AALBORG UNIVERSITY

VISUAL ANALYSIS & PERCEPTION LAB

Visual and Multimodal Learning Systems for Human Behavior Understanding

UNIVERSITAT DE BARCELONA

Dr. Sergio Escalera

www.sergioescalera.com

sergio@maia.ub.es

ELLIS Fellow, IAPR Fellow, AAIA Fellow

University of Barcelona

UNIVERSITAT DE BARCELONA

- Universitat de Barcelona, established in 1450, is one of the top public universities in Barcelona, Spain. It is ranked #168 in QS Global World Rankings 2022.
- 16 faculties & 10 affiliated centres.
- 63K students, 5.2K teachers+researchers, 4.5K trainee researchers, 600 PhD thesis (2018).

University of Barcelona – Faculty of Mathematics and Informatics

UNIVERSITAT DE BARCELONA

- Informatics degree, Prof. Sergio Escalera head of Informatics
- Mathematics degree

- Master involvement (interuniversitary):
  - AI
  - Data Science
  - Computer Vision
  - Behavioral Data Science

Computer Vision Center (CVC-UAB)

Only Center in Europe fully devoted to Computer Vision

- CVC is a legally independent non-profit institution founded in 1995, belonging to the Catalan CERCA network. Located in Bellaterra (Barcelona). Dedicated to research, technology transfer, training, and outreach.
- 38 senior researchers + 52 students (2019)
- 45 JCR indexed journals, 65 international conference papers, 12 thesis (2019).
- **HuPBA**, head Prof. Sergio Escalera, 1 of the 8 strategic research lines

Computer Vision ecosystem en Catalonia

+1.185 professionals

90 Pure Player companies

+160 M€ Total income

+21% Market Growth

Ecosystem of Computer Vision in Catalonia 2019

+ 500 Publications in high impact index journals

+ 200 Funded projects

+ 68 Thesis

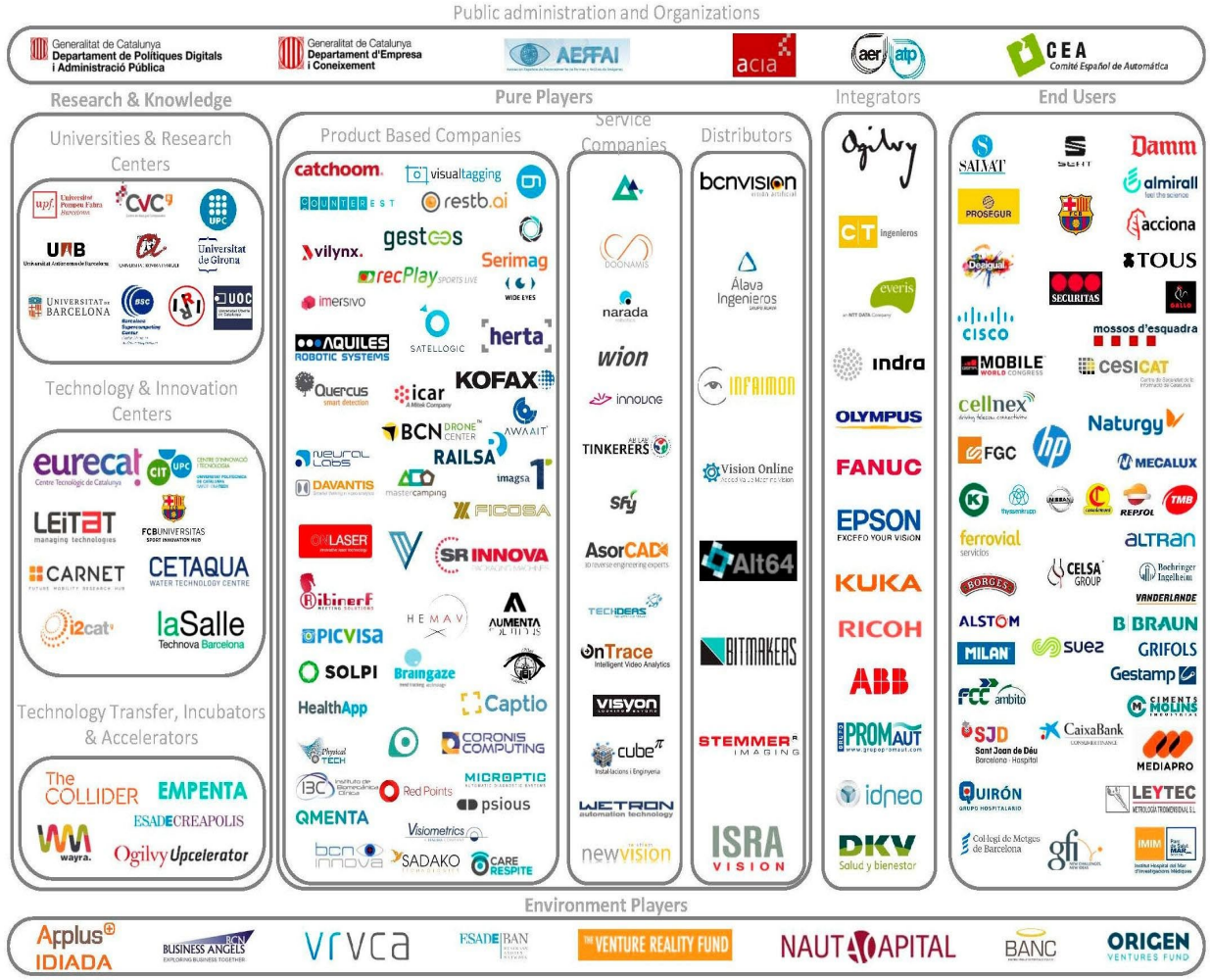
Health & Wellbeing

Security & control

Advanced processes (Industry 4.0)

Marketing/ Retail

Dealers



Research fields • Computer Vision • Machine Learning • Social Signal Processing • Affective Computing • Personality Computing	**Application domains:** • eHealth and well-being • Security • Smart cities • Leisure

Research lines
• Deep Learning
• Domain Adaptation
• Bias and fairness
• Explainability and interpretability
• Spatio-temporal modeling & video understanding
• Multi-modality, multi-view & multi-task learning
• Attention mechanisms
	Main collaborators
	• Medical researchers:
	  • Psychologists
	  • Psychiatrists
	  • Neurologists
	• National (UAB, UOC), and international universities (AAU, Berkeley, Boston)
	• Companies (Google, Microsoft, Disney, Amazon, NVIDIA, and Facebook)

13
YEARS

15+
MEMBERS

400+
TOTAL PUBLICATIONS

50+
COMPETITIVE PROJECTS

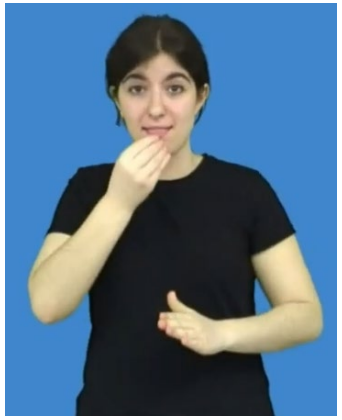
30+
TRANSFER ACTIVITIES

20+
INTERNATIONAL COMPETITIVE PROJECTS

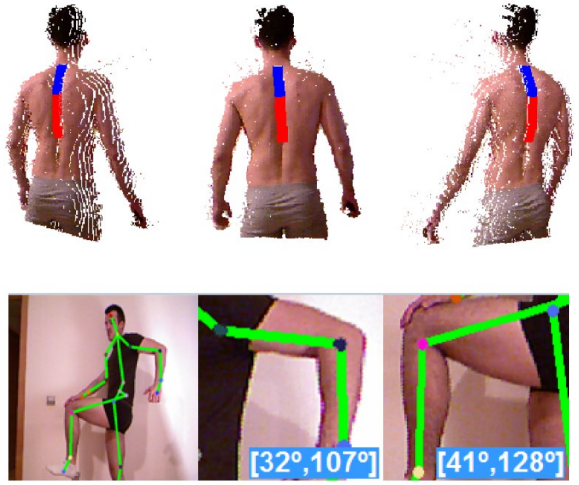
Examples of transfered technology

Visual and multimodal monitoring in neurorehabilitation

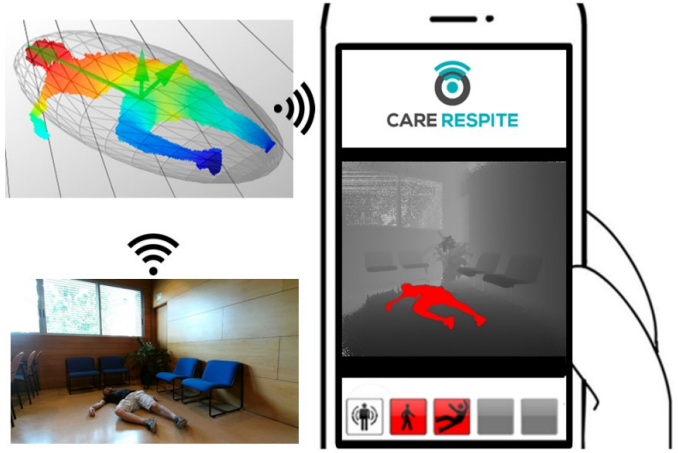
ADHD and mental health monitoring



SLR



Physiotherapy and fitness



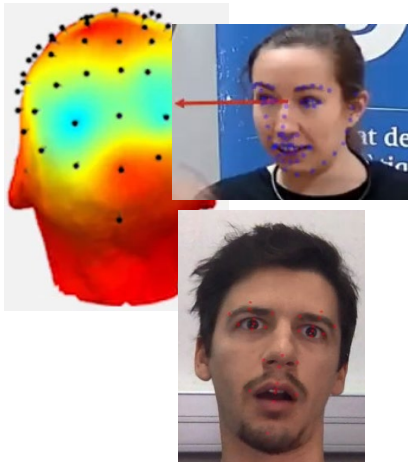
Risk event monitoring

Coaching for the elder

Virtual try ons

Overview of current research lines in LAP

FACE ANALYSIS

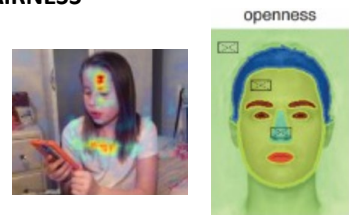


VISUAL (AND MULTIMODAL) MODELING OF HUMANS

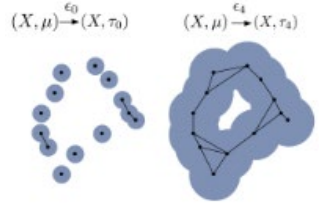
3D (& 4D) POSE, SHAPE, TEXTURE (IN 3D AND FROM 2D)

BEHAVIOR ANALYSIS

UNDERSTANDING AND EXPLAINING HUMAN BEHAVIOR (Affective & Personality Computing)
-INTERPRETABILITY & EXPLAINABILITY
-FAIRNESS



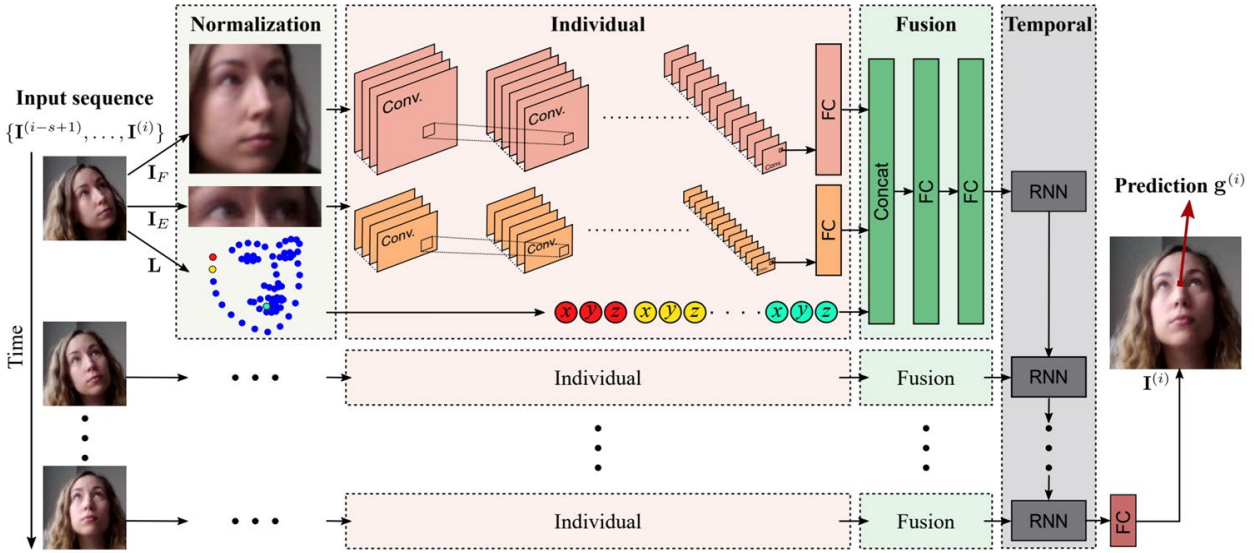
BIAS ANALYSIS VISUALIZATION

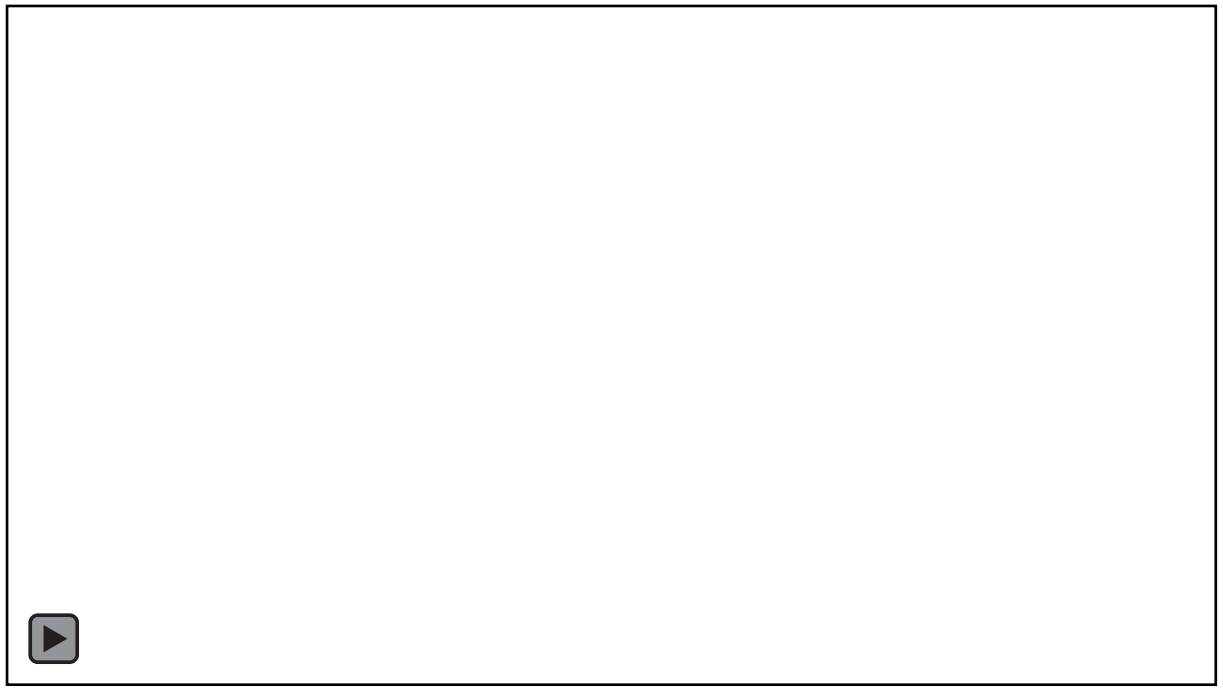


INTERPRETING AND EXPLAINING LEARNING

Face analysis

- Gaze
- Facial Action Units
- Multimodal app.

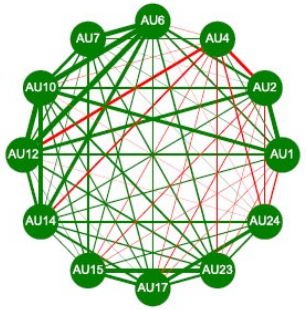




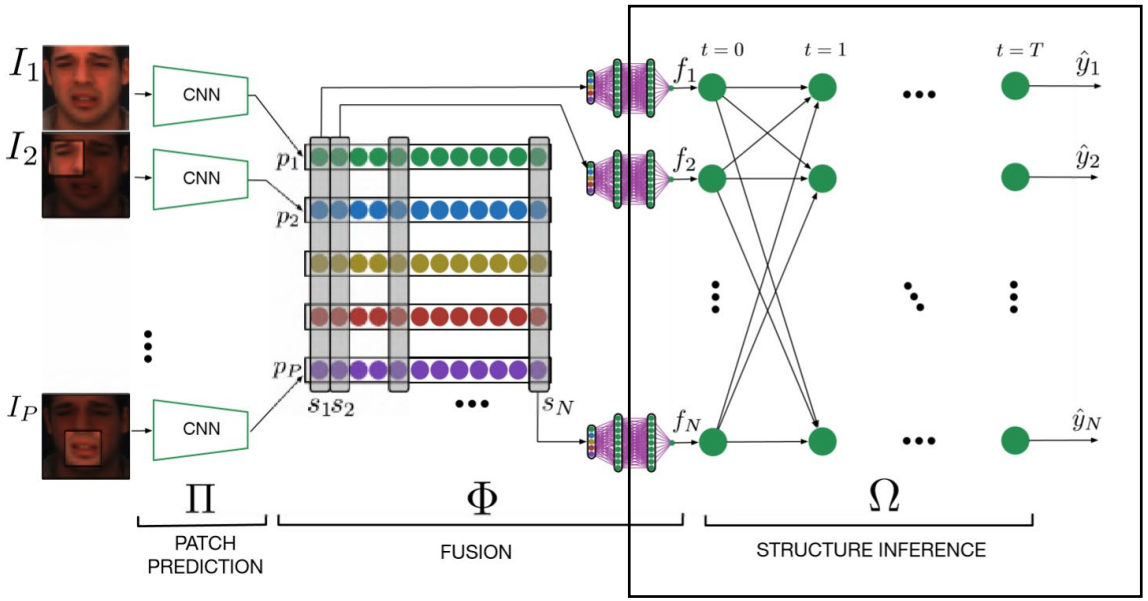
Palmero et.al. BMVC 2018

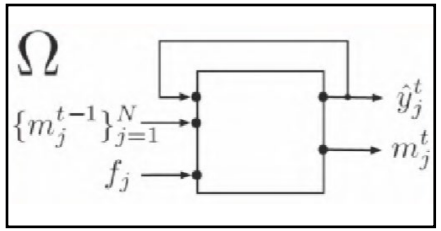
Face analysis

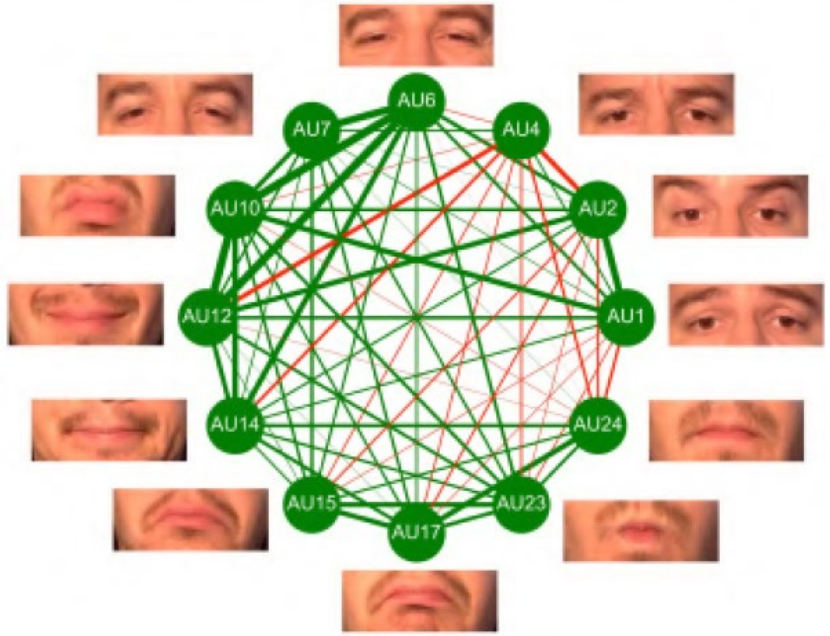
- Gaze
- Facial Action Units
- Multimodal app.



(b)



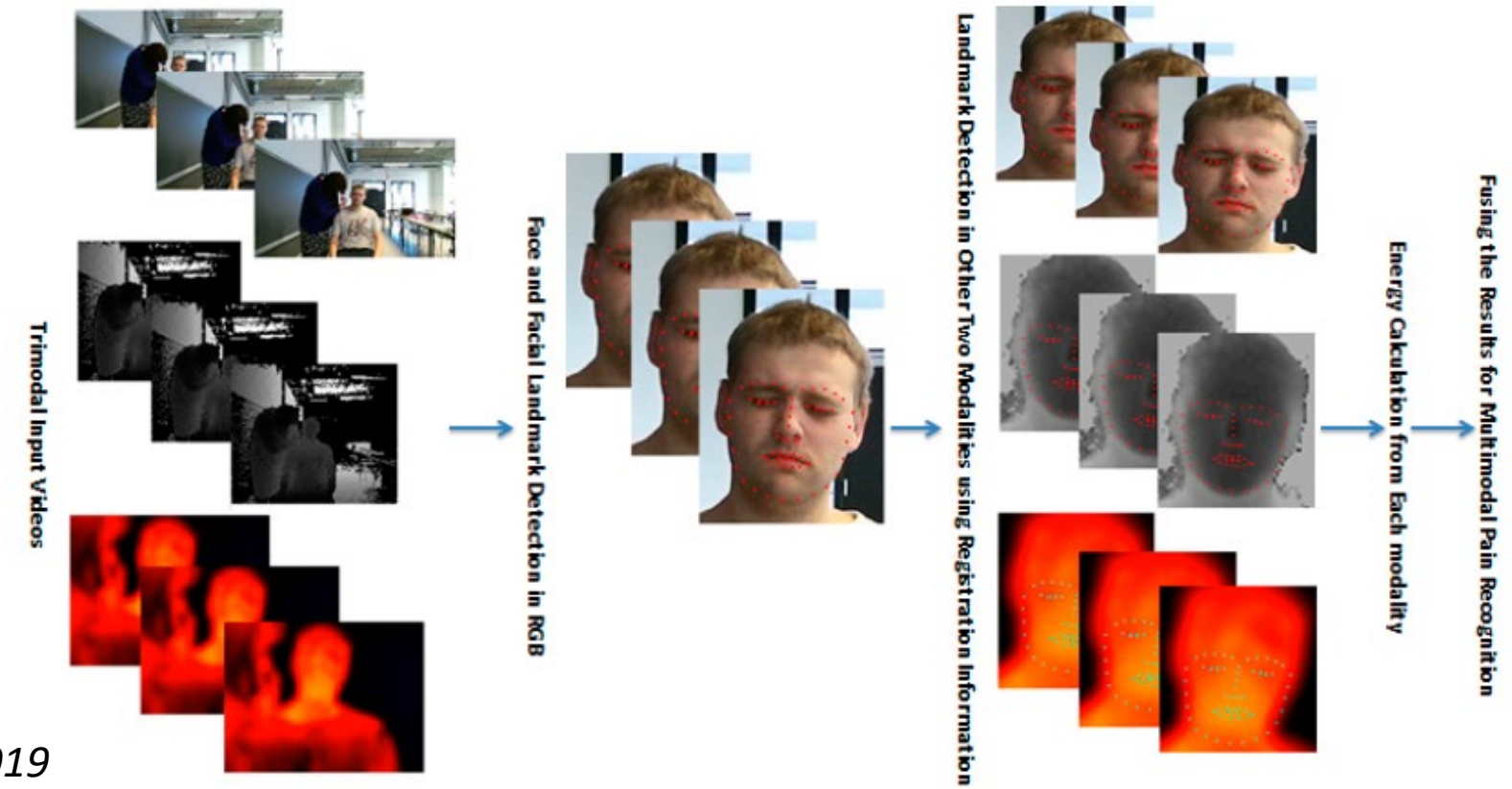




Corneanu et.al. ECCV 2018

Face analysis

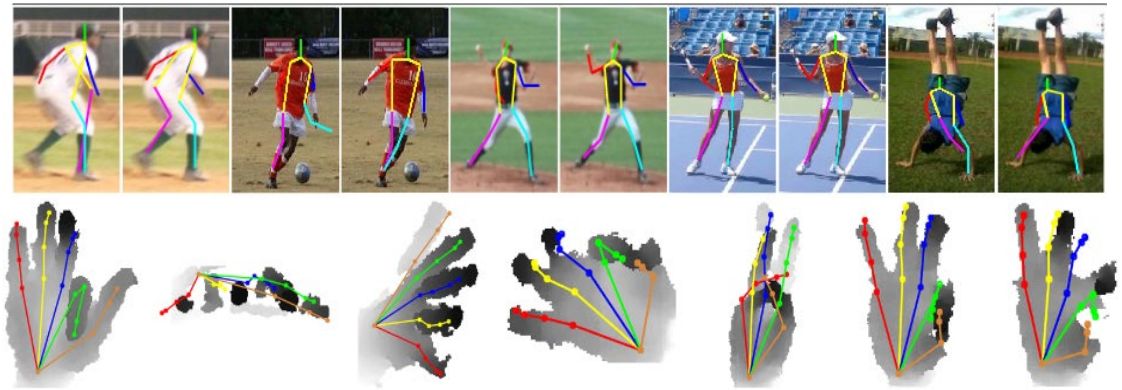
- Gaze
- Facial Action Units
- Multimodal app.

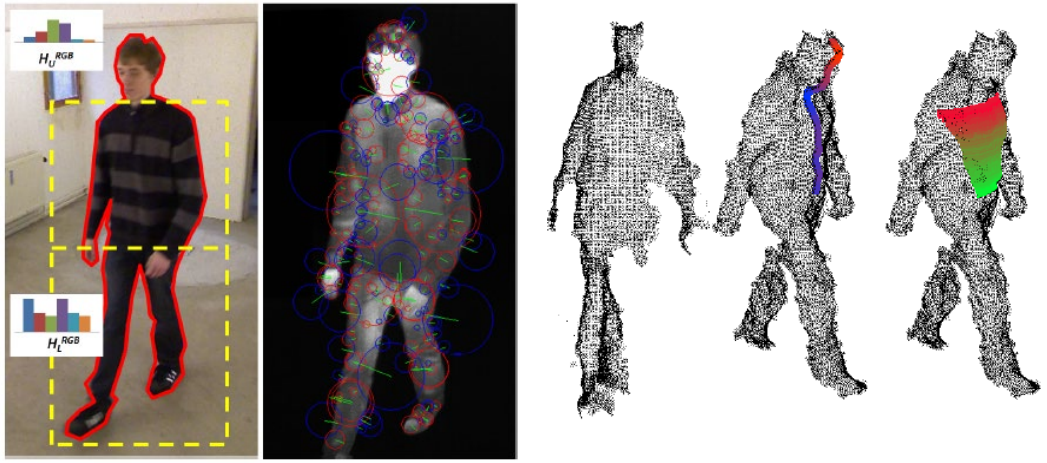


Faces and Gestures, 2019

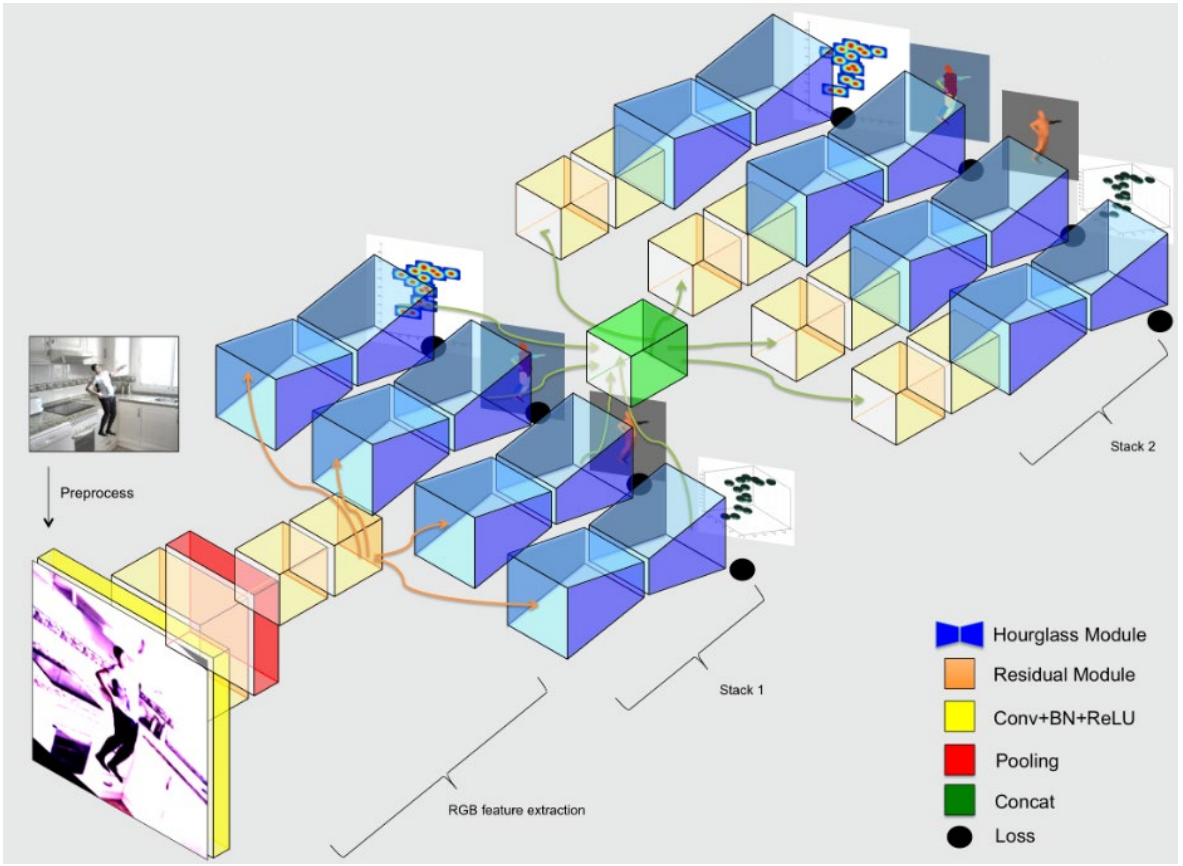
Body Analysis

- Posture and multimodal





Sanchez et.al. FG 2019



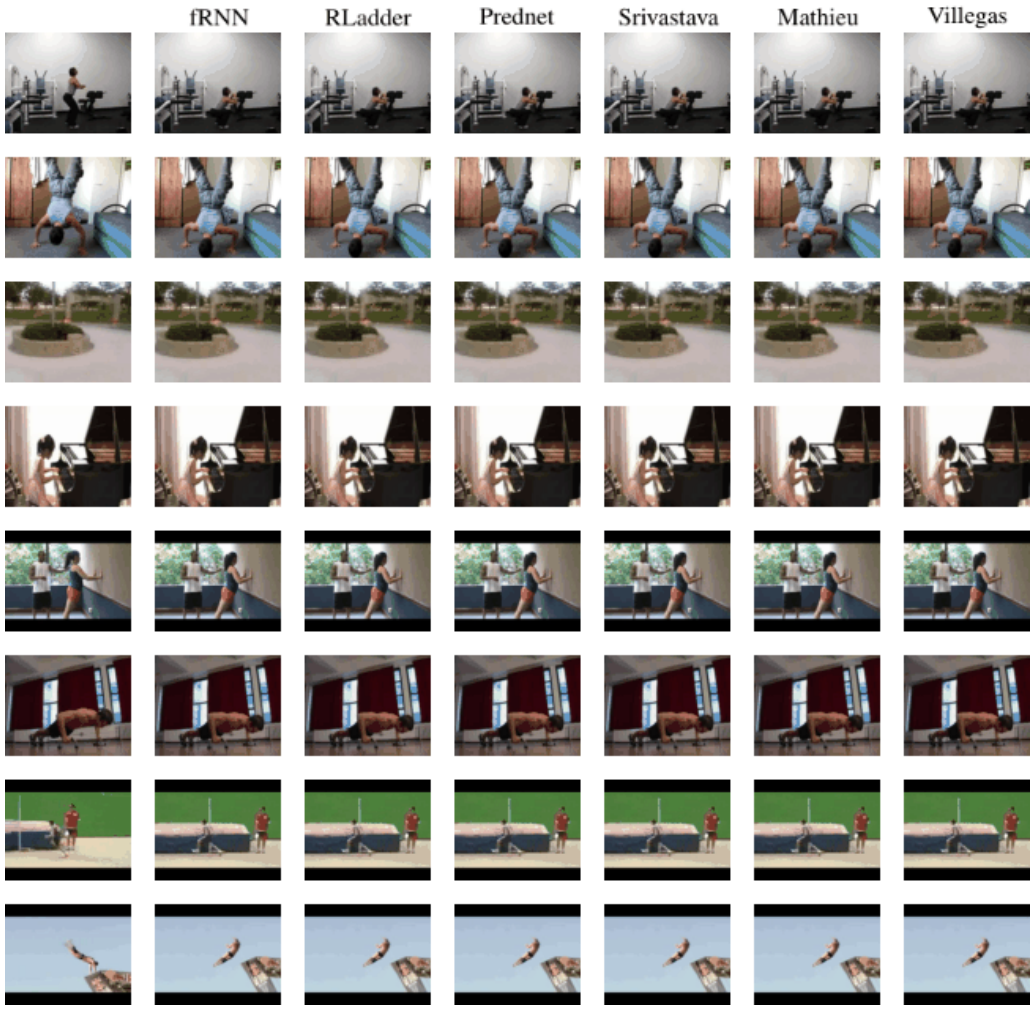
Shanxin Yuan et.al. CVPR 2018

Body Analysis

- Posture and multimodal

Madadi et.al. PR 2020
Bertiche et.al. ICCV 2021, SIGGRAPH ASIA 2021 & 2022

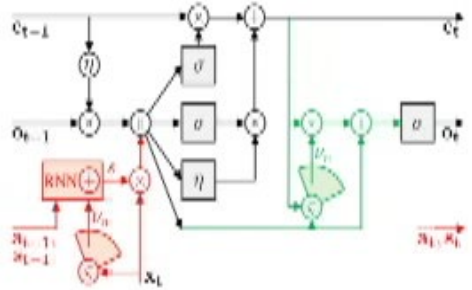
Behavior

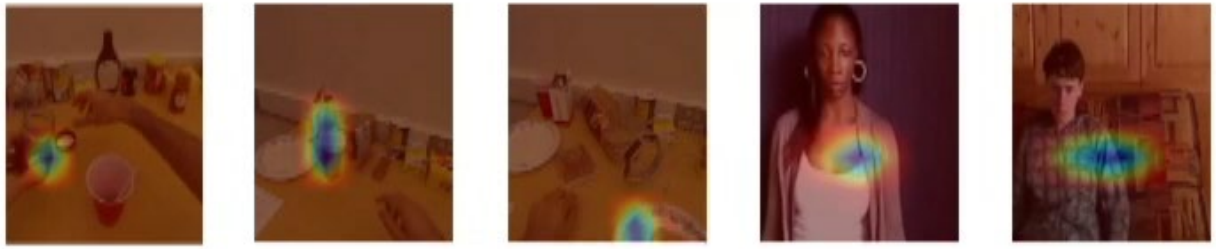


Oliu et.al. ECCV 2018

Long Short-Term Attention

State-of-the-art recognition performance

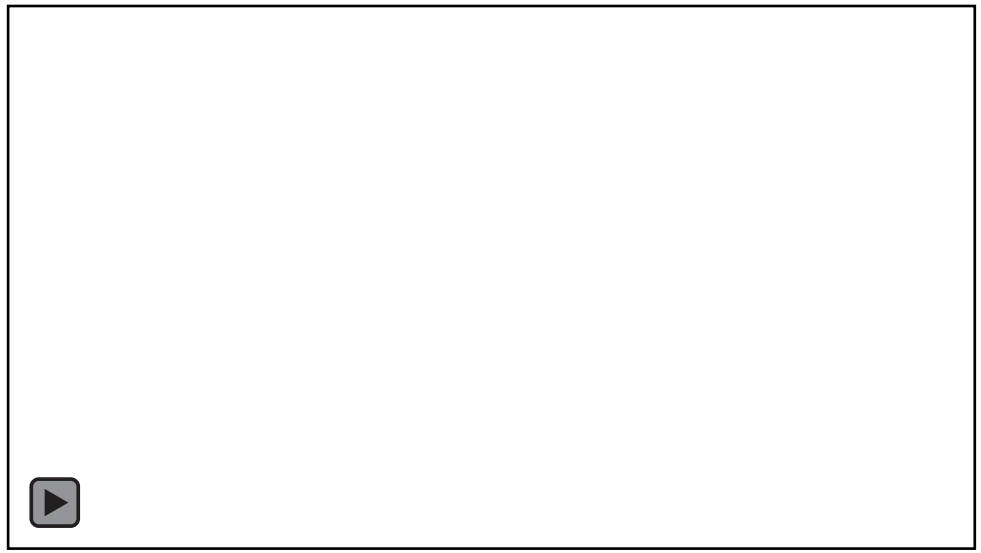




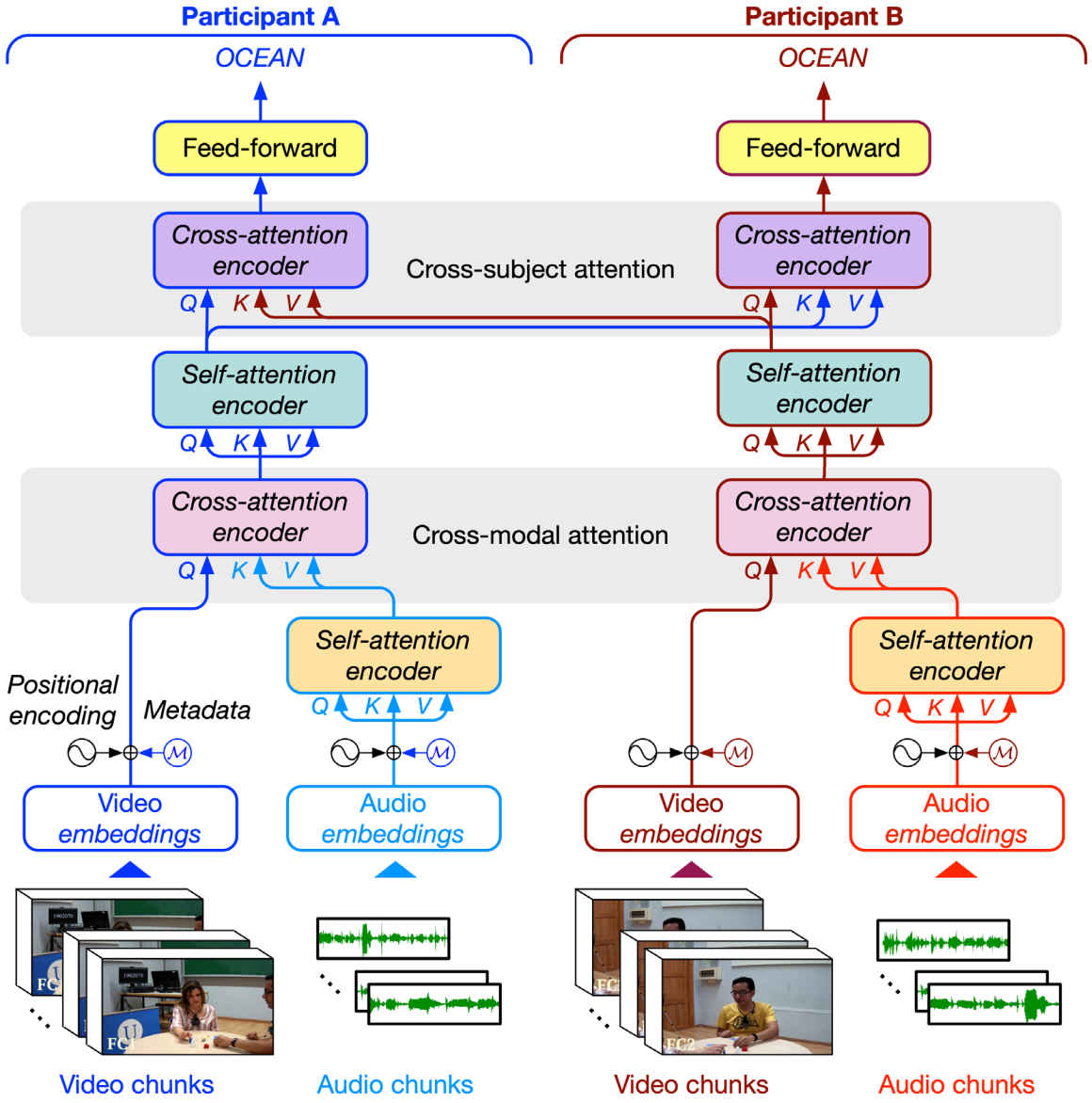
Sudhakaran et.al. CVPR 2019, CVPR 2020
EPIC Kitchens 2019, 2020, Top-3 winning solutions

Behavior

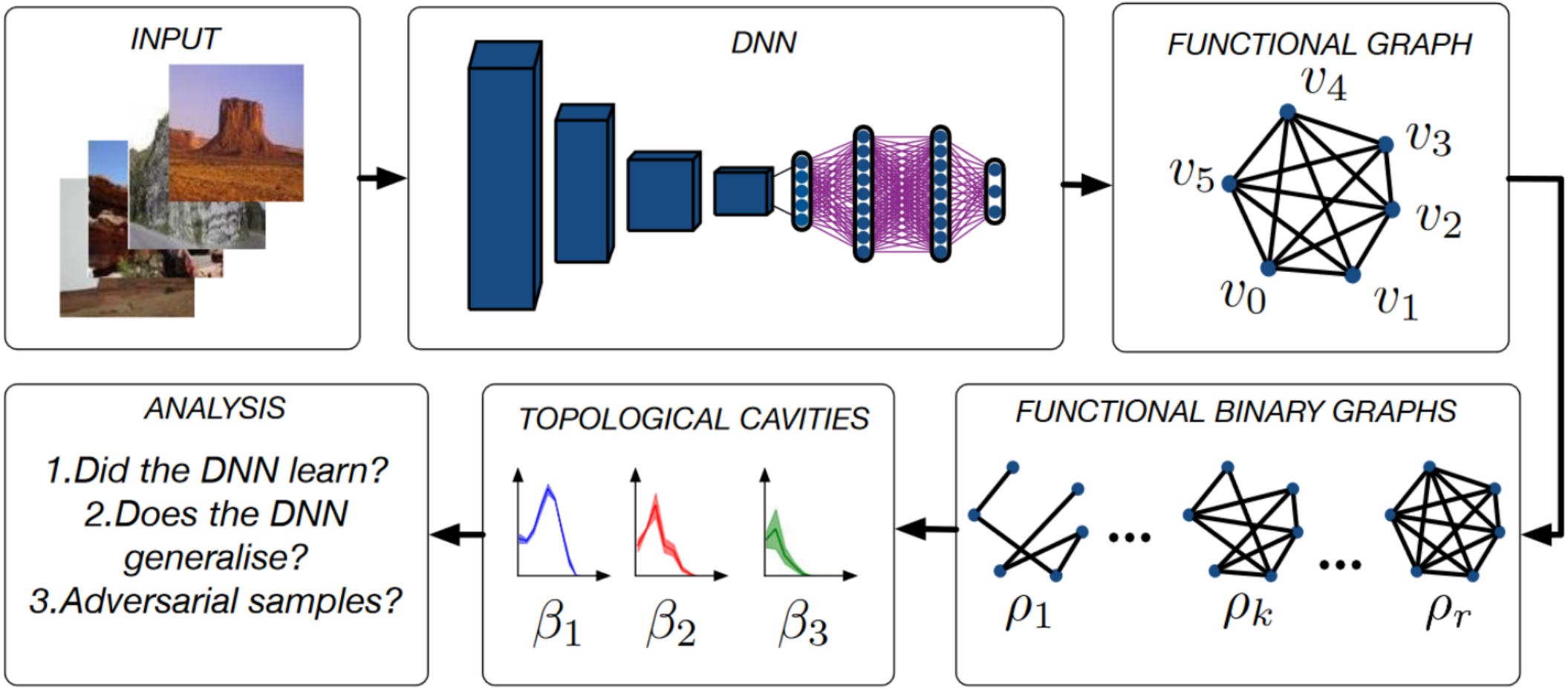
- Moving towards affective HCI and assistive technology
- Regressing self-reported personality from dyadic multimodal video data



Dyadformer
David Curto, Albert Clapés, ICCV 2021



Interpretability

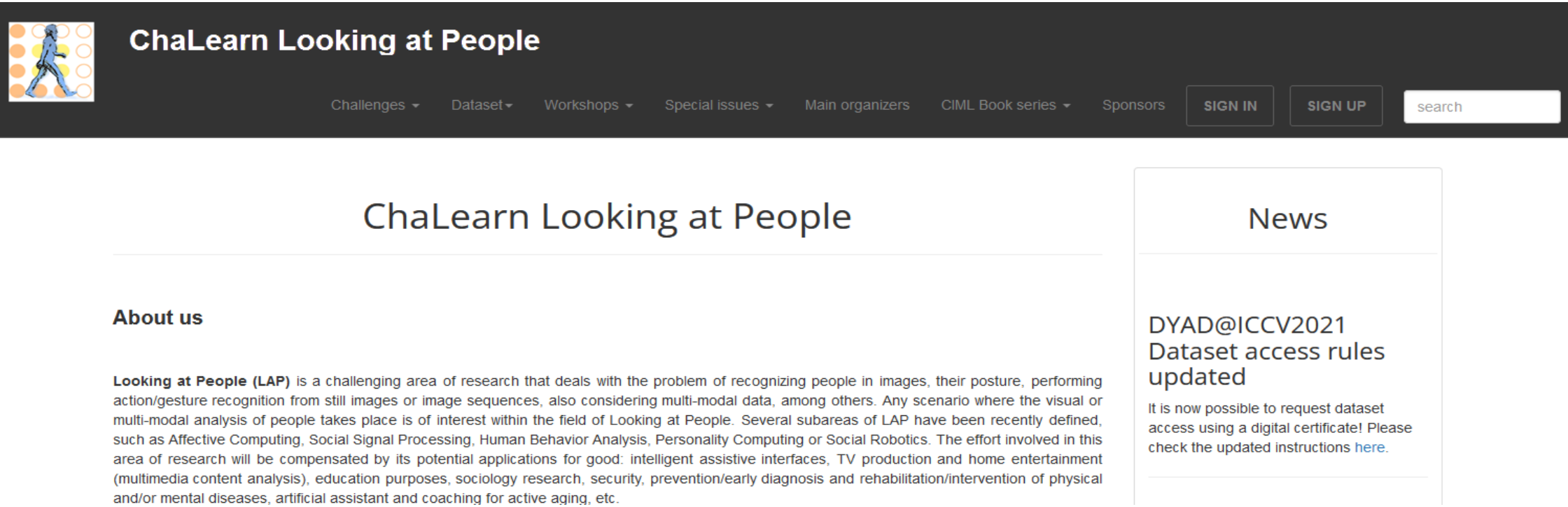


Corneanu et.al. CVPR 2019, CVPR 2020

CVPR best paper award nominee

ChaLearn

- Non-profit organization. Berkeley. We organize challenges to stimulate research in this field. The web sites of past challenges remain open for post-challenge submission as ever-going benchmarks Promoting open data, educational materials, and challenge organization. Link with ChaLearn and Codalab initiatives.
- **President: Isabelle Guyon, Google**, Université Paris-Saclay, France
- **Vice-president: Sergio Escalera,** University of Barcelona, Spain
- Involved in the **implementation of Neurips Competition track and Neurips Benchmarking track**



ChaLearn Looking at People

Challenges ▾ Dataset ▾ Workshops ▾ Special issues ▾ Main organizers CIML Book series ▾ Sponsors SIGN IN SIGN UP search

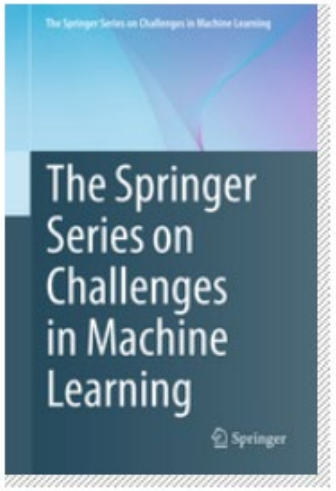
- > 25 new datasets
- > 25 organized challenges at CVPR, ICCV, ECCV, NeurIPS, …
- > 25 organized workshops at CVPR, ICCV, ECCV, NeurIPS, …
- > 10 organized Special Issues to related workshop/challenge topics

Main sponsors

Microsoft Google amazon

IAPR

NVIDIA Disney Research facebook



The Springer Series on Challenges in Machine Learning

Series Editors: **Escalante**, Hugo Jair, **Guyon**, Isabelle, **Escalera**, Sergio

ISSN: 2520-131X

ABOUT THIS SERIES

The books in this innovative series collect papers written in the context of successful competitions in machine learning. They also include analyses of the challenges, tutorial material, dataset descriptions, and pointers to data and software. Together with the websites of the challenge competitions, they offer a complete teaching toolkit and a valuable resource for engineers and scientists.

~10 edited volumes up to date

CodaLab

Accelerating reproducible computational research.

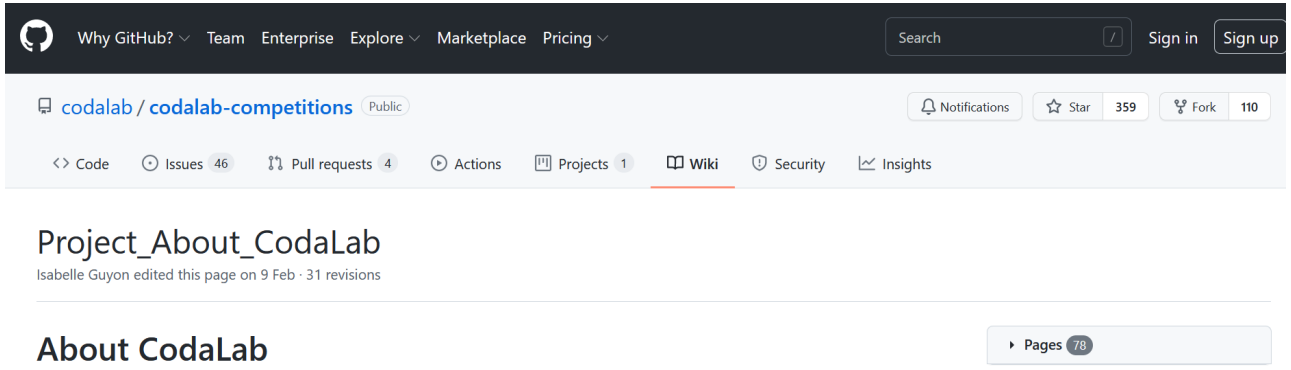
The CodaLab Team

Percy Liang is an assistant professor of Computer Science at Stanford University. His primary research areas are machine learning and natural language processing. He leads the development of CodaLab in close collaboration with Microsoft Research and the rest of the community.

Isabelle Guyon is full professor at UPSud University Paris-Saclay and president of ChaLearn a non-profit organization dedicated to running machine learning competitions. Her research interested include automatic machine learning, transfer learning, and causal discovery. Isabelle served as an advisor in the development of the CodaLab competition platform and pioneered the implementation of several challenges on Codalab.

Evelyne Viegas is a Director at Microsoft Research responsible for the outreach artificial intelligence program. She leads the CodaLab project working in collaboration with Isabelle Guyon, Percy Liang and the machine learning and artificial intelligence communities.

Sergio Escalera is adjunct professor at Universitat Oberta de Catalunya, Aalborg University, and Dalhousie University and a member of the Visual and Computational Learning consolidated research group of Catalonia and a member of the Computer Vision Center at UAB. He is series editor of The Springer Series on Challenges in Machine Learning. He is Editor-in-Chief of American Journal of Intelligent Systems and editorial board member of more than 5 international journals. He is vice-president of ChaLearn Challenges in Machine Learning, leading ChaLearn Looking at People events.



- Preferred open source competition platform by the community.
- More than thousand organized competitions

Stanford University

CHALEARN

Microsoft

université PARIS-SACLAY

Research interests

- Main interest in image, video, and multimodal data analysis for LAP
- Promoting **explainability** and **interpretability** for **transparency**
- **Bias** detection and mitigation for **fairness, context, personalization**
- **Multimodal** learning and with noise and asynchronous data
- **Self-supervised learning** to manage huge amount of data and reduce the need of annotated data
- **Domain adaptation**
- **Uncertainty estimation and human in the loop**
- **LAP challenges** design and implementation
- **Real applications for good!**

Dr. Sergio Escalera, www.sergioescalera.com, sergio@maia.ub.es

Video Transformers

Javier Selva Castelló

javierselva.github.io

UNIVERSITAT DE BARCELONA

PBA

BREAKING NEW GROUND
AALBORG UNIVERSITY

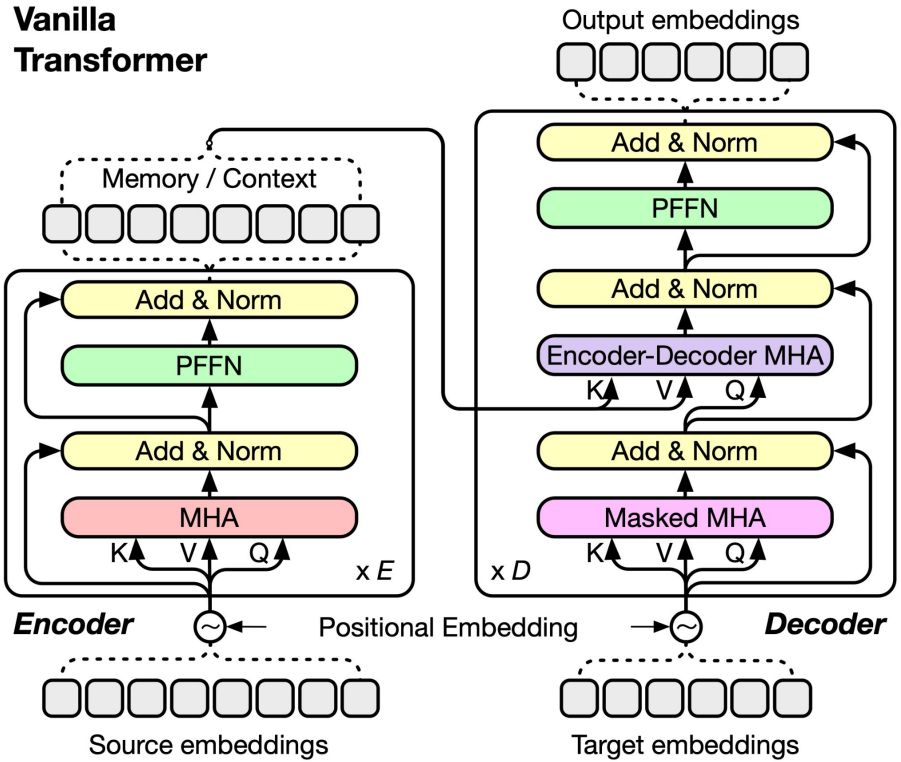
Outline

- Transformers.
- **Video pre-processing.**
- **Architectures for video.**
- **Training Video Transformers.**
- Conclusions.

Transformers

Transformer Overview

- Sequence modeling (tokens).
- Multi-head Attention.
  - Non-local interactions.
- Position-wise FeedForward.
  - Evolve token representations.

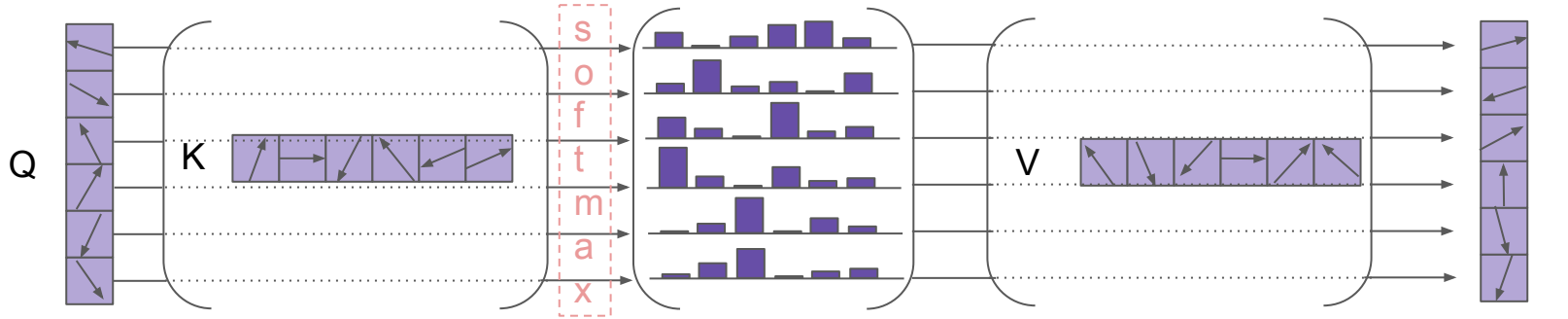


Vaswani, A. et al. *"Attention is all you need."* In NeurIPS (2017).

Non-local Token Mixing

Queries, Keys and Values are linear transformations of the input.

$$\mathbf{Q} = \mathbf{X}\mathbf{W}_Q, \mathbf{K} = \mathbf{X}\mathbf{W}_\mathrm{K}, \mathbf{V} = \mathbf{X}\mathbf{W}_\mathrm{V}$$



$$a_{ij} = \mathrm{softmax}(\frac{\mathbf{q}_i \mathbf{k}_j^\top}{\sqrt{d_k}})$$

$$y_i = \sum_{j=0}^{T} v_i a_{ij}$$

Vaswani, A. et al. *"Attention is all you need."* In NeurIPS (2017).

Long-term Modeling for Video

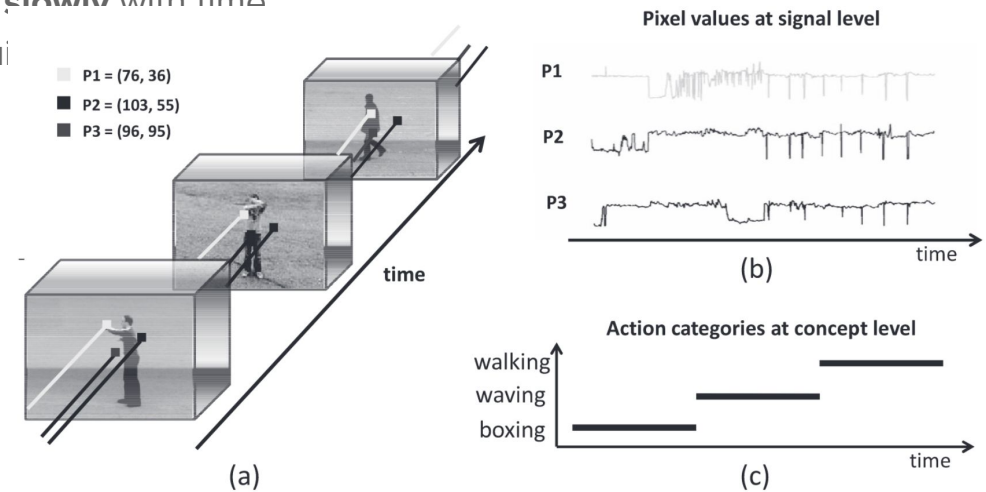
- Video is inherently a sequence.
- Opposed to RNNs or CNNs, allows to model **long-term interactions** in a single operation.
- Benefit motion modeling.

Wang, X. et al. *"Non-local neural networks."* In CVPR (2018).

Challenges

Video:

- Highly **dimensional**.
- Highly **redundant:**
  - Appearance-based semantics vary **slowly** with time
  - Capturing fine-grained motion requi



[1] **Feichtenhofer, C.** et al. "*SlowFast Networks for Video Recognition.*" In ICCV (2019).
[2] **Zhang, Z.** et al. "*Slow Feature Analysis for Human Action Recognition.*" In TPAMI (2012).

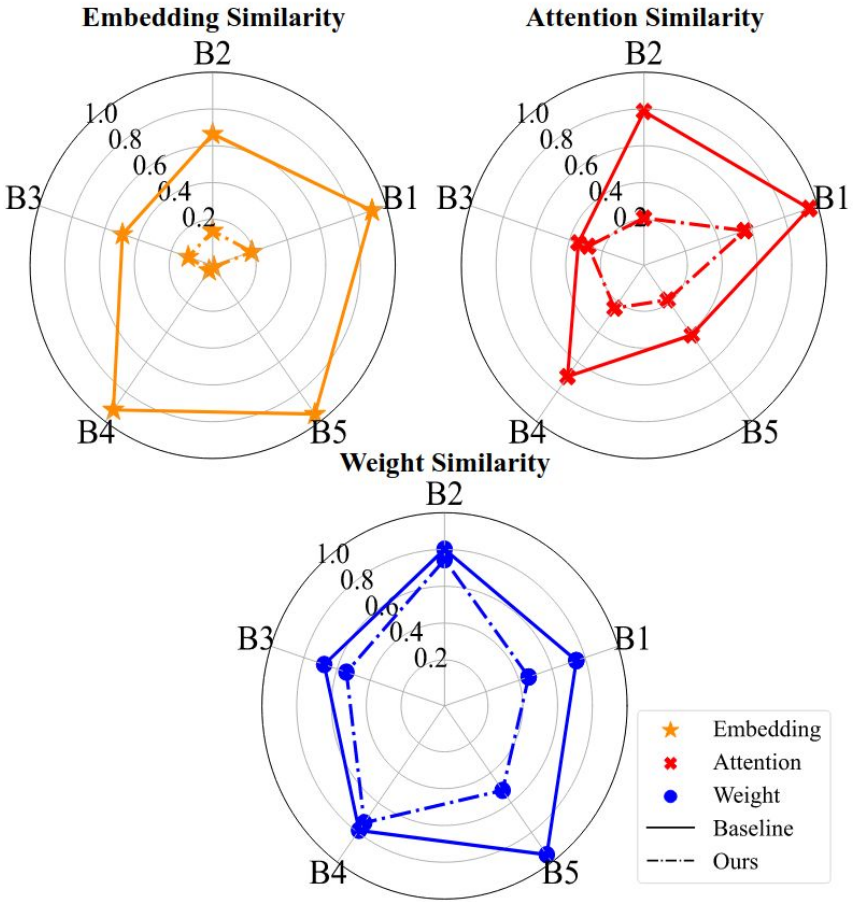
Challenges

Video:

- Highly **dimensional**.
- Highly **redundant.**

Transformers:

- **Token smoothness**



Chen, T. et al. "*The Principle of Diversity: Training Stronger Vision Transformers Calls for Reducing All Levels of Redundancy.*" In CVPR (2022).

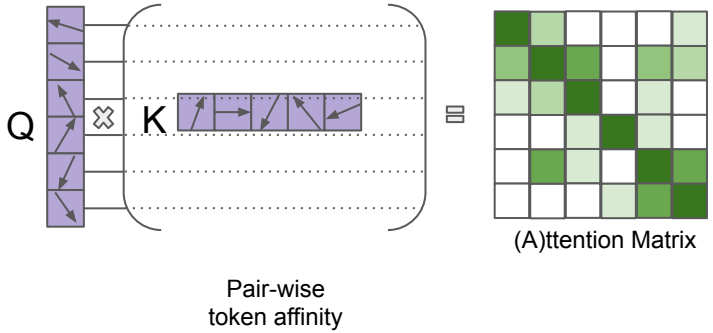
Challenges

Video:

- Highly **dimensional**.
- Highly **redundant.**

Transformers:

- **Token smoothness**
- Quadratic complexity **O(N^2)**



For scale, *25 frames* at *256 x 256 resolution*, tokenized as *1 x 16 x 16* sized patches, results in ~**41M** elements in A.

Challenges

Video:

- Highly **dimensional**.
- Highly **redundant.**

Transformers:

- **Token smoothness**
- Quadratic complexity **$O(N^2)$**
- Lack of **inductive biases**
  - Versatile architecture.
  - Requires large datasets!

Dosovitskiy, A. et al. *"An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale."* In ICLR (2021).

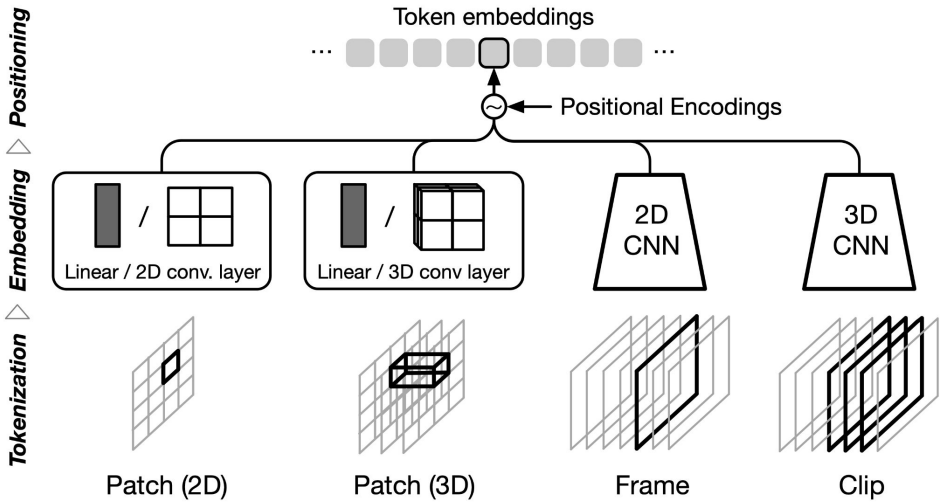
Video pre-processing

Pre-processing

Tokenization: Turned into a sequence of elements.

Embedding: Map raw pixel values into continuous vector representation.

Positional information: Transformers are agnostic to position.



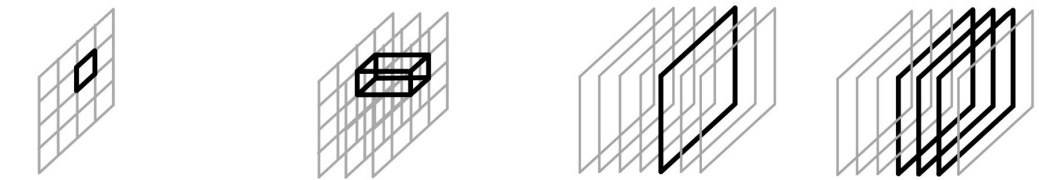
Tokenization

Defines **granularity** at which interactions can be learned.

More token receptive field, less fine-grained feature modeling.

Defines **complexity** of the self-attention computation.

Smaller tokens = More tokens = More compute!



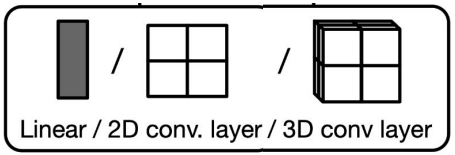
Embeddings

Minimal Embeddings: Single layer

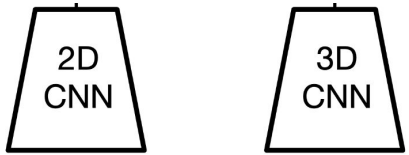
Large Embedding Networks: Complete SOTA architectures

Dimensionality reduction.

CNNs will alleviate Transformer training.



Minimal Embeddings



Large Embedding Networks

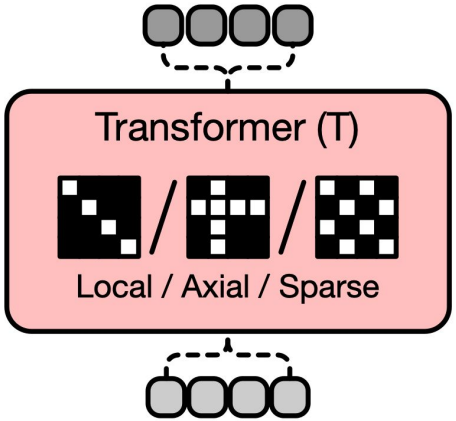
Architectural designs

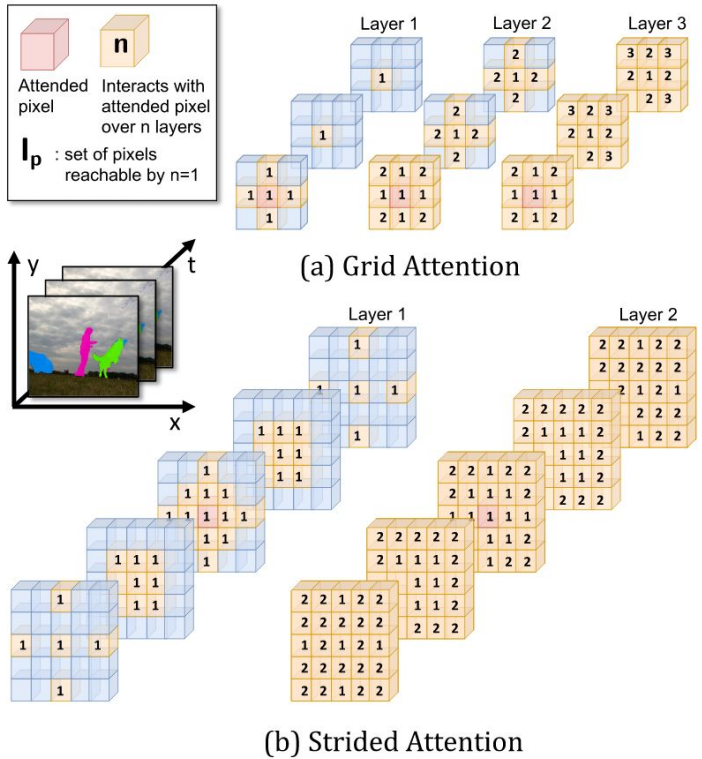
Restricted Approaches

General techniques to reduce **complexity**!

Limit scope of attention operation.

Depend on **stacking** to account for complete input.





Duke, B. et al. "*SSTVOS: Sparse Spatiotemporal Transformers for Video Object Segmentation*". In CVPR (2021).

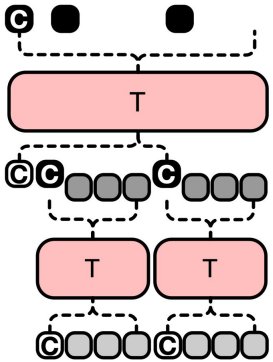
Aggregation

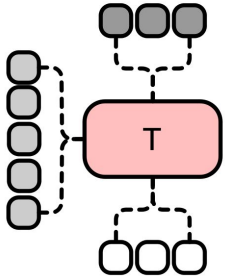
Leverage a **reduced set of tokens**.

May help reduce **redundancy**.

Needs to be done carefully:
- May lose ability to model some dependencies later on!

Hierarchy **vs.** **Query-driven Compression**

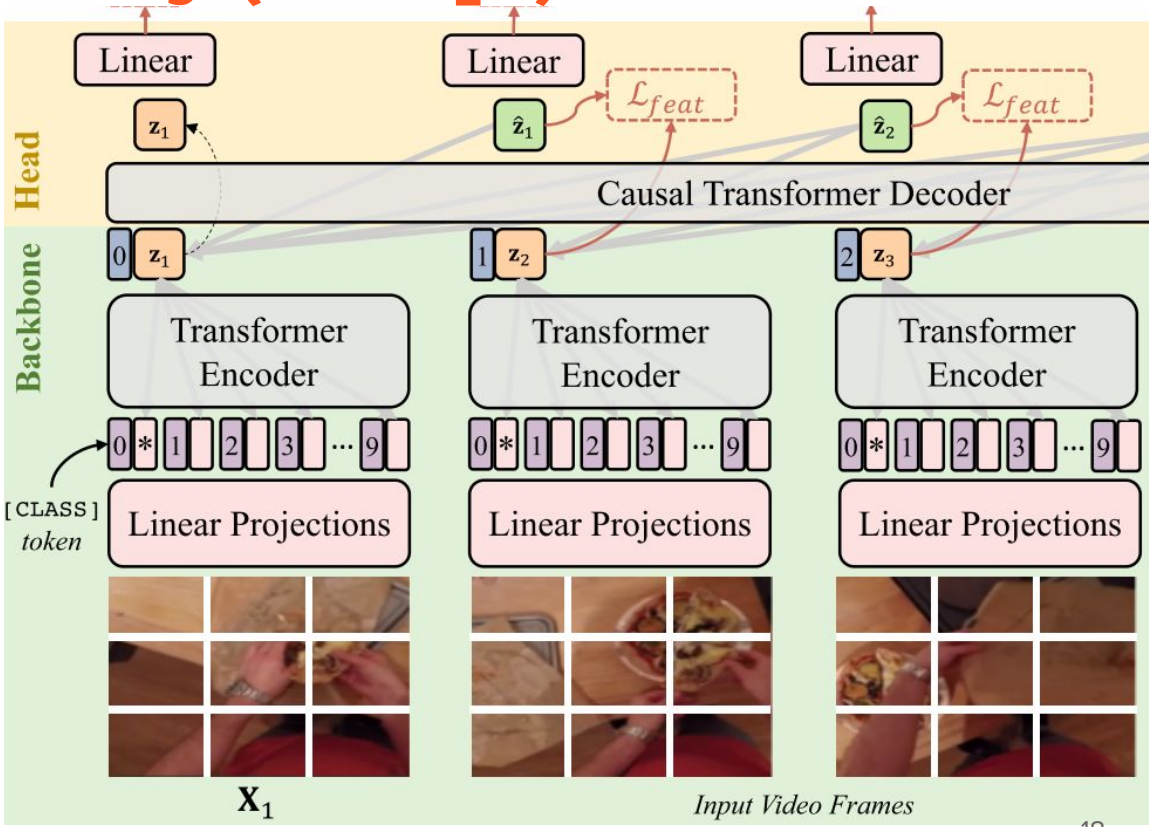




Aggregation: Hierarchy (Abrupt)

Abrupt approaches:

- **Early** aggregation:
  miss fine-grained motion
- **Local** neighborhoods
- **CLS** token

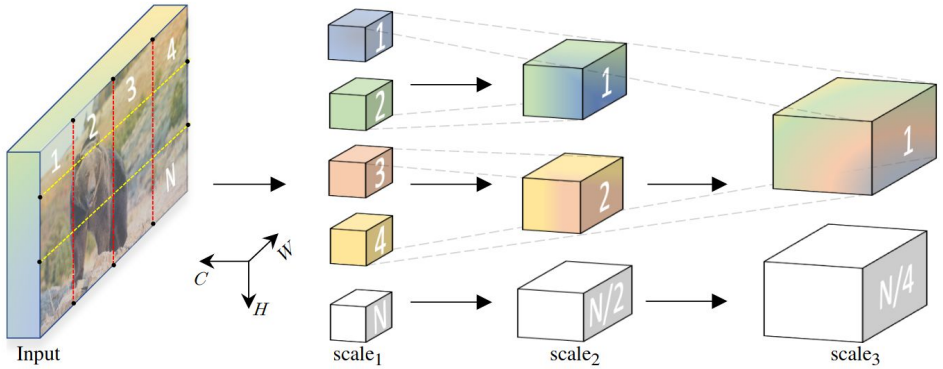


Girdhar, R. et al. "*Anticipative Video Transformer*". In ICCV (2021).

Aggregation: Hierarchy (Progressive)

Progressive approaches:

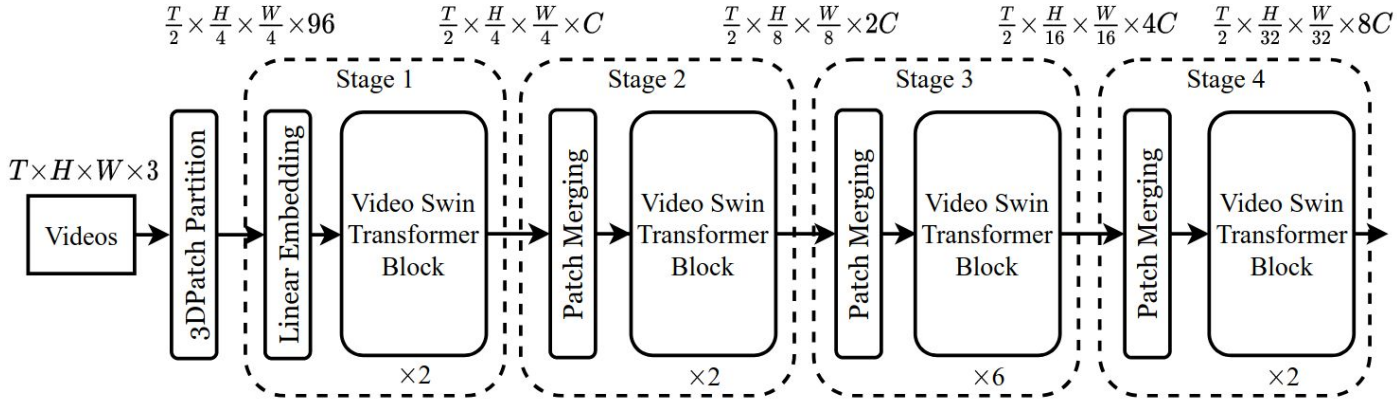
- **Gradual** aggregation.
- **Global** interactions.
- Local learned **pooling.**



Fan, H. et al. "*Multiscale Vision Transformers.*" In ICCV (2021).

Aggregation: Hierarchy (Progressive)

- Maintain **temporal fidelity**.
- Feature **dimensionality increase** with depth.
- Best **performance**/cost ratio!

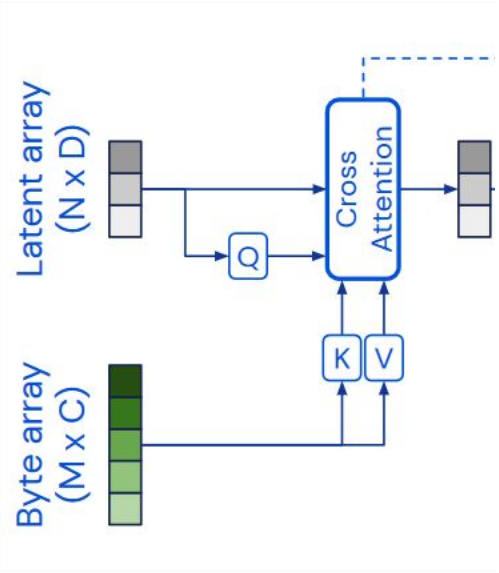


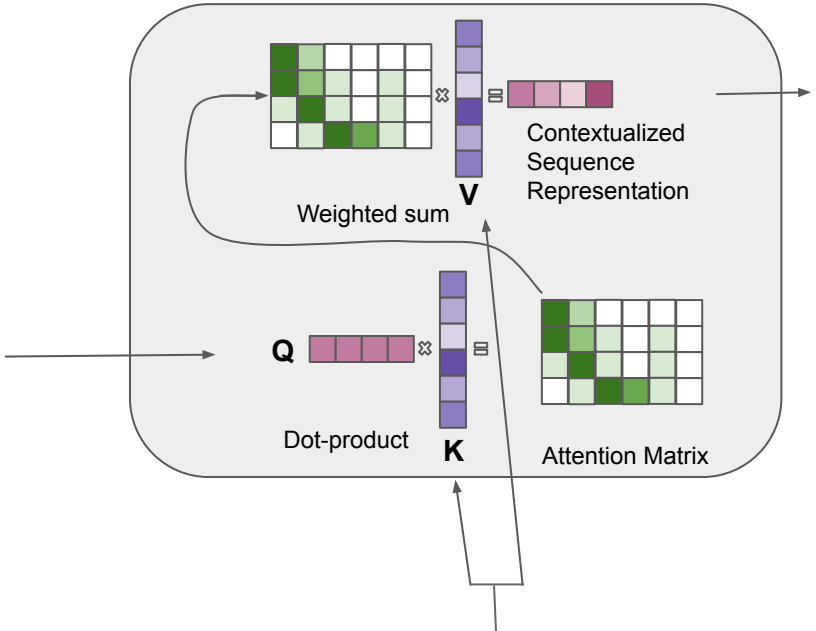
Liu, Z. et al. "*Video Swin Transformer.*" In CVPR (2022).

Aggregation: Query-driven Compression

General technique, allows **many** different **applications**.

Compress full sequence into **fewer tokens**. **Reduce complexity**!

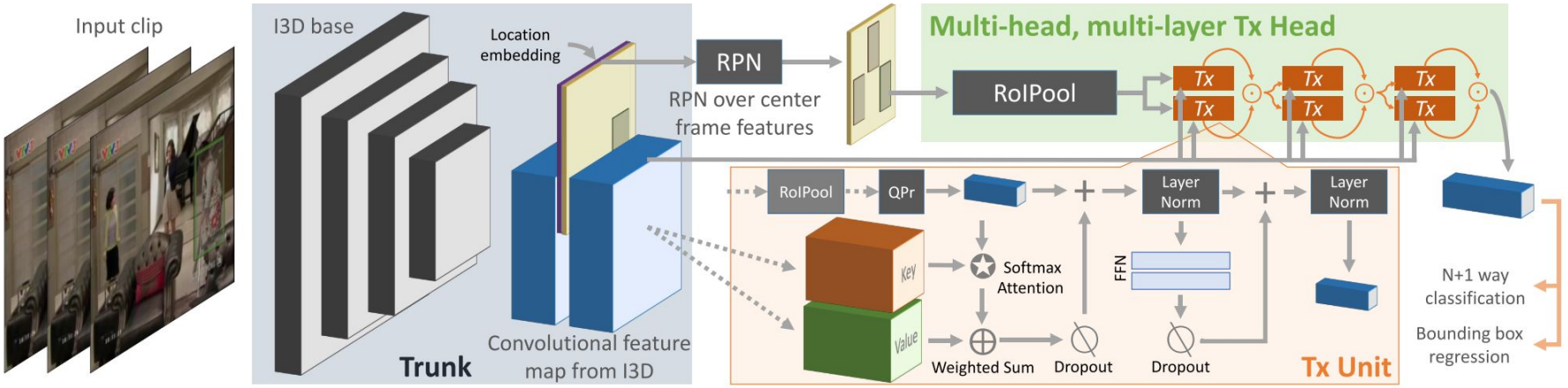




Jaegle, A. et al. "*Perceiver: General Perception with Iterative Attention.*" In ICML (2021).

Aggregation: Query-driven Compression

Data driven

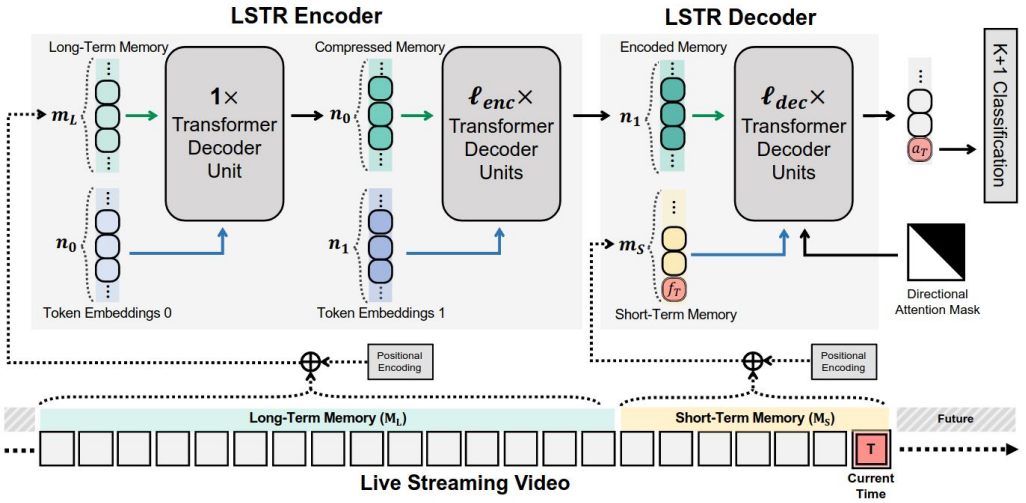


Girdhar, R. et al. "*Video Action Transformer Network Rohit.*" In CVPR (2019).

Long-term modeling

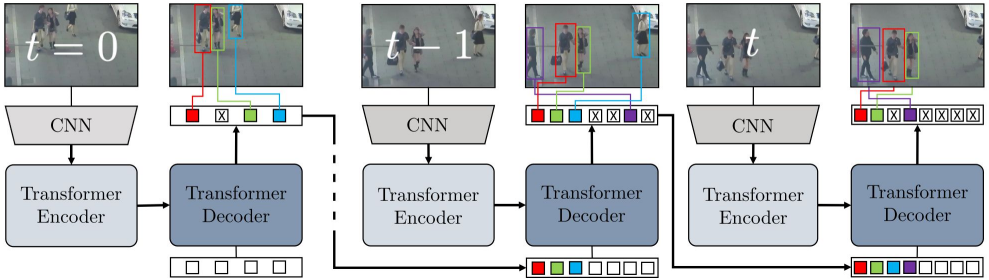
Memory

- Stores **several past observations**.
- Can be accessed:
  - Through Cross-Attention.
  - By concatenation.
- Generally **compressed**:
  - At storing time.
  - Sparse sampling.
  - Query-driven Compression.
- **Discard** old elements:
  - FIFO.
  - Relevance (attention).



Wu, C. et al. "*MeMViT: Memory-Augmented Multiscale Vision Transformer for Efficient Long-Term Video Recognition.*" In CVPR (2022).

Long-term modeling



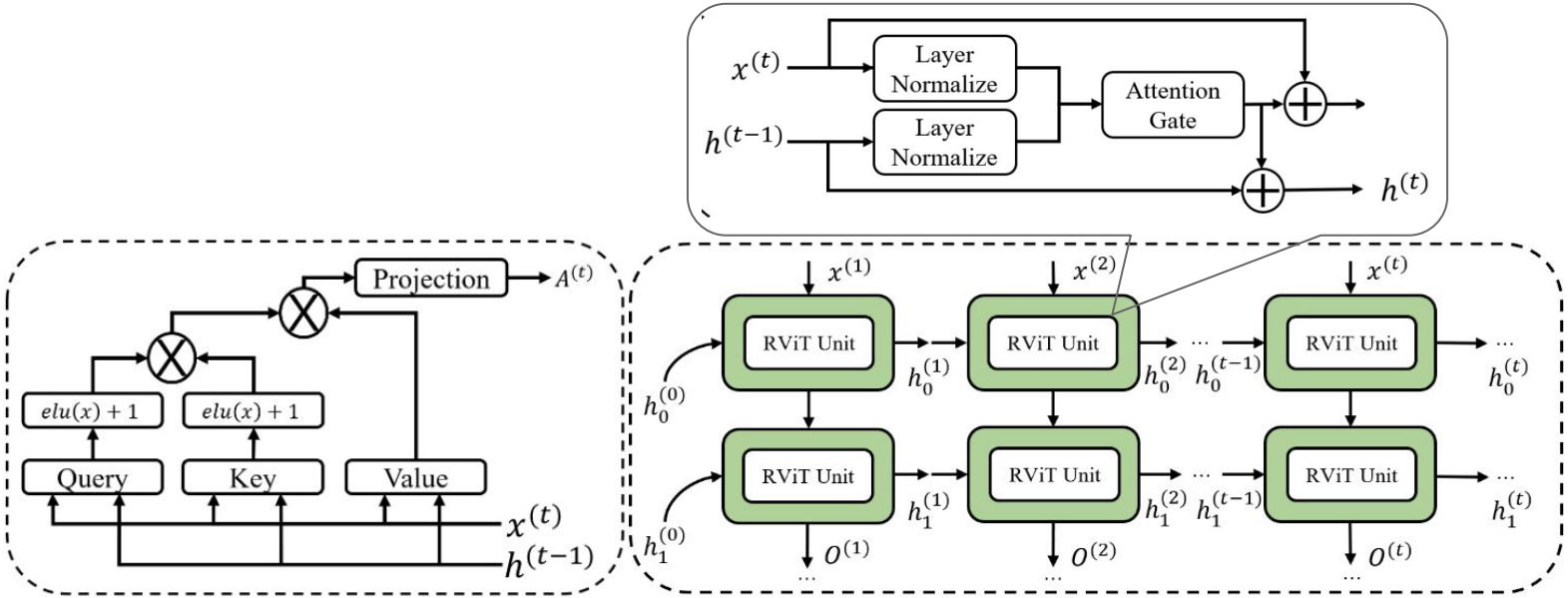
Images source: https://github.com/timmeinhardt/trackformer

Recurrence

- **Collapses** past observations in a single **recurrent state**.
- Send information to next time-step:
  - Within the Attention Operation.
  - Forwarding Output.
- Used when forwarding **local temporal information** is enough.
- Maintain access to **full resolution** at each new observation.

Meinhardt, T. et al. *"TrackFormer: Multi-Object Tracking with Transformers."* In CVPR (2022).

Long-term modeling: Recurrence



Attention Gate
(Aggregates **current input with recurrent state**)

Traditional Recurrence Unrolling
(Replaces **RNN Units** with a **Transformer**)

Yang, J. et al. "*Recurring the Transformer for Video Action Recognition.*" In CVPR (2022).

Training the Transformer

Training Regime

Solving lack of **inductive biases**:

Large datasets.

Self-supervised Learning.

Computational limitations!

End-to-end training of embedding layers.

Large batches, videos, architecture...

Multiple training stages.

- Minimal Embedding Networks:
  - **End-to-end**
  - Require **large-datasets** or SSL.
- End-to-end with Large Emb. Net.:
  - **Small** Transformer.
  - **Efficient** designs.
  - CNN-Transformer **tandem.**
- Frozen embeddings:
  - Most common.
  - Transformer boost!
  - Limited by the pre-trained features.
- Image vs Video pre-training
  - Appearance variability
  - Motion modeling.

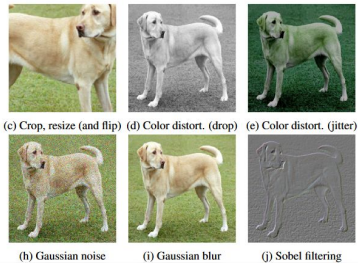
Self-Supervised Learning (SSL)

Allows to leverage un-annotated data.

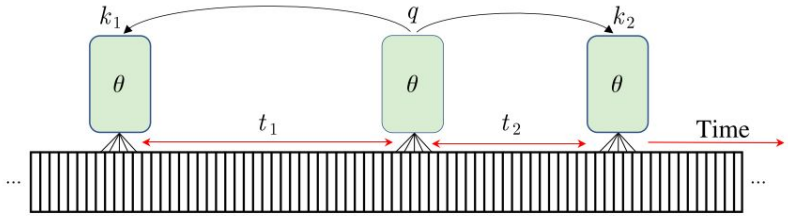
Provides data-specific biases.

Traditional **Instance-based Learning**.

- Spatial data **augmentation**.
- **Align** views through InfoNCE [1].
- Learn:
  - **invariance** to perturbations.
  - instance-based representations.



$$\mathcal{L} = -\sum_{i,k}\left[\log \frac{\exp(\hat{z}_{i,k}^{\top}\cdot z_{i,k})}{\sum_{j,m}\exp(\hat{z}_{i,k}^{\top}\cdot z_{j,m})}\right]$$



May learn **invariance to temporal deformations**!

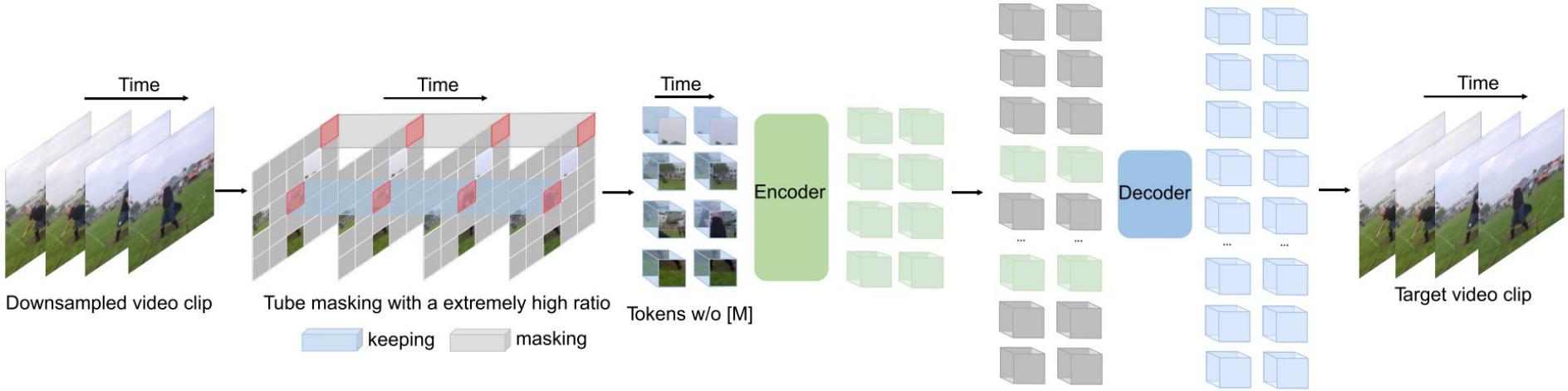
[1] **Oord, A.** et al. "*Representation Learning with Contrastive Predictive Coding.*" In ArXiv (2018).
[2] **Chen, T.** et al. "*A Simple Framework for Contrastive Learning of Visual Representations*". In ICML (2020).
[3] **Purushwalkam, S.** et al. "*Demystifying Contrastive Self-Supervised Learning: Invariances, Augmentations and Dataset Biases.*" In NeurIPS (2019).
[4] **Feichtenhofer, C.** et al. "*A Large-Scale Study on Unsupervised Spatiotemporal Representation Learning.*" In CVPR (2021).

SSL: Masked Token Modeling

Force to leverage **global appearance and motion semantics**...

... to solve **local token-dependent** prediction tasks.

Crucial: **High masking ratios**, mask entire **tubes**!



Tong, Z. et al. "*VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training.*" In NeurIPS (2022).

Conclusions

Conclusions

Complexity:

Reduce appearance **redundancy.**

Careful! Need to keep **fine-grained motion.**

Progressive aggregation.

Memory models (variable length!).

MTM exploits Transformer abilities.


ANY
QUESTIONS
?