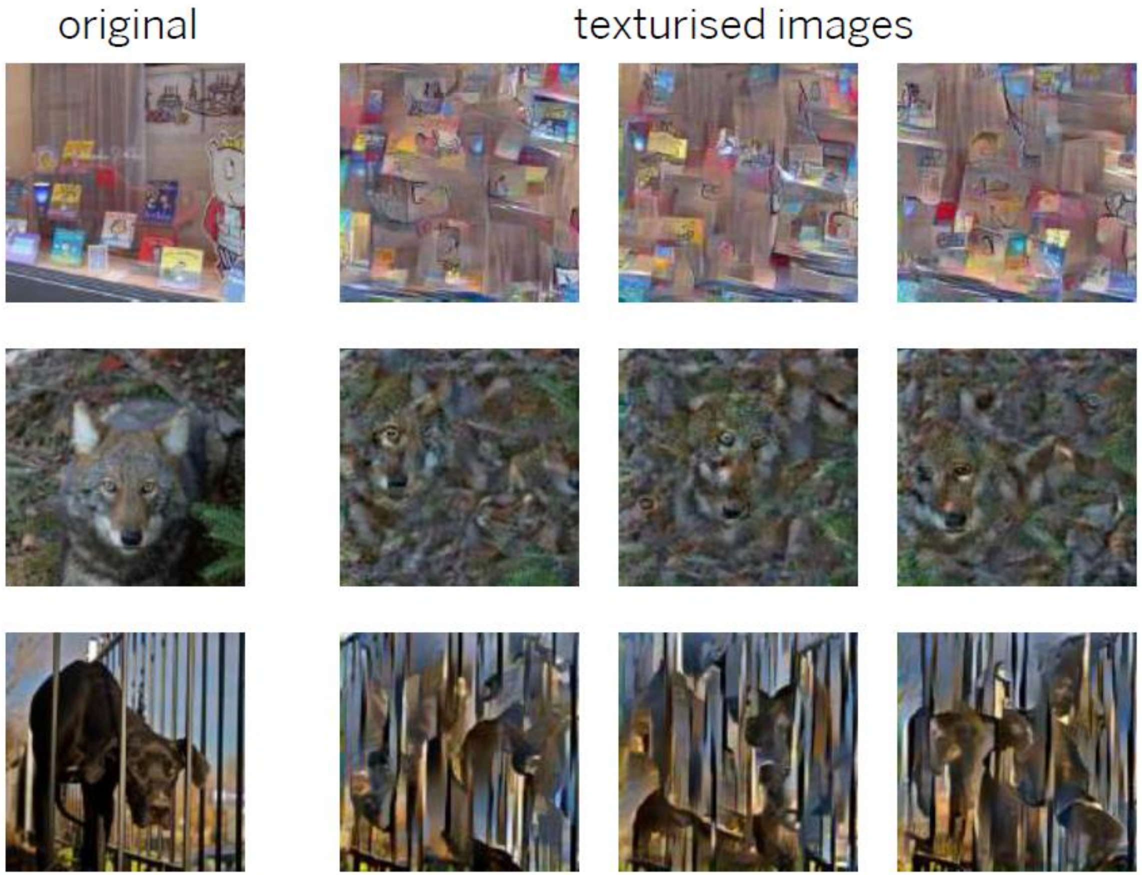
„Wool!“

„Cat!“

„Elephant!“

# Texture versus global shape in DCNN object recognition

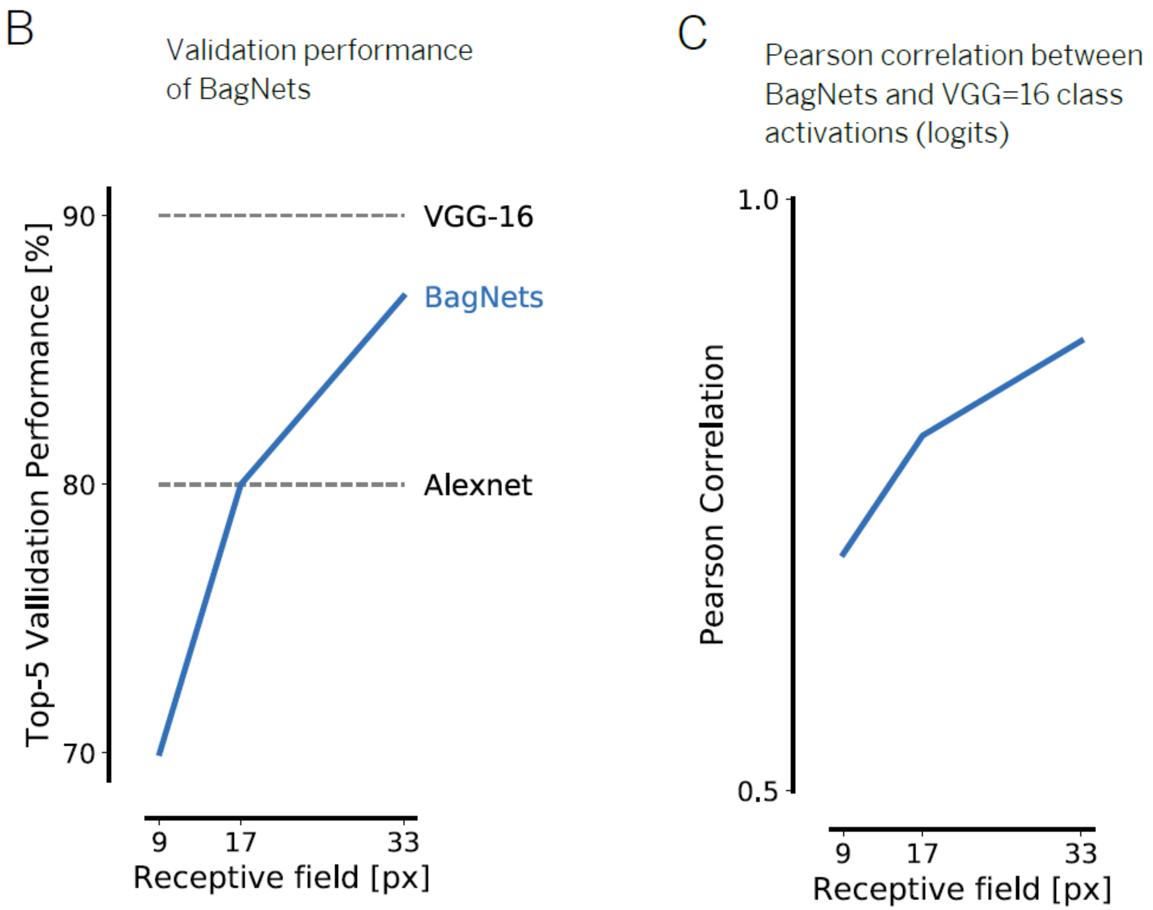
- Creation of texturized images (Brendel & Bethge, 2019)
- Surprisingly high accuracy of VGG-16 on scrambled ImageNet pictures: 79.4% top-5 accuracy (vs. 90.1% on normal pictures)
- But: Texturized images created based on hidden layer activations of VGG-19 ➔ Circular argument



# Local features versus global shape in DCNN object recognition

- Creation of a bag-of-local features model (Brendel & Bethge, 2019)
- Modification of ResNet-50 to reduce size of receptive field of the topmost convolutional layer to q × q pixels with $q \in \{9, 17, 33\}$
- Surprisingly high classification accuracy on ImageNet
- High correlation of class activations between VGG-16 and BagNets

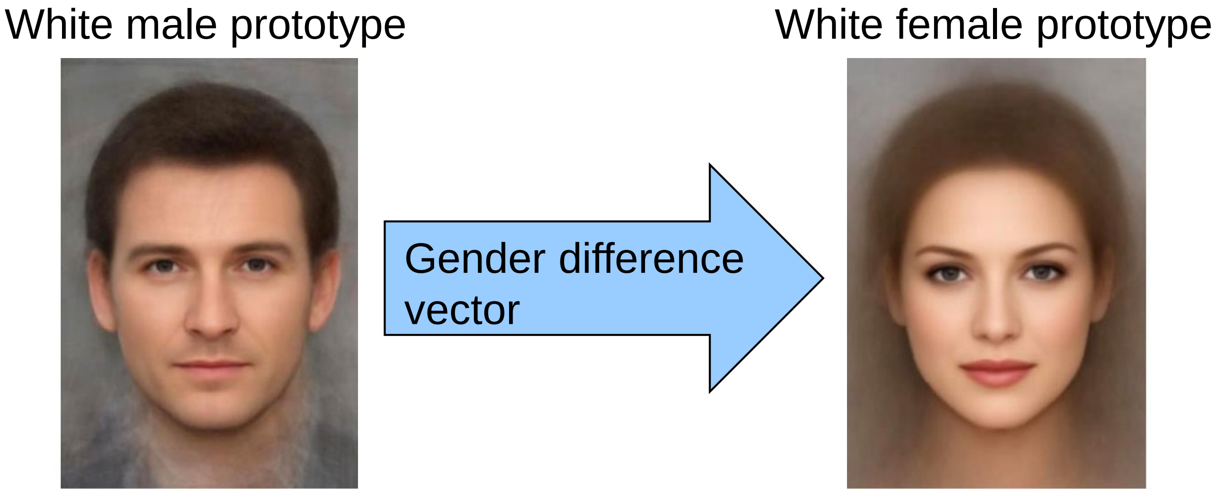
➔ Object recognition in DCNNs mainly contingent upon local features and texture, not on global shape

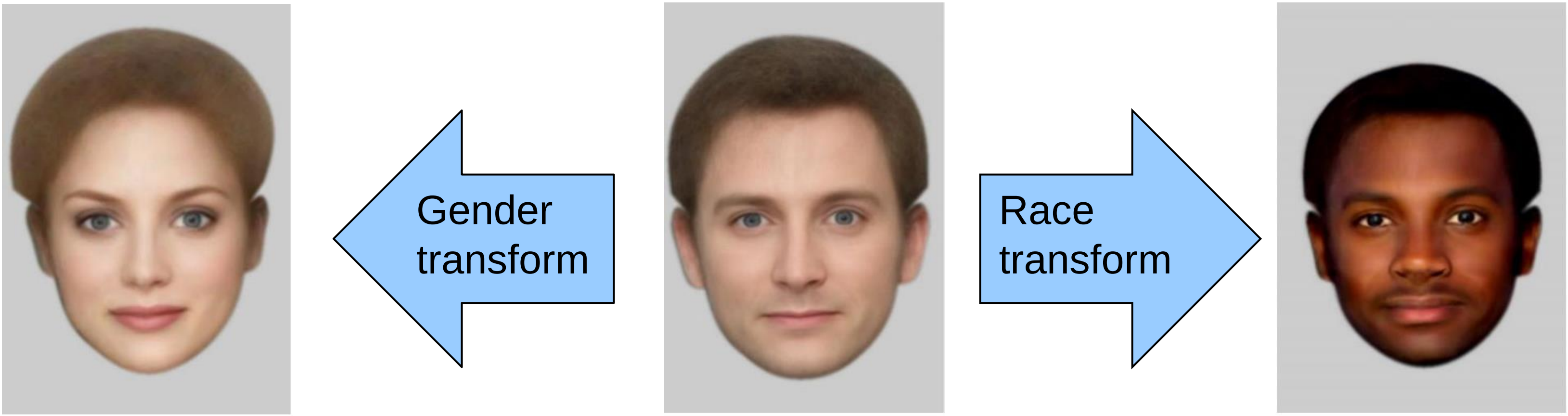


# Face processing in humans and DCNNs

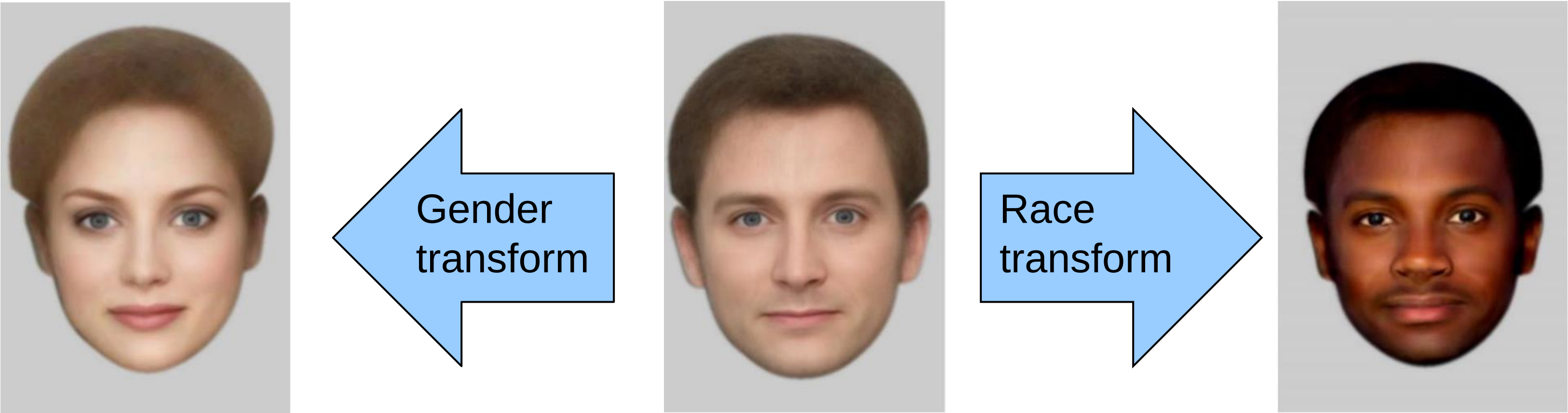
# Gender and race cues

- Investigation of reliance on race and gender cues for face identification in humans and DCNNs (Hancock et al., 2020)
- Creation of four prototypical faces by averaging multiple individual faces:
  - White male
  - White female
  - Black male
  - Black female



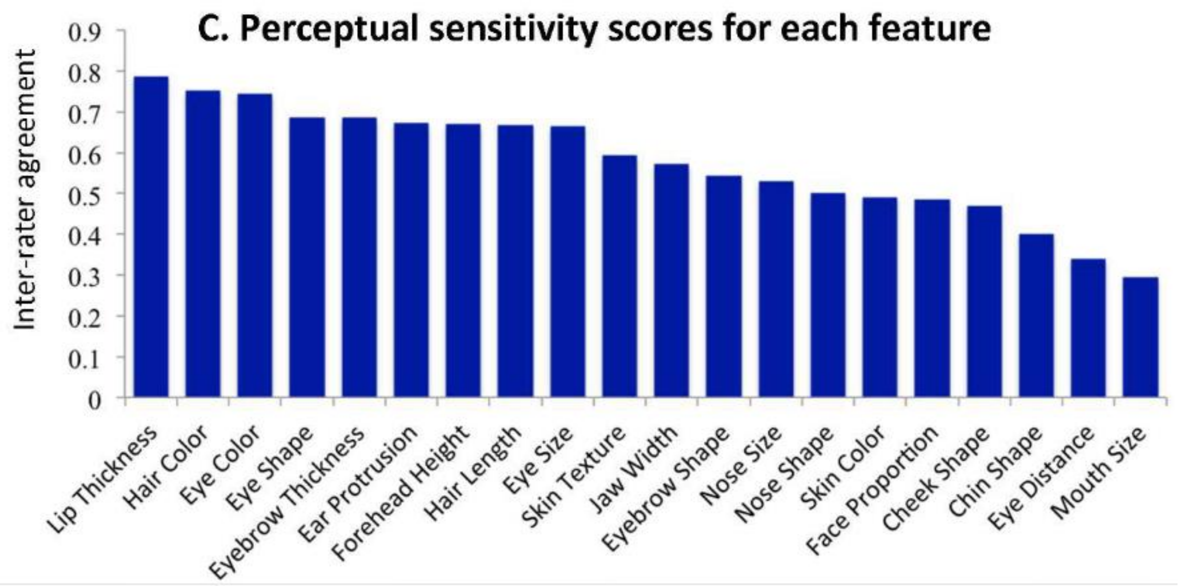
# Gender and race cues




| transform direction | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| male -> female | 100 | 93.5 | 100 | 98.8 | 42.9 | 99.4 |
| female -> male | 99.2 | 88.6 | 98.5 | 91.7 | 30.3 | 97 |
| male; White -> Black | 76.4 | 19.4 | 55.8 | 31.1 | 0.5 | 31.8 |
| female; White -> Black | 43.9 | 6.1 | 44.7 | 11.3 | 0 | 18.3 |

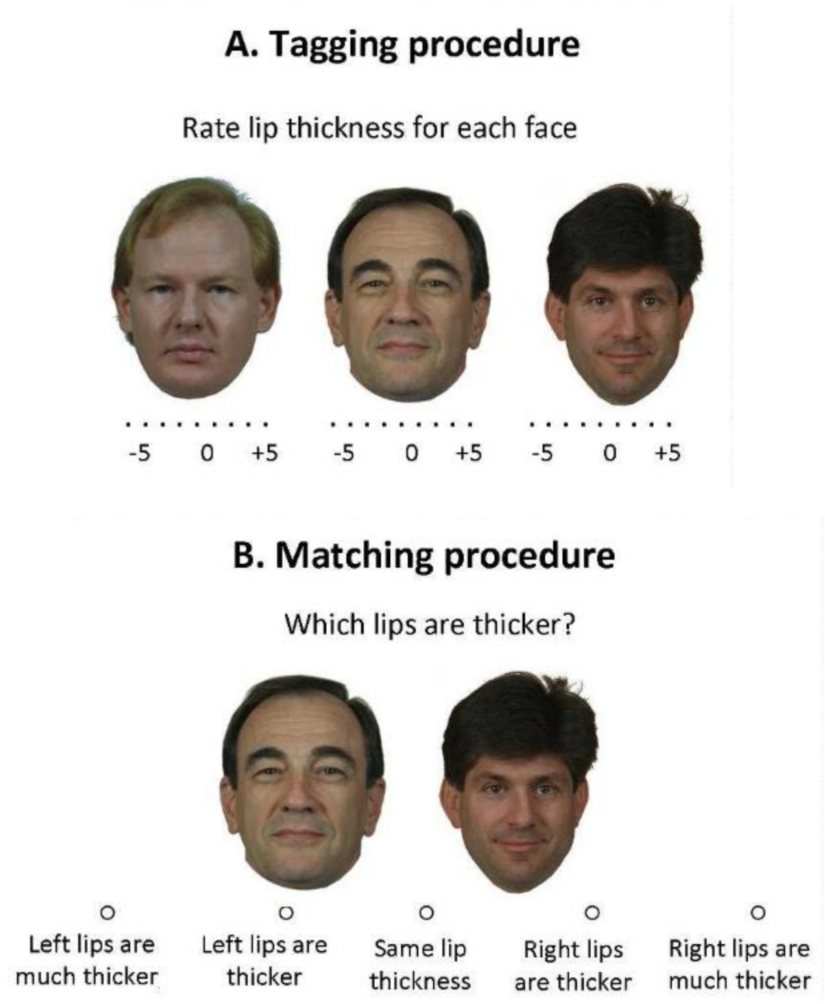
# Diagnostic Features

- Measurement of perceptual sensitivity for different facial features (Abudarham et al. 2016)

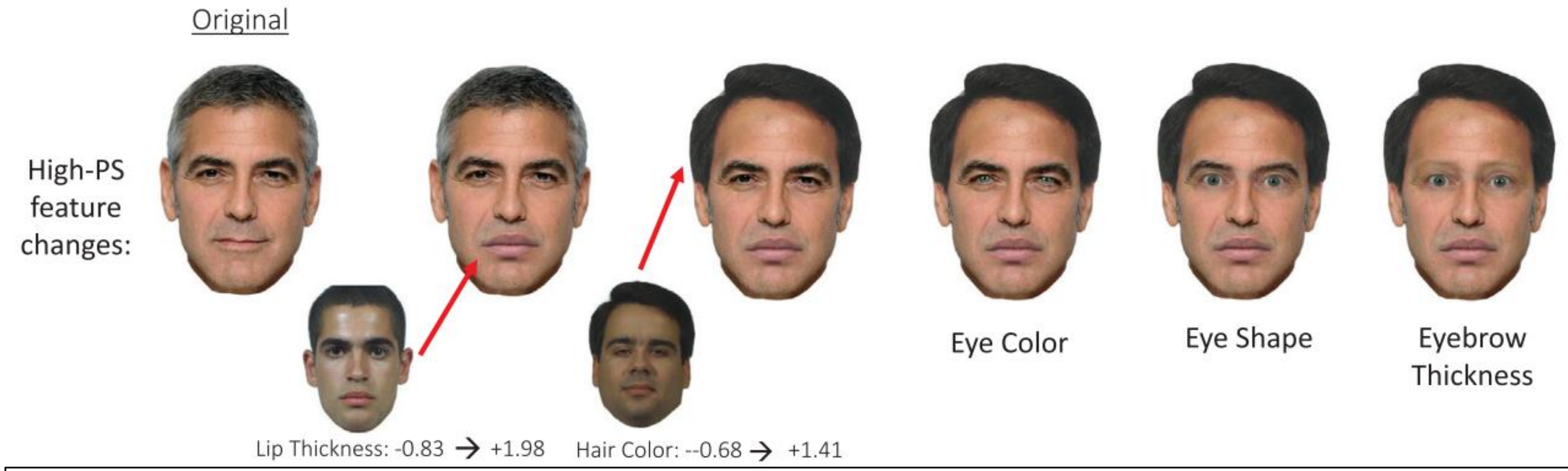



High PS features

Low PS features

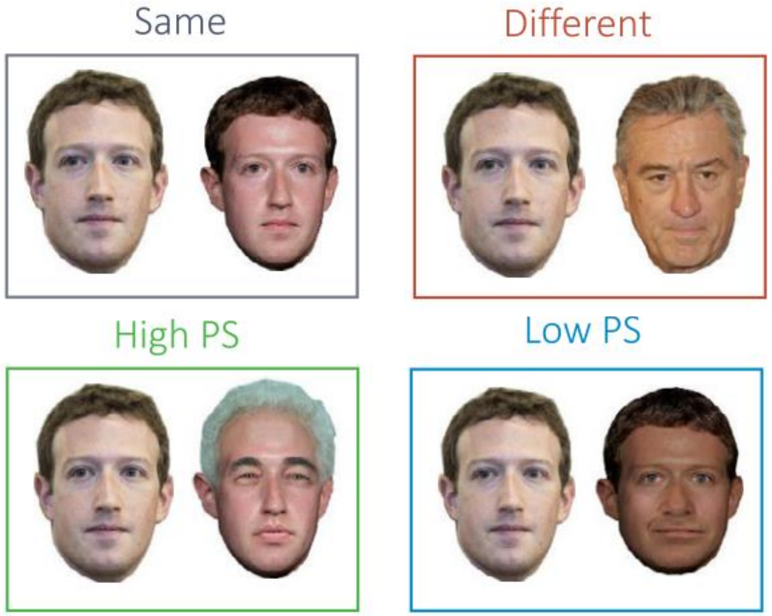


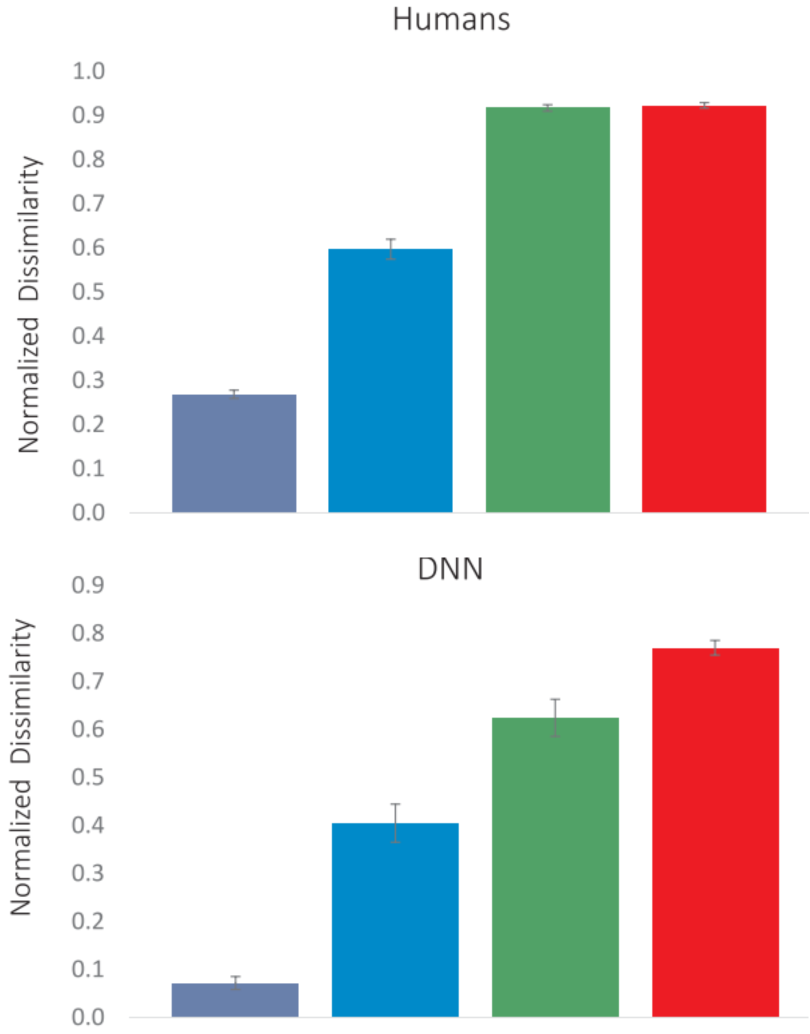
# Diagnostic Features



# Diagnostic Features

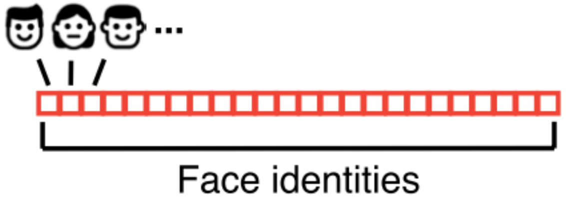
- Presentation of four types of face pairs to human participants and DCNNs (Abudarham et al., 2019)
- Task: Indicate whether two faces belong to the same person on a scale ranging from 1 (definitely not the same person) to 6 (definitely the different people)



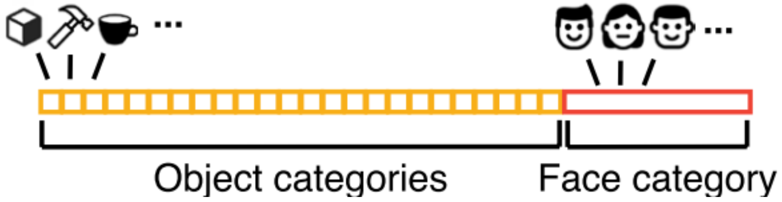



# The inversion effect in DCNNs

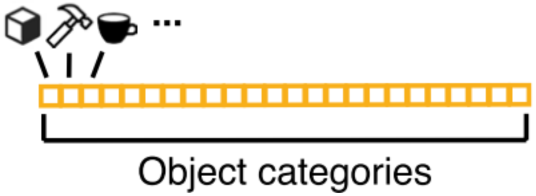
- Comparsion of face matching performance for upright and inverted faces (Dobs et al., 2023)
  - Human participants
  - Face-Identification CNN

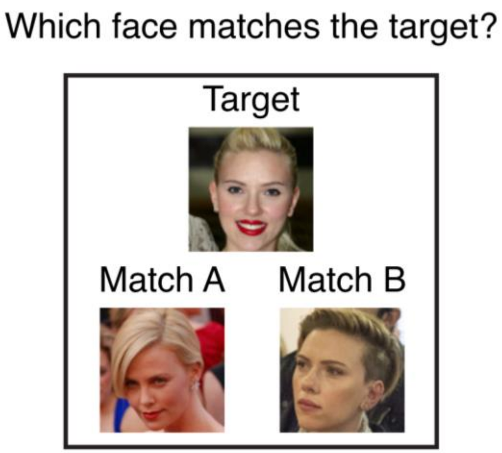


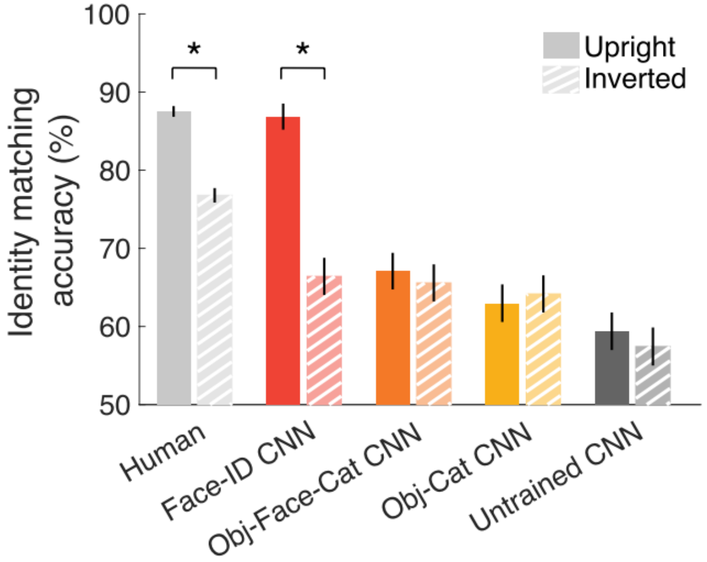

  - Objects-and-Face Categorization CNN




  - Object-Categorization CNN

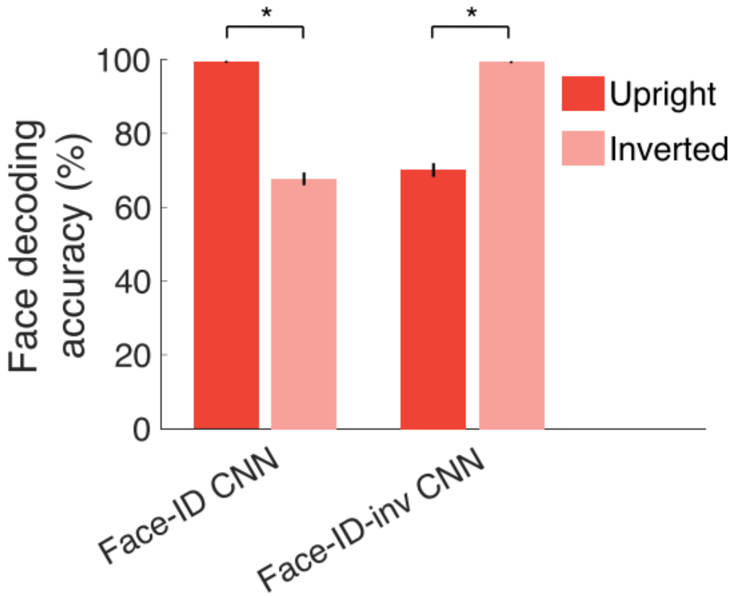


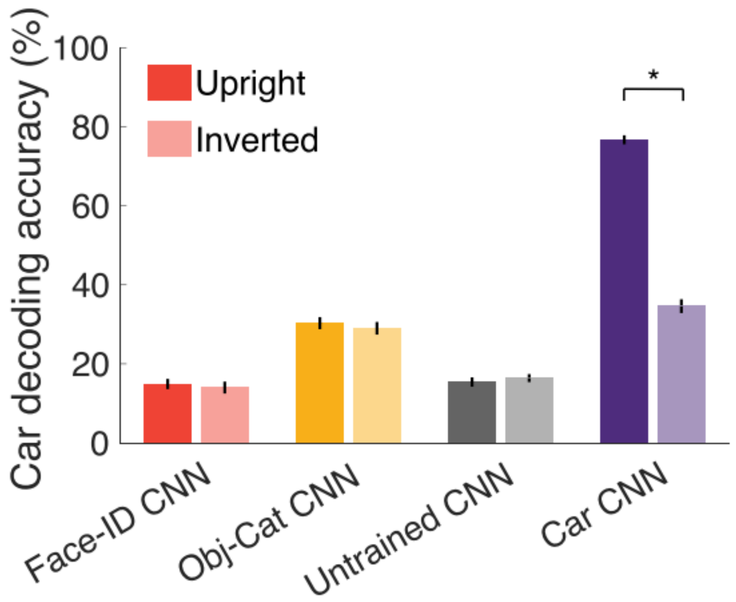



# The inversion effect in DCNNs

- But how specific is this inversion effect to (upright) faces?
- Training of two new CNNs:
  - Inverted-Face-Identification CNN
  - Car-Model-Identification CNN



# Comparing holistic processing between humans and DCNNs

# Experiment 1: Rationale and stimuli/input

- If DCNNs rely mainly on local-feature information to process faces:
  - Performance should be less affected when holistic information is degraded
  - Performance should be more affected when local-feature information is degraded
- Manipulation of test-set images of the VGGFace2 database

Original

Both local-feature and holistic information intact ➔ Baseline

Local-feature information severely degraded, holistic information intact

Mooney

Coarsely scrambled

Local-feature information intact, holistic information somewhat degraded

Local-feature information intact, holistic information substantially degraded

Finely scrambled

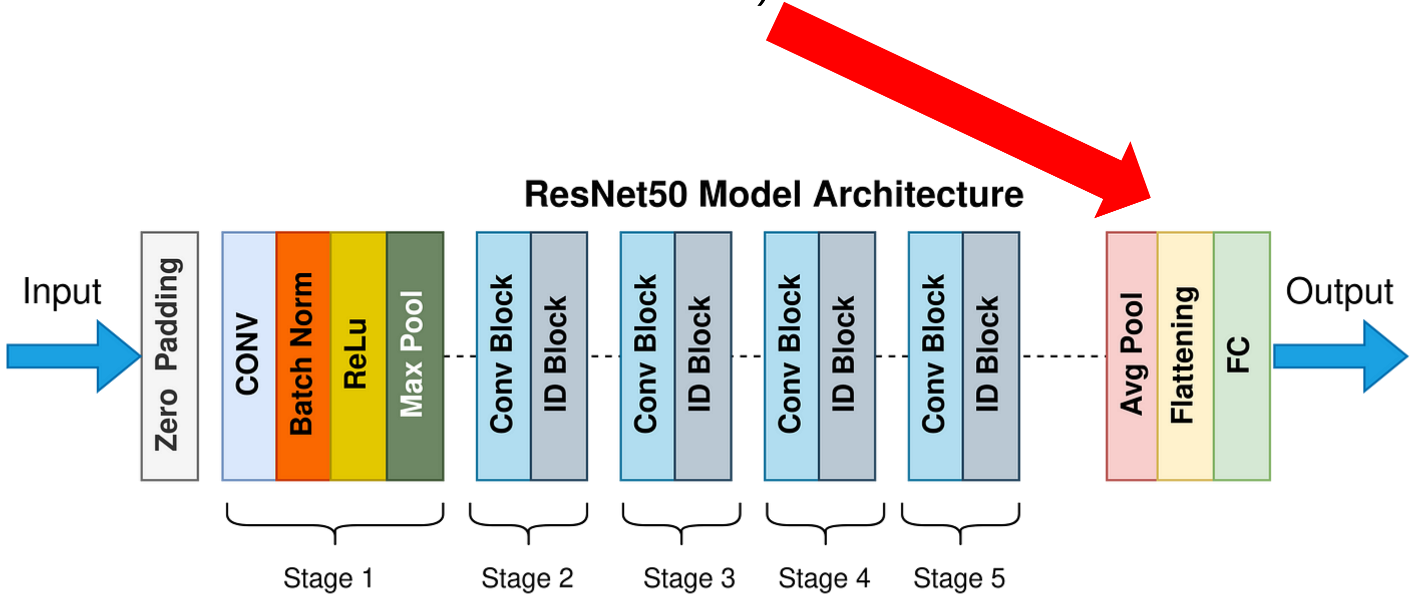
# Experiment 1: Participants and architecture

## Human participants

- *N* = 32
- Recruited via Prolific
  - Age 19-35
  - Living in Germany, Austria or Switzerland
  - Fluent in German
  - Normal or corrected-to-normal vision
  - High reputation on Prolific

## ResNet-50

- DCNN with state-of-the art ImageNet classification accuracy at time of publication (He et al., 2016)
- High accuracy in face recognition (Cao et al., 2018)
- Trained on the training set of VGGFace2 (Cao et al., 2018)
- Feature extraction at penultimate layer (before class activations are calculated)

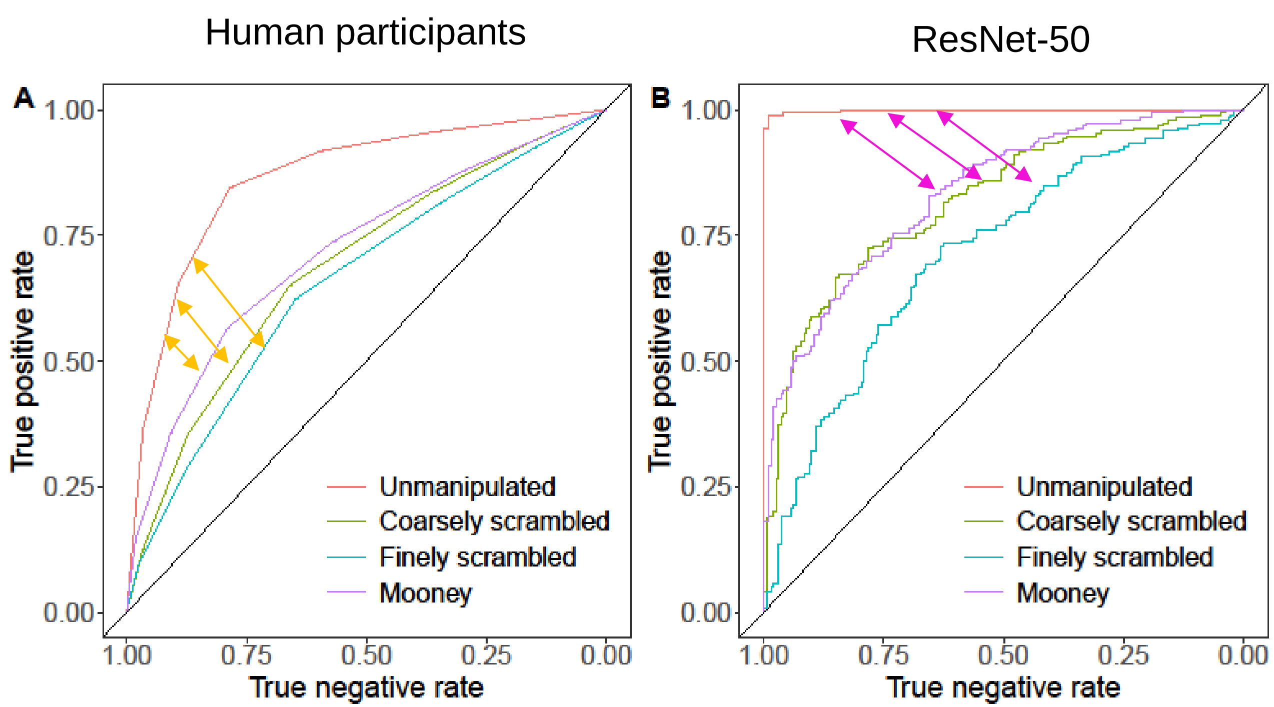


# Experiment 1: Procedure

- Unfamiliar face-matching procedure
- Judge identity of two face photographs using a six-point Likert scale
- 384 experimental trials
  - Original vs. Mooney vs. coarsely scrambled vs. finely scrambled (25% each)
  - Same identity vs. different identities (50% each)
  - Male faces vs. female faces (50% each)

# Experiment 1: Results

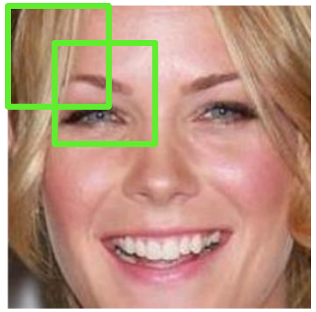
- Statistical procedure
  - Quantify the difference in matching accuracy for each participant (and ResNet-50) between:
    - Original vs. coarsely scrambled
    - Original vs finely scrambled
    - Original vs. Mooney
  - Test whether the pattern of performance decrements for human participants significantly deviates from the pattern of ResNet-50
- Result: All three manipulations more detrimental to ResNet-50 than to human participants



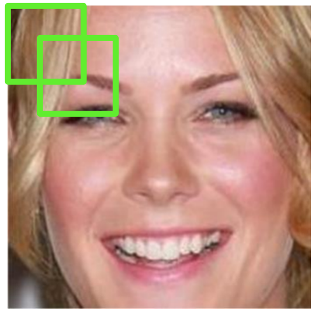
# Experiment 2: Rationale and architecture

- Restriction of holistic processing:
  - Not by manipulation of input images
  - But by manipulation of architecture
- Two new architectures based on BagNets (Brendel & Bethge, 2018)
  - BagNet-73 (receptive field of approximately 1/9 of input image)
  - BagNet-57 (receptive field of approximately 1/16 of input image)

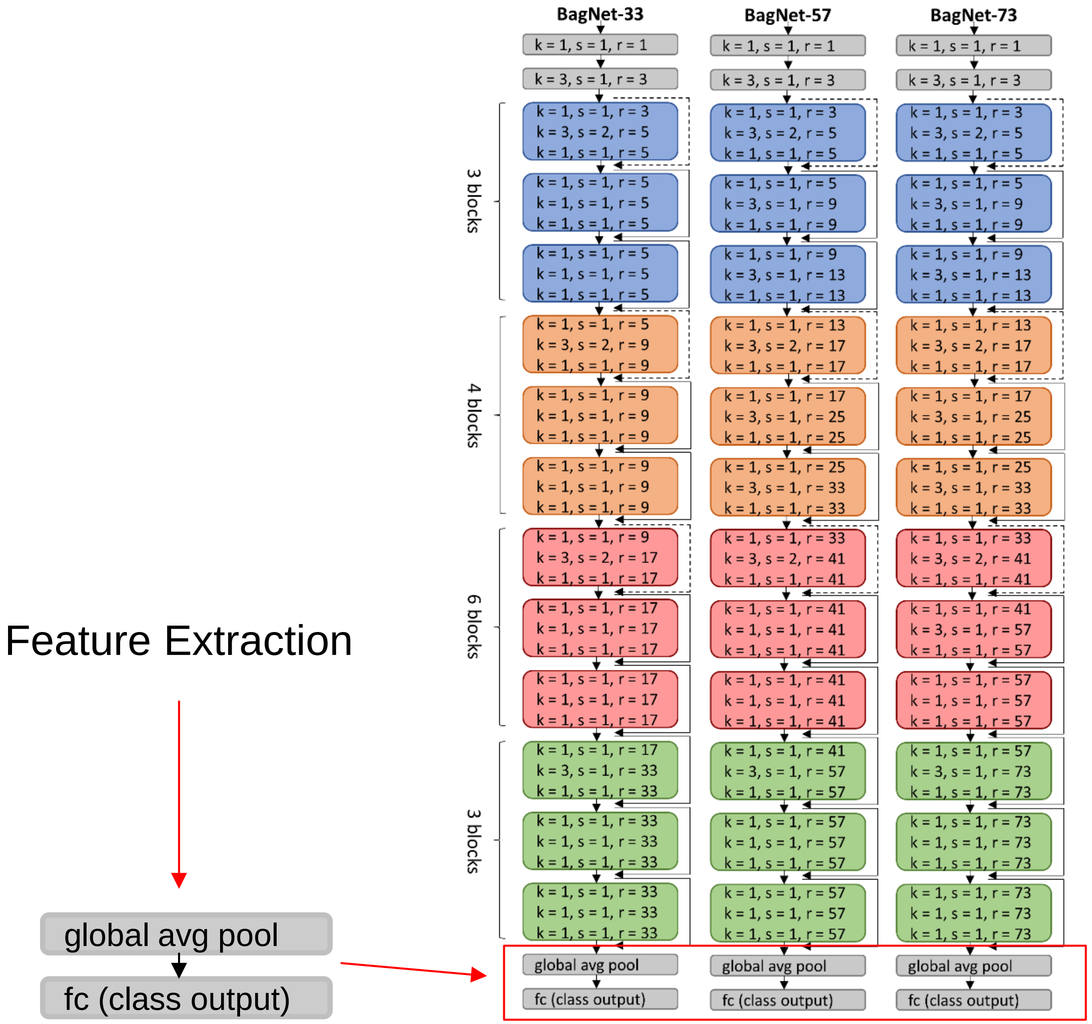
ResNet-50



BagNet-73

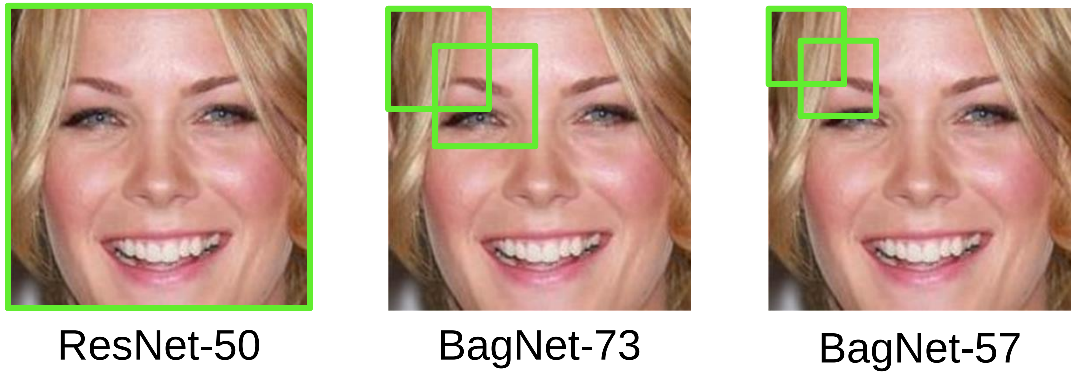


BagNet-57



# Experiment 2: Rationale and architecture

- Restriction of holistic processing:
  - Not by manipulation of input images
  - But by manipulation of architecture
- Two new architectures based on BagNets (Brendel & Bethge, 2018)
  - BagNet-73 (receptive field of approximately 1/9 of input image)
  - BagNet-57 (receptive field of approximately 1/16 of input image)




- Training on the training split of VGGFace2 using the same procedure as Cao et al. (2018):
  - ~3.1 Mio images
  - 8631 individuals
- Training parameters:
  - Three stages with learning rates of 0.1, 0.01, 0.001
  - 22 epochs per stage
  - Batch size: 256
  - Optimizer: Stochastic gradient descent

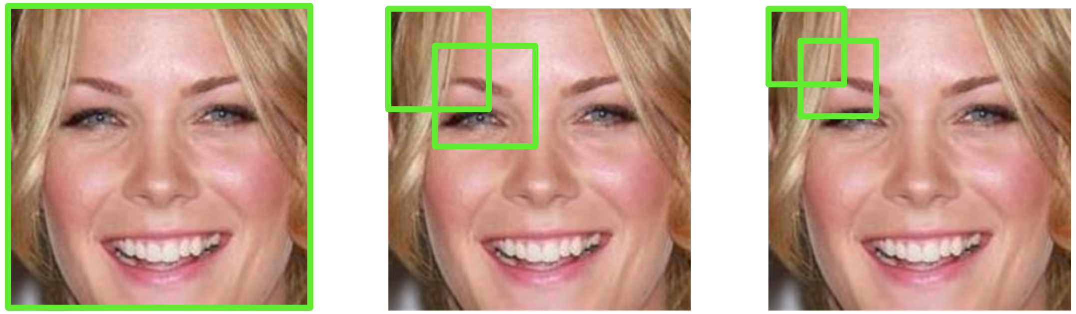
# Experiment 2: Participants and procedure

## Participants

- $N = 36$
- Recruited via Prolific
  - Age 18-35
  - Living in Germany, Austria or Switzerland
  - Fluent in German
  - Normal or corrected-to-normal vision
  - High reputation on Prolific

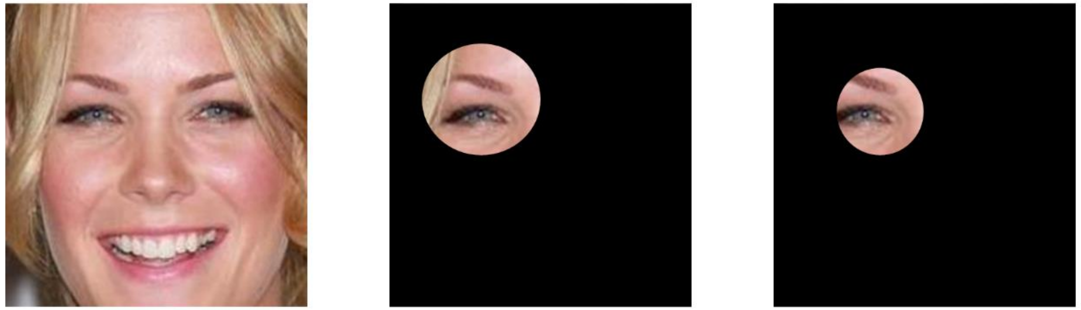
## Procedure

- How to simulate highly overlapping receptive fields?



ResNet-50 BagNet-73 BagNet-57

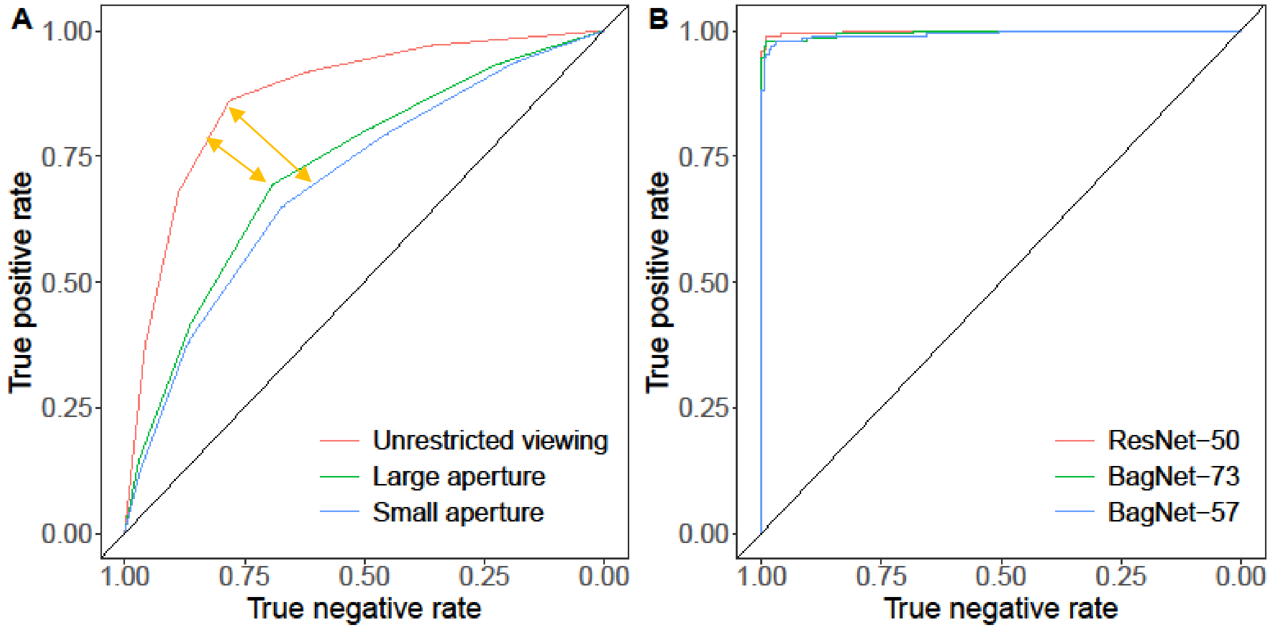
- Movable visual aperture controlled by participants



Unrestricted Large Aperture Small Aperture

# Experiment 2: Results

- Statistical procedure
  - Quantify the difference in matching accuracy for each participant (and DCNNs) between:
    - Whole image vs. receptive field / aperture of approximately 1/9 of the input image size
    - Whole image vs. receptive field / aperture of approximately 1/16 of the input image size
  - Test whether the pattern of performance decrements for human participants significantly deviates from the pattern of DCNNs
  - Result: Substantial decrements for human participants, no effect on DCNNs



# Interpretation and integration of both experiments

**Experiment 1**
Performance decrements due to perturbation of holistic (and local feature-based) information similar (even larger) for DCNNs than for human participants

**Experiment 2**
No performance decrements for DCNNs due to restricted receptive fields, but substantial decrements for human participants due to viewing apertures

**Integration**
Optimisation conditions during training critical factor

The nature of the computations that underlie perception depends more upon the computational problems that have to be solved than on the particular hardware in which their solutions are implemented. (Marr, 1982/2010, p. 29):

# Limitations of comparing cognitive processes between humans and DCNNs

**Limitations**

- Different optimisation conditions between human participants and DCNNs
- susceptibility of DCNNs to low-level image perturbations
- Feedforward DCNNs vs. feedback brains

**Potential solutions**

- High variance in training sets to mimic human developmental conditions (illumination, size, distance, perspective)
- Integration of operations typical for the human primary visual cortex into DCNNs (Dapello et al. 2020; Pogoncheff et al., 2023)
- Integration of recurrent processes in DCNNs (Mnih et al., 2014, Kubilius et al., 2019)

**But:** DCNNs as unique opportunity of modelling and manipulating the optimisation conditions of the human visual system in order to answer „why“ questions regarding the human cognitive system (Kanwisher et al., 2023)

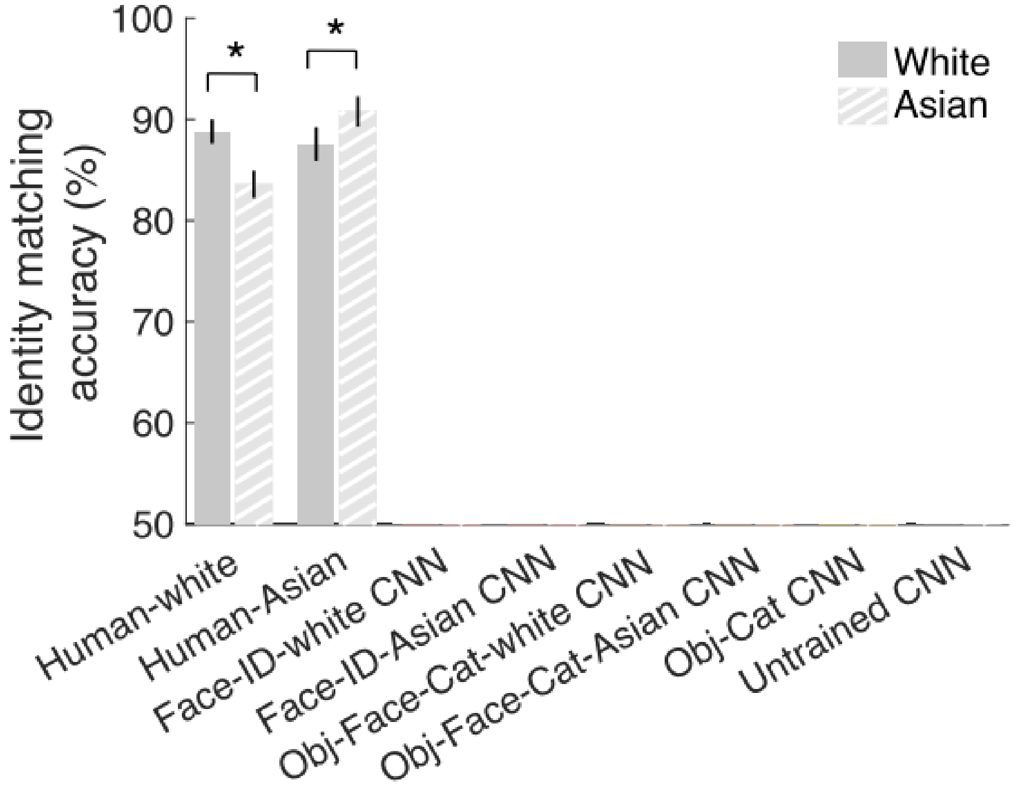
**Why artificial face recognition fails…**

**How it can be improved…**

**And what what it tells about human face recognition**

# The own-race bias in DCNNs

- Face matching task with White and Asian faces (Dobs et al., 2023)
  - White participants
  - Asian participants
  - DCNN trained to classify White faces
  - DCNN trained to classify Asian faces
  - DCNN trained to classify objects with White faces as one class
  - DCNN trained to classify objects with Asian faces as one class
  - DCNN trained to classify objects
  - Untrained DCNN



# Implicit racial bias in common face datasets

- Racial Faces in the Wild Dataset (RFW; Wang et al., 2019)
  - 6000 face pairs for a difficult matching task
  - Comprises faces from four races
    - 25% White
    - 25% Asian
    - 25% Indian
    - 25% African
- Evaluation of models trained on common face datasts on RFW
- Most likely cause of impaired performance for non-White faces: Uneven distribution in common datasets

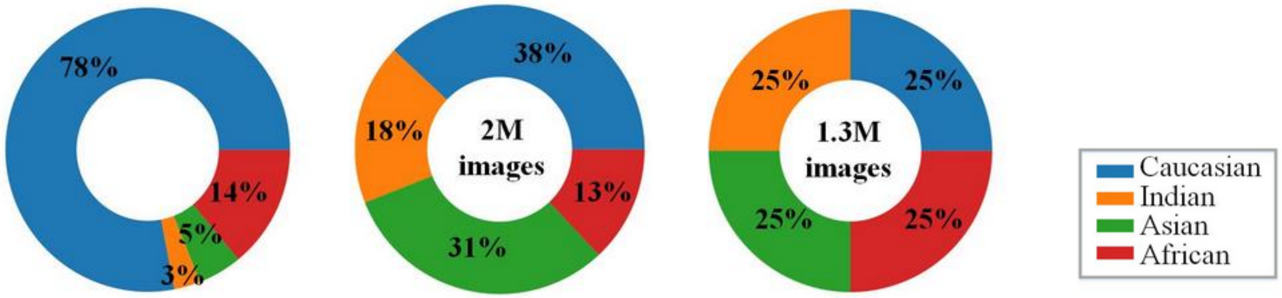
| | Model | RFW Caucasian | RFW Indian | RFW Asian | RFW African |
|---|---|---|---|---|---|
| commercial API | Microsoft [5] | 87.60 | 82.83 | 79.67 | 75.83 |
| | Face++ [4] | 93.90 | 88.55 | 92.47 | 87.50 |
| | Baidu [3] | 89.13 | 86.53 | 90.27 | 77.97 |
| | Amazon [1] | 90.45 | 87.20 | 84.87 | 86.27 |
| | mean | 90.27 | 86.28 | 86.82 | 81.89 |
| SOTA algorithm | Center-loss [65] | 87.18 | 81.92 | 79.32 | 78.00 |
| | Sphereface [39] | 90.80 | 87.02 | 82.95 | 82.28 |
| | Arcface[1] [21] | 92.15 | 88.00 | 83.98 | 84.93 |
| | VGGface2 [15] | 89.90 | 86.13 | 84.93 | 83.38 |
| | mean | 90.01 | 85.77 | 82.80 | 82.15 |

[1] Arcface here is trained on CASIA-Webface using ResNet-34.

| Train/ Test | Database | Racial distribution (%) Caucasian | Asian | Indian | African |
|---|---|---|---|---|---|
| train | CASIA-WebFace [67] | 84.5 | 2.6 | 1.6 | 11.3 |
| | VGGFace2 [15] | 74.2 | 6.0 | 4.0 | 15.8 |
| | MS-Celeb-1M [30] | 76.3 | 6.6 | 2.6 | 14.5 |
| test | LFW [33] | 69.9 | 13.2 | 2.9 | 14.0 |
| | IJB-A [37] | 66.0 | 9.8 | 7.2 | 17.0 |
| | RFW | **25.0** | **25.0** | **25.0** | **25.0** |

# Mitigating racial bias in DCNNs

- Creation of racially balanced face datasets (Wang & Deng, 2020)
- Training of models on BUPT-Globalface using reinforcement learning with adaptive margins
- Evaluation on RFW dataset



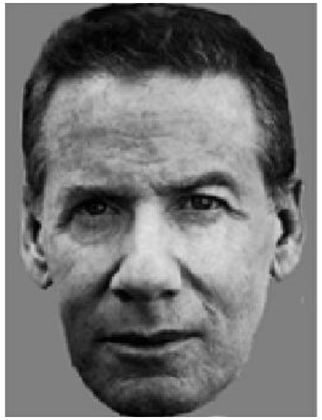

(a) Existing training datasets (b) BUPT-Globalface (c) BUPT-Balancedface

| Methods | Caucasian | Indian | Asian | African |
|---|---|---|---|---|
| Softmax | 95.62 | 91.97 | 90.85 | 89.98 |
| M-RBN(soft) | 93.50 | 94.50 | 90.06 | 93.43 |
| RL-RBN(soft) | 94.53 | 95.03 | 94.20 | 94.05 |
| Cosface [47] | 96.63 | 94.68 | 93.50 | 92.17 |
| M-RBN(cos) | 96.15 | 95.73 | 93.43 | 94.76 |
| RL-RBN(cos) | 96.03 | 95.15 | 94.58 | 94.27 |
| Arcface [13] | 97.37 | 95.68 | 94.55 | 93.87 |
| M-RBN(arc) | 97.03 | 95.58 | 94.40 | 95.18 |
| RL-RBN(arc) | 97.08 | 95.63 | 95.57 | 94.87 |

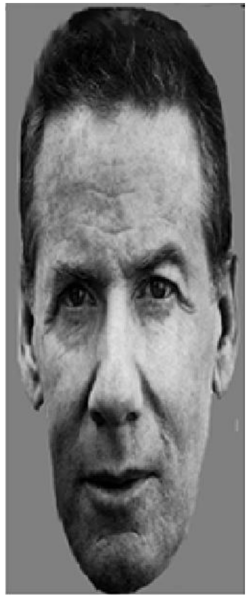
# Discussion & Outlook

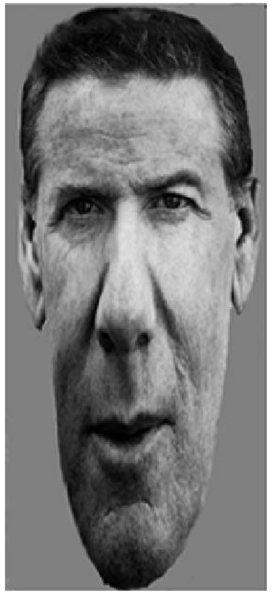
- By comparing face processing between human and DCNNs, we can
  - Advance our understanding of human face processing
  - Improve artificial face processing
  - Potentially better coordinate the collaboration between humans and artifical face recognition

- Further avenues for future resarch could include
  - Testing the limits of holistic face processing
    - How small can receptive field sizes get before substantial performance decrements occur?
    - Can DCNNs trained on scrambled faces achieve similar matching performance as DCNNs trained on whole faces?
  - New innovative ways to manipulate holistic processing in input images



Hole et al. (2002)

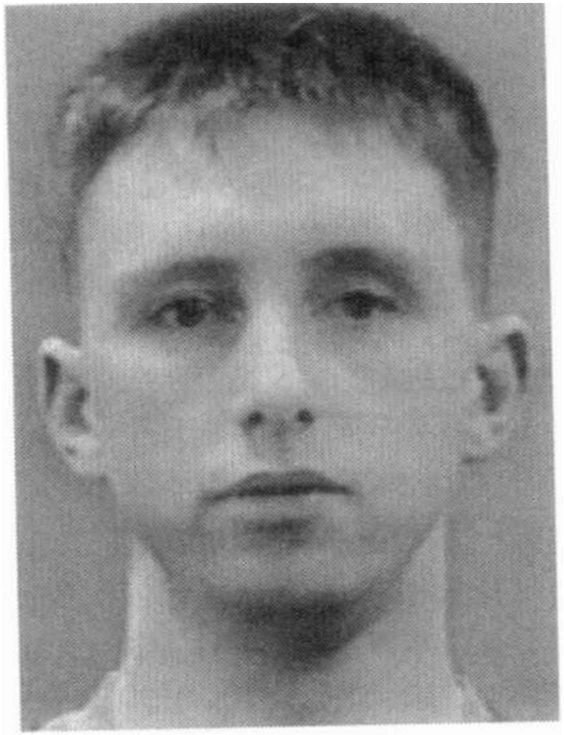


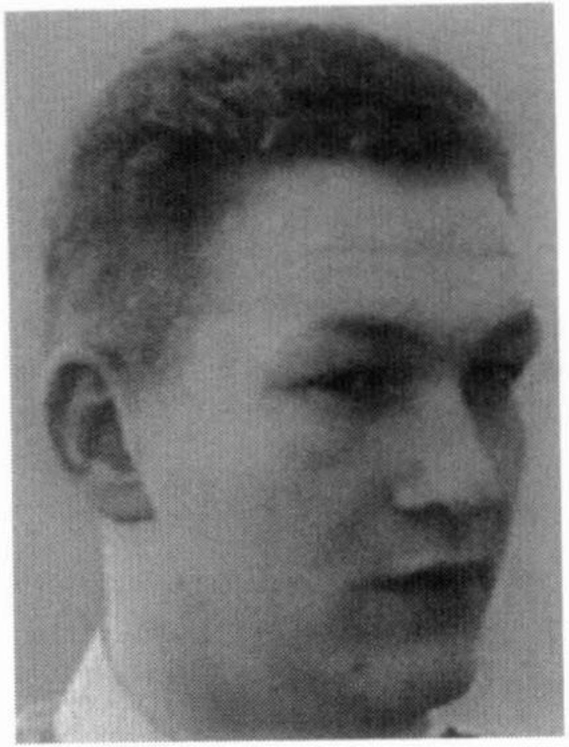


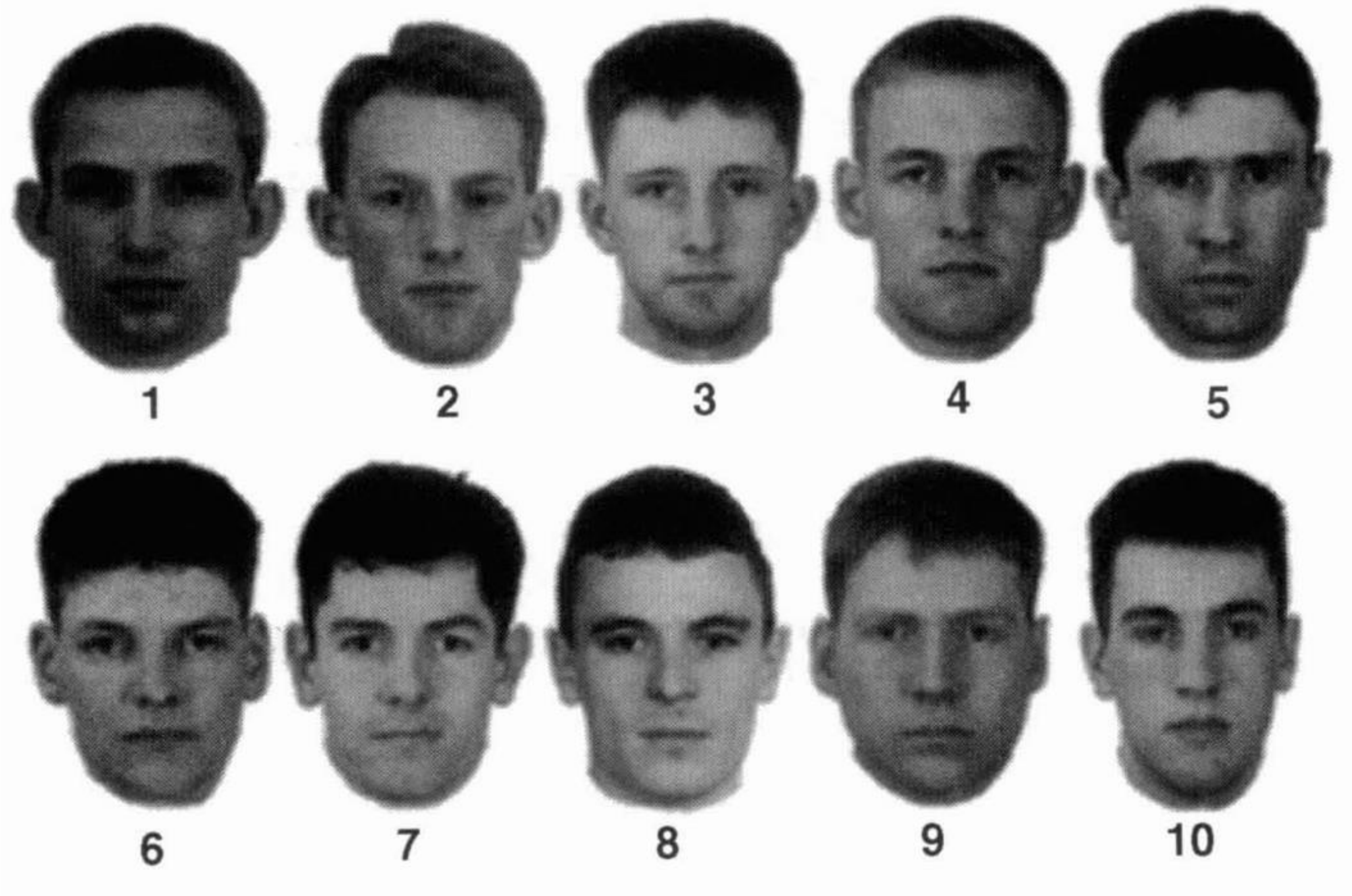
Thanks for your attention!

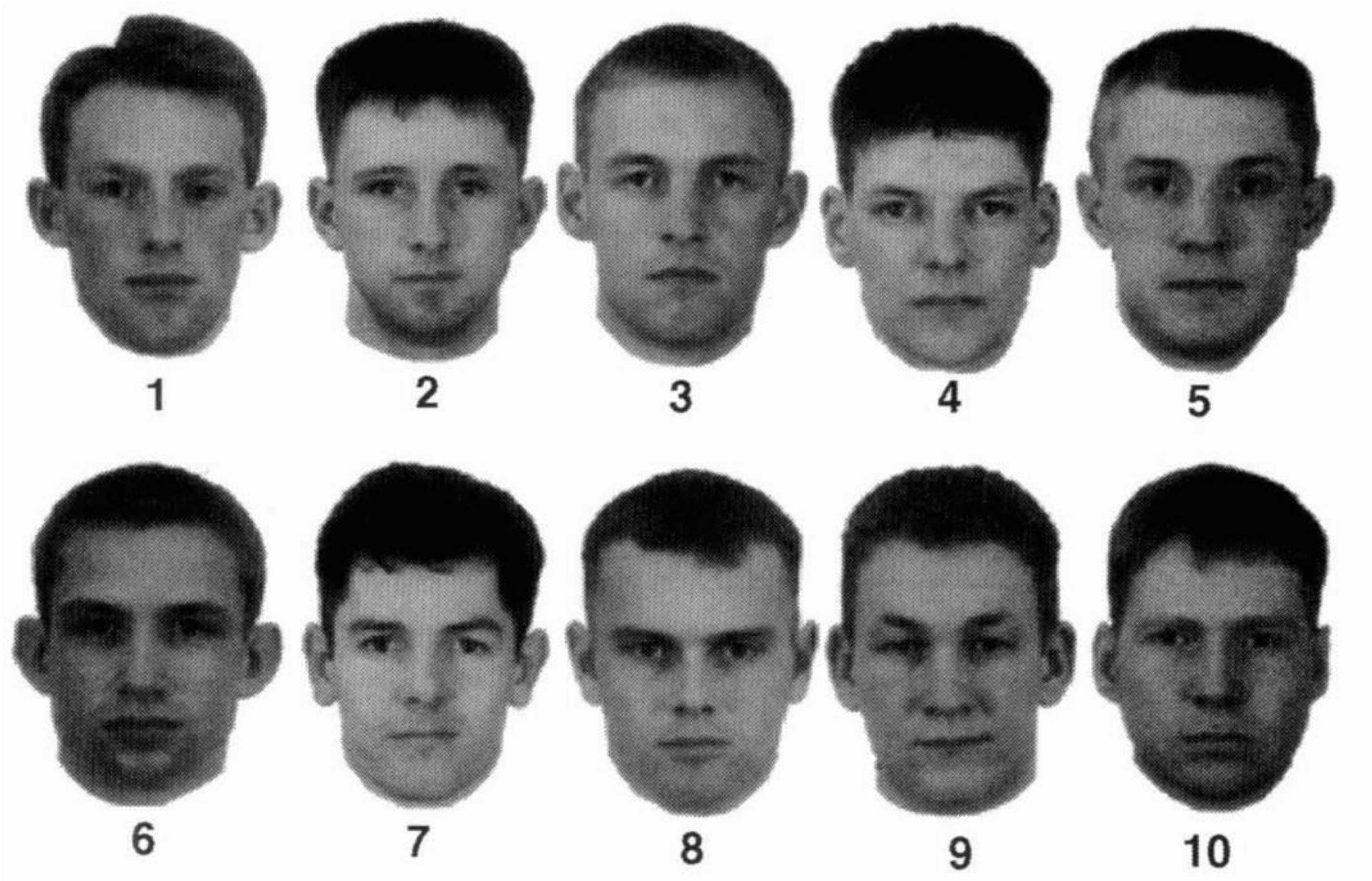
# Group tasks

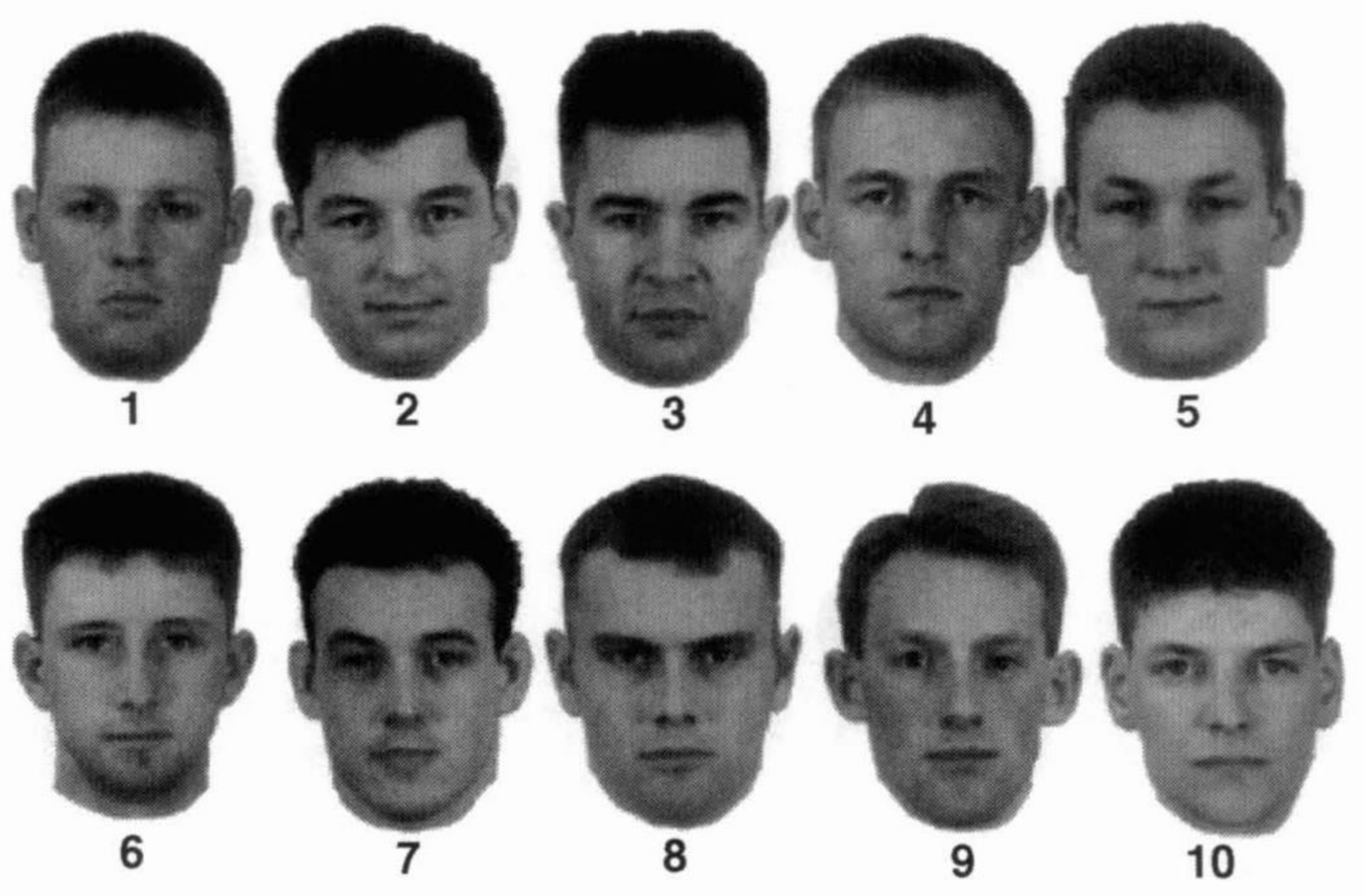
1. It is difficult to describe faces verbally. Why do you think this is the case? How could we use AI to improve the verbal description of faces?
2. We saw that it is easy to tell whether two cars are the same or not, but difficult to tell whether two faces are the same or not. Try to describe in your own words: What are the crucial differences between cars and faces (or other objects)?
3. We are able to recognize familiar faces sometimes after we have not seen them for decades, but it can be difficult to tell whether to faces that are presented simultaneously show the same person or not. Why do you think is that the case?
4. TOP SECRET Task: Meet me in one of the separate rooms to get your assignment





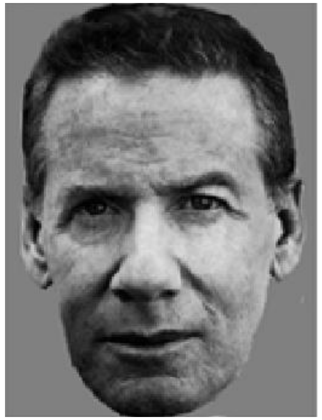

1
2
3
4
5
6
7
8
9
10


1
2
3
4
5
6
7
8
9
10


1
2
3
4
5
6
7
8
9
10

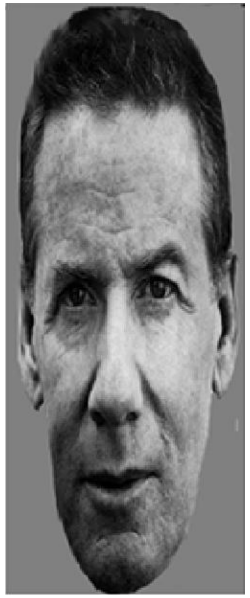
# Outlook

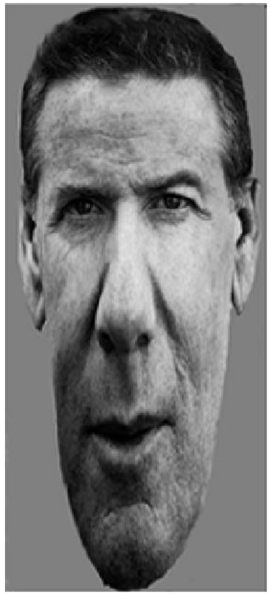
- To what degree is holistic processing necessary for face recognition ?
  - Training of original BagNets (9, 17, 33) on VGGFace2
  - Training of ResNet-50 on scrambled faces



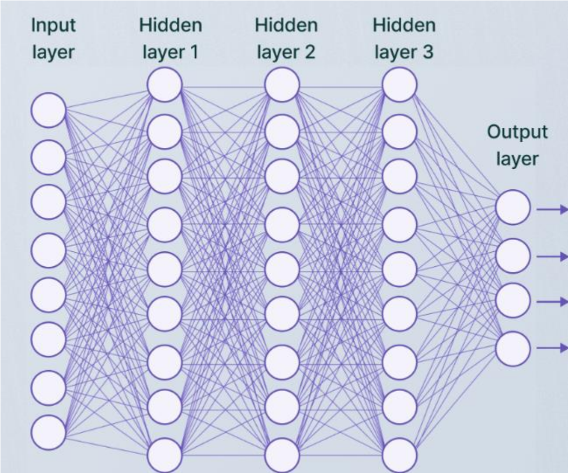
Hole et al. (2002)

- Performance decrements of DCNNS in Experiment 1 due to perturbation of high-level information (holistic, local features) or low-level image properties?
  - Comparison of face matching performance for scrambled faces between ResNet-50 and BagNets
  - Manipulation holistischer/lokaler Merkmals-information ohne inhärente low-level Artefakte



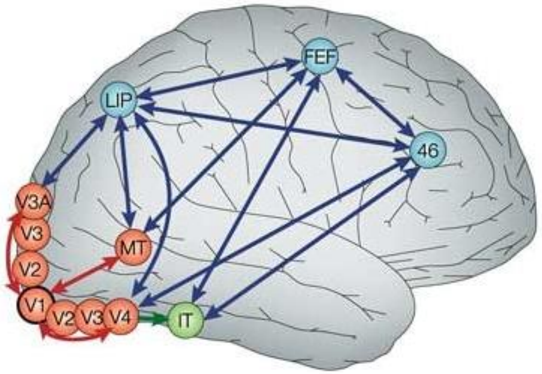


# DCNNs as a model for the human visual system




Banana

abstraction




Banana

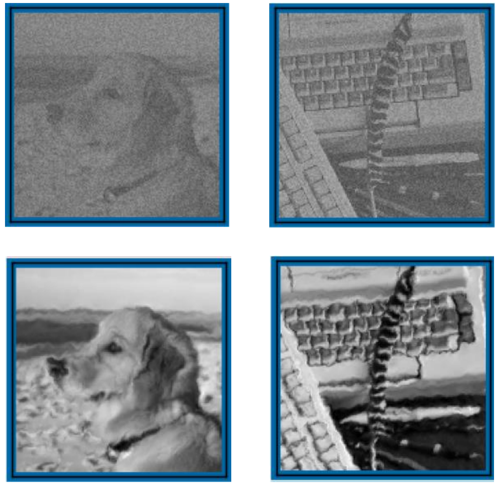
Critics:
DCNNs not a good model of human cognition due to large number of model parameters ➔ replacing one black box with another black box

Proponents:
DCNNs can provide valuable insights under experimental manipulation of model architecture, learning algorithms, and characteristics of the input

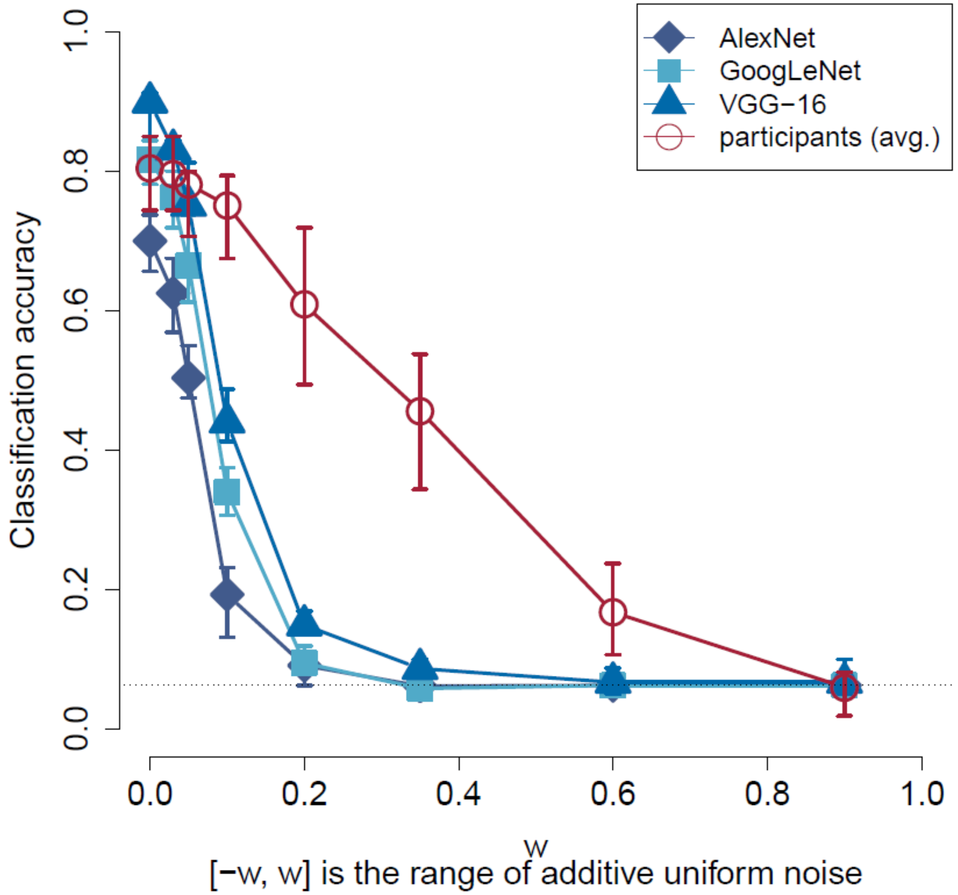
Specific manipulations of the input should lead to similar behaviour for humans and DCNNs.

# Differences between humans and DCNNs in object recognition

- Comparison of recognition accuracy for 16 ImageNet classes (Geirhos et al., 2018)
- Restricted viewing conditions for humans:
  - Image presented for 200 ms
  - Masked with pink noise for 200 ms
- Four different image manipulations
  - Greyscale
  - Reduced Contrast
  - Added noise
  - Eidolon



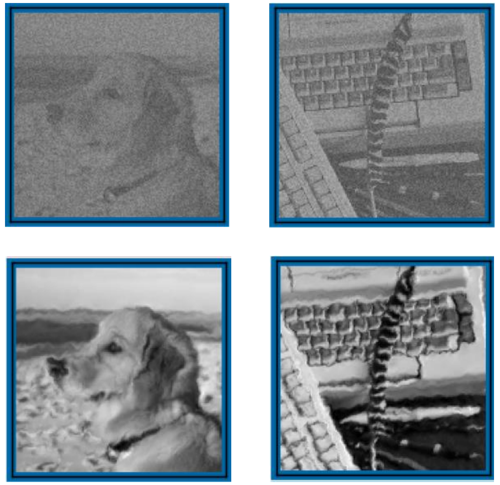
➔ Human object recognition more robust than DCNN recognition



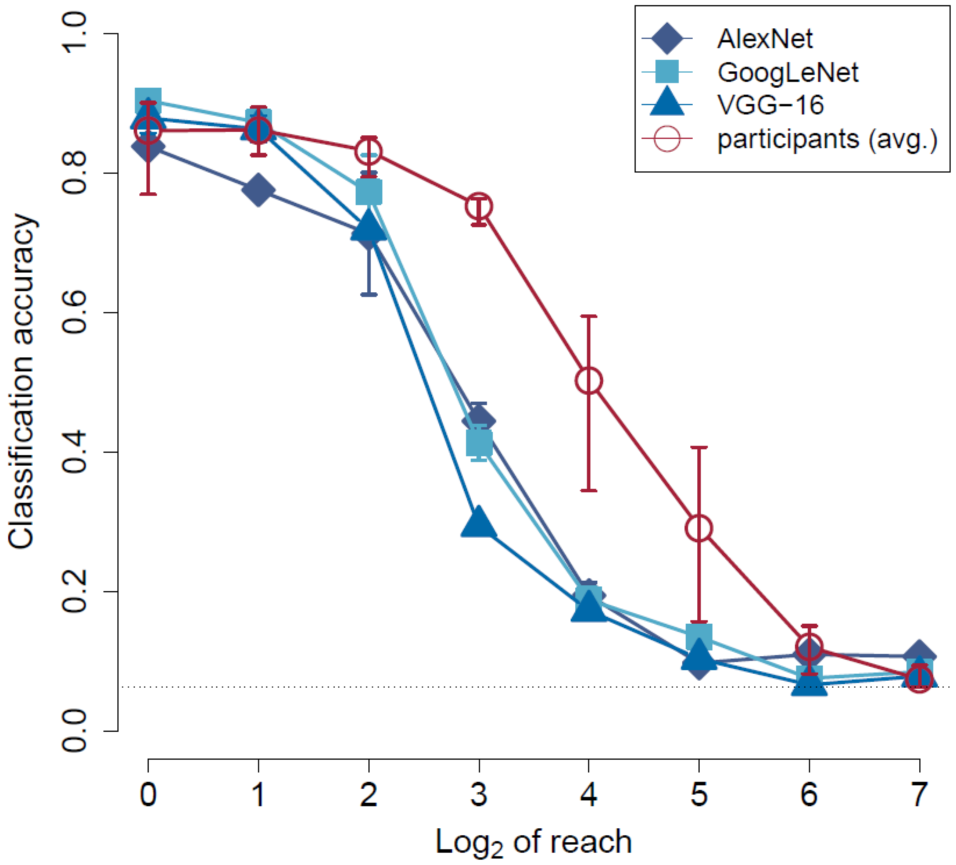

50%-Accuracy examples

# Differences between humans and DCNNs in object recognition

- Comparison of recognition accuracy for 16 ImageNet classes (Geirhos et al., 2018)
- Restricted viewing conditions for humans:
  - Image presented for 200 ms
  - Masked with pink noise for 200 ms
- Four different image manipulations
  - Greyscale
  - Reduced Contrast
  - Added noise
  - Eidolon



➔ Human object recognition more robust than DCNN recognition

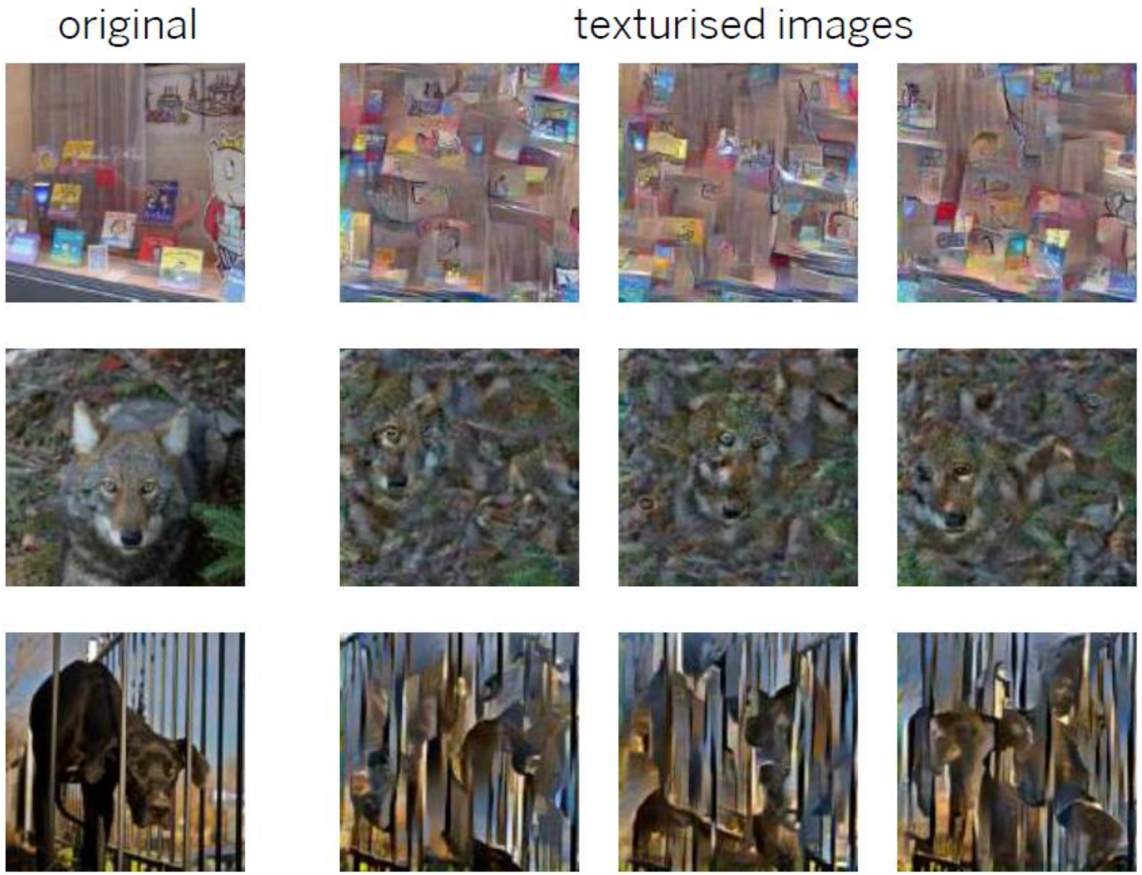



50%-Accuracy examples

# Local features versus global shape in DCNN object recognition

- Creation of a bag-of-local features model „BagNets“ (Brendel & Bethge, 2019)
- Modification of ResNet-50 to reduce size of receptive field of the topmost convolutional layer to q × q pixels with q ∈ {9, 17, 33}
- Surprisingly high classification accuracy on ImageNet
- High correlation between calss activations of VGG-16 and BagNets
- Surprisingly high accuracy of VGG-16 on scrambled ImageNet pictures: 79.4% top-5 accuracy (vs. 90.1% on normal pictures)

➔ Object recognition in DCNNs mainly dependent on local features (image patches); global shape (gestalt) largely ignored



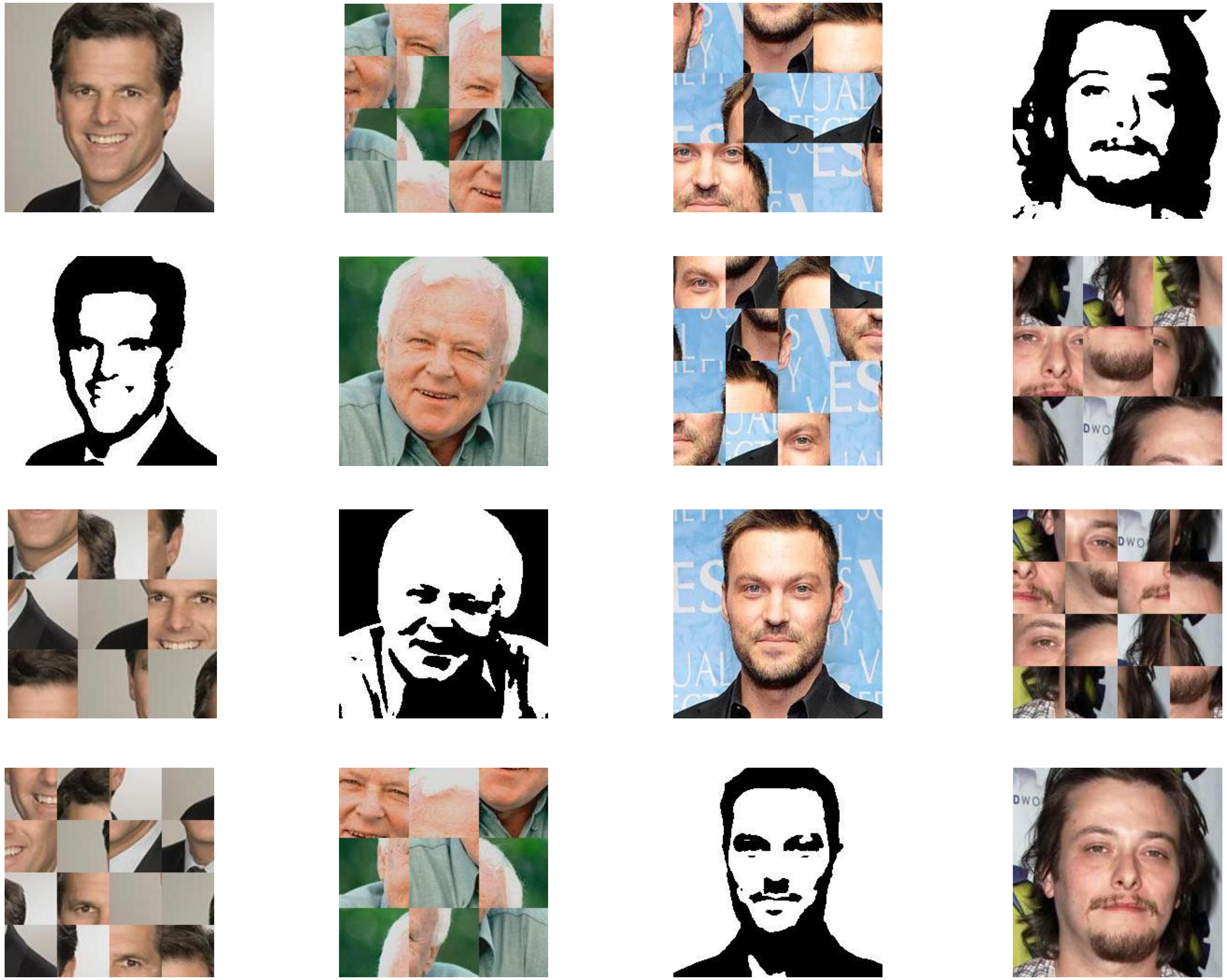
# Stimulus rotation groups

Rotation Group 1

Rotation Group 2

Rotation Group 3

Rotation Group 4



Goal: A Given participant sees each celebrity only in one condition

# Control of human participants‘ motivation and ability

- 40 GFMT trials
- Glasgow Face Matching Test (Burton et al., 2010)
- Psychometric test to measure individual unfamiliar-face-matching ability
- Participants with less than 60% accuracy excluded

Same identity

Different identity

# Creation of Mooney Faces

- Creation of 35 Mooney candidates for each image:
  - Application of Gaussian Filter with $\sigma \in \{10, 15, 20, 25, 30\}$
  - Binarization with threshold $t \in \{0.4, 0.45, 0.475, 0.5, 0.525, 0.55, 0.6\}$
- Selection of optimal candidate:
  - No texture information left
  - Still some information about eye region left