A Spatio-temporal Event Transformer on Versatile Tasks for Human Behavior Analysis

Kejing Xia[1,†], Lixuan Wei[1,†] and Lei Yu[1,*]

[1]*Electronic Information School, Wuhan University, Wuhan, Hubei Province, P.R.China. 430072*

Abstract

Human behavior encompasses a range of physical actions and changes exhibited by individuals, including gesture variations, facial expressions, and physiological state changes. Given the diversity and complexity of these behaviors, a versatile and robust analytical method is highly necessary. We introduce the **S**patio-**t**emporal **E**vent **T**ransformer (STET), the first approach to leverage event streams for multi-task human behavior analysis. By utilizing the unique characteristics of events, we can effectively extract features across different time scales (ranging from 1 μs to extended sequences). Specifically, we employ an E-Voxel Guided Multi-head Transformer (EVGMH) module to fuse event information at various scales, enhancing the representation of spatio-temporal features. Our experimental results on multiple task datasets demonstrate the efficacy of the STET method, establishing a new benchmark for event-based human behavior analysis.

Keywords

Human behavior, Event-based, Multi-task, rPPG, Micro-expression, Micro-gesture

1. Introduction

The field of human behavior analysis has deepened with the continuous development of computer vision [1]. This expansion not only enhances our understanding of human behavior, emotions, and physiological states but also drives advancements in areas such as healthcare, security, and human-computer interaction. However, due to the diversity and complexity of human behavior, various analysis techniques (e.g. heartbeat detection [2], facial expression recognition [3], and gesture understanding [4]) exhibit significant differences in their representation methods. For instance, heartbeat is characterized by periodicity, facial expressions by idiosyncratic and subtle variations, and gestures by idiosyncratic and large movements. Additionally, existing methods [5, 6, 7, 8] typically address only a single type of human behavior detection.

To address these issues and adopt a general approach to physiological measurements, current research has proposed video-based methods. These methods can conveniently replace traditional, highly specific non-visual methods with remote techniques, utilizing the learning capabilities of networks to uncover pixel-level subtle changes caused by physiological signals in images. In particular, some methods [5, 9] achieve remote photoplethysmography (rPPG) measurement through end-to-end network architectures, discarding contact-based ECG and

MiGA@IJCAI24: International IJCAI Workshop on 2nMicro-gesture Analysis for Hidden Emotion Understanding, August 3, 2024, Jeju, Korea.

*Corresponding author.

†These authors contributed equally.

xiakejing.lesia@whu.edu.cn (K. Xia); weilx_selina@whu.edu.cn (L. Wei); ly.wd@whu.edu.cn (L. Yu)

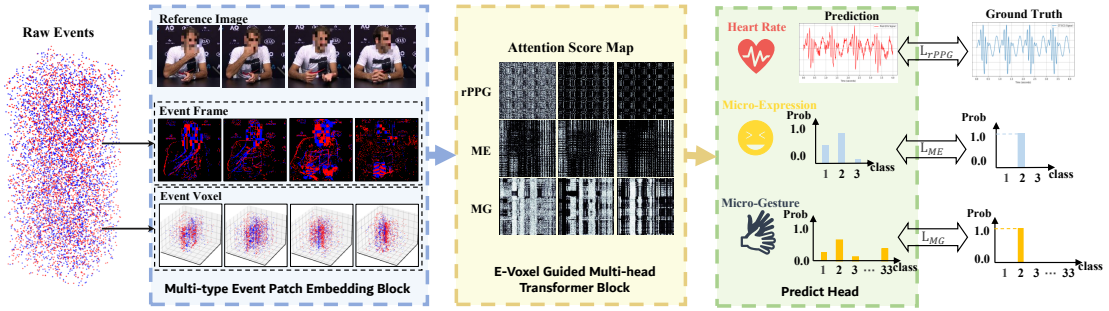

Figure 1: Overview of Spatio-temporal Event Transformer network, and visualization for middle outputs.

PPG methods in the process; In the field of micro-expression research, initial studies [6] employed high-frame-rate (200 FPS) videos to capture the requisite data, subsequently utilizing deep networks to comprehend the micro-level facial alterations; In the field of micro-gesture, JSSE [7] proposed a 3D-CNNs micro gesture recognition network that incorporates skeletal and semantic embedding loss to improve classifying results.

Event cameras are asynchronous visual sensors that sample brightness changes for each pixel and generate a stream of events with timestamps when the intensity change occurs [10]. Each event contains the coordinate, timestamp, and polarity of the brightness-changed pixel. Event cameras possess outstanding features, including high dynamic range (HDR), low latency ($1\mu s$), and low power consumption [10]. In comparison to frame-based cameras, event cameras are capable of capturing a greater quantity of useful behavioral information, such as facial blood vessel pulsations, expression changes, and gesture movements.

In light of these distinctive characteristics of event cameras, we are the first to propose an event-based method for human behavior analysis, i.e. STET. As shown in Figure 1, the end-to-end STET includes a Multi-type Event Patch Embedding (MTE) module, an E-Voxel Guided Multi-head Transformer (EVGMH) module, and task-specific prediction heads. The MTE module extracts event features from event formations to tokens. The EVGMH module is designed to enhance the feature and facilitate spatio-temporal feature representations, effectively decomposing the idiosyncratic and periodic features in the original input events. Finally, the feature representations are fed to the task-specific prediction heads for different tasks.

In summary, our contributions are as follows:

- We propose an end-to-end STET method, the first approach presenting a general framework utilizing events for multi-task human behavior analysis.
- We introduce an E-Voxel Guided Multi-head Transformer Block to make use of features from event frames that focus on global compressed information and event voxels that focus on local temporal information.
- We build a simulated event dataset, E-Human Behaviour (EHB) dataset, for multiple human behavior analysis tasks. Additionally, we conduct multiple experiments and provide an event-based experimental benchmark.

2. Related Works

2.1. Human Behaviour Analysis

Remote photoplethysmography measurement. The minute variations in light reflections from human facial skin can be utilized to extract heart rate [11]. Various HR estimation techniques have utilized information from different color channels and regions of interest (ROI) to recover rPPG signals, employing techniques like blind signal separation (e.g., independent component analysis (ICA) [12, 13]), matrix completion [14], and optimal space transformation [15]. Building on these foundations, several deep learning-based approaches have been developed [16, 17, 18, 19, 20, 21]. Špetlík et al. [20] introduced a two-stage, two-dimensional CNN method, while Chen et al. further advanced the field with convolutional attention networks for physiological measurement [21]. To better utilize spatio-temporal information in facial videos, several end-to-end spatio-temporal attention methods have been developed to directly recover rPPG signals from these videos [5, 9, 22, 23, 24, 25].

Micro-expression recognition. In micro-expression recognition, various methods based on optical flow, such as Main Directional Mean Optical Flow (MDMO) [26] and its improved versions like Facial Dynamics Map [27], Sparse MDMO [28], and Bi-Weighted Oriented Optical Flow [29], extract micro-expression features by calculating the average optical flow of facial regions. Additionally, some CNN and LSTM methods analyze micro-expressions by extracting spatio-temporal features from frames [30, 31, 32]. Due to the short duration and subtlety of micro-expressions, Yang proposed MERTA [33], which employs three attention mechanisms to construct feature maps. To further leverage the temporal and spatial features of micro-expressions, Micron-BERT [6] introduced a self-supervised framework μ-BERT consistently achieves State-of-the-Art result in various ME benchmarks.

Micro-gesture understanding. Body gestures can convey a wide range of emotions and mental states [34, 35, 36, 37]. For the classification task of pre-segmented body gestures, traditional RGB-based methods such as 2D CNN [38, 39, 40] and 3D CNN [41, 42] utilize frame sequence depth and optical flow for classification. Additionally, some approaches directly extract gestures by leveraging spatio-temporal relationships within frame sequences. The SlowFast model [43] captures both long-term and short-term temporal information through dual pathways. Shah et al. [44] employ graph-encoding convolutional networks to identify semantic features, while Huang et al. [8] propose a deep framework with an ensemble hypergraph-convolution Transformer. Recently, the JSSE model [7] introduced a 3D-CNNs micro-gesture recognition network that incorporates skeletal and semantic embedding loss to enhance classification results.

2.2. Event Camera in Computer Vision

Event cameras are bio-inspired visual sensors that generate asynchronous event streams [10] by responding to changes in brightness [45]. Each event consists of pixel location, timestamp, and polarity, distinguishing them from conventional RGB-based data. Due to the paradigm shift in visual information acquisition, event cameras possess unique characteristics such as extremely low latency, high dynamic range (HDR), and low power consumption [10]. These features drive research in various fields, including occlusion removal [46, 47], super-resolving [48, 49], and motion deblurring [50, 51]. In our task, event cameras have significant advantages in capturing subtle behaviors, including rPPG signals, micro-expression, and micro-gesture. The

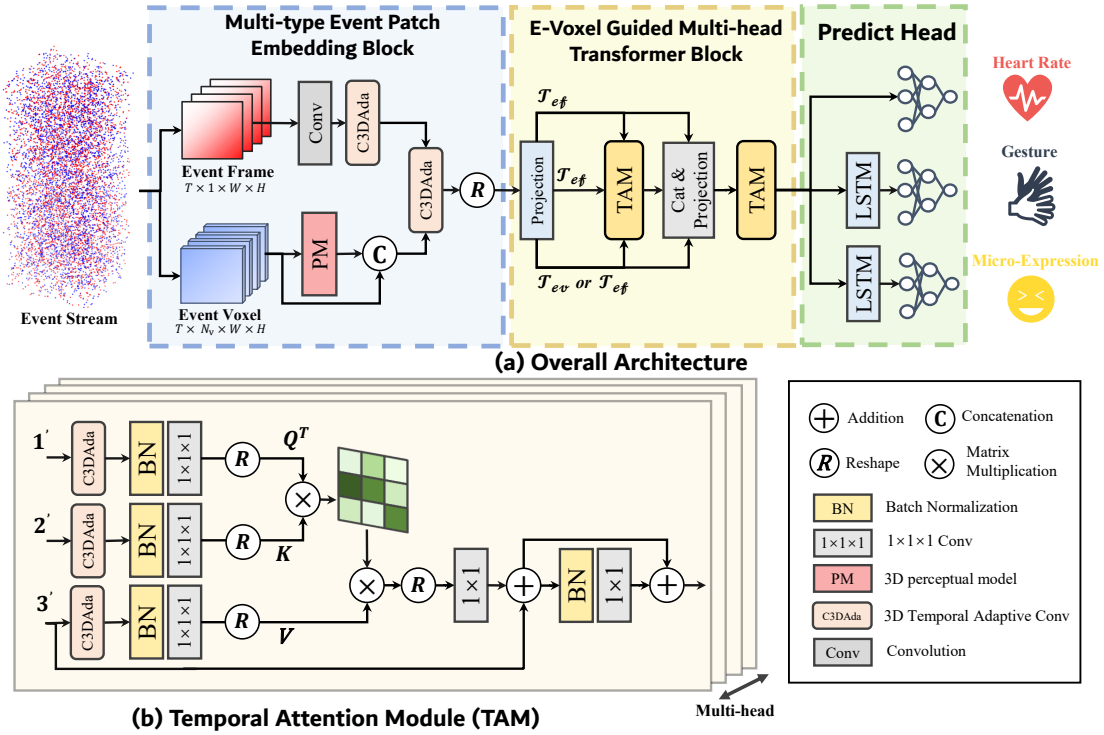

Figure 2: Pipeline of our proposed Spatio-temporal Event Transformer network.

characteristics of event cameras enable the precise capture of brief yet complex micro-behaviors. On the other hand, the asynchronous feature of event data collection offers a degree of privacy protection, reducing personal information (e.g. facial features, physiological conditions) leakage during behavioral measurements.

3. Methods

This section presents a comprehensive overview of our **S**patio-**t**emporal **E**vent **T**ransformer (STET) network, which is depicted in Fig.2.

Our model processes raw event streams as input. Specifically, the i-th event, $\mathbf{e}_i = (\mathbf{x}_i, t_i, p_i)$ is triggered at pixel coordinate $\mathbf{x}_i$ and time t_i whenever the log-scale brightness change exceeds a threshold $c > 0$,

$$\log(I(t, \mathbf{x})) - \log(I(t + \tau, \mathbf{x})) = p \cdot c, \tag{1}$$

where $I(t, \mathbf{x})$ and $I(t + \tau, \mathbf{x})$ are the brightness intensity of position $\mathbf{x}$ at time t and $t + \tau$. The polarity $p \in \{+1, -1\}$ indicates the direction of brightness change.

First, a Multi-type Event Patch Embedding (MTE) module is developed to extract event features as tokens, utilizing both global-local and spatio-temporal event information. Furthermore, an E-Voxel Guided Multi-head Transformer (EVGMH) module is employed to enhance event features and facilitate spatio-temporal feature representations.Finally, the feature representations are fed into task-specific prediction heads for different tasks. Further details will be provided in the subsequent sections.

3.1. Multi-type Event Patch Embedding Block

Multi-type Event Patch Embedding Block (MTE) can be divided into two parts: spatial and temporal (frame and temporal-voxel). First, the long event stream is partitioned into short sequences $\mathcal{S} = \{\mathbf{S}_1, \mathbf{S}_2, ..., \mathbf{S}_k\}$. The event points in each short sequence are projected onto the pixel coordinate system according to their coordinates $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)$ and accumulated sequentially, resulting in a sequence of k event frames $\mathcal{F} = \{\mathbf{F}_1, \mathbf{F}_2, ..., \mathbf{F}_k\}$. Next, the feature dimensions are expanded through a coarsely-to-finely downsample 3D temporal adaptive convolution layer [52], achieving patch embedding to obtain the token $\mathcal{T}_{ef}$. Additionally, to preserve the temporal information in the event frames, the event points in the $\mathcal{S}$ short sequences are projected onto the voxel coordinate system based on $\mathbf{v} = (t, \mathbf{x}_1, \mathbf{x}_2)$, producing an event voxel sequence. However, event voxels are extremely sparse and not suitable for direct representation. Inspired by CEUTrack [53] and typical 3D perceptual models [54, 55, 56], we perform a sparse representation of event voxels. The event voxels are converted into interval voxels $\mathbf{V}_e = (\mathbf{V}_x, \mathbf{V}_y, \mathbf{V}_z, \mathbf{V}_f)$. Here, $\mathbf{V}_x$ and $\mathbf{V}_y$ represent the coordinates $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)$, $\mathbf{V}_z$ represents the time dimension t, and $\mathbf{V}_f$ is the feature representation of the event voxel $\mathbf{V}_e$. $\mathbf{V}_f$ has a dimension of 19, consisting of three coordinate values and voxel features of length 1×16. We then sort the voxel grids by event density and select the top 1024 grids to sparsify the input dimensions, obtaining sparse event voxel $\mathcal{V}$. Moreover, we use the 3D temporal adaptive convolution layer for its patch embedding and get the token $\mathcal{T}_{ev}$.

$$\mathcal{T}_{ef}, \mathcal{T}_{ev} = MLP(MTE(\mathcal{S})), \tag{2}$$

where $\mathcal{T}_{ef}, \mathcal{T}_{ev} \in \mathbb{R}^{B \times T \times D}$, here B, T and D denote batch size, short sequences numbers, and token dimension, respectively. Finally, fully connected projection layers are applied to $\mathcal{T}_{ef}$ and $\mathcal{T}_{ev}$, before inputting them into the transformer architecture.

3.2. E-Voxel Guided Multi-head Transformer Block

Inspired by Physformer [9], we propose the Temporal Attention Module (TAM). The E-Voxel Guided Multi-head Transformer Block (EVGMH) consisting of multiple TAMs that compute the attention between the frame token $\mathcal{T}_{ef}$ and the voxel token $\mathcal{T}_{ev}$ as well as self-attention of the frame token $\mathcal{T}_{ef}$ itself, as shown in Fig.2, Temporal Attention Module.

Based on the self-attention mechanism [57, 58], we utilize 3D temporal adaptive convolution (C3DAda) layer instead of point linear projection for query (Q) and keyword (K) projections, which captures fine-grained local temporal features. For the value (V) projection, we utilize linear projection.

$$\begin{cases} Q, K, V = \text{C3DAda}(\mathcal{T}_{ef}), \text{C3DAda}(\mathcal{T}_{ev}), W_v^T \mathcal{T}_{ef}, TAM \textit{ in the first layer}, \\ Q, K, V = \text{C3DAda}(\mathcal{T}_{ef}), \text{C3DAda}(\mathcal{T}_{ef}), W_v^T \mathcal{T}_{ef}, TAM \textit{ in other layers}, \end{cases}$$

Afterwards, the flattened sequences of Q, K, and V, denoted as $Q, K, V \in \mathbb{R}^{B \times T \times D}$, are split into h heads (with $D_h = D/h$ for each head). For the i-th head ($i \leq h$), the self-attention score (S) can

be calculated using the following formulas:

$$Att_i, Score_i = \text{Softmax}(\frac{Q_i K_i^T}{\sqrt{D_h}}), \text{Softmax}(V_i \cdot Att_i). \tag{3}$$

The output of E-Voxel Guided Multi-head Transformer Block (EVGMH) is obtained by concatenating Att_i from all heads and then applying a linear (full connected layer) projection $\in \mathbb{R}^{D \times D}$. The first layer and following layers of EVGMH operation can be represented as:

$$\mathcal{T}_{EVGMH} = \text{MLP}(\text{Concate}(\text{Att}_1; \text{Att}_2; ...; \text{Att}_h)). \tag{4}$$

As depicted in Fig.2, a residual connection and layer normalization are applied after the attention operation. Subsequently, the result is fed into the feed-forward network, which comprises two convolution layers with batch normalization and nonlinear activation to refine local inconsistencies and reduce noise in the features.

3.3. Prediction Head

Based on different prediction tasks, we designed simple prediction heads. We firstly downsample the output token to obtain a sequence of feature representations. For the rPPG task, a fully connected layer outputs PPG signal results as a time series. For micro-expression or micro-gesture recognition, the feature sequence passes through an LSTM block, followed by a fully connected layer projecting to the classification categories.

3.4. Loss Function

For rPPG prediction, we use PhysFormer's label distribution method [59] and frame HR estimation as a multi-label classification task [60] with $\mathscr{L}_{\text{LD}}$. We also examine the correlation between the predicted rPPG signal and ground truth with $\mathscr{L}_{\text{time}}$. For micro-expression and micro-gesture classification, we use the cross-entropy loss function as $\mathscr{L}_{CE}$. More details for loss function are provided in the Appendix. The final loss function is defined as:

$$\mathscr{L}_{rPPG} = \alpha \cdot \mathscr{L}_{\text{time}} + \beta \cdot \mathscr{L}_{\text{LD}}, \mathscr{L}_{ME} = \mathscr{L}_{CE}, \mathscr{L}_{MG} = \mathscr{L}_{CE}. \tag{5}$$

where the α and β are hyperparameters that can be adjusted to balance $\mathscr{L}_{\text{LD}}$ and $\mathscr{L}_{\text{time}}$.

4. Experiments

4.1. Experimental Settings

Datasets. SMIC: SMIC [61] is made up of 164 samples. Lacking apex frame and action unit labels, the samples span 16 participants of 3 ethnicities. The recordings are taken with a resolution of 640 × 480 at 100 fps. **UBFC**: UBFC [62] is a dataset comprised of RGB video recordings (30fps). Participants are seated and lit with ambient light. The dataset includes ground truth obtained through contact-sensor-based PPG measurements. **iMiGUE**: iMiGUE [63] collected 359 videos of post-match press conferences from Grand Slam tournaments via

online video-sharing platforms. The videos have a resolution of 1280×720 and a frame rate of 25 fps. A total of 18,499 MG samples are labeled and assigned to 32 categories, along with one non-MG category, resulting in an average of approximately 51 MG samples per video.

Our E-Human Behaviour dataset is from simulation. First of all, the dataset is produced from the public rPPG datasets UBFC, micro-expression datasets SMIC, and a part of micro-gesture understanding datasets iMiGUE (because of space constraints), in which we input the reshaped gray (from RGB) channel images into the event simulator DVS-voltmeter [64]. In total, our proposed E-Human Behaviour dataset consists of 584 face region (164 from SMIC and 420 from UBFC) and 3500 upper body region (from iMiGUE) event streams with a resolution of 128×128 and the corresponding PPG signals as well as micro-expression and micro-gesture categories.

Implementation details. We implement the network using PyTorch [65] and the ADAM optimizer [66] with an initial learning rate of 4e-4 and a weight decay of 5e-5. The model is trained for 200 epochs on a NVIDIA TITAN Xp 12G. For $\mathscr{L}_{rPPG}$, like Physformer, a fixed α is set to 0.1, while the parameter β for frequency loss grew exponentially. For EVGMH, we used the settings block number $N = 6$, multi-head $h = 4$, $D = 96$, $D'_{FFN} = 512$. The targeted token patch size Ts × Hs × Ws is set to 1 × 4 × 4. During the training phase, we randomly select event streams from the dataset corresponding to a constant duration and input these streams into the network via the MTE module with a temporal size of 160, corresponding to the subsequent segment inputs.

4.2. Comparison with the State-of-the-arts

For the three tasks, we perform 9-fold cross-validation on the UBFC, SMIC, and iMiGUE datasets. We use the experimental metrics from other methods and present the test results for the UBFC, SMIC, and iMiGUE datasets in Table 1, Table 2, and Table 3, respectively. Although our event-based approach does not show a distinct advantage in the evaluated metrics, it demonstrates the feasibility of our method.

Notably, the training set accuracy reaches nearly 100% during training on the SMIC datasets, but the model performs poorly on the test set, indicating overfitting due to the small dataset size. This issue is less significant for the larger iMiGUE dataset, where there is no significant gap between training and test set performance.

The event data in this experiment is simulated from the original image data. The quality of this simulation depends on the original data, and the information in the simulated events cannot exceed that of the original images. Additionally, our simulation produces a one-channel event stream, resized to 128 × 128 for network bandwidth considerations. Despite these constraints, the event-based method has proven effective in the three human behavior analysis tasks, demonstrating its potential in this field.

4.3. Ablation Study

In this section, we provide the results of ablation studies for micro-gesture on the 9-fold cross-validation of the iMiGUE dataset.

Analysis of network hyperparameters. We perform ablation experiments on the network hyperparameters, specifically investigating the number of feature dimensions in the TAM and

Table 1
Quantitative analyses are performed on the UBFC dataset using various methods. The best results are **bolded**, while the second-best results are underlined. The "Modality/Channel" column specifies the modality and the count of channels used. The "Train sets" column details the dataset and the count of sequences used.

Methods		Modality/Channel	Train sets	MAE	RMSE
UBFC	EfficientPhy [5]	RGB/3	[67] BP4D$^+$ 1400	9.21	17.11
	DeepPhys [21]	RGB/3	BP4D$^+$ 1400	3.36	12.86
	MTTS-CAN [68]	RGB/3	BP4D$^+$ 1400	12.78	22.43
	BigSmall [69]	RGB/3	BP4D$^+$ 1400	**1.03**	**2.55**
	STET(ours)	Event/1	UBFC 36	10.31	22.74

Table 2
Quantitative analyses are conducted on the SMIC dataset using various methods. AMAN [70] and TSCNN [71] use pre-trained models and fine-tune on their mixed dataset.

Methods		Modality/Channel	Train sets	UF1(%)	UAR(%)
SMIC	AMAN [70]	RGB/3	pre-train + fine-tune	77.00	79.87
	TSCNN [71]	RGB/3	pre-train + fine-tune	72.36	72.74
	μ-BERT [6]	RGB/3	[72]CASME3 4599	**85.50**	**83.84**
	STET(ours)	Event/1	SMIC 144	60.33	59.34

Table 3
Quantitative analyses are conducted on the iMiGUE dataset using various methods.

Methods		Modality/Channel	Train sets	Top-1(%)	Top-5(%)
iMiGUE	TSN [39]	RGB/3	iMiGUE 245	51.54	85.42
	TSM [38]	RGB/3	iMiGUE 245	61.10	91.24
	EHCT [8]	Skeleton/1	iMiGUE 245	63.02	**91.36**
	JSSE [7]	Skeleton/1	iMiGUE 245	**64.12**	91.10
	STET(ours)	Event/1	iMiGUE 63	56.75	86.75

the number of EVGMHs. As presented in Table.4, the evaluation metrics exhibit improvement with an increase in feature dimensions and a higher number of EVGMHs. This enhancement can be attributed to the larger network model's superior capability to capture spatio-temporal features within the event stream, resulting in more precise predictions.

Table 4
Quantitative analyses of ablation experiments, "OOM" means CUDA out of memory

dim+block	24+3	24+6	48+3	48+6	48+12	96+6	96+12(ours)	144+6	144+12
Top-1(%)	39.25	36.00	40.50	36.50	51.25	46.50	**56.75**	50.00	OOM
Top-5(%)	76.50	72.75	71.25	65.75	81.25	73.75	**86.75**	81.75	OOM

5. Conclusion

This paper introduces STET, a novel event-based multi-task architecture for rPPG measurement, micro-expression recognition, and micro-gesture understanding. Our approach combines events compressed into frames and transformed into voxel representations, fully leveraging the potential of events to capture the spatio-temporal features of human behavior. The results demonstrate that our method effectively models spatially and temporally diverse event signals. While the current STET does not surpass state-of-the-art methods, event-based approaches offer promising new solutions for the field, indicating that further exploration into event-based analysis of human behavior is warranted.

References

[1] P. V. K. Borges, N. Conci, A. Cavallaro, Video-based human behavior understanding: A survey, IEEE transactions on circuits and systems for video technology 23 (2013) 1993–2008.

[2] D. C. Mack, J. T. Patrie, P. M. Suratt, R. A. Felder, M. Alwan, Development and preliminary validation of heart rate and breathing rate detection using a passive, ballistocardiography-based sleep monitoring system, IEEE Transactions on information technology in biomedicine 13 (2008) 111–120.

[3] Y. Li, J. Wei, Y. Liu, J. Kauttonen, G. Zhao, Deep learning for micro-expression recognition: A survey, IEEE Transactions on Affective Computing 13 (2022) 2028–2046.

[4] J. J. Ojeda-Castelo, M. d. L. M. Capobianco-Uriarte, J. A. Piedra-Fernandez, R. Ayala, A survey on intelligent gesture recognition techniques, IEEE Access 10 (2022) 87135–87156.

[5] X. Liu, B. Hill, Z. Jiang, S. Patel, D. McDuff, Efficientphys: Enabling simple, fast and accurate camera-based cardiac measurement, in: Proceedings of the IEEE/CVF winter conference on applications of computer vision, 2023, pp. 5008–5017.

[6] X.-B. Nguyen, C. N. Duong, X. Li, S. Gauch, H.-S. Seo, K. Luu, Micron-bert: Bert-based facial micro-expression recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 1482–1492.

[7] K. Li, D. Guo, G. Chen, X. Peng, M. Wang, Joint skeletal and semantic embedding loss for micro-gesture classification, arXiv preprint arXiv:2307.10624 (2023).

[8] H. Huang, X. Guo, W. Peng, Z. Xia, Micro-gesture classification based on ensemble hypergraph-convolution transformer, Micro-gesture Analysis for Hidden Emotion Understanding 2023 (2023) 9.

[9] Z. Yu, Y. Shen, J. Shi, H. Zhao, P. H. Torr, G. Zhao, Physformer: Facial video-based physiological measurement with temporal difference transformer, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 4186–4196.

[10] G. Gallego, T. Delbrück, G. Orchard, C. Bartolozzi, B. Taba, A. Censi, S. Leutenegger, A. J. Davison, J. Conradt, K. Daniilidis, et al., Event-based vision: A survey, IEEE transactions on pattern analysis and machine intelligence 44 (2020) 154–180.

[11] C. Takano, Y. Ohta, Heart rate measurement based on a time-lapse image, Medical engineering & physics 29 (2007) 853–857.

[12] M.-Z. Poh, D. J. McDuff, R. W. Picard, Advancements in noncontact, multiparameter phys-

iological measurements using a webcam, IEEE Transactions on Biomedical Engineering 58 (2011) 7–11.
[13] Ming-Zher, Poh, Daniel, J., McDuff, Rosalind, W., Picard, Non-contact, automated cardiac pulse measurements using video imaging and blind source separation, Optics Express 18 (2010).
[14] S. Tulyakov, X. Alameda-Pineda, E. Ricci, L. Yin, J. F. Cohn, N. Sebe, Self-adaptive matrix completion for heart rate estimation from face videos under realistic conditions, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2396–2404.
[15] W. Wang, S. Stuijk, G. De Haan, A novel algorithm for remote photoplethysmography: Spatial subspace rotation, IEEE Transactions on Biomedical Engineering (2015) 1–1.
[16] H. Lu, H. Han, Nas-hr: Neural architecture search for heart rate estimation from face videos, Virtual Reality & Intelligent Hardware 3 (2021) 33–42.
[17] H. Lu, H. Han, S. K. Zhou, Dual-gan: Joint bvp and noise modeling for remote physiological measurement, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 12404–12413.
[18] X. Niu, H. Han, S. Shan, X. Chen, Synrhythm: Learning a deep heart rate estimator from general to specific, in: 2018 24th international conference on pattern recognition (ICPR), IEEE, 2018, pp. 3580–3585.
[19] G.-S. Hsu, A. Ambikapathi, M.-S. Chen, Deep learning with time-frequency representation for pulse estimation from facial videos, in: 2017 IEEE International Joint Conference on Biometrics (IJCB), 2017, pp. 383–389.
[20] R. Špetlík, V. Franc, J. Matas, Visual heart rate estimation with convolutional neural network, in: Proceedings of the british machine vision conference, Newcastle, UK, 2018, pp. 3–6.
[21] W. Chen, D. McDuff, Deepphys: Video-based physiological measurement using convolutional attention networks, in: Proceedings of the european conference on computer vision (ECCV), 2018, pp. 349–365.
[22] X. Liu, B. L. Hill, Z. Jiang, S. Patel, D. McDuff, Efficientphys: Enabling simple, fast and accurate camera-based vitals measurement, arXiv preprint arXiv:2110.04447 (2021).
[23] E. M. Nowara, D. McDuff, A. Veeraraghavan, The benefit of distraction: Denoising camera-based physiological measurements using inverse attention, in: Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 4955–4964.
[24] Z. Yu, X. Li, X. Niu, J. Shi, G. Zhao, Autohr: A strong end-to-end baseline for remote heart rate measurement with neural searching, IEEE Signal Processing Letters 27 (2020) 1245–1249.
[25] Z. Yu, X. Li, G. Zhao, Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks, arXiv preprint arXiv:1905.02419 (2019).
[26] Y.-J. Liu, J.-K. Zhang, W.-J. Yan, S.-J. Wang, G. Zhao, X. Fu, A main directional mean optical flow feature for spontaneous micro-expression recognition, IEEE Transactions on Affective Computing 7 (2015) 299–310.
[27] F. Xu, J. Zhang, J. Z. Wang, Microexpression identification and categorization using a facial dynamics map, IEEE Transactions on Affective Computing 8 (2017) 254–267.
[28] Y.-J. Liu, B.-J. Li, Y.-K. Lai, Sparse mdmo: Learning a discriminative feature for micro-

expression recognition, IEEE Transactions on Affective computing 12 (2018) 254–261.
[29] S.-T. Liong, J. See, K. Wong, R. C.-W. Phan, Less is more: Micro-expression recognition from video using apex frame, Signal Processing: Image Communication 62 (2018) 82–92.
[30] D. H. Kim, W. J. Baddar, Y. M. Ro, Micro-expression recognition with expression-state constrained spatio-temporal feature representations, in: Proceedings of the 24th ACM international conference on Multimedia, 2016, pp. 382–386.
[31] J. Li, Y. Wang, J. See, W. Liu, Micro-expression recognition based on 3d flow convolutional neural network, Pattern Analysis and Applications 22 (2019) 1331–1339.
[32] H.-Q. Khor, J. See, R. C. W. Phan, W. Lin, Enriched long-term recurrent convolutional network for facial micro-expression recognition, in: 2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018), IEEE, 2018, pp. 667–674.
[33] B. Yang, J. Cheng, Y. Yang, B. Zhang, J. Li, Merta: micro-expression recognition with ternary attentions, Multimedia Tools and Applications 80 (2021) 1–16.
[34] K. Carlstead, J. L. Brown, W. Strawn, Behavioral and physiological correlates of stress in laboratory cats, Applied Animal Behaviour Science 38 (1993) 143–158.
[35] C. Mohiyeddini, S. Bauer, S. Semple, Displacement behaviour is associated with reduced stress levels among men but not women, PloS one 8 (2013) e56355.
[36] C. F. Sharpley, A. Sagris, When does counsellor forward lean influence client-perceived rapport?, British Journal of Guidance & Counselling 23 (1995) 387–394.
[37] H. G. Wallbott, Bodily expression of emotion, European journal of social psychology 28 (1998) 879–896.
[38] J. Lin, C. Gan, S. Han, Tsm: Temporal shift module for efficient video understanding, in: Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 7083–7093.
[39] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, L. Van Gool, Temporal segment networks for action recognition in videos, IEEE transactions on pattern analysis and machine intelligence 41 (2018) 2740–2755.
[40] M. Xu, M. Gao, Y.-T. Chen, L. S. Davis, D. J. Crandall, Temporal recurrent networks for online action detection, in: Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 5532–5541.
[41] J. Carreira, A. Zisserman, Quo vadis, action recognition? a new model and the kinetics dataset, in: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6299–6308.
[42] K. Hara, H. Kataoka, Y. Satoh, Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2018, pp. 6546–6555.
[43] C. Feichtenhofer, H. Fan, J. Malik, K. He, Slowfast networks for video recognition, in: Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 6202–6211.
[44] A. Shah, H. Chen, G. Zhao, Representation learning for topology-adaptive micro-gesture recognition and analysis, in: IJCAI-MIGA Workshop & Challenge on Micro-gesture Analysis for Hidden Emotion Understanding (MiGA) July 21, 2023 Macao, China, Redaktion Sun SITE, 2023.
[45] P. Lichtsteiner, C. Posch, T. Delbruck, A 128× 128 120 db 15 μ s latency asynchronous

temporal contrast vision sensor, IEEE journal of solid-state circuits 43 (2008) 566–576.
[46] X. Zhang, W. Liao, L. Yu, W. Yang, G.-S. Xia, Event-based synthetic aperture imaging with a hybrid network, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 14235–14244.
[47] W. Liao, X. Zhang, L. Yu, S. Lin, W. Yang, N. Qiao, Synthetic aperture imaging with events and frames, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 17735–17744.
[48] C. Zhang, X. Zhang, M. Lin, C. Li, C. He, W. Yang, G.-S. Xia, L. Yu, Crosszoom: Simultaneous motion deblurring and event super-resolving, IEEE Transactions on Pattern Analysis and Machine Intelligence (2024).
[49] L. Yu, B. Wang, X. Zhang, H. Zhang, W. Yang, J. Liu, G.-S. Xia, Learning to super-resolve blurry images with events, IEEE Transactions on Pattern Analysis and Machine Intelligence (2023).
[50] X. Zhang, L. Yu, W. Yang, J. Liu, G.-S. Xia, Generalizing event-based motion deblurring in real-world scenarios, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 10734–10744.
[51] K. Chen, L. Yu, Motion deblur by learning residual from events, IEEE Transactions on Multimedia (2024).
[52] Z. Huang, S. Zhang, L. Pan, Z. Qing, M. Tang, Z. Liu, M. H. Ang Jr, Tada! temporally-adaptive convolutions for video understanding, arXiv preprint arXiv:2110.06178 (2021).
[53] C. Tang, X. Wang, J. Huang, B. Jiang, L. Zhu, J. Zhang, Y. Wang, Y. Tian, Revisiting color-event based tracking: A unified network, dataset, and metric, arXiv preprint arXiv:2211.11010 (2022).
[54] Y. Yan, Y. Mao, B. Li, Second: Sparsely embedded convolutional detection, Sensors 18 (2018) 3337.
[55] S. Shi, C. Guo, L. Jiang, Z. Wang, J. Shi, X. Wang, H. Li, Pv-rcnn: Point-voxel feature set abstraction for 3d object detection, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 10529–10538.
[56] T. Yin, X. Zhou, P. Krahenbuhl, Center-based 3d object detection and tracking, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 11784–11793.
[57] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).
[58] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929 (2020).
[59] X. Niu, H. Han, S. Shan, X. Chen, Continuous heart rate measurement from face: A robust rppg approach with distribution learning, in: 2017 IEEE international joint conference on biometrics (IJCB), IEEE, 2017, pp. 642–650.
[60] B.-B. Gao, C. Xing, C.-W. Xie, J. Wu, X. Geng, Deep label distribution learning with label ambiguity, IEEE Transactions on Image Processing 26 (2017) 2825–2838.
[61] X. Li, T. Pfister, X. Huang, G. Zhao, M. Pietikäinen, A spontaneous micro-expression database: Inducement, collection and baseline, in: 2013 10th IEEE International Conference

and Workshops on Automatic face and gesture recognition (fg), IEEE, 2013, pp. 1–6.
[62] S. Bobbia, R. Macwan, Y. Benezeth, A. Mansouri, J. Dubois, Unsupervised skin tissue segmentation for remote photoplethysmography, Pattern Recognition Letters 124 (2019) 82–90.
[63] X. Liu, H. Shi, H. Chen, Z. Yu, X. Li, G. Zhao, imigue: An identity-free video dataset for micro-gesture understanding and emotion analysis, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 10631–10642.
[64] S. Lin, Y. Ma, Z. Guo, B. Wen, Dvs-voltmeter: Stochastic process-based event simulator for dynamic vision sensors, in: European Conference on Computer Vision, Springer, 2022, pp. 578–593.
[65] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al., Pytorch: An imperative style, high-performance deep learning library, Adv. Neural Inform. Process. Syst. 32 (2019).
[66] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).
[67] Z. Zhang, J. M. Girard, Y. Wu, X. Zhang, P. Liu, U. Ciftci, S. Canavan, M. Reale, A. Horowitz, H. Yang, et al., Multimodal spontaneous emotion corpus for human behavior analysis, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 3438–3446.
[68] X. Liu, J. Fromm, S. Patel, D. McDuff, Multi-task temporal shift attention networks for on-device contactless vitals measurement, Advances in Neural Information Processing Systems 33 (2020) 19400–19411.
[69] G. Narayanswamy, Y. Liu, Y. Yang, C. Ma, X. Liu, D. McDuff, S. Patel, Bigsmall: Efficient multi-task learning for disparate spatial and temporal physiological measurements, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024, pp. 7914–7924.
[70] M. Wei, W. Zheng, Y. Zong, X. Jiang, C. Lu, J. Liu, A novel micro-expression recognition approach using attention-based magnification-adaptive networks, in: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2022, pp. 2420–2424.
[71] B. Song, K. Li, Y. Zong, J. Zhu, W. Zheng, J. Shi, L. Zhao, Recognizing spontaneous micro-expression using a three-stream convolutional neural network, Ieee Access 7 (2019) 184537–184551.
[72] J. Li, Z. Dong, S. Lu, S.-J. Wang, W.-J. Yan, Y. Ma, Y. Liu, C. Huang, X. Fu, Cas (me) 3: A third generation facial spontaneous micro-expression database with depth information and high ecological validity, IEEE Transactions on Pattern Analysis and Machine Intelligence 45 (2022) 2782–2800.

6. Appendix

6.1. Loss Function

As mentioned in [9], the facial rPPG signals with similar HR values exhibit similar periodicity. Motivated by this observation, we consider each event stream as an instance associated with a label distribution. This label distribution covers a range of class labels, indicating the extent to which each label describes the instance. By doing so, an event stream contributes not only to its target HR value but also to its neighboring HR values. To incorporate similarity information among HR classes during training, the rPPG-based HR estimation problem is formulated as a specific L-class multi-label classification task, where L is 139 in our case (each integer HR value within the range of 42 to 180 bpm is considered a class). Each event stream is assigned a label distribution $\mathrm{p} = \{p_1, p_2, ..., p_L\} \in \mathbb{R}^L$. We assume that each entry of p is a real value between 0 and 1, satisfying $\sum_{k=1}^{L} p_k = 1$. To construct the corresponding label distribution p, a Gaussian distribution function centered at the ground truth HR label Y_{HR}, with a standard deviation of σ as follows was employed,

$$p_k = \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{(k-(Y_{HR}-41))^2}{2\sigma^2}). \tag{6}$$

The label distribution loss, denoted as $\mathscr{L}_{\mathrm{LD}}$, is formulated as the Kullback-Leibler (KL) [60] divergence between the label distribution $\mathbf{p}$ and the softmax of the predicted rPPG signal's power spectral density (PSD), represented as $\hat{\mathbf{p}}$:

$$\mathscr{L}_{\mathrm{LD}} = \mathrm{KL}(\mathbf{p}, \mathrm{Softmax}(\hat{\mathbf{p}})). \tag{7}$$

Moreover, in the time domain, we analyze the accuracy of the response prediction by realizing the correlation between the predicted rPPG signal and ground truth. We define the loss function $\mathscr{L}_{\mathrm{time}}$ in the time domain based on the negative Pearson loss [25]. To adapt the Pearson correlation for use as a loss function, the loss function $\mathscr{L}_{\mathrm{time}}$ can be defined as:

$$\mathscr{L}_{\mathrm{time}} = \begin{cases} |\rho|, & \text{if } \rho < 0, \\ 1-\rho, & \text{if } \rho \geq 0, \end{cases} \tag{8}$$

where ρ is the Pearson correlation coefficient between the predicted rPPG signal and ground truth. Finally, the dynamic loss $\mathscr{L}_{rPPG}$ is formulated as follows:

$$\mathscr{L}_{rPPG} = \alpha \cdot \mathscr{L}_{\mathrm{time}} + \beta \cdot \mathscr{L}_{\mathrm{LD}}, \tag{9}$$

where the value of β is dynamically adjusted based on the current epoch number n_i and the total number of epochs n:

$$\beta = \beta_0 \cdot \eta^{\frac{n_i-1}{n}}. \tag{10}$$

Then, in the micro-expression and micro-gesture classification task, the cross-entropy loss function is used to measure the difference between the predicted probability distribution and the true distribution. Given an input sample x with its true label y, and the model's predicted probability distribution $\hat{prob}$, the cross-entropy loss function $\mathscr{L}_{CE}$ is defined as follows:

$$\mathscr{L}_{CE} = -\sum_{i} prob(y_i) \log(\hat{prob}(y_i)), \tag{11}$$

where $prob(y_i)$ is the true label distribution for class i (one-hot encoded for classification tasks), and $\hat{prob}(y_i)$ is the predicted probability for class i. Eventually, we adopt cross-entropy loss as $\mathscr{L}_{ME}$ (micro-expression) and $\mathscr{L}_{MG}$ (micro-gesture) as follows:

$$\mathscr{L}_{ME} = \mathscr{L}_{CE}, \mathscr{L}_{MG} = \mathscr{L}_{CE}. \tag{12}$$