SMU SINGAPORE MANAGEMENT UNIVERSITY | School of Computing and Information Systems

SaccadeX

Directed Acyclic Graph-based Semi-Supervised Learning of Continuous Ocular Dynamics from Sparse Neuromorphic Streams

Nuwan Bandara, Thivya Kandappu, Archan Misra

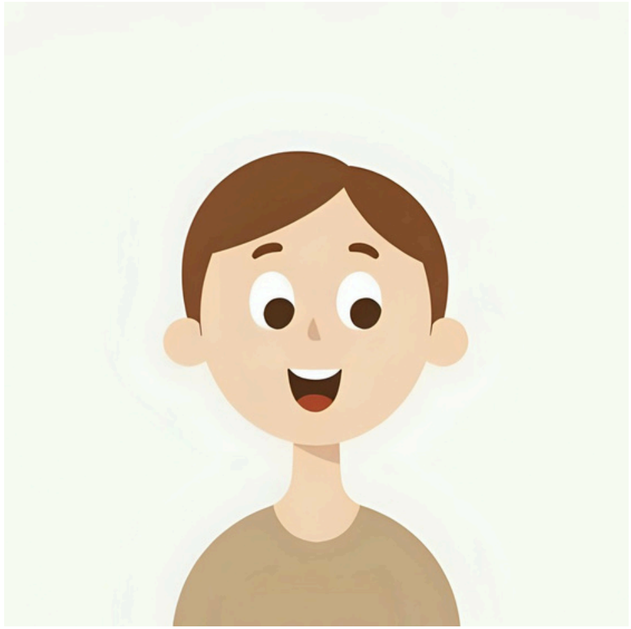
School of Computing and Information Systems, Singapore Management University

IEEE/CVF Winter Conference on Applications of Computer Vision 2026

WACV 2026 Tucson, Arizona • Mar 6 - 10, 2026

High-Fidelity Eye Tracking

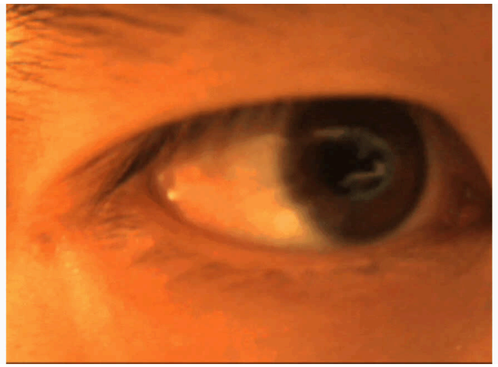
Why **Fine-grained, High-frequency** Eye Tracking?



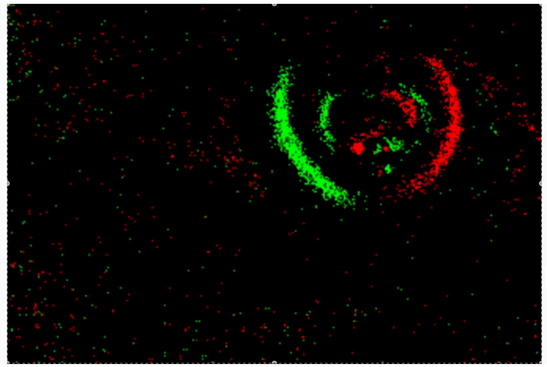
*Images, except for the authentication logo, are generated using Canva AI image generator

RGB vs Event Vision

Eye, being the **fastest mechanical organ** in human body, exhibits **rapid** and **intricate** movements including **pupillary acceleration reaching values as high as 24, 000 deg/s²**



- Poor temporal resolution
- Higher susceptibility to motion blur and low lighting conditions
- Higher power consumption
- Synchronous and **Dense**



- **Higher temporal resolution**
- **Lower susceptibility to motion blur** and **low lighting** conditions
- **Lower power** consumption
- **Asynchronous** and Sparse

Event-based Eye Tracking

Approach Set 1: **RGB-guided** Pupil Localization in Events

Approach Set 2: **Exclusive Event-based** Pupil Localization

Suffers From

Mismatch with asynchronous event data

Commonly used **Dense 2D** frames are **inadequate to capture full temporal** info.

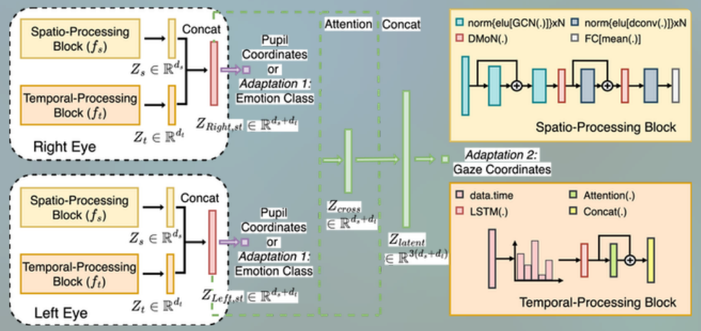
Further, both approaches have shortcomings of:

- Guided by coarser labels present at fixed timesteps: **label sparsity***
- **Sub-optimal**** event accumulation, prohibiting efficient online streaming
- Fail to capture fine-grained temporal relationships **within event volumes**

*Especially pronounced in ocular datasets, where annotations are sparse, and the underlying dynamics are subtle and high-frequency

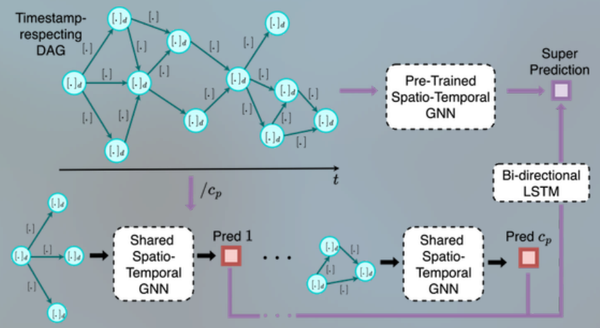
**Computationally Expensive

To address the temporally inadequate 2D representations



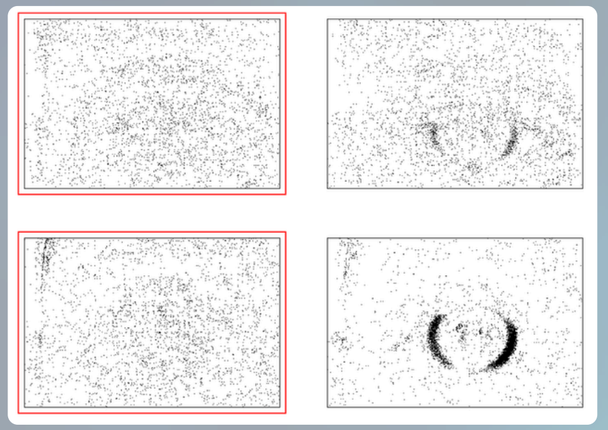
Directed Acyclic Graph (DAG) representation that preserves spatiotemporal structure and causality

To address the label sparsity via explicit temporal modelling within event volumes



Semi-supervised label propagation via a teacher–student pipeline

To address the sub-optimal accumulation for efficient streaming



Density-based event accumulation for online tracking

SaccadeX

Evaluated on three ocular tasks using four benchmark datasets, demonstrating the benefits of event-centric learning for continuous ocular dynamics modelling

DAG-based Spatio-Temporal GNN:
DAG Construction

Objective:

To **preserve both temporal resolution** and **local connectivity** of events while ensuring **causal consistency** via explicit **temporal ordering**

Implementation:

Given an event sequence: $E^v = \{v_i\}_{i=1}^{L}$ with each event: $v_i = (x_i, y_i, t_i, p_i)$

We construct a DAG*: $G = (V, E)$ under **conditions** below:

- Nodes: $V = \{v_1, \ldots, v_n\};\ |V| = n$

  $\mathbf{p}_{v_i} = [\lambda_1 t_i, \lambda_2 x_i, \lambda_3 y_i]$

  $\mathbf{x}_{v_i} = [\lambda_1 t_i, \lambda_2 x_i, \lambda_3 y_i, p_i] \in \mathbb{R}^4$

- Directed Edges: $E \subseteq V \times V;\ |E| = m$

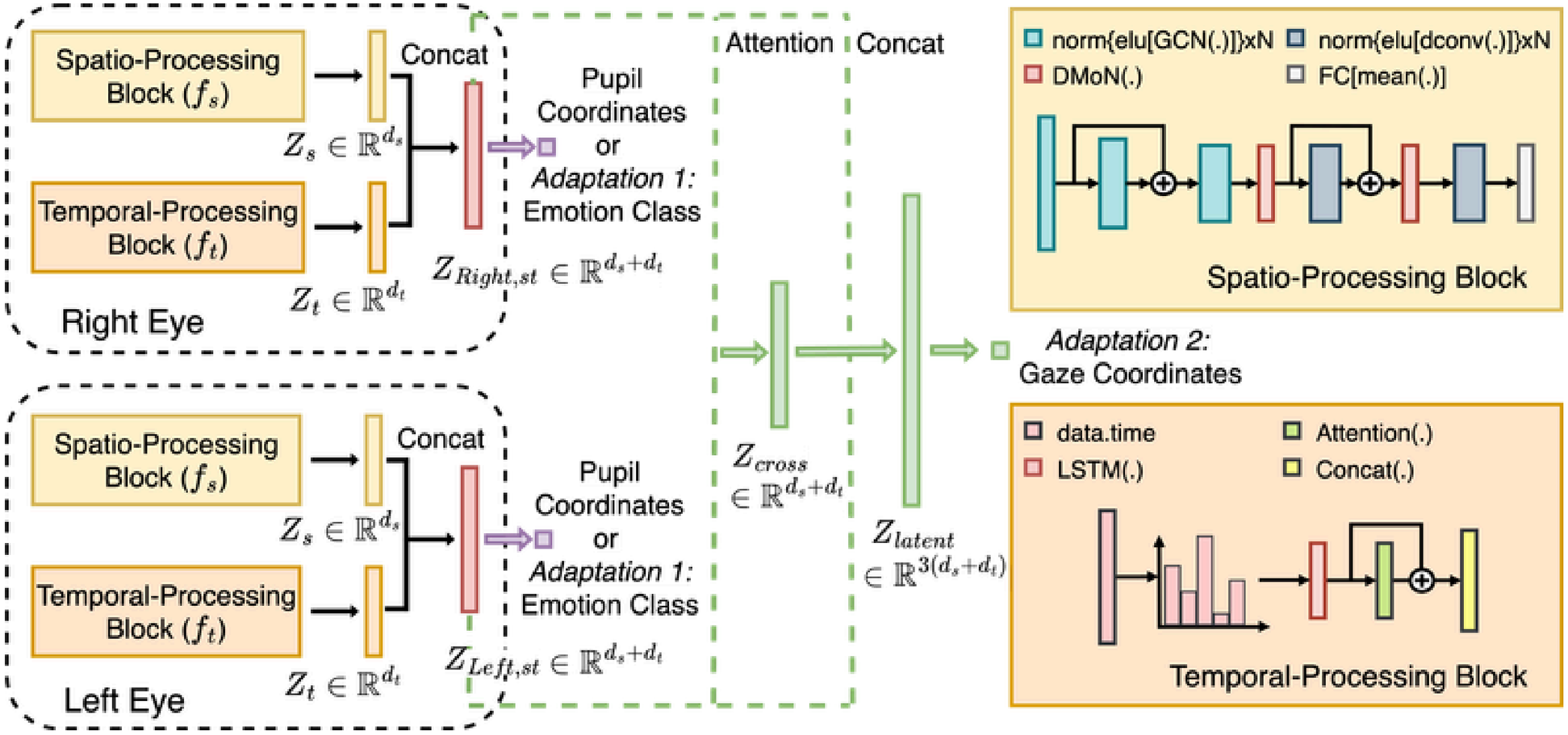
  v_i to $v_j \ \leftrightarrow\ \|\mathbf{p}_{v_i} - \mathbf{p}_{v_j}\|_d \leq \xi$ and $t_{v_i} < t_{v_j}$

  $D_{\min} \leq \sum_{j=1}^{n} A_{ij} \leq D_{\max}$ and Hawkes-encodings[1]

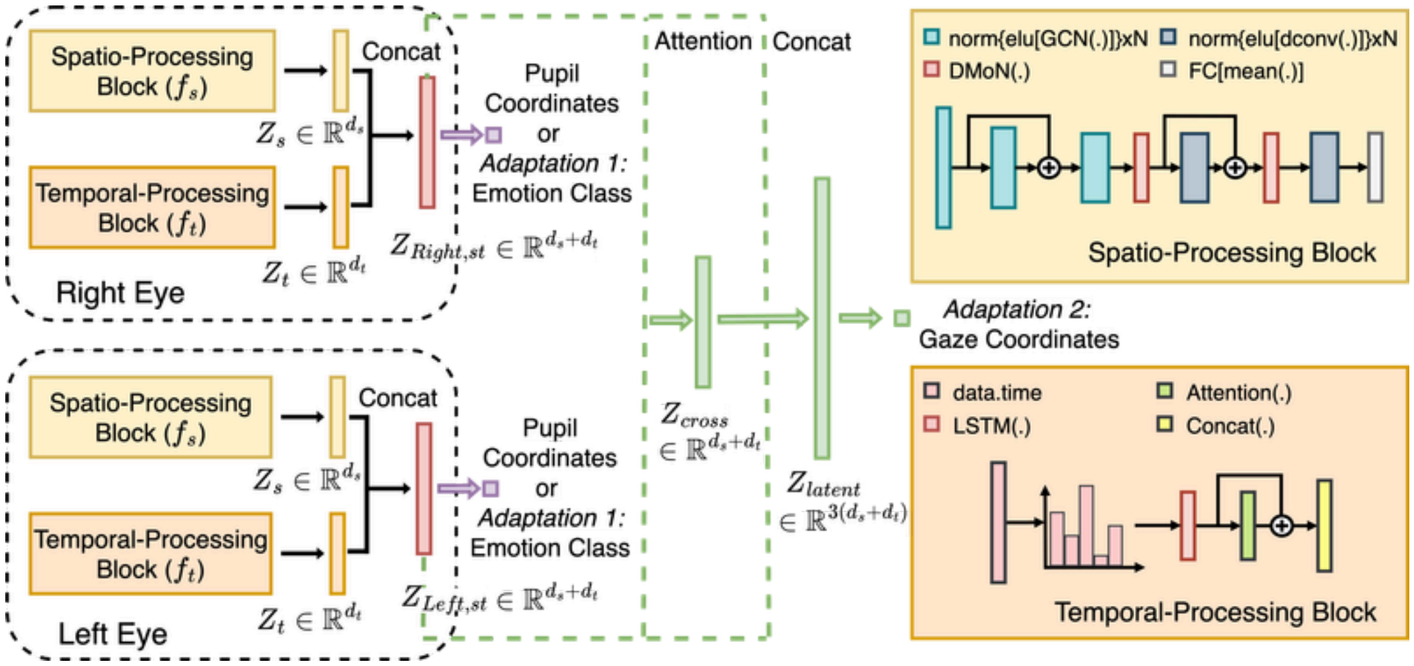
*Theoretical justification of DAG's appropriateness to represent event data is provided in the supplementary

[1] N. Bandara, et al., EyeGraph: Modularity-aware spatiotemporal graph clustering for continuous event-based eye tracking, NeurIPS 2024

DAG-based Spatio-Temporal GNN:
Dual-Branch Architecture

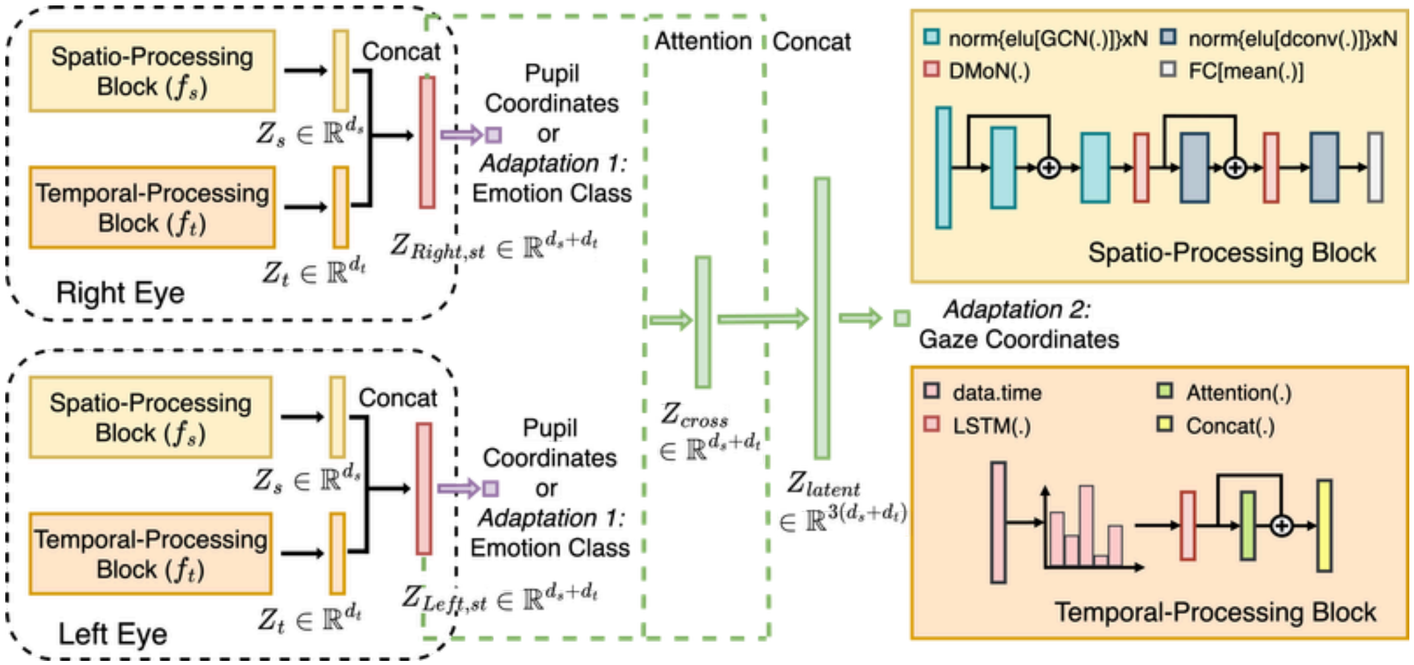


DAG-based Spatio-Temporal GNN:
Dual-Branch Architecture



Based on **Spatial modularity**
(i.e., events cluster around anatomical regions)
and thus,
the encoder, f_s, which is a multi-layer GCN with DMoN,[1]
is to yield embeddings
$Z_s \in \mathbb{R}^{d_s}$
with **anatomically coherent clusters**

DAG-based Spatio-Temporal GNN:
Dual-Branch Architecture

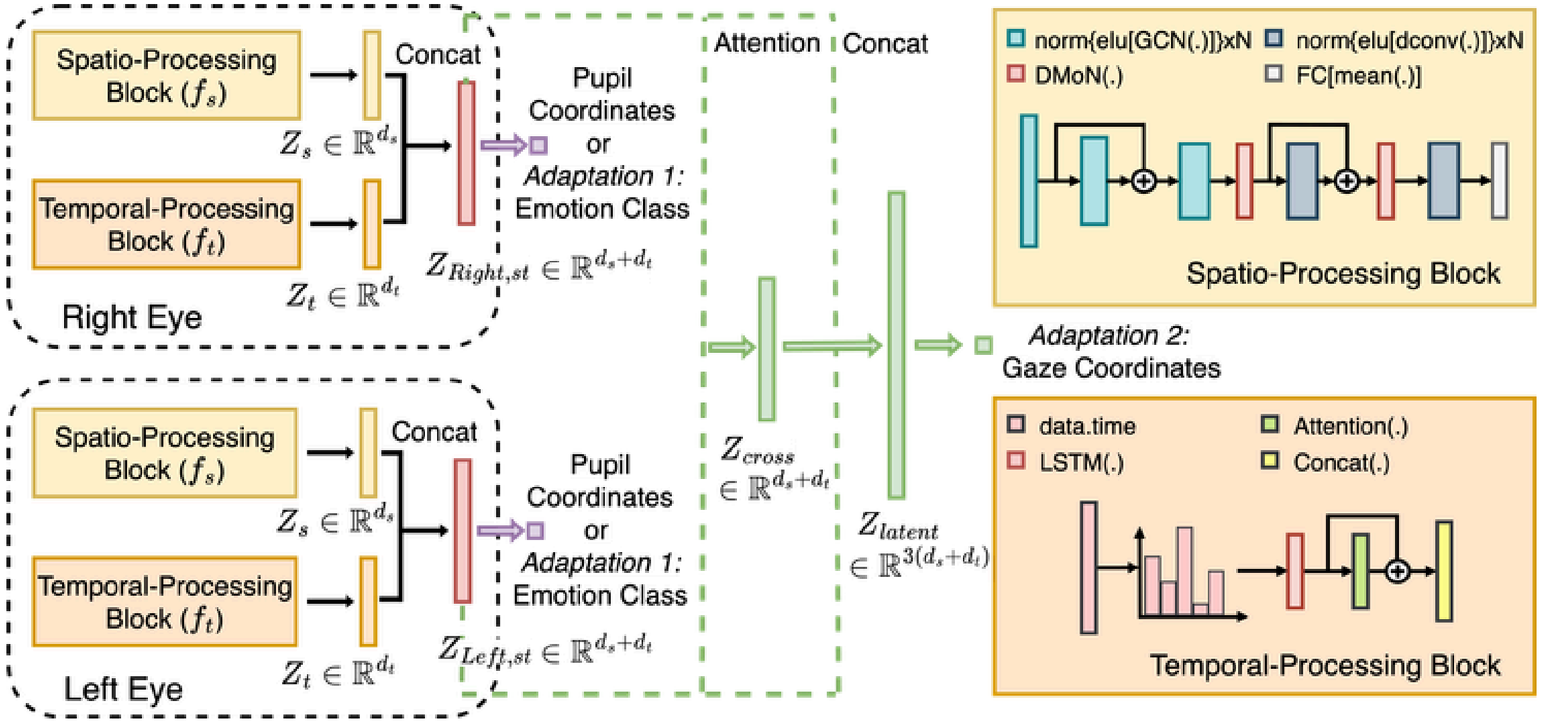


Based on **Temporal Continuity** (i.e., motion patterns evolve smoothly, reflecting fine-grained physical dynamics) and thus, f_t encodes cumulative timestamps into embeddings

$$Z_t \in \mathbb{R}^{d_t}$$

preserving motion continuity

DAG-based Spatio-Temporal GNN:
Composite Loss Function



Optimize for both **prediction accuracy** and **modular structure** using:

$$\mathcal{L} = \gamma_1 \left\| \mathbf{y} - \hat{\mathbf{y}} \right\|_2^2 - \gamma_2 \frac{\mathrm{Tr}(\mathbf{B}\mathbf{X}\mathbf{X}^\top)}{2m} + \gamma_3 \frac{\sqrt{C}}{n} \left\| \sum_i \mathbf{C}_i^\top \right\|_F - 1$$

DAG-based Spatio-Temporal GNN: Curriculum Learning to Address Graph Heterogeneity

- Event-based motion data exhibits **inherent variability in the size and structure*** of constructed DAGs, resulting in **training instability**

- To address this, we adopt curriculum learning with **progressive training stages**. The model first learns from simpler graphs (i.e., those with higher node counts) before incorporating more **complex graphs that are smaller, noisier, or spatially incomplete**.

Algorithm 1 Curriculum learning-based training pipeline

Require: Continuous event stream $E^{v}_{(t_i,x_i,y_i,p_i)}$ $\forall i \in N_{samples}$ in graph space $\mathbf{G}$, the labels y_i $\forall i \in N_{samples}$ DAG-based learning model f, the number of steps N_{st}

Initialize complexity flags $C = \{c_i \mid \forall i \in N_{samples}\}$ with zeros for graphs G_i $\forall i \in N_{samples}$

Calculate the number of nodes range $\min(n_{G_i})$, $\max(n_{G_i})$ $\forall i \in N_{samples}$

Map G_i $\forall i \in N_{samples}$ to complexity level: $f_{cp}[\min(n_{G_i}), \max(n_{G_i}), n_{G_i}, N_{st}] : G_i \mapsto \{1, .., N_{st}\}$ s.t. $n_{G_i} > n_{G_{i-1}} \longrightarrow c_i > c_{i-1}$

Re-order $\mathbf{G}$: $\mathbf{G} \longleftarrow f_{sort}[\mathbf{G}, C]$

while Training **do**

  if $c_i = N_{st}$ **then**

    $f : G_i \mapsto y_i$

  end if

  ...

  if $c_i = 1$ **then**

    $f : G_i \mapsto y_i$

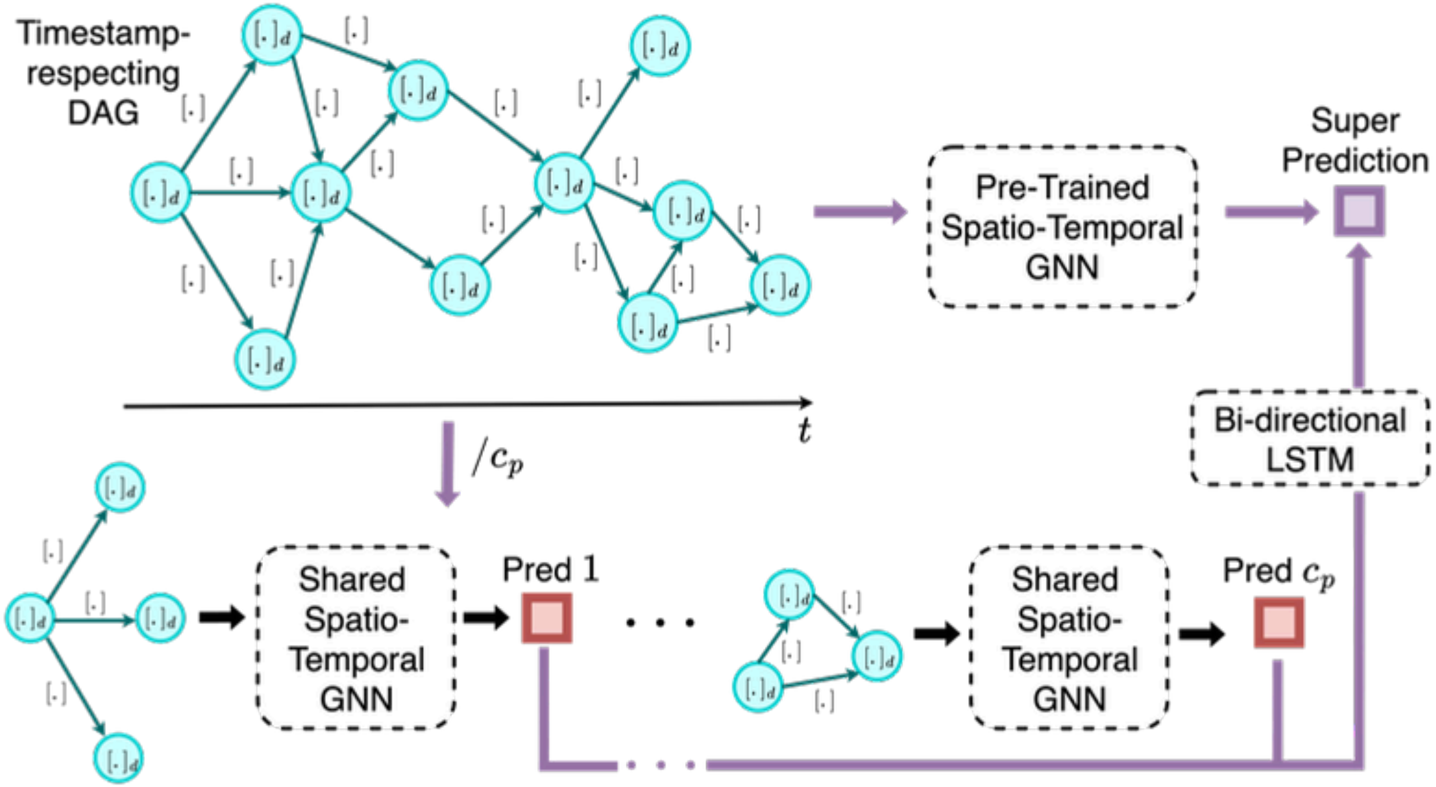
  end if

end while

*Empirical evidence is provided in the paper and supplementary

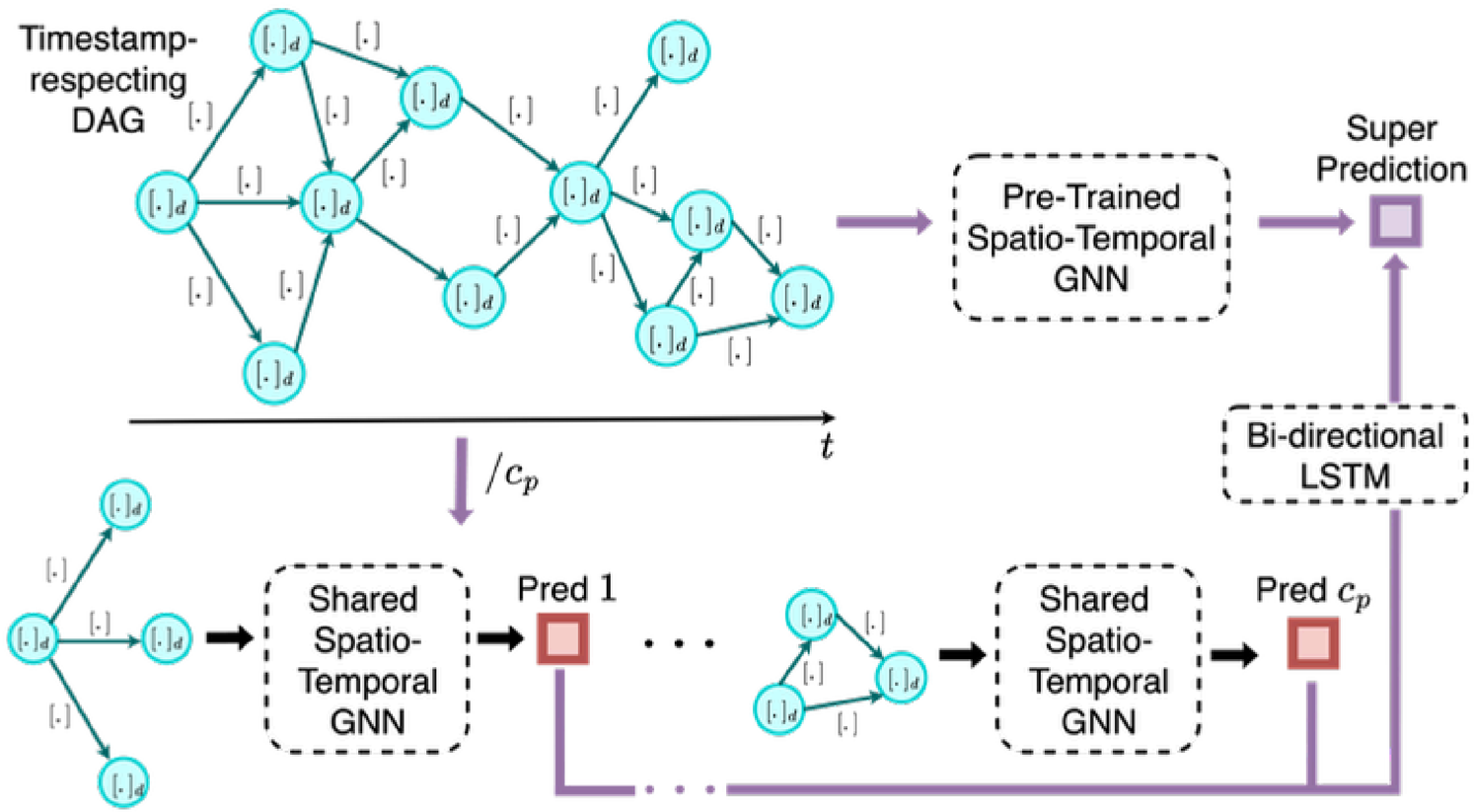
Semi-Supervised Learning Framework:
Self-Training Pipeline

- To facilitate **high-frequency ocular dynamics modelling** beyond scarce RGB-aligned labels, we propose a semi-supervised learning strategy, where **labels are interpolated across intermediate events within accumulated volumes** while ensuring temporal consistency



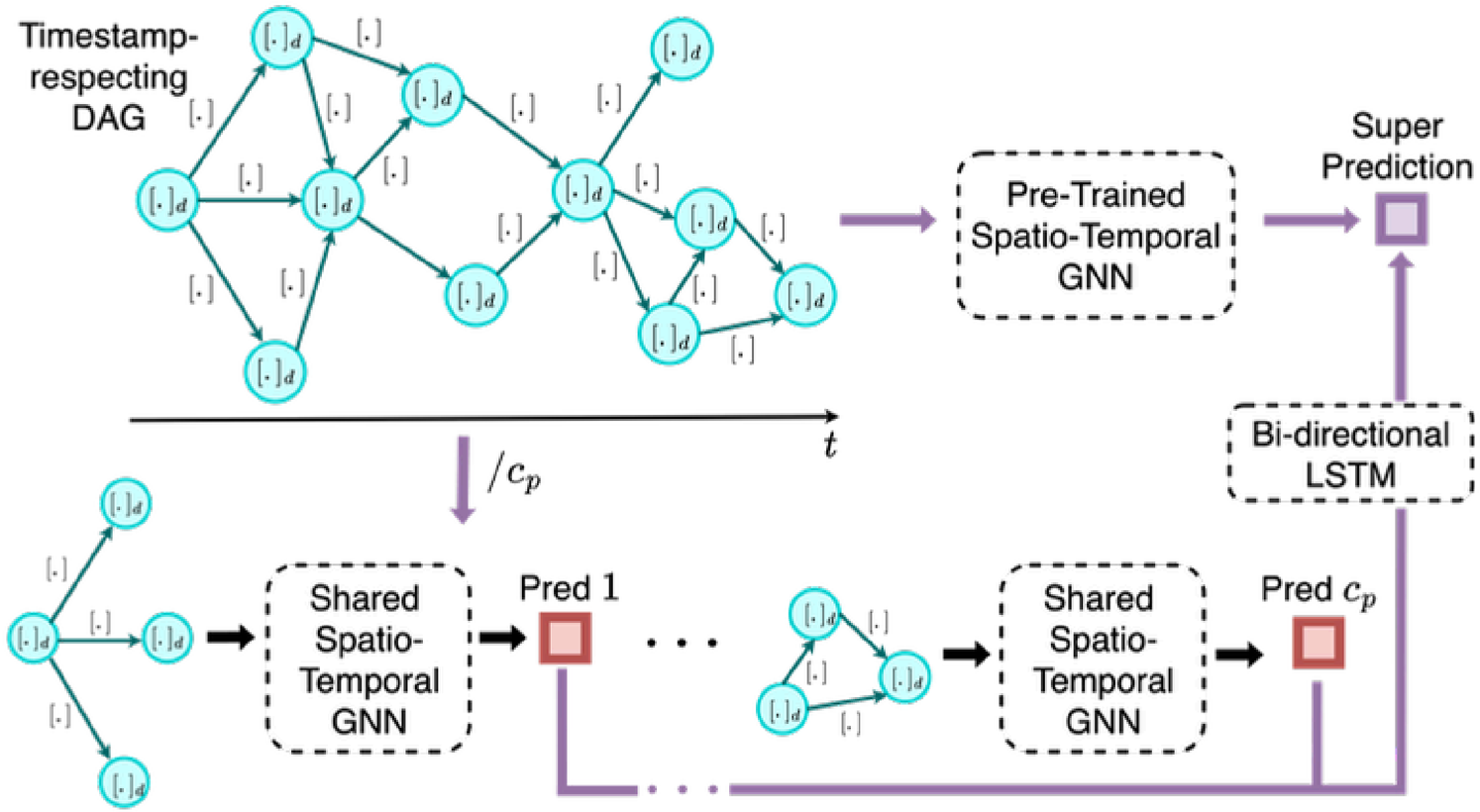
Semi-Supervised Learning Framework:
Self-Training Pipeline

- **Event Volume Partitioning**
- Teacher-Student Architecture
- Bidirectional Training Process



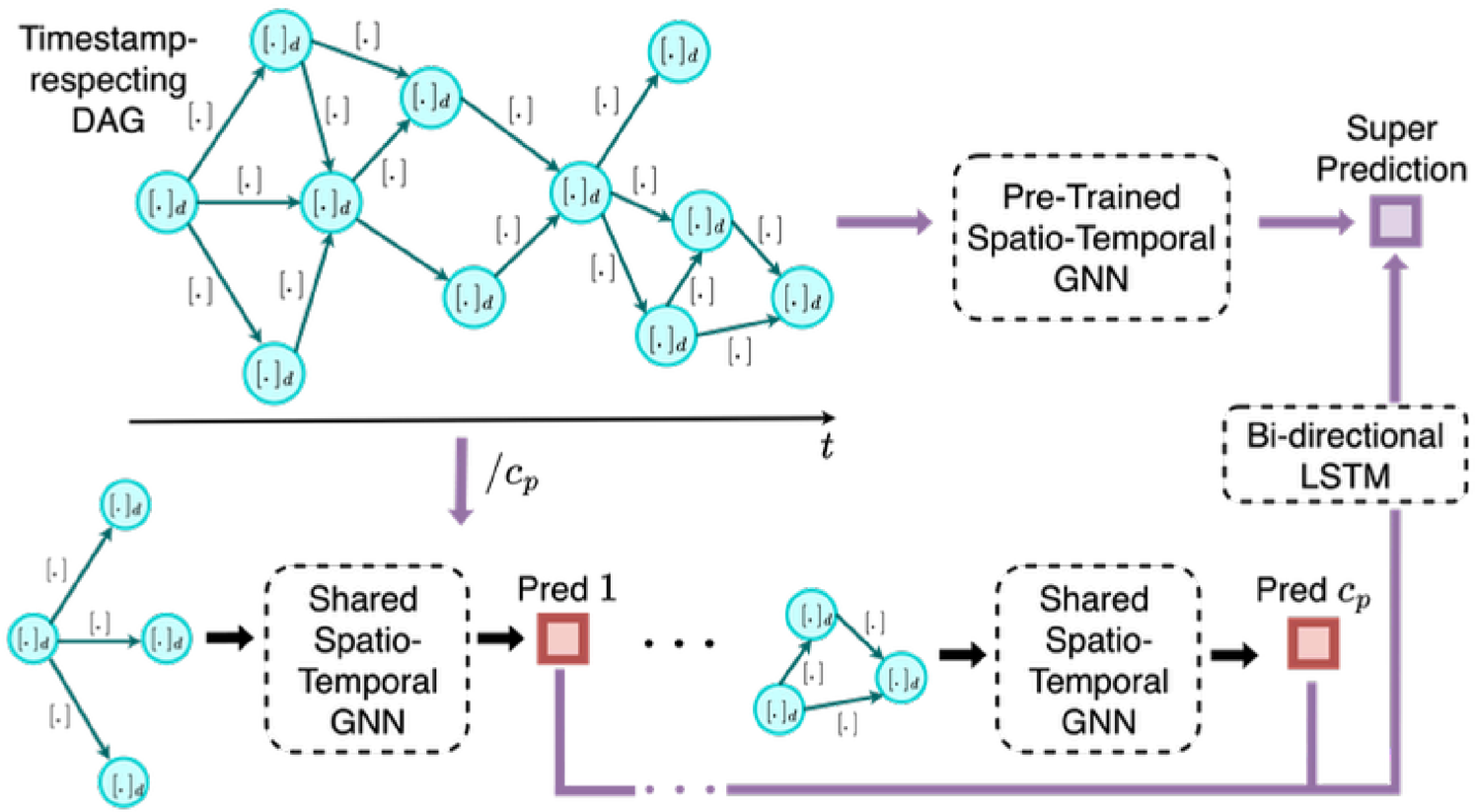
Semi-Supervised Learning Framework:
Self-Training Pipeline

- Event Volume Partitioning
- **Teacher-Student Architecture**
- Bidirectional Training Process



Semi-Supervised Learning Framework:
Self-Training Pipeline

- Event Volume Partitioning
- Teacher-Student Architecture
- **Bidirectional Training Process**



Density-based Event Accumulation for Online Tracking

- Existing strategies face unique challenges:
  - **Fixed time-window:** Sensitive to motion speed and illumination
  - **Fixed event volume:** Risk resulting in blurred frames
  - **Adaptive:** Prohibitive processing overhead

- In response, we modify the fixed volume approach by introducing a **lightweight density-based pre-processing step**
  - The key intuition is that **local spatio-temporal variations in event density**, when compared against global patterns, **often signal underlying physical dynamics**

Algorithm 2 Online density-based event accumulation

Require: Continuous event stream $E^v_{(t,x,y,p)}$, density threshold Δ_d, time bin T_b, number of spatial partitions N_P and event spatial resolution $[x_{max}, y_{max}]$

Set (spatial) partition width: $W_P = x_{max}//\sqrt{N_P}$, height: $H_P = y_{max}//\sqrt{N_P}$ {Here, // refers to integer division}
Set number of partitions in x: $N_x = x_{max}//W_P$, in y: $N_y = y_{max}//H_P$
Initialize time $t = 0$
while Inference **do**
  while $t\%T_b \; != \; 0$ **do**
    Initialize partition density frame: $D_P = [0]_{N_x \times N_y}$
    Collect events $E^v \longleftarrow e_i(t, x, y, p)$
  end while
  for each $e_i(t, x, y, p) \in E^v$ **do**
    Map $e_i(.)$ to partition: $E^v \mapsto \{1, ..., p_i, ..., N_P\}$
    $D_{P\{i\}}[p_i] \longleftarrow D_{P\{i-1\}}[p_i] + 1$
  end for
  Calculate $\mu(D_P)$, $\max(D_P)$
  if $\max(D_P) > (1 + \Delta_d) \times \mu(D_P)$ **then**
    Accumulate E^v
  end if
end while

Experiments & Results:
Supervised Learning Results

Pupil Tracking Results on 3ET+ Dataset

Method	# Prm.	p5↑	p3↑	p1↑	l_2↓	l_1↓
[42]	8.5M	97.05	90.73	33.75	1.67	2.11
[41]	7.1M	96.31	83.83	23.91	2.03	2.56
[31]	1.1M	**97.79**	**94.58**	45.50	1.44	**1.82**
[47]	465K	77.20	47.97	7.32	3.51	4.63
Ours††	**379K**	95.34	92.37	40.88	1.41	2.08
Ours†	**379K**	97.14	94.55	**46.80**	**1.40**	2.07

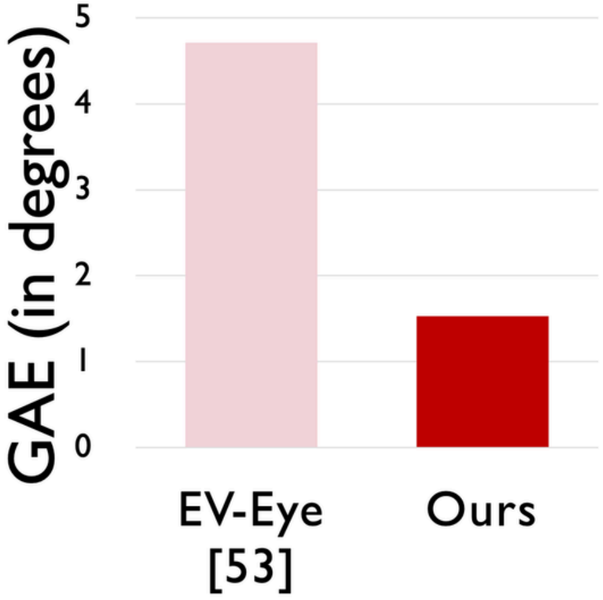
- Our method achieves competitive performance while using **significantly fewer parameters**. Key improvements include **38.75% improvement in p1 metric** over MambaPupil and **95.5% parameter reduction**

[42] Zhong Wang et al. Mambapupil: Bidirectional selective recurrent model for event-based eye tracking. CVPR, 2024
[41] Zuowen Wang et al. Event-based eye tracking. AIS 2024 challenge survey. CVPR, 2024
[31] Yan Ru Pei et al. A lightweight spatiotemporal network for online eye tracking with event camera. CVPR, 2024
[47] Baoheng Zhang et al. Co-designing a sub-millisecond latency event-based eye tracking system with submanifold sparse CNN. CVPR, 2024

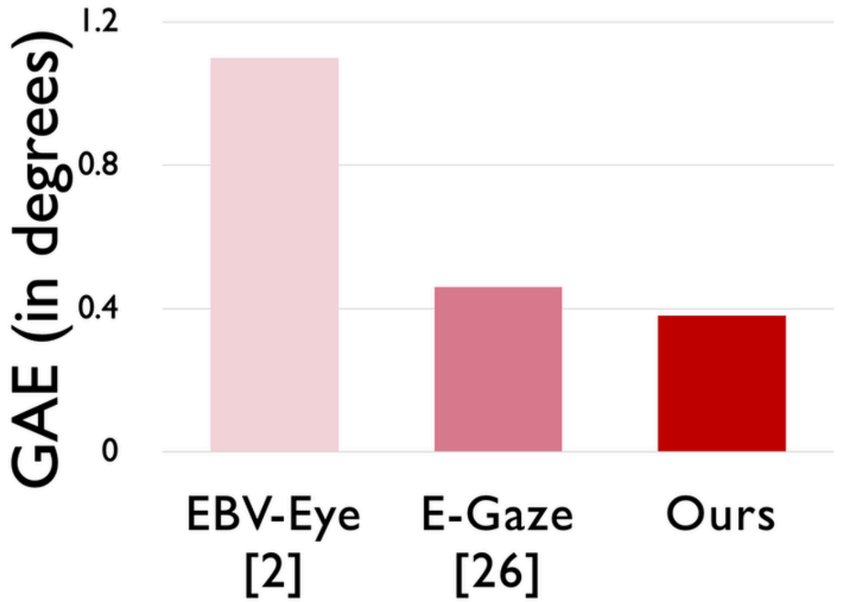
† With curriculum learning, †† Without curriculum learning

Experiments & Results:
Supervised Learning Results

Gaze Estimation Results



On EV-Eye Dataset



On EBV-Eye Dataset

- Our method achieves a mean GAE of 1.53 degrees on EV-Eye and 0.38 degrees on EBV- Eye, representing **68% and 63% error reductions** compared to existing SOTA methods

[53] Guangrong Zhao et al. EV-Eye: Rethinking high-frequency eye tracking through the lenses of event cameras. NeurIPS, 2024.
[2] Anastasios Angelopoulos et al. Event-based near-eye gaze tracking beyond 10,000 Hz. IEEE TVCG 2021
[26] Nealson Li et al. E-gaze: Gaze estimation with event camera. IEEE TPAMI, 2024

Experiments & Results:
Supervised Learning Results

Emotion Recognition Results on SEE[48] Dataset

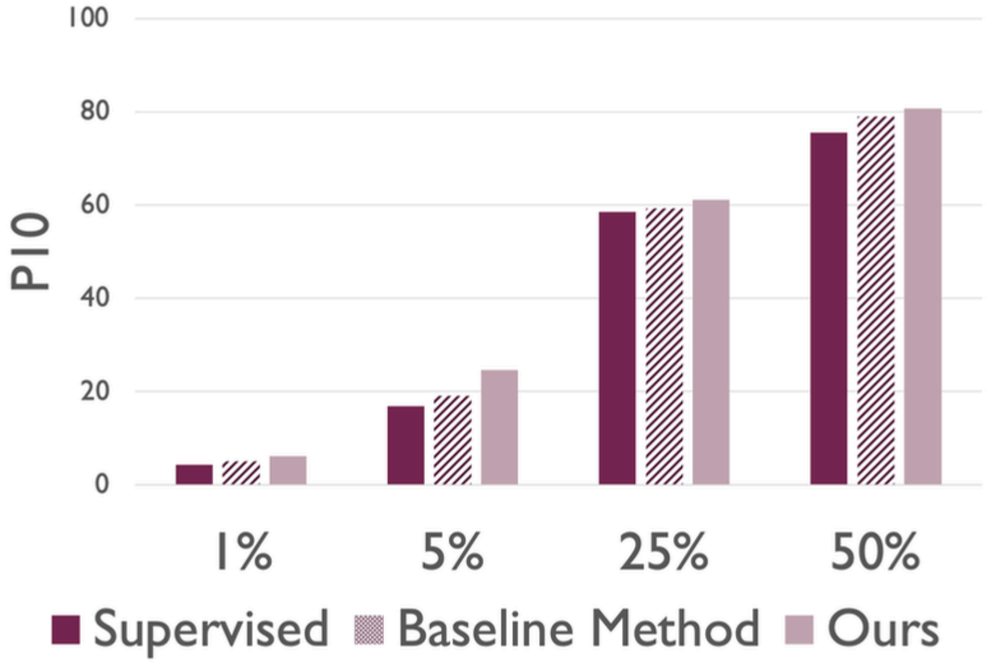
Method	Type	$WAR \uparrow$	$UAR \uparrow$
Former DFER [51]	Face	65.8	67.2
EyeMotion [16]	Eye	78.8	79.5
EMO [44]	Eye	63.1	63.3
SEE [48]	Eye	<u>83.6</u>	<u>84.1</u>
Ours	Eye	**86.9**	**87.4**

- Our method consistently **outperforms all baselines with at least 3.3% improvement** in both WAR (86.9%) and UAR (87.4%) metrics compared to the previous best method, SEE [48]

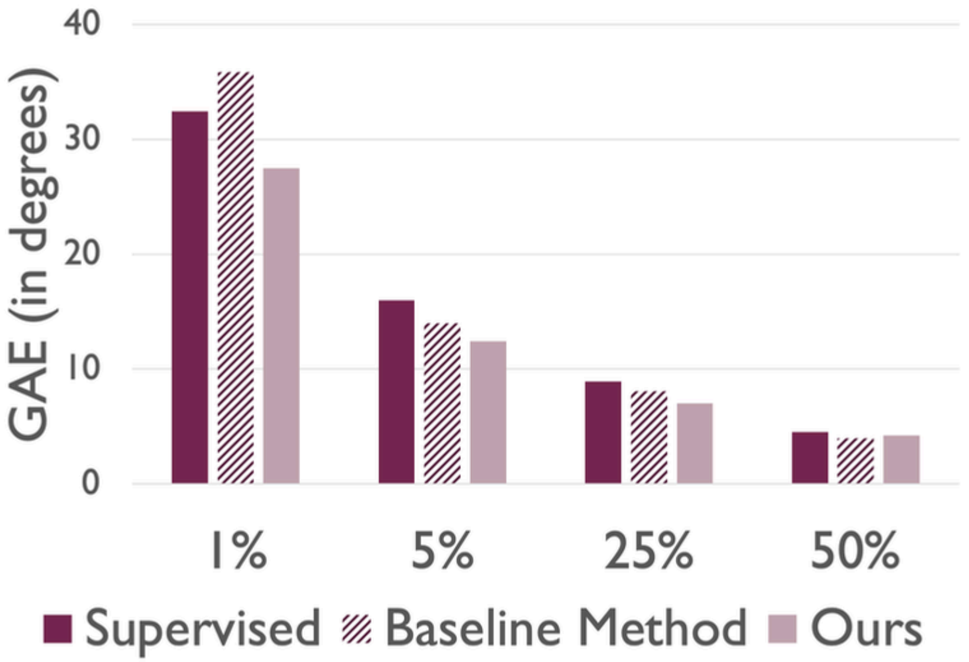
[51] Zengqun Zhao and Qingshan Liu. Former-dfer: Dynamic facial expression recognition transformer. ACM Multimedia, 2021
[16] Steven Hickson et al. Eyemotion: Classifying facial expressions in vr using eye-tracking cameras. WACV, 2019
[44] Hao Wu et al. EMO: Real-time emotion recognition from single-eye images for resource-constrained eyewear devices, MobiSyS, 2020.
[48] Haiwei Zhang et al. In the blink of an eye: Event-based emotion recognition, ACM SIGGRAPH, 2023

Experiments & Results:
Semi-Supervised Learning Results

Under varying label settings*



Pupil Tracking on 3ET+ dataset



Gaze Tracking on EBV-Eye dataset

- In both tasks, our semi-supervised pipeline **consistently outperforms both supervised counterparts and vanilla self-training baselines** across all label ratios

*Refer to the section 6 in the paper for more details

Experiments & Results:
More Results

- More results on:
  - **Computational Efficiency**
  - **Density-based Event Accumulation**
  - Ablations on:
    - Event representation efficiency
    - Graph architecture
    - Curriculum learning
    - RGB frame interpolation

  can be found in the paper and supplementary

Task	Method	# Prm.	TT/Ep.†	Lt.*	FLOPs
Pupil	[9]	417K	7.9 s	–	1.09T
	[42]	8.5M	5.4 s	–	2.61T
	[31]	1.1M	–	119ms	0.95T
	Ours	379K	3.3 s	29 ms	0.90T
Gaze	[50]	17.27M	–	700 ms	1.29T
	Ours	701K	1.4h	37 ms	0.42T

Computational Efficiency Comparison

Limitations & Future Work

- While our framework optimizes spatio-temporal learning, processing large-scale event streams still **remain computationally demanding**, particularly for embedded and resource-constrained devices

- Any inaccuracies or biases in the ground truth (or teacher labels) **propagate through the pseudo-labeling process**, potentially affecting model generalization

- Future works will explore further optimizing these (or similar) pipelines to **integrate them into resource-constrained devices**

THANK YOU

Check out more details of SaccadeX at

https://eye-tracking-for-physiological-sensing.github.io/SaccadeX/

If any clarifications, please contact us at

pmnsribandara@gmail.com