Video Understanding of Complex Human Activities

Michal Balazia

bichal.balazia@inria.fr

INRIA Sophia Antipolis – **STARS team**

Nice University Hospital - **CoBTeK**,

INRIA-STARS Research Team

Video Understanding for Human Behavior Analysis

Objectives:

- to measure objectively human behaviors by recognizing their everyday activities, their emotion, eating habits and lifestyle,
- to improve and optimize the quality of life of people suffering from behavior disorders.

Method:

- Designing vision systems for the recognition of human activities
- Human behavior can be modeled by learning from a large number of data from a variety of sensors.

Applications :

- Safety & Health & Well-being (CoBTeK : Behavior Disorder)
- and many other applications (e.g. Sport)

3iA Côte d'Azur
Interdisciplinary Institute for Artificial Intelligence

Inria informatics mathematics

People Tracking on MOT

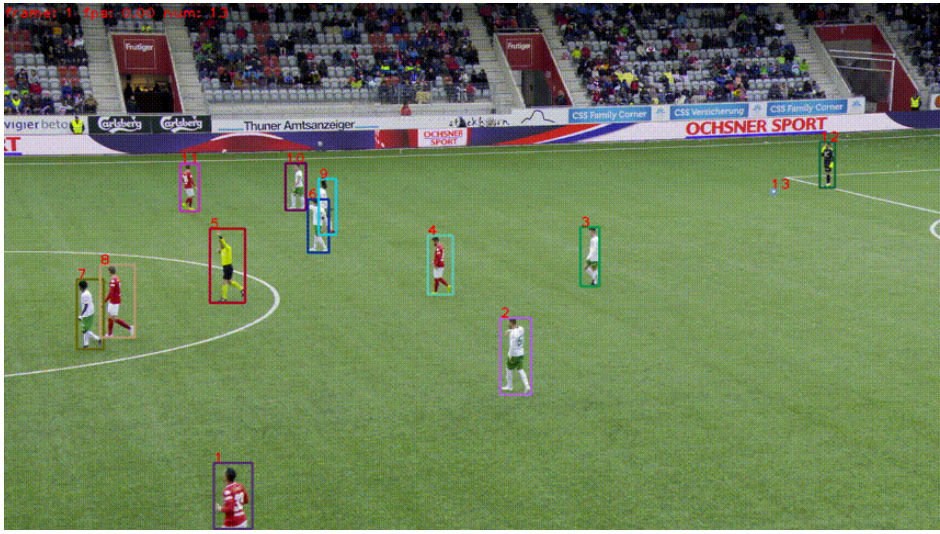
Foundation models:

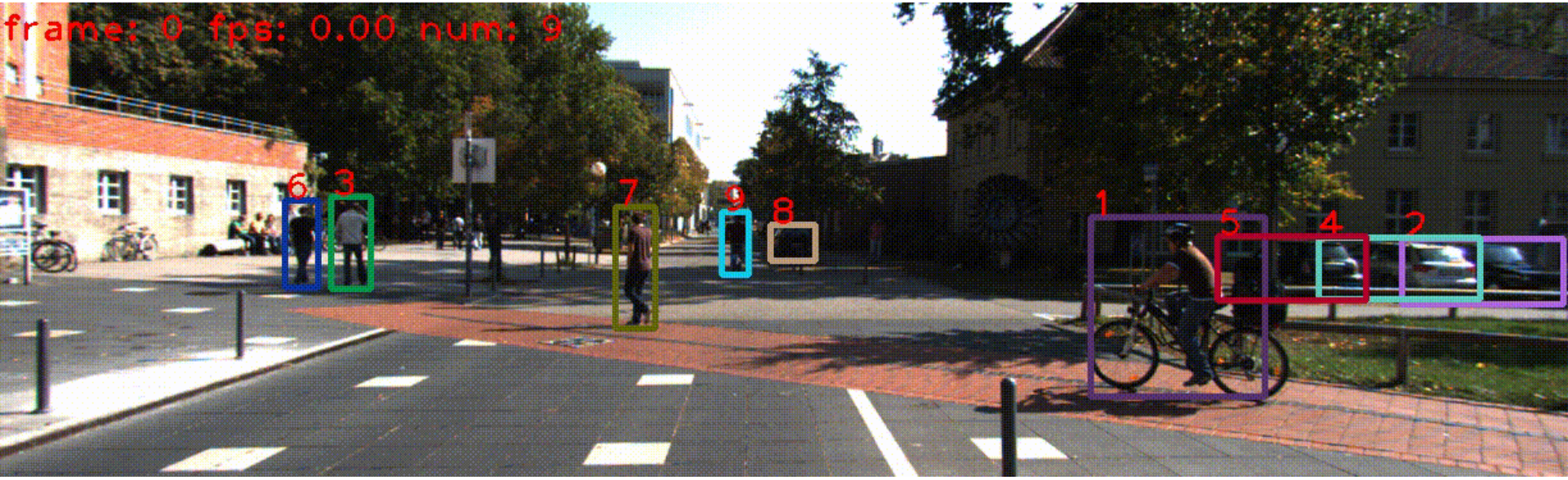
Grounded DINO + Segment Anything (SAM) + Track Anything (Cutie, DEVA, Massa)

People Tracking in real world situations

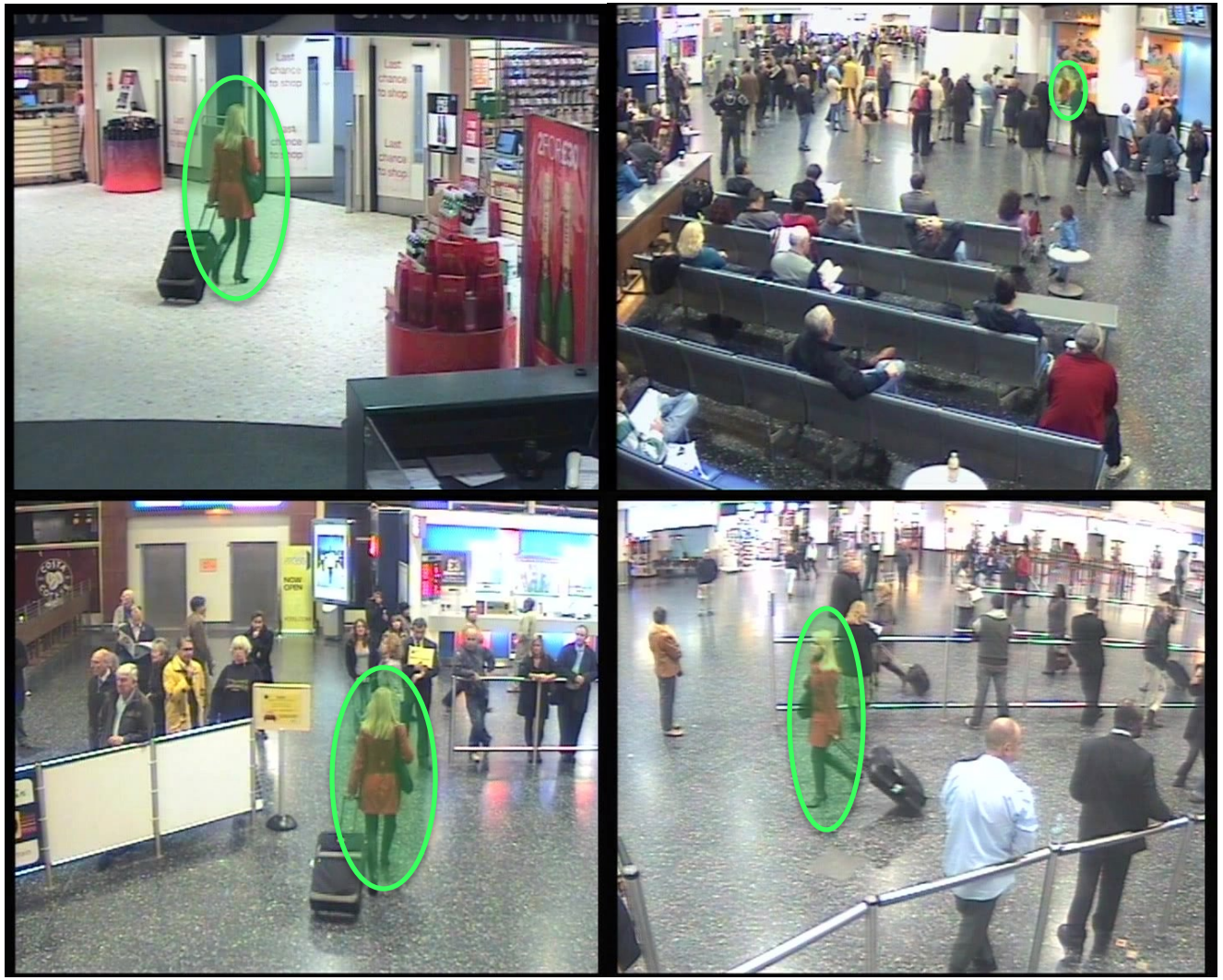
[Tomasz Stańczyk] MOT Multi-object tracking challenge: [MOT17]

People Tracking in real world situations





Person Re-identification



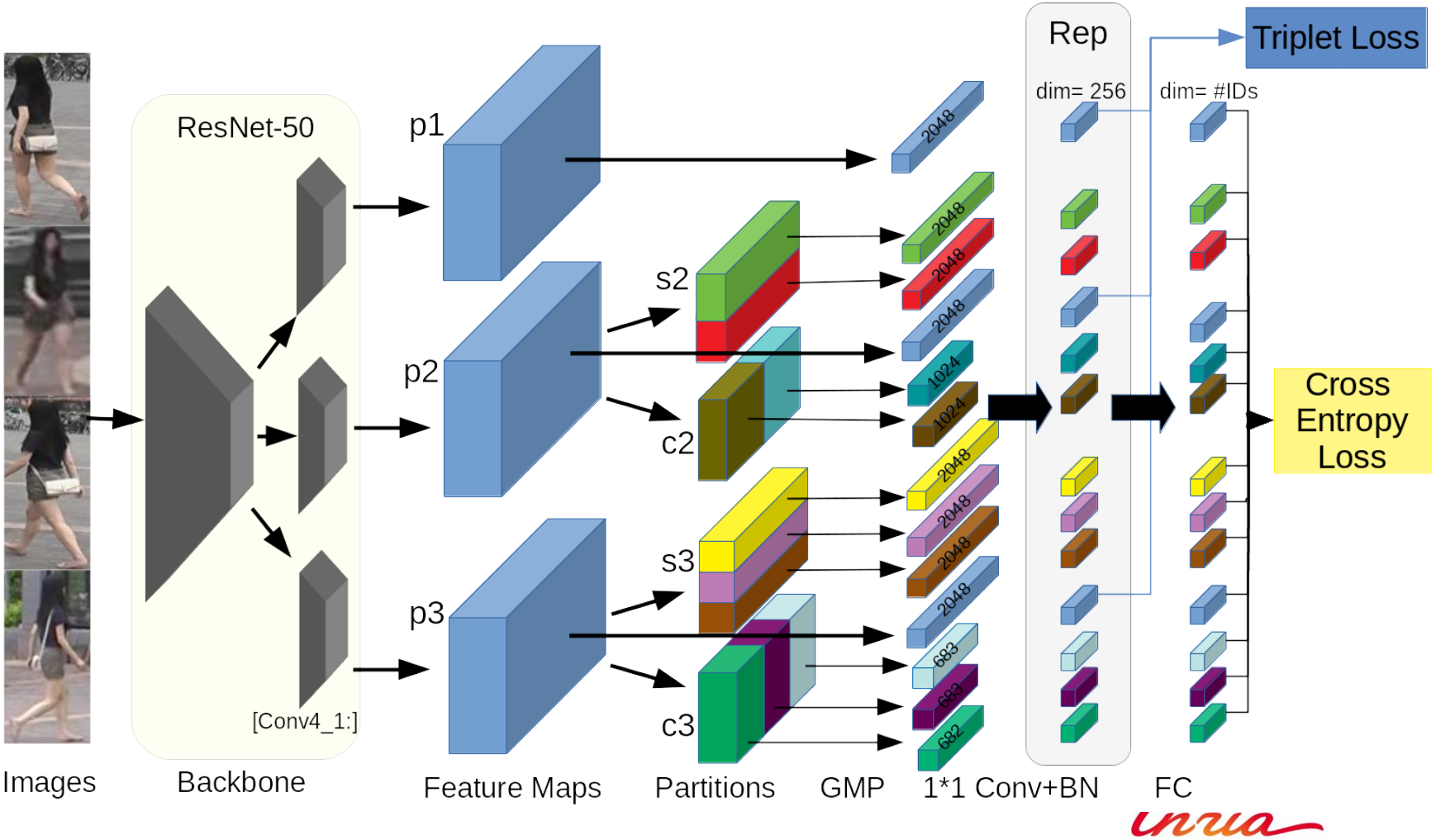
Person Re-Identification

Learning image features (large datasets)

$$L_{triplet} = \sum_{i=1}^{P}\sum_{a=1}^{K}[\max_{p=1,\ldots,K} \|\mathbf{a_i} - \mathbf{p_i}\|_2 - \min_{\substack{n=1,\ldots,K \\ j=1,\ldots,P \\ j\neq i}} \|\mathbf{a_i} - \mathbf{n_j}\|_2 + \alpha]_+$$

a: anchor
p: pseudo positive
n: pseudo negative

General Architecture of SCR: Spatial and Channel partition CNN Representations :



Person Re-Identification Results on Market-1501 [Hao Chen, PAMI22-24]

Qualitative results

- Success cases

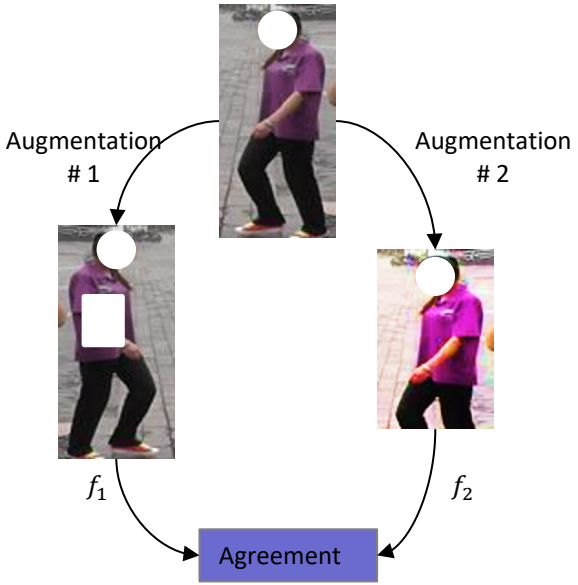
- Failure cases

High accuracy, but SCR (ours) requires **large** amount of labeled training data
-> unsupervised learning
-> domain transfer
-> life long (continuous) learning

Person Re-identification
Self-Supervised (Contrastive) Learning

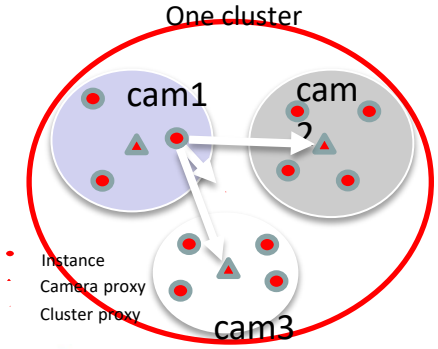
1. Add data augmentation (perturbation) to create a positive pair

2. Maximize the representation similarity between the positive pair



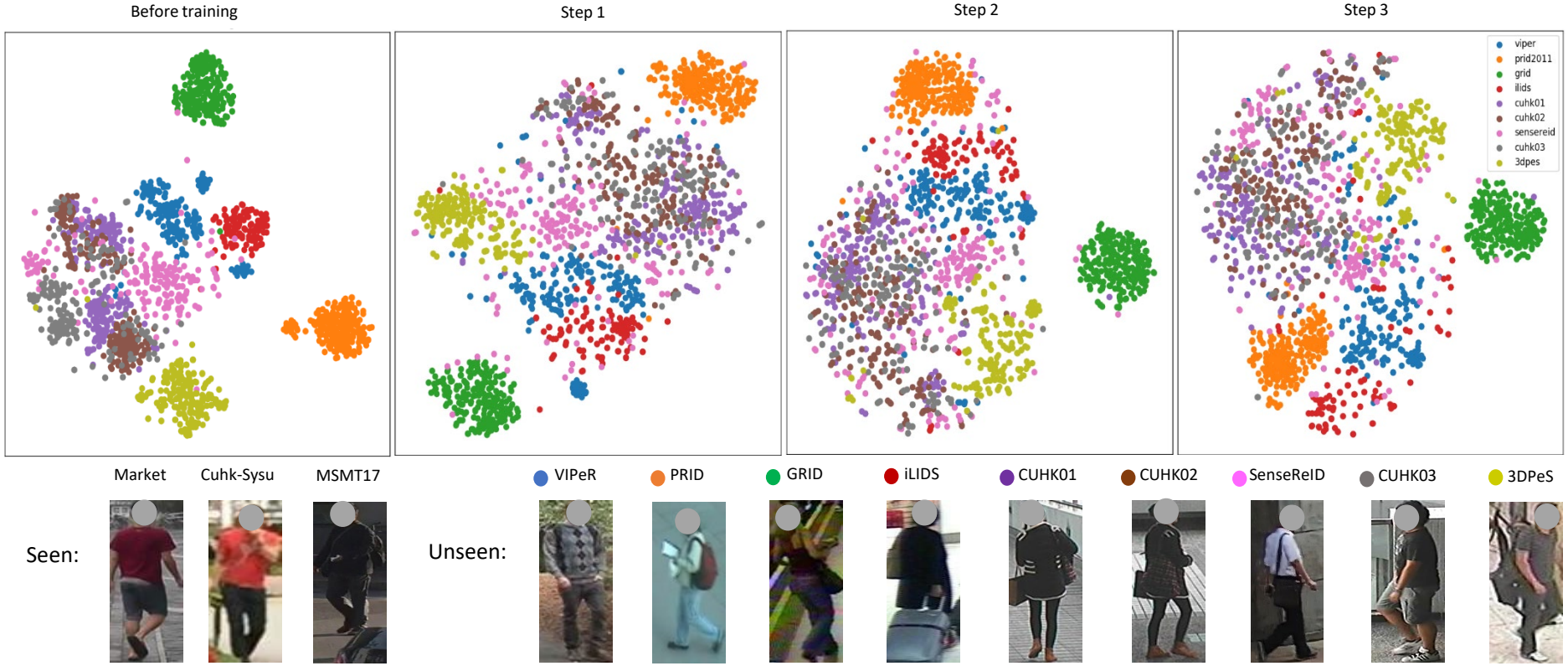
A model representation can be invariant/resilient to perturbations.

$$\mathcal{L}_{vi} = \mathbb{E}[\log\,(1 + \frac{\sum_{i=1}^{K} \exp\,(sim(f'_{new}, k_i)/\tau)}{\exp\,(sim(f, f_{pos})/\tau)})]$$



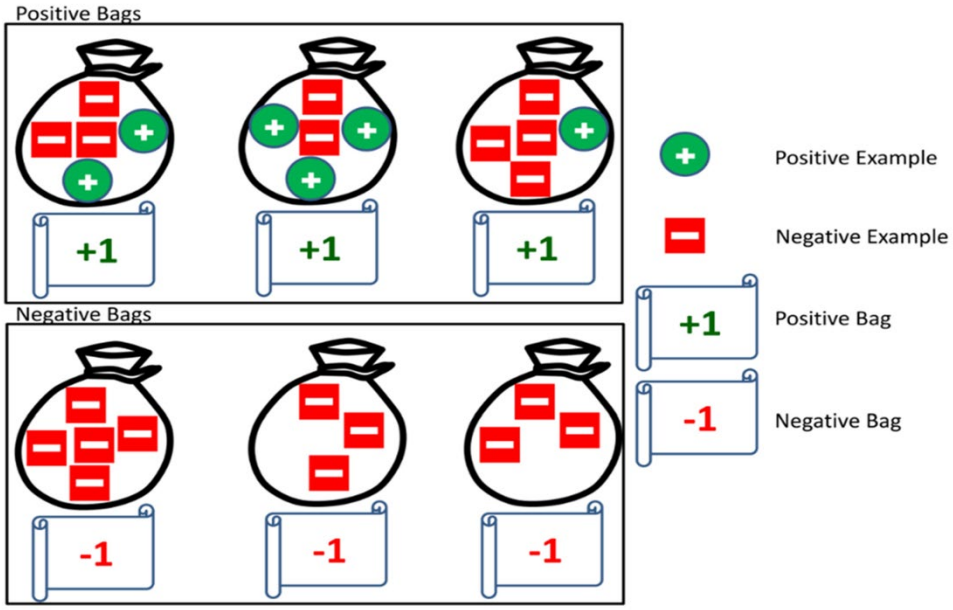
Lifelong (continuous) Domain Adaptation

Dataset Visualization



Weakly-supervised Video Anomaly Detection

[Snehashis Majhi]



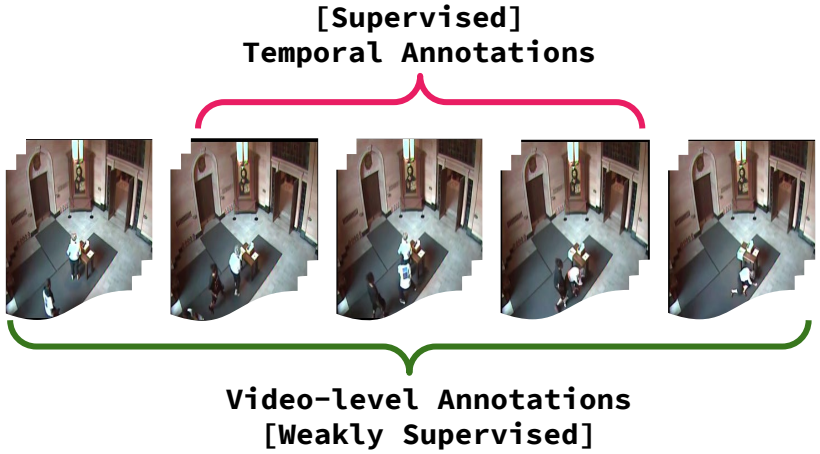
MIL [Multiple Instance Learning] loss = 1 – (max_abnormal – max_normal)

Weakly-supervised Video Anomaly Detection

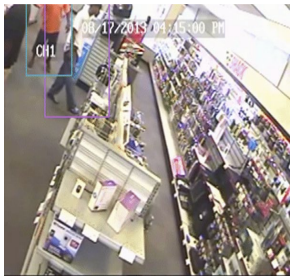
[Snehashis Majhi]

Challenges

- **Lack of temporally annotated videos.**



- **Human Centric fine-grained Anomalies**



- **Sparsity of Anomaly**

- **Long and Short Anomalies**

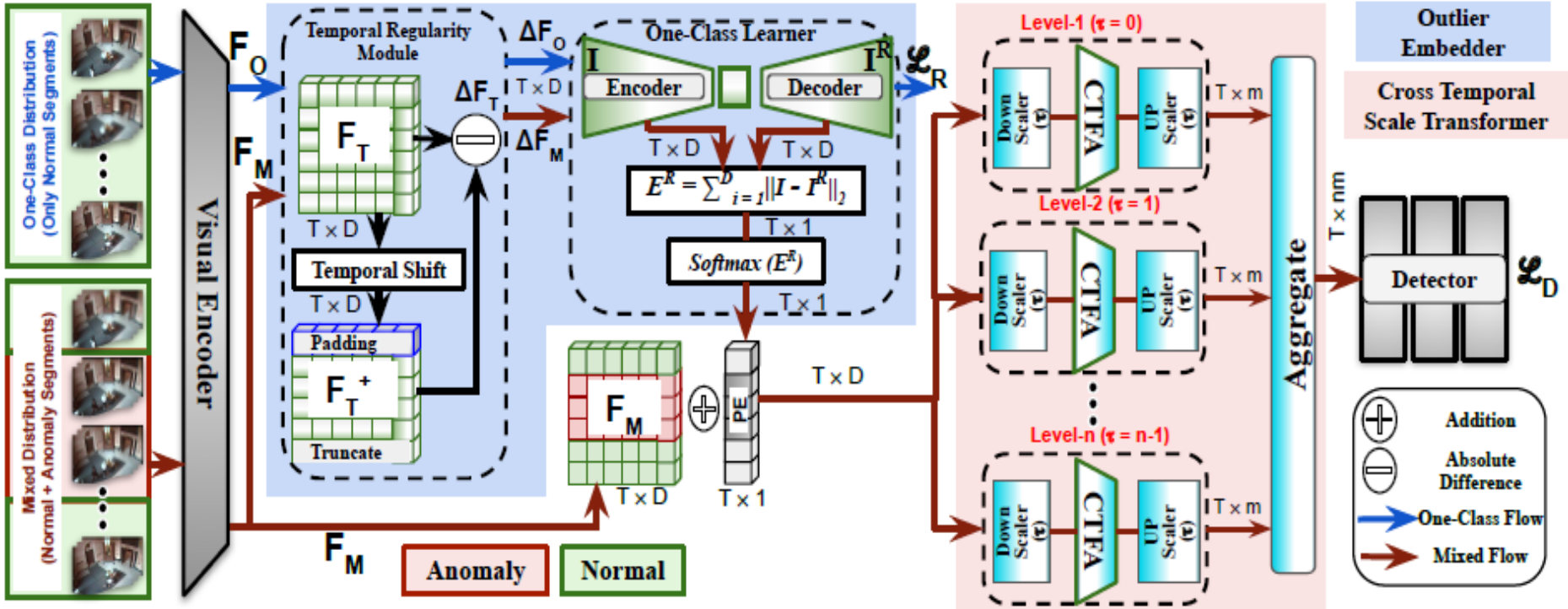
Long

Short

Weakly-supervised Video Anomaly Detection

Approach 1 : OE-CTST

Outlier-Embedded Cross Temporal Scale Transformer



Temporal pseudo labels: start and end of events

Multi Instance Learning: MIL loss

A Poly-modal Inductor for Weakly-supervised Video Anomaly Detection

- **RGB only features** are not sufficiently distinctive enough to distinguish **complex anomalies like shoplifting and visually similar normal events**.

- **Towards robust complex real-world anomaly detection, it is essential to augment RGB with additional modalities.**

- **But how many additional modalities?**
  - **FIVE**

We will see in this work !!

(a) Complex real-world anomalies with multi-modal saliencies

Activity Recognition for Daily-living activities

Web and movies datasets:
(Kinetics, UCF101, ActivityNet,…)

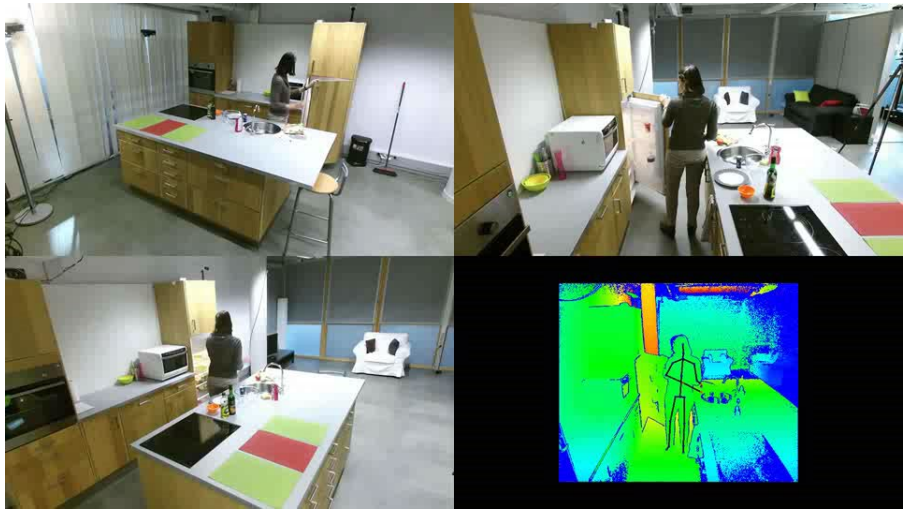
- Large number of classes
- High inter-class variation
- Camera motion
- Different environments
- Short actions

Different challenges compared to Fine-grained video datasets:

(Toyota smart home, Dahlia, NTU,…)

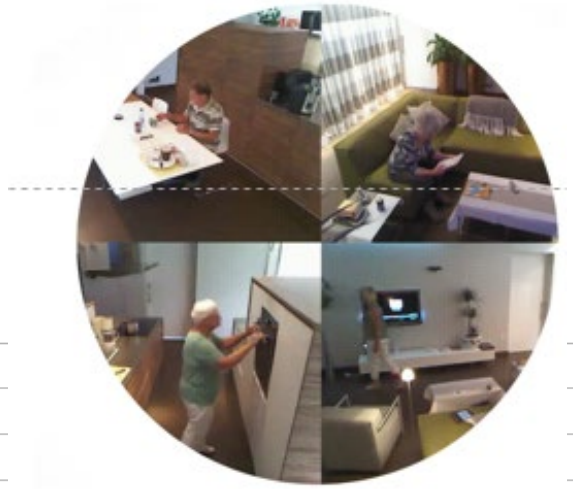
- **Real-time** recognition
- High intra-class variation
- **Low inter-class** variation
- Same environment, background
- Long and Composed actions

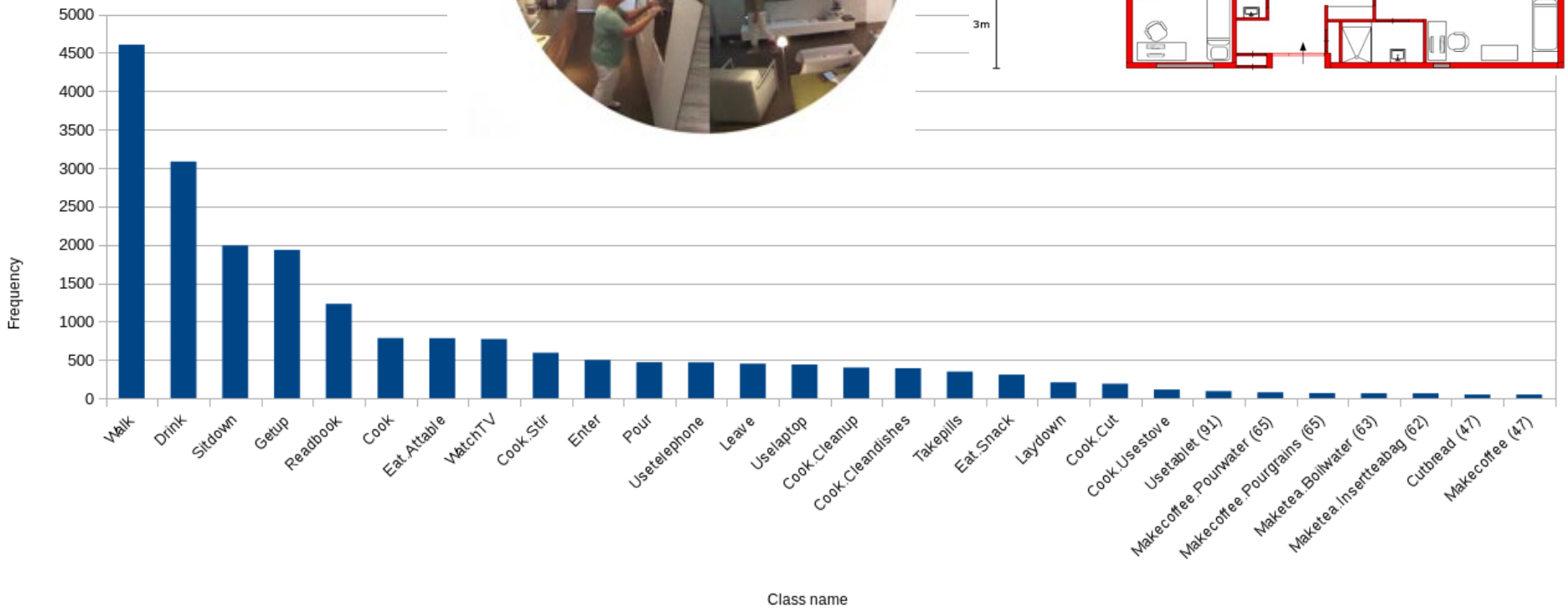
Need to model spatio-temporal relations



Toyota Smart-Home
Large scale daily living dataset

Partner Robot for
Assisted living

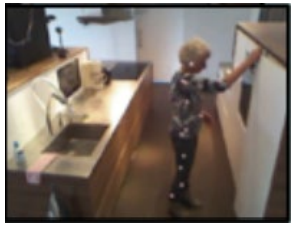




Toyota Smart-Home
Large scale daily living dataset

Privileged Modalities

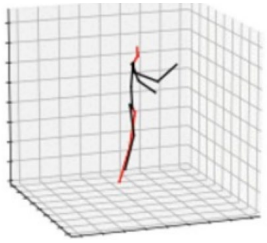
Different input modalities : RGB based and others: Audio, Text, Bio Signals (EEG, ECG, EDA, HR) …



RGB

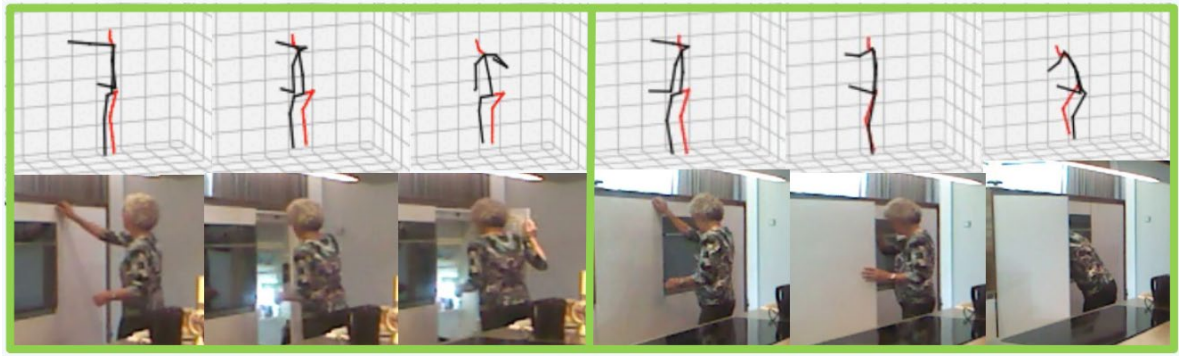
Depth

Optical Flow



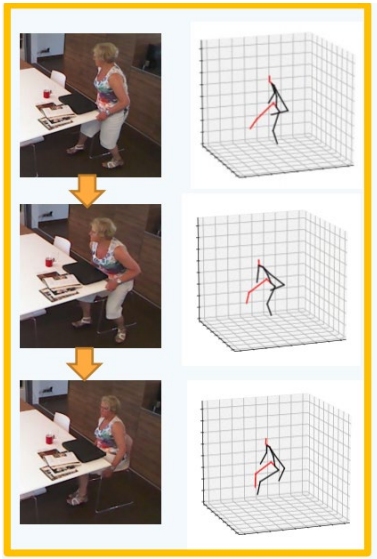
2D/3D skeleton

Complementary Nature: RGB vs skeleton



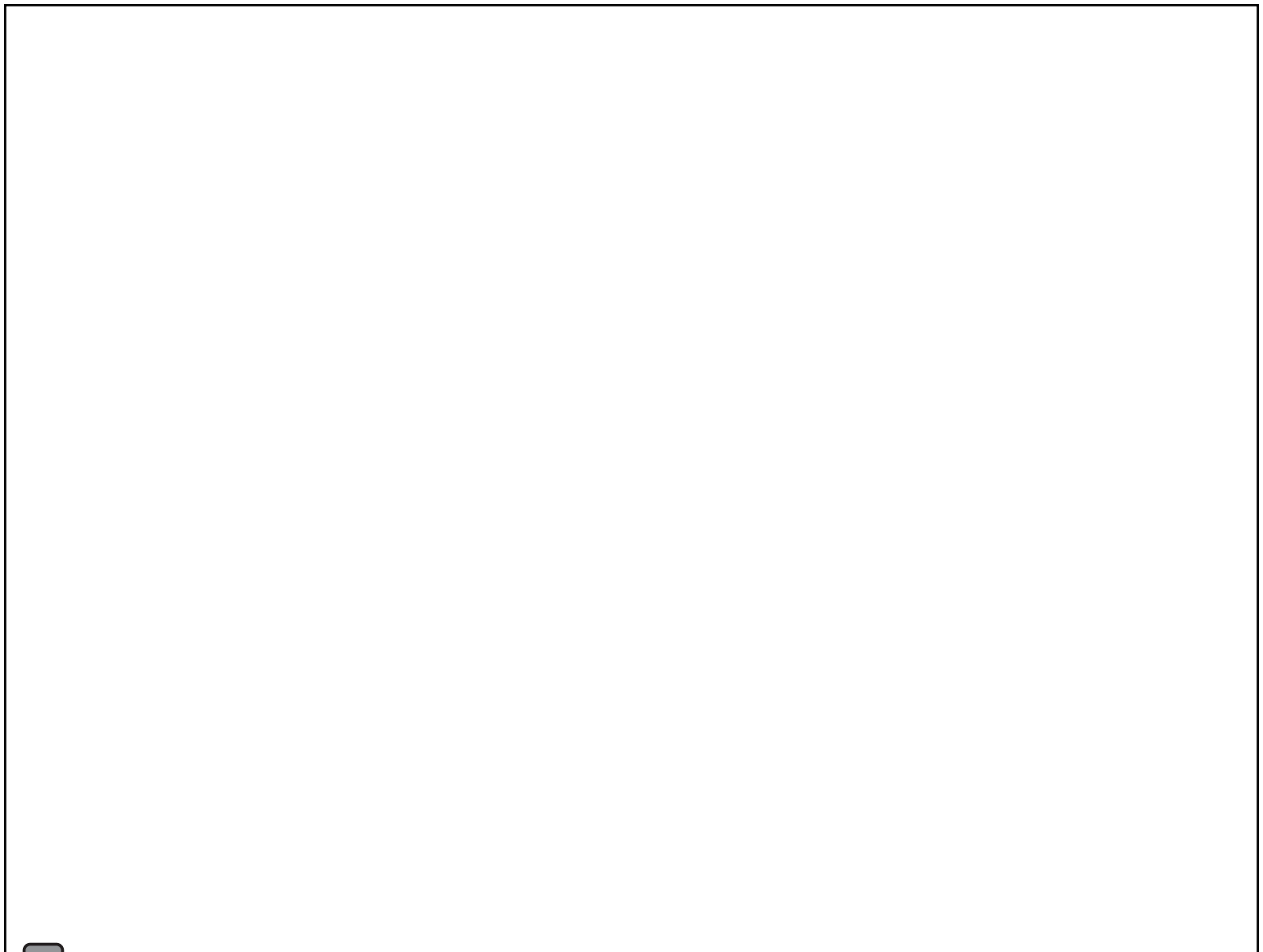
Open fridge Open cupboard

Filtering the noisy appearance patterns, distractors
Help capturing the body motion



Sit down

Toyota Smart-Home
Large scale daily living dataset



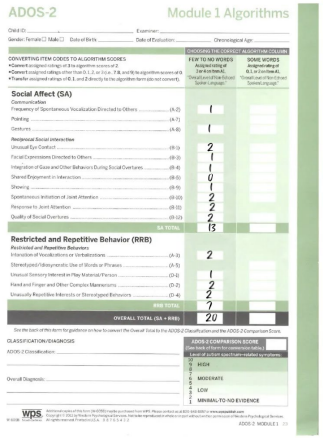
Analysis of human interactions for Autism Spectrum Disorder (ASD) [Abid Ali]

ASD : Lifelong Neurodevelopmental disorder

- Deficit in attention, reciprocal interactions, Communications
- Stereotypies - restrictive and repetitive behaviors & interests

Semi-automated ASD Analysis

Detect/recognize ASD behaviors



ADOS-2 Scores

Action Understanding Models

Coarse-interactive Activity Recognition [Dyadic interactions] (by Abid Ali Khan)

Tight interactions

- Proper symmetry & synchronization
- Physical contact
- Short activities
- Minimal mobility

Conversational interactions

- No mobility
- Shot from front-facing camera
- Activities → talking, eye-gaze aversion

Loose interactions

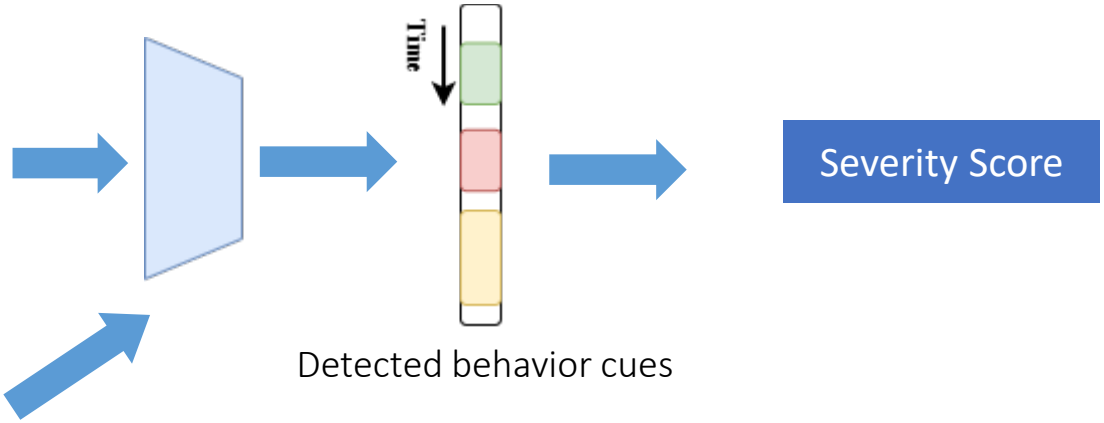
- Complex mobility
- Asynchronous & asymmetrical
- Without direct physical contact
- More than 2 min long actions

Ali et al. "Loose Social-Interaction Recognition in Real-world Therapy Scenarios", WACV'2025

Autism Severity Score Assessment

Untrimmed video

Action Detection



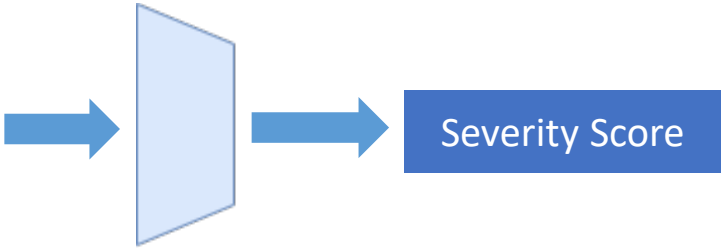
Detected behavior cues

Fine-grained annotations:
- Repetition
- Stereotypies
- Gaze
- Facial expressions

Fully-supervised method

Untrimmed video

Deep learning model

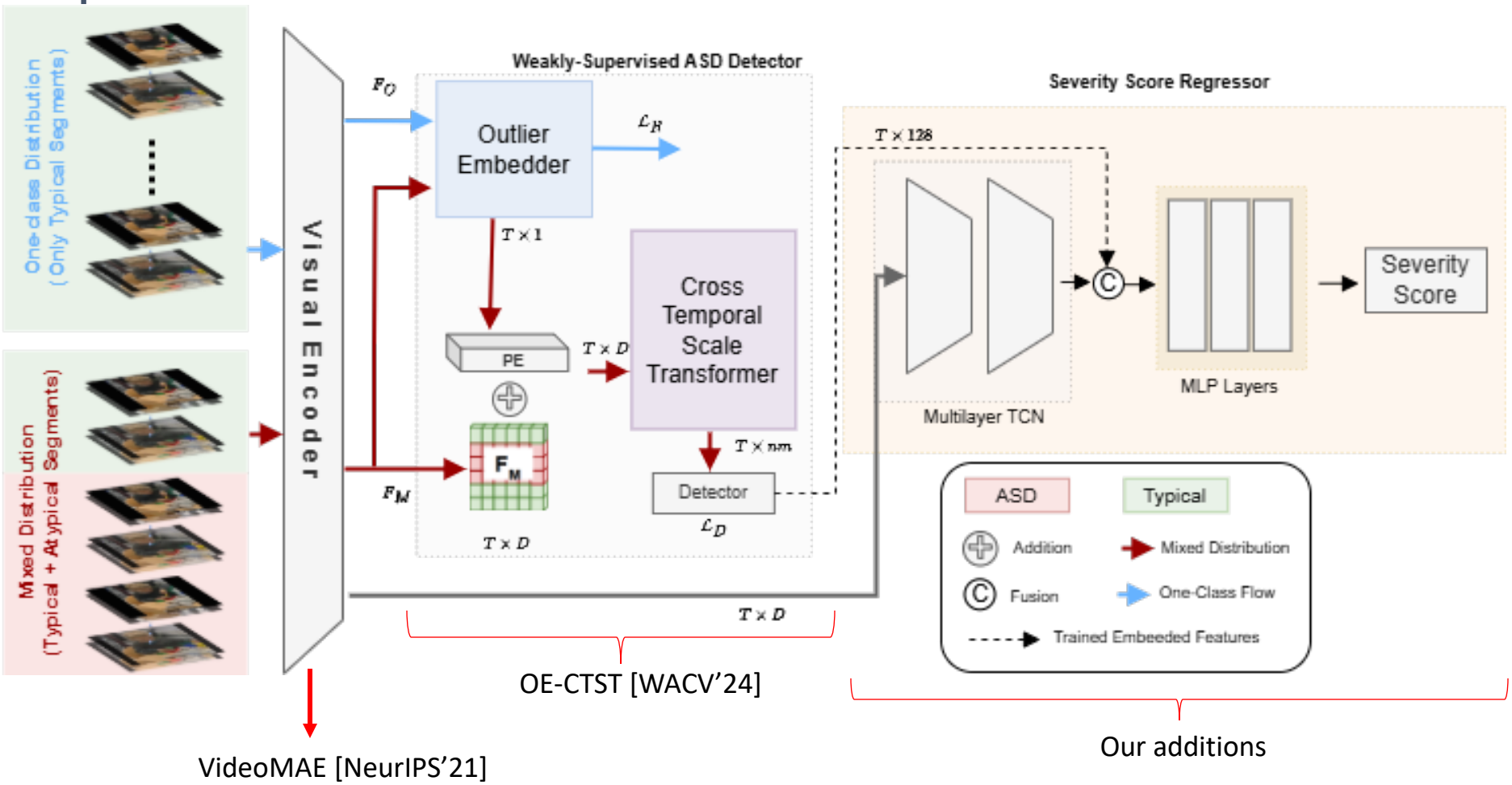


Weakly-supervised method

Abid, et al. "Weakly-supervised Autism Severity Assessment in Long Videos" CBMI'24

Autism Severity Score Assessment

Proposed Framework



Autism Severity Score Assessment

Dataset and ASD scoring

RESTRICTED AND REPETITIVE BEHAVIOR (RRB)	Module	score
Restricted and Repetitive Behaviors		
Intonation of Vocalizations or Verbalizations	(A-3)	2
Unusual Sensory Interest in Play Material/Person	(D-1)	0
Hand and Finger and Other Complex Mannerisms	(D-2)	3
Unusually Repetitive Interests or Stereotyped Behaviors	(D-4)	1

Severity Levels	No. of hour-long Videos	No. of segmented modules/long video	Train/Test
No-autism	14	35	27/8
Weak	6	19	16/3
Moderate	20	52	40/12
High	35	110	87/23

Dataset

Autism Severity Score Assessment

Summary and Limitations

Summary

- Weakly-supervised way → severity score estimation.
- Estimate overall autism score.
- **Discover new gestures biomarkers** for autism.

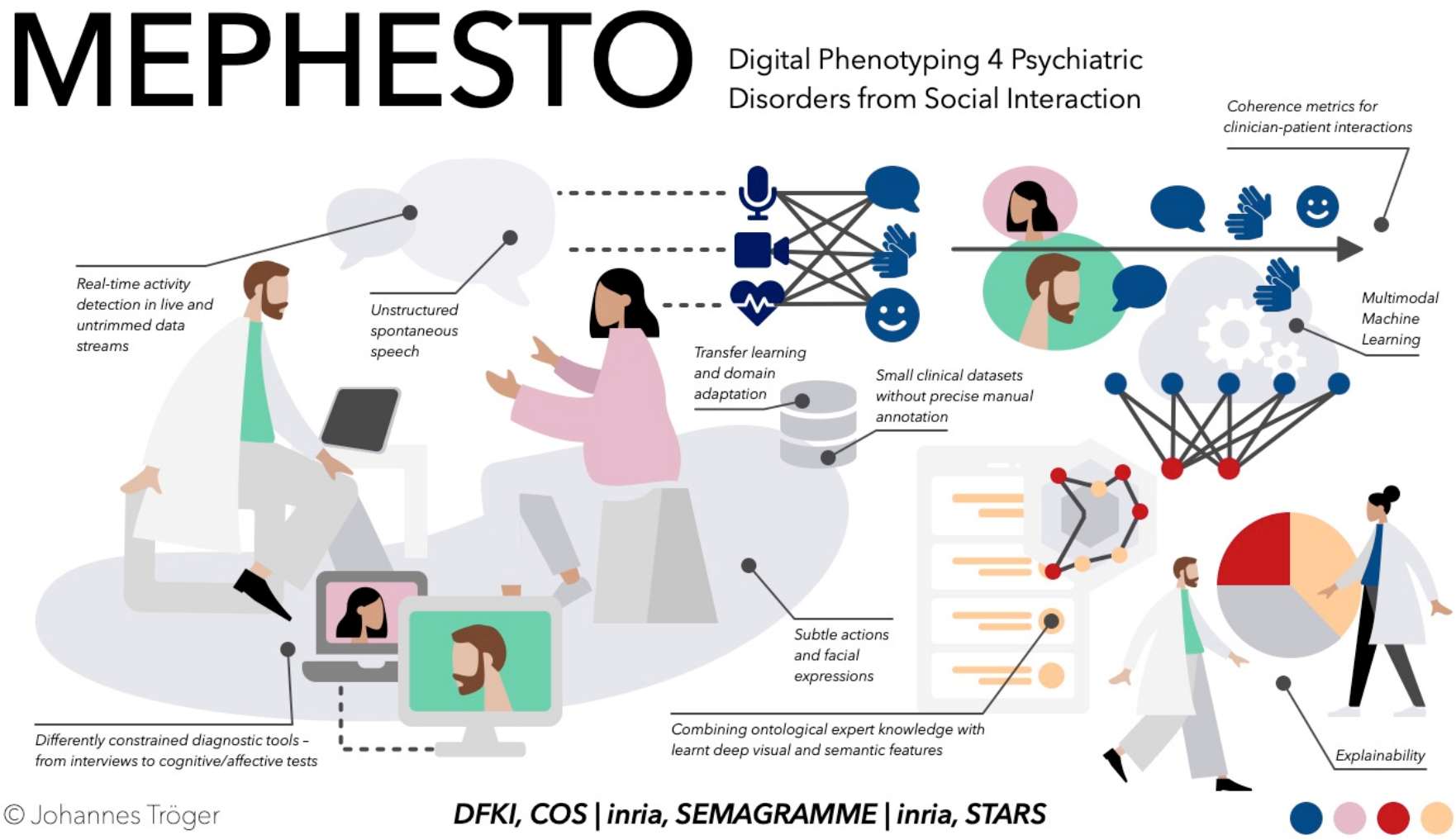
Limitations

- Only overall severity score
- Dataset bias → high severity
- Fine-grained complex biomarkers → facial expressions, eye-gaze, emotions

Perspective for autism severity score

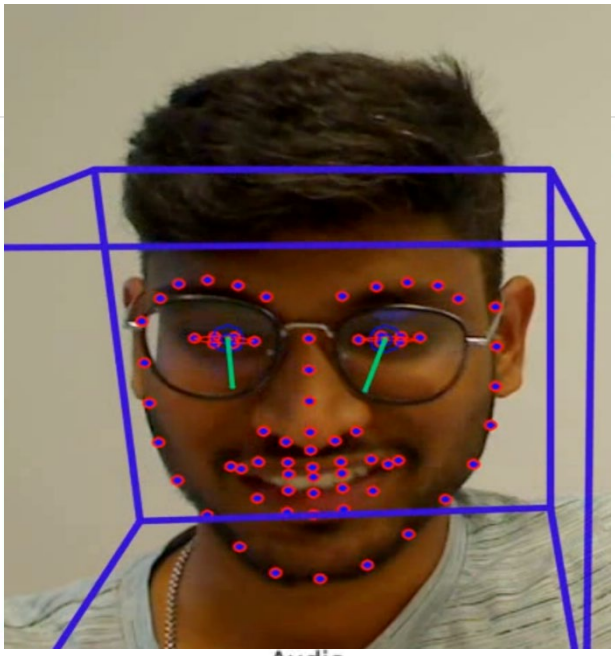
- Severity analysis → whole dataset
- Per module severity score analysis
- Design more robust models to discover novel biomarkers responsible for autism.

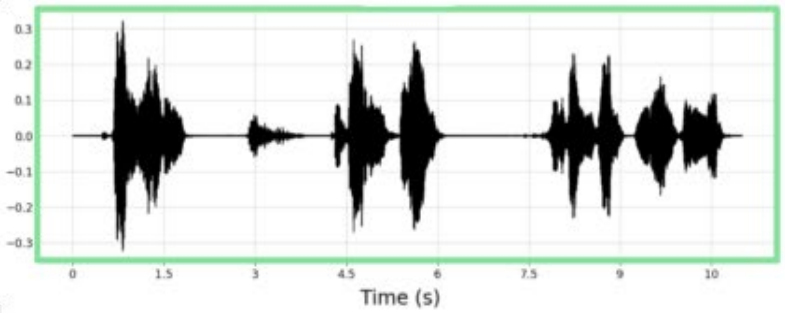
Digital Phenotyping for Psychiatric Disorders from Social Interaction : audiovisual + physiological

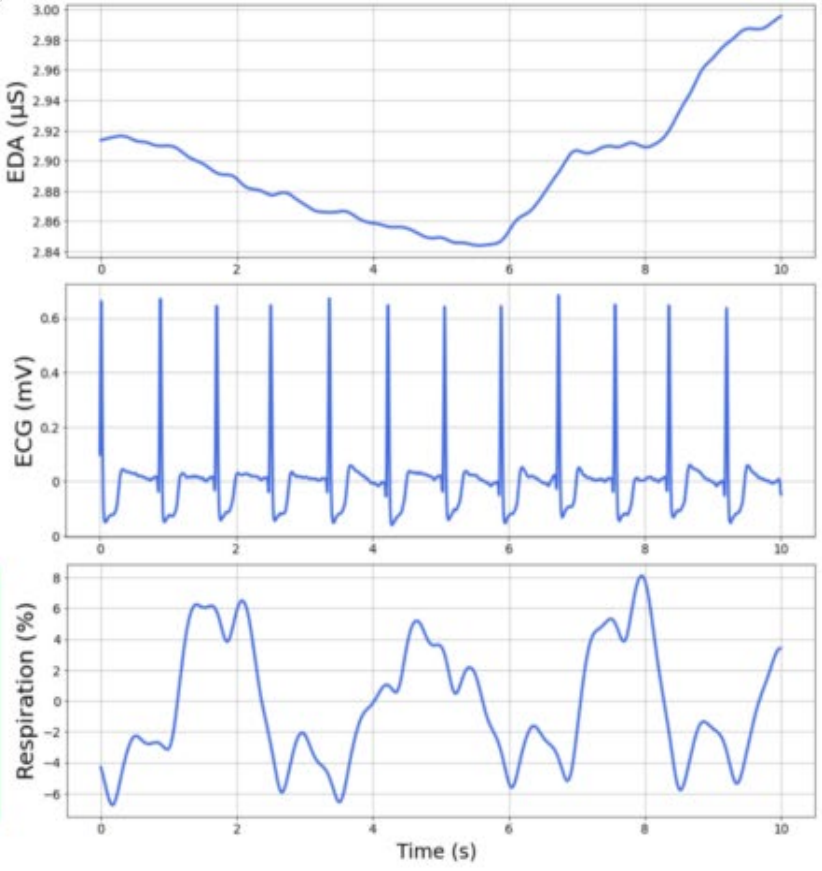


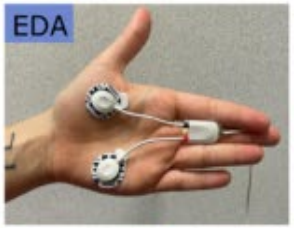
Emotion Recognition : Facial Expression Recognition (by Valeriya Strizhkova)

Characterizing Emotion using Facial Motion and Physiological signals









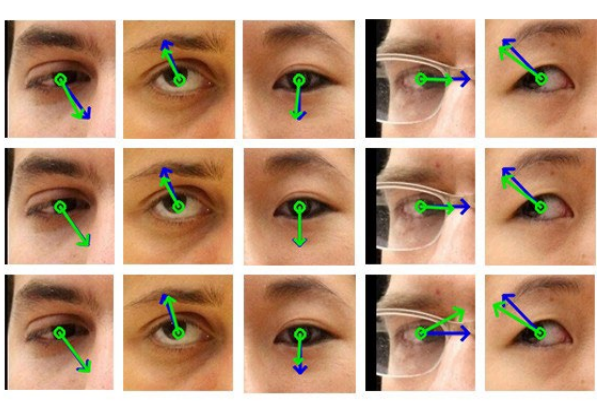
Emotion Recognition : Facial Expression Recognition

Characterizing Emotion using Facial Motion and Physiological signals

Emotion Recognition : gaze estimation

Characterization of gaze (attention) during speech: case of schizophrenia (rupture of content).

Green dot: eye tracker



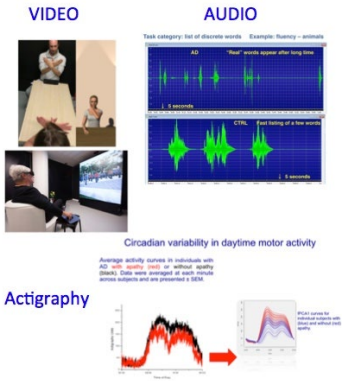
Multimodal Recognition of Human Interactions

Characterization of emotion: with Video (Gesture, gaze) + Bio-signals (EDA, ECG) case of Schizophrenia, Depression, Bipolar, Post-Traumatic Stress Disorder.

Conclusion

A **global framework** for building real-time video understanding systems:

- **Activity Monitoring** Systems to measure levels of everyday activities: from handcrafted to (un)supervised learned models of activity
- Robust for **long term** video monitoring
- Online and real-time recognition with limited user interaction during training



Perspectives:

- View-point invariant - Real-world settings
- Generate totally unsupervised models
- Generic semantic activity models (cross scenes), Adaptive learning
- Use finer features as input for the algorithm (head, posture, facial, hand, gesture…)
- More semantics, emotion, mental states.
- Multi-modalities (e.g. speech)
- Reaction to Stimulation : Serious Games