# Cognitive systems

PERCEPTION: VISION

HUMAN VS. AI (COMPUTER VISION)

**Vnímání** (též percepce) zachycuje to, co v daný okamžik působí na smysly, informuje o vnějším světě (barva, chuť) i vnitřním (bolest, zadýchání). **Vnímání** je subjektivním odrazem objektivní reality v našem vědomí prostřednictvím receptorů. Umožňuje základní orientaci v prostředí, respektive v aktuální situaci.

Otázka: Proč subjektivní? Co vše je dělá subjektivní?

# Perception

Example image that illustrates the complexity of recognizing chairs as no single technique in terms of contours, appearance, components etc. is adequate to correctly allow 'counting of the number of chairs'. (Source: Bülthoff, Max Planck Institute for Biological Cybernetics (MPIK), Tübingen, Germany.)

# Vision

# Vision

# Vision

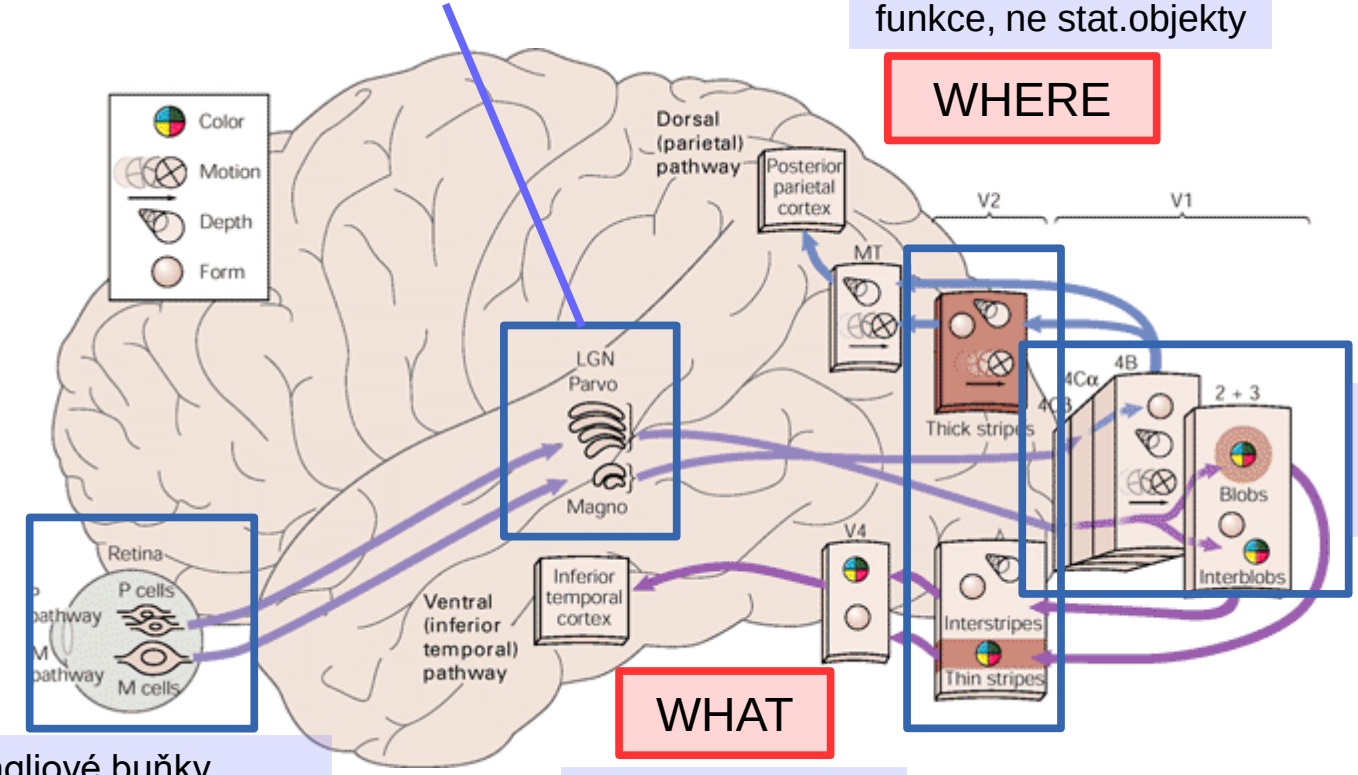
Human vision:

- 50% neural tissue
- 66% activity

Thalamus, Corpus geniculatum laterale
První reorganizace info z nerv.vzruchu

**Dorsální dráha**
Visuoprostorové funkce, ne stat.objekty

WHERE




Primární viz.kůra
**V1**

Gangliové buňky
M buňky – 8% → M dráhy ,
P buňky 80% → P dráhy

WHAT

**Ventrální dráha**
Citlivé na obrysy,
Velké rozlišení

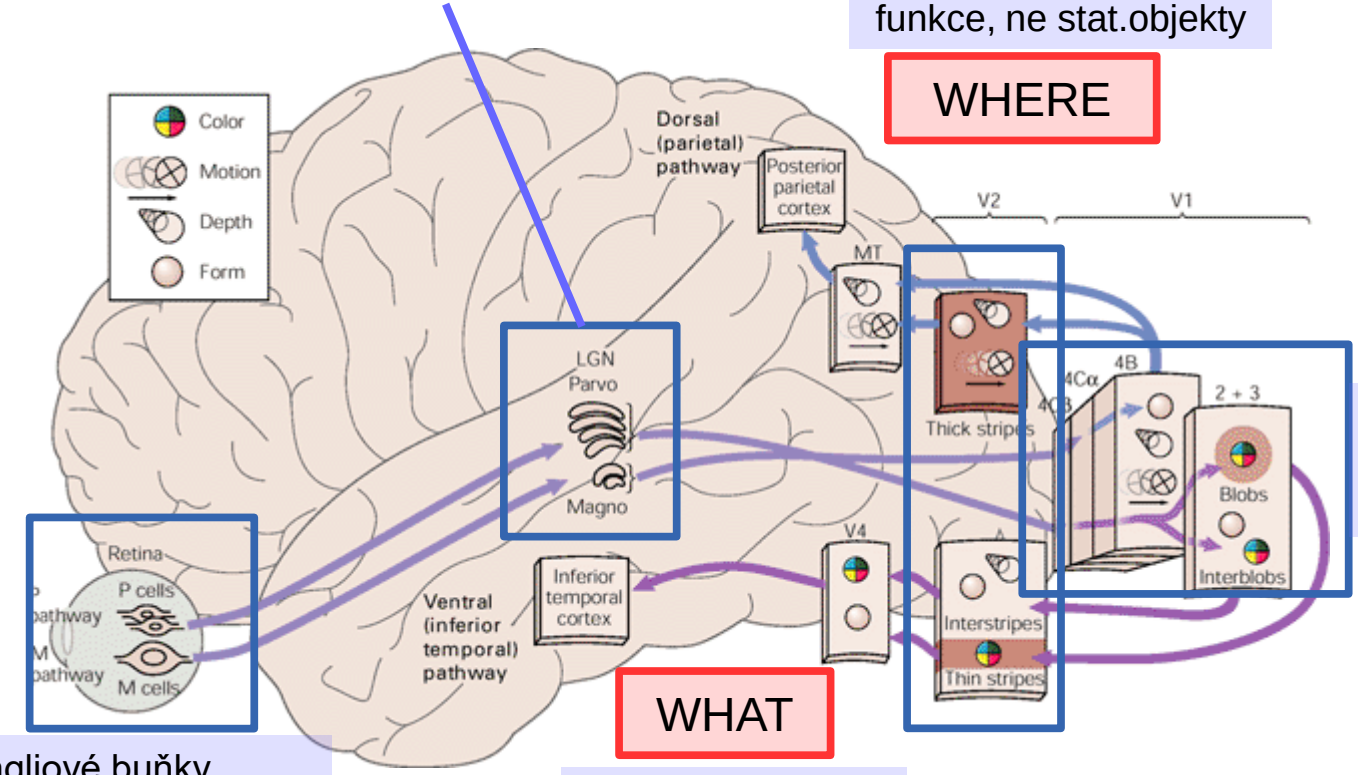
Human vision:

- 50% neural tissue
- 66% activity

Thalamus, Corpus geniculatum laterale
První reorganizace info z nerv.vzruchu

**Dorsální dráha**
Visuoprostorové funkce, ne stat.objekty

WHERE



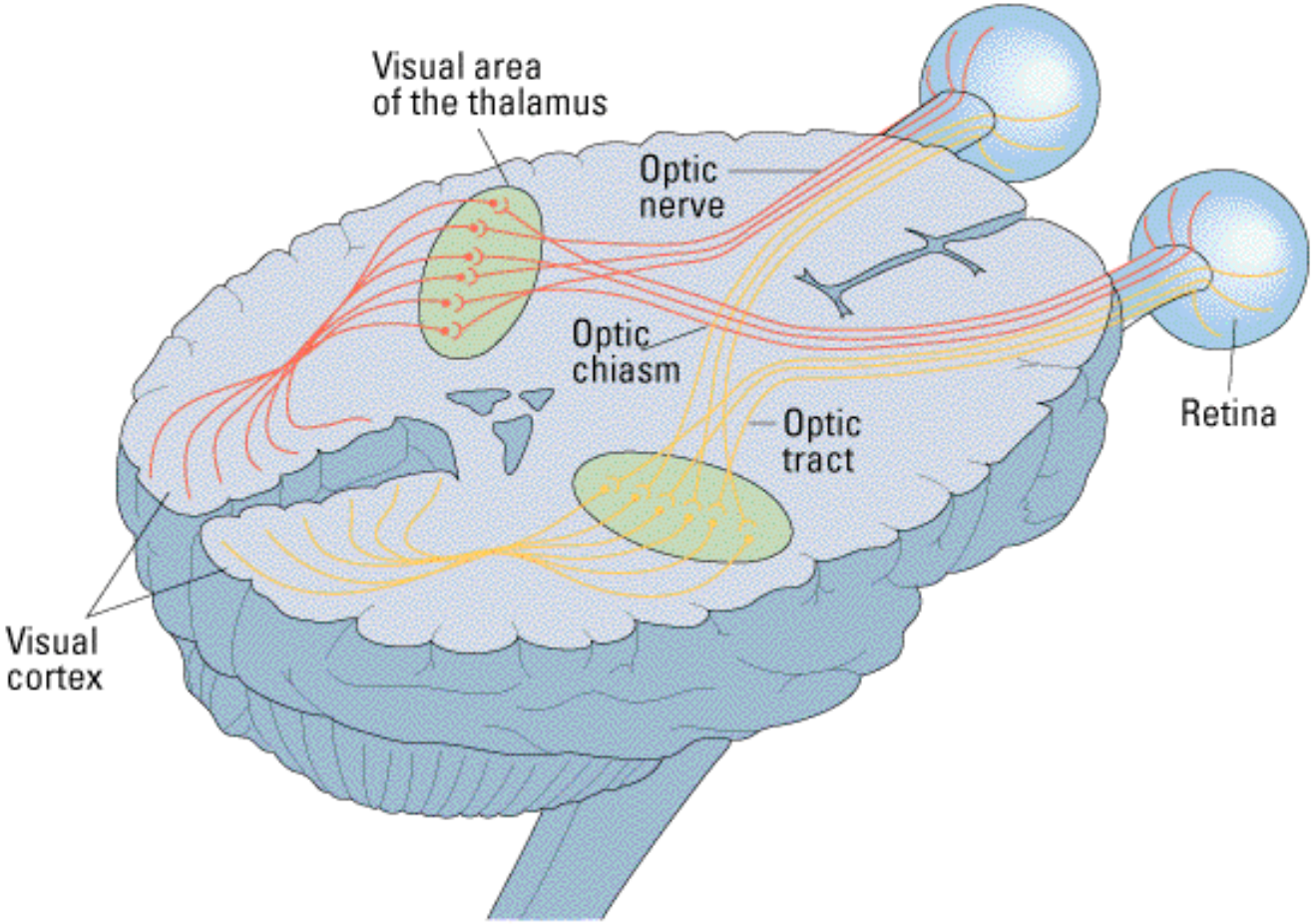

Primární viz.kůra **V1**

Gangliové buňky
M buňky – 8% → M dráhy ,
P buňky 80% → P dráhy

WHAT

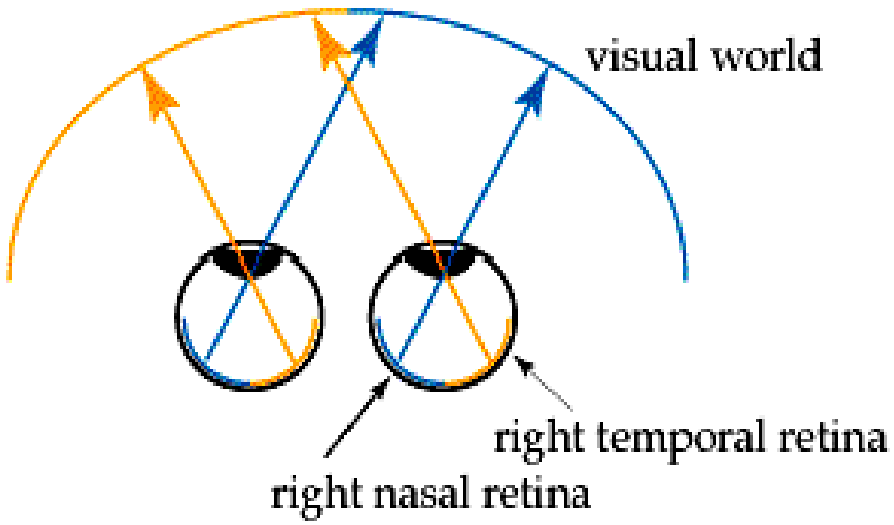
**Ventrální dráha**
Citlivé na obrysy,
Velké rozlišení

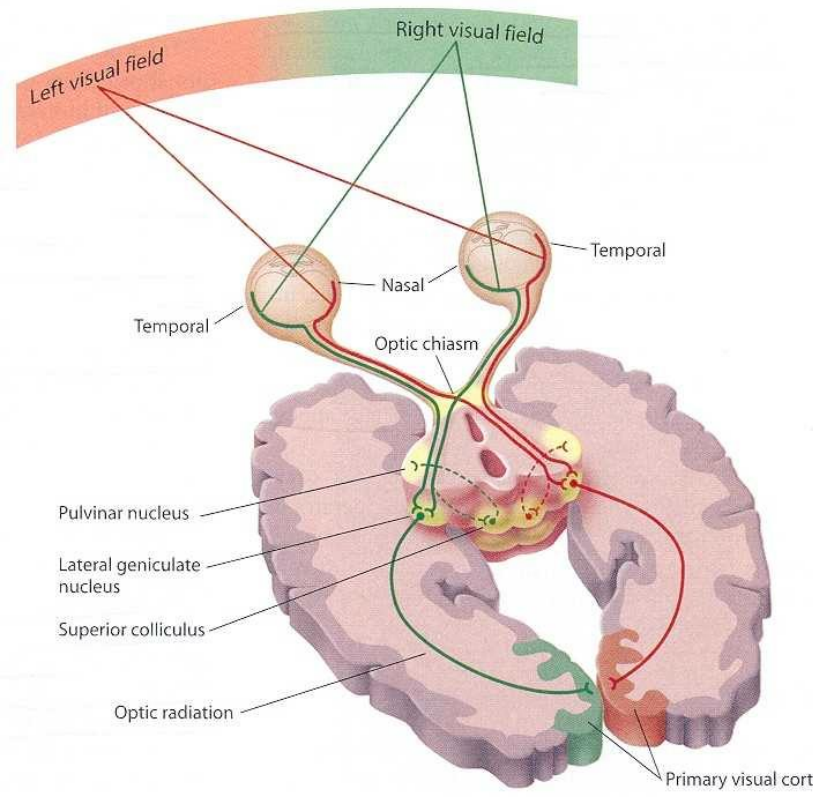
# Nervové zpracování

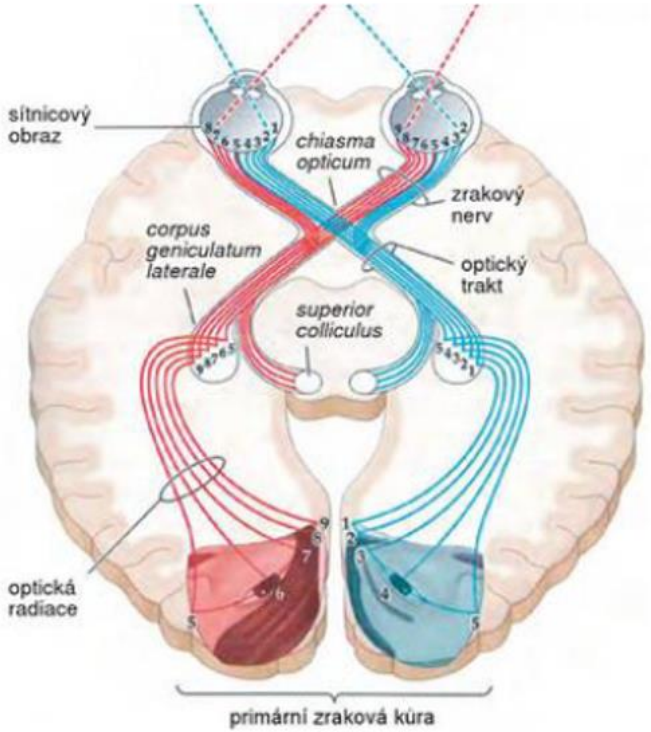


# Chiasma optikum

Překřížení drah optického nervu, aby byla zpracovávána separátně pravá a levá část zorného pole



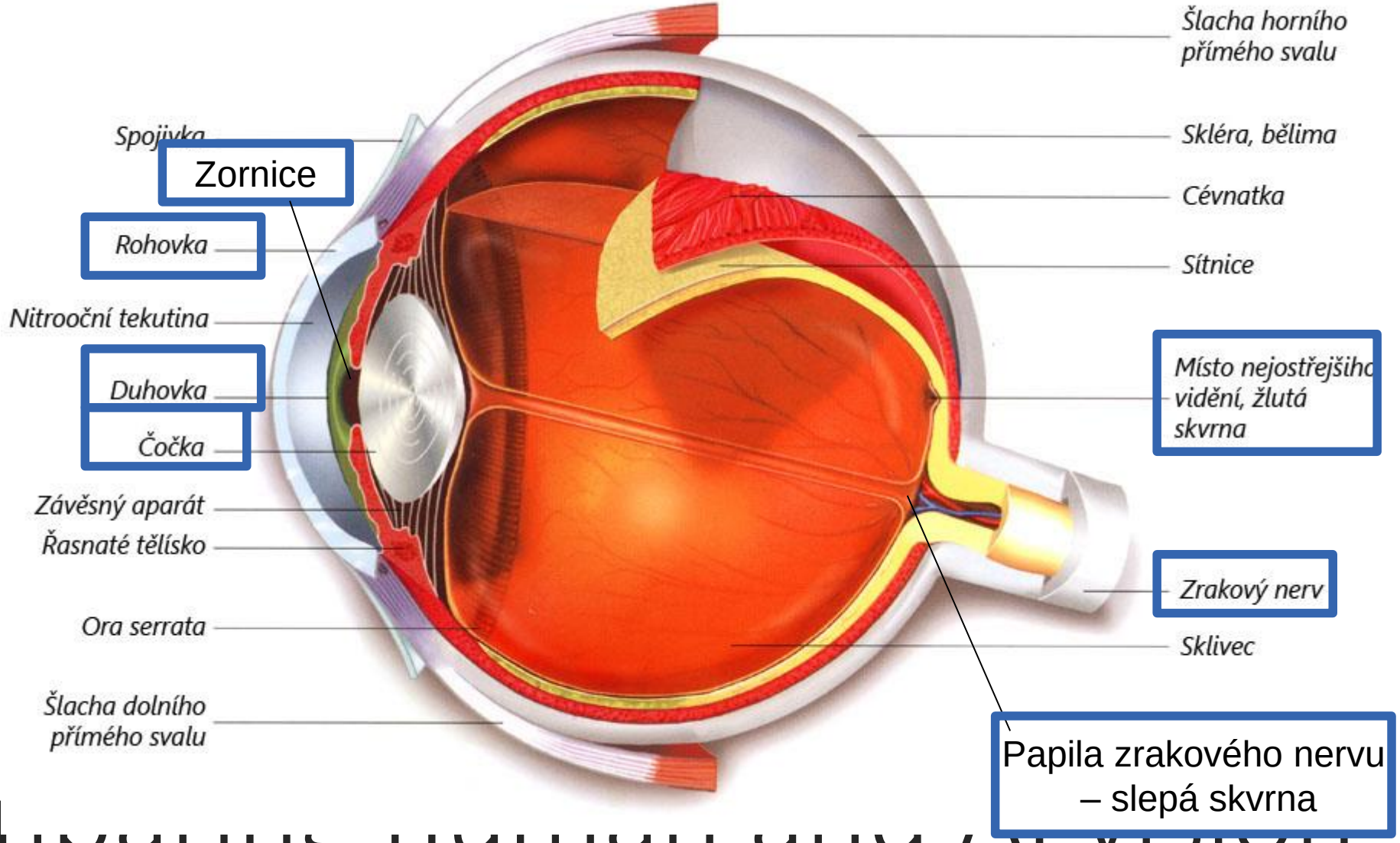



(1)the neuron fired only the line was in a particular place on the retina,
(2)the activity of the neuron changed depending on the orientation of the line, and
(3) sometimes the neuron fired only when the line was moving in a particular direction (different neurons respond to different angles).

Simple and complex cells

https://www.youtu

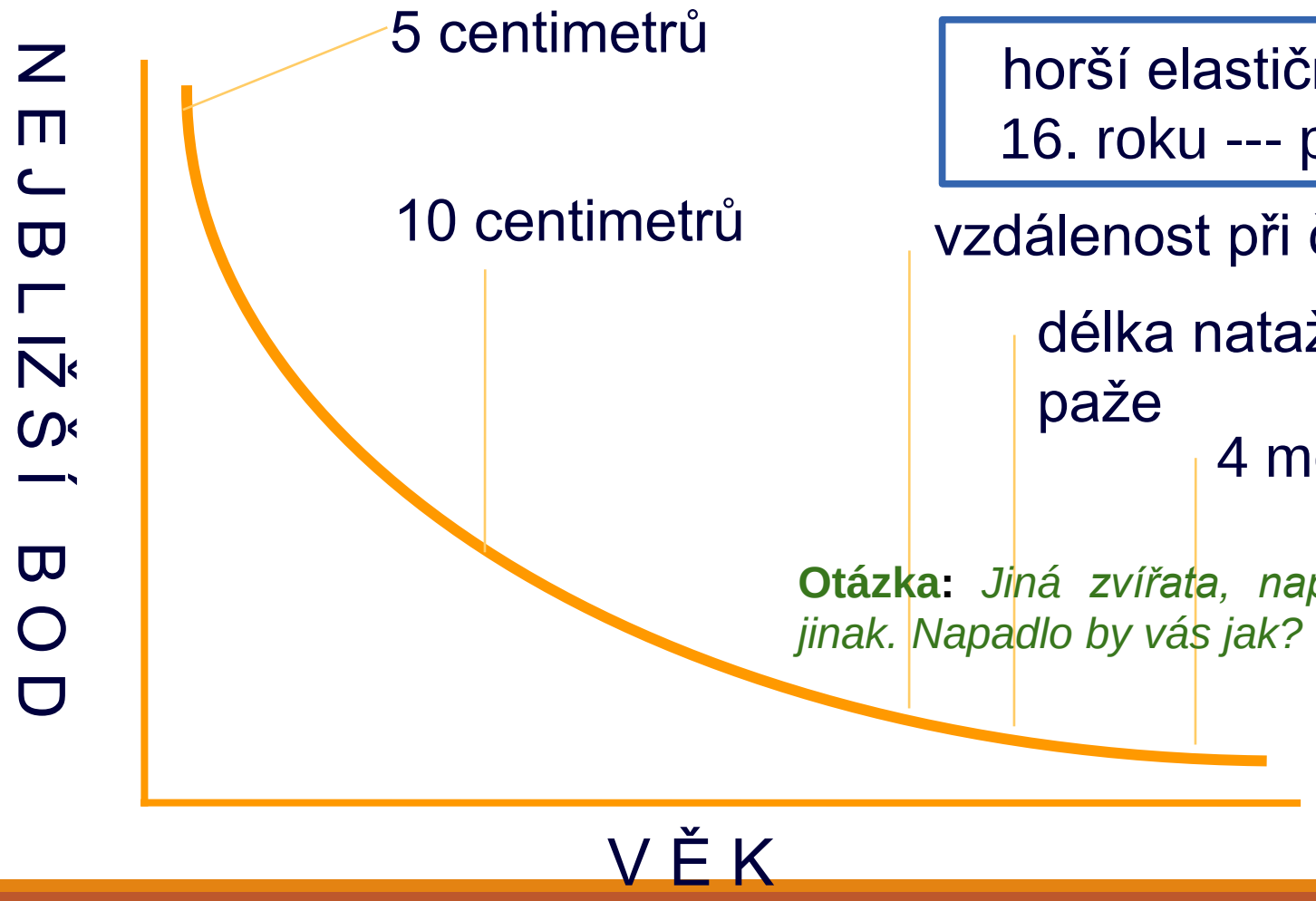
# Human vision –




# Comparing human and AI vision

# Akomodace čočky

**Optická soustava:**
rohovka (usměrnění) → komorový mok → zornice → čočka → sklivec → sítnice




horší elastičnost už od 16. roku --- presbyopie

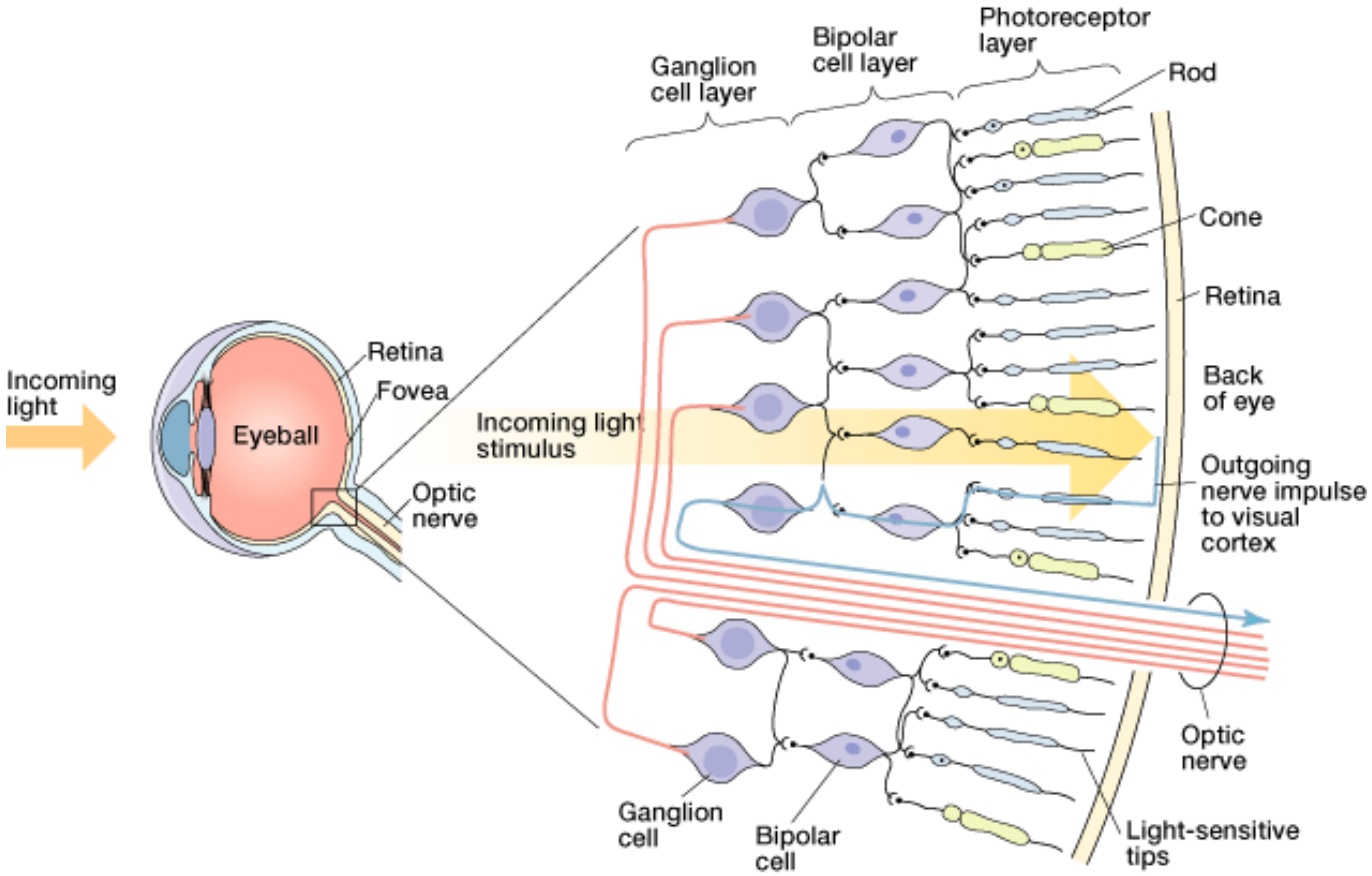
**Otázka:** *Jiná zvířata, například ryby, akomodují čočku jinak. Napadlo by vás jak?*

# Akomodace čočky

**Optická soustava:**
rohovka (usměrnění) → komorový mok → zornice → čočka → sklivec → sítnice

## Problémy:

- Krátkozrakost
- Dalekozrakost
- Strabismus Co ho způsobuje?
    - různá délka očních svalů
    - → ukázka úloh na léčbu https://cyber.felk.cvut.cz/novakpe/Osobni/Info/Strabismus.pdf




# Fotoreceptors

- Za sítnicí pigmentový epitel – absorpce (melanin)
- Axony nemyelizované → transparentní
- ve žluté skvrně ostatní buňky na straně

# Fotoreceptory

## Tyčinky

- 100-120 miliónů
- Umožňují rozlišení světla a tmy, resp. odstínů šedi
- Funkční při nižším osvětlení – 6 fotonů (větší zesílení)
- Nejcitlivější asi 20° od žluté skvrny
- Ve žluté skvrně absentují
- Plná adaptace na tmu trvá 20-30 minut
- Min.12 Hz změny, 100ms

## Čípky

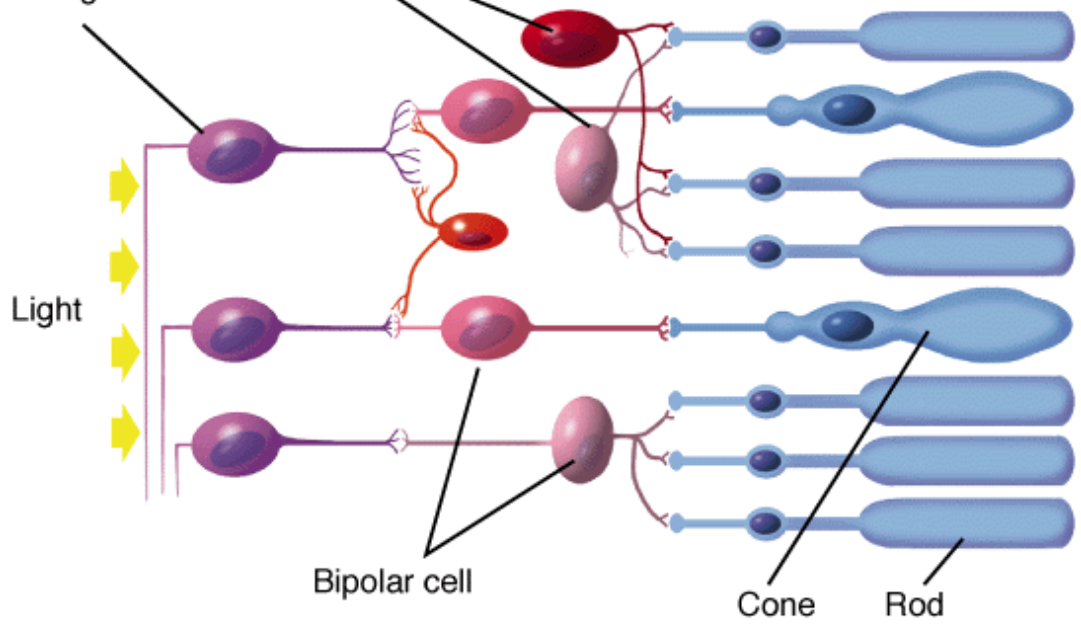
- 7 miliónů
- Umožňují rozlišení barev - 3 druhy fotopigmentu
- Funkční při vyšším osvětlení – méně citlivé
- Nejcitlivější na žluté skvrně
- Ve žluté skvrně výhradně čípky
- Plná adaptace na světlo nepřesáhne 1 minutu
- Změny až 55 Hz

→ čípky ostřejší, lepší rozlišení změn, barvy, jen v šeru špatné

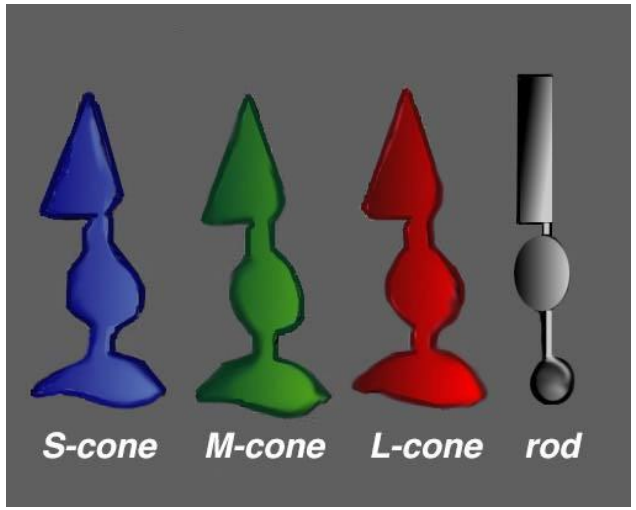
# Fotoreceptory

Akční potenciály

Změny membránového potenciálu

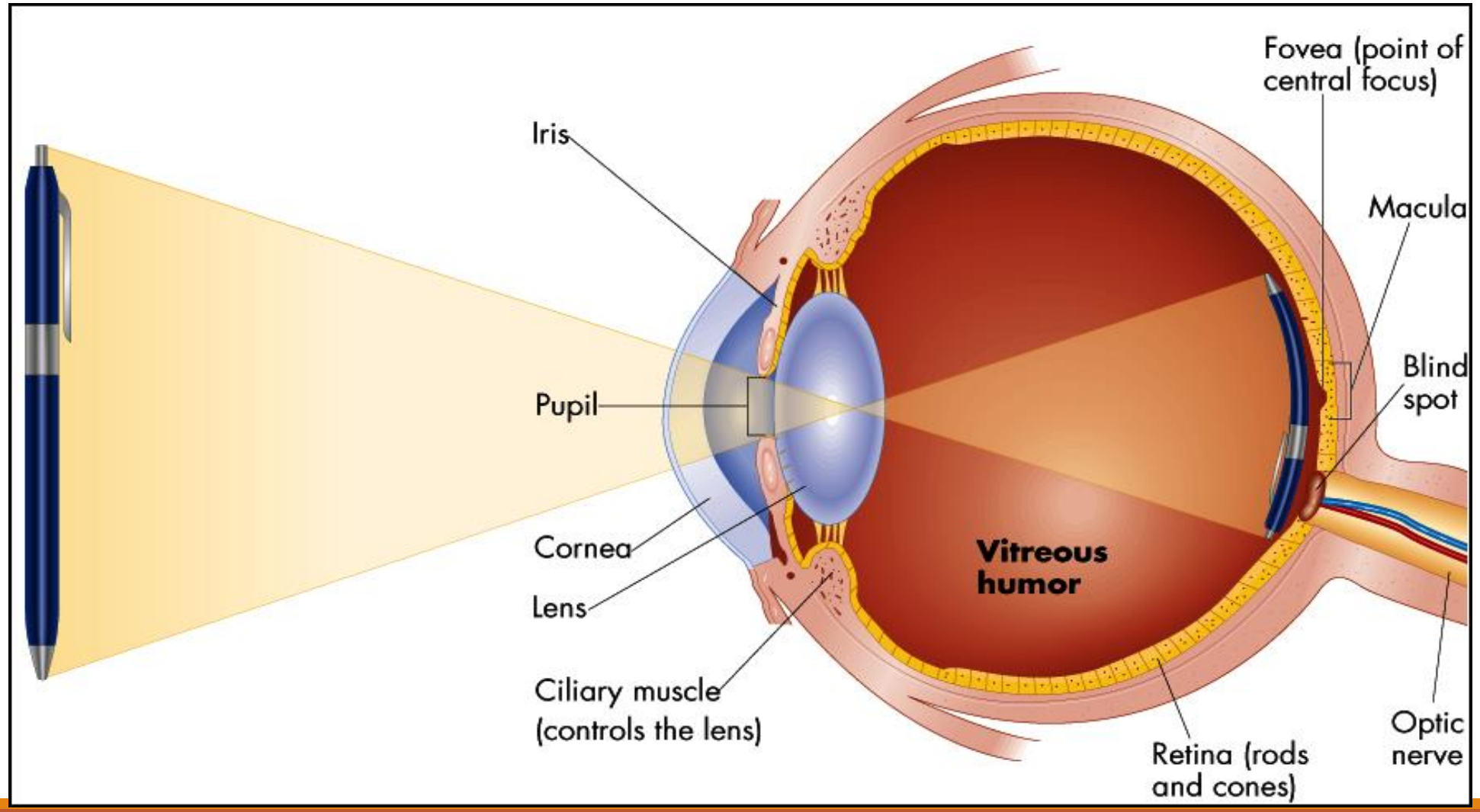



Vis.pigment rhodopsin

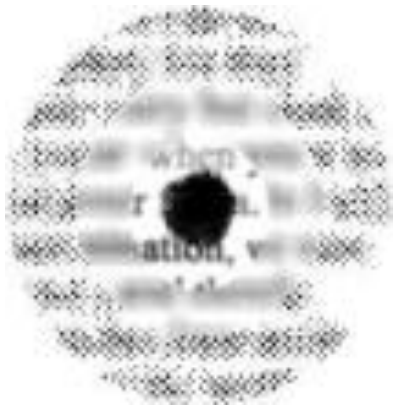



Proč má systém čípků lepší rozlišení než systém tyčinek, i když je poměr tyčinek:čípkům 20:1?

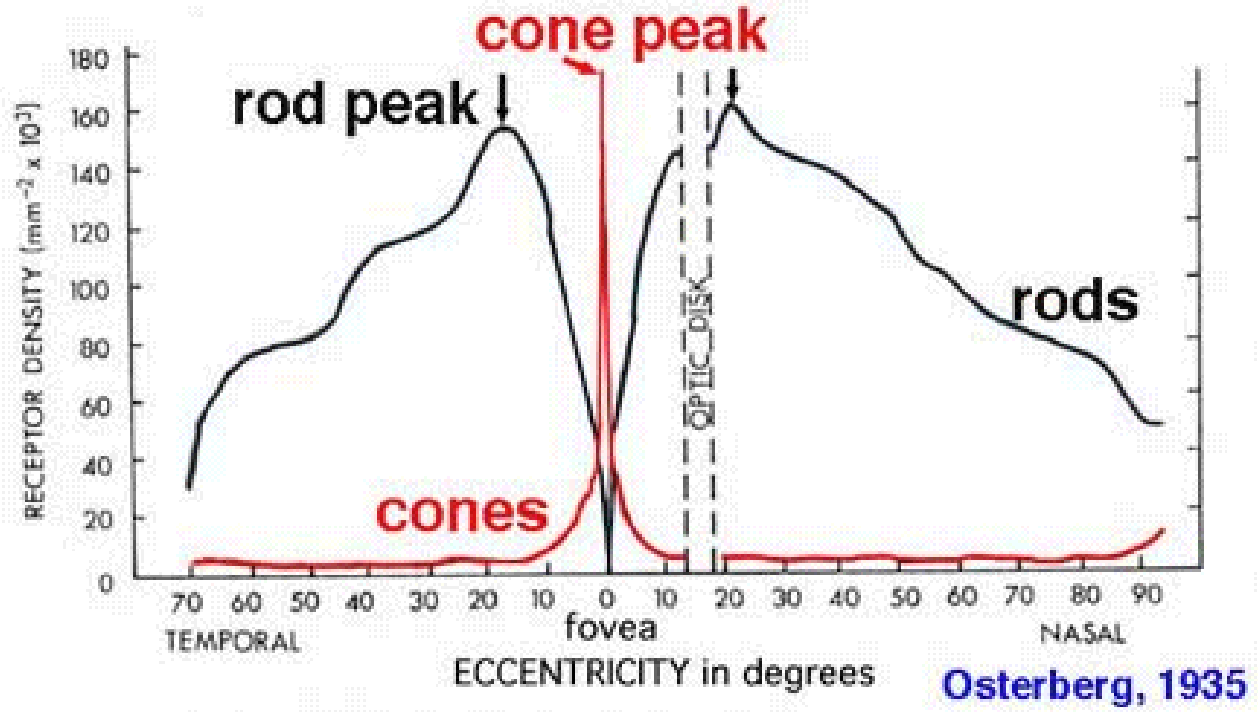
# Sítnicový obraz



# Sítnicový obraz



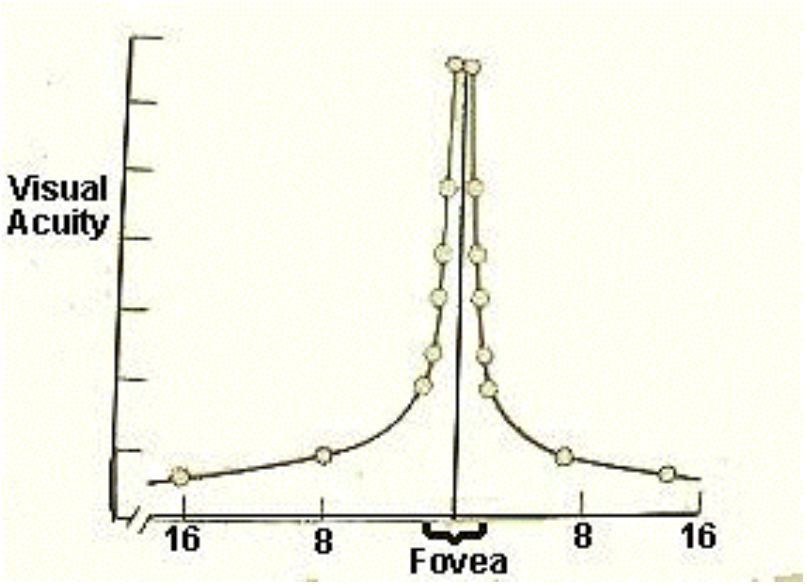
# Distribuce tyčinek a čípků

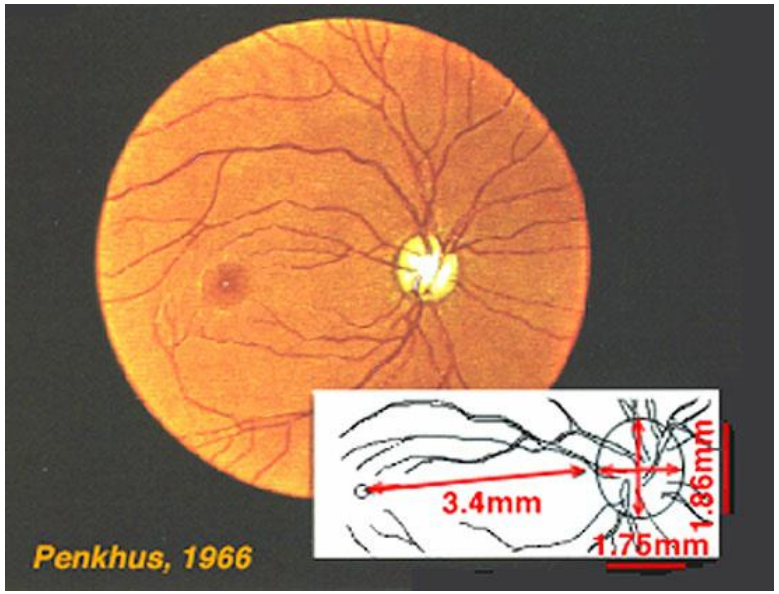


Klesající citlivost se vzdáleností podnětu od žluté skvrny

# Žlutá skvrna

- zde se promítá obraz sledovaného objektu
- místo nejostřejšího vidění
- výhradně čípky



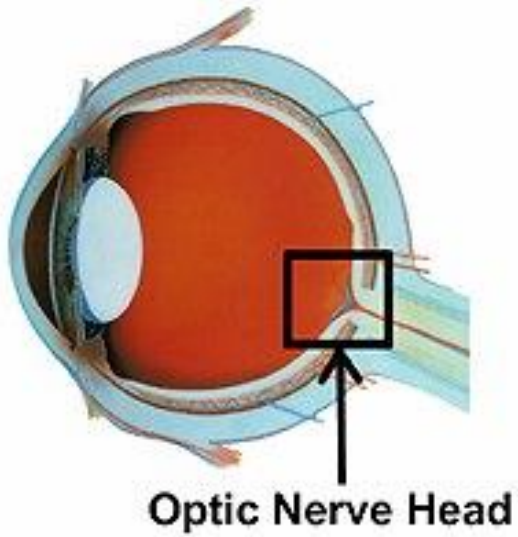



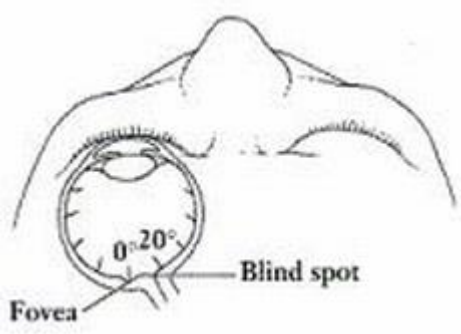
# Slepá skvrna

Ústí optického nervu

17° od žluté skvrny

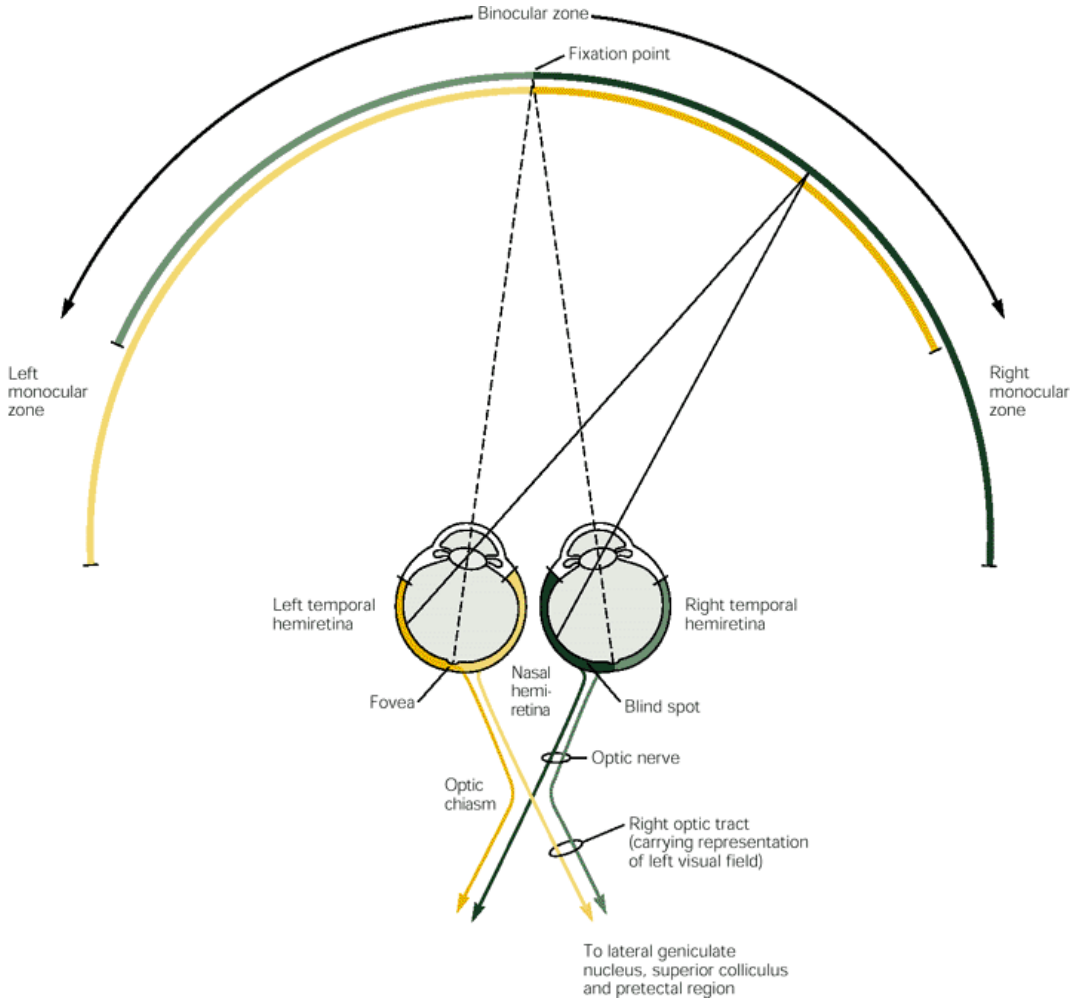
Proč naše vidění nelimituje?



# Zorné pole

# Zorné pole


Binocular zone
Fixation point
Left monocular zone
Right monocular zone
Left temporal hemiretina
Right temporal hemiretina
Fovea
Nasal hemi-retina
Blind spot
Optic nerve
Optic chiasm
Right optic tract (carrying representation of left visual field)
To lateral geniculate nucleus, superior colliculus and pretectal region

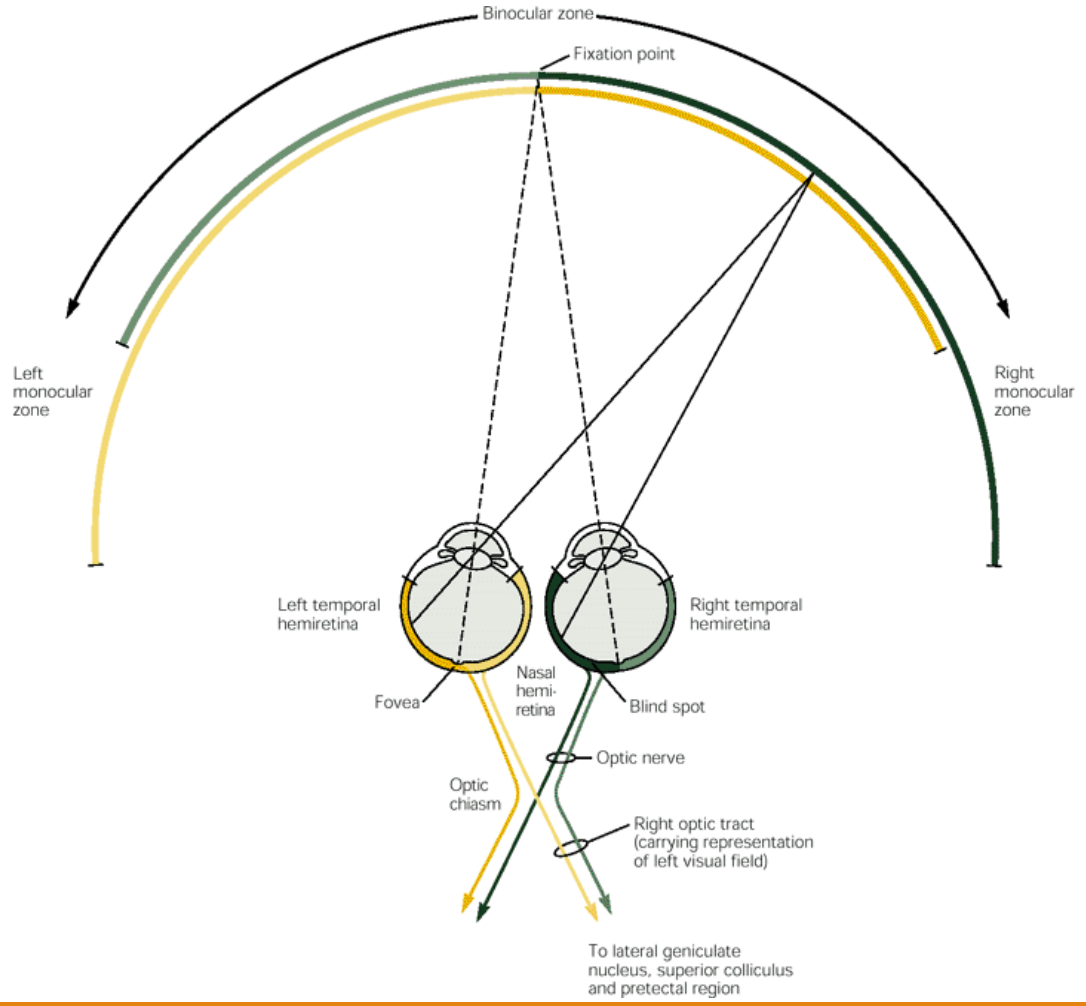
# Zorné pole

**Šířka zorného pole**

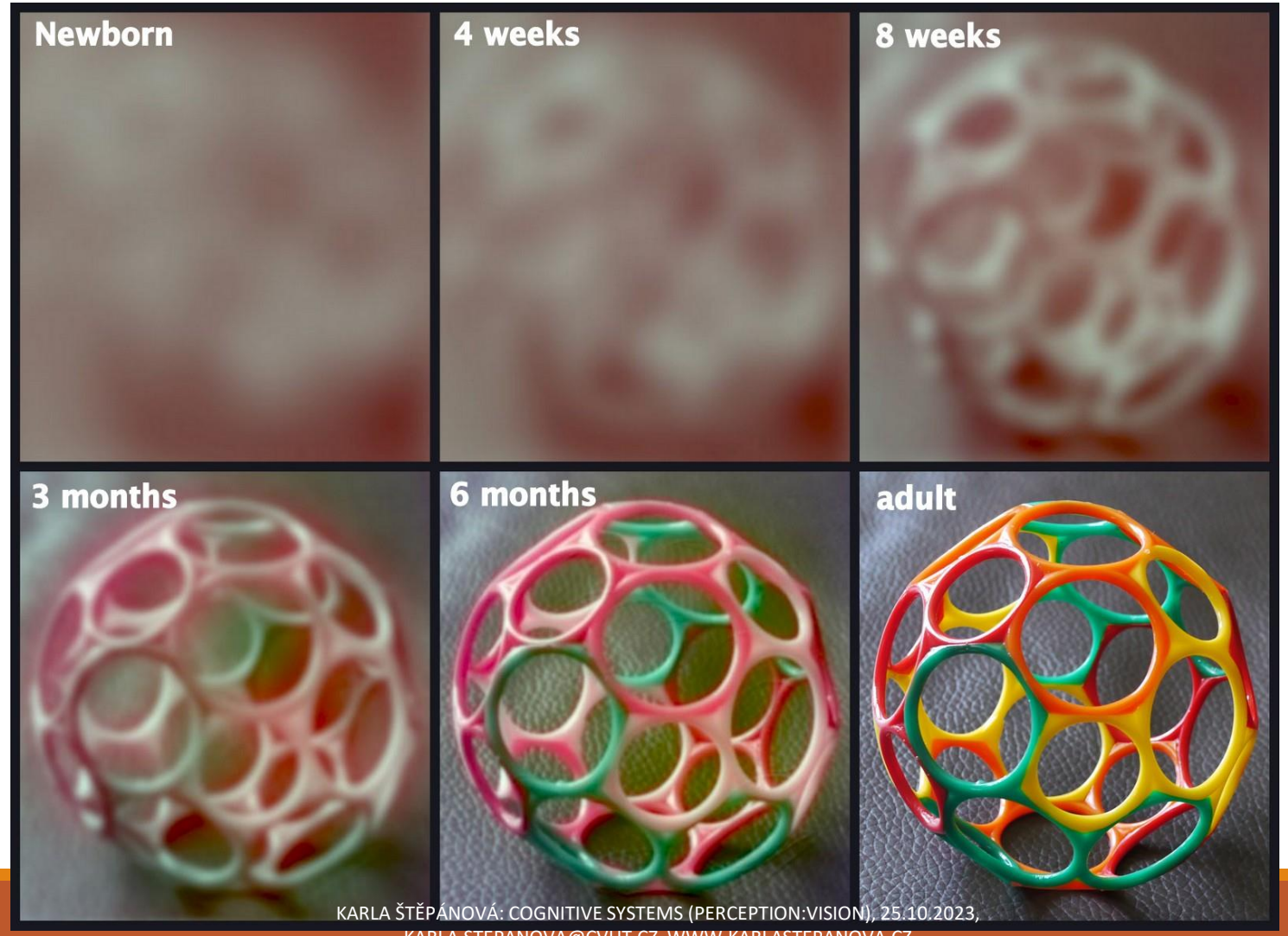
* Jedno oko:
160° (š) x 175° (v)

* Obě oči:
200° (š) x 135° (v)

* Oblast binokulárního překrytí:
120° (š) x 135° (v)



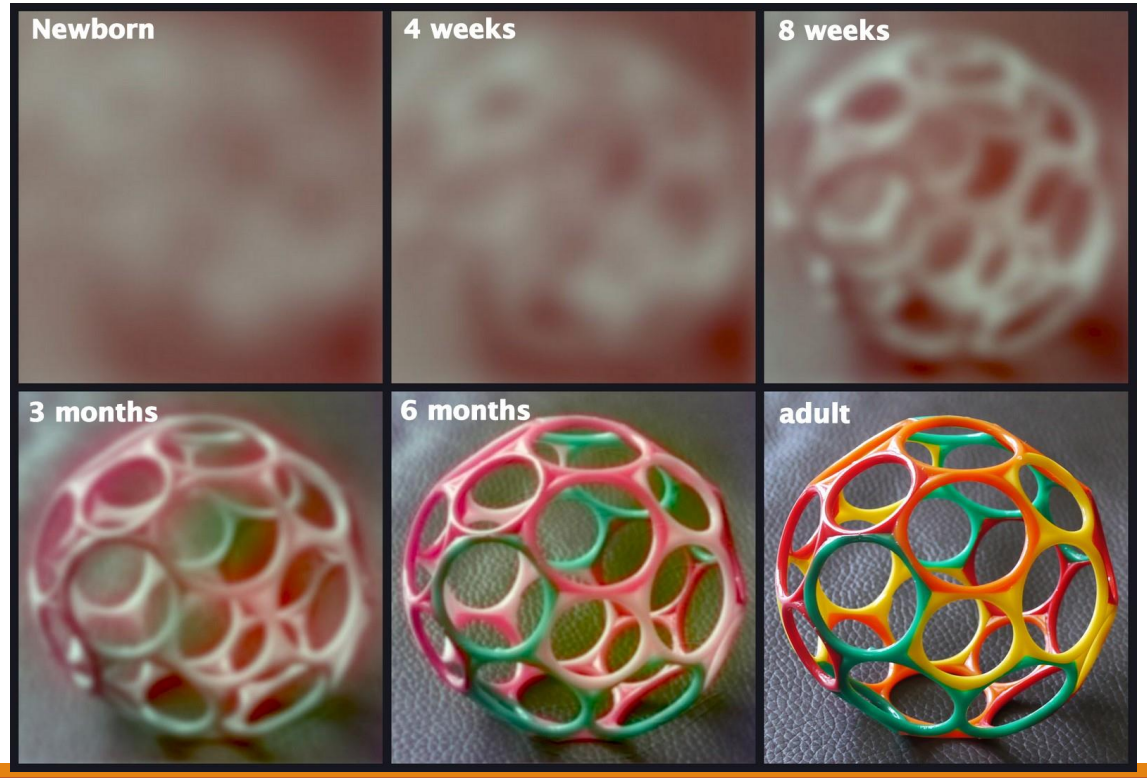
# Vývoj vidění



# Vývoj vidění

- **Ostrost vidění**
  - oční svaly (2 měsíce)
  - Vzdálenost rohovky a sítnice
  - Vyvinutí sítnice, mozkových drah (6m)
- **Tváře**
  - test matka x cizinec
  - Matka v šále (1 měsíc)
- **Barva**
  - Sílení čípků,raději delší vlnové délky (x dospělí)
  - Až po 3měsících
- **Citlivost na světlo**
  - 50x výše práh než dospělý
  - Prodlužování fotoreceptorů, vývoj sítnice
- **Vnímání hloubky**
  - Nápovědi – důležitost očních svalů
  - Vizuální hrana – Eleanor J. Gibson

https://www.youtube.com/watch?v=p6cqNhHrMJA



# Oční pohyby

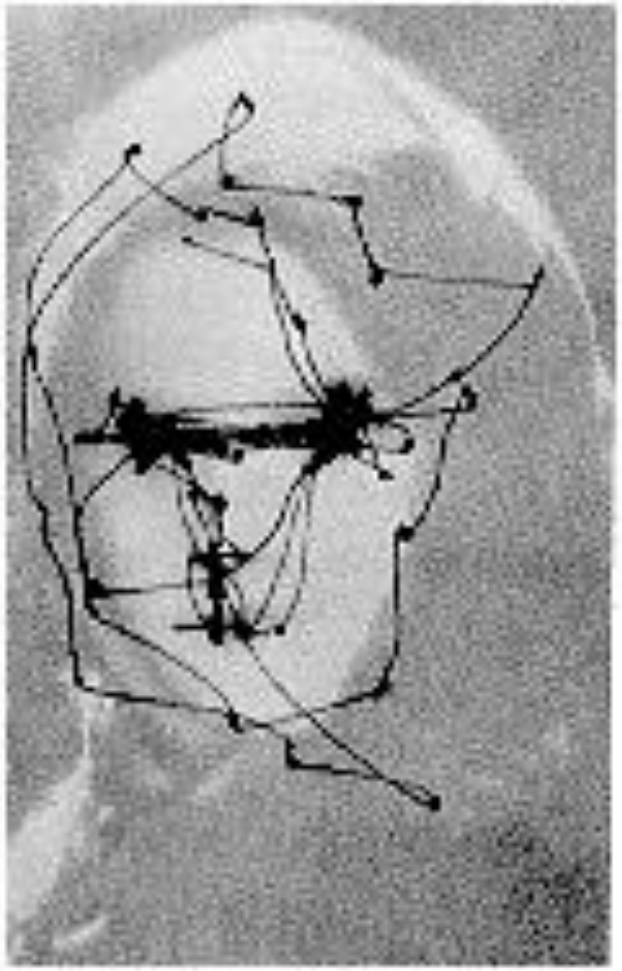
Jejich úkolem je udržet předmět našeho zájmu v oblasti žluté skvrny

Vyhasínání vjemu při stabilizaci promítaného obrazu objektu

## Druhy očních pohybů

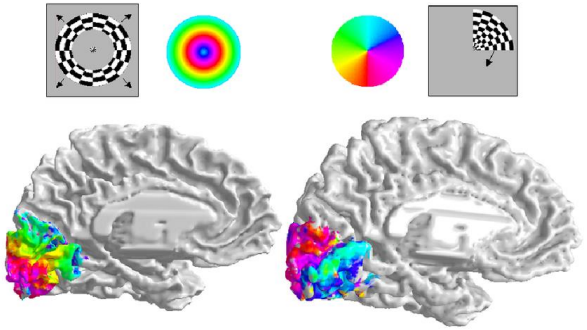
- **Volní** – parietální kůra -> frontální oční pole (premotorická část frontálního laloku) -> střední mozek -> most -> retikulární formace ->hlavové nervy -> okohybné svaly
- **Nevolní**
  - Sakadické oční pohyby – neplynulé, 3-4x/s
    - Hladké sledovací pohyby – fovea
    - Rychlé trhavé pohyby
  - Vestibulookulomotorický reflex – s natočením hlavy
- Provázející oční pohyby
- Kon/Divergence

# Sakadické oční pohyby



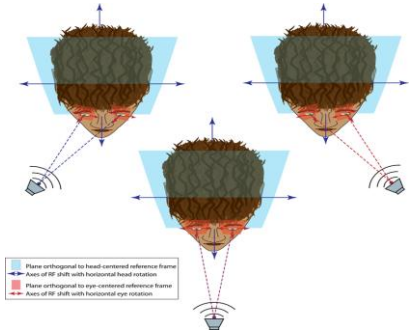
# Vizuální informace zpracována v různých oblastech mozku

- Hierarchický systém X oddělená centra
- fotoreceptory → bipolární b. → gangliové b. → optický nerv → thalamus → primární viz.kůra → extrastriate oblasti
- Retinotopic map, retinotopy (zač.20 st. na lidech s lézemi, 1941 experimentálně)

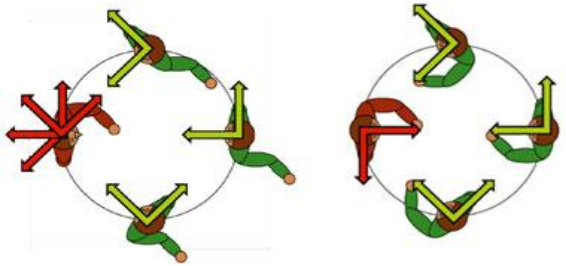


**Head – centered ref. rámec** **X** **body – centered ref. rámec**

Retinotopic map + Pozice očí



Retinotopic map + oči + hlava



- **Makak**
  - přes 50% neokortexu vizuální informaci, 11% somatosenzorickou, 3% sluch
  - centra pro viz.info se liší velikostí
  - aktivitu buněk jednotlivých oblastí, odpovědi neuronů na podnět různé orientace http://youtu.be/y_l4kQ5wjiw (1:30)

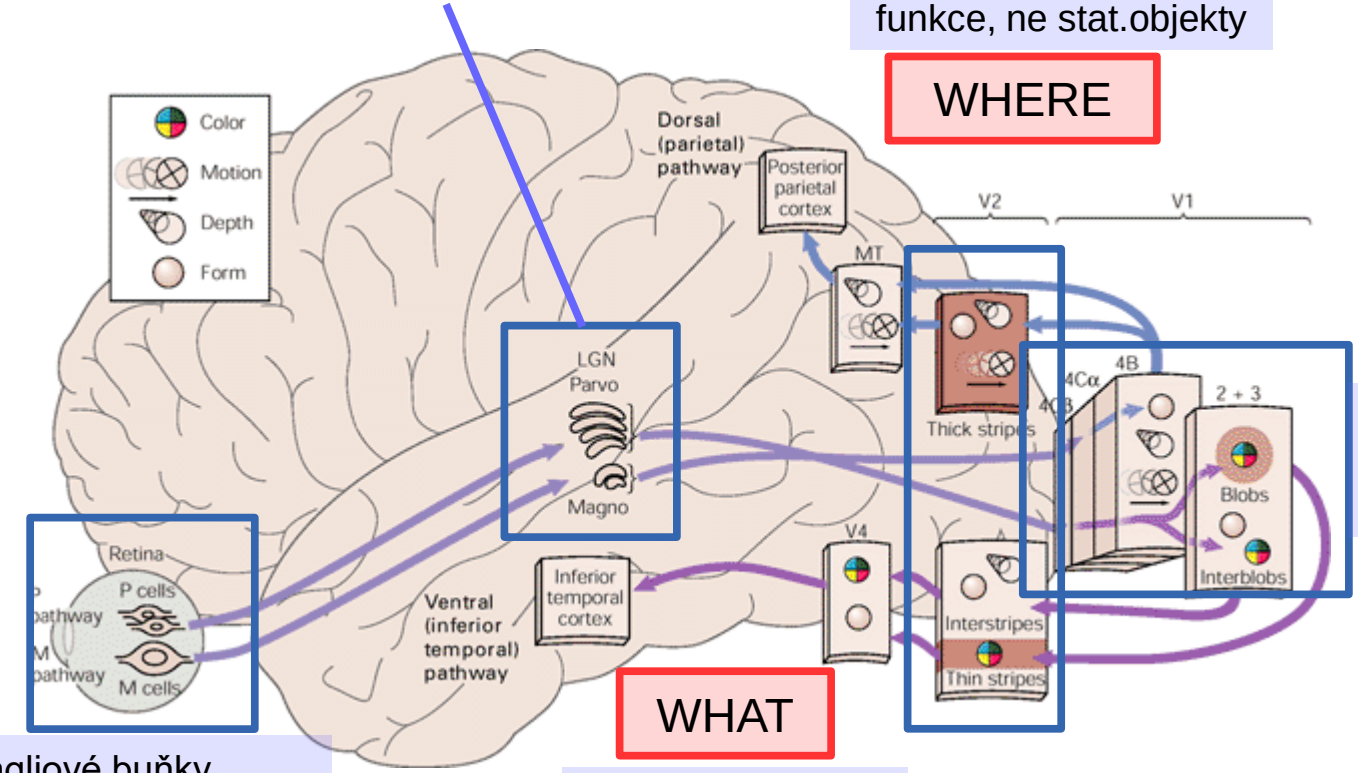
# Vizuální informace zpracována v různých oblastech mozku

| | Parietální kůra | Temporální kůra |
|---|---|---|
| **Léze u lidí** | Visual neglect – ne slepou skvrnu (léze V1), ale kontralaterální neglect<br>→ **prostorová reprezentace** | Problém rozlišení objektů, vizuální paměť, rozpoznání tváří<br>→ **object recognition** |
| **Léze u opic** | Problém lokace objektů (dát ruce)<br>**Neovlivní identifikaci** | Problém identifikace objektů – barva, orientace, vzor, tvar<br>**Neovlivní lokaci** |
| **BOLD - PET** | Lateral occipital extrastriate region<br>Object location, při pozornosti na **rychlost** objektu | Rozpoznání objektů, pozornost na **barvu, tvar** |

Thalamus, Corpus geniculatum laterale
První reorganizace info z nerv.vzruchu

**Dorsální dráha**
Visuoprostorové funkce, ne stat.objekty

WHERE




Primární viz.kůra
**V1**

Gangliové buňky
M buňky – 8% → M dráhy ,
P buňky 80% → P dráhy

WHAT

**Ventrální dráha**
Citlivé na obrysy,
Velké rozlišení

# Korespondence postupných úrovní

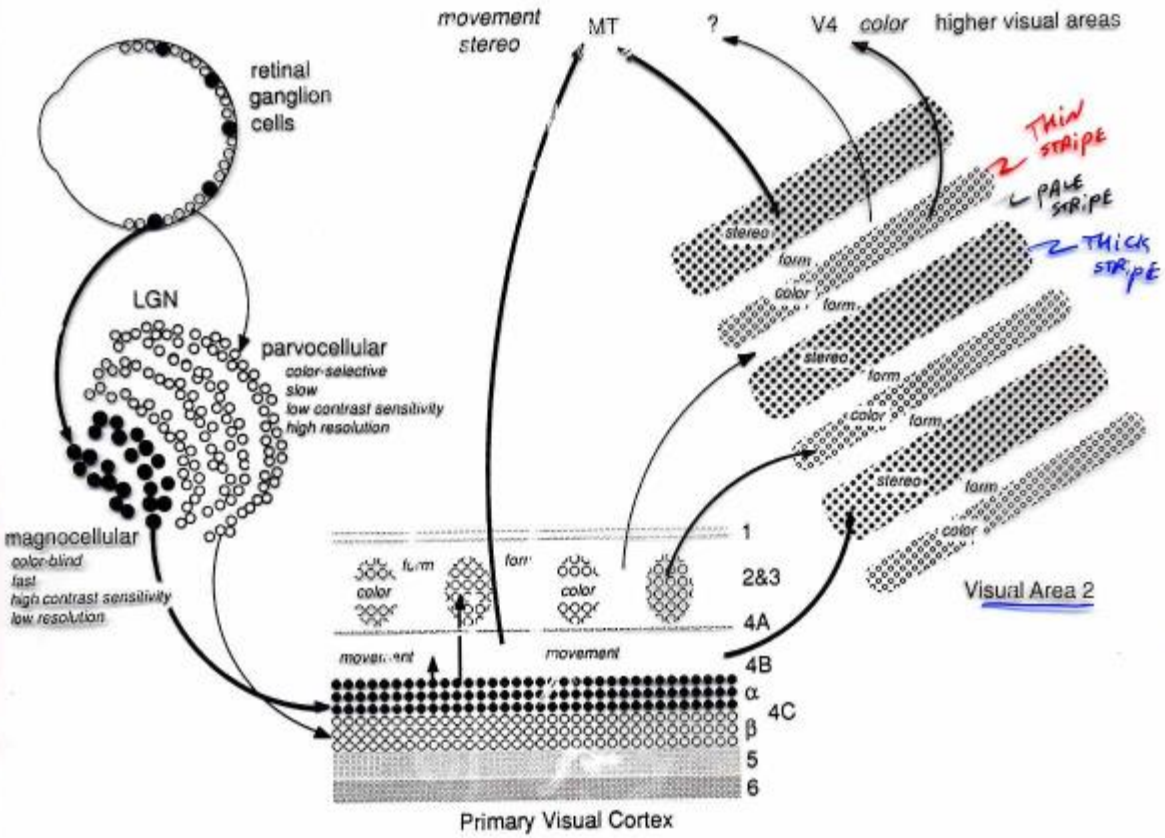
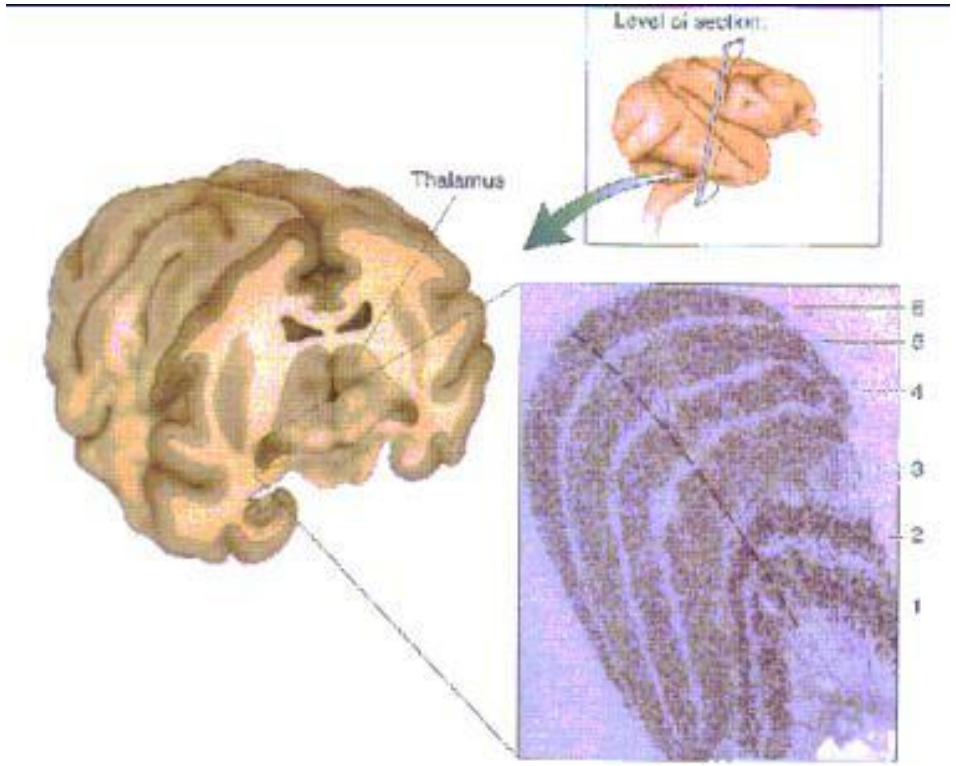



**Fig. 2.10:** Anatomical interconnection scheme of the neurons of the human color-coding system. The retinal ganglion cells project into the parvocellular lateral geniculate nucleus (LGN) and via the β-band of the primary visual cortex and visual area 2 (V2) to visual area V4. (After Livingstone and Hubel, 1988.)

# Corpus geniculatum laterale

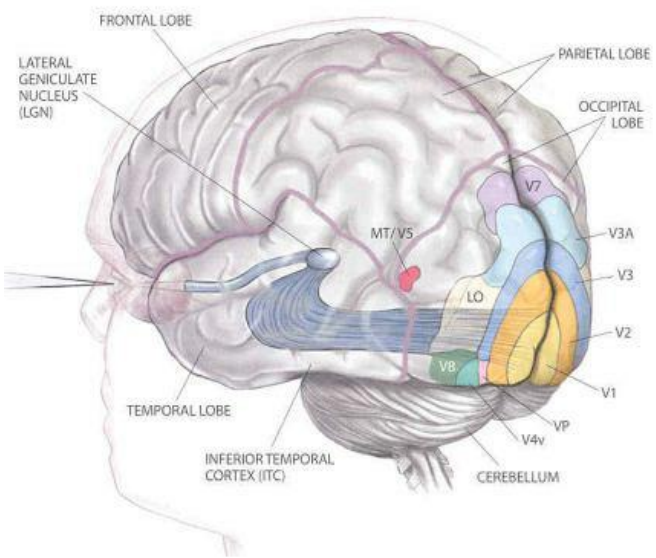


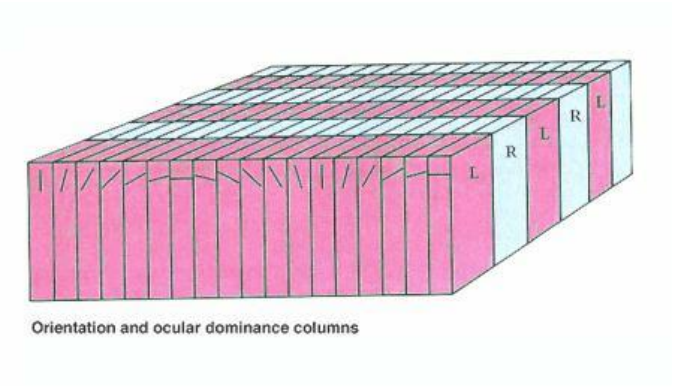

thalamus

první reorganizace informací obsažených v nervovém vzruchu

magno- a parvocelulární systém

# Primární zrakový kortex (V1)



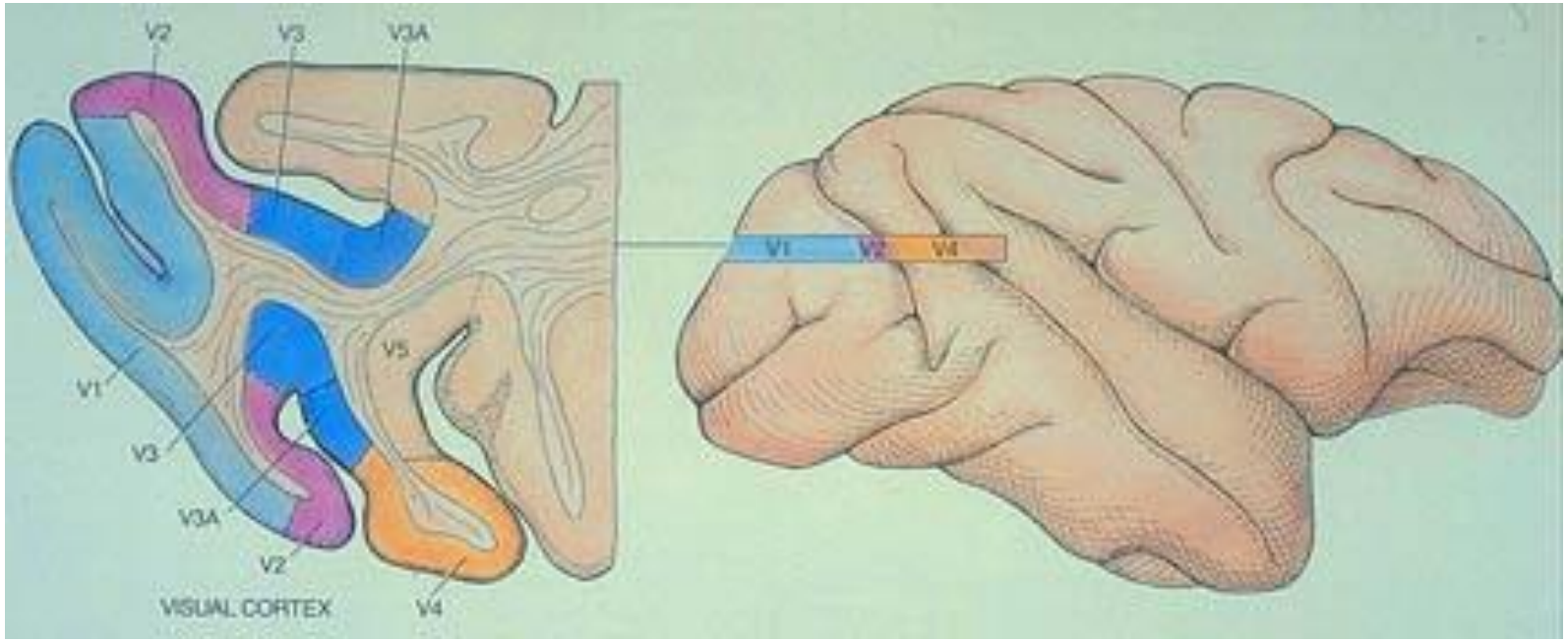

Primární analýza a detekce jednotlivých elementů

separátní zpracování obou stran zorného pole

magnifikace obrazu na žluté skvrně

Hypersloupec – ze sloupců pro všechny orientace z jednoho regionu prostoru

# Další zpracování




zpracování různých modulů podnětu v různých úrovních

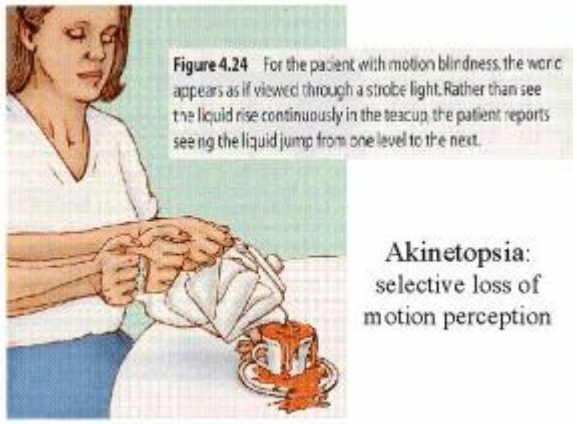
s postupující úrovní rostoucí specialisace nervových buněk

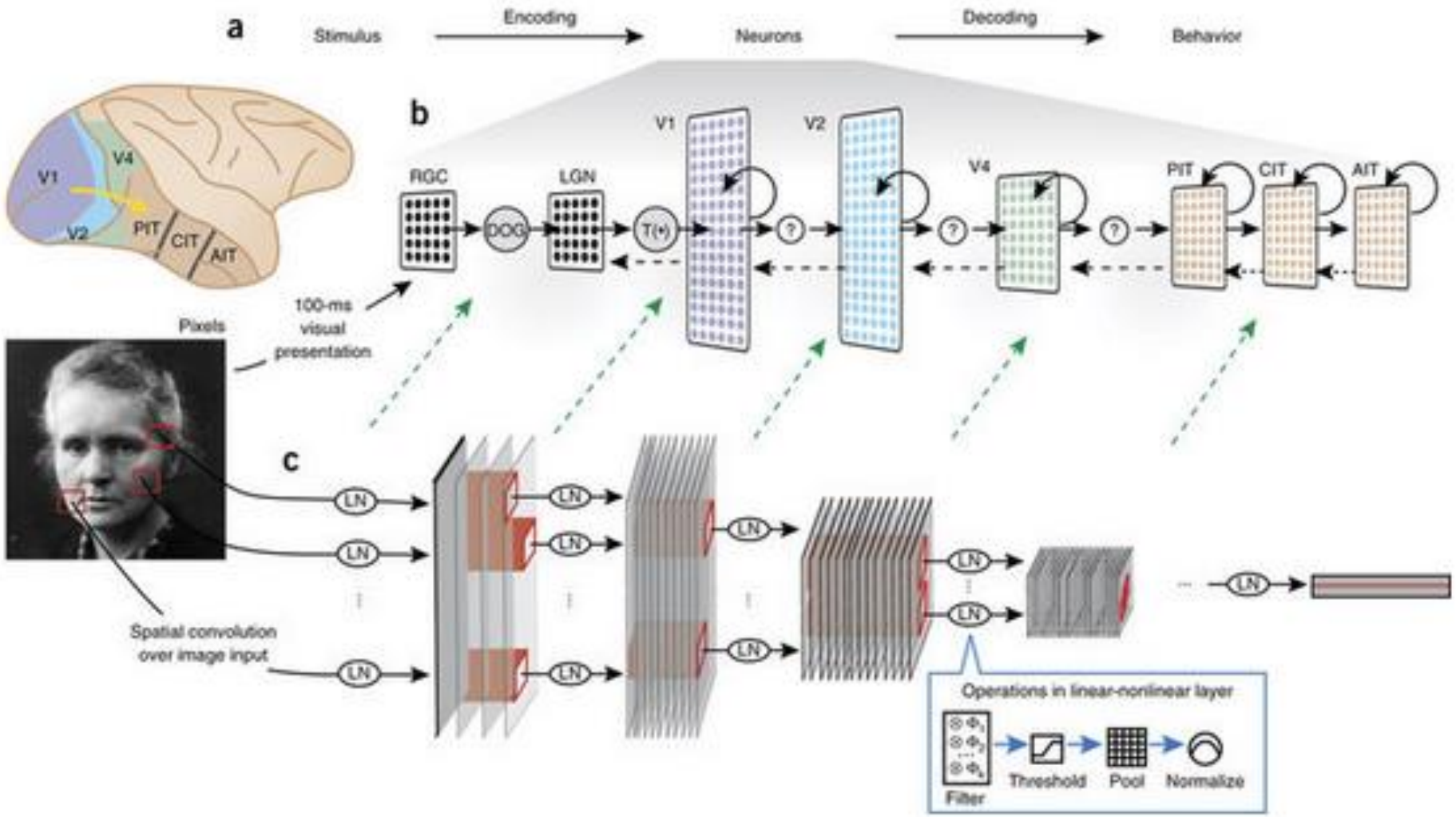
# Další zpracování (V3, V4, V5, IT)

**Barva** : V4, achromatopsie

**Tvar** : více sekvencí (V3, IT, V4); speciální zpracování tváří;agnosie, prosopagnosie

**Pohyb** : V5, akinetopsie



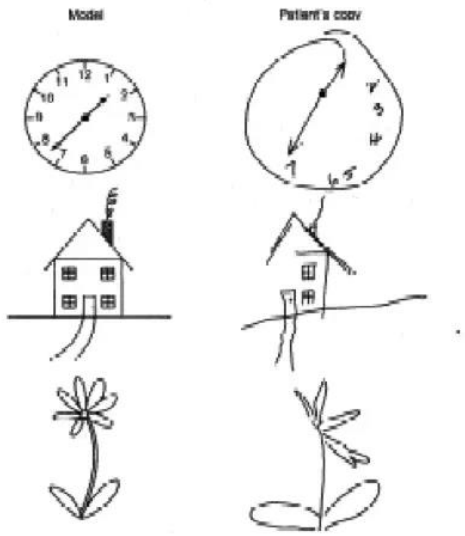



https://becominghuman.ai/from-human-vision-to-computer-vision-convolutional-neural-network-part3-4-24b55ffa7045

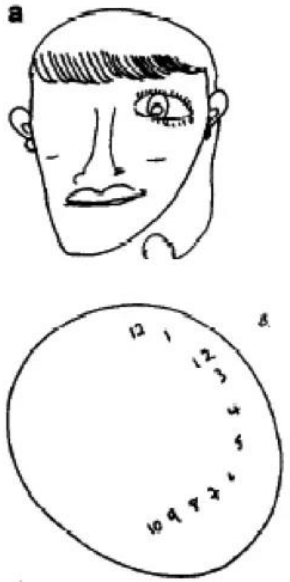
# Poruchy vnímání (hemispatial neglect)

- Left visual neglect after stroke

Copying:




Spontaneous drawing:



# Poruchy vnímání

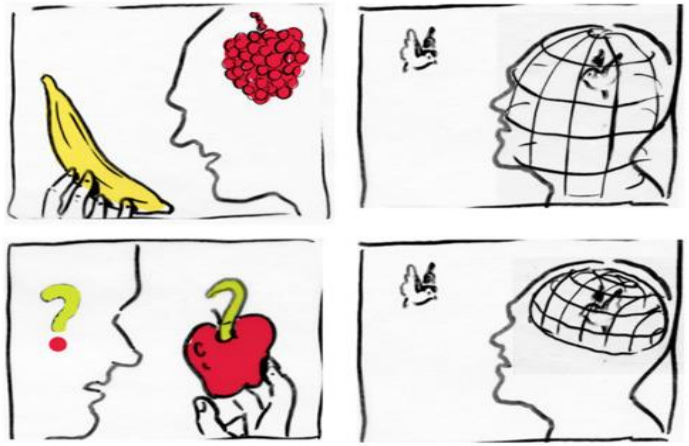
- Agnosie

  = ztráta schopnosti, znalosti

  – S.Freud – ne problém senzorů

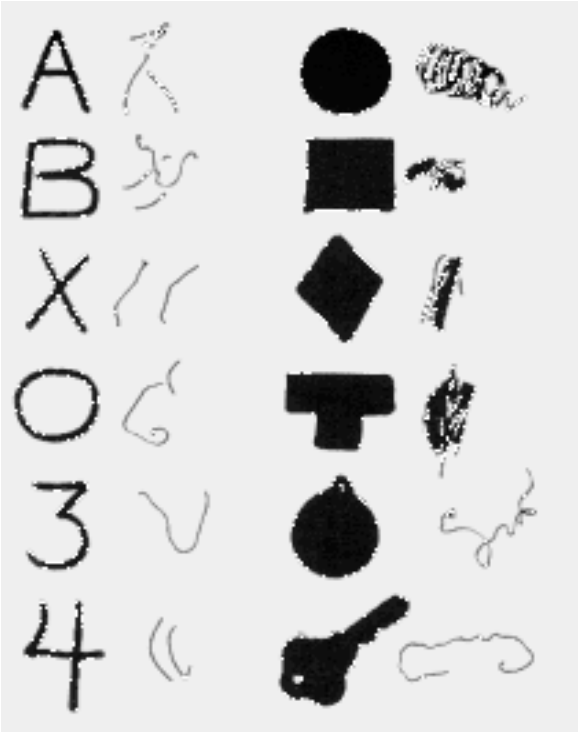
  – Neschopnost rozpoznat a identifikovat objekty a osoby, přestože o nich má subjekt předchozí znalost

  – Poukazuje na specifická centra perceptuálního systému



# Vizuální agnosie

- **Aperceptivní agnosie**
  - Neschopnost pojmenovat, napodobit nebo rozpoznat objekty prezentované vizuálně.
  - Je zachována schopnost vnímání barev, identifikace objektu a nevizuálních nápovědí.
- **Asociační agnosie**
  - Porušená identifikace objektů
  - Rozpoznají objekt, ale ne mu dát význam



# Prosopagnosie

- deficit v oblasti vnímání tváří
- Mrtvice, degenerativní onemocnění
- Typy:
  - aperceptivní typ ~ vnímání
  - amnestický typ ~ poruchy paměti na tváře
- Problém při sledování filmu – neudrží děj

Konkrétní tvář ← X → Identifikace

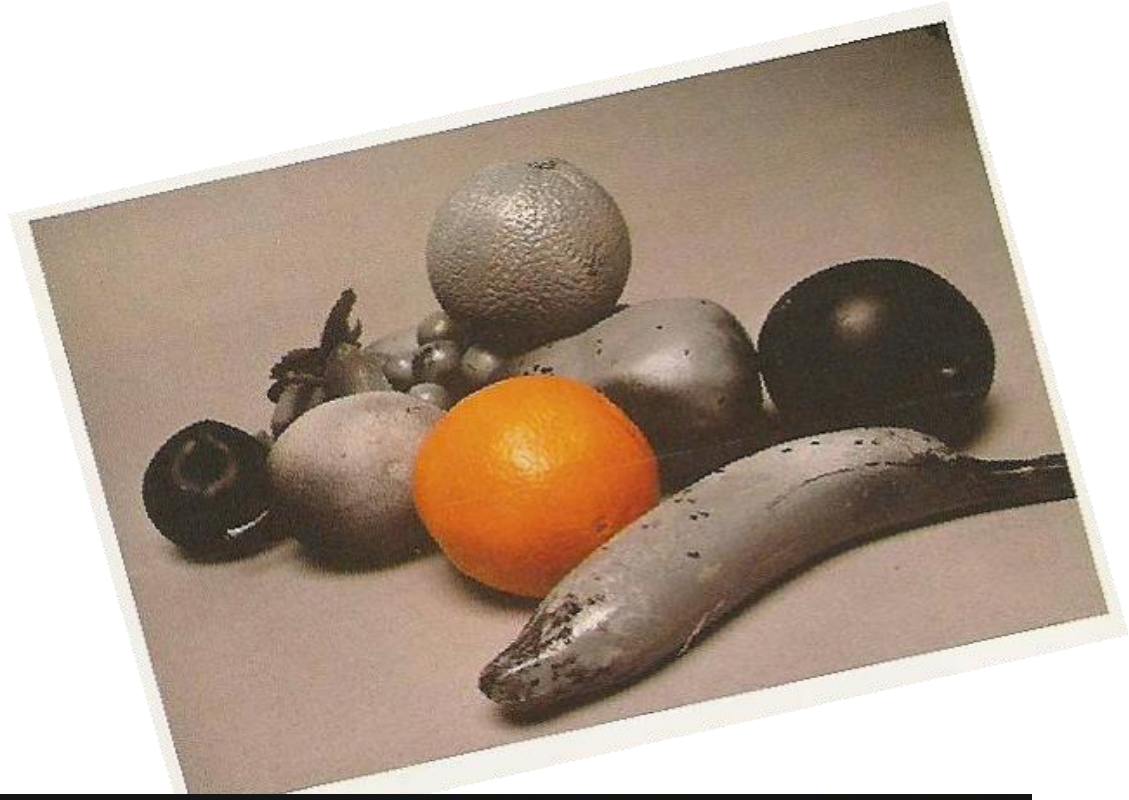
**Bilaterální** inferior temporal cortex

# Achromatopsia

- Poškození vnímání barev v celém spektru

+ fotofobie, malá ostrost

Bilateral temporal cortex



**Příčiny**:
- nefunkčnost čípků
- dědičná čípková slepota
- některá onemocnění sítnice

# Akinetopsie

- **Neschopnost vnímat pohyb** díky poškození dorzální vizuální dráhy (V5/MT).

- 2 typy
  - Pohyb~ série fotografií
  - Nevidí pohyb – jako by ztuhl

http://youtu.be/B47Js1MtT4w

# Simultagnosie

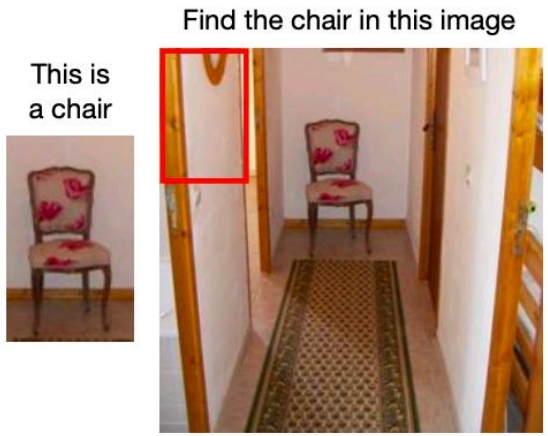
Schopnost vnímat pouze
jedno slovo nebo objekt
v jeden okamžik

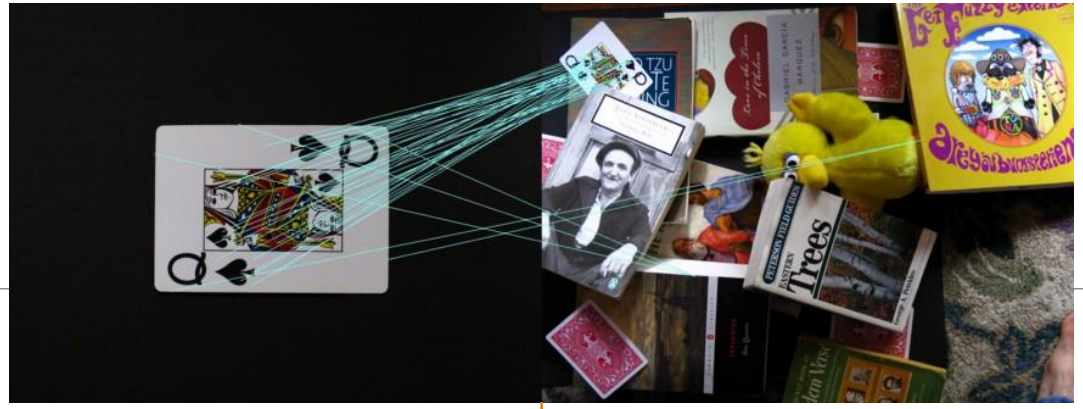
# Poruchy vnímání

- slepota ke slovům
- Agnosie k orientaci objektů
- Slepota k hloubce
- Slepota ke gestům
- Slepota k prostředí (rozpoznat v jakém se nachází prostředí)

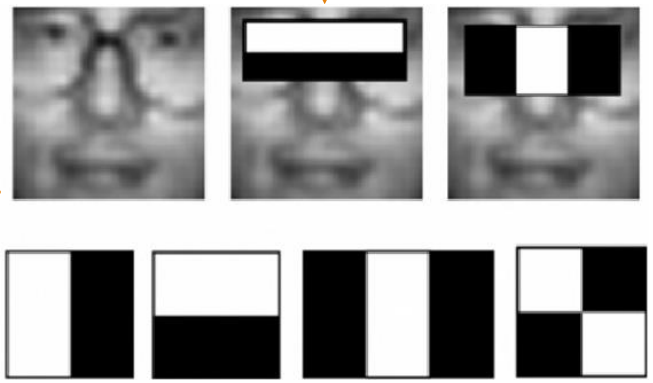
...

- sběratel známek nepozná známky
- pozorovatel ptáků nerozpozná ptáky
- může být selektivní pro živé/neživé/pro slova/pohyb...



1999, David Lowe

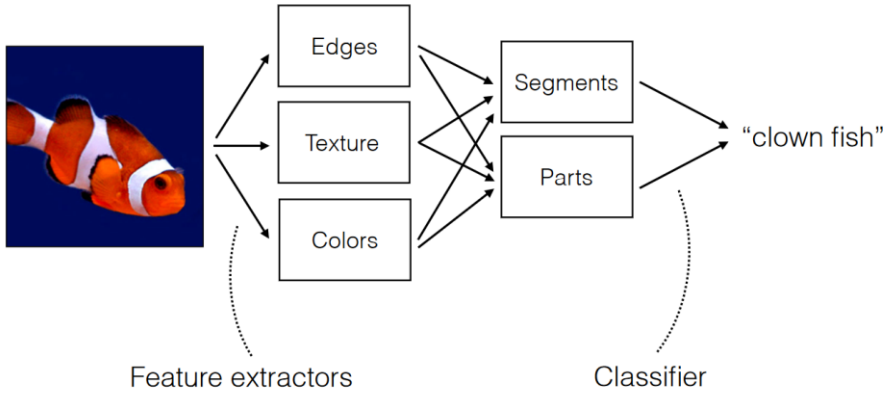


2000, Viola-Jones face detection

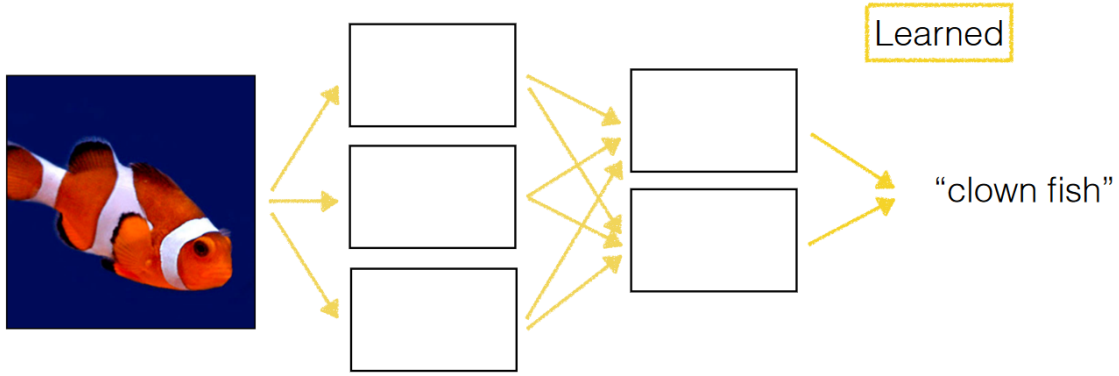
AlexNet, 2012

# Recent advances in computer vision

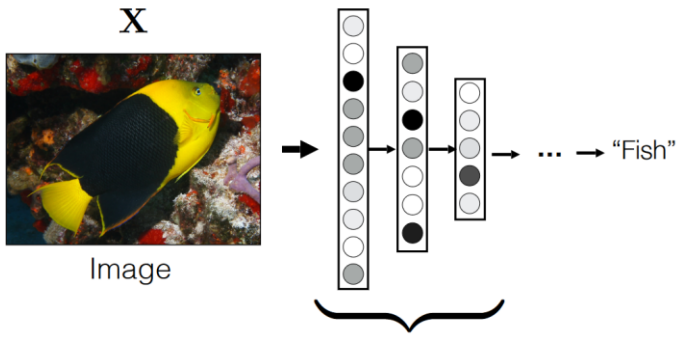
## Classical object recognition




## Deep learning



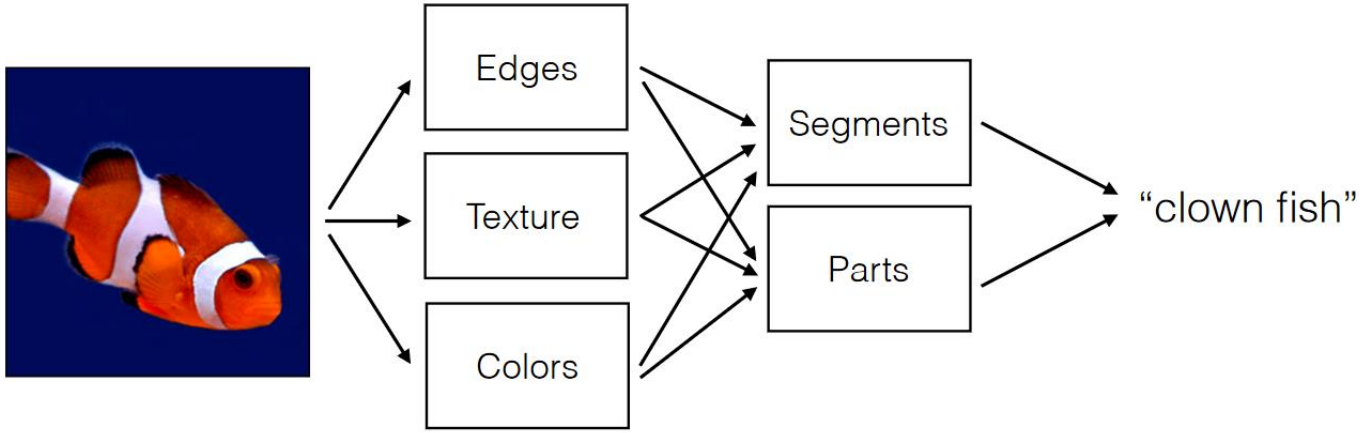

## What do deep nets internally learn?

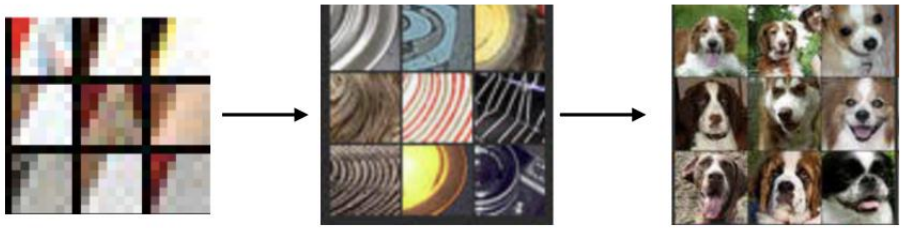



Representations!

A CNN is a multiscale, hierarchical representation of data

# CNNs *learned* the classical visual recognition pipeline!

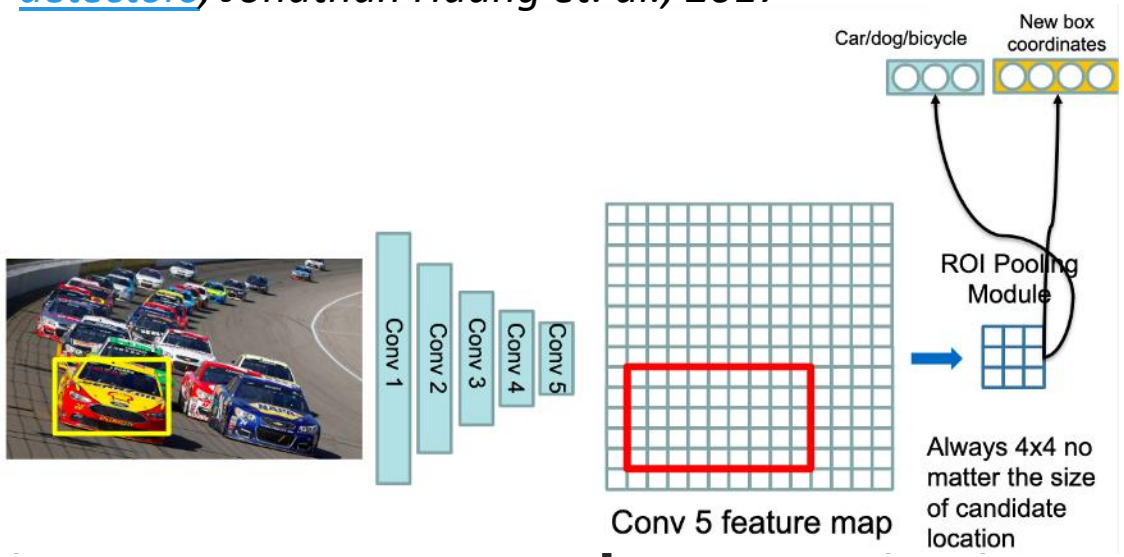


AlexNet, 2012 https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

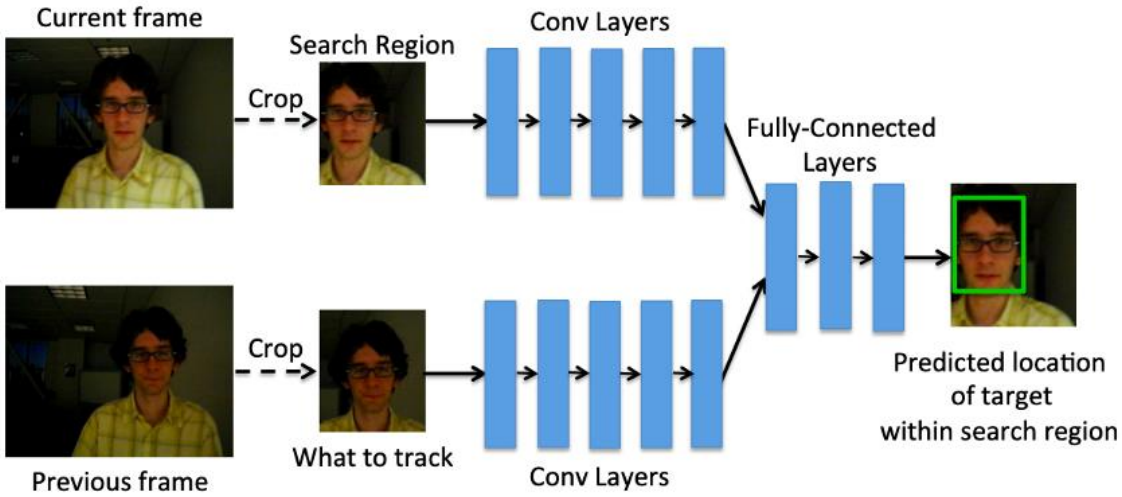
| Detection Frameworks | Train | mAP | FPS |
|---|---|---|---|
| Fast R-CNN [5] | 2007+2012 | 70.0 | 0.5 |
| Faster R-CNN VGG-16[15] | 2007+2012 | 73.2 | 7 |
| Faster R-CNN ResNet[6] | 2007+2012 | 76.4 | 5 |
| YOLO [14] | 2007+2012 | 63.4 | 45 |
| SSD300 [11] | 2007+2012 | 74.3 | 46 |
| SSD500 [11] | 2007+2012 | 76.8 | 19 |

*Speed/accuracy trade-offs for modern convolutional object detectors*, *Jonathan Huang et. al., 2017*




# Recent advances in computer vision

https://towardsdatascience.com/recent-advances-in-modern-computer-vision-56801edab980




D. Held, S. Thrun, and S. Savarese “Learning to Track at 100 FPS with Deep Regression Networks”, ECCV 2016.

MOTS: Multi-Object Tracking and Segmentation — Paul Voigtlaender et. al., CVPR 2019

# Recent advances in computer vision

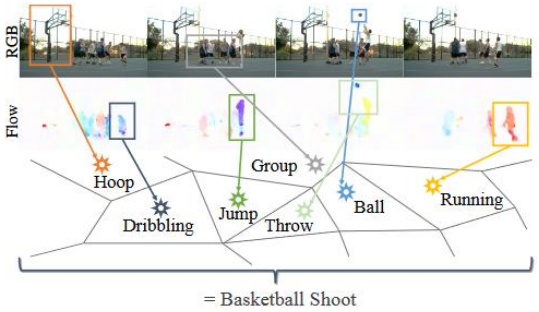
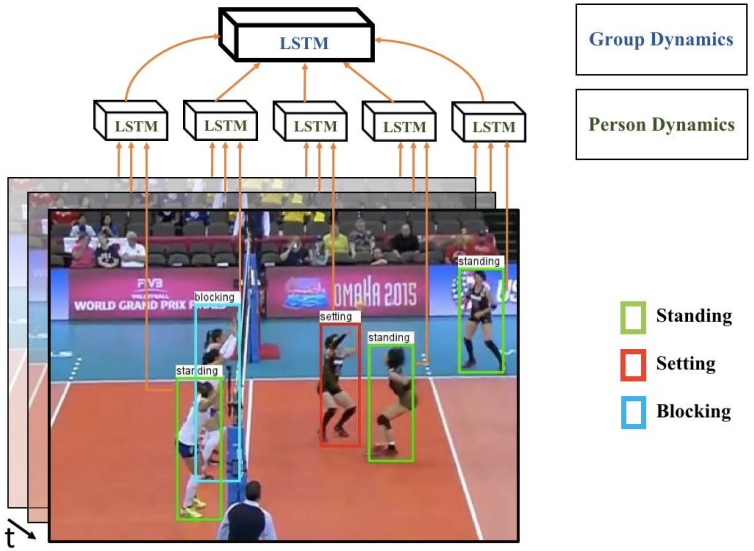


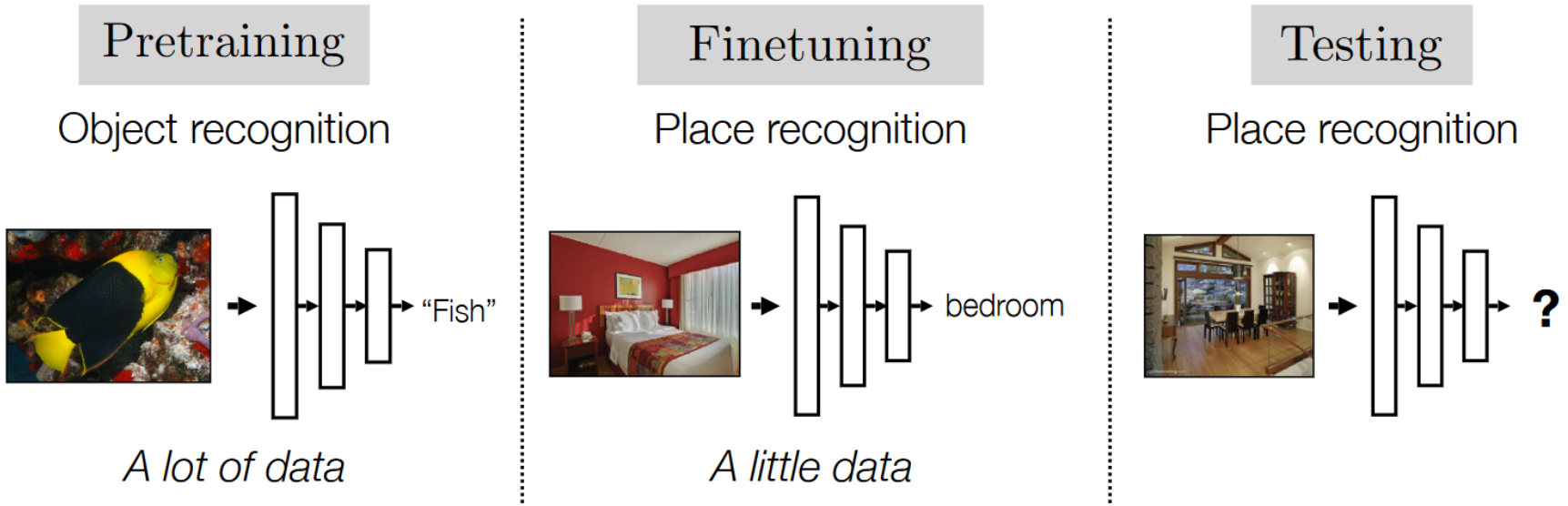

Figure 1: How do we represent actions in a video? We propose ActionVLAD, a spatio-temporal aggregation of a set of action primitives over the appearance and motion streams of a video. For example, a basketball shoot may be represented as an aggregation of appearance features corresponding to 'group of players', 'ball' and 'basketball hoop'; and motion features corresponding to 'run', 'jump', and 'shoot'. We show examples of primitives our model learns to represent videos in Fig. 6.




https://blog.qure.ai/notes/deep-learning-for-videos-action-recognition-review
https://towardsdatascience.com/literature-survey-human-action-recognition-cc7c3818a99a
https://towardsdatascience.com/deep-learning-architectures-for-action-recognition-83e5061ddf90

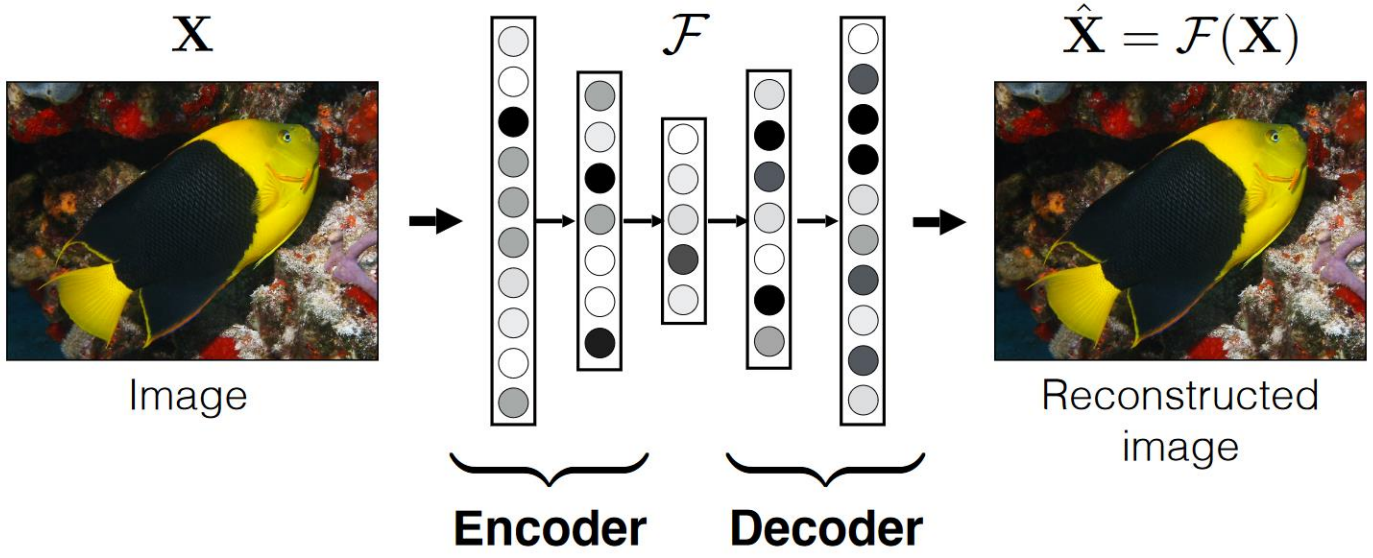
# Recent advances in computer vision

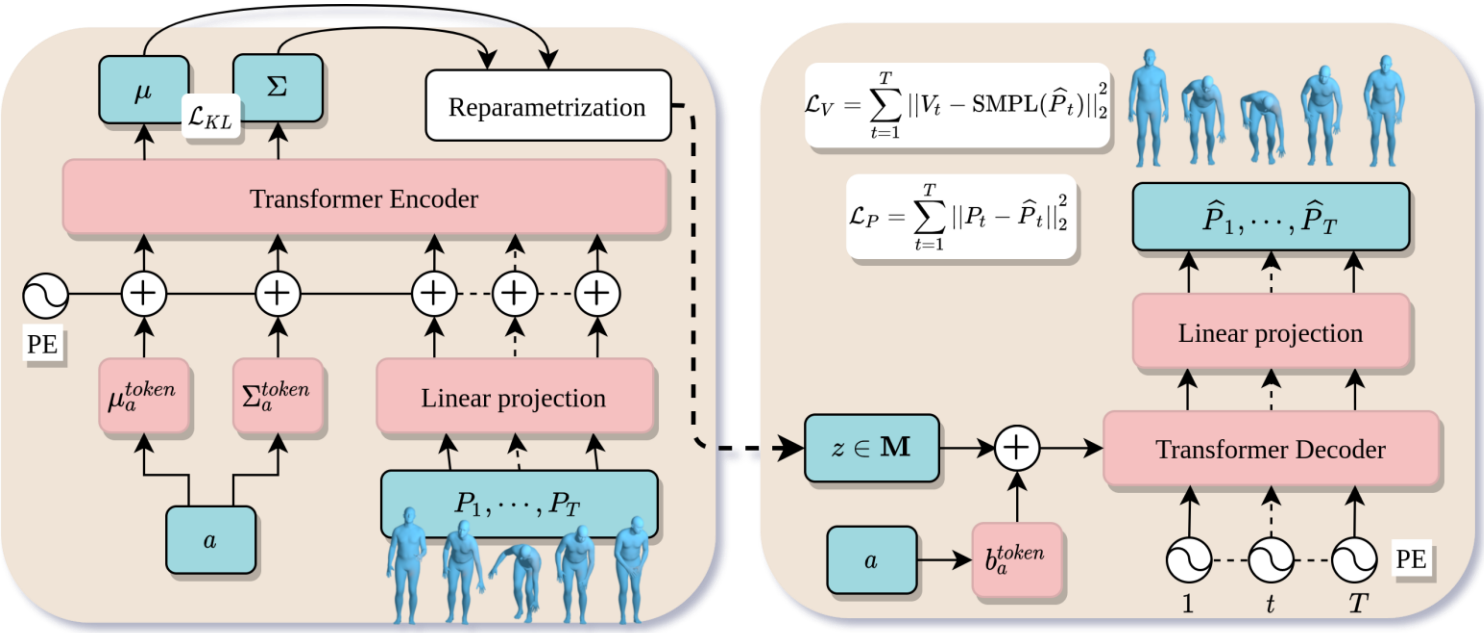
# Recent advances in computer vision – transfer learning



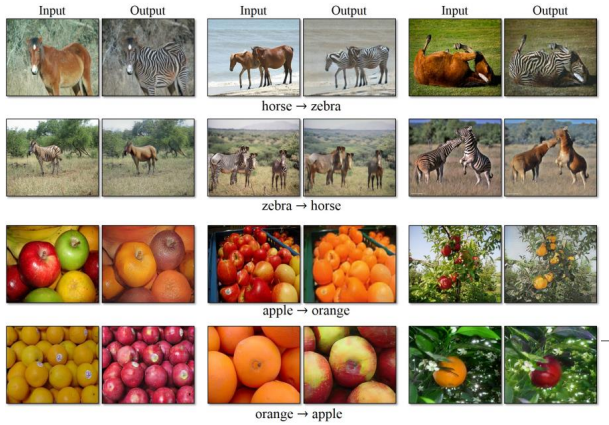

**Finetuning** starts with the representation learned on a previous task, and adapts it to perform well on a new task.

# Recent advances in computer vision – autoencoders, GANs, VAEs

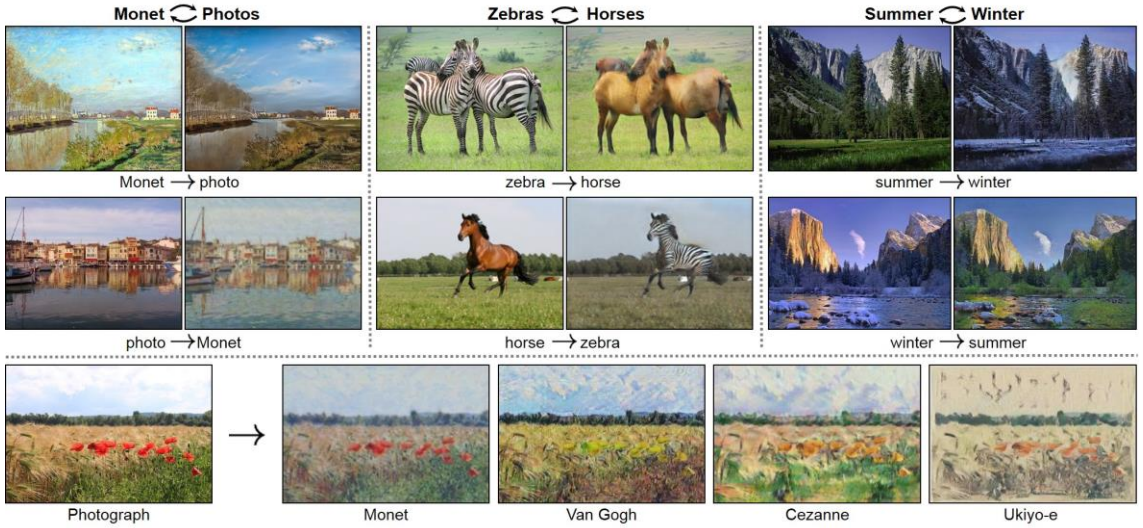





Petrovich, Mathis, Michael J. Black, and Gül Varol. "Action-conditioned 3d human motion synthesis with transformer vae." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021.

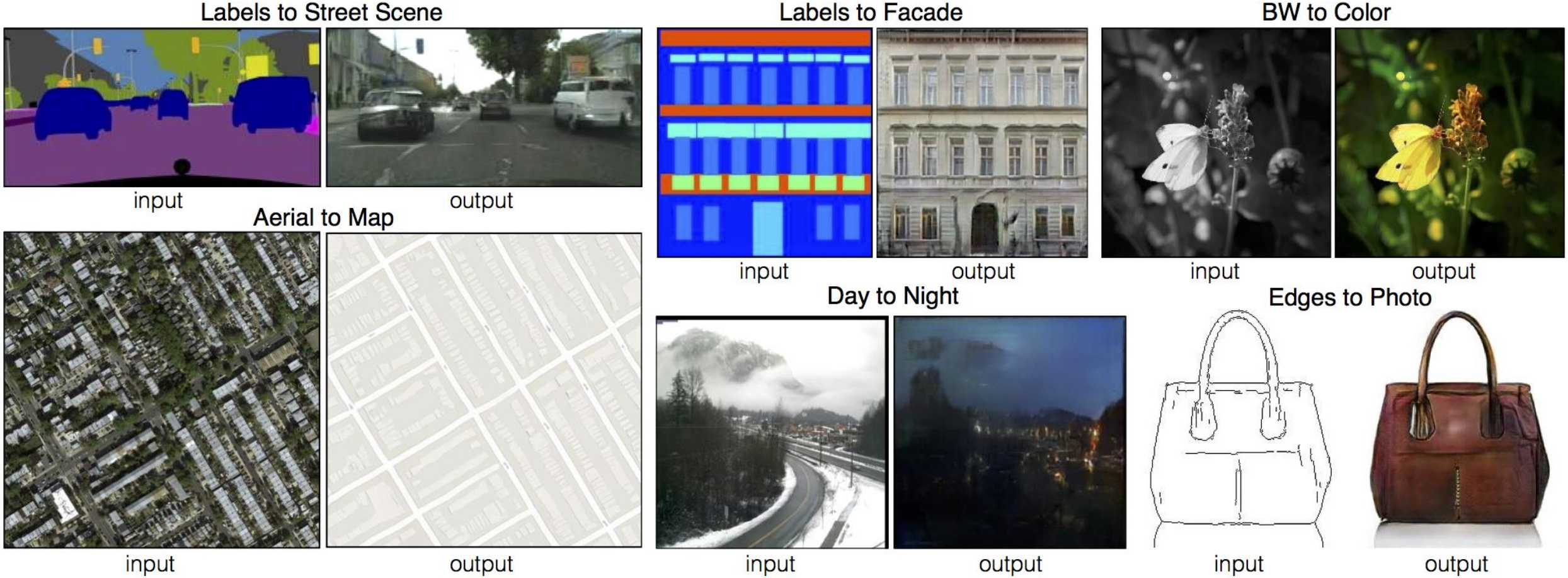



CycleGAN



# Recent advances in computer vision

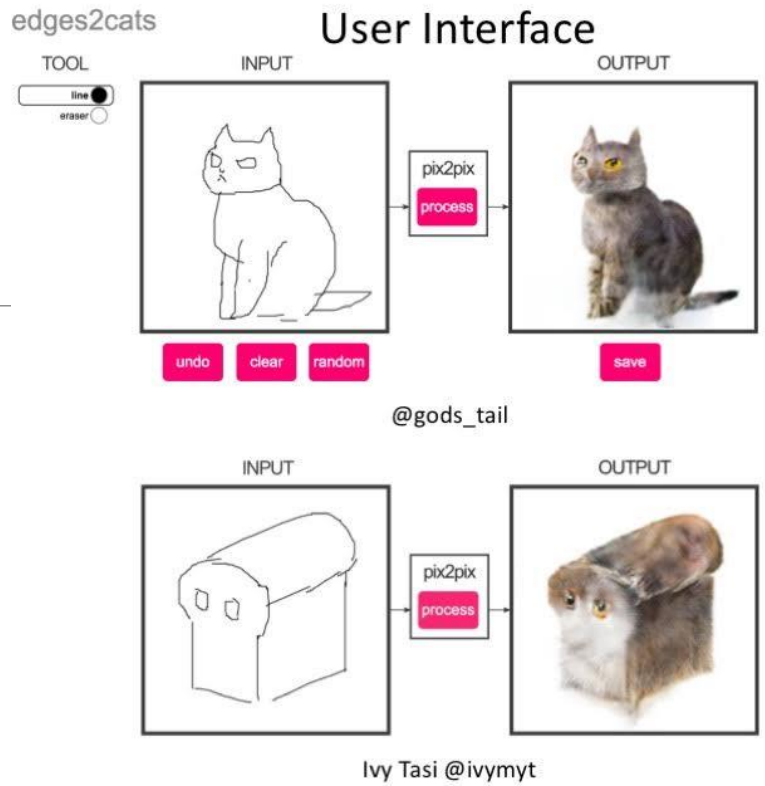
https://towardsdatascience.com/recent-advances-in-modern-computer-vision-56801edab980

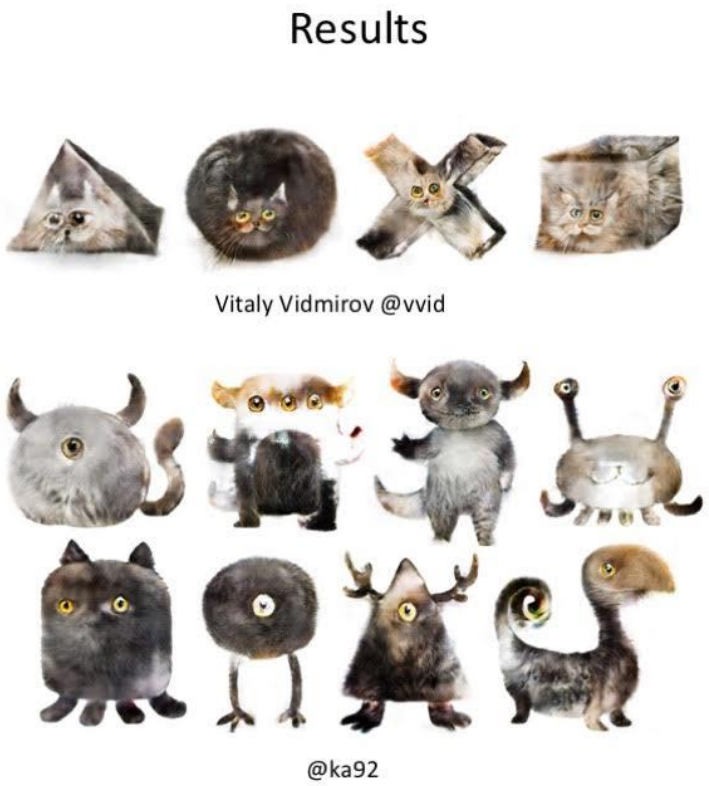



Pix2Pix

# Recent advances in computer vision

https://github.com/lood339/pytorch-two-GAN

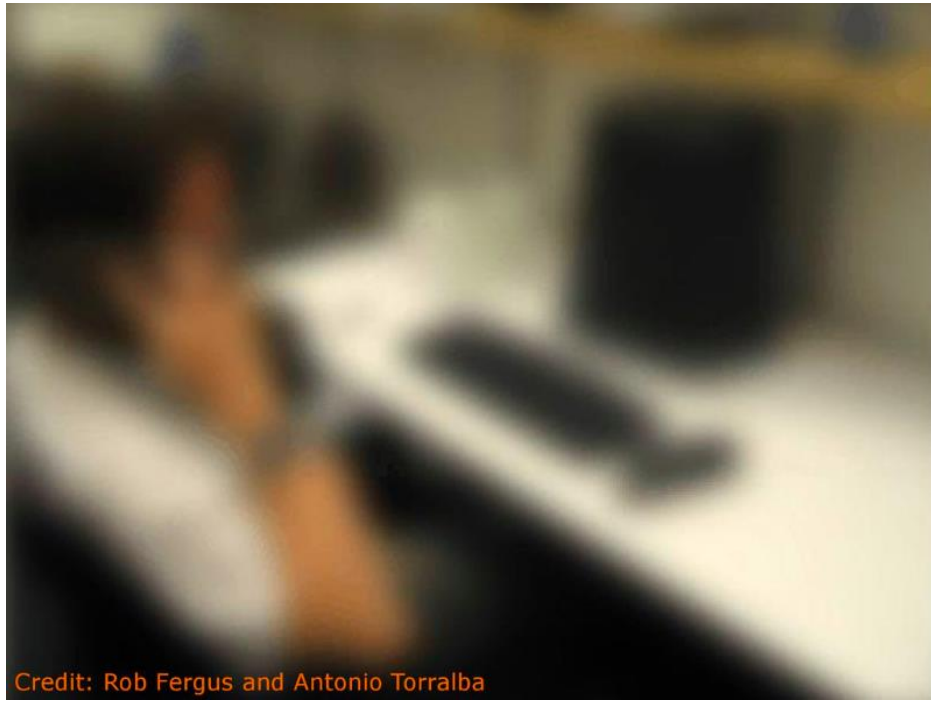


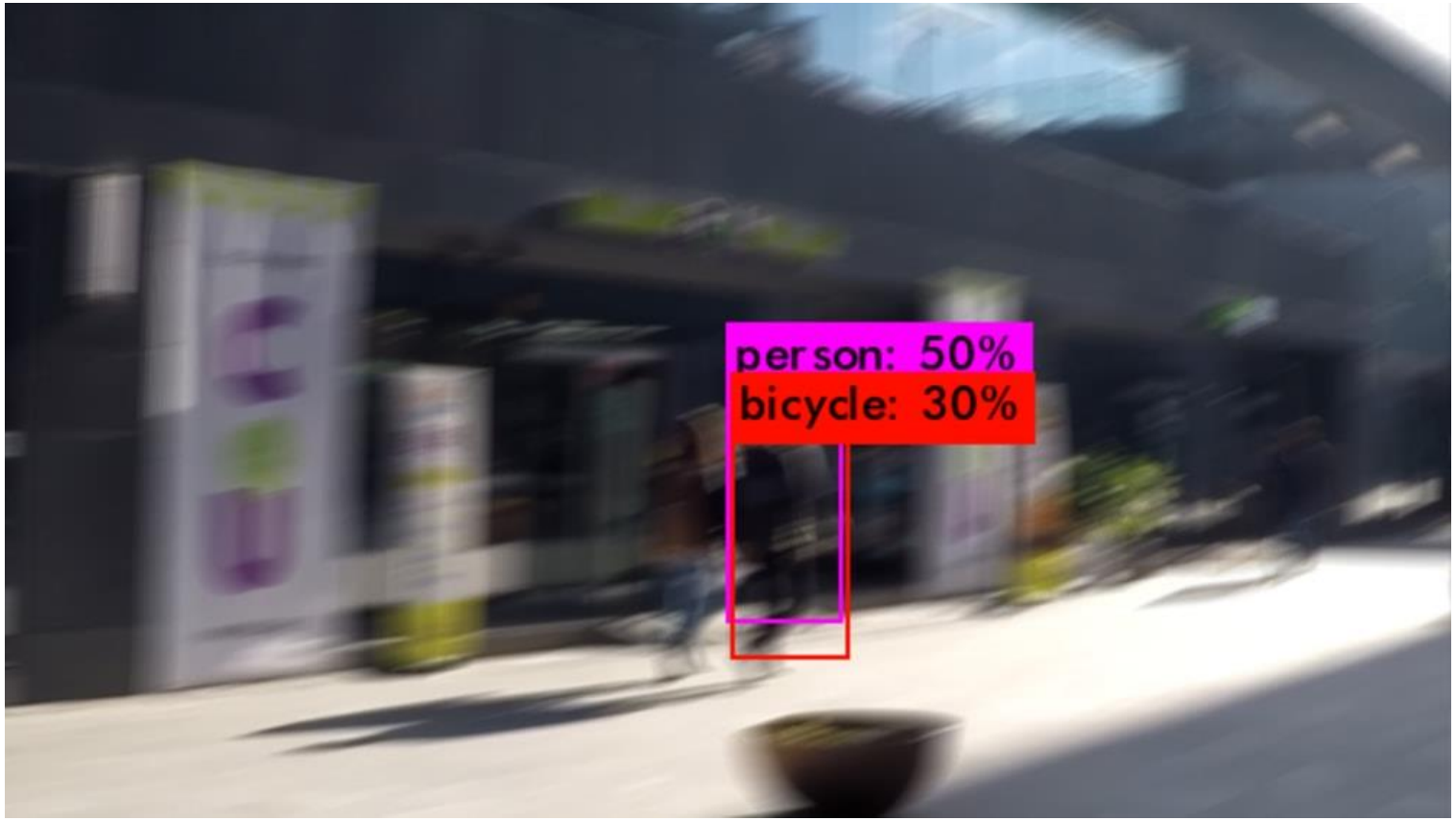




**Demo**
https://affinelayer.com/pixsrv/

# Recent advances in computer vision

https://github.com/lood339/pytorch-two-GAN




https://www.arxiv-vanity.com/papers/1711.07064/
DeBlurGAN

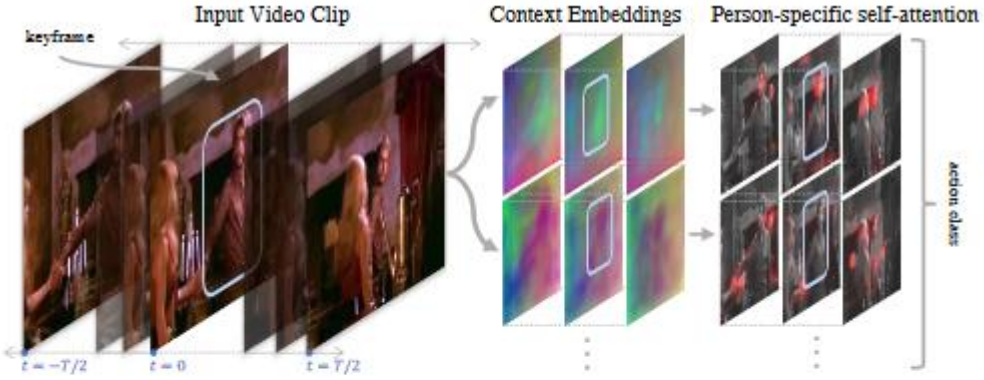
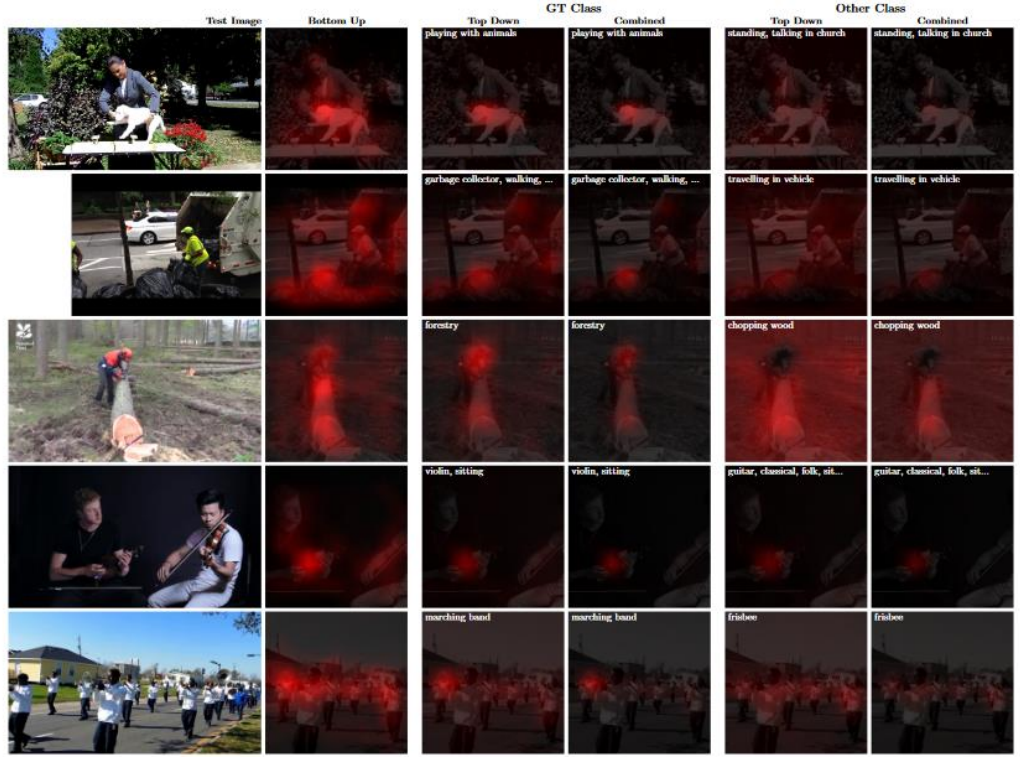
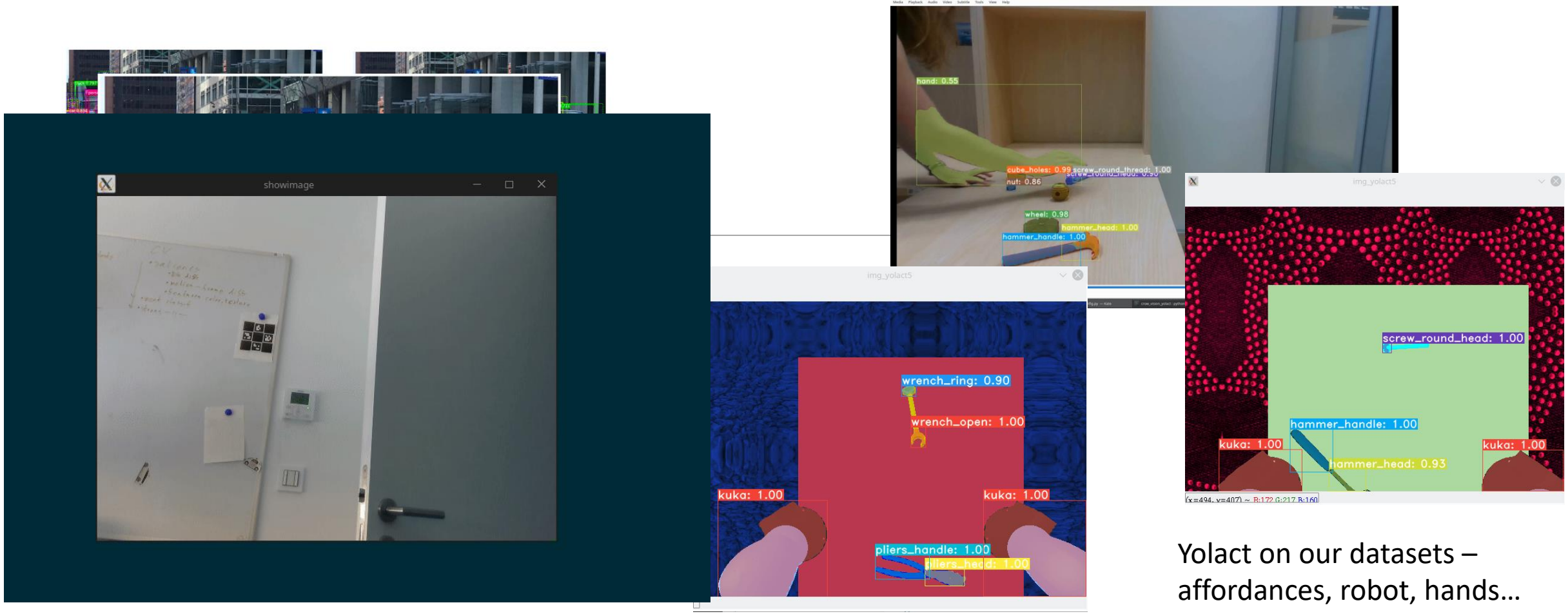



Figure 1: **Action Transformer in action.** Our proposed multi-head/layer Action Transformer architecture learns to attend to relevant regions of the person of interest and their context (other people, objects) to recognize the actions they are doing. Each head computes a clip embedding, which is used to focus on different parts like the face, hands and the other people to recognize that the person of interest is 'holding hands' and 'watching a person'.




https://arxiv.org/pdf/1812.02707.pdf

Video Action Transformer Network




Yolact on our datasets – affordances, robot, hands…

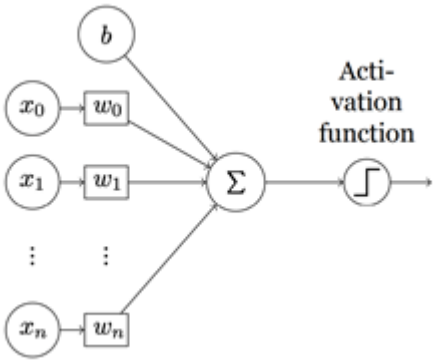
# Recent advances in computer vision

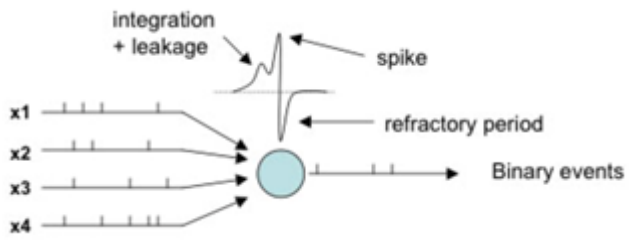
**The General problems of VA in machine vision:**

1. How can the system know what information is significant enough?
2. How does the visual system know when and how to direct attention and select significant information rather than doing so randomly?
3. Where is (are) the next potential target(s) of visual attention shifts? That is, how does it know where to actually focus its attention to?

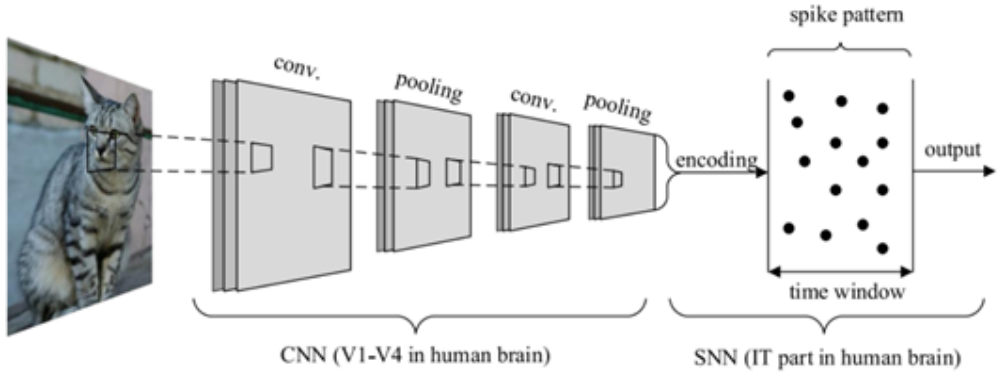
Event-based vision

Visual attention, Spiking neural networks…



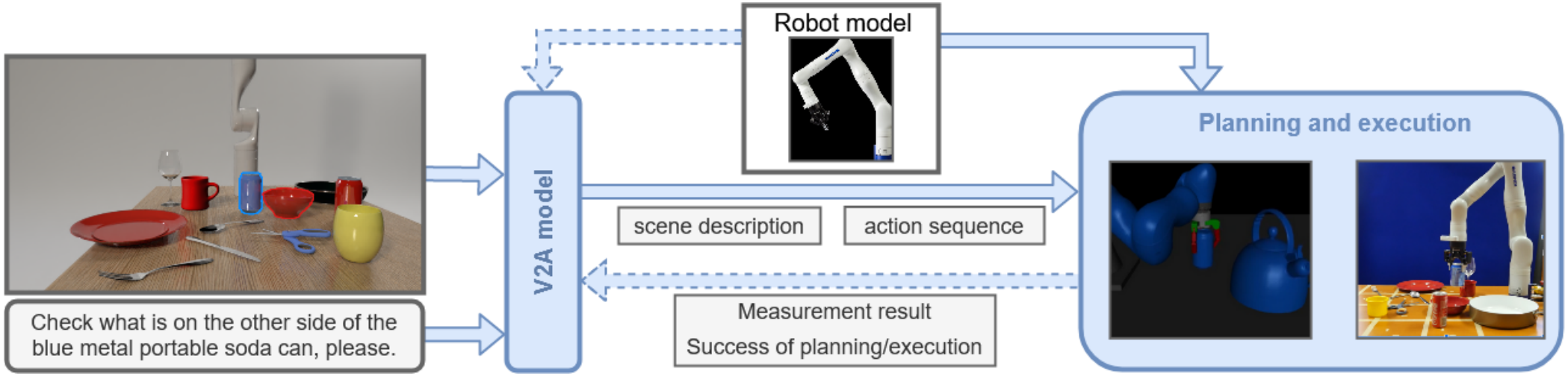

*Therefore, it can be said that the core objective of visual attention is to achieve the least possible amount of visual information to be processed to solve complex high-level tasks, e.g., object recognition.*




# Where next?

https://becominghuman.ai/from-human-vision-to-computer-vision-towards-spiked-based-visual-intelligence-and-neuromorphic-913e5de21bf9

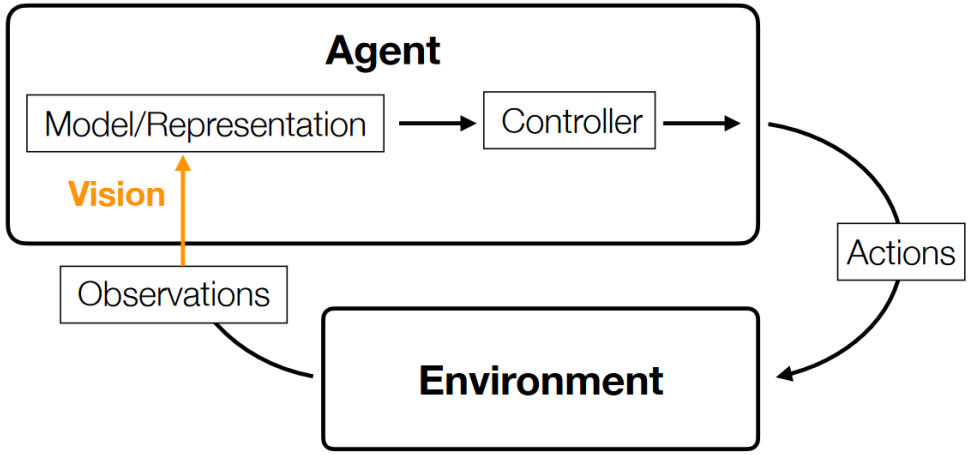
# Computer vision for embodied agents




Behrens, Jan Kristof, et al. "Embodied Reasoning for Discovering Object Properties via Manipulation." *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021.

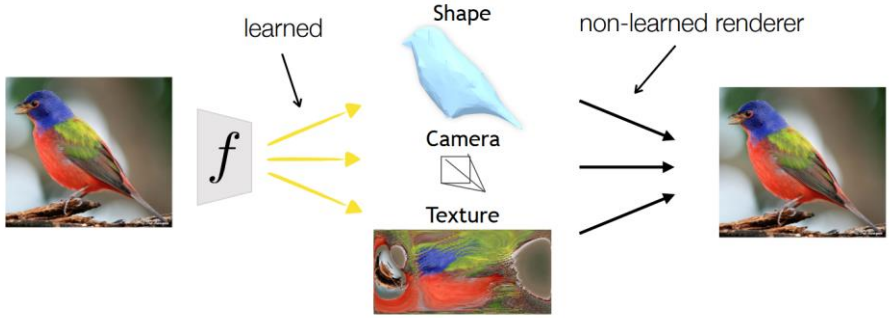
# Computer vision for embodied agents

Intelligent agents



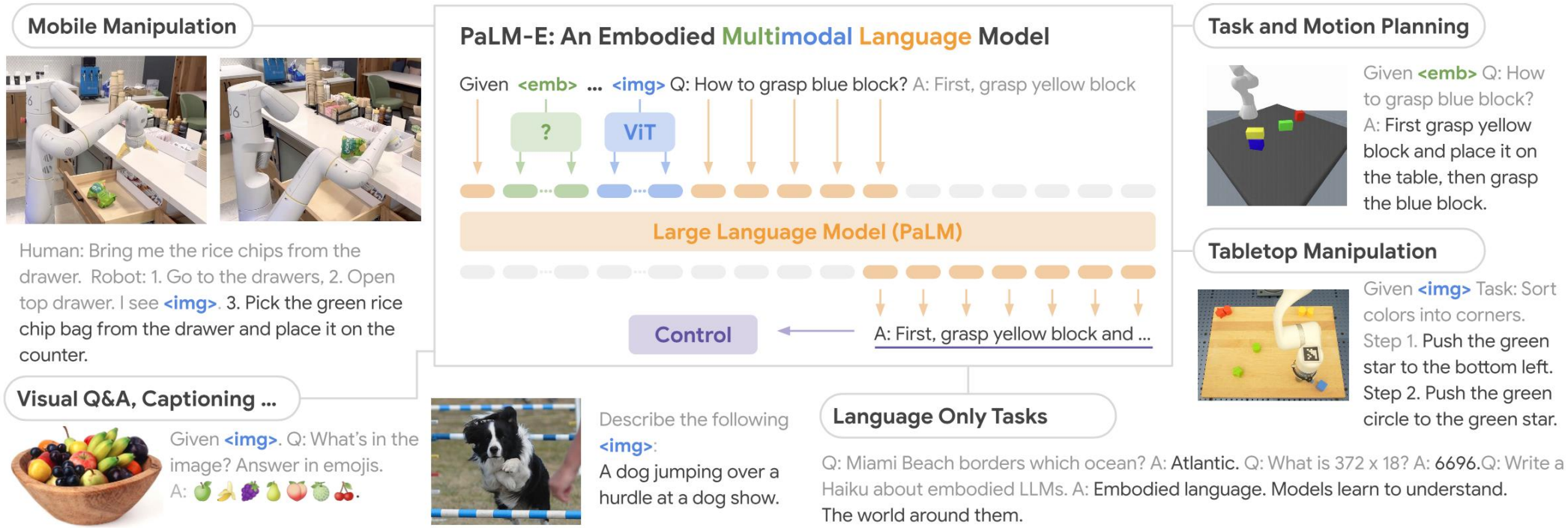

Funny looking autoencoder

Find a [shape, camera, texture] combination that renders to the image.

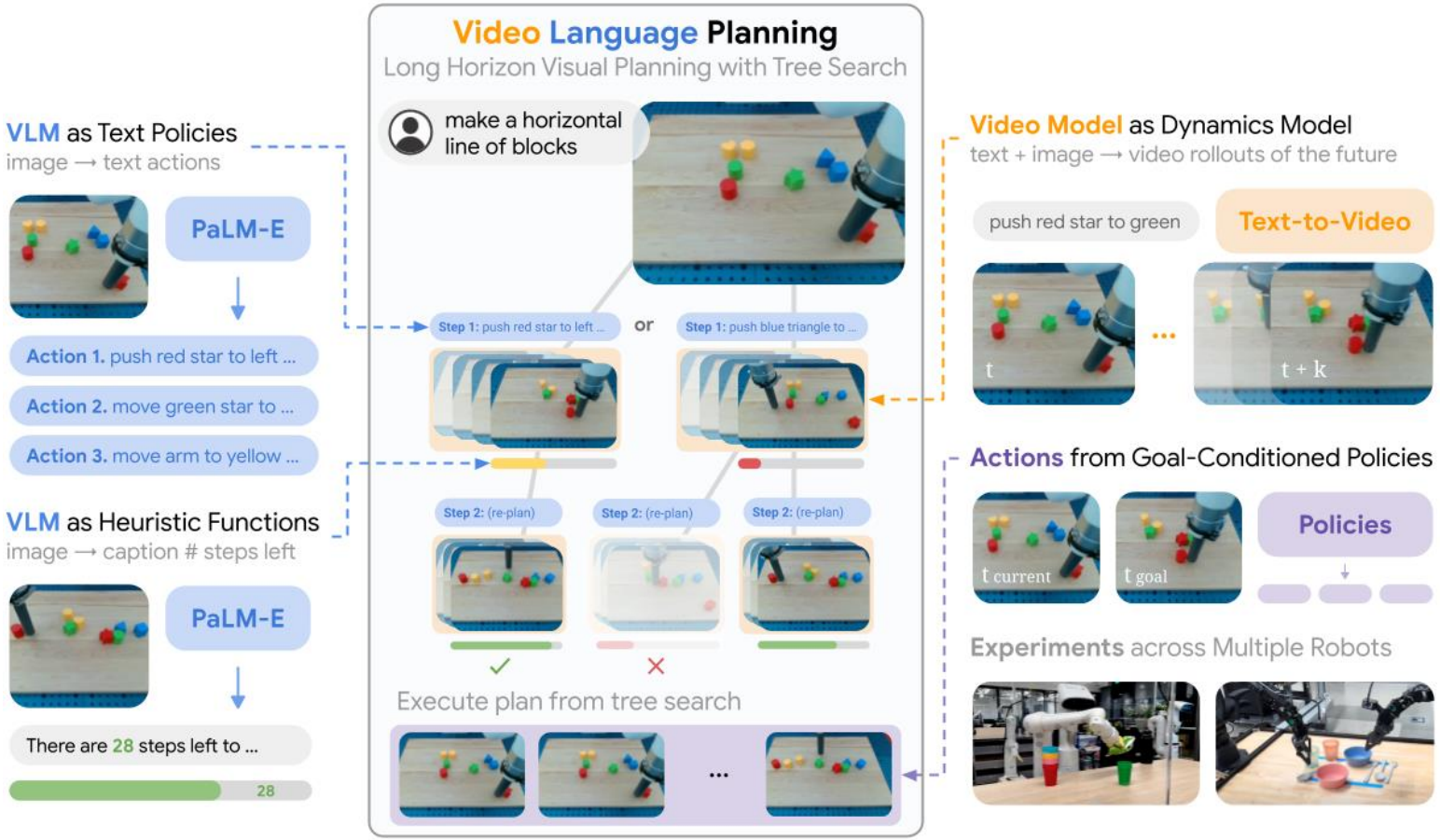



http://6.869.csail.mit.edu/sp22/lectures/L16/16_vision_for_embodied_agents.pdf

# Computer vision for embodied agents



# Computer vision for embodied agents



**Checklist for Comparison Studies**

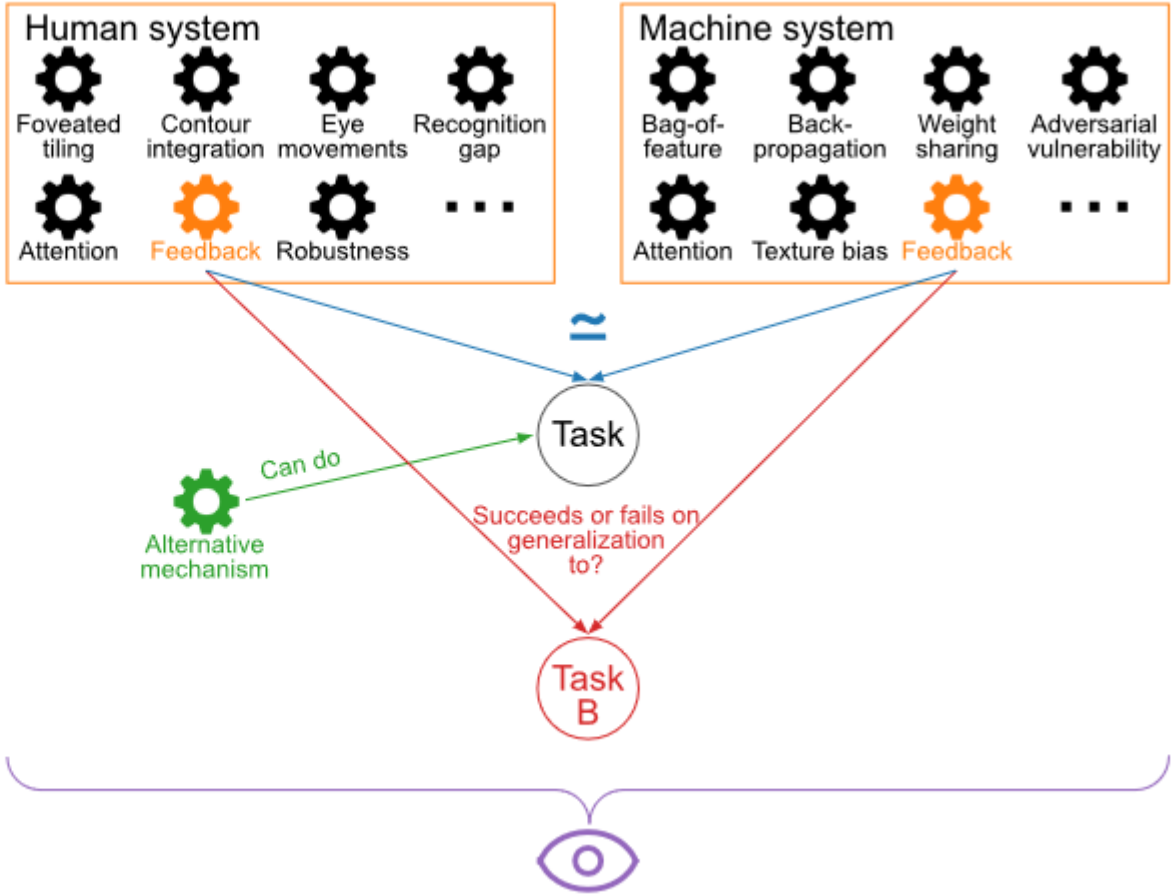
i. Isolate implementational or functional properties
Ex: BagNets[1] isolate local feature processing

ii. Align experimental conditions
Ex: Stimulus timing for humans (short presentation time = core object recognition[2])

iii. Distinguish necessary vs. sufficient
Ex: Both shape and texture features are sufficient but not necessary on ImageNet[3,4]

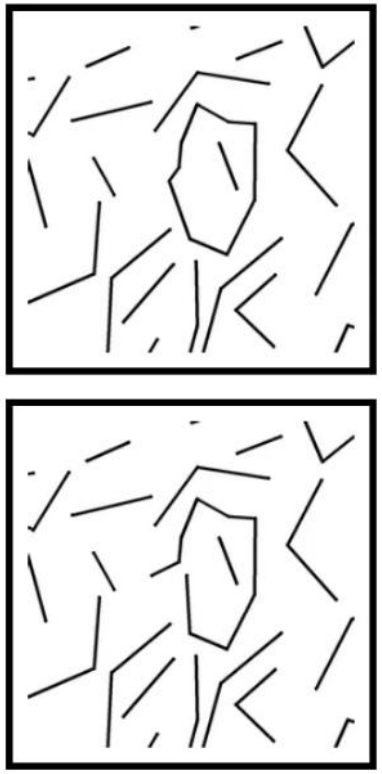
iv. Test generalization
Ex: Controversial stimuli[5] are useful out-of-distribution data that reveal human inductive bias

v. Resisting human bias
Ex: Stimulus selection and data analysis influence identified similarities b/w human and machine perception of adversarial examples[6]

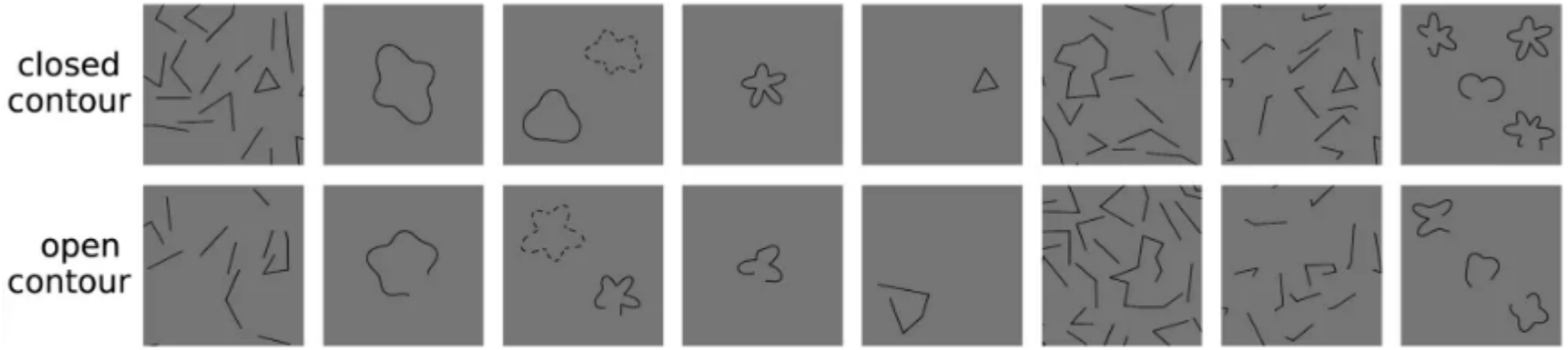



# Comparing human and AI vision

https://bdtechtalks.com/2020/08/10/computer-vision-deep-learning-vs-human-perception/
https://arxiv.org/abs/2004.09406



*Can you tell which one of the above images contains a closed shape?*

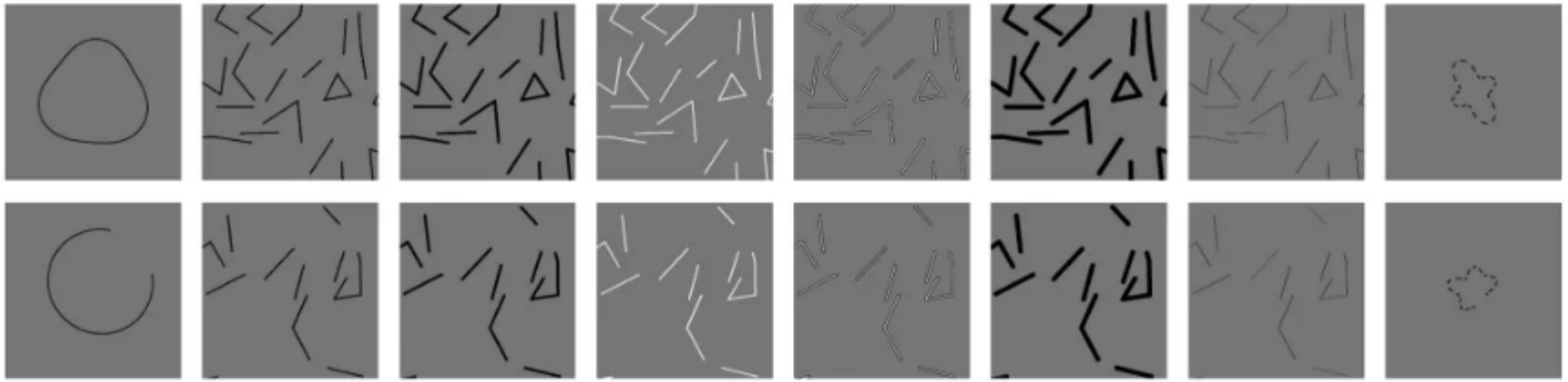



*The ResNet neural network was able to detect various open- and closed-contour images, despite having been trained on straight-line examples only.*

# Comparing human and AI vision

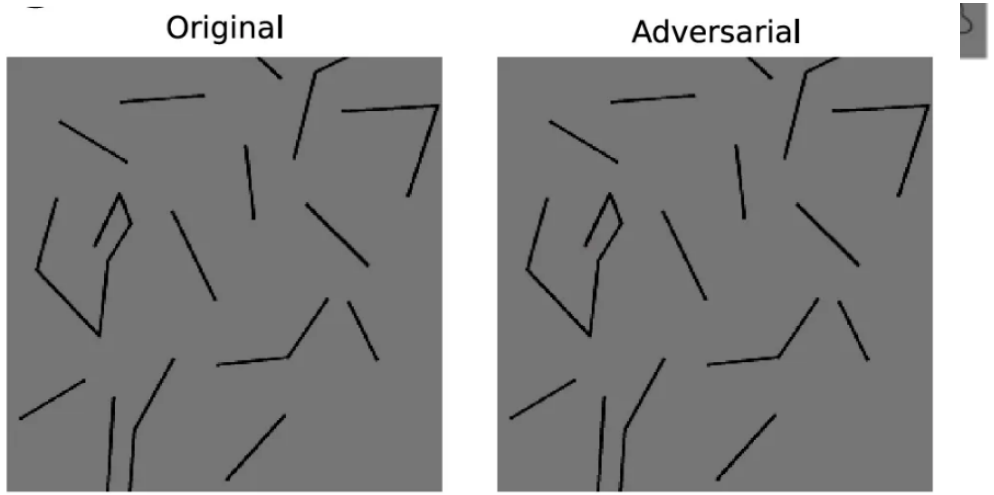
https://bdtechtalks.com/2020/08/10/computer-vision-deep-learning-vs-human-perception/
https://arxiv.org/abs/2004.09406

*Can you tell which one of the above images contains a closed shape?*



*The ResNet-50 neural network struggled when presented with images that contained lines of different colors and thickness, and where shapes were larger than the training set.*

*The ResNet neural network was able to detect varic straight-line examples only.*

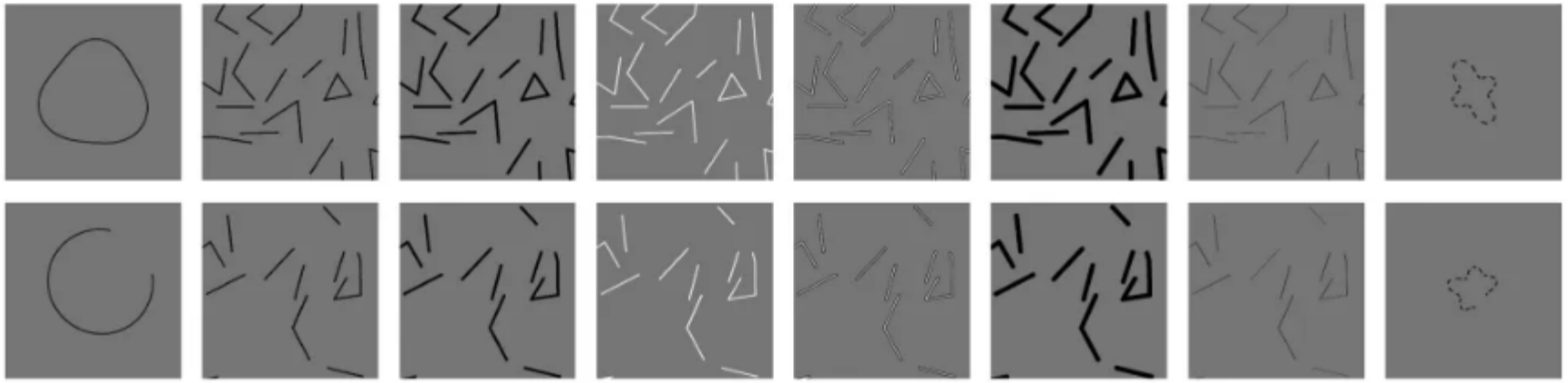



*The image on the right has been modified with adversarial perturbations, noise that is imperceptible by humans. To the human eye, both images are identical. But to a neural network, they are different images.*

# Comparing human an

https://bdtechtalks.com/2020/08/10/computer-vision-deep-learning-vs-human-perception/
https://arxiv.org/abs/2004.09406

*Can you tell which one of the above images contains a closed shape?*

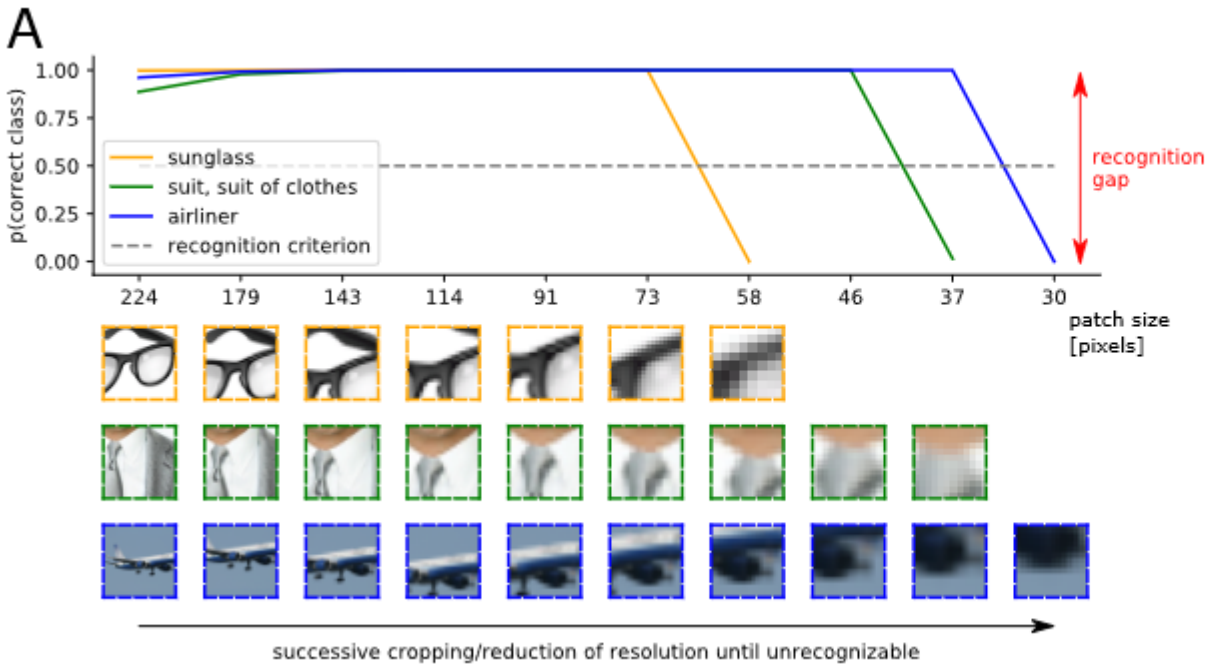


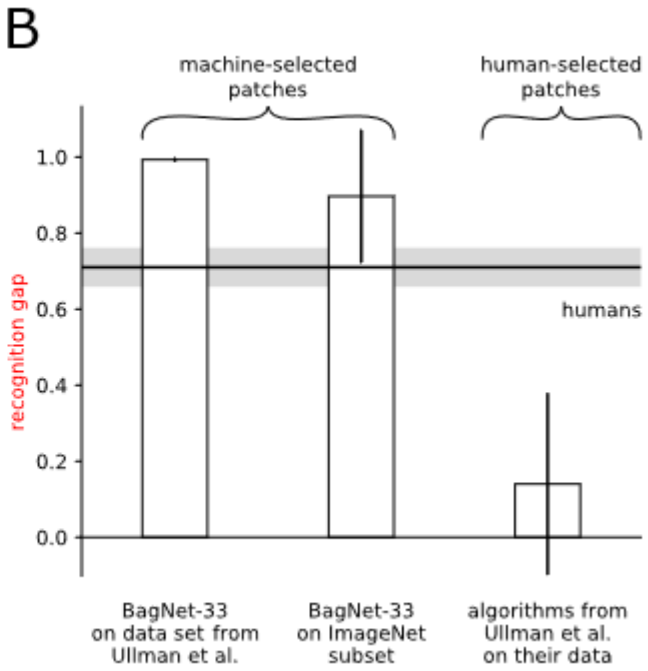
*The ResNet-50 neural network struggled when presented with images that contained lines of different colors and thickness, and where shapes were larger than the training set.*

As our AI systems become more complex, we will have to develop more complex methods to test them. Previous work in the field shows that many of the popular benchmarks used to measure the accuracy of computer vision systems are misleading.

# Comparing human and AI vision

https://bdtechtalks.com/2020/08/10/computer-vision-deep-learning-vs-human-perception/
https://arxiv.org/abs/2004.09406




# Comparing human and AI vision

https://bdtechtalks.com/2020/08/10/computer-vision-deep-learning-vs-human-perception/
https://arxiv.org/abs/2004.09406

- Survailence, face detection
- Medical imaging – cancer detection, etc.
- Traffic cameras
- Quality assurance – infrared thermocameras
- Same pictures as trained
- …

http://www.cs.cmu.edu/~stein/BSIFT/example_objects.jpg

- Invariant object recognition
- Attention
- Emotions, memories
- Occlusions – human takes into account embodiments of agents
- Similarities, but not overdetecting (e.g. we see the chair is on the picture, not in reality easily)
- Semantic retrieval
- New objects

# Comparing human and AI vision

Example image that illustrates the complexity of recognizing chairs as no single technique in terms of contours, appearance, components etc. is adequate to correctly allow 'counting of the number of chairs'. (Source: Bülthoff, Max Planck Institute for Biological Cybernetics (MPIK), Tübingen, Germany.)