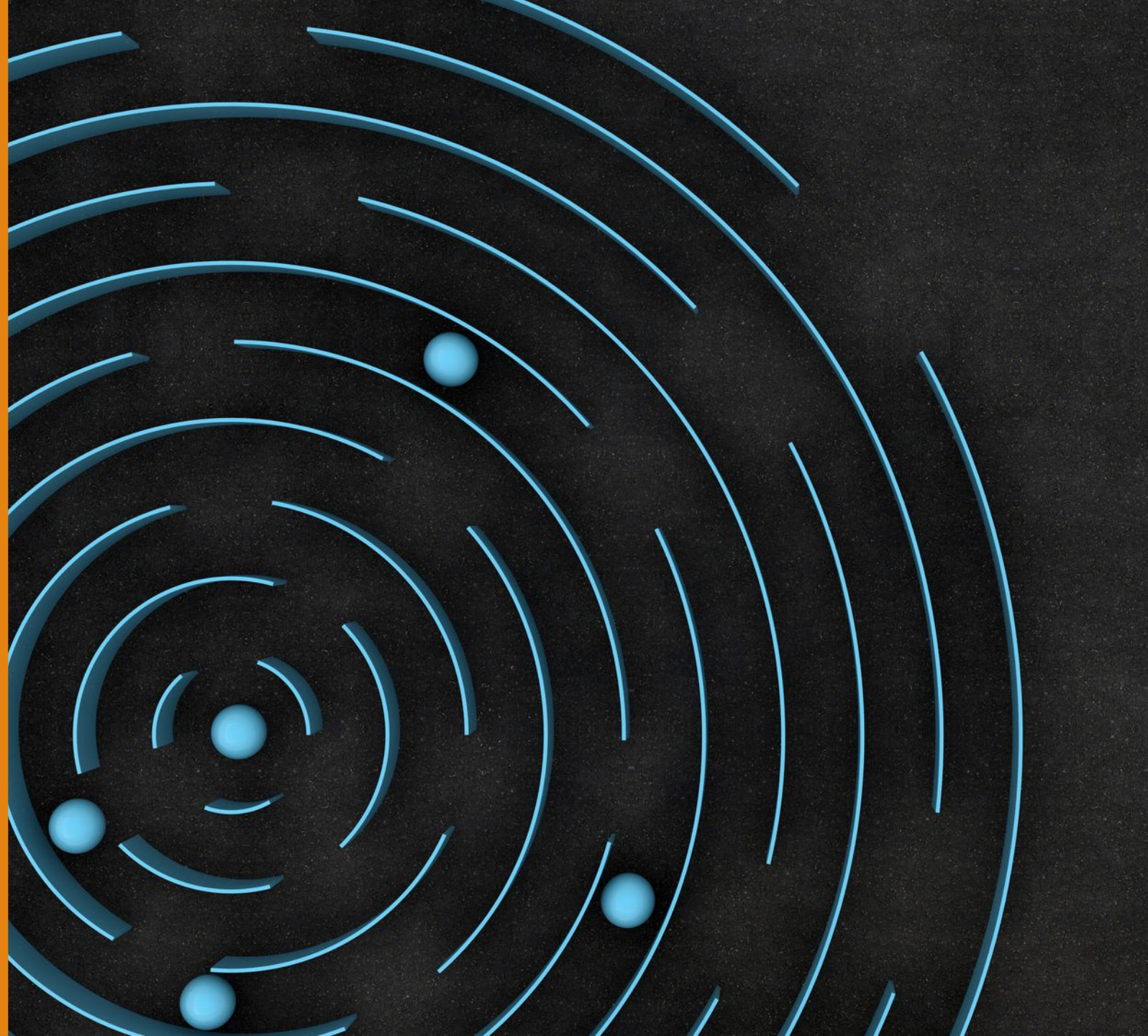
Cognitive systems

PERCEPTION: VISION

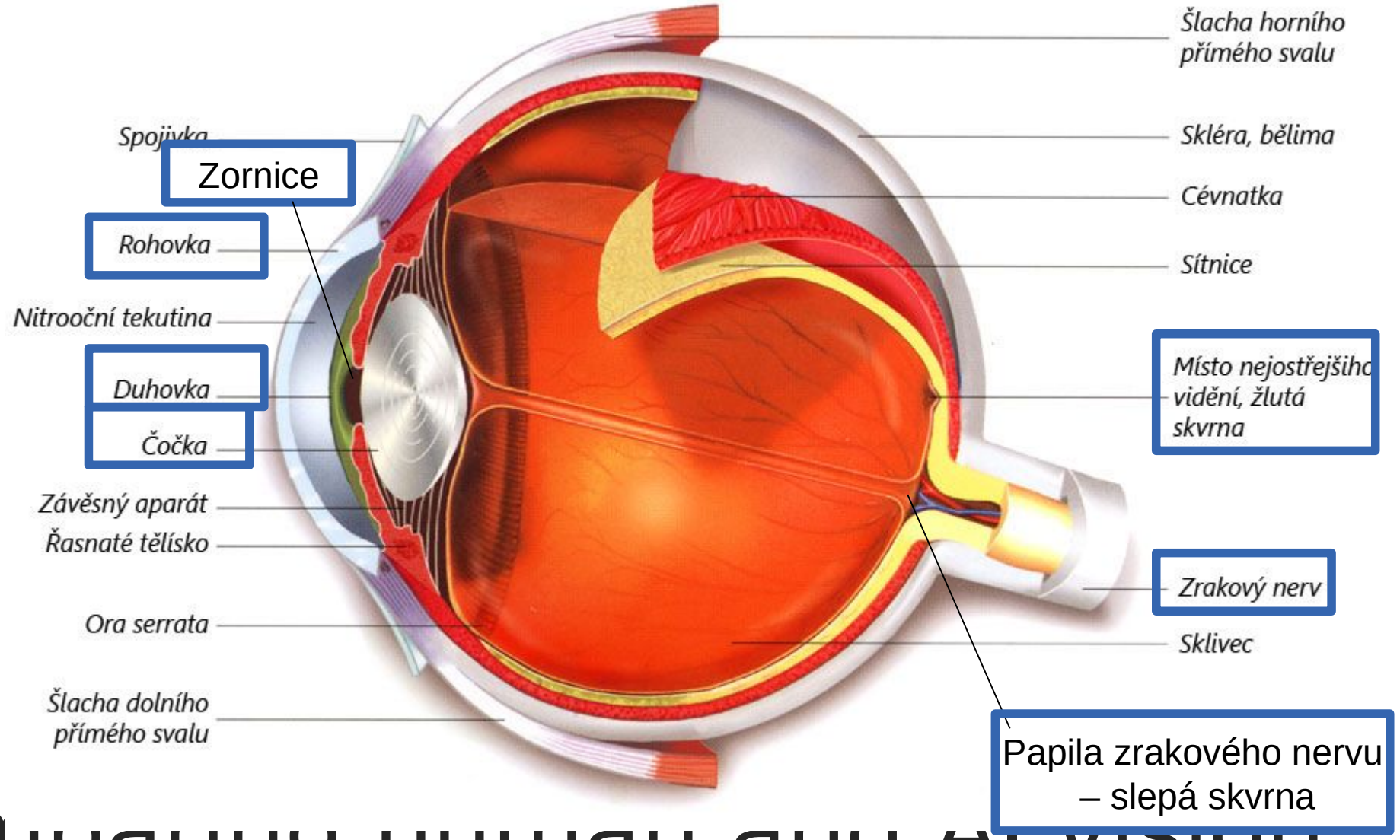
HUMAN VS. AI (COMPUTER VISION)



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?

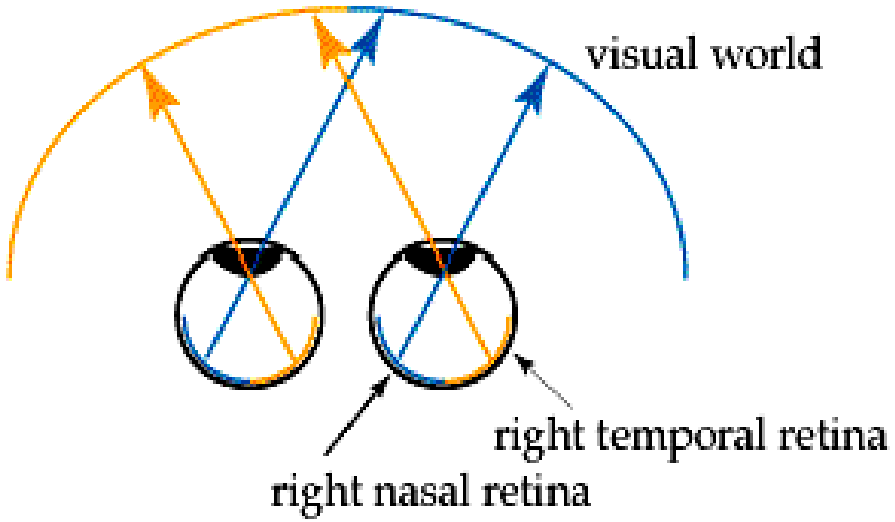
Perception

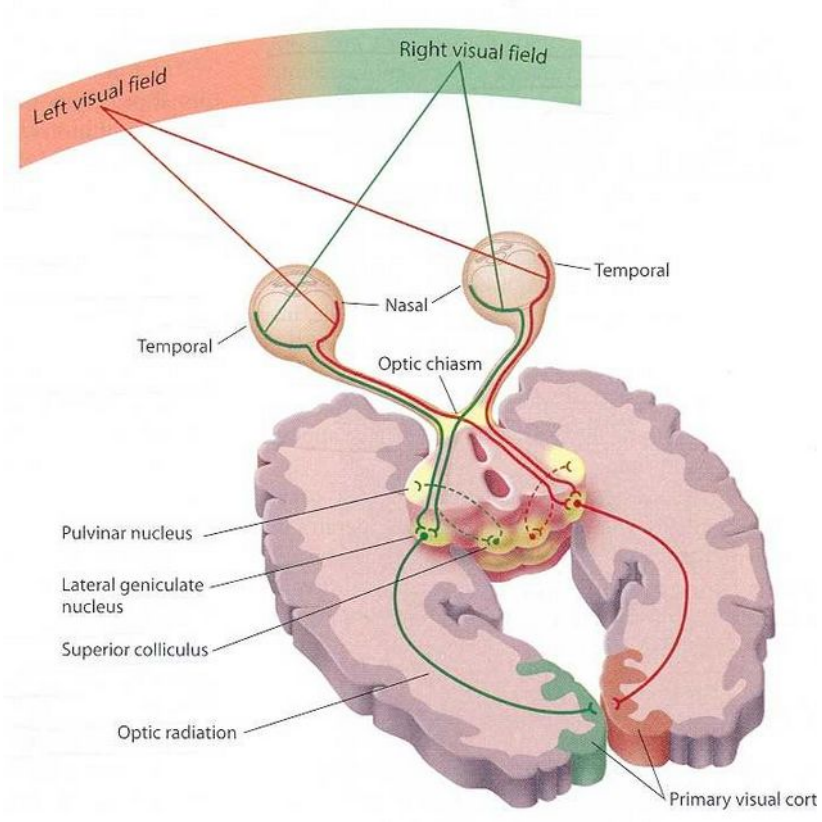


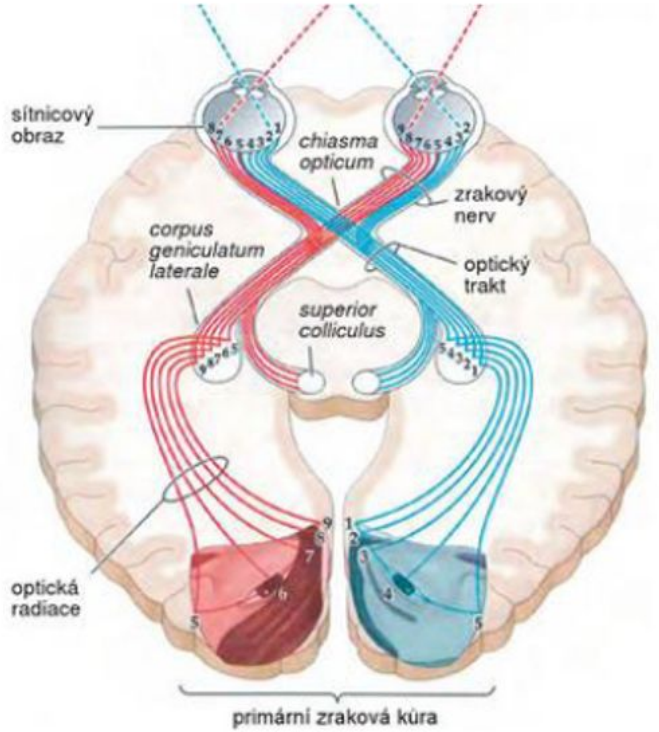
Comparing human and AI vision

Chiasma optikum

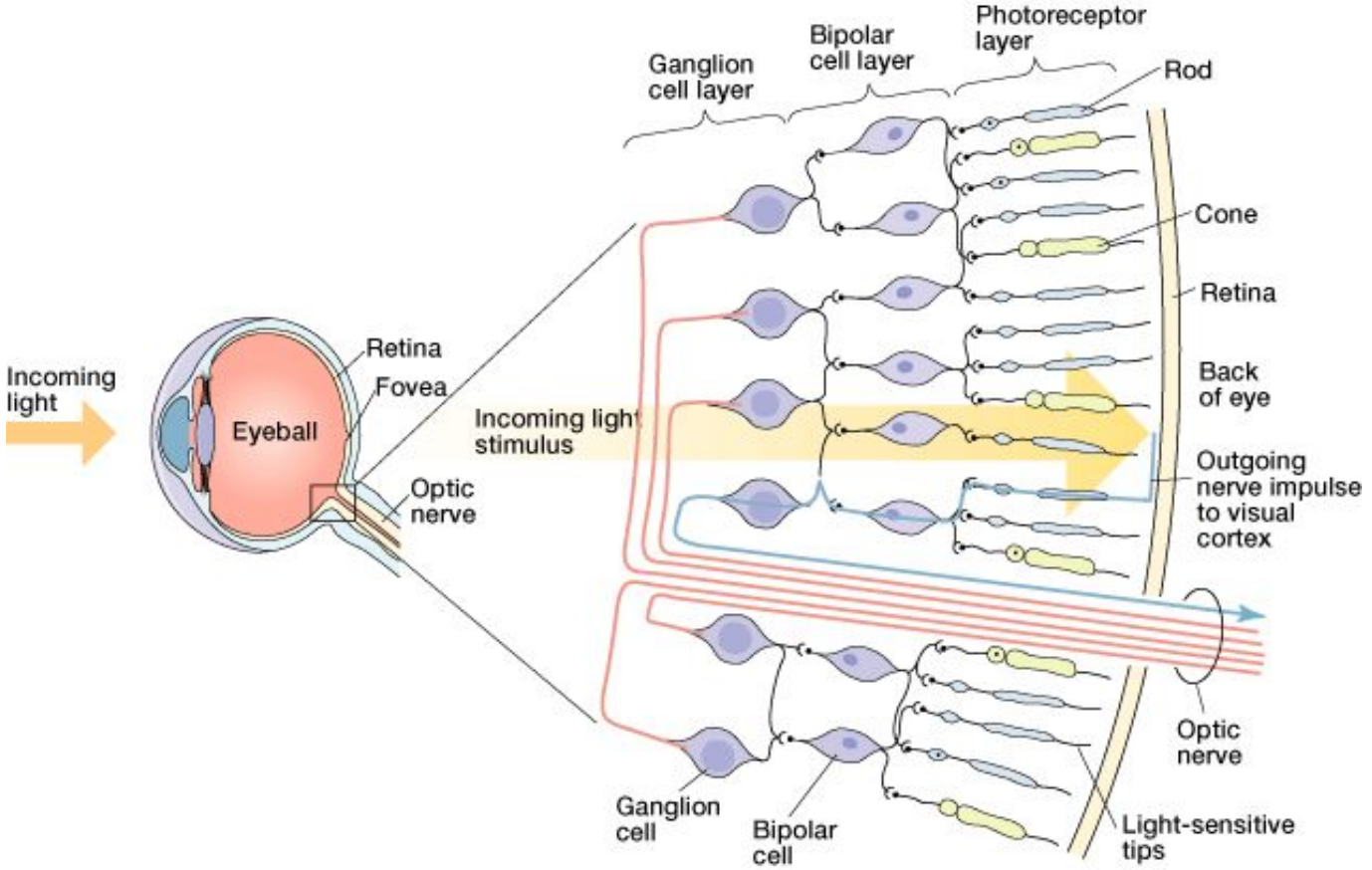
Překřížení drah optického nervu, aby byla zpracovávána separátně pravá a levá část zorného pole







Fotoreceptors

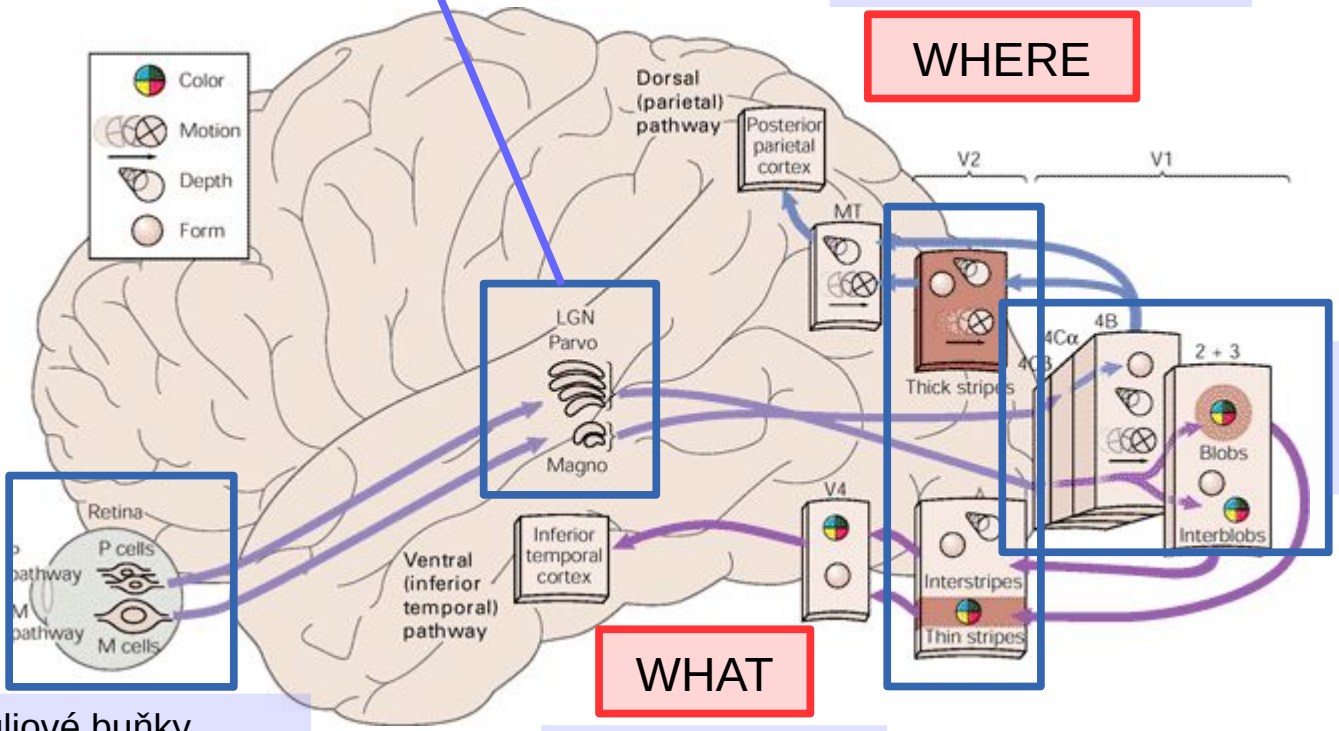


- Za sítnicí pigmentový epitel – absorpce (melanin)
- Axony nemyelizované → transparentní
- ve žluté skvrně ostatní buňky na straně

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

Dorsální dráha
Visuoprostorové funkce, ne stat.objekty

WHERE

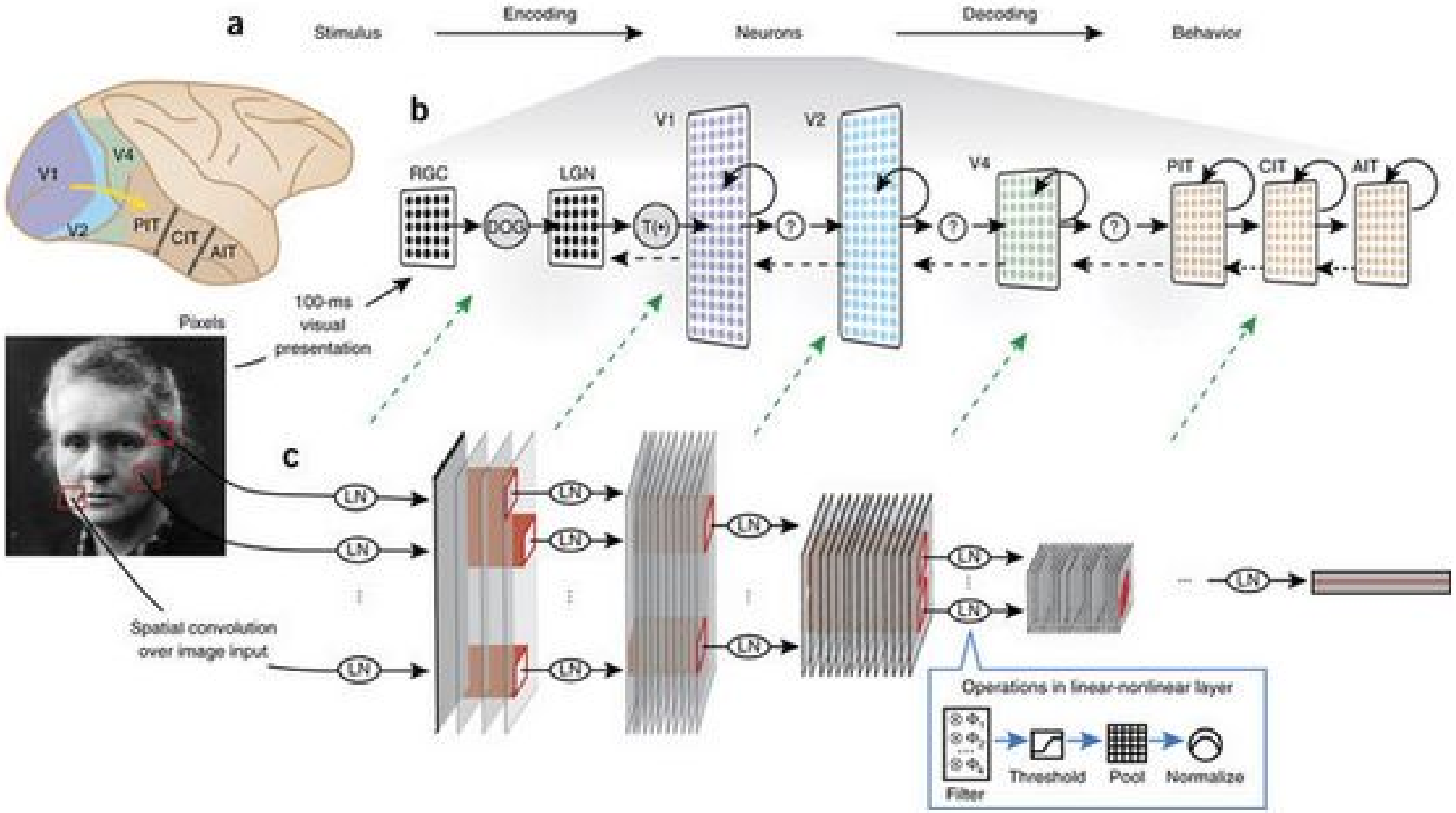


Primární viz.kůra
V1

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

WHAT

Ventrální dráha
Citlivé na obrysy,
Velké rozlišení

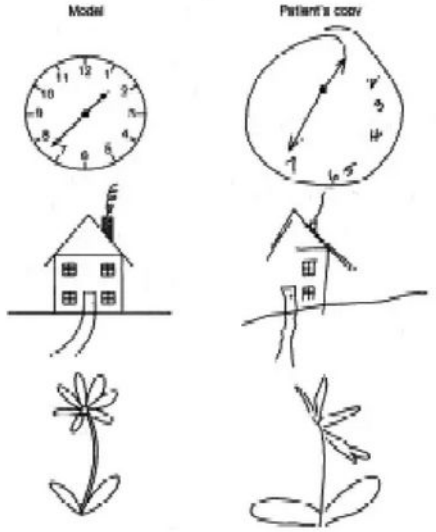


https://becominghuman.ai/from-human-vision-to-computer-vision-convolutional-neural-network-part3-4-24b55ffa7045

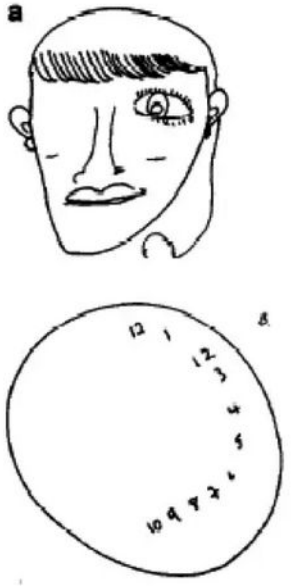
Poruchy vnímání (hemispatial neglect)

- Left visual neglect after stroke

Copying:

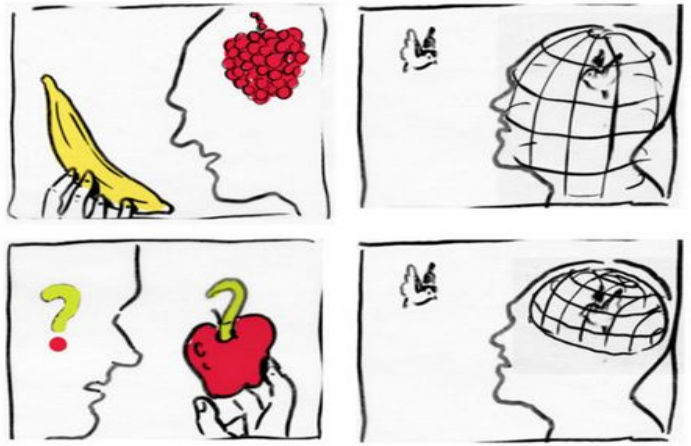


Spontaneous drawing:



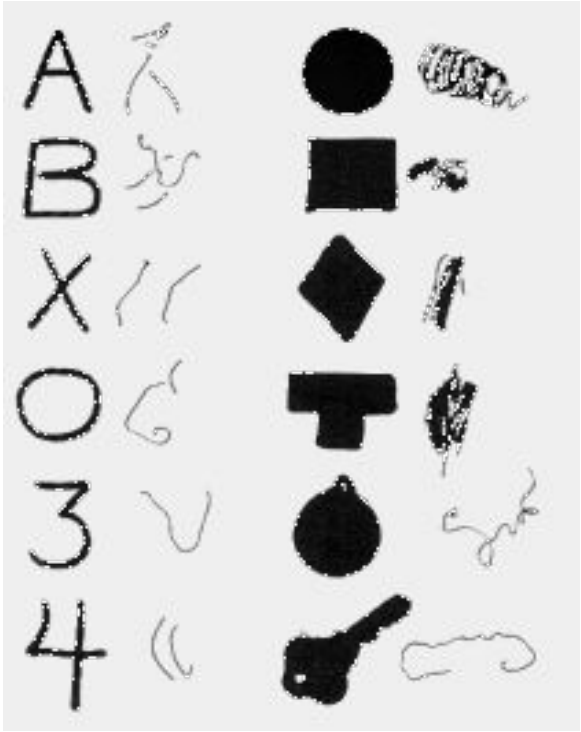
Poruchy vnímání

- 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

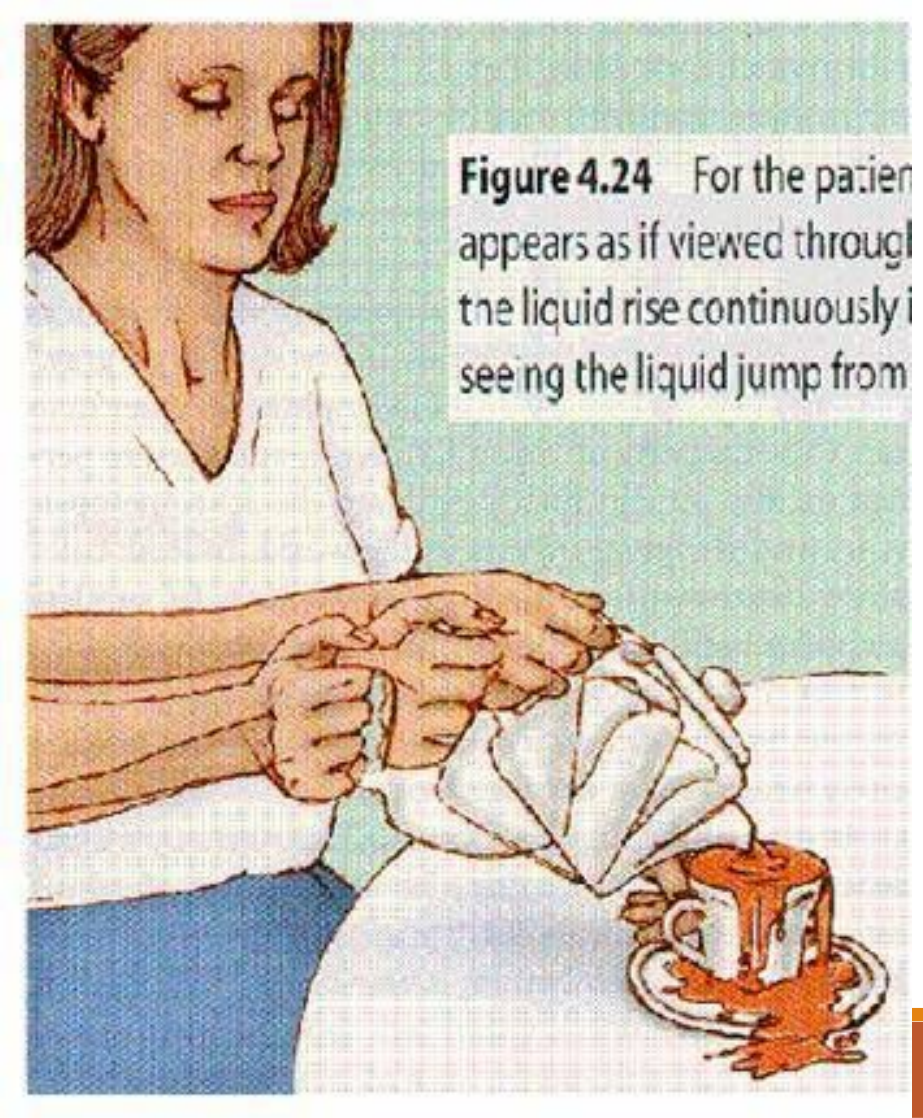


Akinetopsie

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

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

http://youtu.be/B47Js1MtT4w

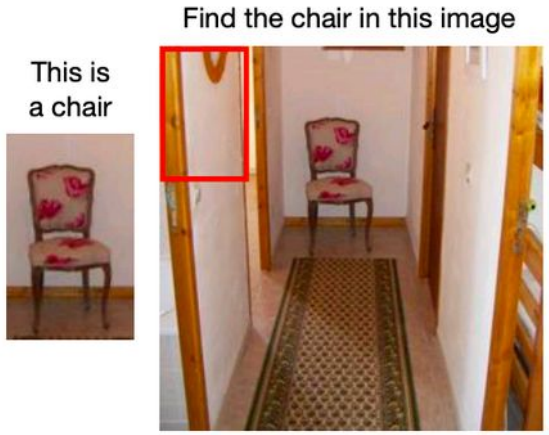


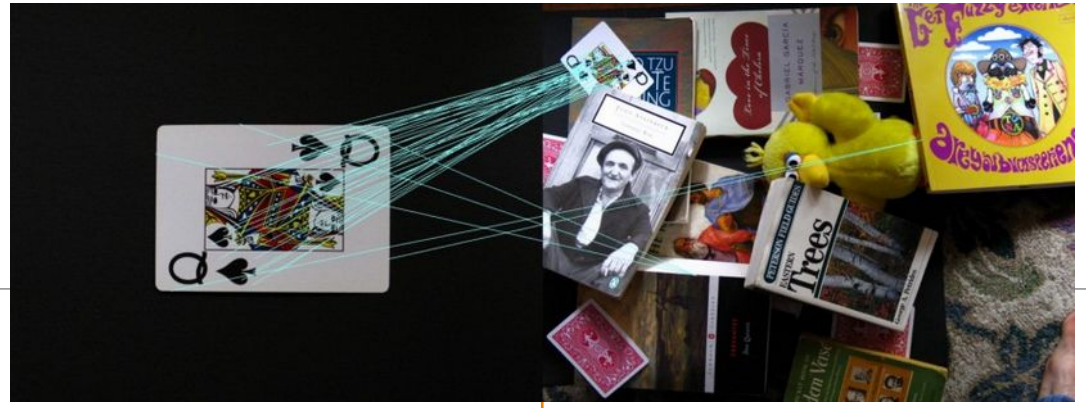
Poruchy vnímání

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

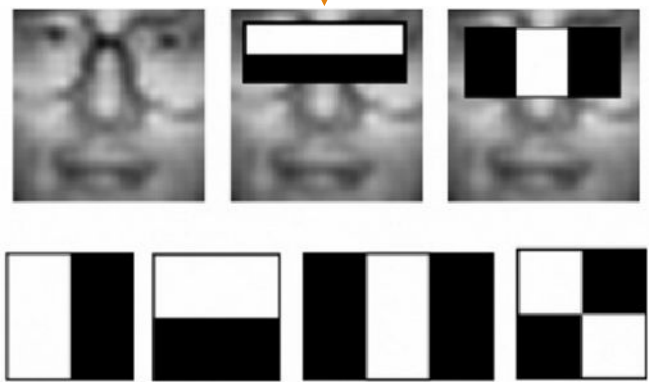
...

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





1999, David Lowe

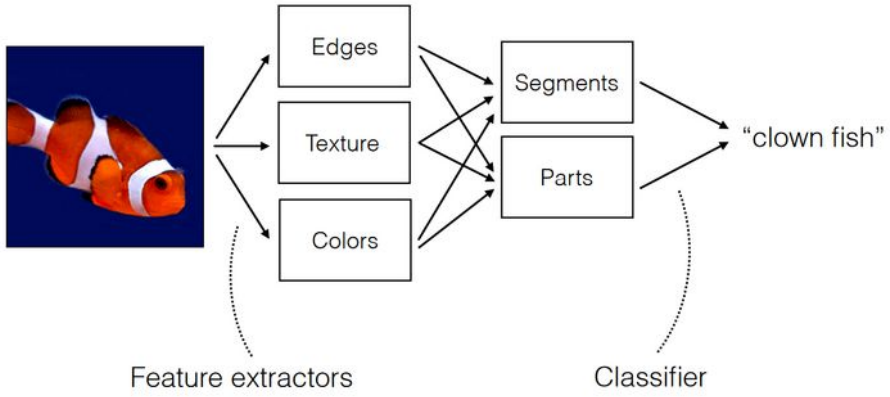


2000, Viola-Jones face detection

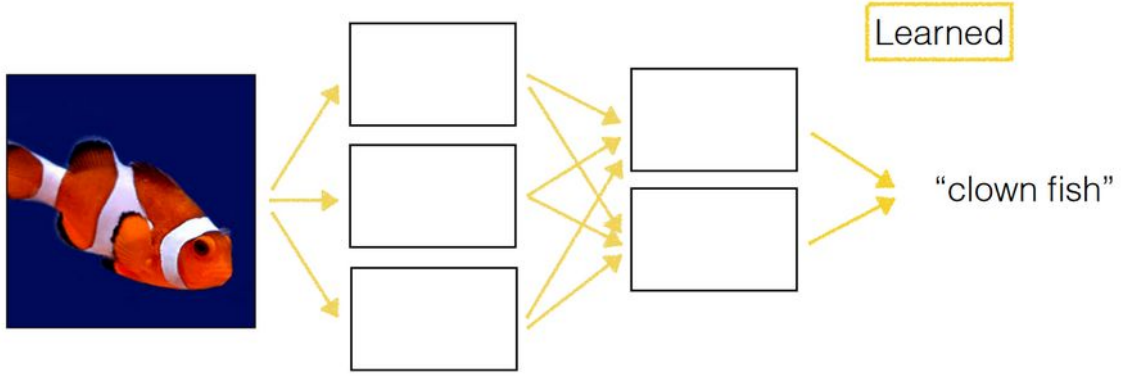
AlexNet, 2012

Recent advances in computer vision

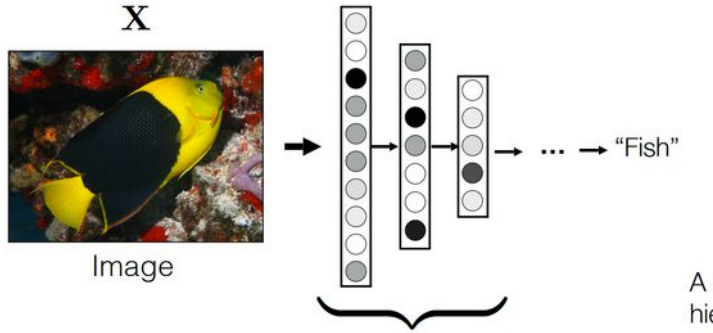
Classical object recognition



Deep learning



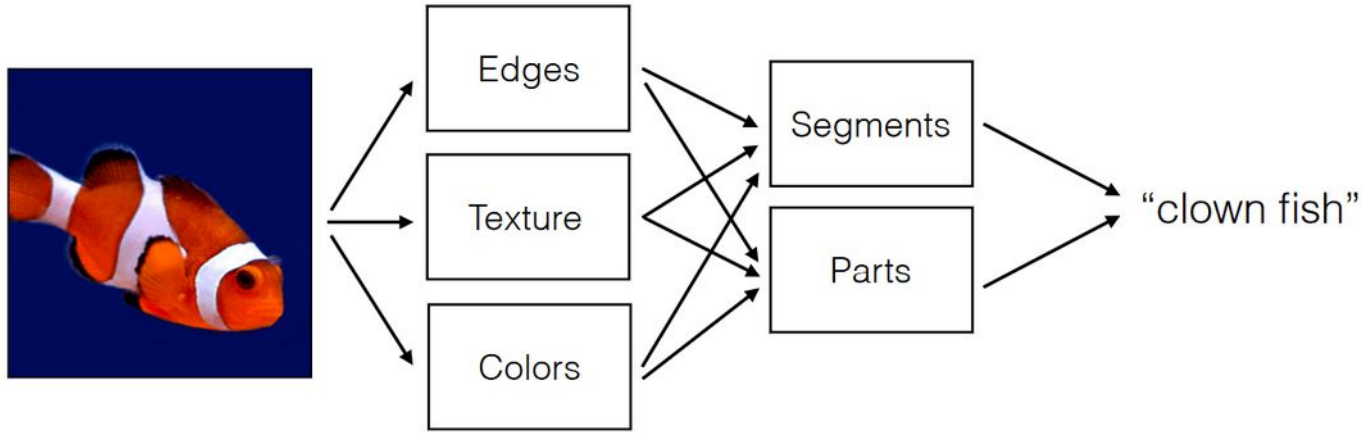
What do deep nets internally learn?

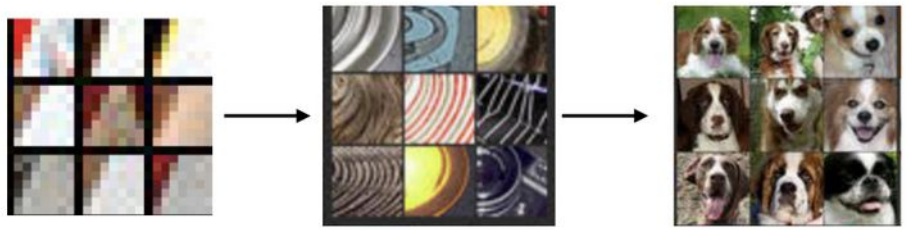


Representations!

A CNN is a multiscale, hierarchical representation of data

CNNs *learned* the classical visual recognition pipeline!

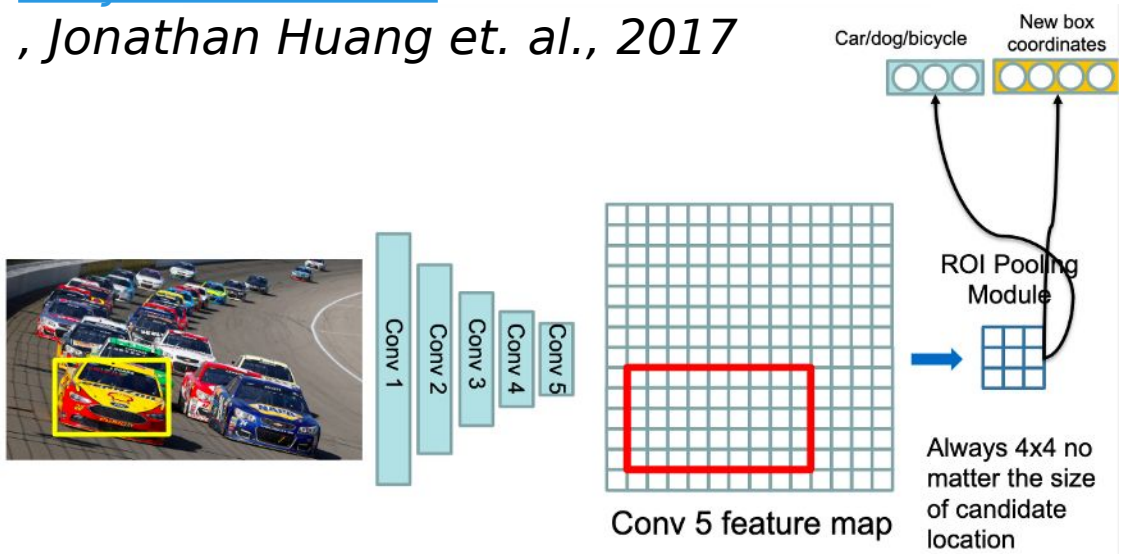




AlexNet, 2012
https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

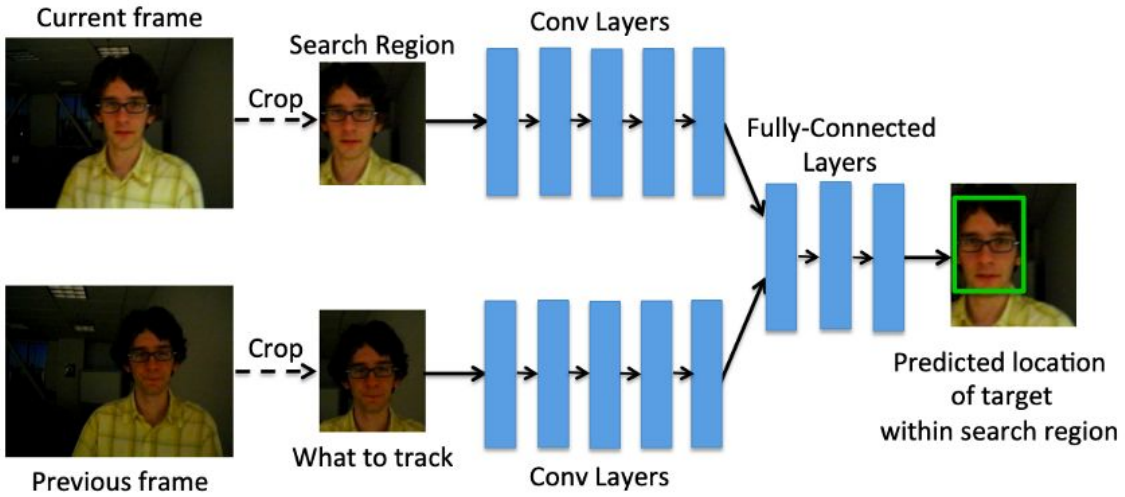
Detection Frameworks	Train	mAP	FPS
Fast R-CNN [5]	2007+2012	70.0	0.5
Faster R-CNN VGG-16[15]	2007+2012	73.2	7
Faster R-CNN ResNet[6]	2007+2012	76.4	5
YOLO [14]	2007+2012	63.4	45
SSD300 [11]	2007+2012	74.3	46
SSD500 [11]	2007+2012	76.8	19

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

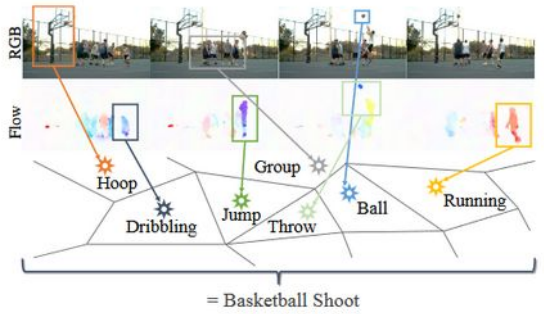
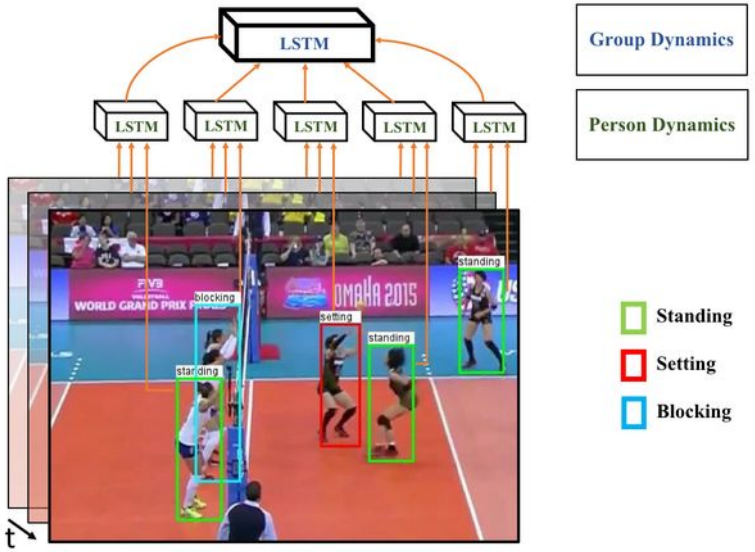
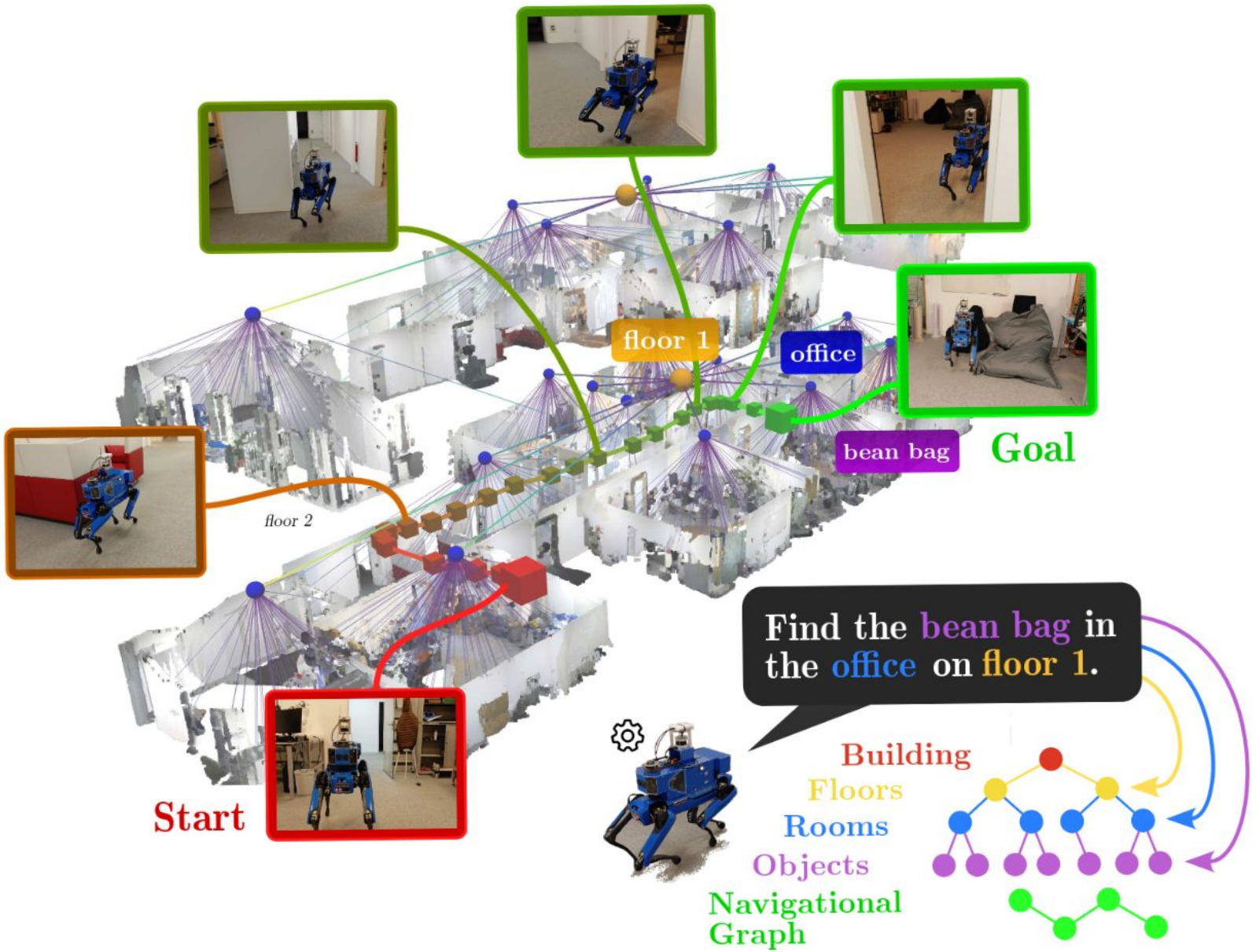


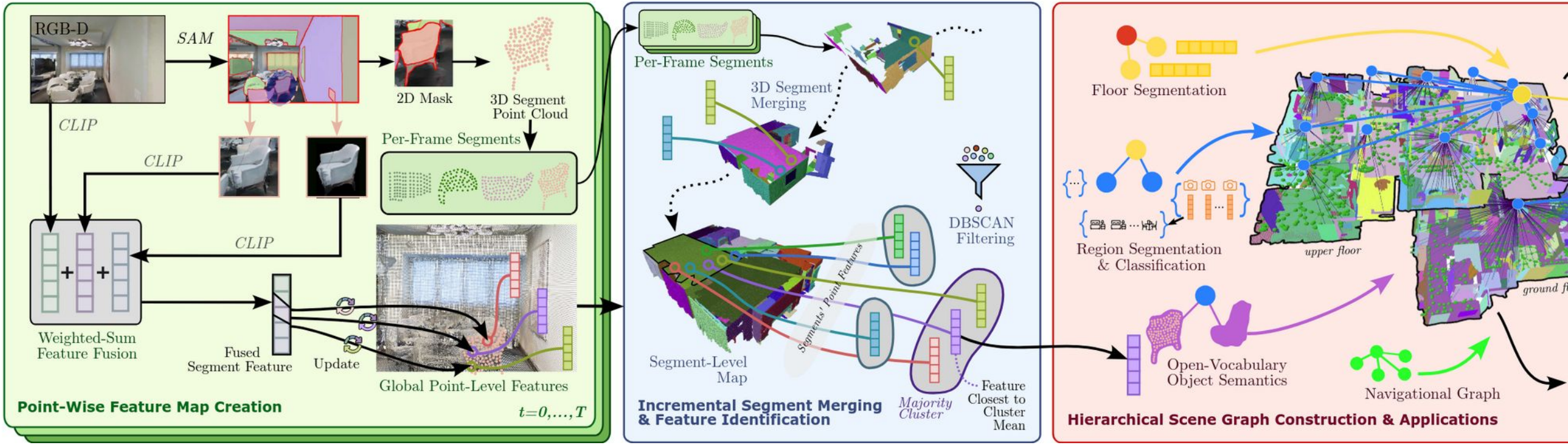
Figure 1: How do we represent actions in a video? We propose ActionVLAD, a spatio-temporal aggregation of a set of action primitives over the appearance and motion streams of a video. For example, a basketball shoot may be represented as an aggregation of appearance features corresponding to 'group of 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

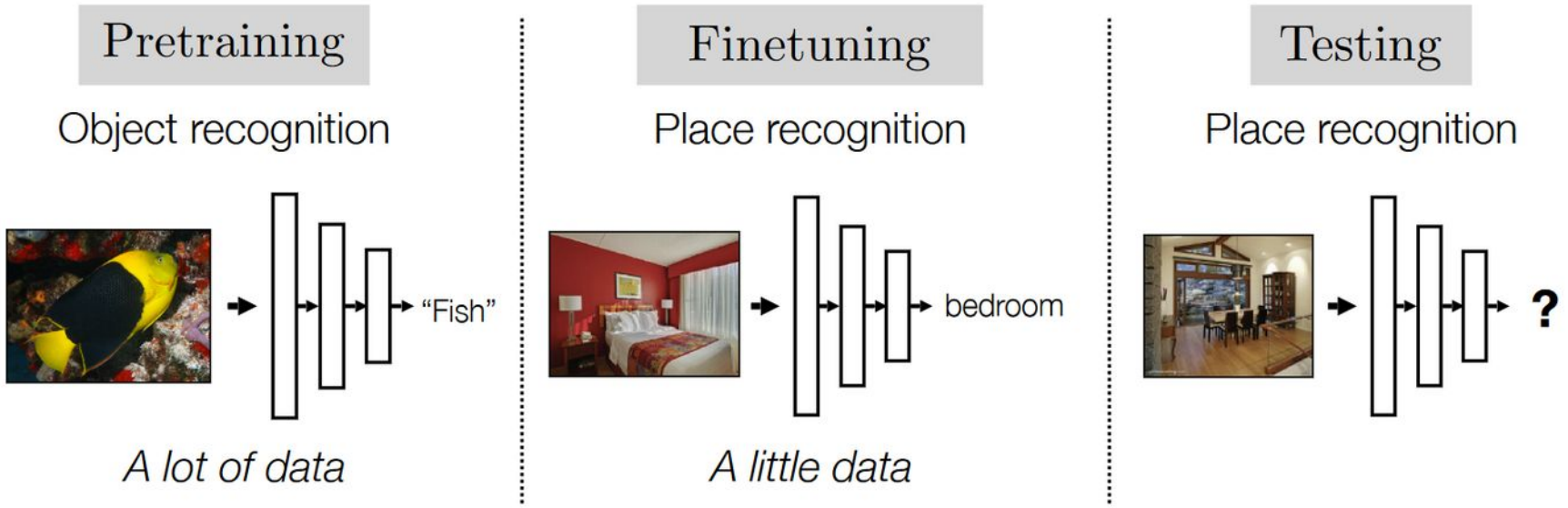
Recent advances in computer vision


floor 1
office
bean bag
Goal
floor 2
Start
Find the bean bag in the office on floor 1.
Building
Floors
Rooms
Objects
Navigational Graph



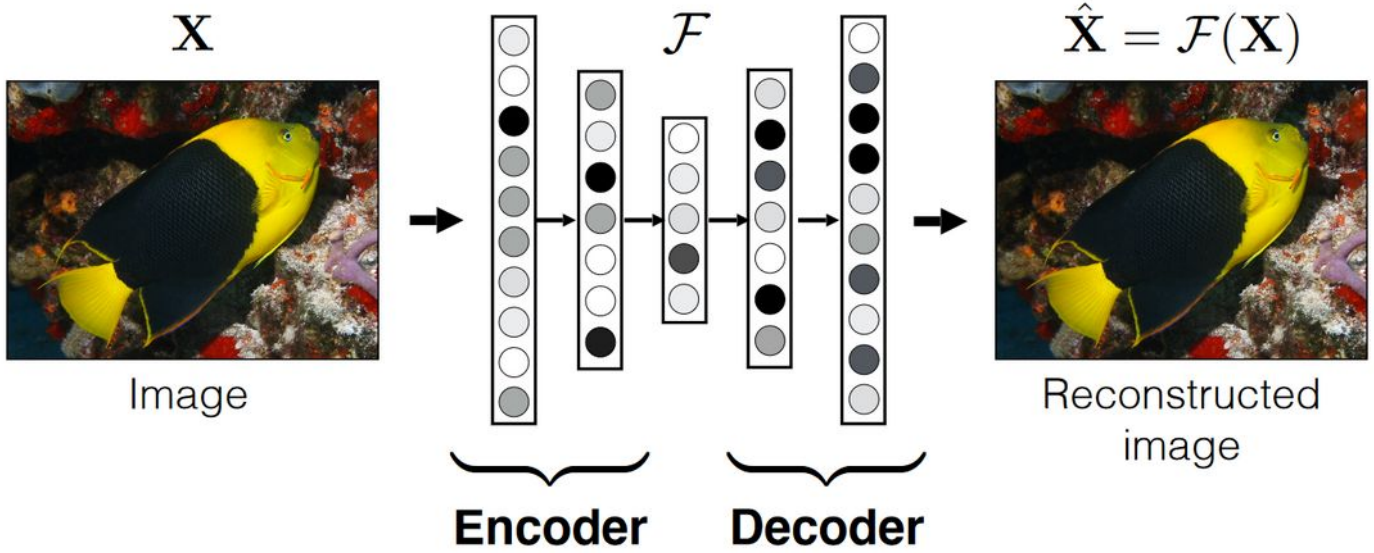
Werby, A., Huang, C., Büchner, M., Valada, A., & Burgard, W. (2024, March). Hierarchical Open-Vocabulary 3D Scene Graphs for Language-Grounded Robot Navigation. In First Workshop on Vision-Language Model for Navigation and Manipulation at ICRA 2024.

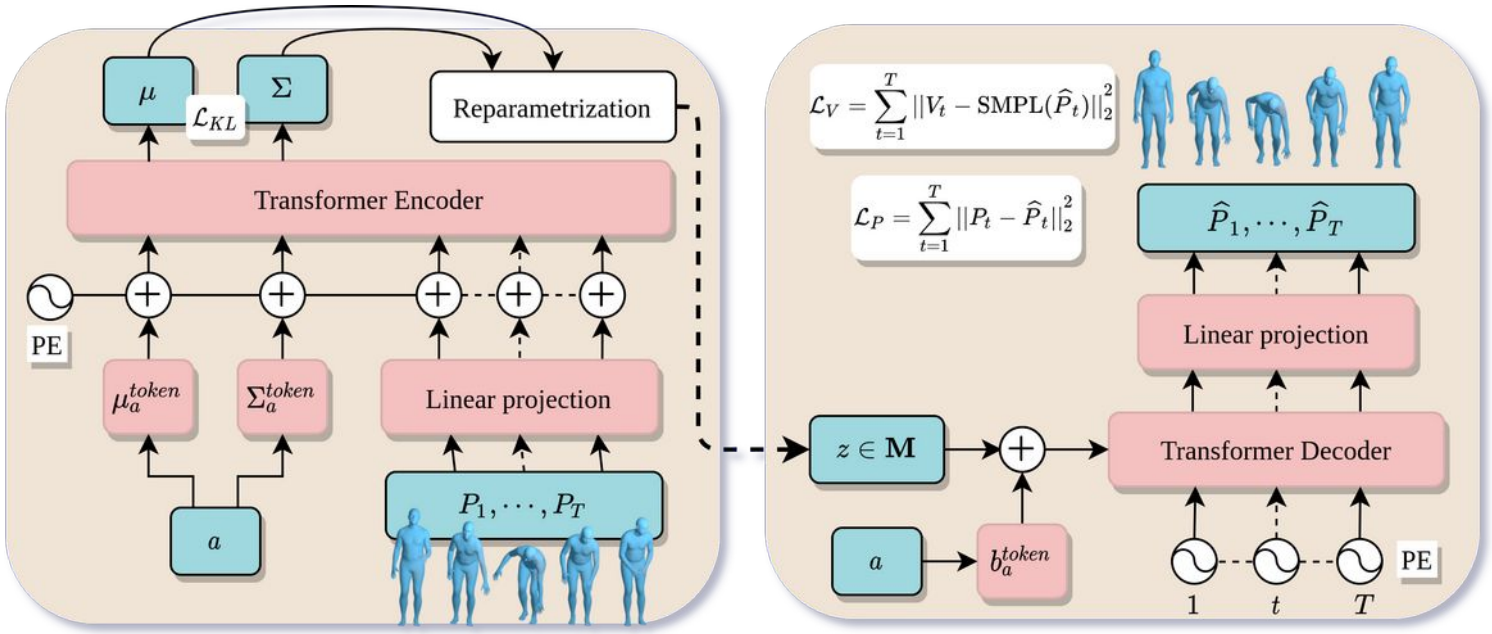
Recent advances in computer vision – transfer learning



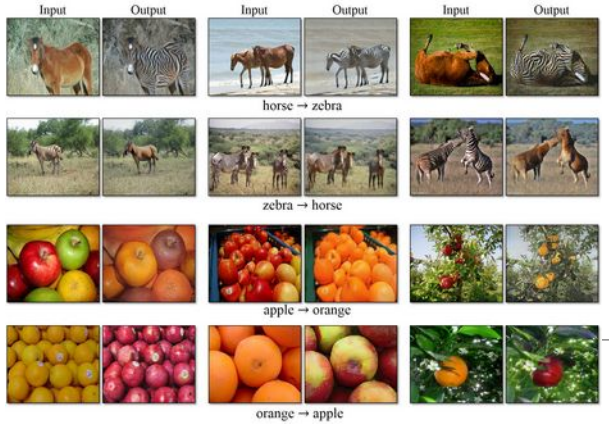
Finetuning starts with the representation learned on a previous task, and adapts it to perform well on a new task.

Recent advances in computer vision – autoencoders, GANs, VAEs

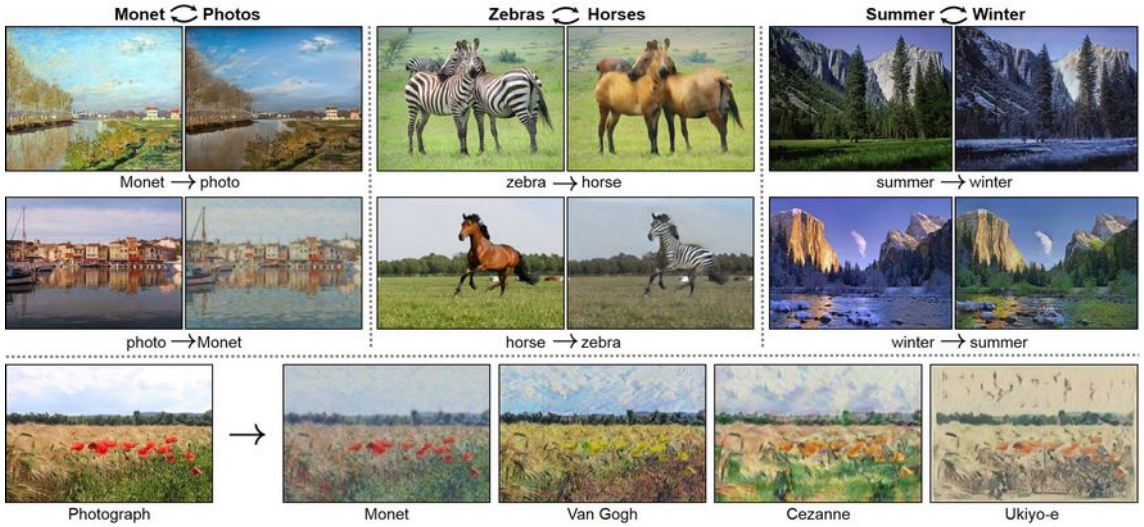




Petrovich, Mathis, Michael J. Black, and Gül Varol. "Action-conditioned 3d human motion synthesis with transformer vae." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021.

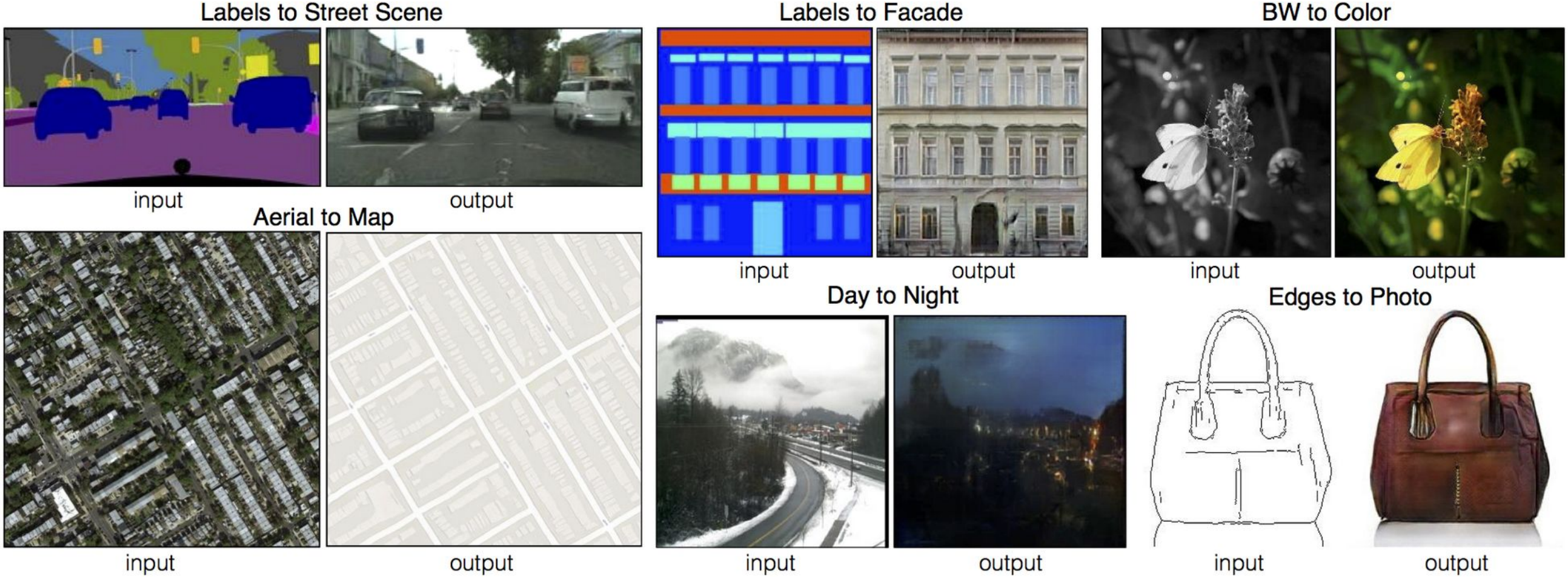


CycleGAN



Recent advances in computer vision

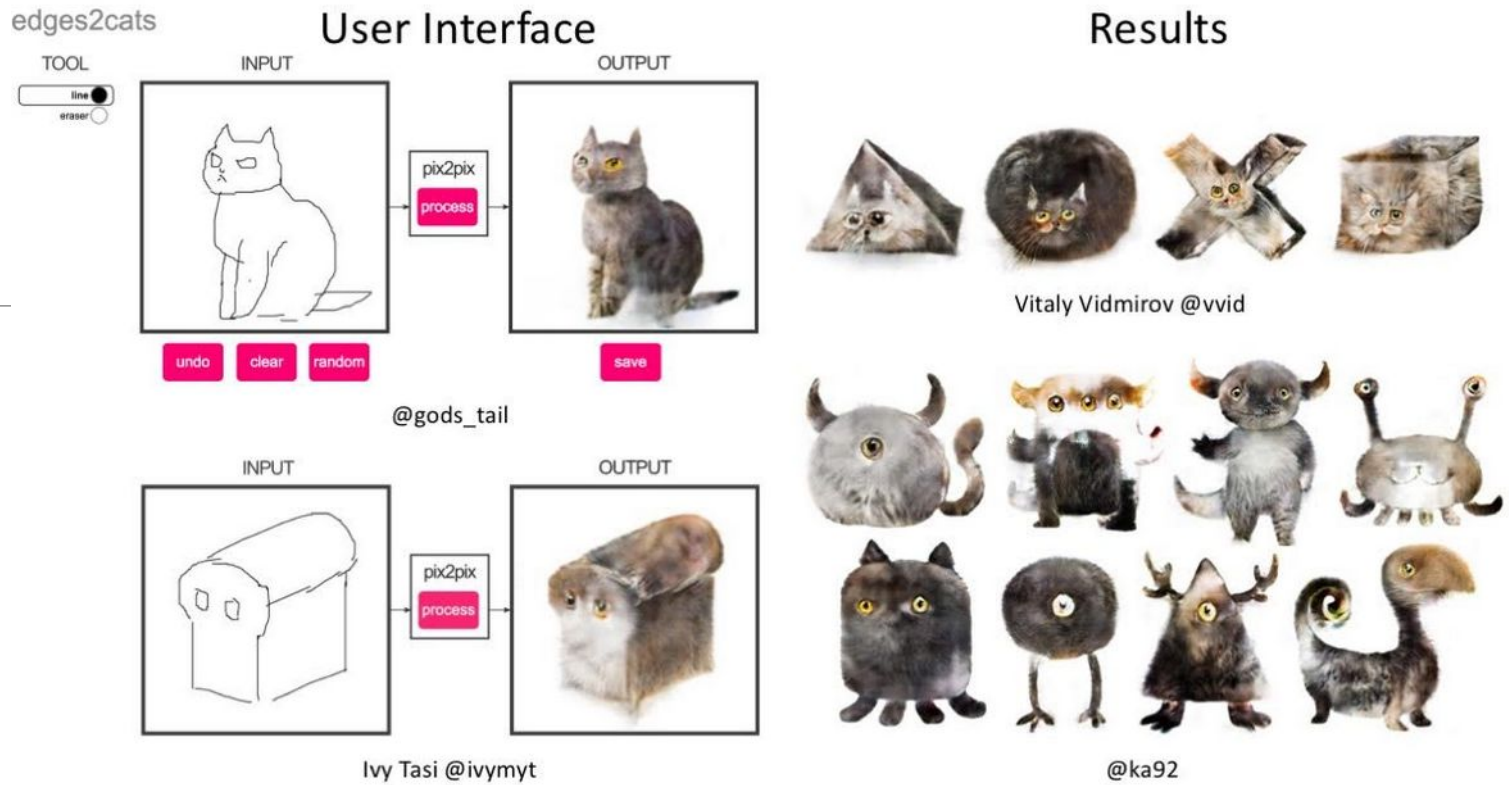
https://towardsdatascience.com/recent-advances-in-modern-computer-vision-56801edab980



Pix2Pix

Recent advances in computer vision

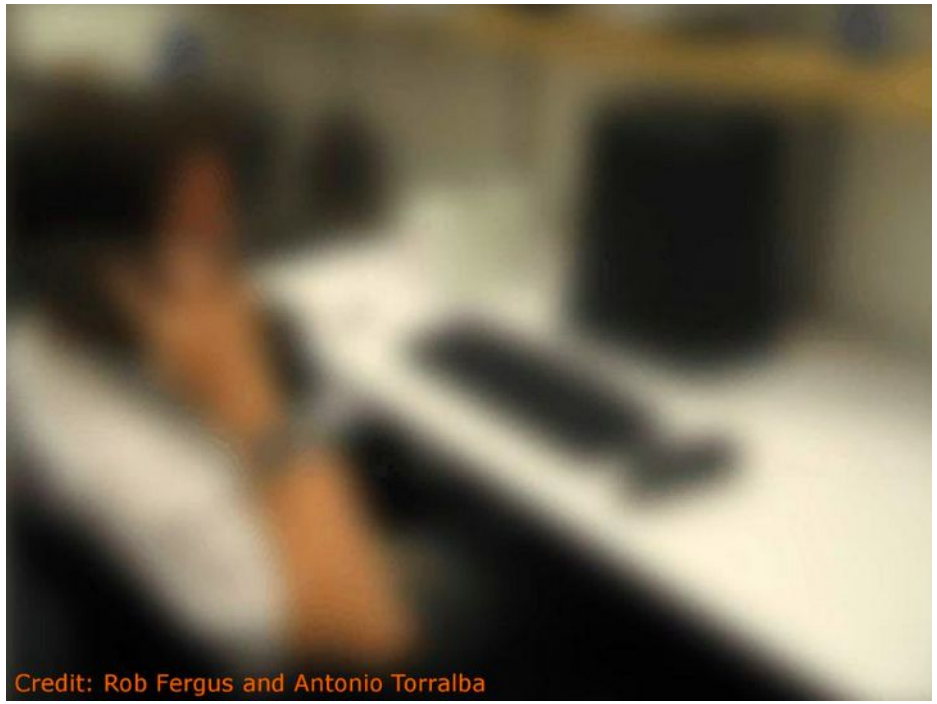
https://github.com/lood339/pytorch-two-GAN

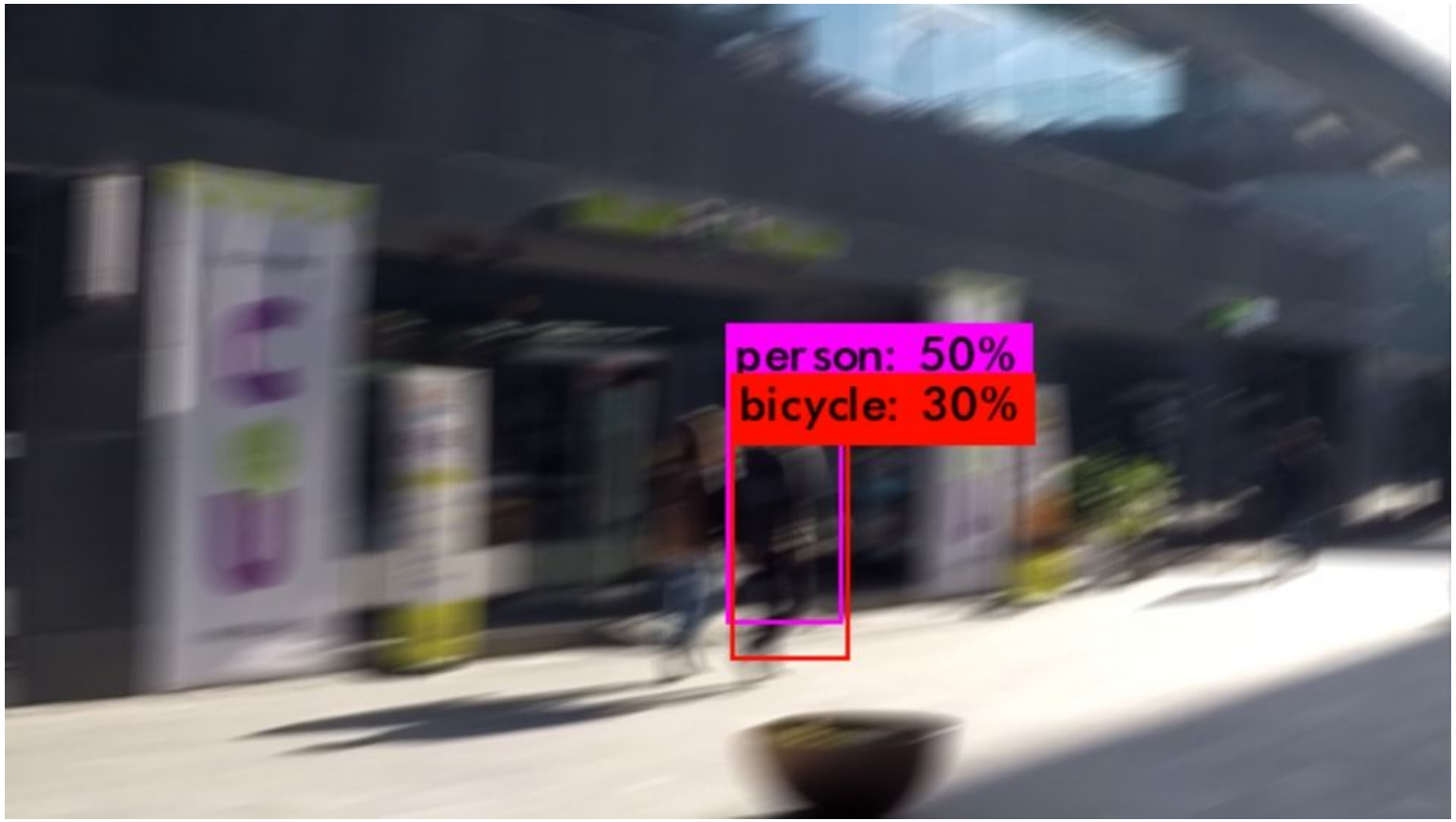


Demo
https://affinelayer.com/pixsrv/

Recent advances in computer vision

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





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

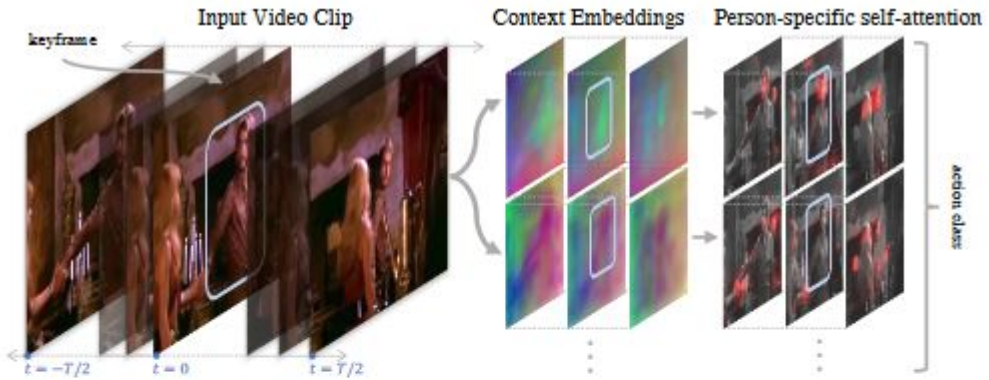
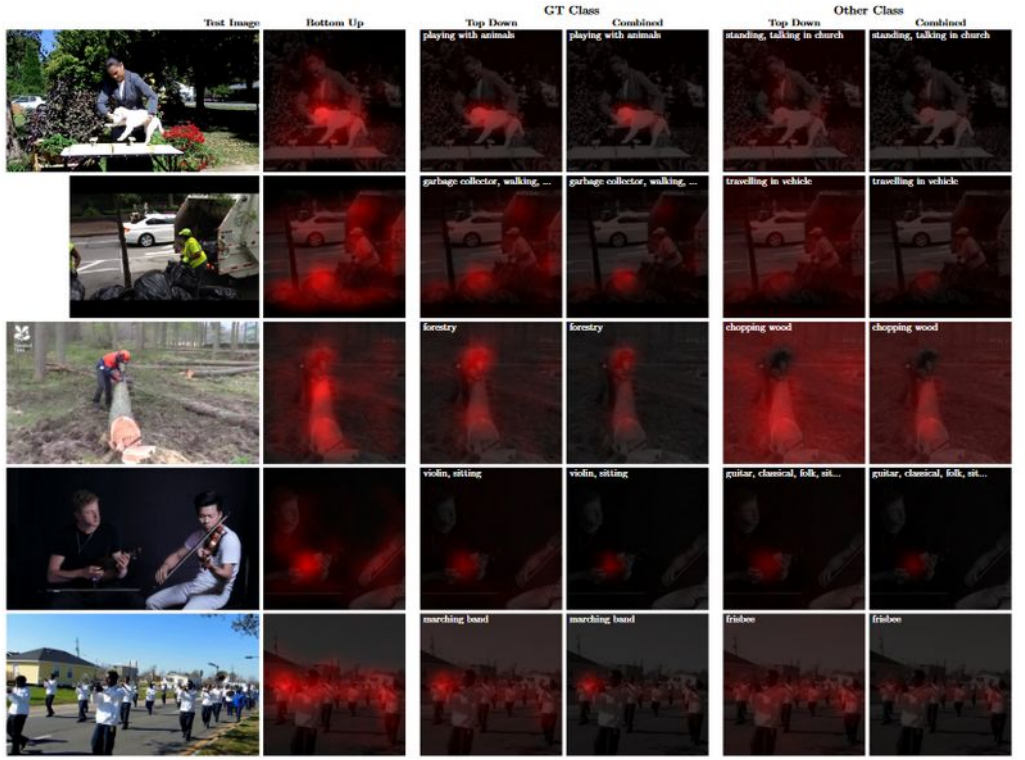
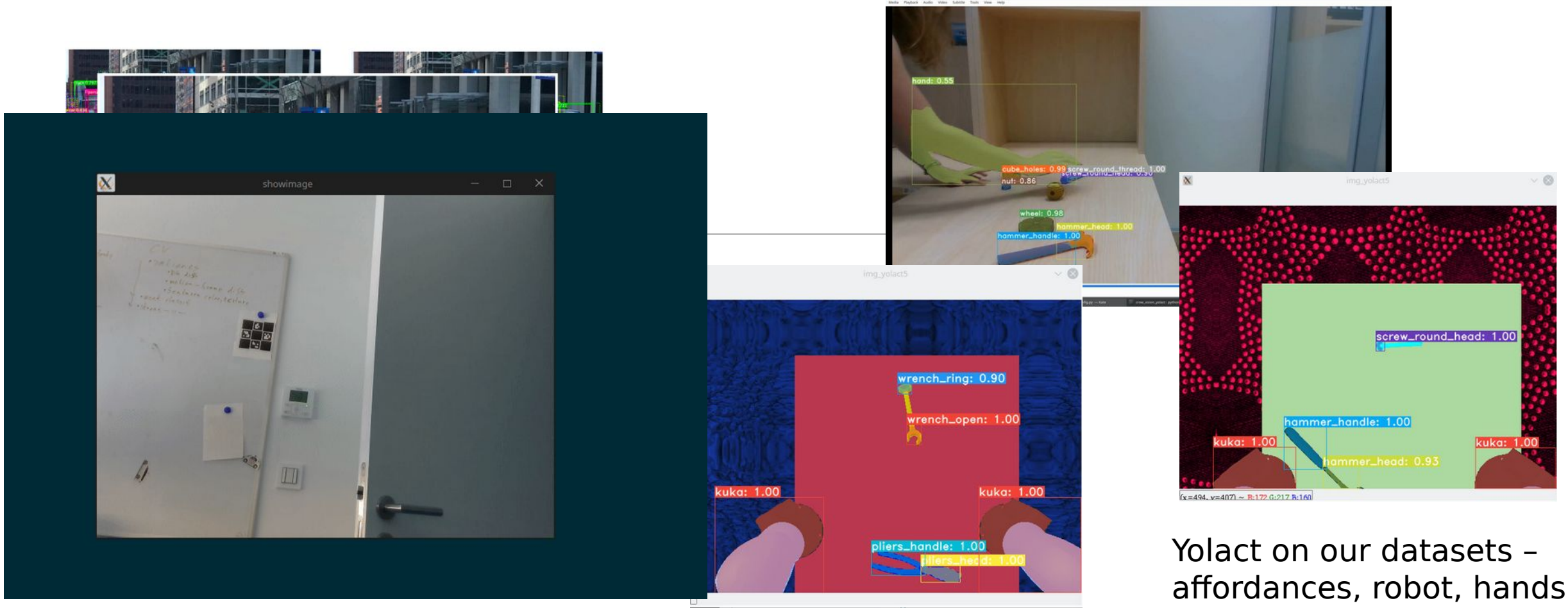


Figure 1: **Action Transformer in action.** Our proposed multi-head/layer Action Transformer architecture learns to attend to relevant regions of the person of interest and their context (other people, objects) to recognize the actions they are doing. Each head computes a clip embedding, which is used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

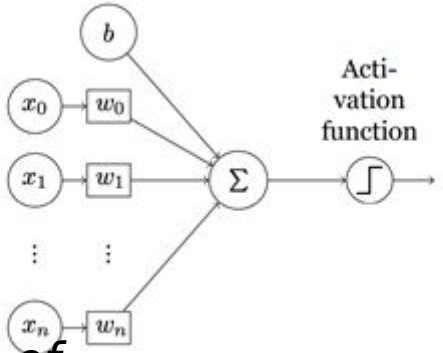
1.How can the system know what information is significant enough?

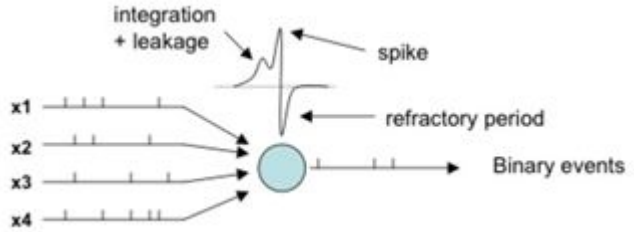
2.How does the visual system know when and how to direct attention and select significant information rather than doing so randomly?

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?

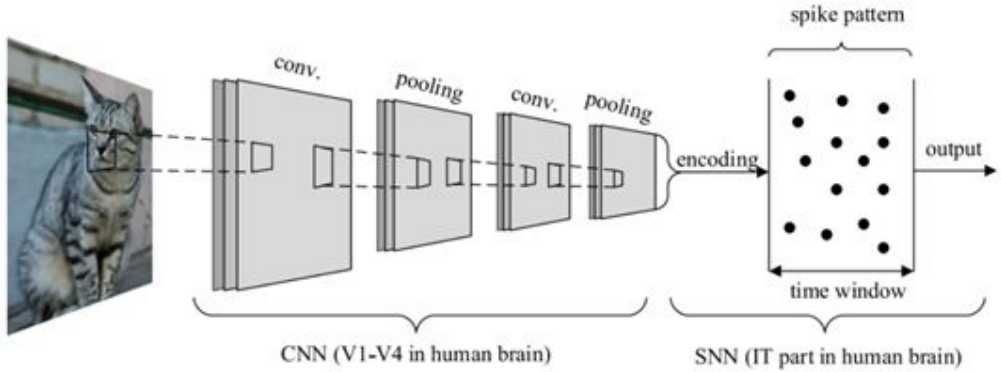
Event-based vision

Visual attention, Spiking neural net





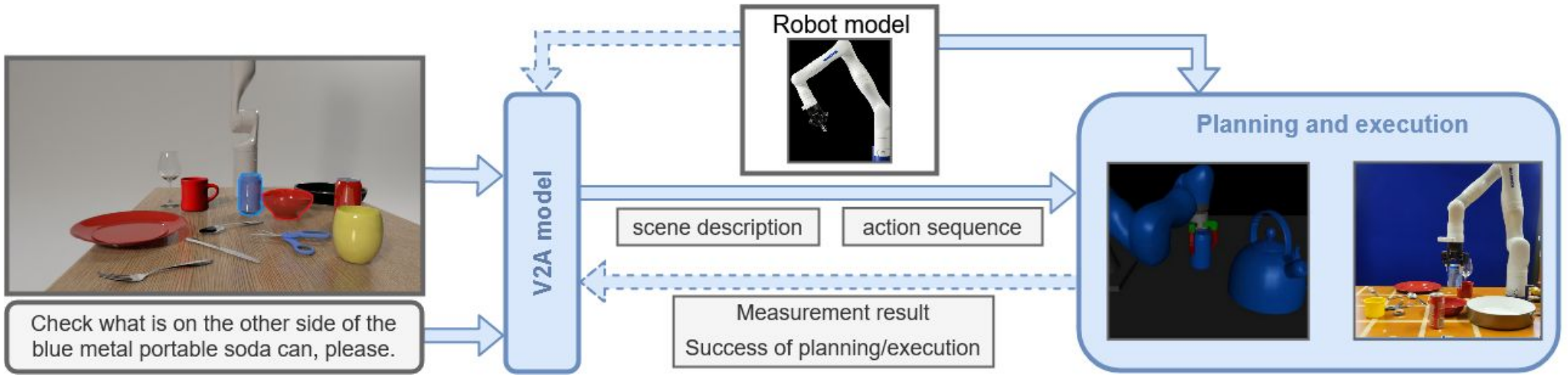
Therefore, it can be said that the core objective of visual attention is to achieve the least possible amount of visual information to be processed to solve complex high-level tasks, e.g., object recognition.



Where next?

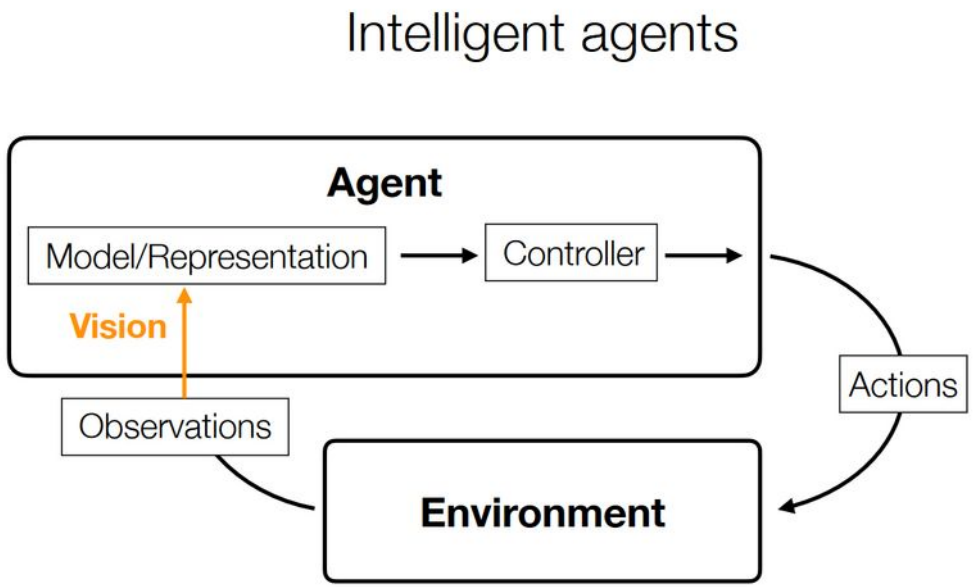
https://becominghuman.ai/from-human-vision-to-computer-vision-towards-spiked-based-visual-intelligence-and-neuromorphic-913e5de21bf9

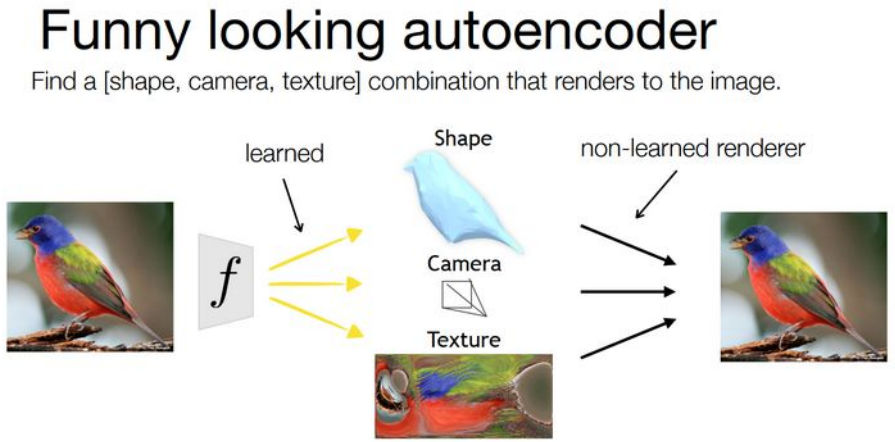
Computer vision for embodied agents



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

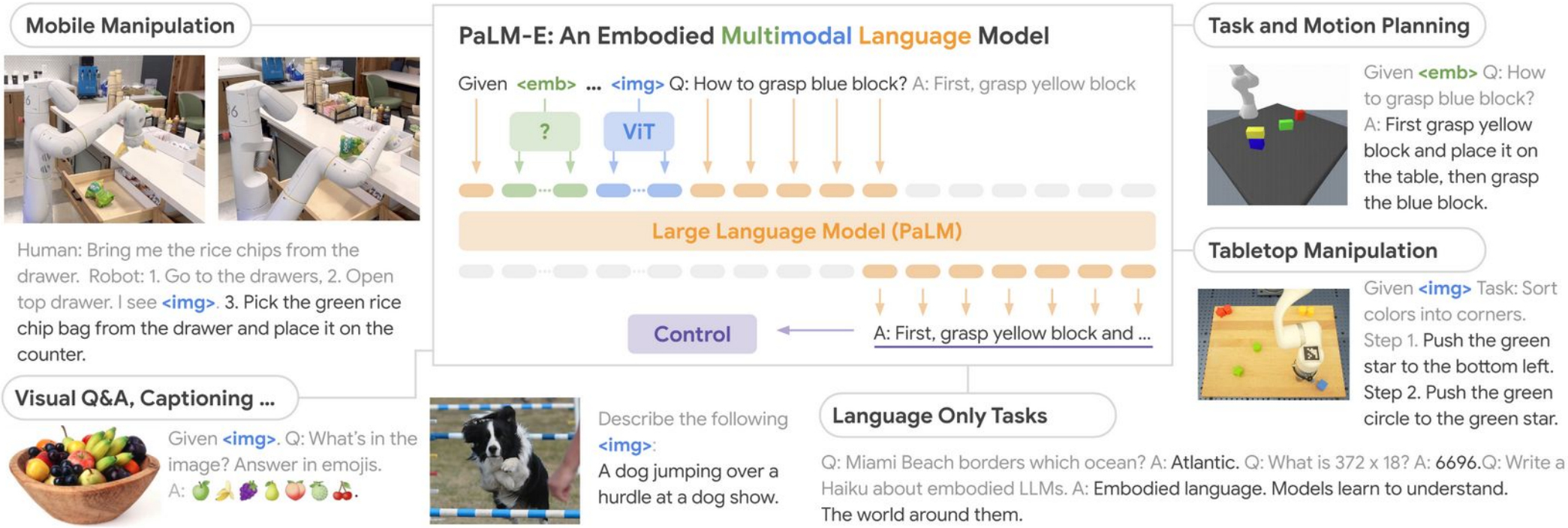
Computer vision for embodied agents



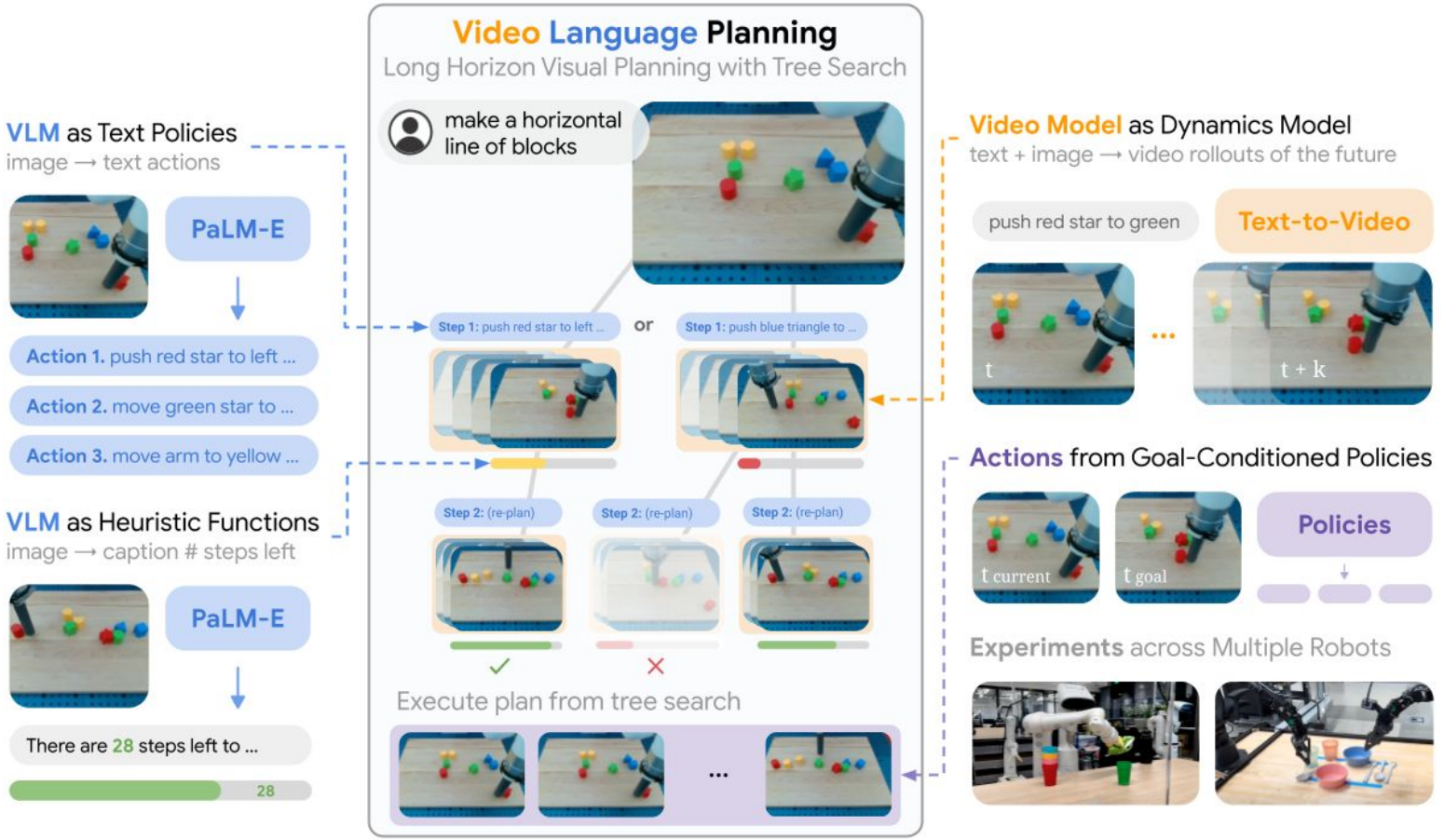


http://6.869.csail.mit.edu/sp22/lectures/L16/16_vision_for_embodied_agents.pdf

Computer vision for embodied agents



Computer vision for embodied agents



Checklist for Comparison Studies

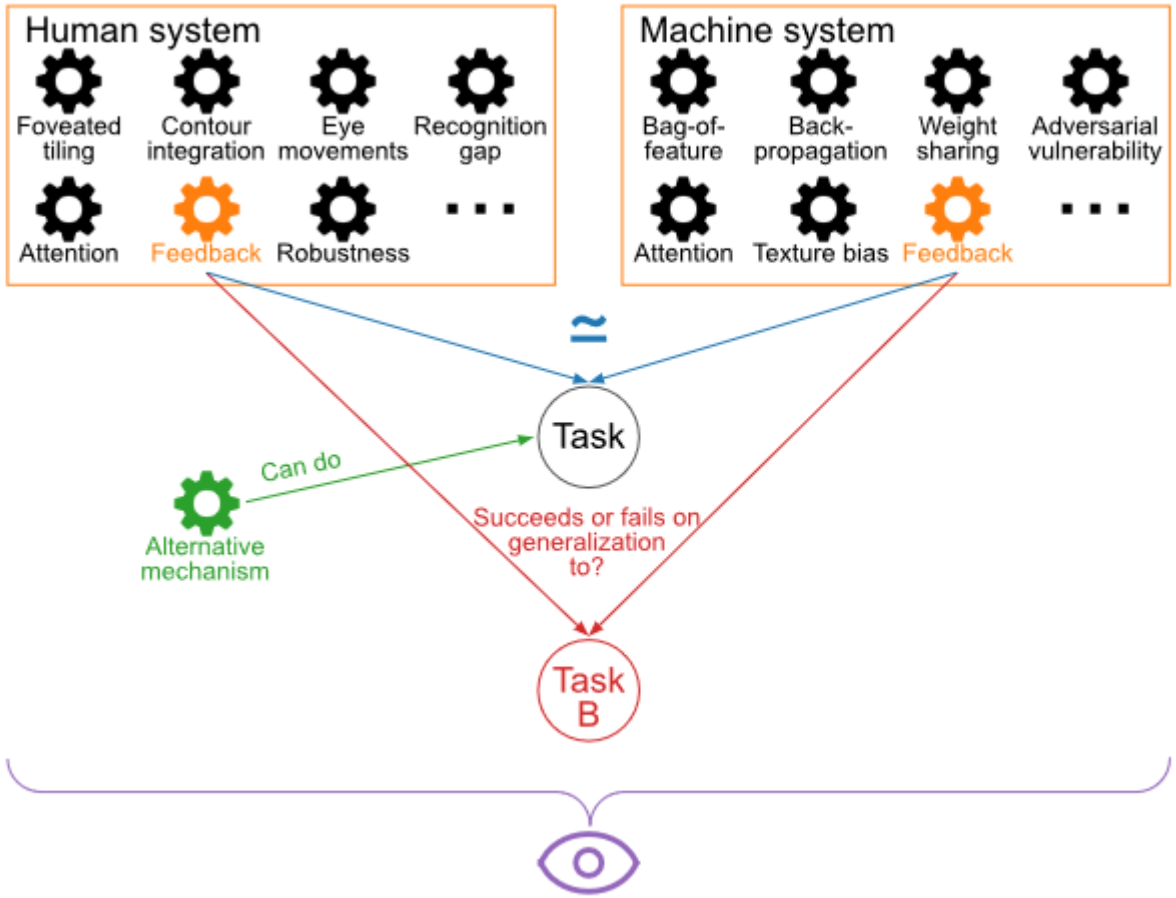
i. **Isolate implementational or functional properties**
Ex: BagNets[1] isolate local feature processing

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]

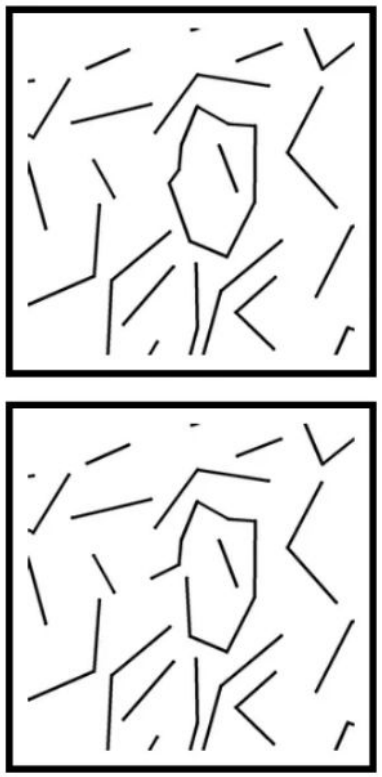
iv. **Test generalization**
Ex: Controversial stimuli[5] are useful out-of-distribution data that reveal human inductive bias

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



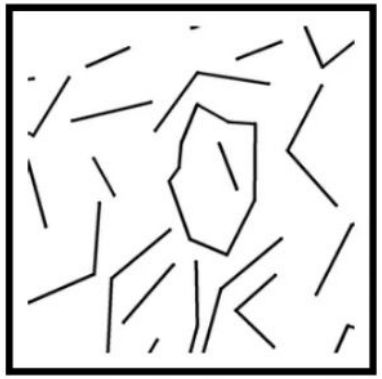
Can you tell which one of the above images contains a closed shape?

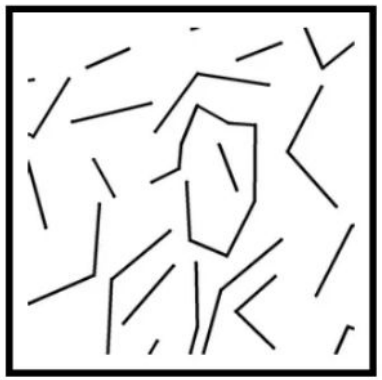
The ResNet neural network was able to detect various open- and closed-contour images, despite having been trained on straight-line examples only.

Comparing human and AI vision

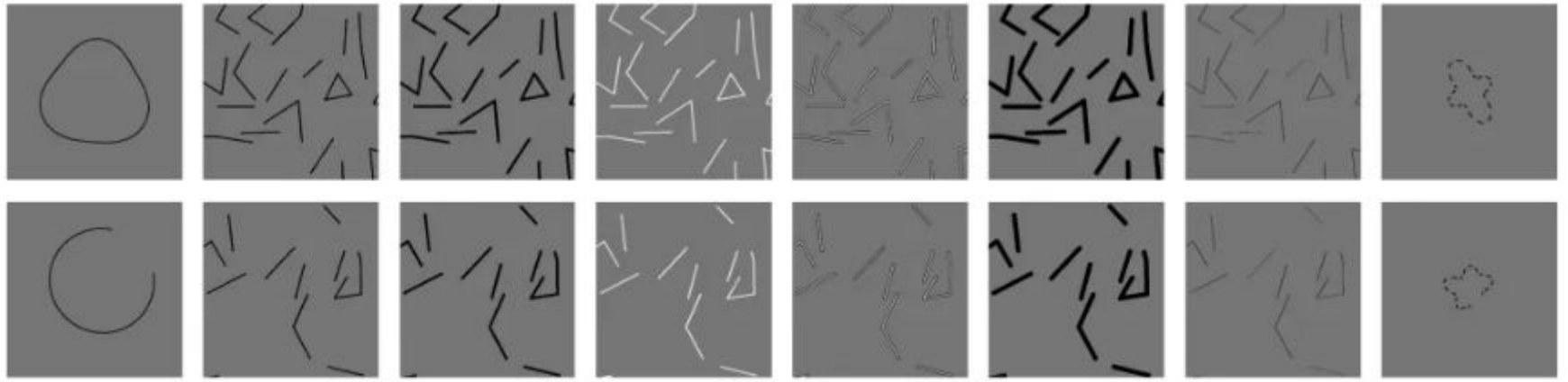
https://bdtechtalks.com/2020/08/10/computer-vision-deep-learning-vs-human-perception/

https://arxiv.org/abs/2004.09406



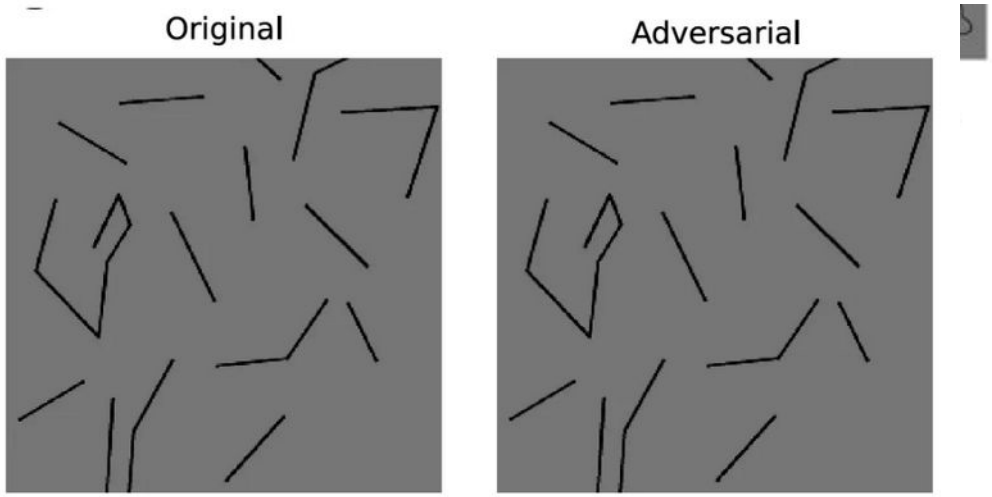


Can you tell which one of the above images contains a closed shape?



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.

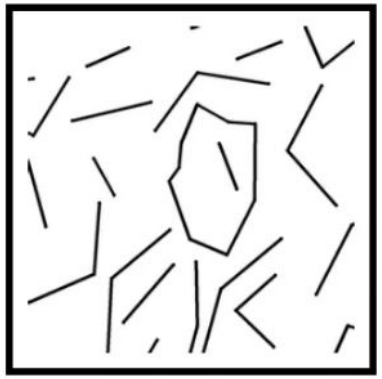


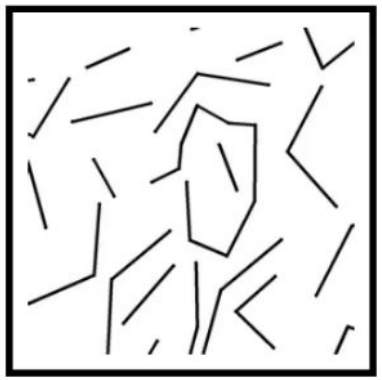
The image on the right has been modified with adversarial perturbations, noise that is imperceptible by humans. To the human eye, both images are identical. But to a neural network, they are different images.

Comparing human anc

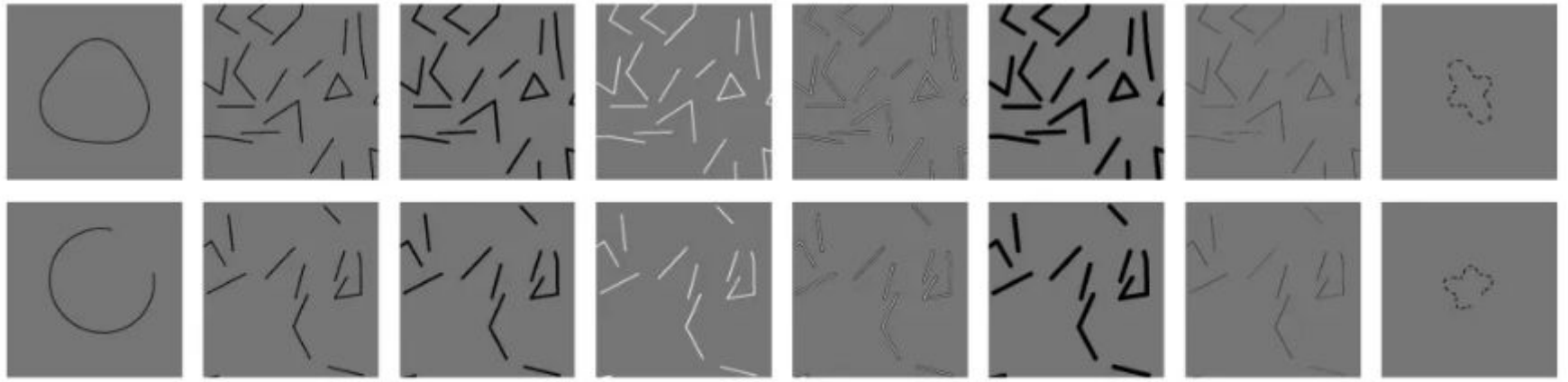
https://bdtechtalks.com/2020/08/10/computer-vision-deep-learning-vs-human-perception/

https://arxiv.org/abs/2004.09406





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

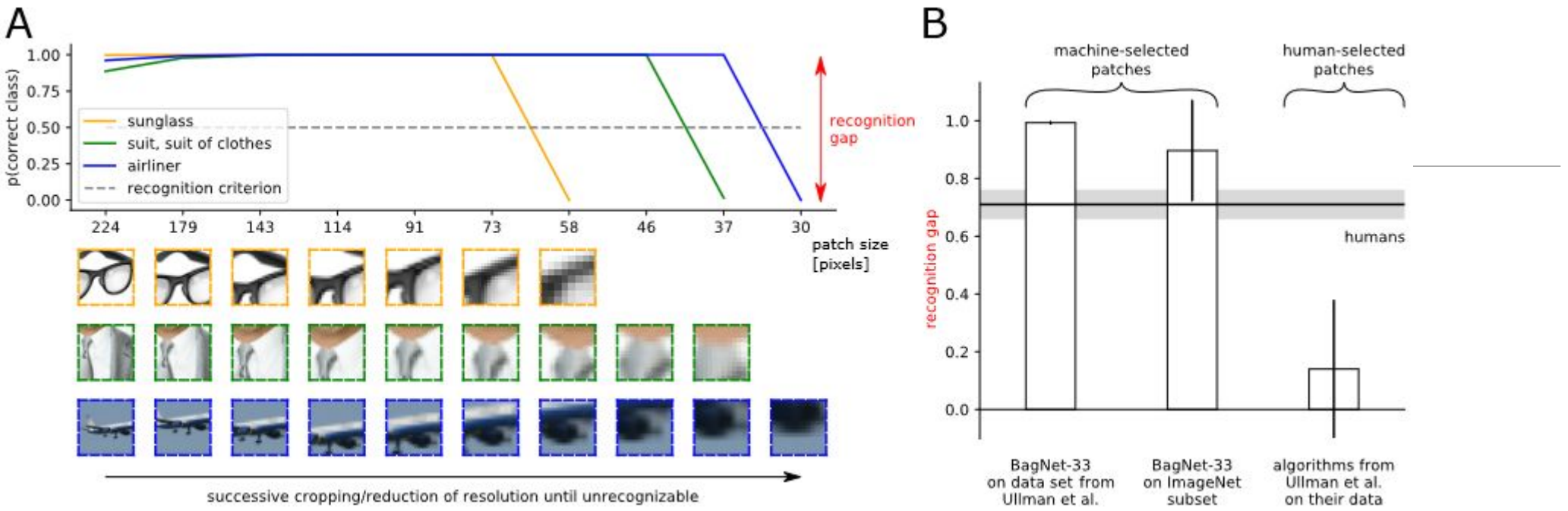


The ResNet-50 neural network struggled when presented with images that contained lines of different colors and thickness, and where shapes were larger than the training set.

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

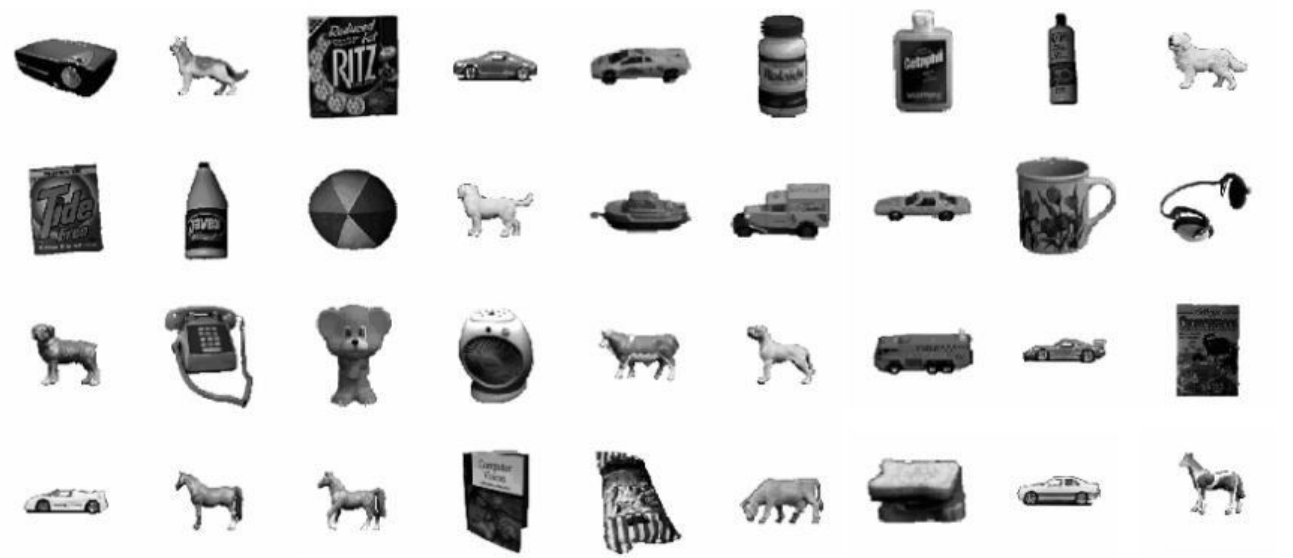
https://bdtechtalks.com/2020/08/10/computer-vision-deep-learning-vs-human-perception/ https://arxiv.org/abs/2004.09406



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, et
- Traffic cameras
- Quality assurance – infrared thermoca
- Same pictures as trained
- …



- 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

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

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.)

Gestalt psychologie

Gestalt = uspořádání/forma

- Nesouhlasná reakce na **strukturalismus** (Wundt, Titchener)
- Vjem není pouze agregát elementů
- **3D zkušenost z 2D obrazů** – organizace počitků do stabilních vzorů (I přes změny vstupních informací)

vjem jako stavba poskládaná z elementárních stavebních prvků

- Identifikovali principy perceptuální organizace
  → ilustrovali viz.iluzemi, percept.konstantami

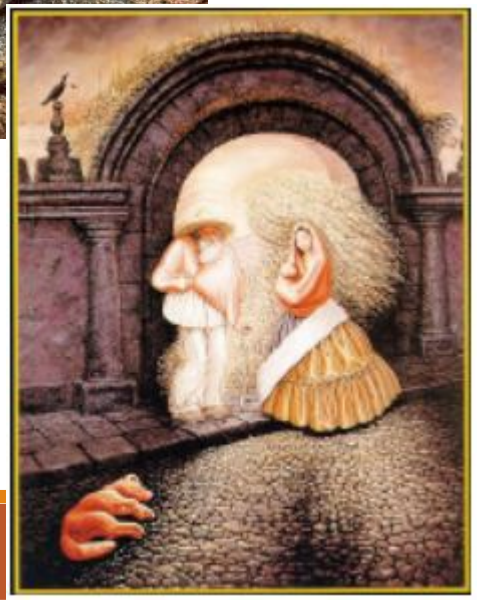
Rozlišení pole na figuru a pozadí

Asociace částí scény → obraz (zbytek pozadí)

Čeho si v zorném poli všimneme ---

--- Vlastnosti figury

Strukturování zorného pole občas odvádí od pravdy - kamufláž, reversibilní, ilusorní kontury, ...



“Our eyes are accustomed to fixing on specific objects. The moment this happens everything around is reduced to background.” Maurits Escher

Co má větší šanci se stát figurou?

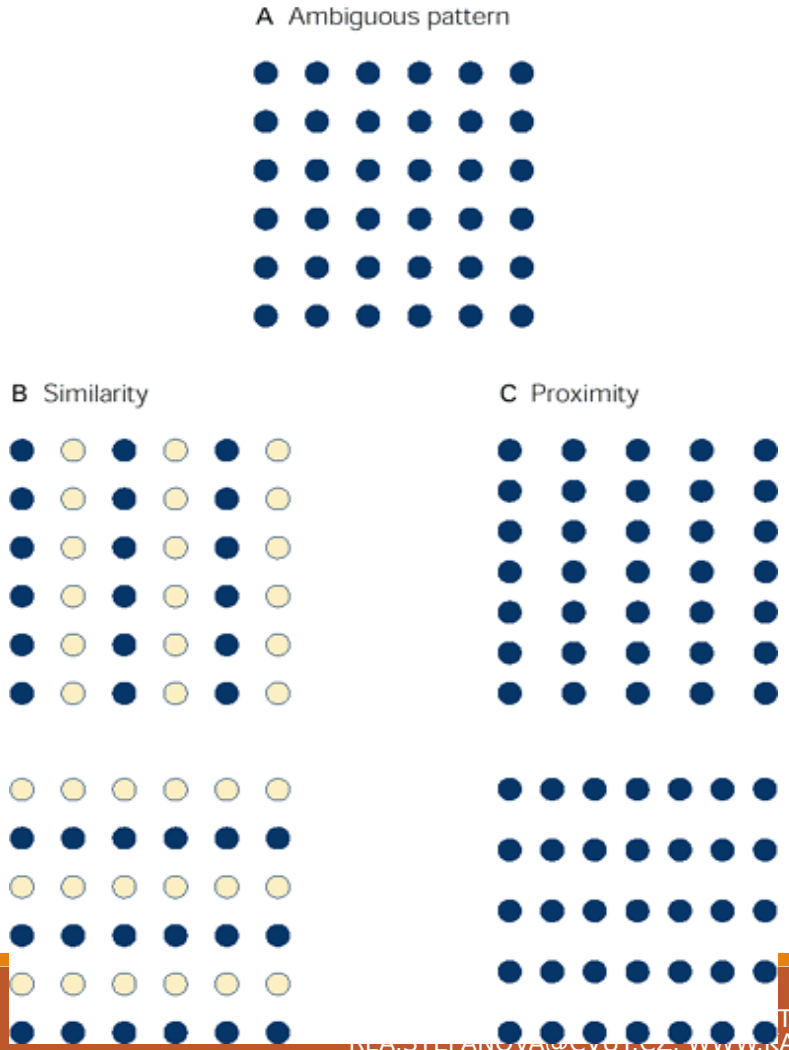
- Figura
  - blízká
  - ohraničená
  - detailní
  - určitá
  - nasycená
  - uzavřená
  - symetrická
- Pozadí
  - vzdálené
  - spojité
  - povšechné
  - neostré
  - splývající
  - uzavírající
  - bez symetrie

Percepční rozdíly mezi figurou a pozadím

Ostrost kontur, přináleží k figuře, sytější barvy

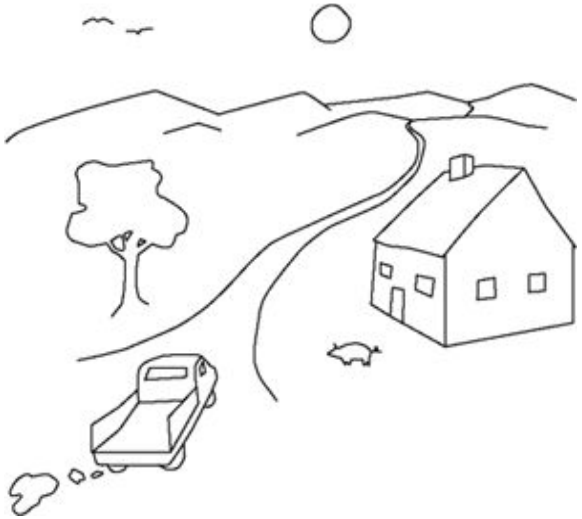
Pattern recognition

Tendence mozku nalézt ve vstupních datech vzor

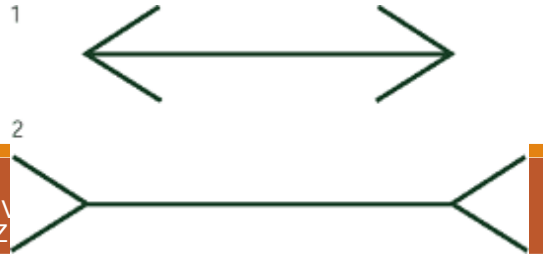


t.šablon X t.rysů
(globální vs. lokální vyhledávání – M a P b.)

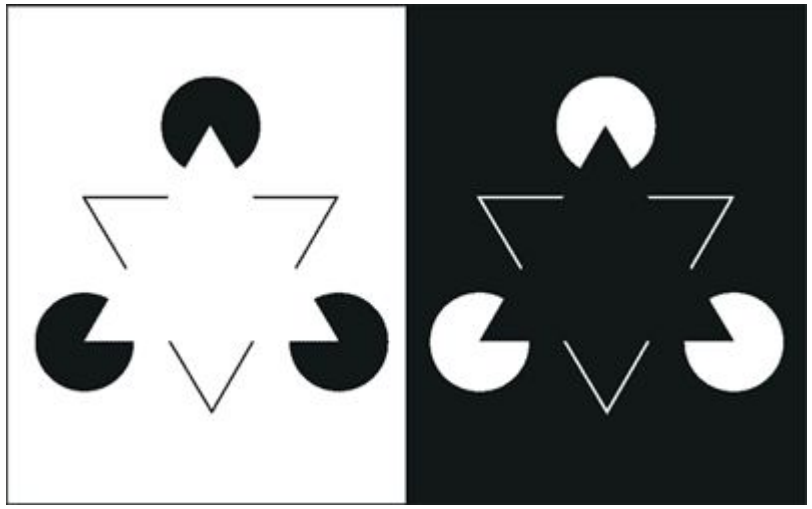
Kontury – nápovědi pro hrany objektů



Tvar jako indikátor velikosti

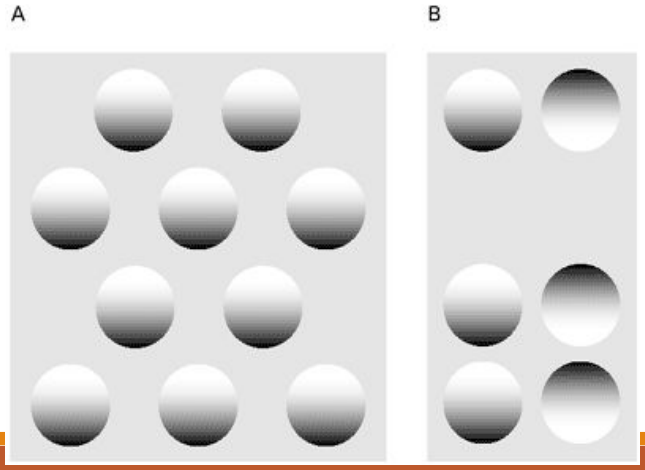


Vyplňování



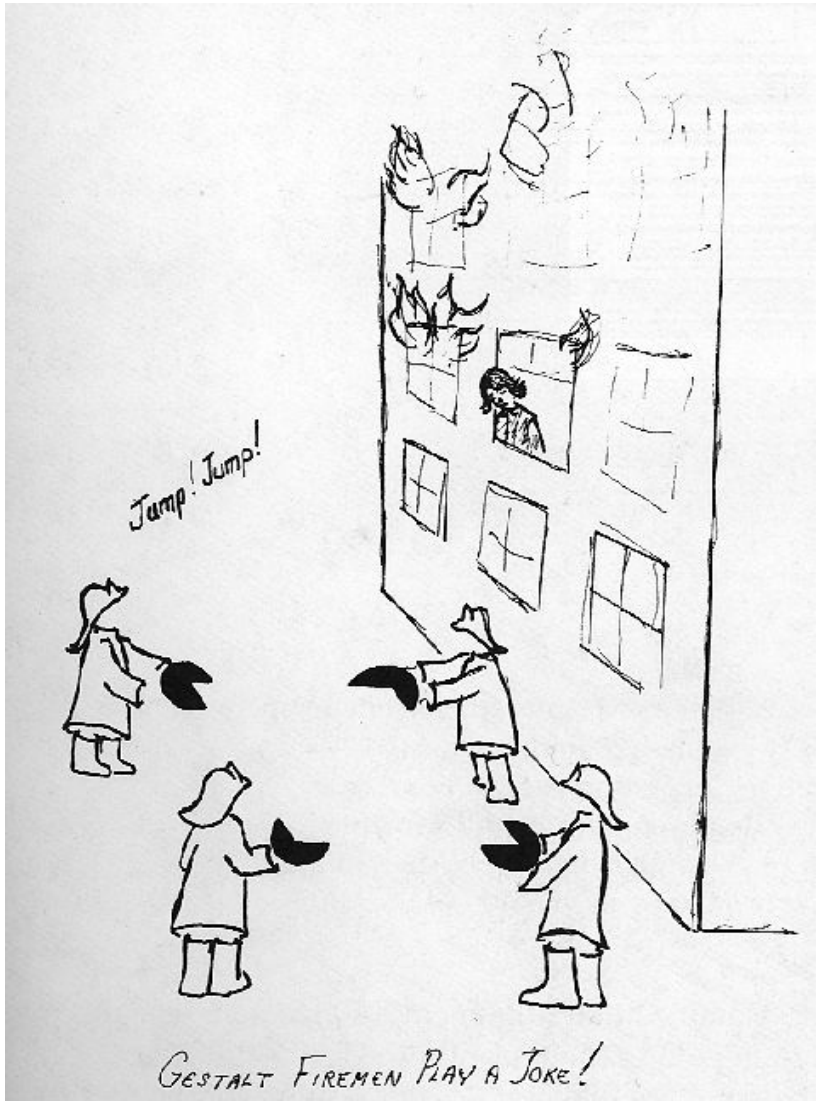
- Z neúplných kontur vidíme jasně objekt
- Neexistující kontury
- Bližší zakrývají vzdálenější
- Vzor v jinak nevztažených tvarech, pokud tvary viděny jako části překrytého objektu

- Jen 1 zdroj světla



Velikost objektů

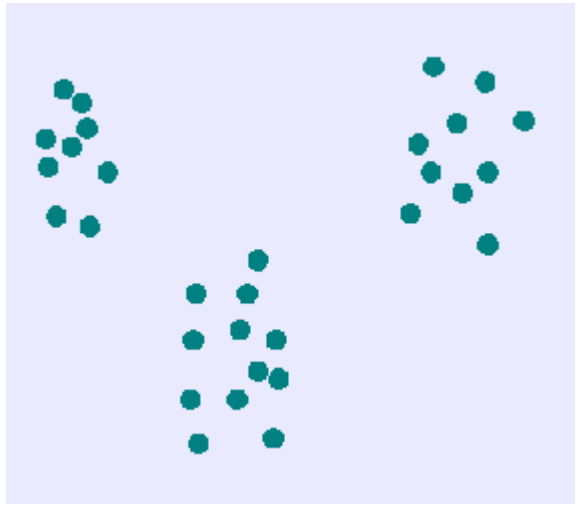
Velikost objektu vnímáme dle ostatních objektů ve scéně

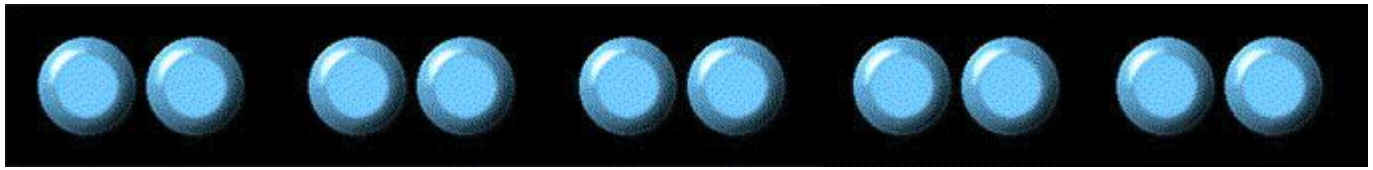

Jump! Jump!
Gestalt Firemen Play a Joke!

Zákony organisace

Blízkost

Blízké předměty/elementy máme sklon si při vnímání shlukovat.

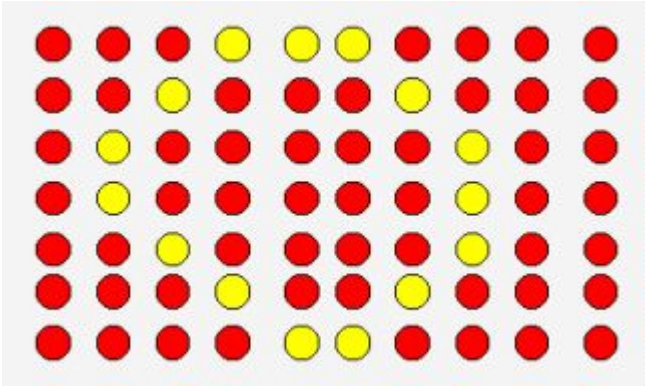


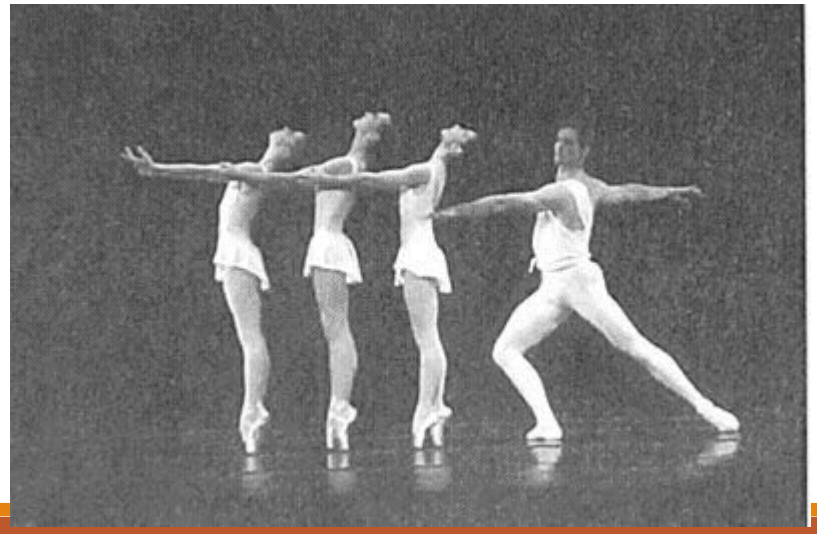


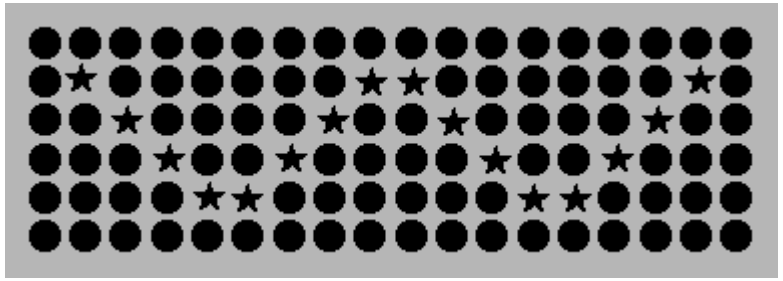
Zákony organisace

P o d o b n o s t

Podobné předměty/elementy
máme sklon si spojovat



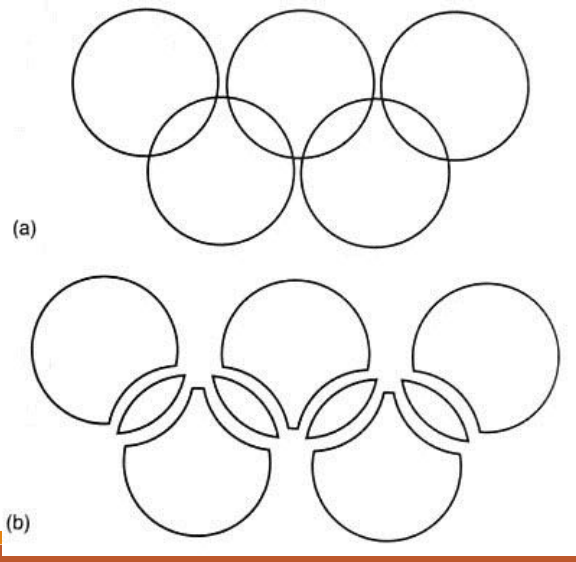




Zákony organisace

V h o d n é
p o k r a č o v á n í

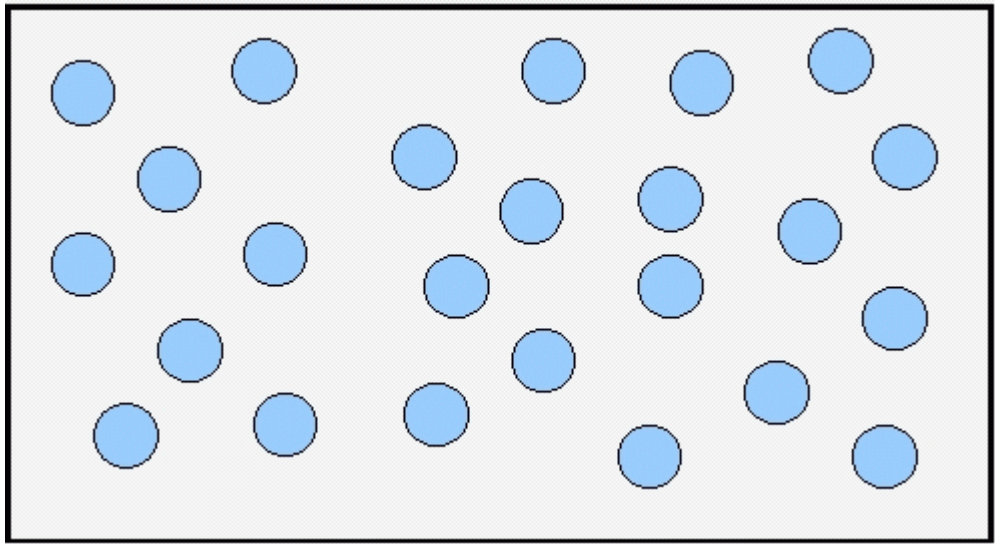
Předměty/elementy navazující na předešlý trend máme sklon sdružovat



Zákony organisace

S p o l e č n ý

o s u d

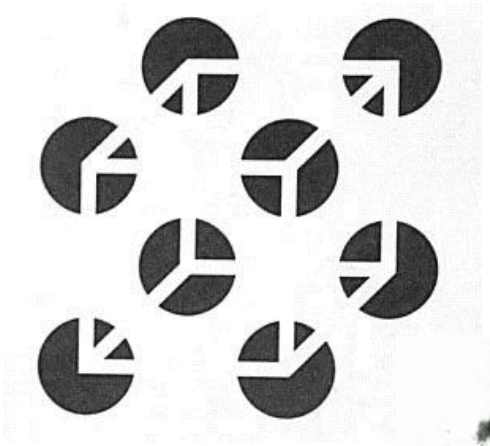


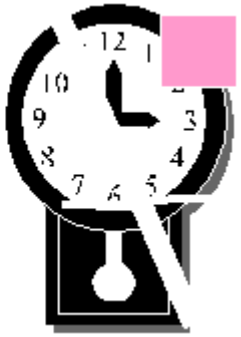
Předměty/elementy se společným pohybem (směrem, rychlostí) máme sklon si spojovat.

Zákony organisace

U z a v ř e n í

Předměty/elementy vytyčené ne do posledního detailu máme sklon ucelovat.





Shrnutí

- Spíš než zákony to jsou **inklinace, tendence**
- Všechny zákony odráží obecnou lidskou tendenci k **preferování jednoduššího, pravidelnějšího, předvídatelnějšího, stabilnějšího** … prostě lepšího tvaru = Prägnanz
- Nicméně **nevylučují lidskou flexibilitu** a schopnost adaptovat se i na nečekané

Gestalt psychologie

- Gestalt “zákony” nepopisují proces vnímání
- Jsou to deskripce percepčního dění
- Vypovídají o lidské tendenci, upřednostňování pravidelností světa
- V mnoha situacích nelze specifikovat, co je lepší nebo pravidelnější
- Lidská schopnost vypořádat se s nepravidelnostmi a jinými nedostatky sledované scenérie
- Novější přístupy jsou analytičtější

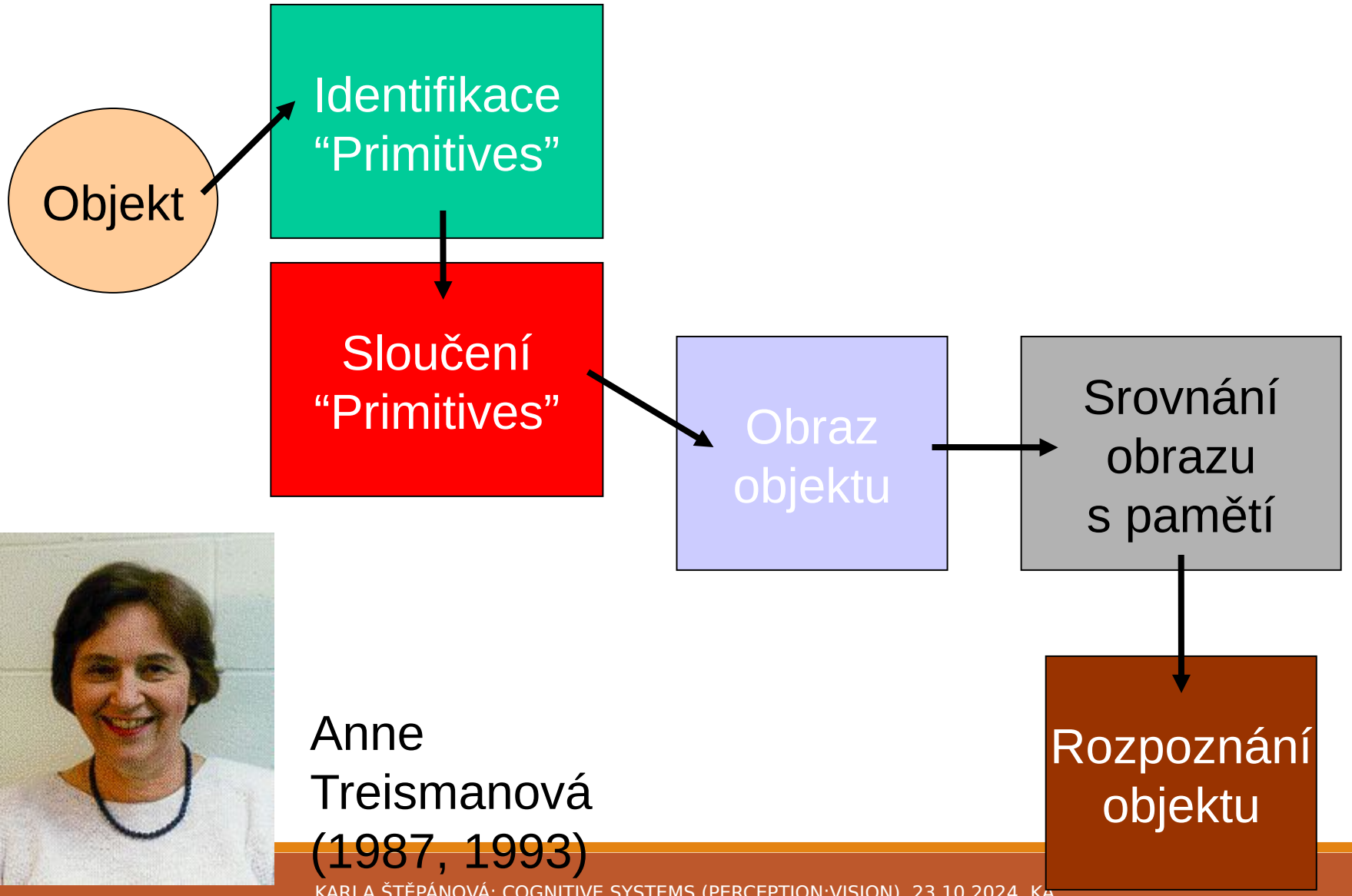
Rozpoznávání objektů

Teorie integrace vlastností (Treismanová)

Výpočetní model (Marr)

Rozpoznávání prostřednictvím komponent (Biederman)

Teorie integrace vlastností



Anne Treismanová (1987, 1993)

Teorie integrace vlastností

Preatentivní stadium

- Dekomposice obrazu.
- Detekce základních vlastností (“primitives”).
- Všímání si nestejností.

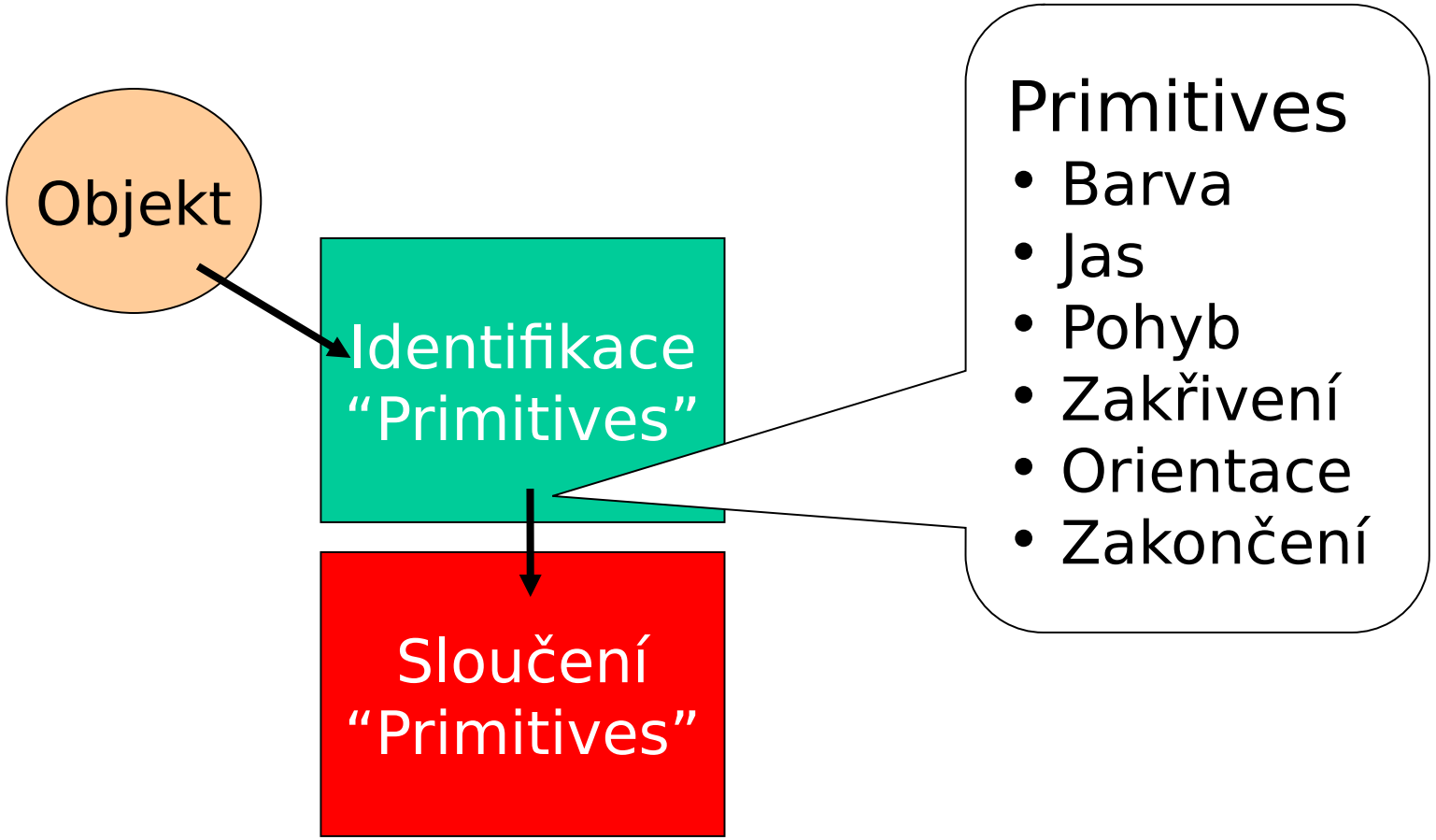
Zaměřené stadium

- Seskládání a společné zpracování informací z jednotlivých kanálů.

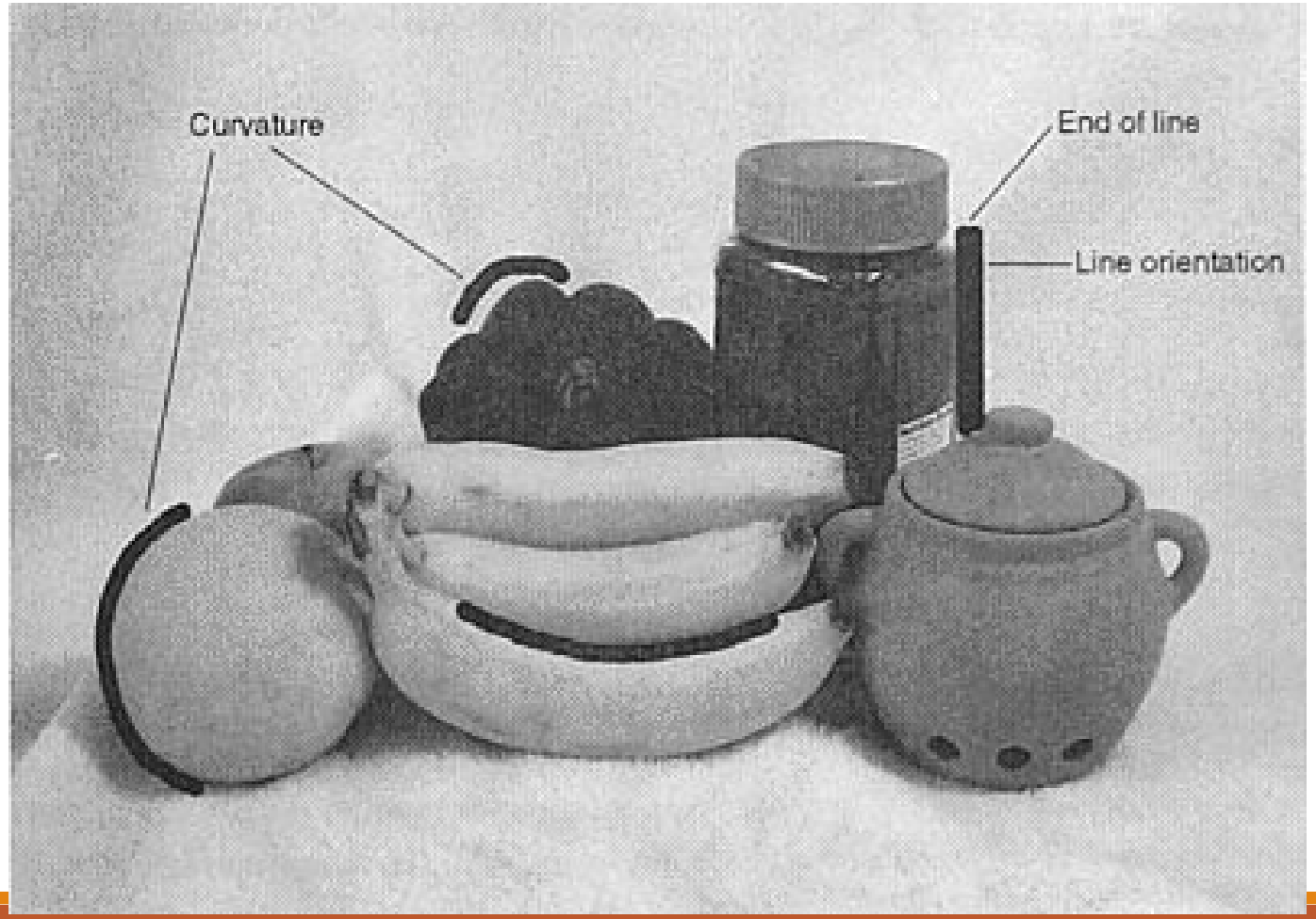
Paměť

- Konfrontace vzniklého obrazu s obrazy uloženými v paměti.

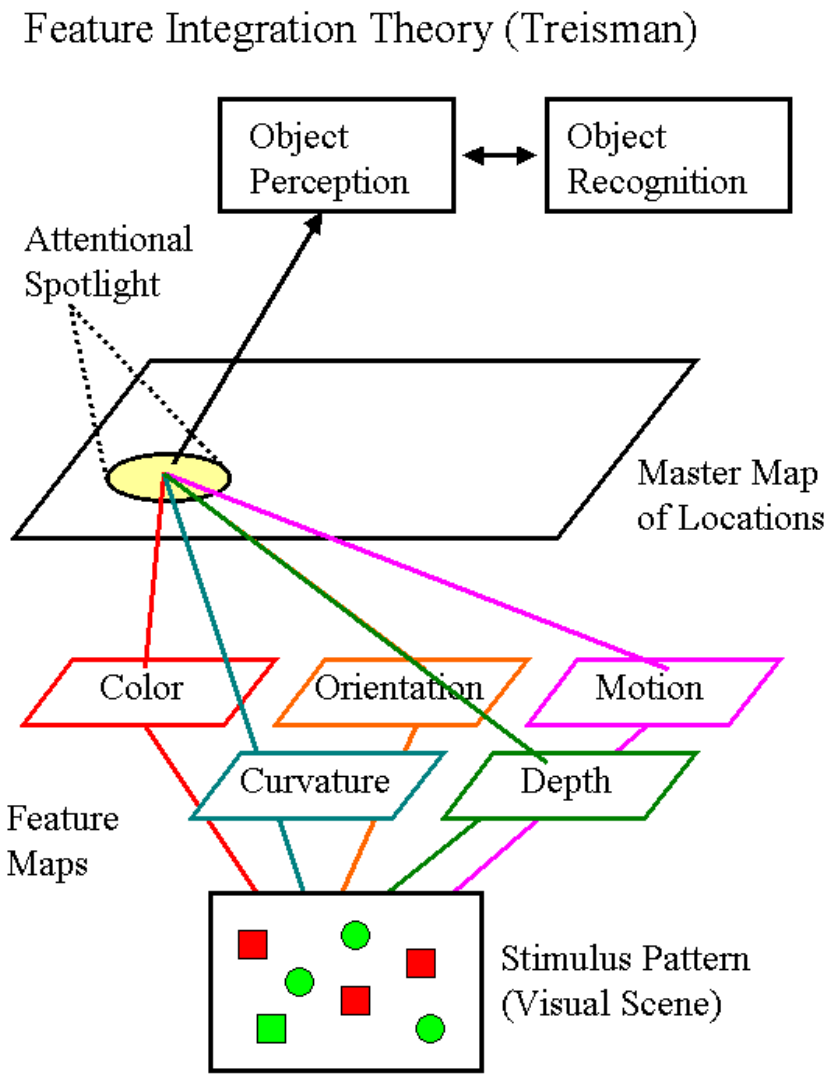
Teorie integrace vlastností



Teorie integrace vlastností



Teorie integrace vlastností



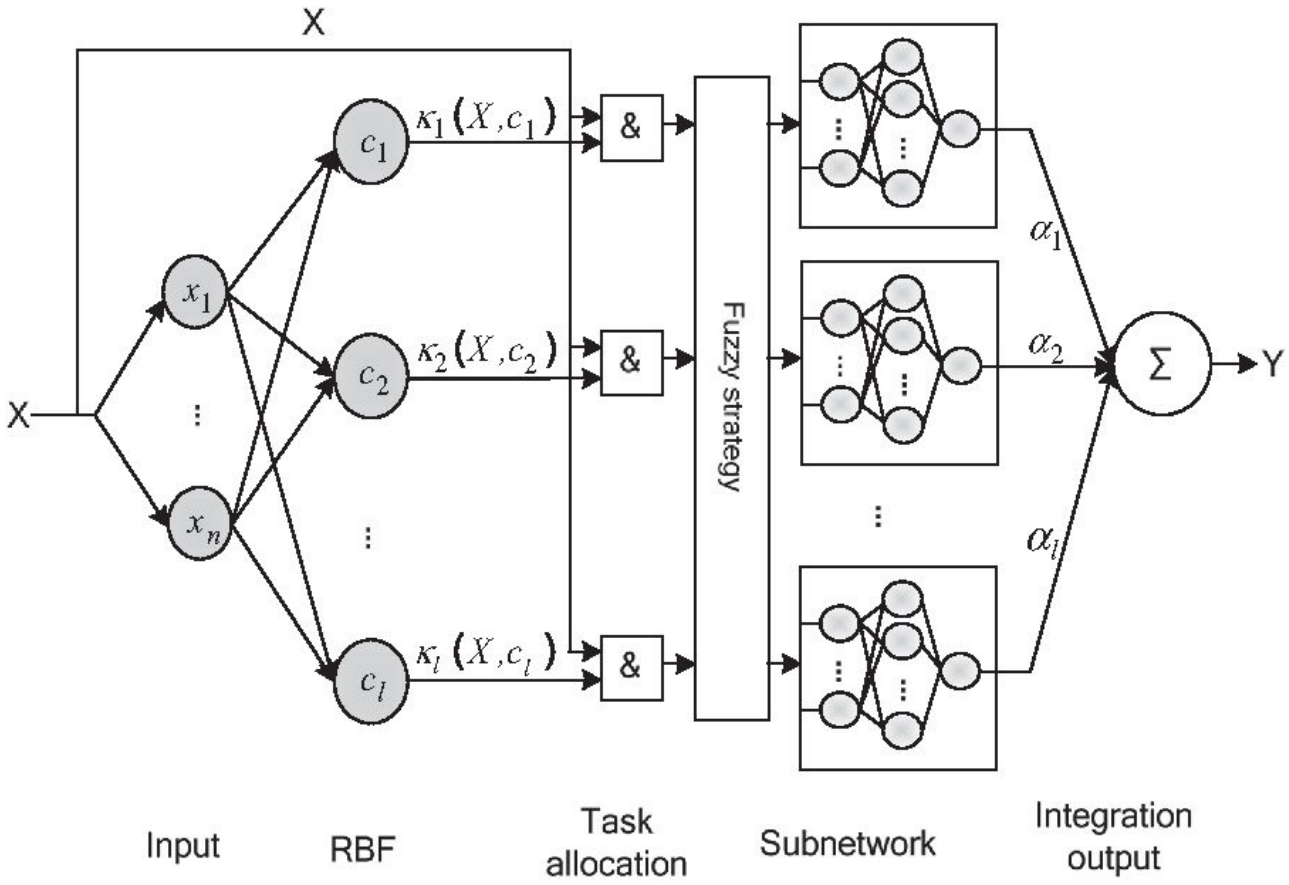
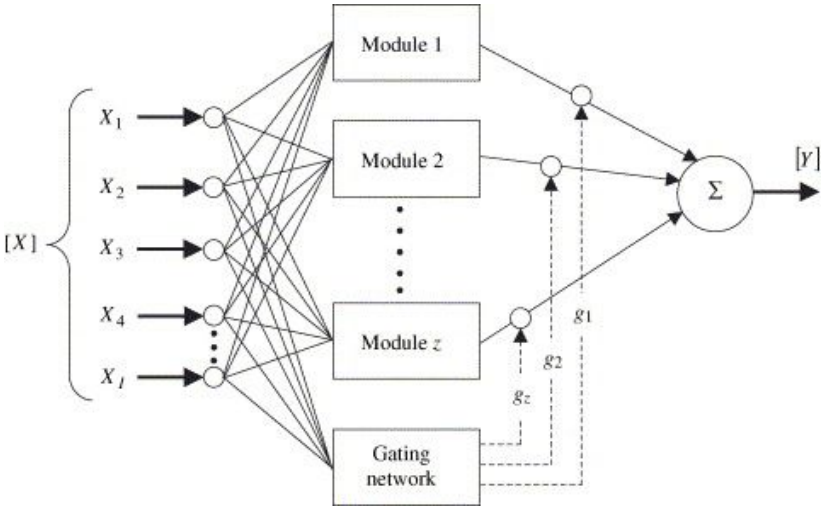


Fig. 1. Structure of OSAMNN



https://vtechworks.lib.vt.edu/bitstream/handle/10919/27998/etd.pdf?sequence=1&isAllowed=y

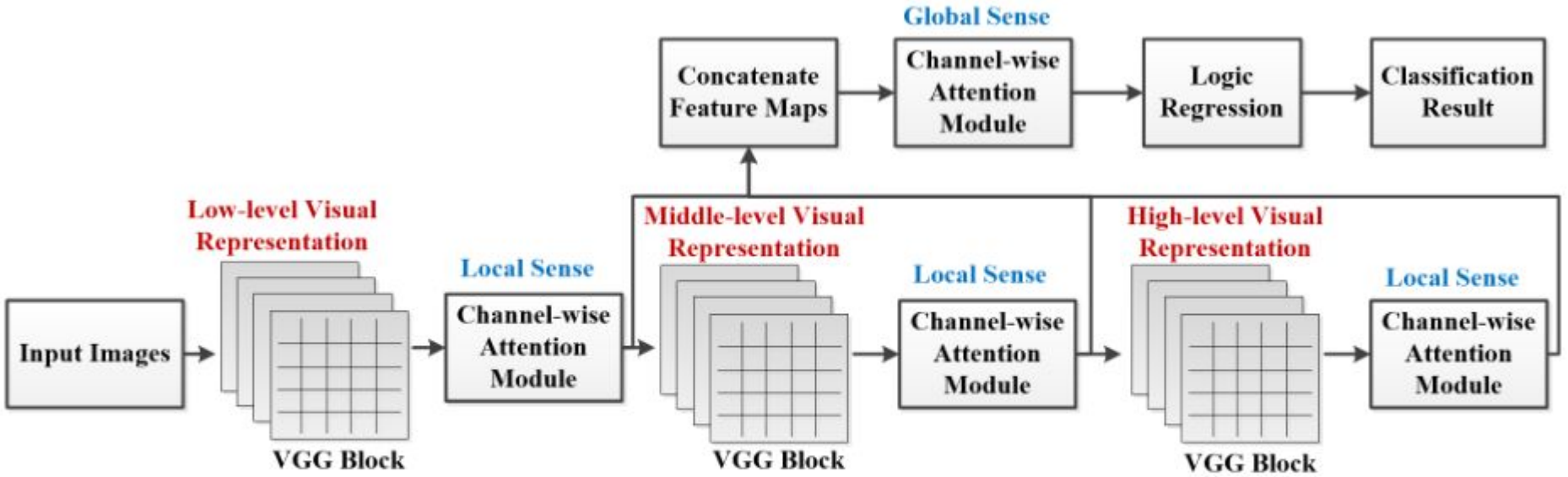


Figure 2. Structure design of attention neural network, where VGG block is designed to extract feature map, channel-wise attention module is utilized to involve channel-wise context information, and multi-layers of attention modules are used to build hierarchy structure.

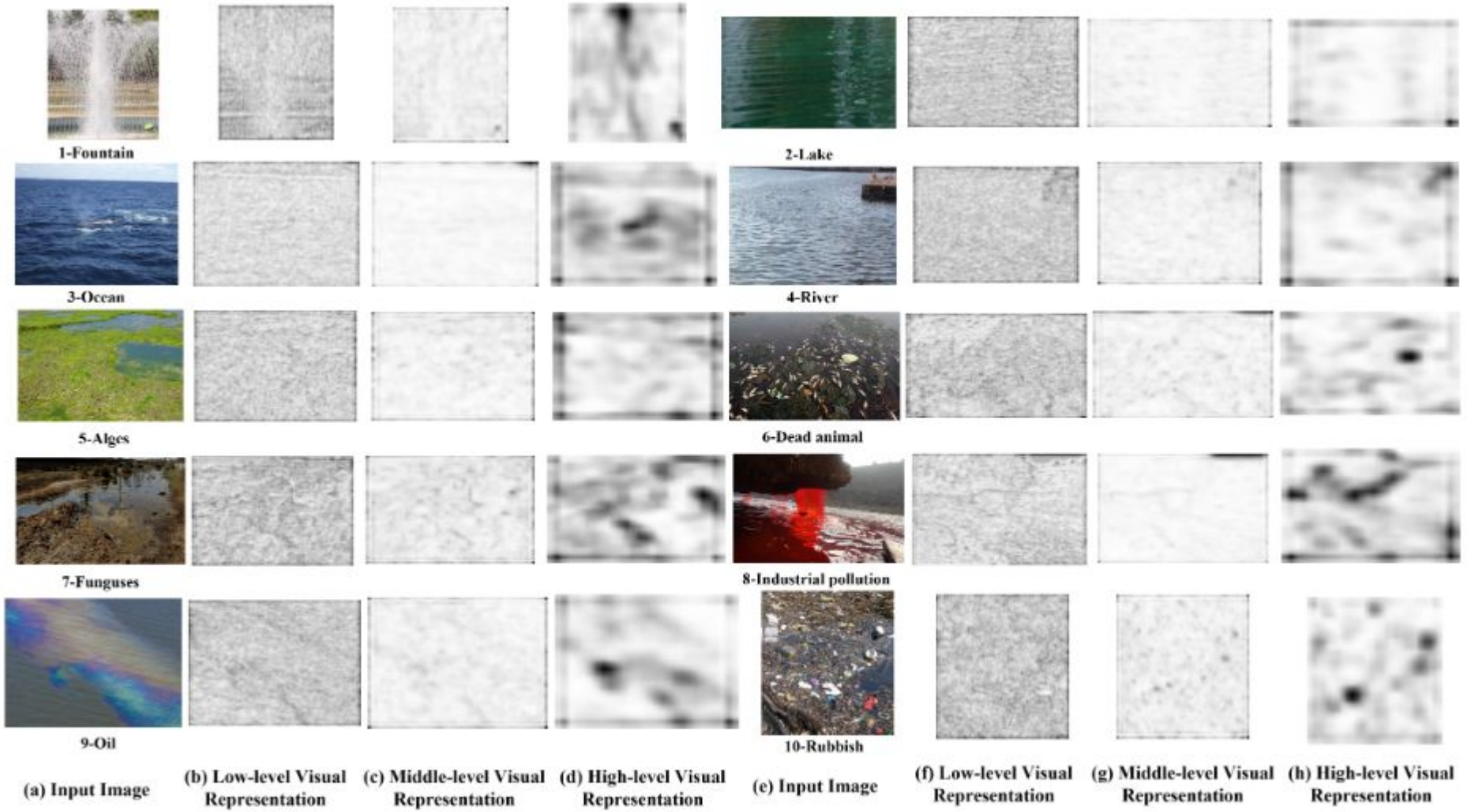
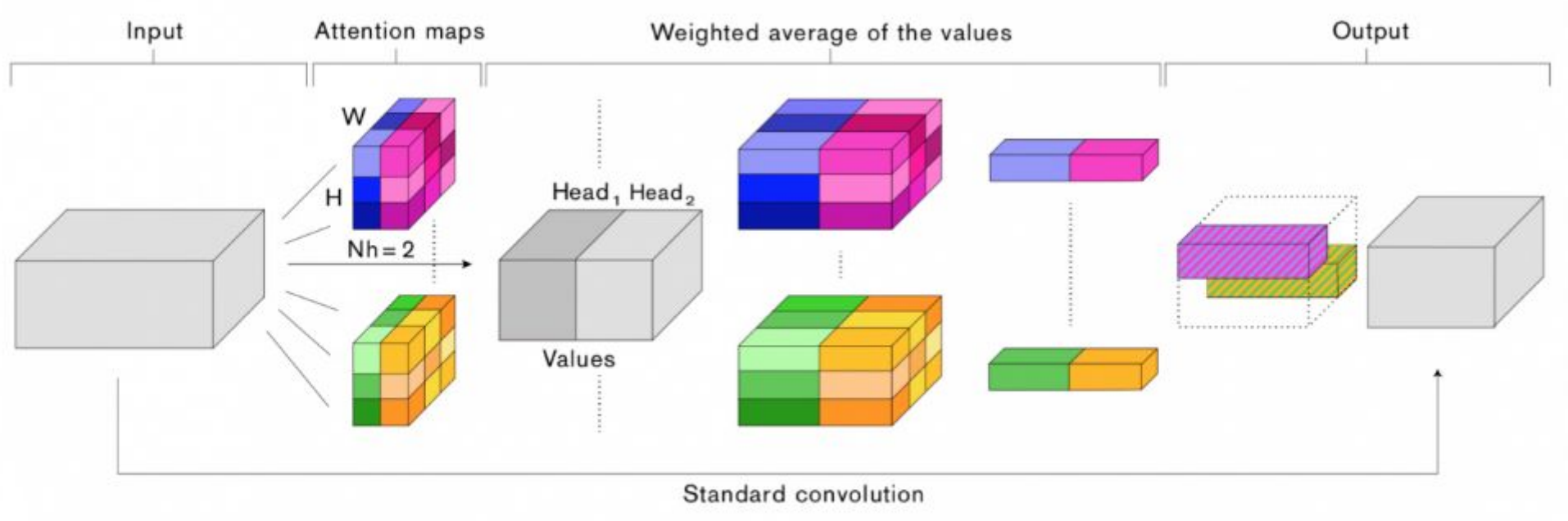


Figure 3. Input samples and corresponding intermediate results for low-, middle-, and high-level visual representations of feature maps extracted by the proposed network.

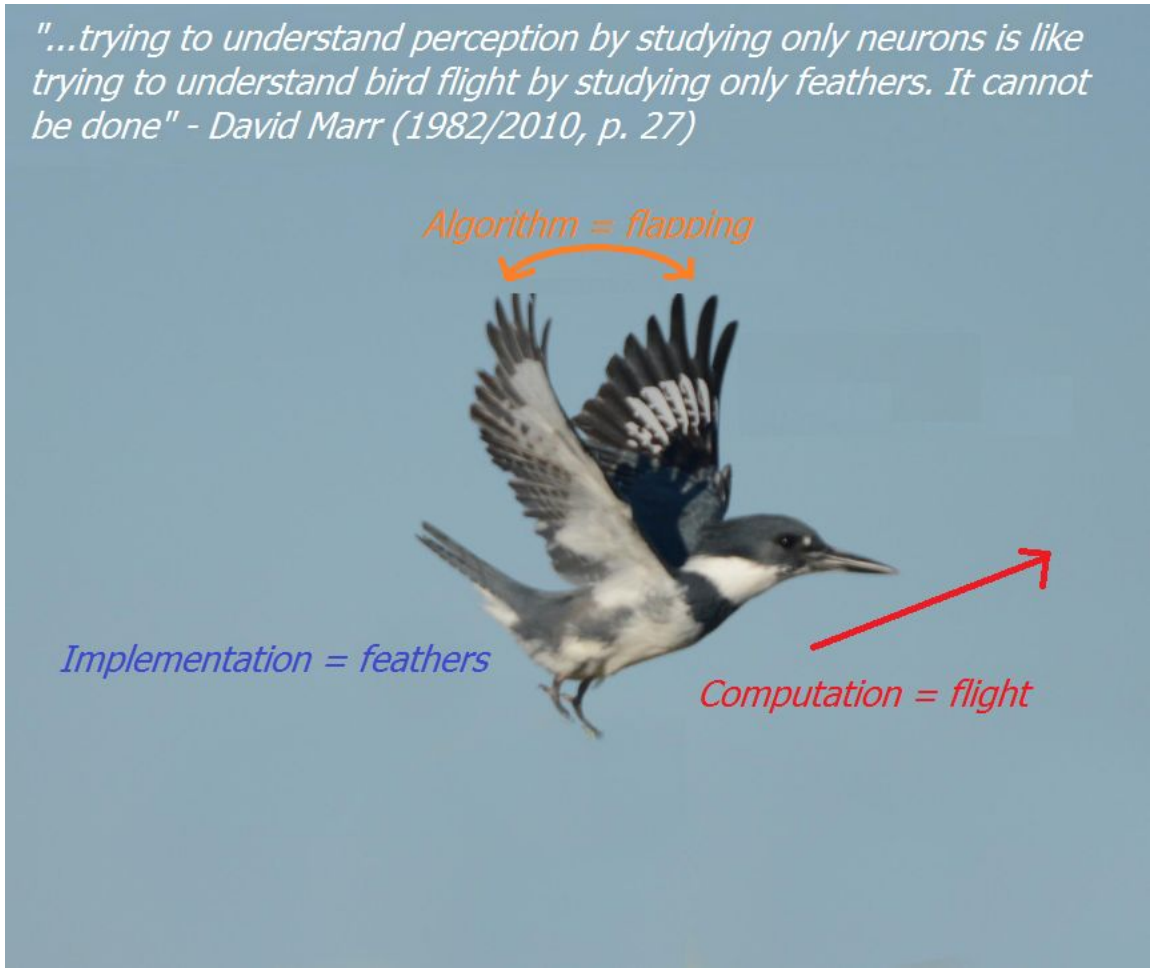


The attention model run in parallel to the convolution operation. The blue/pink and orange/green maps represent the different results for each head. (Image: Bello et al.)

https://www.lyrn.ai/2019/05/03/attention-augmented-convolutional-networks/

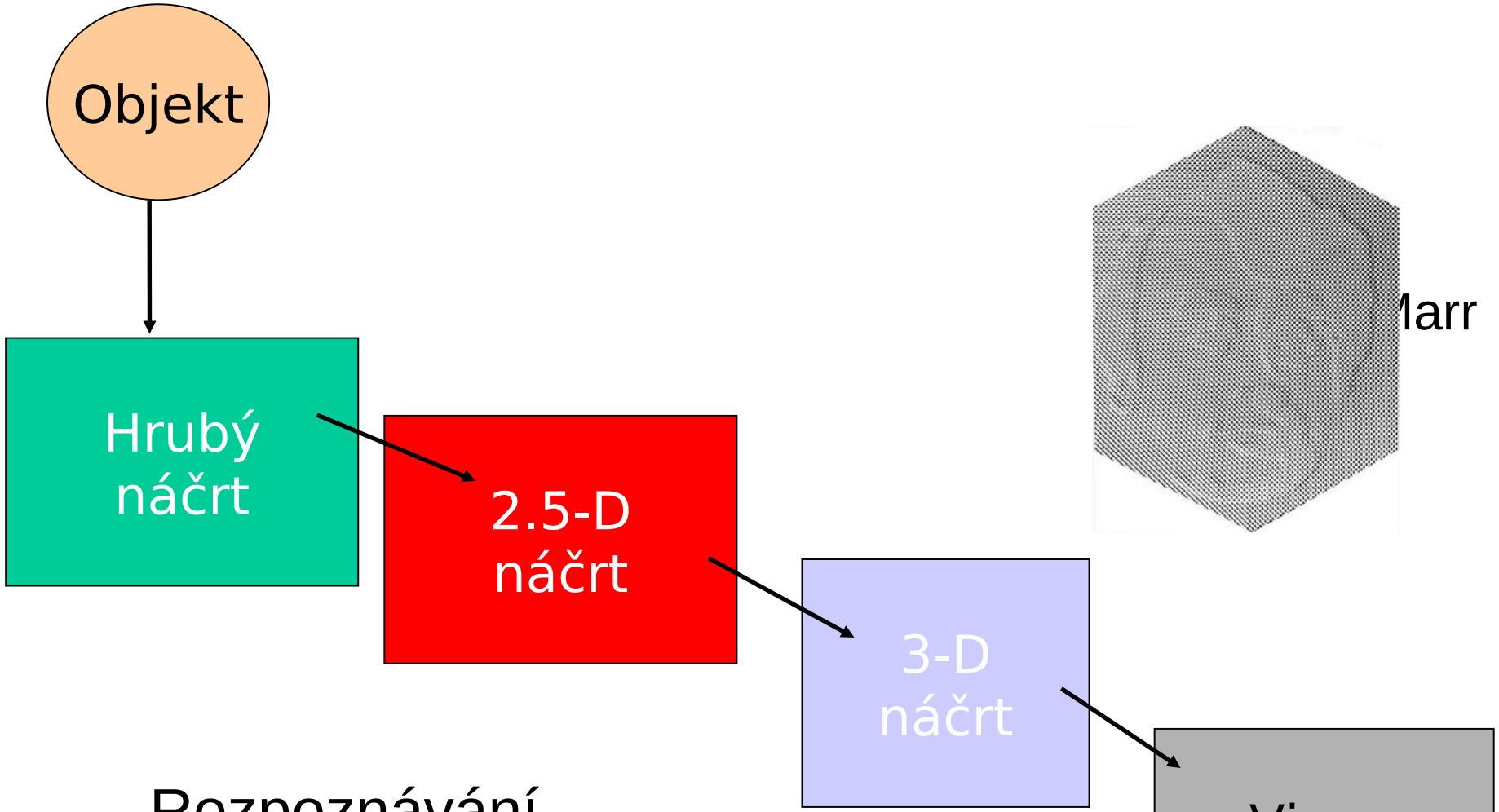
https://arxiv.org/pdf/1706.03762.pdf

Marrův model



source:

Marrův model



Marr

Rozpoznávání ve třech krocích

Marrův model

Hrubý náčrt

- Nalezení oblastí diskontinuity v intensitě světla. Indikují kontury.
- Vztah mezi konturami.
- Hotový prvotní náčrtek je popisem vztahů na sítnici.

2.5-D náčrt

- Rekonstrukce třetí dimenze.
- Závislost na momentální podobě.

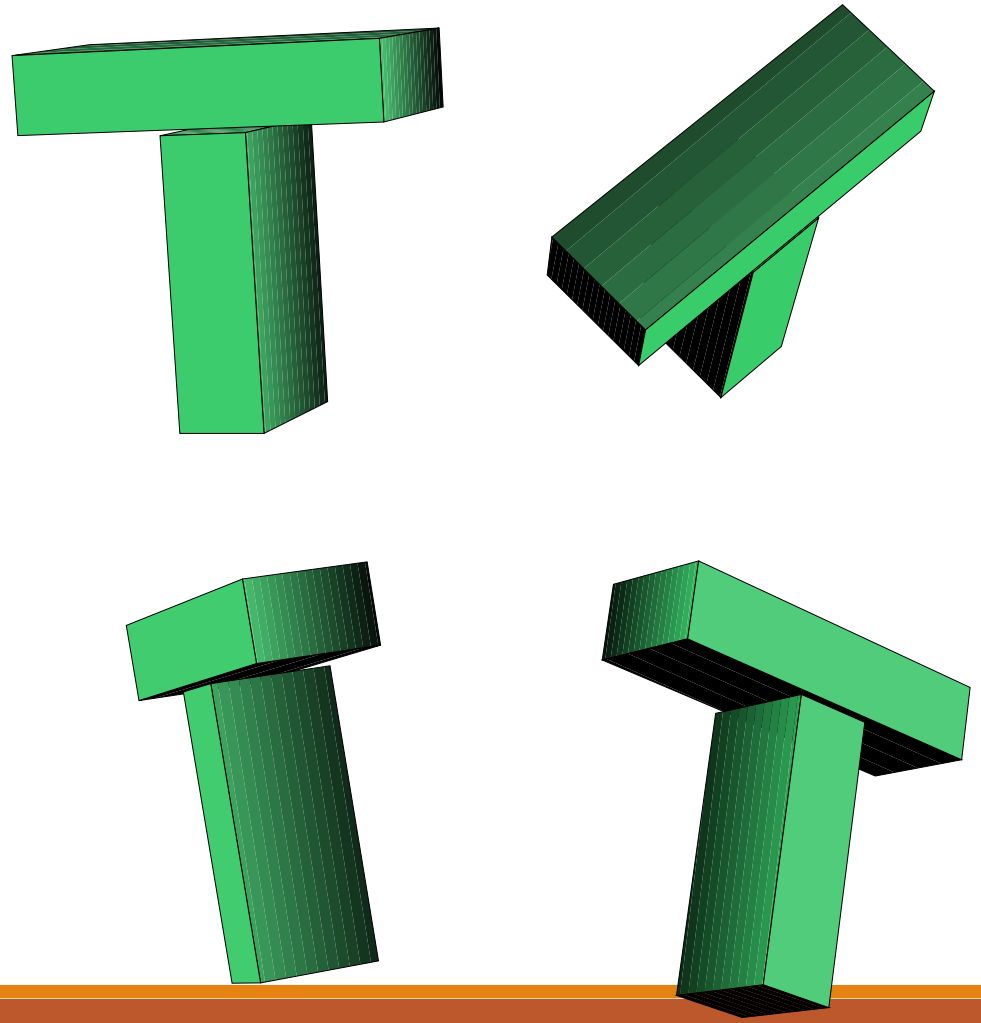
3-D náčrt

- Rozpoznání objektu z libovolného úhlu.
- Definován vztah vzhledem k pozorovateli.

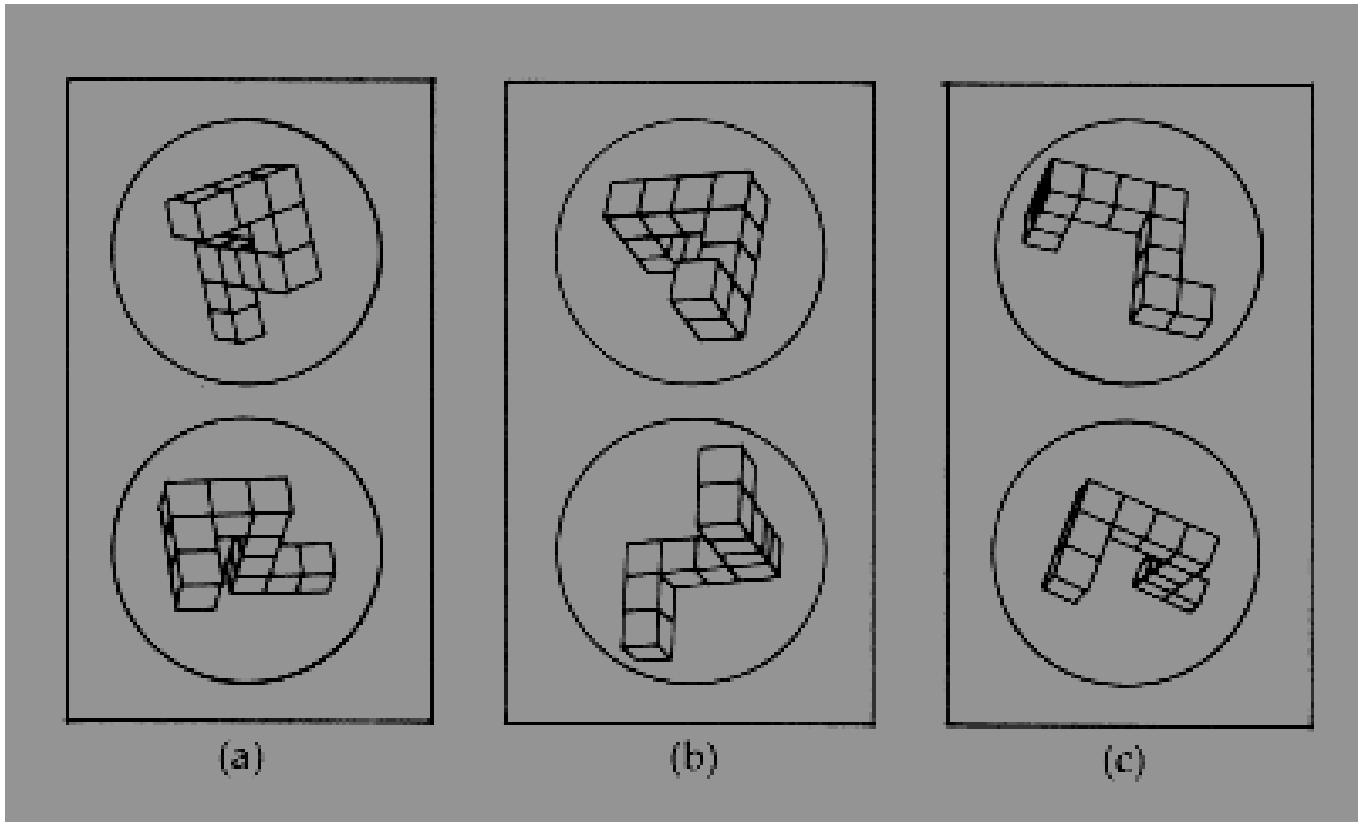
Hrubý náčrt

Hrubý náčrt

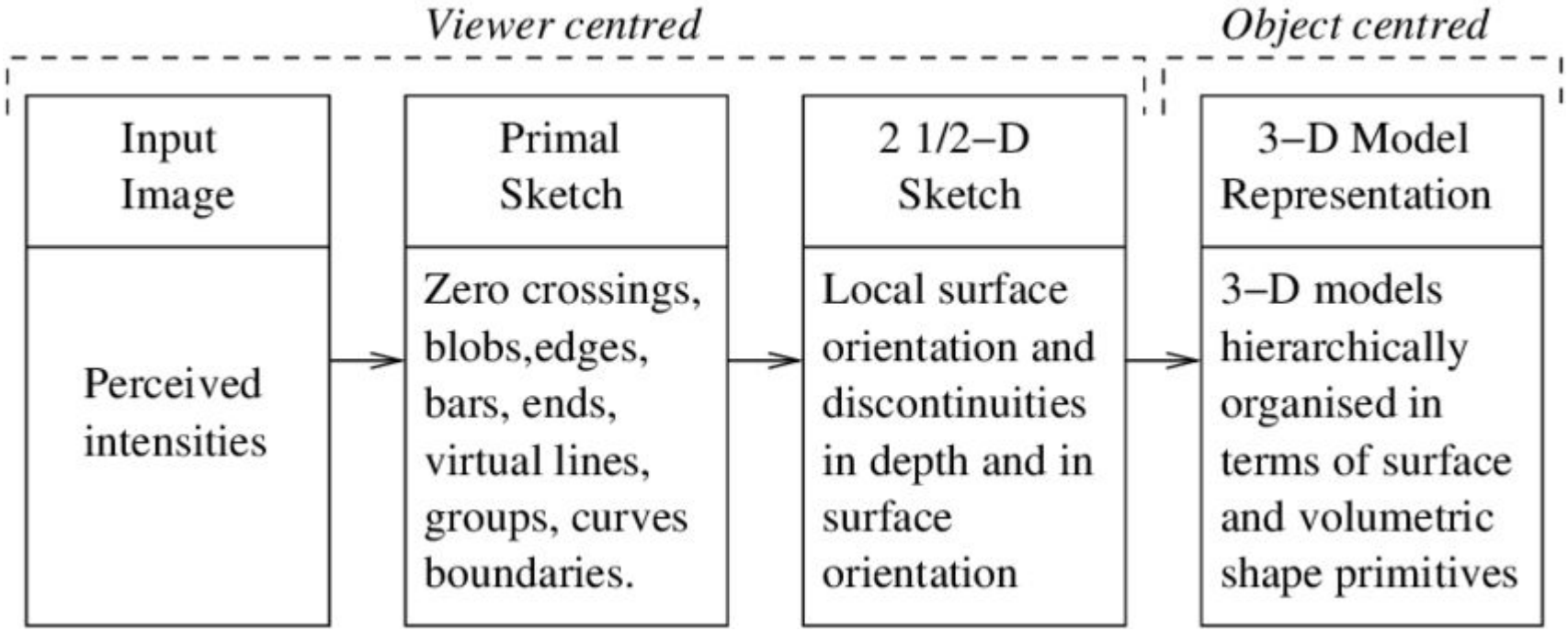
2,5 D náčrt



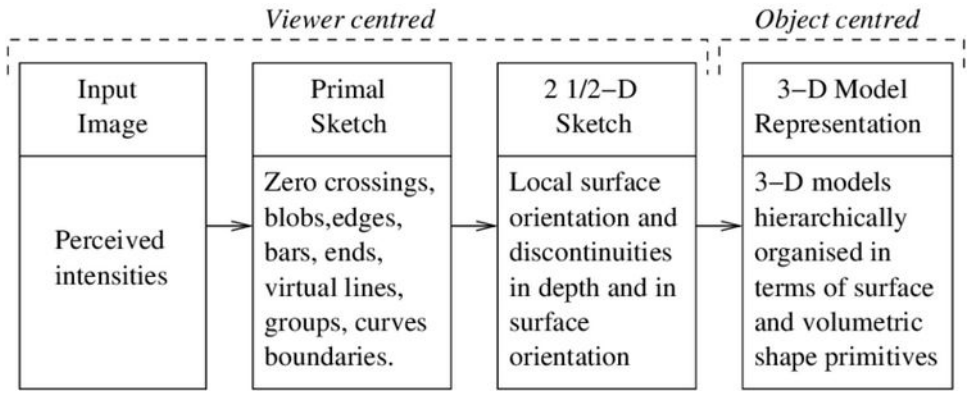
3D náčrt



Marrův model



Marrův model



1.Image-based Processing (low lev
2.Surface-based Processing
3.Object-based Processing
4.Category-based Processing (high

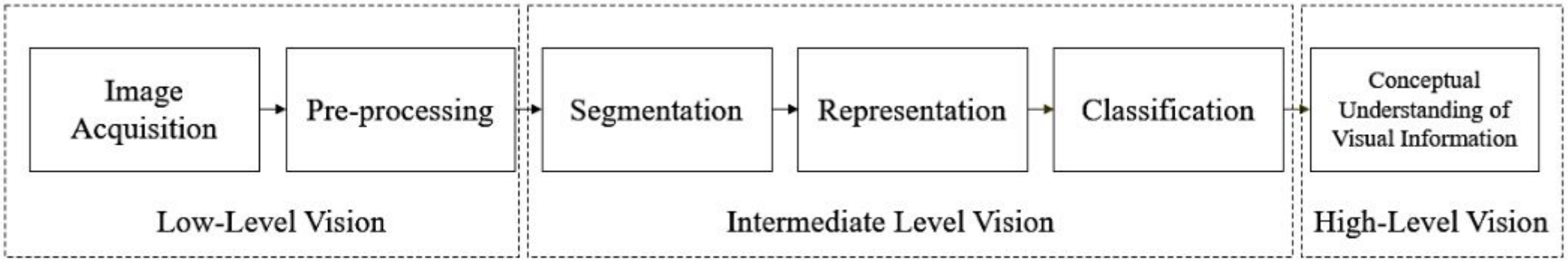


Fig. 1 – Block diagram of a typical cognitive vision system

Rozpoznání prostřednictvím komponent

Irving Biederman (1987)

Partial Tentative Geon Set Based on Nonaccidentalness Relations

	Geon	Edge: Straight S, Curved C	Symmetry: Rot & Ref ++, Ref +, Asymm –	Size: Constant ++, Expanded –, Exp & Cont – –	Axis: Straight +, Curved –
1.		S	++	++	+
2.		C	++	++	+
3.		S	+	–	+
4.		S	++	+ +	–
5.		C	++	–	+
6.		S	+	+ +	+

Rozpoznání prostřednictvím komponent

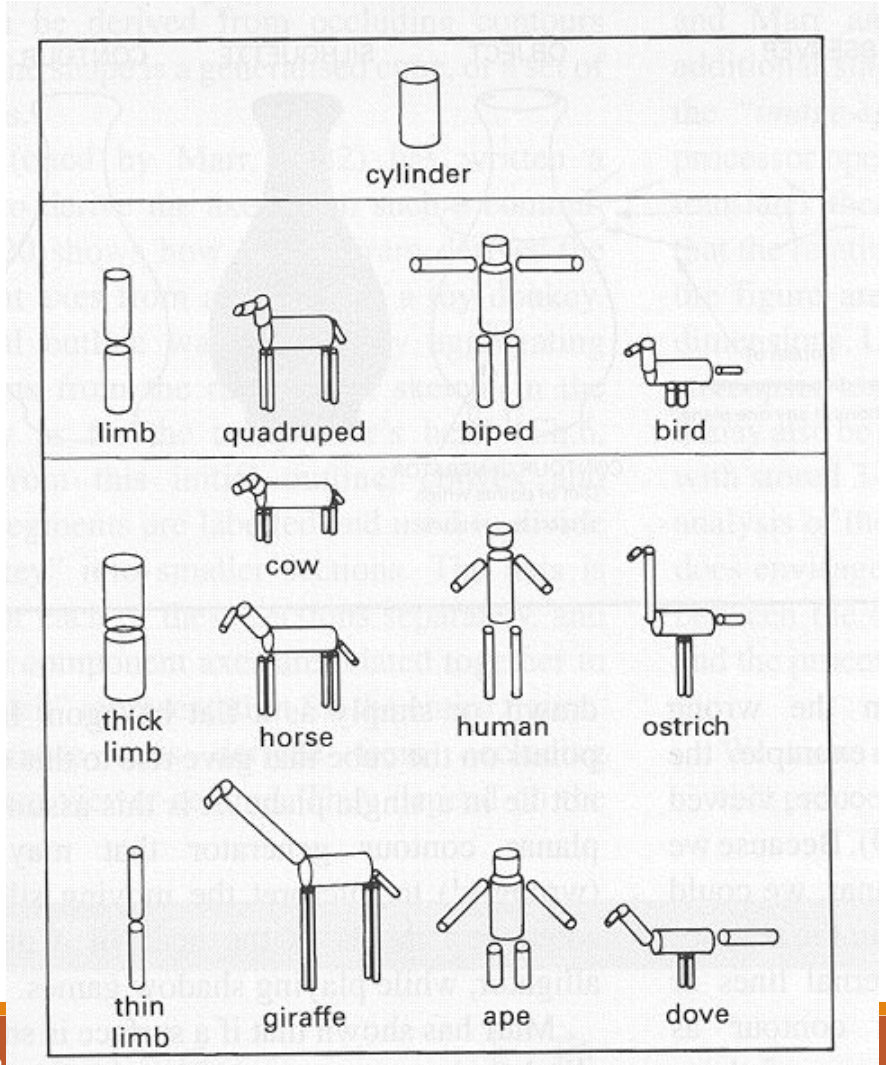
Objekty poznáváme tak, že detekujeme jednoduché 3-D tvary a nalezneme vztah mezi nimi panující

Primitives se liší v geometrických vlastnostech

Jejich kombinací vzniká spousta tvarů - objektů

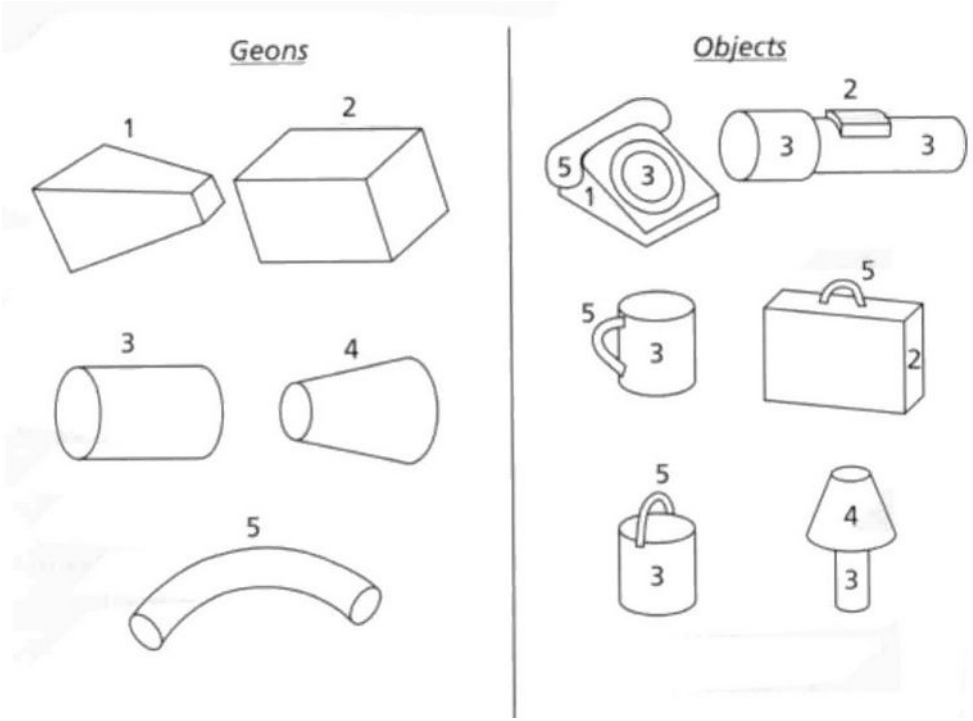
Předpoklad plné automatisace (heslo: úspornost vnímání) – není místo pro vliv zkušenosti

Rozpoznání prostřednictvím komponent



Rozpoznání prostřednictvím komponent

Geons

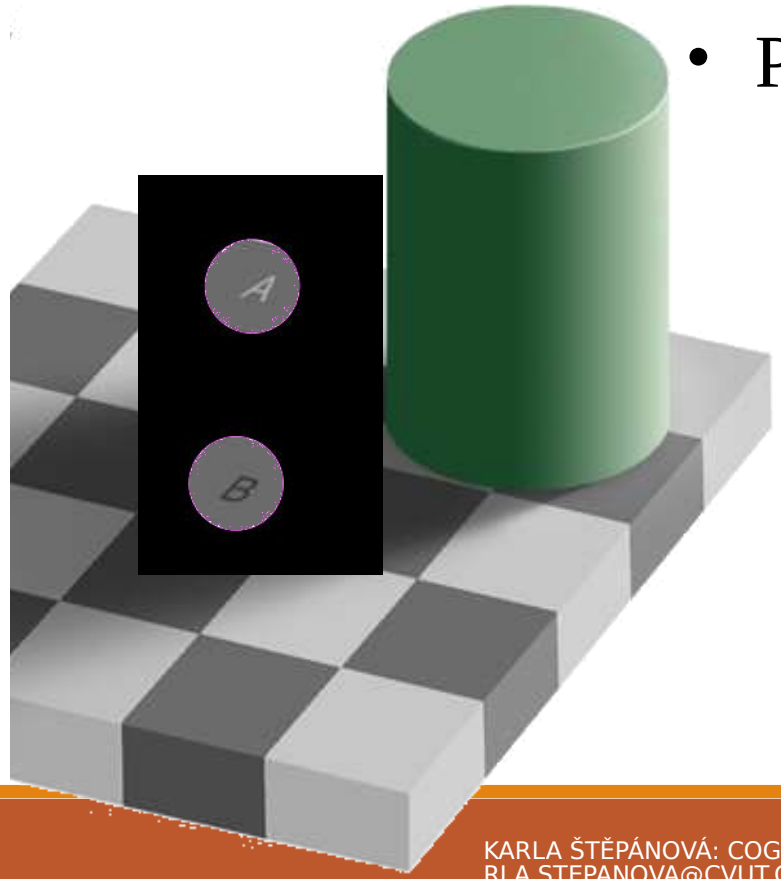


Percepce a vnímání

- Jaký je mezi nimi rozdíl?
- Percepce je činnost smyslových orgánů.
- Vnímání je proces rozpoznání, organizace a interpretace informací.

Základy vnímání

- Perceptuální konstanty
  - Objekt zůstává identický přestože je vnímán odlišně.
- Příklad:
  - Tvarová konstanta



Jak vnímáme hloubku

- Nápovědi při vnímání hloubky
  - Obrazové
    - Vzájemná pozice
    - Velikost
    - Gradient textury
    - Lineární perspektiva
  - Pohybové
  - Binokulární

Vzájemná pozice

Velikost

Amesova komora

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

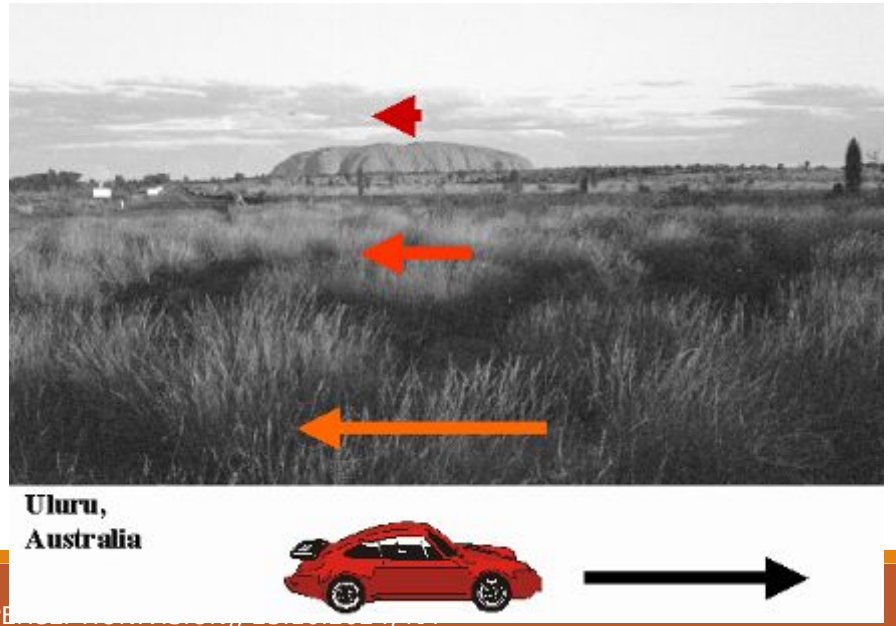
http://youtu.be/gJhyu6nlGt8

Gradient textury

Lineární perspektiva

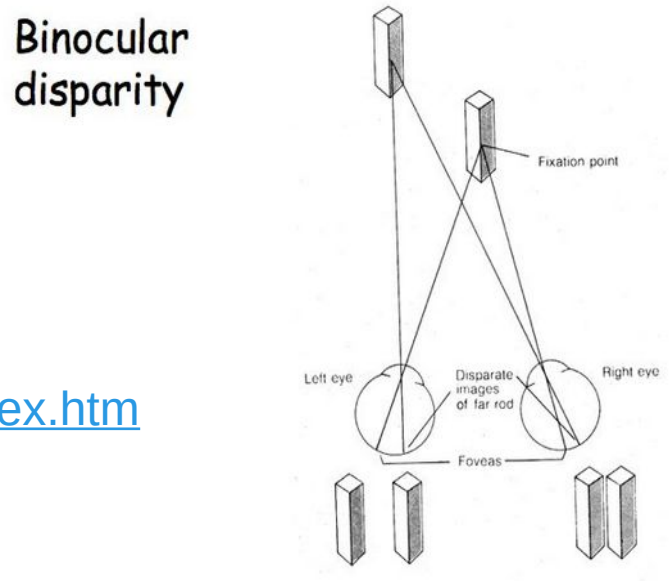
Pohybové nápovědi

- Pohybová paralaxa- Během pohybu se okolní objekty pohybují rozdílnou rychlostí vzhledem ke vzdálenosti od pozorovatele



Binokulární nápovědi

- Binokulární disparita – oči jsou od sebe vzdáleny asi 8 cm a vytvářejí dva odlišné pohledy na okolí. Z nich mozek vytváří 3D představu.
  - Palce za sebou
  - Stereogramy



http://www.netaxs.com/~mhmyers/rds-ex.html

Teorie vnímání

- Přímé teorie vnímání
  - Vnímá vzniká na základě informací z prostředí
  - Zpracování zdola nahoru
  - Části jsou identifikovány, složeny dohromady a rozpoznány
- Konstruktivní teorie vnímání
  - Lidé aktivně konstruují své vjemy na základě očekávání
  - Zpracování zhora dolu

Gibsonova ekologická teorie

- Všechny informace nutné ke tvorbě vjemu jsou obsaženy v prostředí
- Vnímání je okamžité a spontánní
- Není potřeba zpracování shora
- Vnímání a jednání nelze oddělit
- Vnímání determinuje jednání a to vytváří nové podněty pro vnímání

Pár otázek

- Pokud si člověk zapamatoval nějakou informaci opilý, bude si jí lépe vybavovat v opilosti než za střízliva.

- Ano – Výzkum učení závislého na situaci dokazuje vliv prostředí na proces zapamatování a vybavování informací. Pokud obě probíhají ve stejném kontextu, je výsledek lepší.

Pár otázek

- Pozpátku nahrané zpravy umístěné v hudební nahrávce ovlivňují chování posluchače.

- Ne – Nebylo prokázáno působení těchto zpráv na chování posluchače .

Pár otázek

- Technika rychlého čtení může zvýšit schopnost čtení textu, při zachování porozumění čtenému.

- Ne– Výkonnost při čtení je ovlivněna faktory rychlosti a přesnosti. Čím rychleji čteme, tím menší přesnost v pochopení. Některé techniky mohou rychlost čtení pouze mírně zvýšit.

Pár otázek

- Freudova technika volných asociací nám poskytuje informace o organizaci paměti.

- Ano – Metoda je podobná sémantickému primingu, založeném na teorii šíření aktivace. Výzkumy nám mohou odhalit individuální rozdíly v organizaci paměti

Pár otázek

- Reklama používající podprahové vnímání je velmi efektivní.

- Ne –Efekt podprahového vnímání je minimální. Výzkumy neprokázaly významný rozdíl oproti klasické prezentaci.

Pár otázek

- Kapacita dlouhodobé paměti je neomezená.

- Ano – Nikomu se dosud nepovedlo zaplnit svou dlouhodobou paměť. Existují omezení při zapamatování informací, způsobené omezenou pozorností, ale materiál v dlouhodobé paměti je uložen nastálo, pokud nedojde k poškození mozku.

Pár otázek

- Rozdíl mezi 500 Kč a 1000 Kč je psychologicky větší než rozdíl mezi 10500 Kč a 11000 Kč.

- Ano – mentální reprezentace velikosti je v horní části škály zhuštěná, proto je rozdíl jiný, přestože se jedná o matematicky stejnou hodnotu.

Pár otázek

- Jestliže je někdo slepý na jedno oko, nemůže vnímat hloubku.

- Ne – existují nápovědy při vnímání (velikost, vzájemná pozice, atmosférická perspektiva), které poskytují informaci o hloubce.