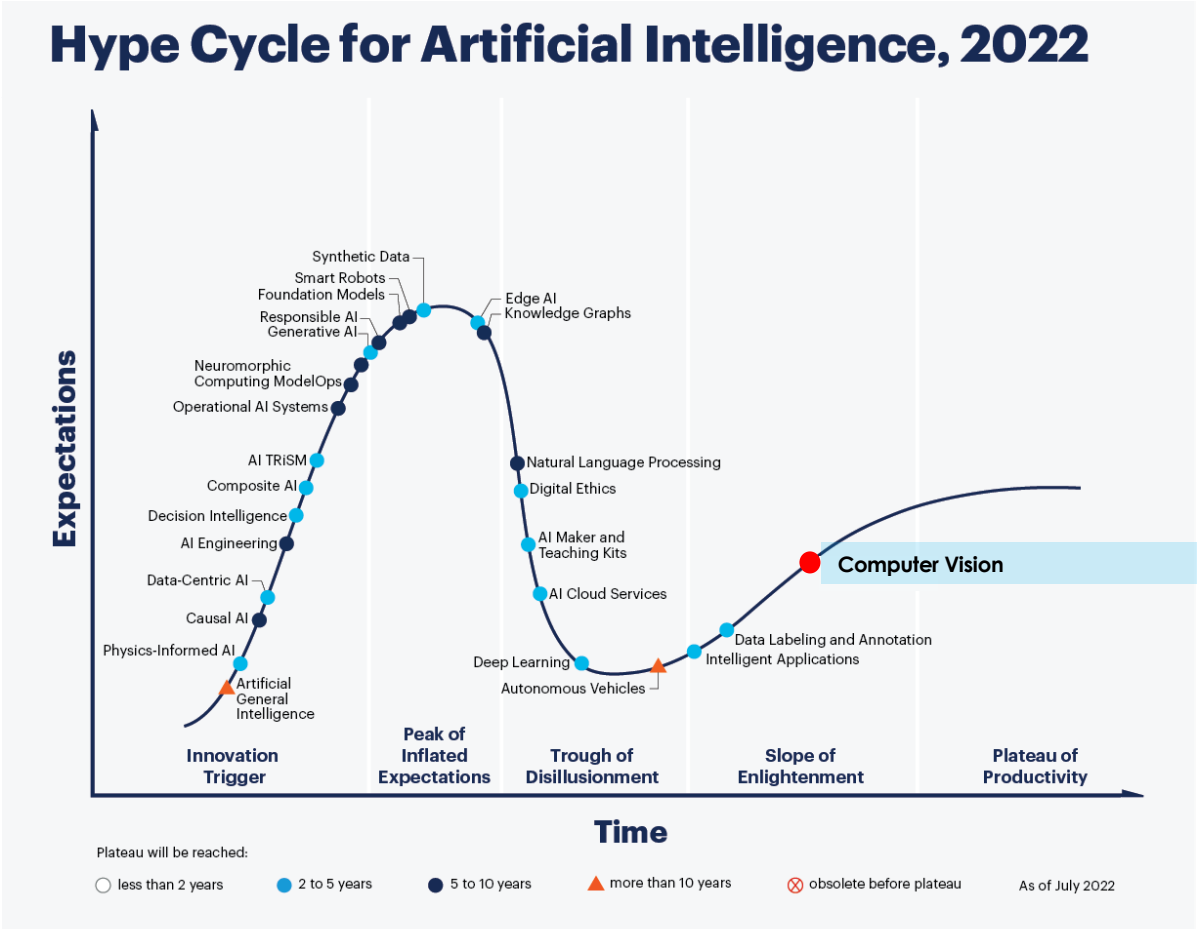
VISMAC 23

Padova 4-8 Settembre 2023

COMPUTER VISION
where it goes with GPT4

Alberto Del Bimbo
Università di Firenze

Computer Vision 2022

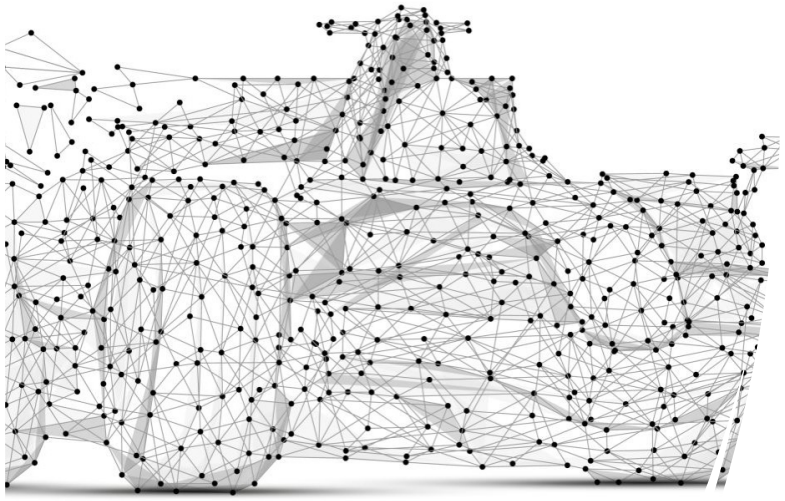


Credits Gartner

Computer Vision is the field where most AI research is being conducted.

Computer Vision is currently past the highest peak on the *slope of enlightenment*, preparing to reach *platea productivity*, where mainstream adoptions will start to take off.

Computer Vision research trends at CVPR23



- *Render the real.*

  Research emphasis this year was on bringing technology closer to reality. Top paper categories included:
  3D Computer Vision; *Image and Video Generation*; *Understanding Humans in Video*.

- *Create an autonomous ecosystem.*

  Research is shifting from how the vehicle may react in an environment to planning for how the environment may respond to it.

- *Converge image and language for more sophisticated techniques.*

  The technology is expanding from an image or text-based approach to a combined effort. Yet converging these modalities with accuracy creates complicated research challenges.

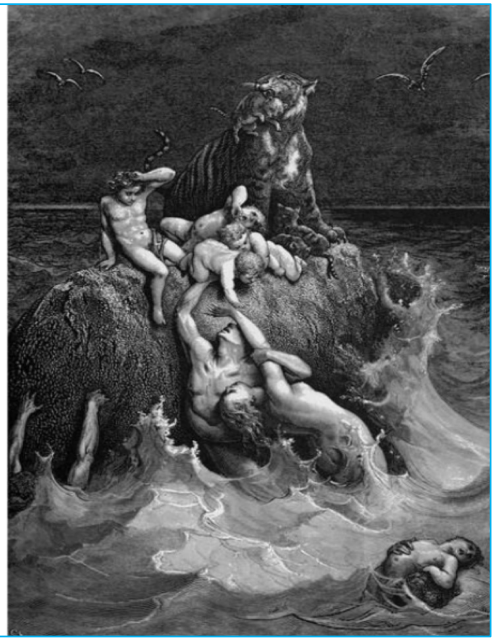
- *Collaborate to meet market demand.*

  More widespread collaboration between academia, government, and industry to meet growing technological demands.

Alexei A. Efros, CVPR 2023

Things are changing fast

- When **GPT4 Image** model is released, it will likely render **80%** of this CVPR's papers obsolete!
  - I've been urging to focus on data for 25 years, and even I am surprised!



Credits A. Efros

Features of Chat GPT4

User What is unusual about this image?

Credits Barnorama

GPT-4 The unusual thing about this image is that a man is ironing clothes on an ironing board attached to the roof of a moving taxi.

Conversational abilities
Can analyze and classify text
Can create websites
Can code flawlessly
Can respond to natural language queries
Can answer trivial questions
Can provide language translations

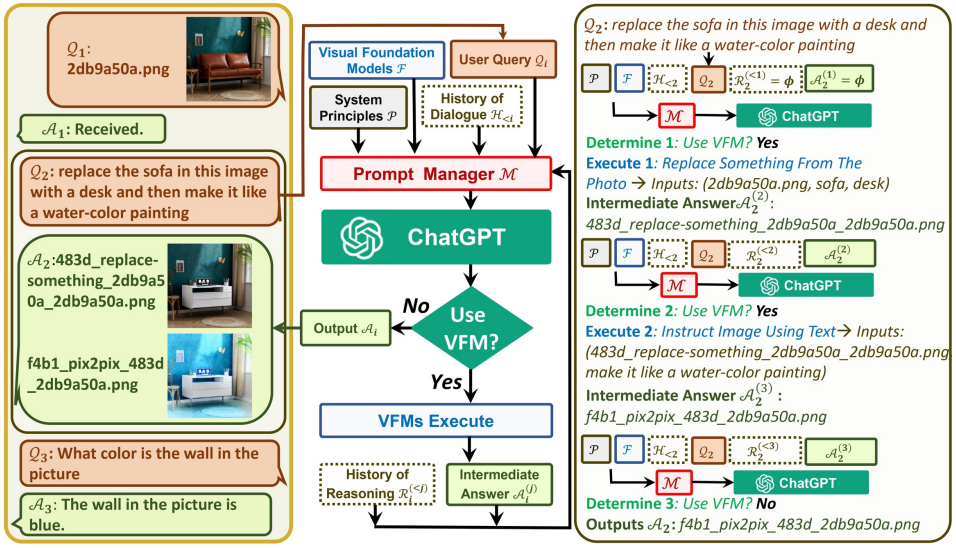
Can create creative content
Can describe images In detail
Can answer questions based on the image *

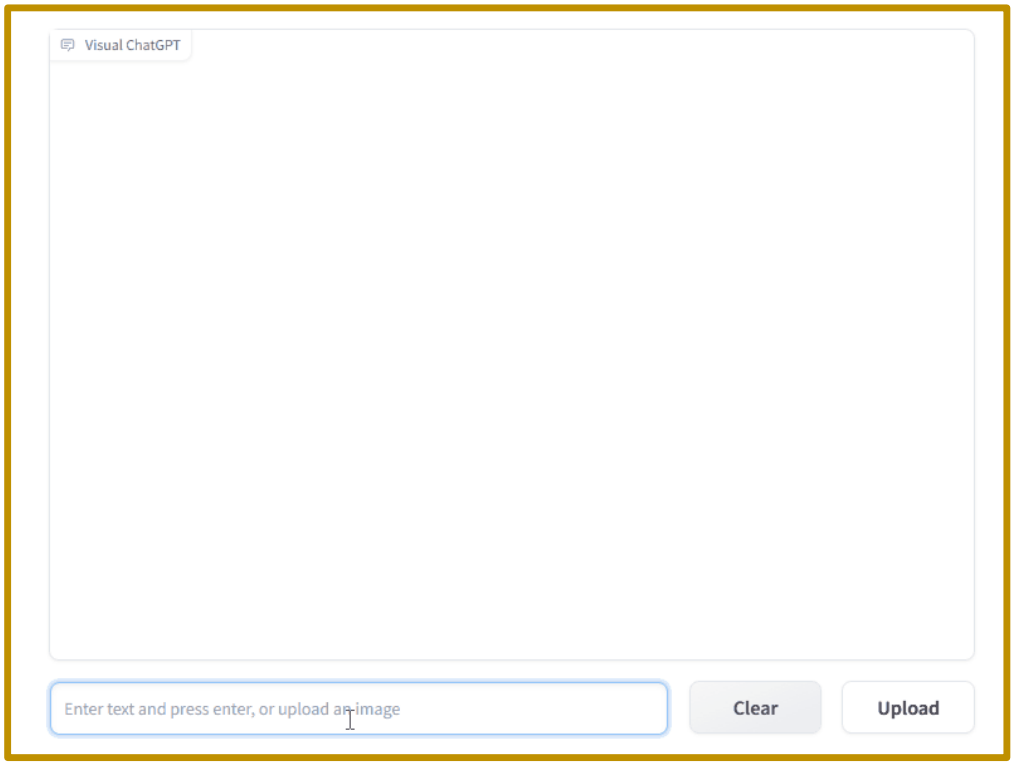
* Not yet available in the plus version of ChatGPT
Only possible in GPT-4 API, waitlist to join

Visual ChatGPT

Microsoft Research recently open-sourced Visual ChatGPT, a chatbot system that can generate and manipulate images in response to human textual prompts.

The system combines OpenAI's ChatGPT with 22 different *visual foundation models* to support multi-modal interactions.



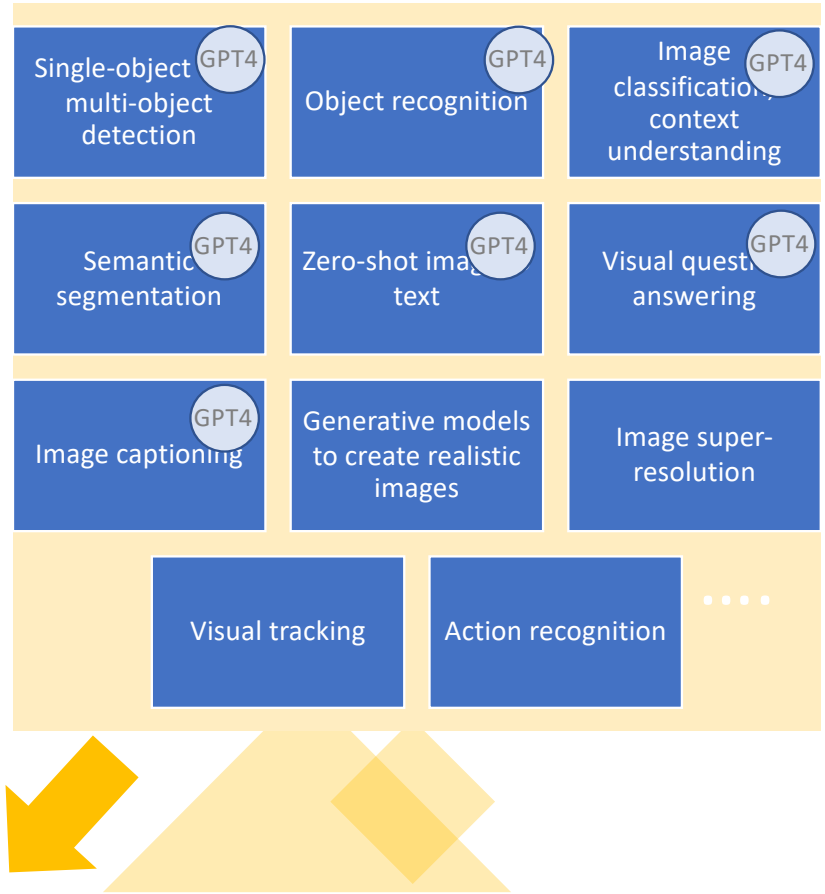


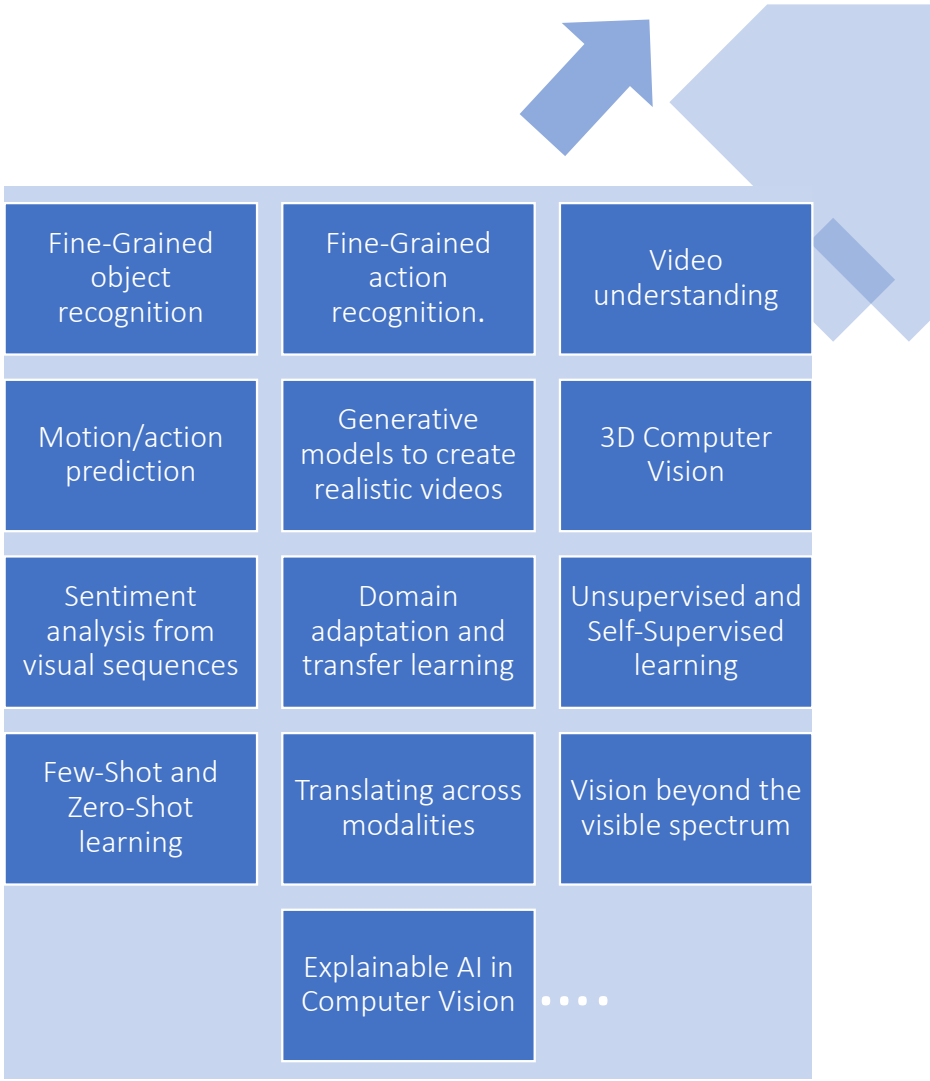
Credits Microsoft Research

Trending topics in CV and GPT4

The main use case for GPT-4's multimodality might be for general consumer use rather than for industrial-grade CV tasks.

Task-specific CV models still vastly outperform GPT-4.
Downstream training of GPT





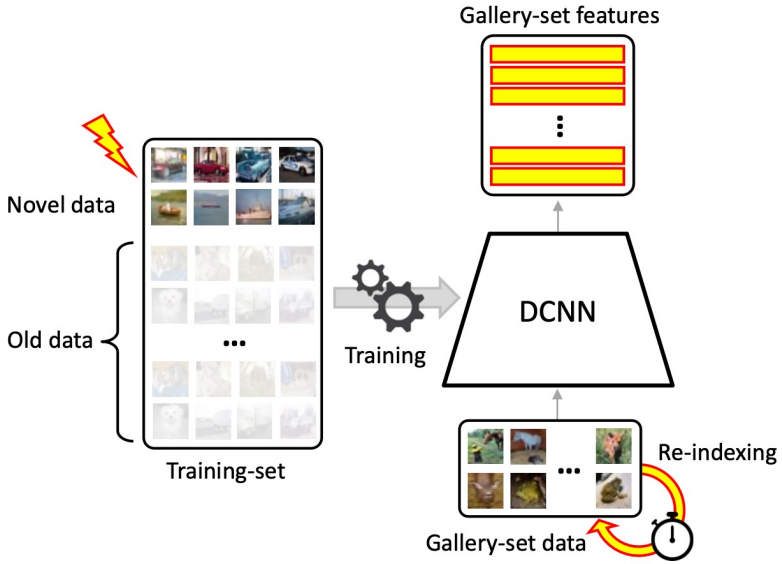
4 captivating research avenues to consider

(from my perspective)

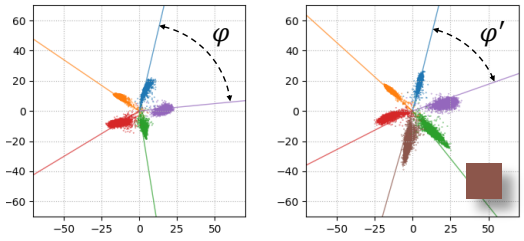
Learning compatible representations

Compatible representation learning aims to learn representations that can be used interchangeably in time.

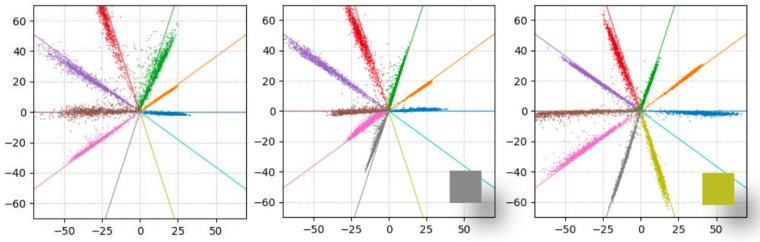
In Visual Search it enables the seamless inclusion of novel training data into the representation, sidestepping the necessity of re-indexing the gallery.



Non-compatible learning



Compatible learning



Credits F. Pernici, UNIFI

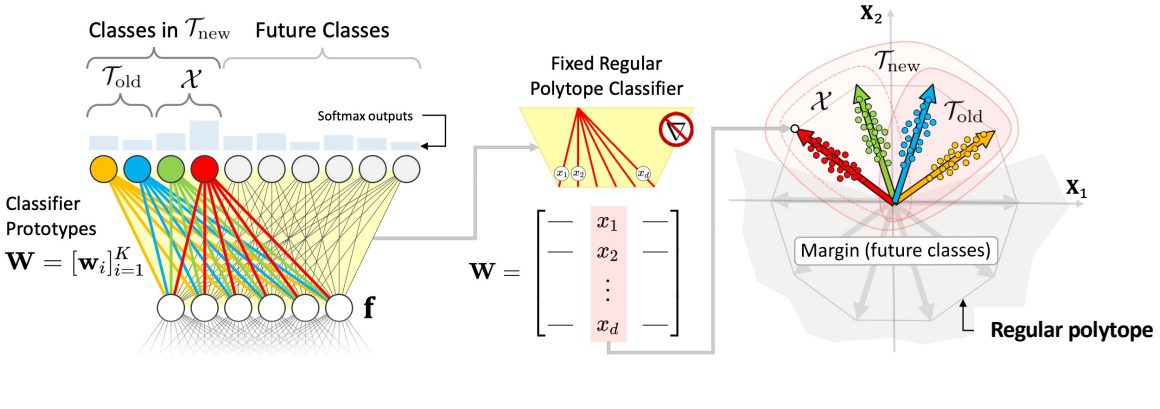
Learning compatible representations

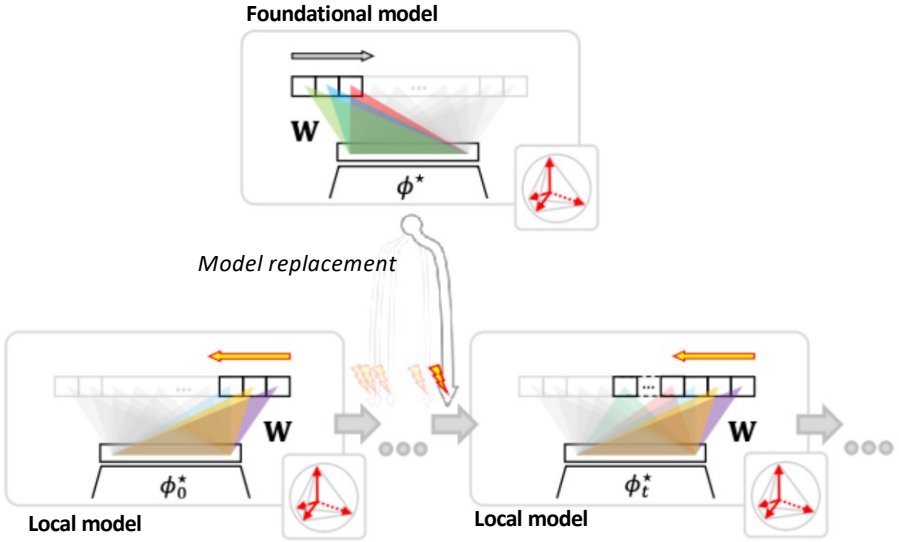
Cores: Compatible representations via stationarity
N. Biondi, F. Pernici, M. Bruni, A. Del Bimbo
IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023

Stationarity and therefore compatibility representation is all you need
N. Biondi, F. Pernici, S. Ricci, A. Del Bimbo
ArXiv 2023

AI paradigm is shifting with the rise of foundational models trained on broad data that can be adapted to a wide range of downstream tasks.

Compatible learning is a pivotal strategy for addressing the task of aligning local fine-tuning with an enhanced iteration of the fundational model.

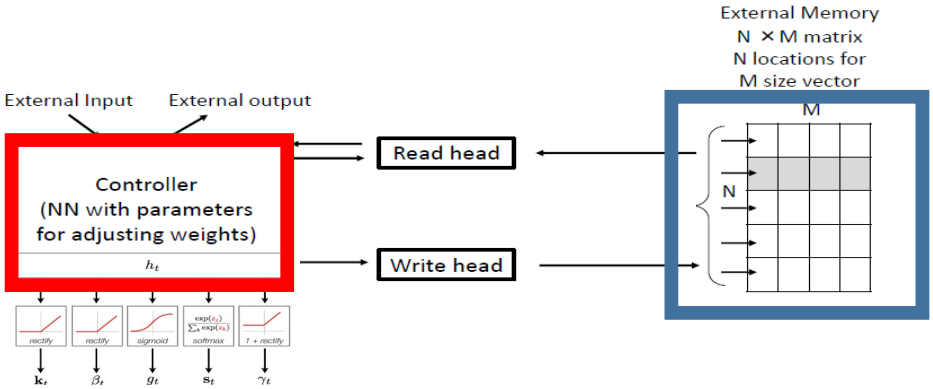




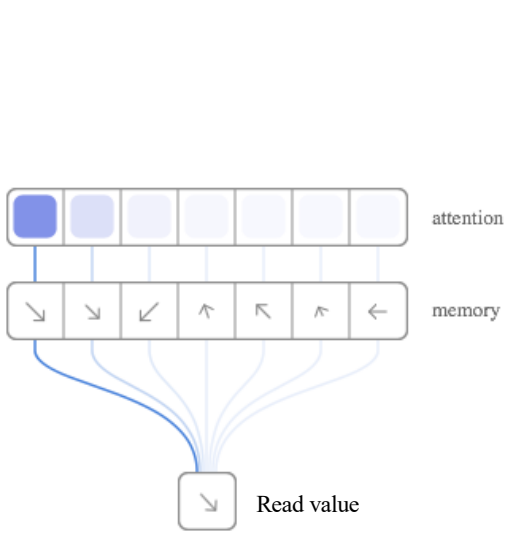
Credits F. Pernici, UNIFI

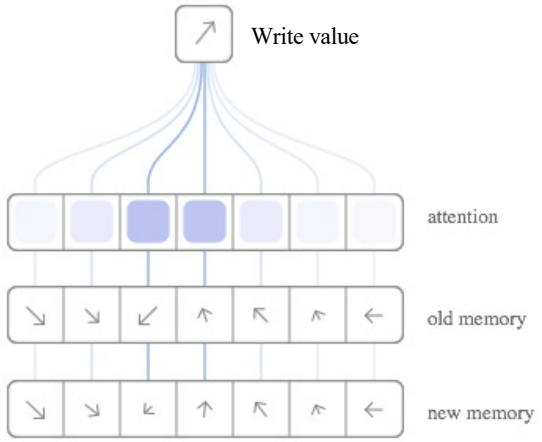
Networks with external memory

Distinguishing between explicit facts — that can be stored in an external memory storage, and implicit knowledge — that is reflected through the networks' trainable weights.



Credits S. Malekmohammadi, Keimyung Univ





Credits A. Graves, Google

Networks with external memory

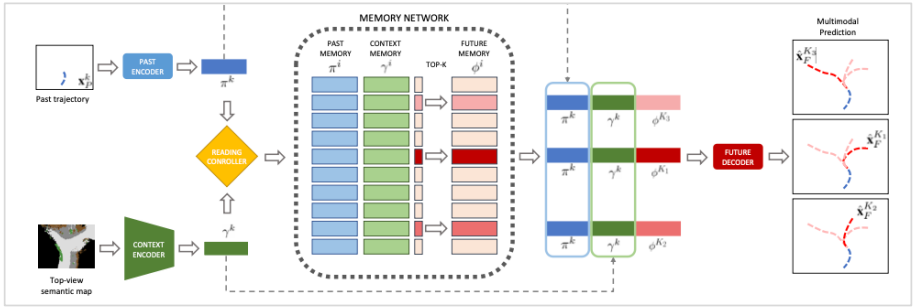
Multiple Trajectory Prediction of Moving Agents with Memory Augmented Networks
F. Marchetti, F. Becattini, L. Seidenari, A. Del Bimbo
IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020

Smemo: Social Memory for Trajectory Forecasting
F. Marchetti, F. Becattini, L. Seidenari, A. Del Bimbo
arXiv preprint, 2022

Memory Augmented Neural Networks to provide predictions reading attentions in the trainable memory during the episodes.

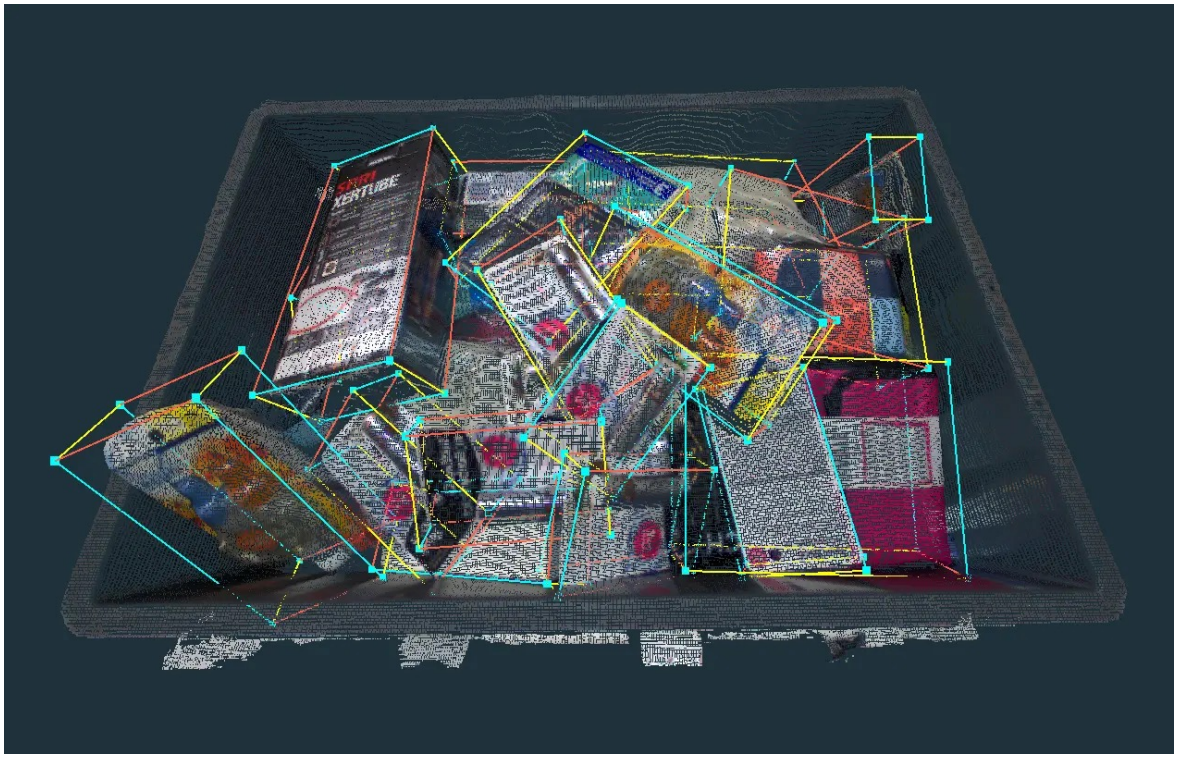
Credits F. Marchetti, UNIFI

Mitigates the inflation in number of parameters needed to store large knowledge, extends the temporal context, improves model explainability.



3D Vision

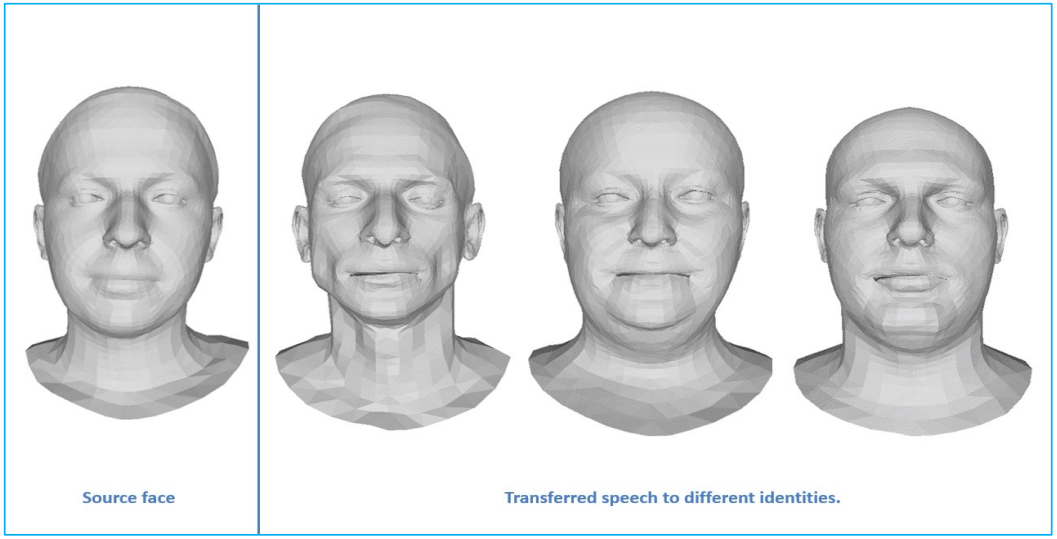
Understanding and interpreting the details of our surroundings,
3D Computer Vision introduces depth information and a level of understanding out of reach for traditional 2D Computer Vision systems.



Credits A. Liu, Covariant

3D Vision

Generating Multiple 4D Expression Transitions by Learning Face Landmark Trajectories
N. Otberdout, C. Ferrari, M. Daoudi, S. Berretti, A. Del Bimbo
IEEE Transactions on Affective Computing, 2023

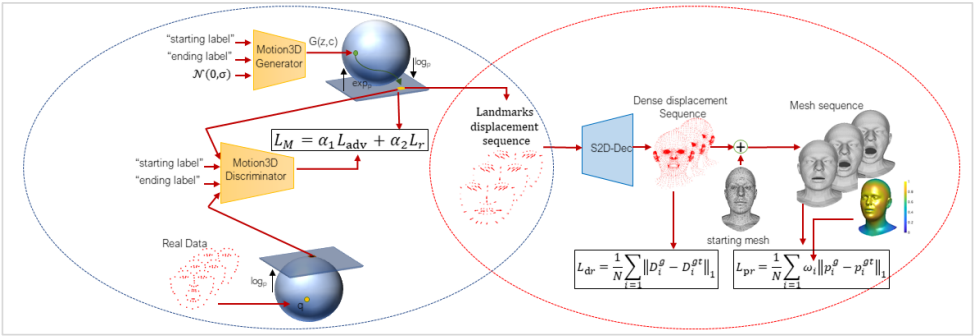


Credits S. Berretti, UNIFI

Modeling 3D faces and face temporal dynamics disentangling structural face elements related to the identity from deformations related to the movable face parts.

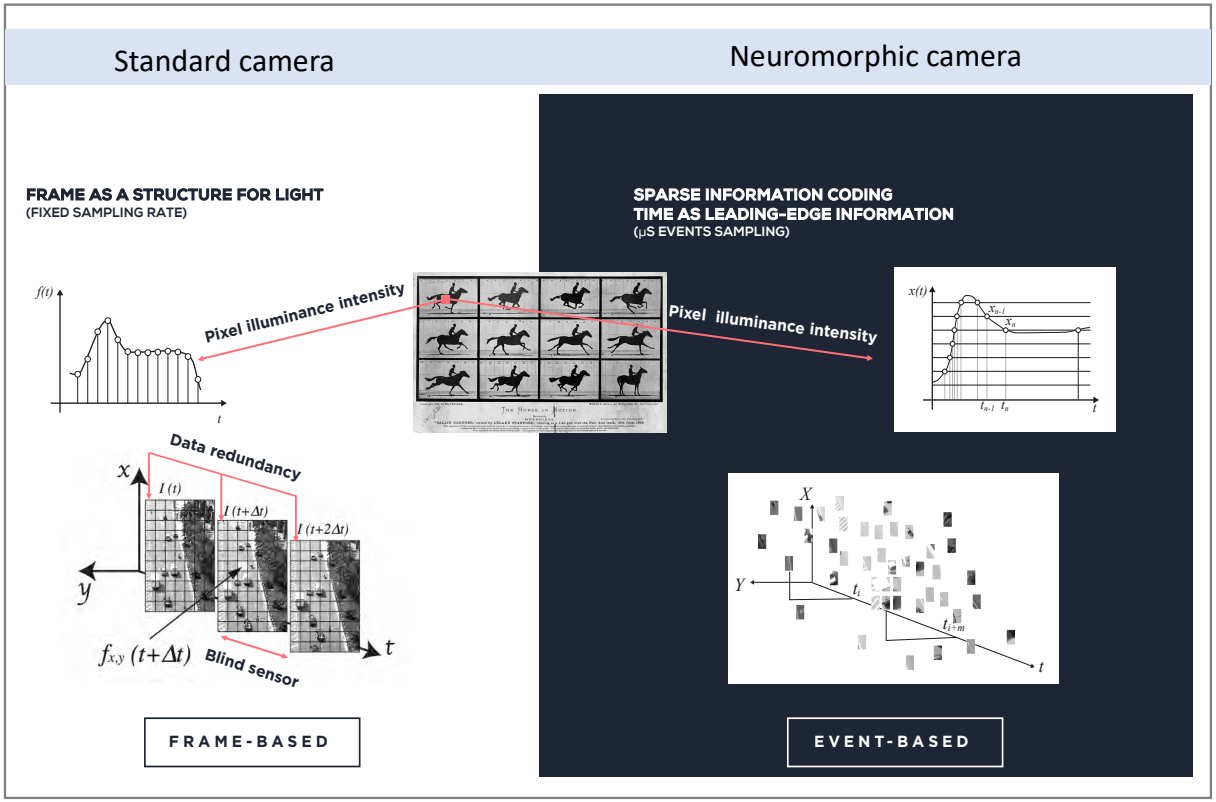
Generate diverse motions even for the same expression. Switch from one expression to another dynamically.

.



Neuromorphic Vision

Bio-inspired sensors that, instead of generating streams of synchronous frames, produce asynchronous events for single pixels where illumination changes occur.



- Asyncronous sampling of order of μ-sec
- Equivalent temporal precision >10000 fps
- Very low power operation <10mW vs 1W
- Higher dynamic range >120dB vs 60dB

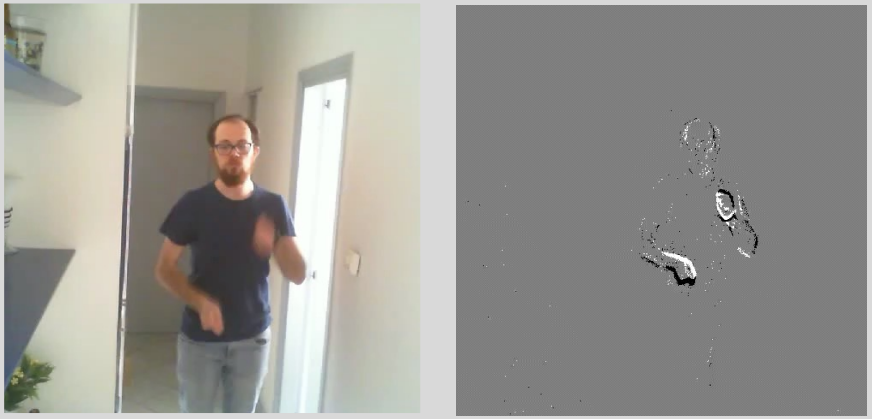
Neuromorphic Vision

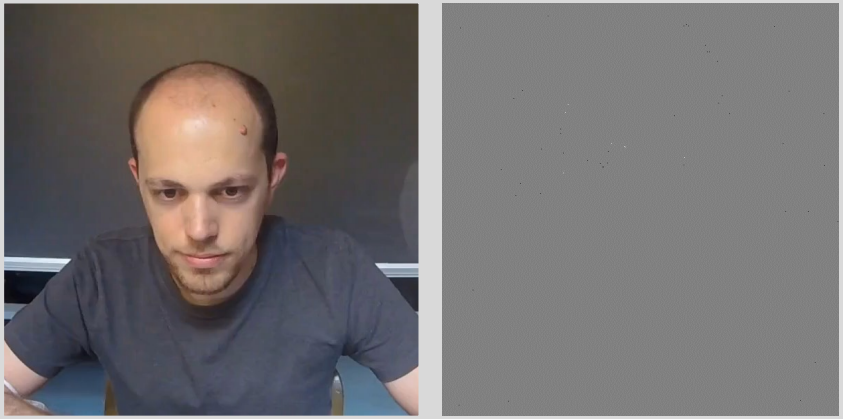
Neuromorphic vision for motion classification at rates that exceed regular cameras, even at a microsecond granularity.

Neuromorphic vision as a privacy preserving tool beyond humans.

Temporal binary representation for event-based action recognition
S Undri Innocenti, F Becattini, F Pernici, A Del Bimbo
Proc. ICPR'20, 2021

Understanding human reactions looking at facial microexpressions with an event camera
F Becattini, F Palai, A Del Bimbo
IEEE Transactions on Industrial Informatics, 2022





Credits F. Becattini, UNISI

Many other research avenues….