

Events and Frames: challenges and opportunities of multimodality

Federico Becattini
University of Siena



UNIVERSITÀ
DI SIENA
1240

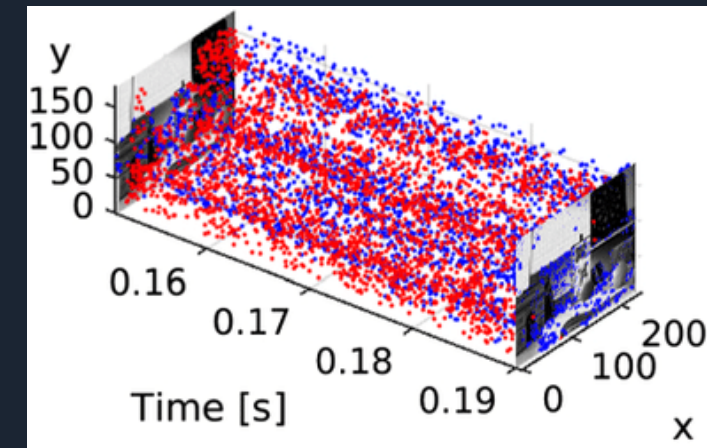
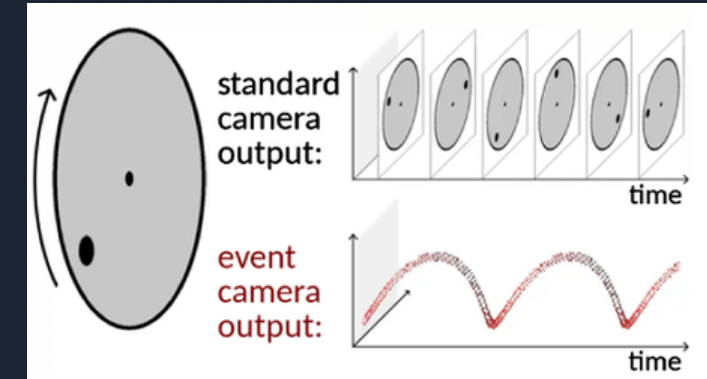
RGB Cameras and Event Cameras

RGB cameras

- fixed frame-rate (e.g., 30 FPS)
- capture dense semantic information and color
- no high-speed dynamics
- motion blur

Event cameras

- asynchronous
- microsecond resolution
- wide dynamic range



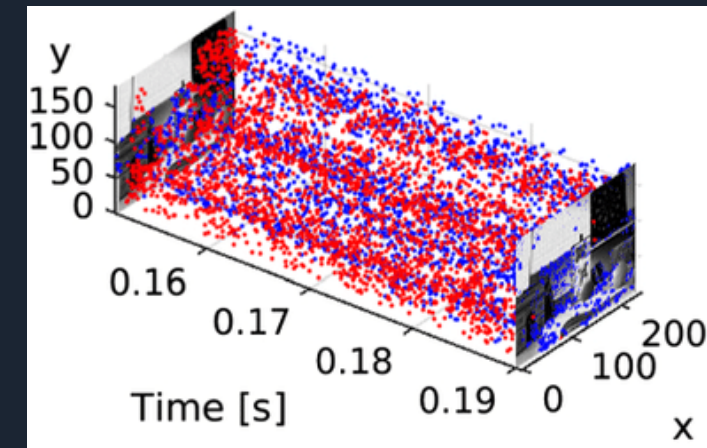
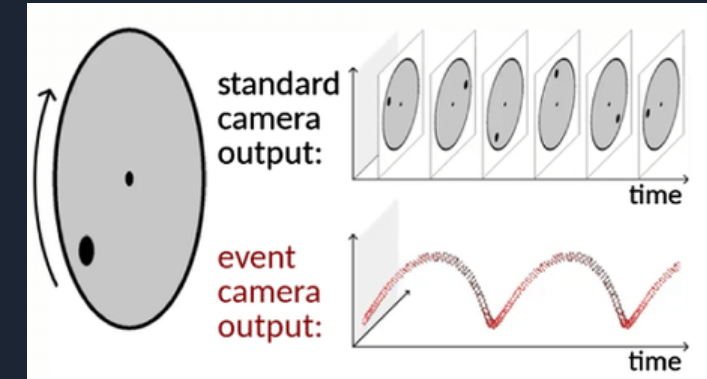
RGB Cameras and Event Cameras

RGB cameras

- fixed frame-rate (e.g., 30 FPS)
- capture dense semantic information and color
- no high-speed dynamics
- motion blur
- well established computer vision literature and codebase

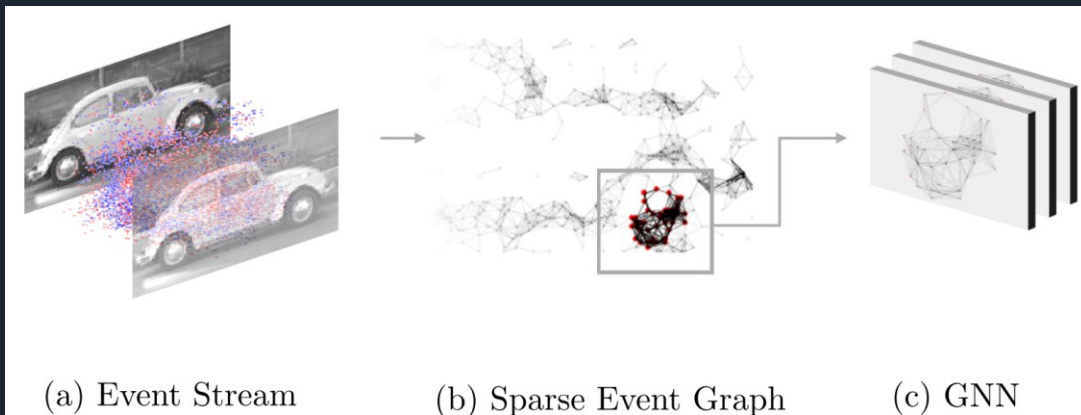
Event cameras

- asynchronous
- microsecond resolution
- wide dynamic range
- standard feature extractors cannot easily process the sparse, binary-like nature of event data

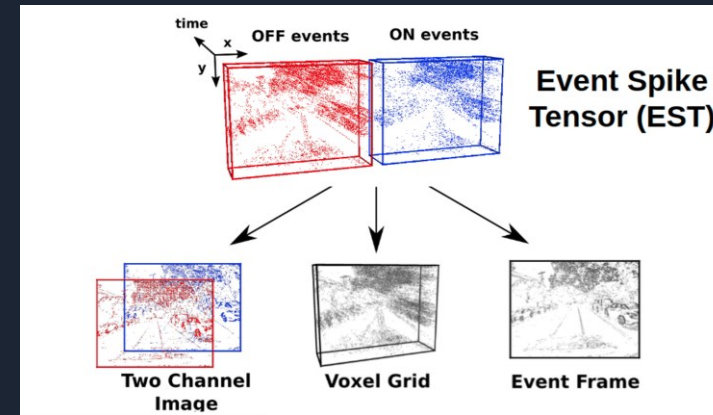


Computer Vision and Event Cameras

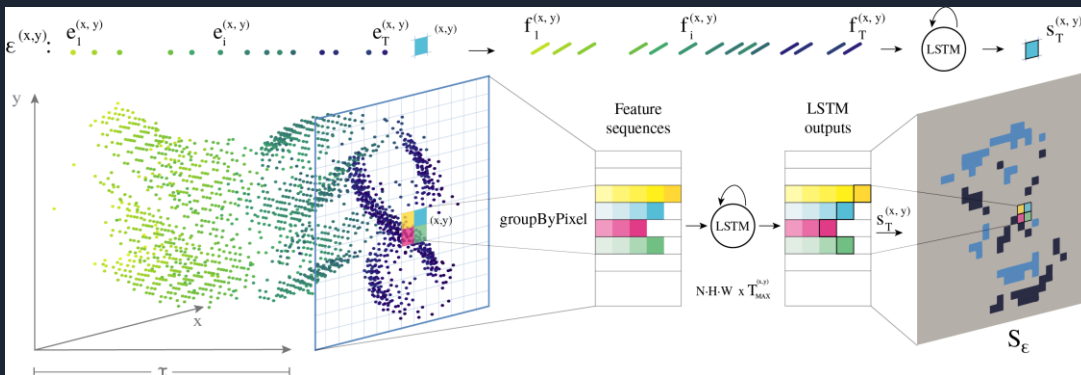
To be able to process events with computer vision techniques, a simple approach is to aggregate events into event frames or other data structures



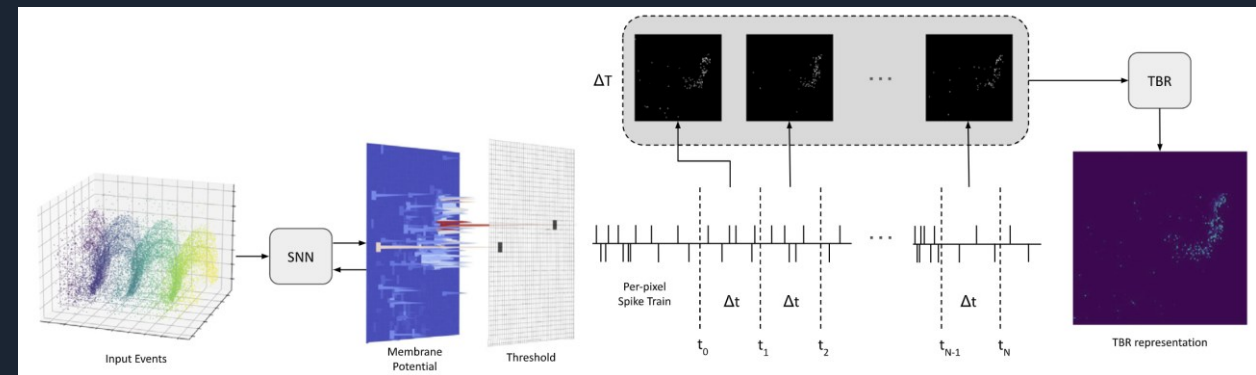
[Schaefer et al. CVPR 2022]



[Gehrig et al. ICCV 2019]



[Cannici et al. ECCV 2020]



[Magrini et al. PRL 2025]

Limitations of Neuromorphic Vision

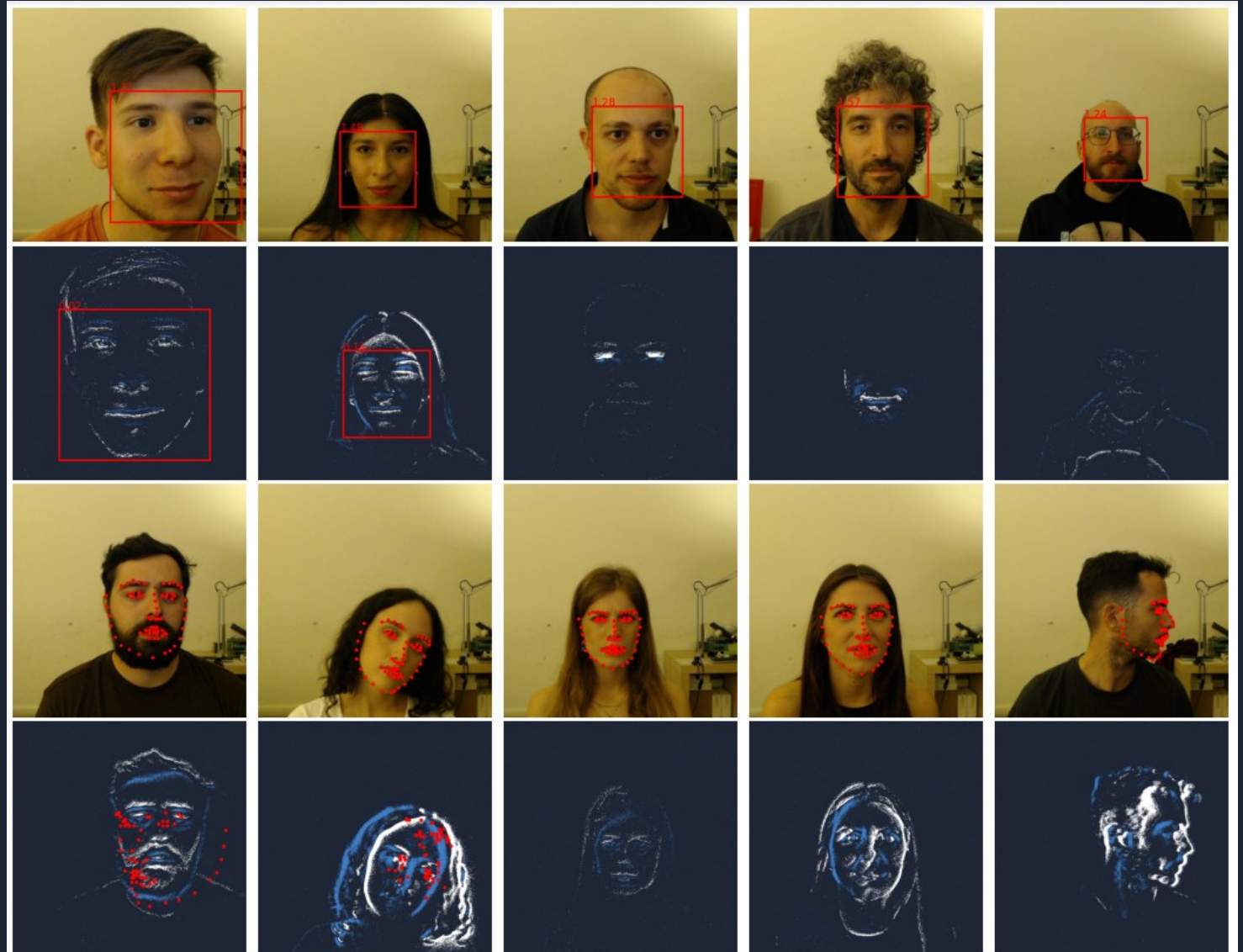
Domain shift

Directly porting models from one modality to another is often impossible due to the severe domain shift

Data Annotation burden

Labeling event data is significantly more complex and counter-intuitive than labeling traditional images

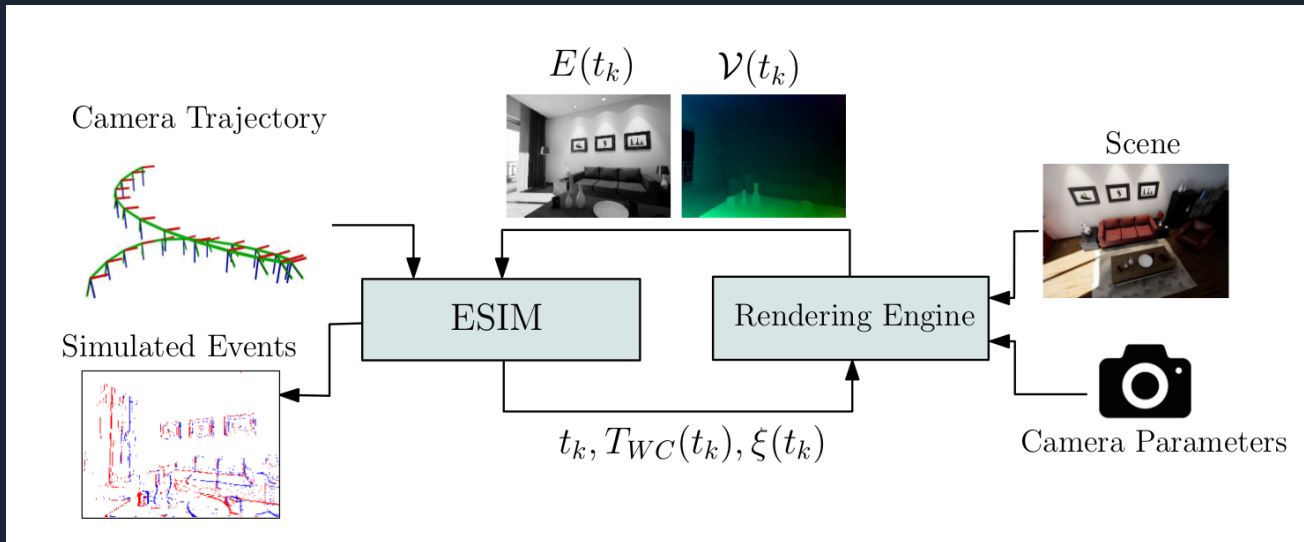
Large-scale annotated datasets for event-based tasks are rare



Event Camera Simulators

A possible solution is to use event camera simulators to generate synthetic versions of annotated datasets

This allows us to train new models on event data



[Rebecq et al. Corl 2018]

ESIM: an Open Event Camera Simulator

Henri Rebecq, Daniel Gehrig, Davide Scaramuzza

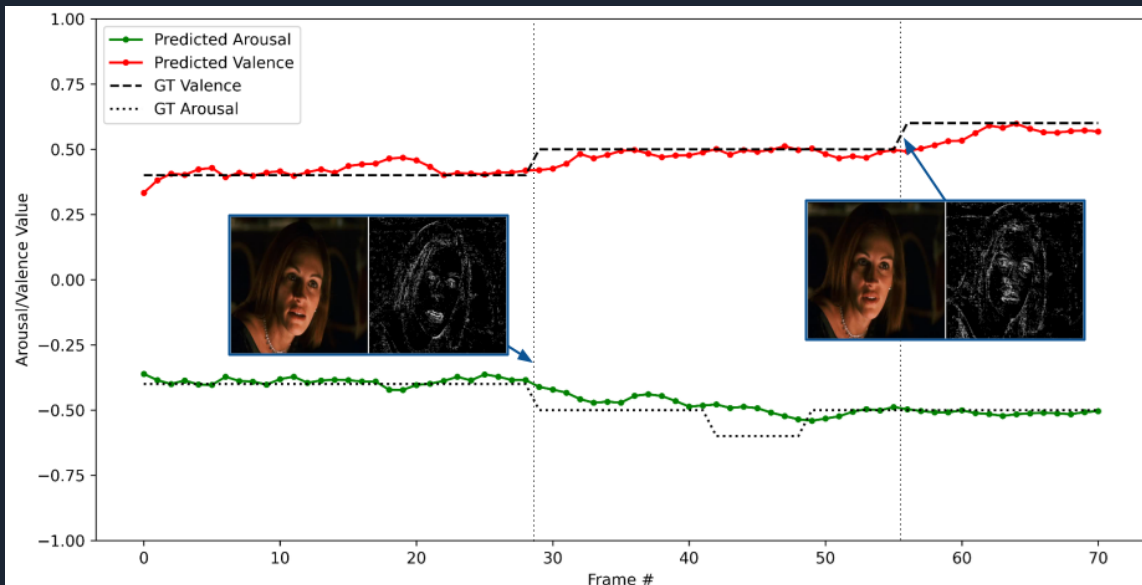


Simulating Faces

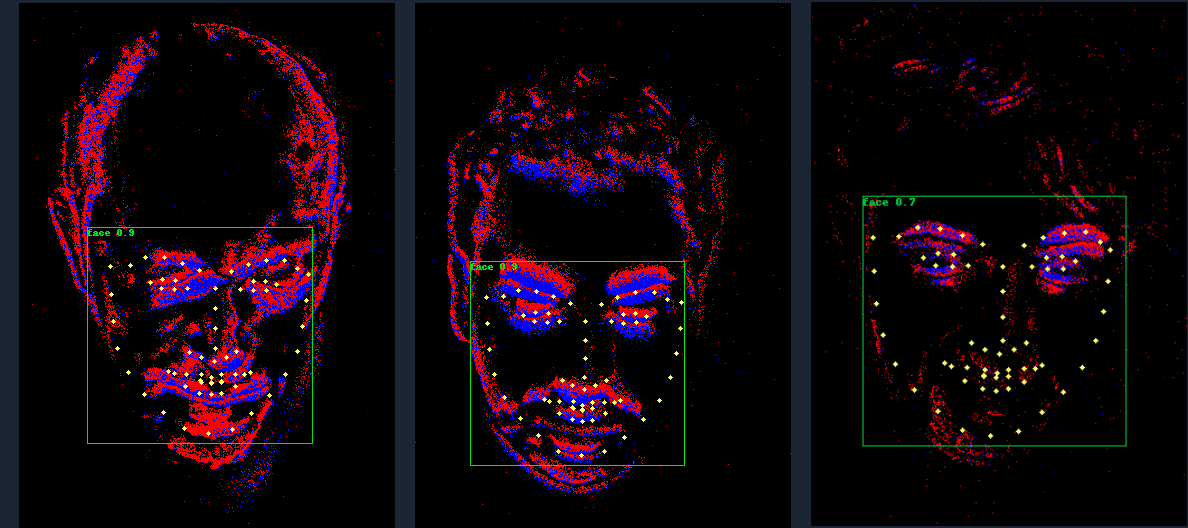
Using event camera simulators we trained simple models for:

- Face detection
- Landmark estimation
- Valence/Arousal estimation

The models can be applied to real event frames



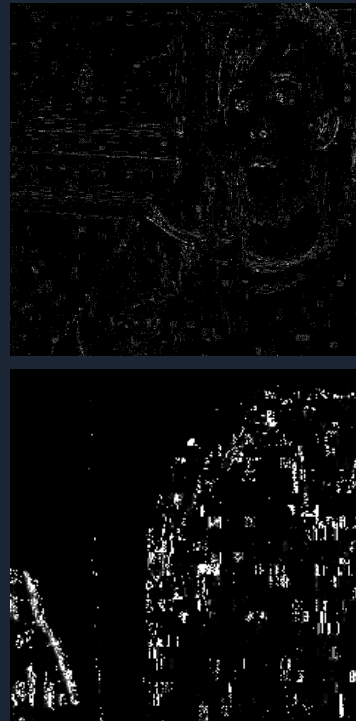
[Berlincioni et al. JAIHC 2024]



[Berlincioni et al. CVPRw 2023]

However, simulators have limitations:

- Usually slow
- Missing information in the source is not recovered
- Compression artifacts in the RGB video might generate spurious events



[Berlincioni et al. JAIHC 2024]

FACEMORPHIC

We collected FACEMORPHIC, a dataset of temporally synchronized RGB and Event streams with the goal of detecting Action Units

We asked each user to perform 24 Action Units and read some sentences (future work: lip reading)

Statistics:

- 64 users
- 446.400 frames
- 3148 temporally synchronized RGB+Event videos
- Approximately 4.13 hours of multimodal recordings



Neuromorphic Facial Analysis with Cross-Modal Supervision

Video-level supervision: we process event frames with a neural network in order to classify the video

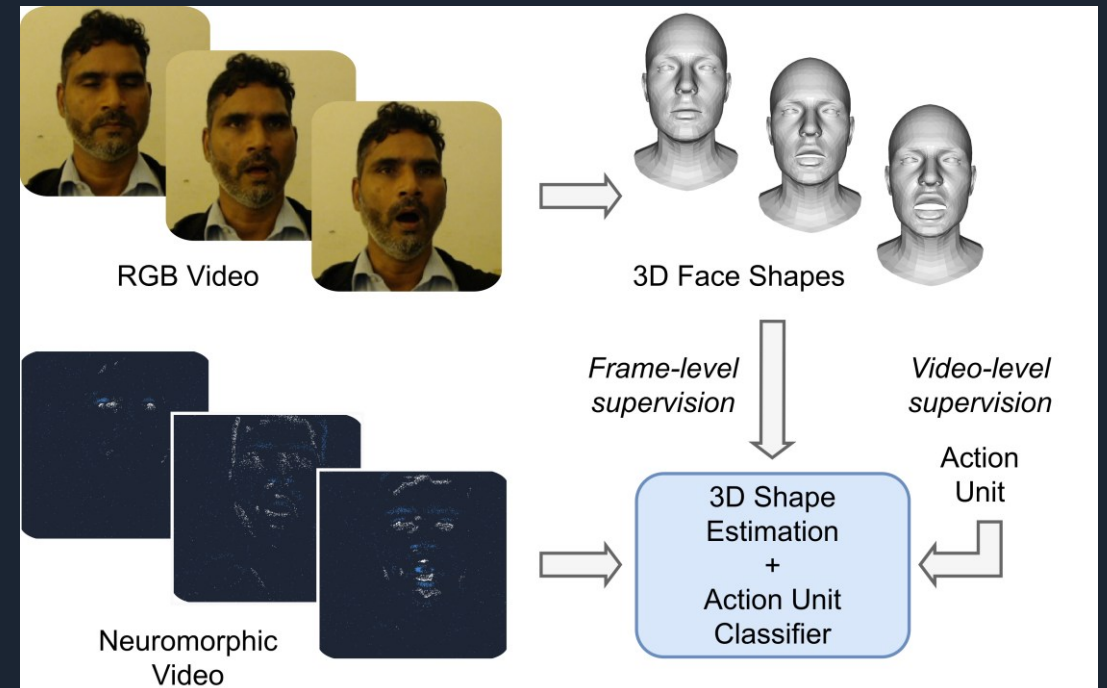
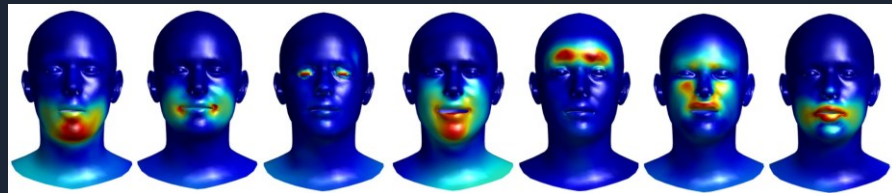
Frame-level supervision: we ask the model to predict, frame-per-frame, the 3D shape of the face

We obtain 3D information from the RGB stream using standard computer vision tools:

- Facial landmark estimation
- 3D Morphable Model (3DMM) Fitting

We discard identity information from the 3DMM and only retain deformations regarding expressions

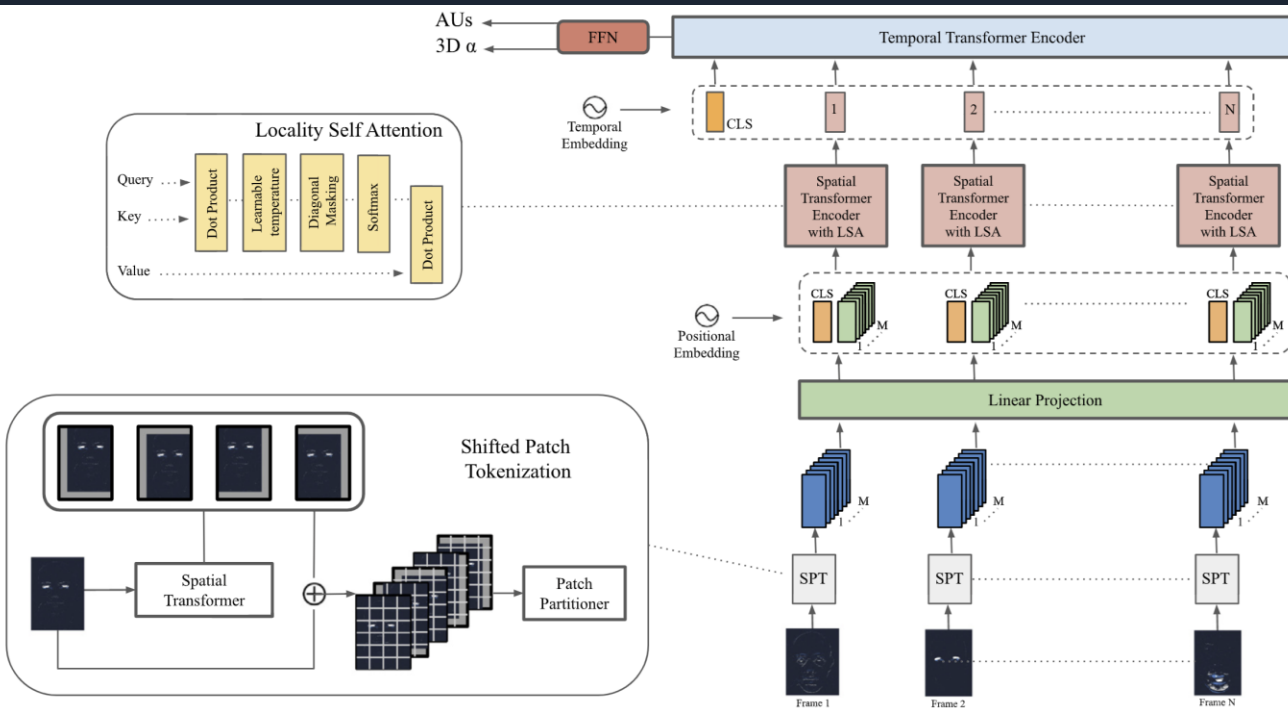
Deformations are expressed as a combination of local deformations from a learned dictionary



Neuromorphic Facial Analysis with Cross-Modal Supervision

Event-based models outperform RGB counterparts

Model	Mod	Accuracy	Top 3 accuracy	Top 5 accuracy
ResNet18+LSTM	Event	46.23	62.91	68.58
	RGB	25.65	38.86	46.12
ResNet18+Transf.	Event	31.74	40.63	52.08
	RGB	4.16	12.50	20.83
I3D	Event	47.08	69.58	80.66
	RGB	29.86	72.00	82.05
SWIN3D-T	Event	37.42	49.17	61.47
	RGB	35.62	42.19	56.10
SWIN3D-S	Event	38.89	51.48	63.29
	RGB	33.99	41.73	54.91
Ours	Event	48.78	69.40	78.26
	RGB	44.93	61.60	71.73



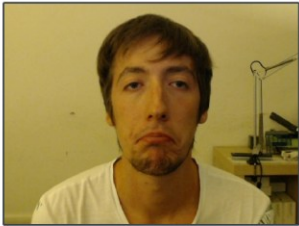
Cross-Modal supervision always improves the models

Model	$\mathcal{L}_\alpha$	Accuracy
ResNet18+LSTM	✓	50.21
ResNet18+Transf.	✓	43.35
I3D	✓	49.58
SWIN3D-T	✓	38.24
SWIN3D-S	✓	40.85
Ours	✓	50.56
ResNet18+LSTM	✗	46.23
ResNet18+Transf.	✗	31.74
I3D	✗	47.08
SWIN3D-T	✗	37.42
SWIN3D-S	✗	38.89
Ours	✗	48.78

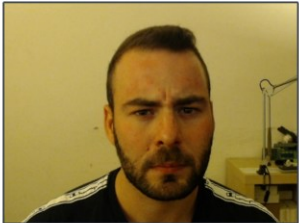
Neuromorphic Facial Analysis with Cross-Modal Supervision

As a byproduct, we are also capable of estimating 3D shape deformation coefficients

AU 15



AU 4



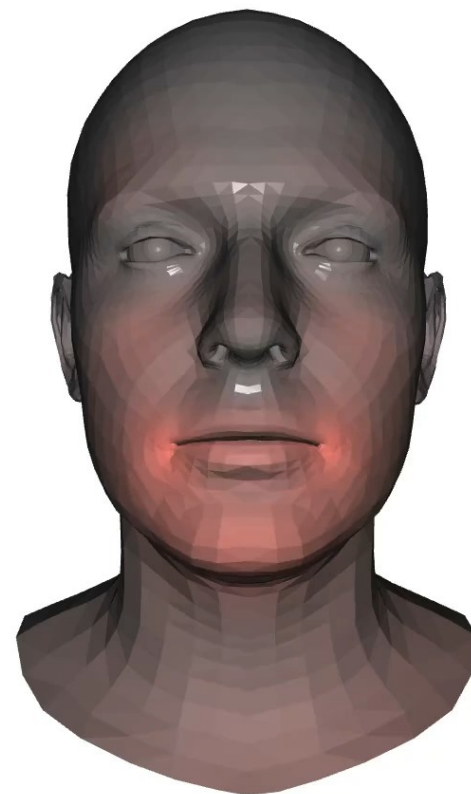
AU 26

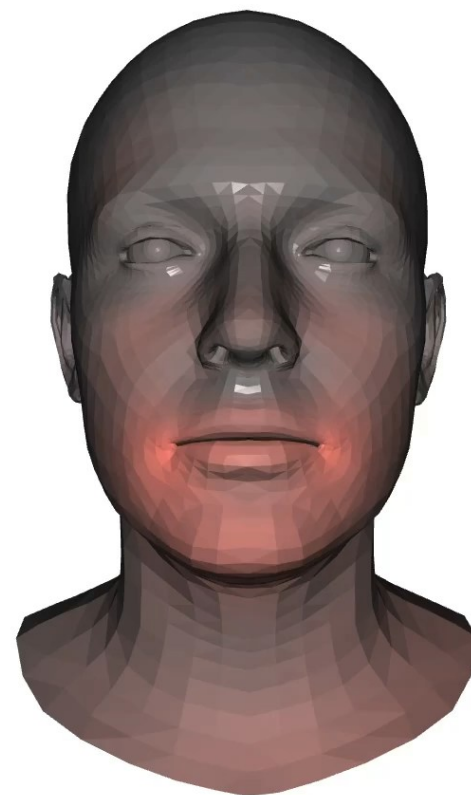


AU 12









Towards multimodal models

Lots of applications can benefit from multimodal RGB+Event processing.

In fact, Event and RGB features can be decomposed into three components [Magrini et al., CVPR 2026]:

- s : information that are shared between modalities
- u : information that is RGB-specific
- v : information that is Event-specific

u and v are independent given s .

Multimodal models usually require a strict synchronicity: both temporal and spatial.

FRED: The Florence RGB-Event Drone Dataset

A field of where multimodality can help is drone detection:

- Targets are small
- Targets can be extremely fast as well as stationary
- Variable sizes and models





Challenging conditions

Night



Shadows



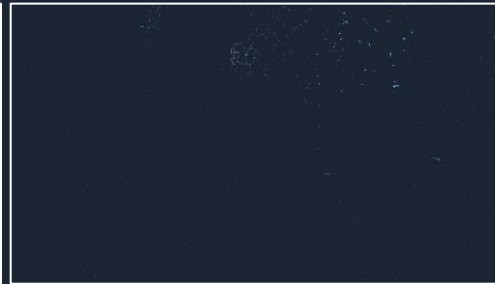
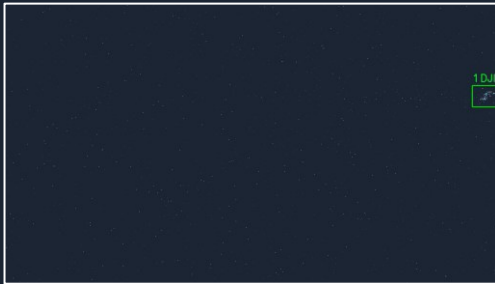
Lens Flare



Insects



Rain



Several challenging conditions:

- RGB-based models struggle with night, shadows, lens flare
- Insects and rain act as distractors for Event-based models

Challenging Split

- Presence of domain shifts: Day → Night, Day → rain, etc.
- Changes affect either RGB, event, or both modalities
- Designed to test model generalization under real-world conditions



1 Betafpv air75



Drone Detection

Drone Tracking



Drone Forecasting



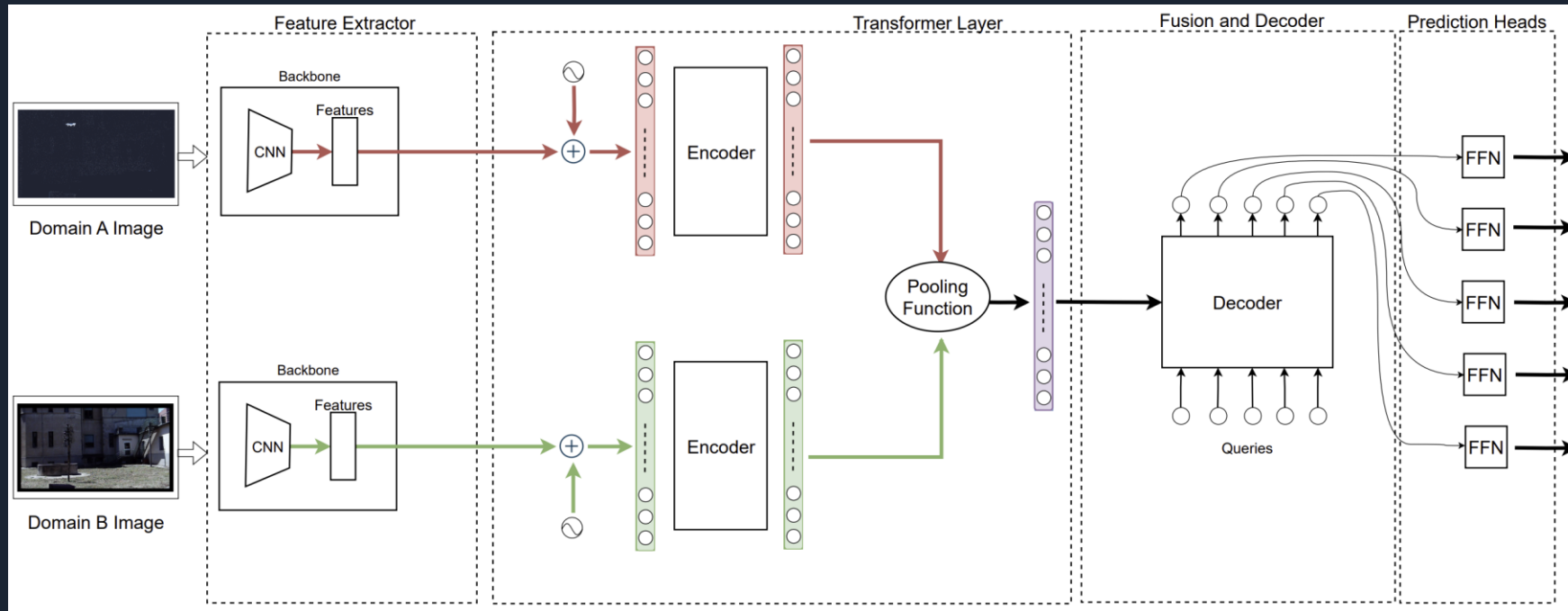
Statistics

1280x720 resolution

Spatiotemporally synchronized RGB+Event streams

7h:07m per modality

Multimodal Drone Detection



[Magrini et al., ECCVw 2024]

We adapted the DETR architecture to process multimodal streams of synchronized frames.

Event-based models largely outperform RGB ones.

Multimodal models improve upon single-modality ones.

Method			Canonical Split		Challenging Split	
	Event	RGB	mAP ₅₀	mAP _{50:95}	mAP ₅₀	mAP _{50:95}
ER-DETR	✓	×	68.42	21.80	62.20	19.81
ER-DETR	×	✓	28.56	7.73	15.84	4.11
ER-DETR	✓	✓	78.59	32.21	75.73	27.87

Can multimodality help single modality models?

Event frames and RGB frames capture complementary signals:

- Events are more robust to illumination changes (domain shift)
- RGB capture more semantic information

Given a single modality, is it possible to mimick the other one?

A modality can be accessed at training time (as Privileged Information) and not used at inference time.

Feature Hallucination

We train a two-branched model, where one modality tries to hallucinate the other.

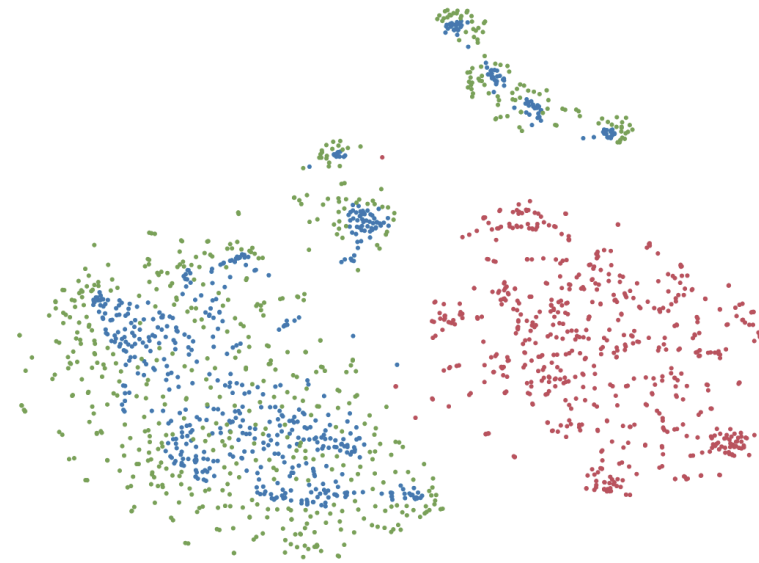
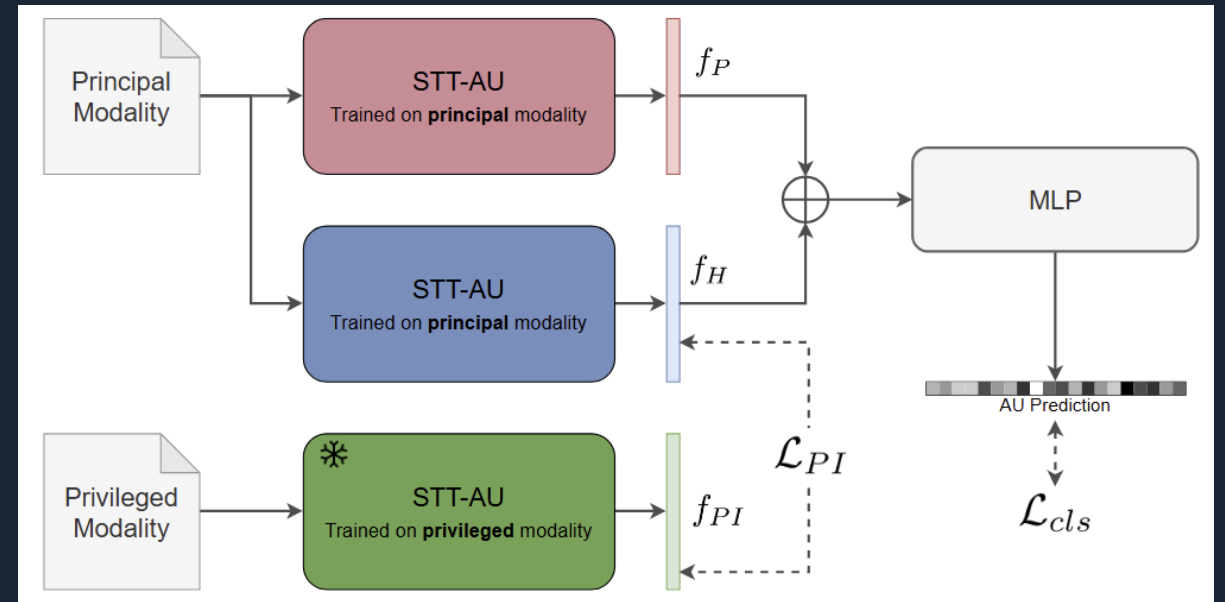
The number of network parameters doubles.

Model	RGB	Event	Accuracy
STT-AU	✓	✗	44.93
	✗	✓	48.78
	✓	✓	56.84
	✓	hallucinated	49.99
	hallucinated	✓	51.05

Results on the FACEMORPHIC dataset

Method	DETR		RT-DETR	
	mAP ₅₀	mAP _{50:95}	mAP ₅₀	mAP _{50:95}
Baseline	15.8	4.1	21.1	7.1
with Feat. Hallucin.	18.9	5.3	25.0	6.5

Results on the FRED dataset for RGB models



Red: RGB

Green: Event

Blue: hallucinated Event

PEPR: Privileged Event-based Predictive Regularization



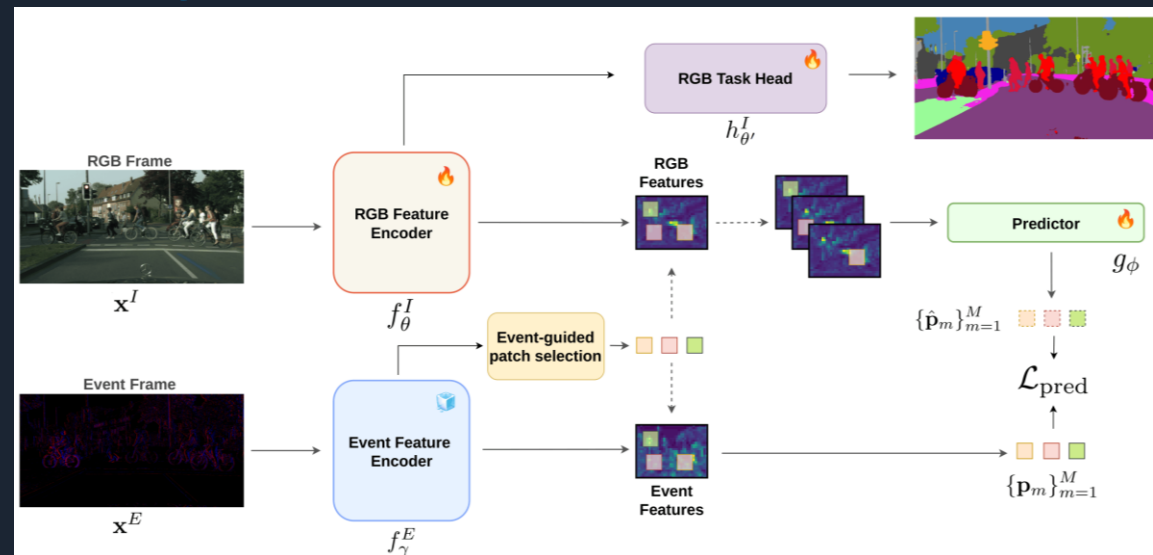
We would like to keep the inference architecture unaltered.

Training time

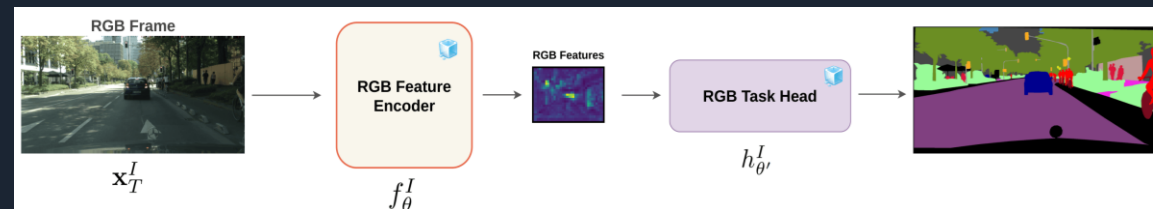
We let the RGB model predict latent Event patches (JEPA-style [Assram 2023]) at training time.

Direct Event-RGB alignment (whole image) can be detrimental as RGB is dense and Events are sparse.

Perceptive task loss prevents modality collapse.

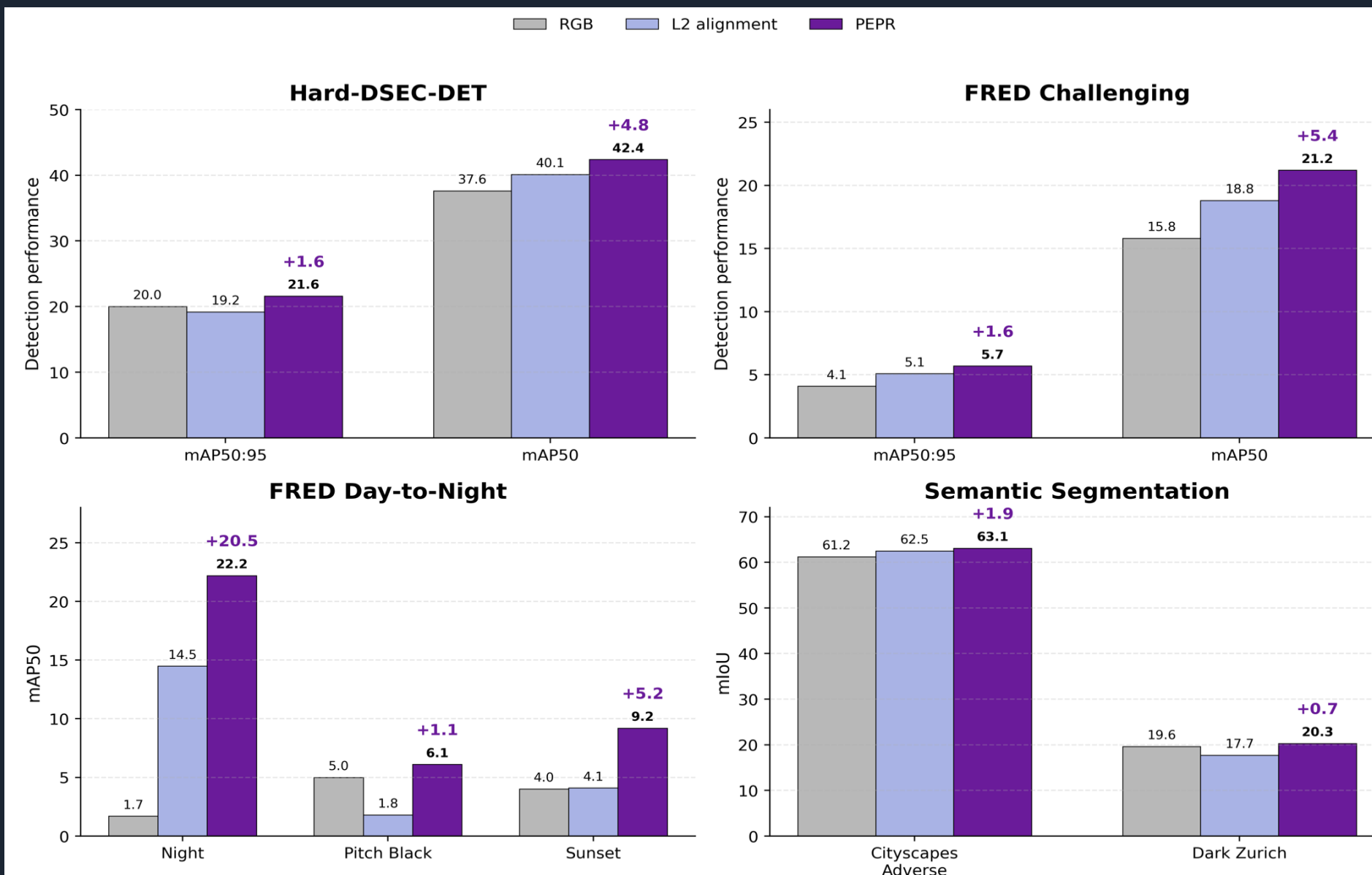


Inference



Event Guided Patch Selection: PEPR samples event-active and low-activity regions for predictive supervision

PEPR: Privileged Event-based Predictive Regularization



Conclusions

RGB and Event streams capture complementary information.

Effectively leveraging both domains is not always straightforward:

- Different ways of fusing modalities
- Depending on data alignment (temporal/spatial synchronicity) some forms of fusion might not be possible

Even if a modality is missing at test time, it can be used at training time as privileged information.

Making the most out of events and frames

advantages and challenges of multimodal applications

Federico Becattini
University of Siena



**UNIVERSITÀ
DI SIENA**
1240