In the Blink of an Eye: Event-based Emotion Recognition

Haiwei Zhang*
Dalian University of Technology
Dalian, China
haiweizhang32009182@mail.dlut.edu.cn

Jiqing Zhang*
Dalian University of Technology
Dalian, China
jqz@mail.dlut.edu.cn

Bo Dong†
Princeton University
Princeton, USA
bo.dong@princeton.edu

Pieter Peers
College of William & Mary
Williamsburg, USA
ppeers@siggraph.org

Wenwei Wu
Dalian University of Technology
Dalian, China
wuwenwei0206@mail.dlut.edu.cn

Xiaopeng Wei†
Dalian University of Technology
Dalian, China
xpwei@dlut.edu.cn

Felix Heide
Princeton University
Princeton, USA
fheide@cs.princeton.edu

Xin Yang†
Key Laboratory of Social Computing and Cognitive Intelligence of Ministry of Education, Dalian University of Technology
Dalian, China
xinyang@dlut.edu.cn

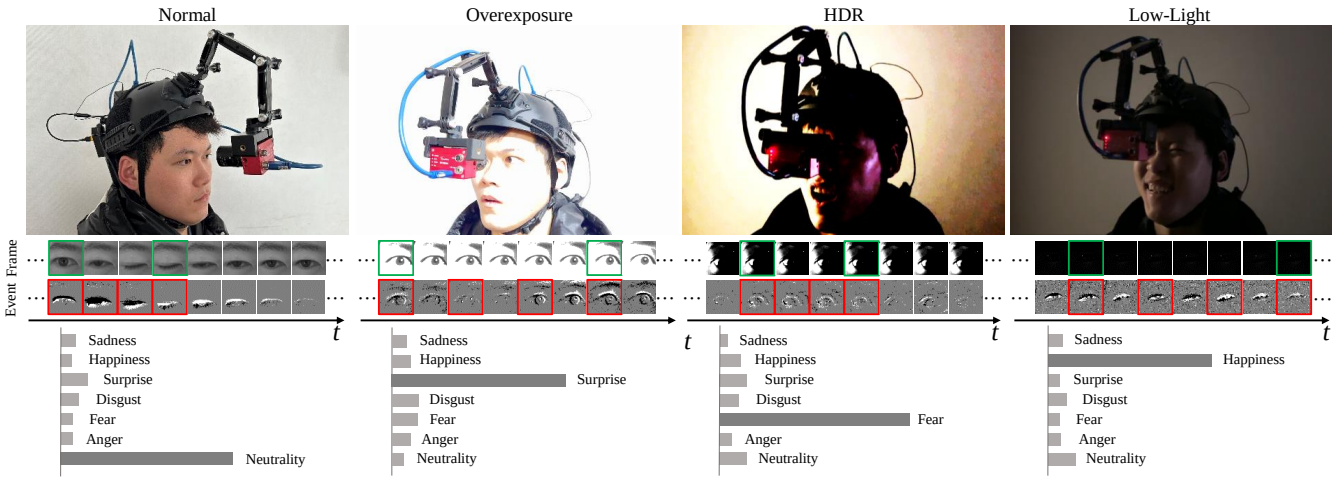


Figure 1: Demonstration of a wearable single-eye emotion recognition prototype system consisting with a bio-inspired event-based camera (DAVIS346) and a low-power NVIDIA Jetson TX2 computing device. Event-based cameras simultaneously provide intensity and corresponding events, which we input to a newly designed lightweight Spiking Eye Emotion Network (SEEN) to effectively extract and combine spatial and temporal cues for emotion recognition. Given a sequence, SEEN takes the start and end intensity frames (green boxes) along with n intermediate event frames (red boxes) as input. Our prototype system consistently recognizes emotions based on single-eye areas under different lighting conditions at 30 FPS.

ABSTRACT

We introduce a wearable single-eye emotion recognition device and a real-time approach to recognizing emotions from partial observations of an emotion that is robust to changes in lighting conditions. At the heart of our method is a bio-inspired event-based

*Equal contribution.
†Corresponding authors.

SIGGRAPH '23 Conference Proceedings, August 6–10, 2023, Los Angeles, CA, USA

ACM ISBN 979-8-4007-0159-7/23/08…$15.00
https://doi.org/10.1145/3588432.3591511

camera setup and a newly designed lightweight Spiking Eye Emotion Network (SEEN). Compared to conventional cameras, event-based cameras offer a higher dynamic range (up to 140 dB vs. 80 dB) and a higher temporal resolution (in the order of μs vs. 10s of *ms*). Thus, the captured events can encode rich temporal cues under challenging lighting conditions. However, these events lack texture information, posing problems in decoding temporal information effectively. SEEN tackles this issue from two different perspectives. First, we adopt convolutional spiking layers to take advantage of the spiking neural network's ability to decode pertinent temporal information. Second, SEEN learns to extract essential spatial cues from corresponding intensity frames and leverages a novel weight-copy scheme to convey spatial attention to the convolutional spiking layers during training and inference. We extensively validate and demonstrate the effectiveness of our approach on a specially collected Single-eye Event-based Emotion (SEE) dataset. To the best of our knowledge, our method is the first eye-based emotion recognition method that leverages event-based cameras and spiking neural networks.

CCS CONCEPTS

• **Computing methodologies → Computer vision; Supervised learning by classification; Spiking neural networks**.

KEYWORDS

Event-based cameras, eye-based emotion recognition

ACM Reference Format:
Haiwei Zhang, Jiqing Zhang, Bo Dong, Pieter Peers, Wenwei Wu, Xiaopeng Wei, Felix Heide, and Xin Yang. 2023. In the Blink of an Eye: Event-based Emotion Recognition. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Proceedings (SIGGRAPH '23 Conference Proceedings), August 6–10, 2023, Los Angeles, CA, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3588432.3591511

1 INTRODUCTION

Real-time emotion recognition in uncontrolled environments is a challenging problem that forms the cornerstone of many *in-the-wild* human-centered interactive computer graphics experiences such as interactive storytelling that adapts to the users emotions, and emotion-aware virtual avatars. Predicting emotions from regular RGB video streams is a challenging and ambiguous endeavor; informative spatial and temporal emotive cues can be adversely affected by head pose and partial occlusions. To help classify emotions in RGB video frames, existing facial emotion recognition models build on complex CNN-based models such as ResNet 50 [Deng et al. 2020b], Transformer [Zhao and Liu 2021], and Inception-based methods [Hickson et al. 2019]. Robustly handling varying lighting conditions and rapid user movements further complicates emotion recognition, and existing methods rely on cumbersome large network enhancement modules [Zhao and Liu 2021] or impose active IR lighting [Wu et al. 2020]. Despite all these innovations, emotion recognition from RGB video streams remains difficult and fragile.

In this paper, we introduce a novel wearable emotion recognition prototype in which a bio-inspired event-based camera (DAVIS346) is affixed in front of a user's right eye. An event-based camera can provide more robust temporal cues for emotion recognition under adverse lighting conditions as it offers a higher dynamic range (up to 140 dB vs. 80 dB) and a higher temporal resolution (in the order of μs vs. 10s of *ms*) than a conventional camera. Even though this setup provides a stable fixed perspective of a right eye and it can robustly handle various lighting conditions, estimating emotion from a single eye still poses unique challenges.

A key issue is that event-based cameras do not capture texture information effectively (see Figure 1). These spatial features are not only essential for emotion recognition but also important for inferring more informative temporal features. For example, while pupil motion and blinking are dominant temporal cues, they are less informative for emotion classification. In contrast, the subtle movements related to the facial units, such as raising the outer brow and squinting, are stronger cues for eye-based emotion recognition.

To address these challenges, we devise a lightweight SEEN, which combines the best from both events and intensity frames to *guide* emotion recognition from asynchronous events with spatial texture cues from corresponding intensity frames. In particular, SEEN consists of a spatial feature extractor and a temporal feature extractor that partially share the same convolutional architecture. During training, the shared convolutional parts are only learned in the spatial feature extractor, and the updated weights are copied to the temporal feature extractor. Consequently, spatial attention can be effectively conveyed to the temporal decoding process. As such, the temporal feature extractor learns to associate spatial and temporal features, resulting in a consistent emotion classification.

To train our lightweight Spiking Eye Emotion Network (SEEN) and to stimulate research in event-based single-eye emotion recognition, we introduce a new Single-eye Event-based Emotion (SEE) dataset. We validate our approach on the SEE dataset and demonstrate state-of-the-art emotion recognition under different challenging lighting conditions, outperforming the runner-up method by a significant margin, 4.8% and 4.6% in WAR and UAR, respectively. The prototype system with an NVIDIA Jetson TX2 operates at 30 FPS in real-world testing scenarios.

Specifically, our work makes the following contributions:

- a novel real-time emotion recognition method based on event camera measurements and a spiking neural network suited for in-the-wild deployment;
- a weight-copy training scheme to enforce learned weights awareness of both spatial and temporal cues; and
- the first publicly available single-eye emotion dataset containing both intensity frames and corresponding raw events, captured under four different lighting conditions.

Limitations. SEEN partially relies on spatial features extracted from intensity frames, which can be adversely affected by extremely degraded lighting conditions, resulting in a significant performance drop. While our method robustly handles most lighting conditions effectively, as evidenced by our experimental results, further improving robustness by solely leveraging events forms an exciting avenue for future research in eye-based emotion recognition.

The code and dataset are available on github.

2 RELATED WORK

We focus our discussion on related work in emotion recognition on measuring emotions (wearable emotion sensing systems) and recognition (facial emotion recognition).

Wearable Emotion Sensing Systems. Emotions impact the human body in subtle ways. However, not all of these signals are equally robust indicators of emotional state, and not all are easily measured. Various bio-signals have been investigated for convenient measurement of indicators of emotional state. Long-term heart rate variability (HRV) has been shown to strongly correlate with emotional patterns [Appelhans and Luecken 2006; Costa et al. 2019]. Similarly, brain activity recorded by electroencephalogram (EEG) sensors also correlates to different emotions [Li et al. 2018; Liu et al. 2020]. Inspired by human perception of emotions, Electromyogram (EMG) measurements of facial muscle contractions [Lucero and Munhall 1999] map to emotions, making wearable emotional detection devices possible [Gruebler and Suzuki 2014]. A disadvantage of these methods is that they require the sensors to make *direct skin contact*, dramatically restricting freedom of activity. Furthermore, due to the displacement of sensors and muscular cross-talk during movement, the results can be of low reliability. An alternative to contact-based measurement is pupillometry, *i.e.,* the measurement of pupil size and reactivity, as a potential indicator of emotion [Mathôt 2018; Nie et al. 2020]. However, pupilometry requires expensive equipment, and the reliability is significantly impacted by ambient lighting [Couret et al. 2019]. Similar to pupilometry, our method also focuses on the eye as an indicator of emotional state. However, in contrast to prior work, we employ an event-based camera that does not require direct skin contact and which can operate in challenging lighting conditions.

Facial Emotion Recognition. Facial emotion recognition has received significant attention in computer graphics and computer vision, with applications ranging from driving facial expressions [Hickson et al. 2019; Ji et al. 2022] to facial reenactment for efficient social interactions [Burgos-Artizzu et al. 2015; Li et al. 2015]. A significant portion of prior work in facial emotion recognition requires observations of the entire face, and several methods have been introduced for effective facial feature learning [Ruan et al. 2021; Xue et al. 2021], dealing with uncertainties in facial expression data [Zhang et al. 2021a], handing partial occlusions [Georgescu and Ionescu 2019; Houshmand and Khan 2020], and exploiting temporal cues [Deng et al. 2020b; Sanchez et al. 2021]. Combinations with other modalities such as contextual information [Lee et al. 2019] and depth [Lee et al. 2020] have also been explored to further improve facial recognition accuracy.

However, observing the entire face is not feasible in many practical situations. Alternatively, several methods focus on the eye area only for emotion recognition. Hickson *et al.* [2019] infer emotional expressions based on images of both eyes captured with an infrared gaze-tracking camera inside a virtual reality headset. Wu *et al.* [2020] rely on infrared single-eye observations to reduce camera synchronization and data bandwidth issues when monitoring both eyes. Both systems require a personalized initialization procedure; Hickson *et al.* require a personalized neutral image, and Wu *et al.* require a reference feature vector of each emotion. The need for a personalized setup makes these systems intrusive and non-transparent to the user and could raise privacy concerns. Furthermore, neither system leverages temporal cues, which are essential for robust emotion recognition [Sanchez et al. 2021]. Our approach does not require personalization, and it leverages temporal and spatial cues to improve emotion recognition accuracy.

3 BACKGROUND

Before detailing our method, we first review work related to the two key components of our event-based emotion recognition method: event-based cameras and spiking neural networks.

Event-based Cameras. An event-based camera differs from a conventional camera in that it does not measure pixel intensities, but instead, an event-based camera records asynchronous (log-encoded) per-pixel brightness changes [Gallego et al. 2022; Gehrig et al. 2021]. Event-based cameras offer a significantly higher dynamic range (up to 140 dB) and a higher temporal resolution (in the order of μs) than conventional cameras. Each event e is characterized by three pieces of information: the pixel location, (x, y); the event triggering time, t; and a polarity, $p \in \{-1, 1\}$ which reflects the direction of the brightness change. Formally, a set of N events can be defined as:

$$\mathcal{E} = \{e_k\}_{k=1}^{N} = \{[x_k, y_k, t_k, p_k]\}_{k=1}^{N}. \tag{1}$$

Under static lighting, a stationary event-based camera only records scene motion, and events are typically triggered by moving edges (*e.g.*, object contours, and texture boundaries). Since the events predominately stem from the motion of edges, the measured events are inherently sparse and devoid of texture information. Furthermore, since the captured events are triggered asynchronously, events are incompatible with CNN-based architectures. Instead, events are aggregated into a frame or grid-based representation [Gehrig et al. 2019; Lagorce et al. 2017; Maqueda et al. 2018; Wang et al. 2022] before neural processing. In our implementation, we adopt the aggregation algorithm of Zhang *et al.* [2021b], which currently offers the highest performance for single object tracking under normal and degraded conditions. We refer to the Supplementary Material for additional details.

Spiking Neural Network (SNN). Spiking neural networks (SNNs) closely mimic biological information processes. An SNN incorporates the concept of time and only exchanges information (*i.e.,* spike) when a *membrane potential* exceeds some potential threshold. Mathematically an SNN neuron simulates the properties of a cell in a nervous system with varying degrees of detail, which models three states of a biological neuron: rest, depolarization, and hyperpolarization [Ding et al. 2022]. When a neuron is at rest, its membrane potential remains constant; typically set to 0. When not at rest, the change in the membrane potential can either decrease or increase. An increase in membrane potential is called depolarization. In contrast, hyperpolarization describes a reduction in membrane potential. When a membrane potential is higher than a potential threshold, an action potential, *i.e.,* spike, is triggered, which for an SNN is a binary value. We refer the interested reader to Ding *et al.* [2022] for an in-depth discussion of these concepts.

In this paper, we use the *leaky integrate-and-fire (LIF)* spiking neuron model [Gerstner and Kistler 2002], one of the most widely used spiking models. When a LIF neuron receives spikes from other

neurons, the spikes are scaled accordingly based on learned synaptic weights. Depolarization is achieved by summing over all the scaled spikes. A decay function over time is used to drive the potential membrane to hyperpolarization. We refer to the Supplemental Material for a detailed formal definition of LIF.

4 SPIKING EYE EMOTION NETWORK (SEEN)

Existing facial emotion recognition methods typically only identify the *"peak"* states of emotions [Hickson et al. 2019] or a single emotion state over a whole sequence [Zhao and Liu 2021], making these methods unsuitable for applications that also require a robust estimate of the in-between states. We introduce a lightweight Spiking Eye Emotion Network (SEEN) that is able to effectively recognize emotions from various states of emotions.

Instead of only memorizing the peak phase of an individual's facial emotion, SEEN is designed to leverage temporal cues to distinguish different phases of emotions using sparse events input captured with an event-based camera (DAVIS346 camera). Compared to a conventional camera, an event-based camera has a number of advantages: it is more sensitive to motion, less sensitive to ambient lighting, and it offers a high dynamic range. Hence, an event-based camera is capable of providing stable temporal information under different lighting conditions. While this makes event-based cameras, in theory, an attractive input modality for motion-based measurements, in practice, a major drawback of existing event-based cameras is that the recorded events are noisy and lack texture information. We address this drawback with a hybrid system that leverages both spatial cues together with conventional intensity frames to guide temporal feature extraction during training and inference. Most commercial event-based cameras are capable of simultaneously capturing both intensity frames and events through spatially-multiplexed sensing.

4.1 SEEN Architecture

As illustrated in Figure 2(a), at its core, the architecture of SEEN consists of a spatial feature extractor, S (described in detail in subsection 4.2), and a temporal feature extractor, T (detailed in subsection 4.3). Given two intensity frames, I^1 and I^n, SEEN interpolates the asynchronous captured events between both intensity frames in n synchronous event frames. Next, the spatial feature extractor S distills spatial cues from the intensity frames I^1 and I^n, and the temporal feature extractor T processes each of the n event frames sequentially in time order. Finally, the temporal features and the spatial cues are then combined to predict n emotion scores. The final predicted emotion is based on the average of the n scores. The core component of the temporal feature extractor T is the SNN layers that make decisions based on membrane potentials to remember temporal information from previous event frames. Unlike RNNs [Kag and Saligrama 2021; Nah et al. 2019], SNNs can effectively learn temporal dependencies of arbitrary length without any special treatment.

4.2 Spatial Feature Extractor S

To make spatial feature extraction independent from the intensity sequence length, we only use the first and last frames of a sequence as the input to the spatial feature extractor, thereby fixing the input size regardless of the sequence length, *i.e.,* two frames. The spatial feature extractor S (Figure 2(b)) leverages a multiscale self-attention perception module, Ω, to obtain discriminative features from different-sized neighborhoods. The extracted spatial features are then transferred into the spiking format, J_s, via a spiking layer, which is subsequently combined with temporal features to enhance feature discrimination. Formally, the spatial feature extractor can be defined as:

$$J_s = \Phi^1(F_s), \tag{2}$$

$$F_s = C_3(C_3(\Omega_{(3,5,7)}(l_s))), \tag{3}$$

$$\Omega_{(x_1,\ldots,x_n)}(\cdot) := C_1([\omega^{s_1}_{(x_1,\ldots,x_n)} C_{x_1}(\cdot), \ldots, \omega^{s_n}_{(x_1,\ldots,x_n)} C_{x_n}(\cdot)]), \tag{4}$$

$$\omega^{s_i}_{(x_1,\ldots,x_n)} = \sigma\left(\left\langle \Upsilon(C_{x_1}(l_s)), \ldots \Upsilon(C_{x_n}(l_s))\right\rangle\right)_i, \tag{5}$$

$$\Upsilon(\cdot) := C_1(\mathcal{BR}(C_1(\mathcal{A}(\cdot)))), \tag{6}$$

$$l_s = C_1([I^1, I^n]), \tag{7}$$

where $[\cdot]$ and $\langle\cdot\rangle$ indicate channel-wise concatenation and a vector, respectively; C_i and σ denote an $i\times i$ convolution layer and a softmax function, respectively; $\mathcal{A}$ denotes an adaptive pooling layer; $\mathcal{BR}$ is a fused batch normalization layer with a ReLU activation function; Φ^t is a spiking layer that keeps membrane potential from the previous time step, $t-1$. The initial membrane potential, *i.e.,* $t=0$, is set to 0 (see Equation 13).

4.3 Temporal Feature Extractor T

The basis building blocks of the temporal feature extractor T are SNN layers. An SNN neuron outputs signals based on a membrane potential accumulation, decay, and reset mechanisms to capture the temporal trends in an input sequence [Ding et al. 2022]. When the membrane potential exceeds a threshold, an action potential (*i.e.,* spike) is triggered and the membrane potential is reset. The trigger process itself is non-differentiable, prohibiting training via conventional stochastic gradient descent optimization methods. Instead, we adopt spatio-temporal backpropagation (STBP) along with a CNN-SNN layer [Wu et al. 2018] to circumvent this issue. This CNN-SNN layer employs a CNN-based layer for the aggregation process and a LIF-based SNN neuron [Gerstner and Kistler 2002] for managing the potential decay and reset processes. This modification takes advantage of CNN-based layers that enable learning of diverse accumulation strategies, resulting in more effective SNN neurons in the temporal domain.

Intensity Attention-Guided Temporal Features. Purely relying on events does not yield a robust solution due to the lack of reliable texture information in the event domain. We, therefore, leverage spatial features from S to inject rich texture cues. Figure 2(c) illustrates the architecture of the temporal extractor T.

The feature extractor T takes n event frames, E^1 to E^n, as input and processes each frame sequentially in time order. Formally, given the spatial feature J_s, the temporal feature extraction of E^t is defined

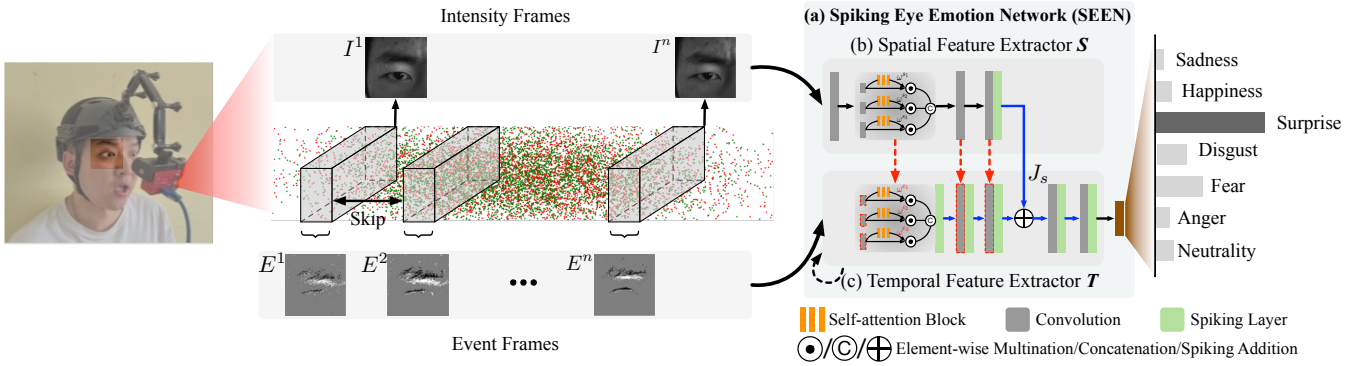


Figure 2: Our Spiking Eye Emotion Network (a) leverages a CNN-SNN-based temporal feature extractor, T, (c) to process n accumulated event frames, *i.e.,* E^i, in time order sequentially. During the process, based on two intensity frames, I^1 and I^n, the spatial feature extractor, S, (b) relies on a multiscale self-attention module to extract spatial features, J_s, which are combined with temporal cues to estimate emotions. The convolutional blocks before spiking-addition operator in the temporal feature extractor T fail to properly train due to lack of texture information in the event frames. Instead, we copy the updated weights from the corresponding blocks of the spatial feature extractor S. During inference, the attention weights are also copied directly from S to T to increase inference speed. The copying operations are marked by the red dashed arrows.

by:

$$O^t = \mathcal{M}(\Gamma(\Gamma(J_c^t))), \tag{8}$$

$$J_c^t = J_e^t \oplus J_s, \tag{9}$$

$$J_e^t = \Phi^t(F_e^t), \tag{10}$$

$$F_e^t = C_3(\Phi^t(C_3(\Phi^t(\Omega_{(3,5,7)}(E^t))))), \tag{11}$$

$$\Gamma(\cdot) := \Phi^t(\Psi(\cdot)), \tag{12}$$

where $\mathcal{M}$ is an operator for obtaining membrane potentials from an SNN layer, and Ψ represents a fully connected layer; $\Phi^t(\cdot)$ indicates an SNN layer, which records the previous spiking status, P^{t-1}, and potential value, V^{t-1}. When receiving membrane potentials X^t, this SNN layer outputs updated spikes, P^t, and updates the recorded membrane potential V^t as follows:

$$P^t = h(V^t - \Theta),$$
$$V^t = \alpha V^{t-1}(1 - P^{t-1}) + X^t,$$
$$h(x) = \begin{cases} 1 & x >= 0 \\ 0 & x < 0 \end{cases}, \tag{13}$$

where Θ is the membrane potential threshold set to 0.3 in all our experiments. The parameter α is a decay factor used for achieving hyperpolarization. The potential value V^t is updated such that, for a spike at timestamp $t-1$, the membrane potential should be reset to 0 by scaling $1 - P^{t-1}$, and X^t is the corresponding item here.

Finally, the emotion is the average of O^t, $t \in [1, n]$:

$$R = \sigma(\frac{1}{n}\sum_{t=1}^{n} O^t), \tag{14}$$

where σ is a Softmax activation function.

4.4 Weight-Copy Scheme

Intuitively, we want temporal information extraction to focus on informative spatial positions, such as facial action units [Ekman and Friesen 1978]. However, events lack sufficient texture information, which impedes the temporal feature extractor from considering spatial information. To alleviate this problem, we propose a weight-copy scheme that copies the weights from the spatial feature extractor to the temporal feature extractor. Thus, during training, only the fully connected layers in T are trained. The weight-copy scheme requires that all convolutional blocks before the spiking-addition operator, *i.e.,* Equation 9, are of the same architecture in S and T; see Equation 3 and Equation 11. Note that the supervised loss conveys the impact from both the spatial and temporal domains enabled by the spiking-addition. Since the weights are fixed before the spiking-addition in the temporal feature extractor T, the training of the spatial features must also account for temporal cues. Therefore, the weight updating implicitly bridges the domain gap between intensity and event frames.

Weight copying is also applied to the self-attention weights, *i.e.,* the self-attention weights in Equation 11 are replaced by the weights from Equation 5; see Figure 2(a). As we will show in our experimental results, this design is more effective than inferring the self-attention weights based on input events (row E in Table 2 except E4-S0) and it yields a more efficient inference.

4.5 Loss Function

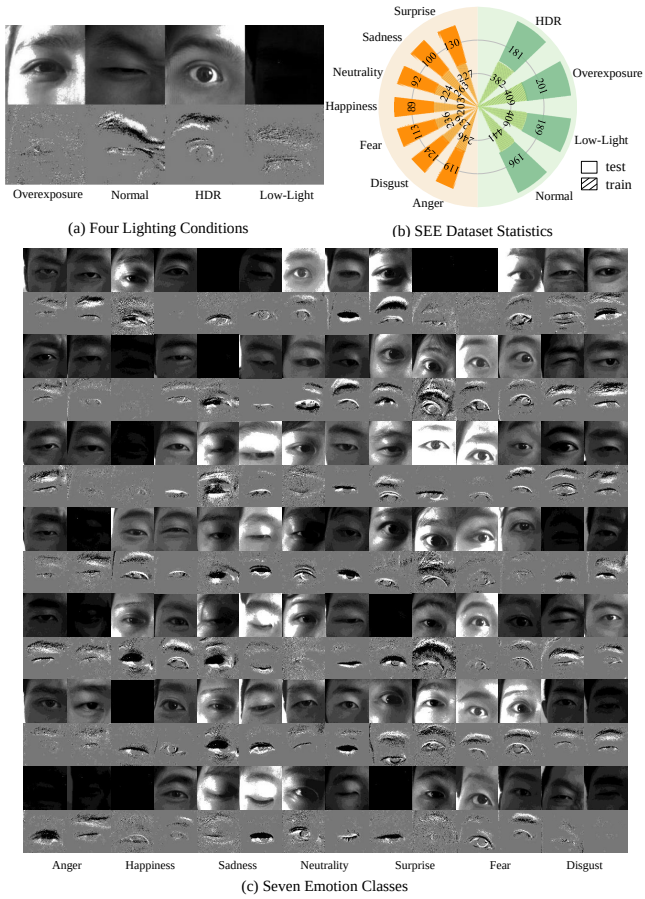
Because emotion recognition is a classification task, we use a regular cross-entropy loss for supervised training of SEEN:

$$\ell = -\frac{1}{7}\sum_{i=1}^{7} y_i \log(\hat{y}_i), \tag{15}$$

where y_i and $\hat{y}_i$ are the predicted i-th emotion's probability and corresponding ground truth probability, respectively.

5 DATASET

To the best of our knowledge, there does not exist an event-based dataset for single-eye emotion recognition. The two most related are



(a) Four Lighting Conditions

(b) SEE Dataset Statistics

(c) Seven Emotion Classes

Figure 3: The newly collected Single-eye Event-based Emotion (SEE) dataset covers seven emotion classes (c) under four lighting conditions (a). The detailed statistics of the SEE dataset are illustrated in (b).

the active infrared lighting/camera datasets Eyemotion [Hickson et al. 2019] (both eyes) and EMO [Wu et al. 2020] (single eye).

To address this lack of training data for event-based emotion recognition, we collect a new Single-eye Event-based Emotion (SEE) dataset; see Figure 3. SEE contains data from 111 volunteers captured with a DAVIS346 event-based camera placed in front of the right eye and mounted on a helmet; see Figure 1. The DAVIS346 camera is equipped with a dynamic version sensor (DVS) and an active pixel sensor (APS), providing both raw events and conventional frames simultaneously. Unlike Eyemotion and EMO, our approach does not require any active lighting source, thereby simplifying installation, testing, and maintenance of the hardware setup. A summary of the technical differences between SEE and the existing emotion datasets is provided in Supplementary Materials.

SEE contains videos of 7 emotions (see Figure 3(c) for an example) under four different lighting conditions: normal, overexposure, low-light, and high dynamic range (HDR) (Figure 3(a)). The average video length ranges from 18 to 131 frames, with a mean frame number of 53.5 and a standard deviation of 15.2 frames, reflecting the differences in the duration of emotions between subjects. In total, SEE contains 2,405/128,712 sequences/frames with corresponding raw events for a total length of 71.5 minutes (Figure 3(b)), which we split in 1,638 and 767 sequences for training and testing, respectively.

6 ASSESSMENT

The main goal of SEEN is to recognize an emotion for any phase of the emotion. Consequently, when evaluating a test sequence, we choose a uniformly distributed random starting point and corresponding testing length. A start point is selected such that the rest sequence is longer than the corresponding testing length. The testing length is defined as the total accumulation time of all included event frames, x, and a skip time, y, between two adjacent event frames, denoted as Ex-Sy. The skip time defines a window in the time domain where all events are ignored; see "skip" in Figure 2. Note that the skip time is not associated with event-based cameras but an experimental setting. Without loss of generality, the accumulation time and skip time are expressed as a multiple of 1/30 s. Thus, Ex-Sy indicates a testing length equal to $(x+(x-1)\times y)/30$ s. To reduce the impact of the randomness, we evaluate all competing methods 20 times for different randomly selected start points for each testing sequence; we use the same random starting points for single-frame competing methods, where only the random start frame is used. To evaluate the proposed approach and compare it to competing methods, we adopt two widely used metrics: Unweighted Average Recall (UAR) and Weighted Average Recall (WAR) [Schuller et al. 2011]. UAR reflects the average accuracy of different emotion classes without considering instances per class, while WAR indicates the accuracy of overall emotions; we refer to the Supplementary Materials for formal definitions of both metrics.

6.1 Training Setup

SEEN is implemented in PyTorch [Paszke et al. 2019] and trained with stochastic gradient descent (SGD) with a momentum of 0.9 and a weight decay of 1e−3. We train SEEN for 180 epochs with a batch size of 32 on an NVIDIA TITAN V GPU. We use the StepLR scheduler to moderate the learning rate. Specifically, the initial learning rate is set to 0.015, the step size is set to 1, and the decay rate is set to 0.94. For the SNN settings, we use a spiking threshold of 0.3 and a decay factor of 0.2 for all SNN neurons.

6.2 Qualitative and Quantitative Evaluation

We compare the effectiveness of SEEN to existing emotion recognition methods relying on conventional intensity images only, including whole-face, single-eye, and double-eye based methods. Of these prior methods, Eyemotion [Hickson et al. 2019] and EMO [Wu et al. 2020] are single-frame methods for predicting an emotion, while all other methods require the full video sequence. As shown in Table 1, SEEN for E4-S3 offers the best performance, outperforming the runner-up method, Eyemotion, by significant margins, 4.8% and 4.6% higher in WAR and UAR, respectively. Under normal, overexposure, and HDR lighting conditions, our approach with the same setting also outperforms Eyemotion by at least 4% in accuracy. However, Eyemotion offers slightly better performance under low-light

Table 1: Quantitative comparison against the state-of-the-art. All methods are retrained and tested on the SEE dataset. The abbreviations are defined as Ha → Happiness; Sa → Sadness; An → Anger; Di → Disgust; Su → Surprise; Fe → Fear; Ne → Neutrality; Nor → Normal; Over → Overexposure; Low → Low-Light. The first and second best results are highlighted in bold and underline, respectively.

Methods		Acc. of Emotion Class (%)							Acc. under Light Conditions (%)				Metrics (%)			
		Ha	Sa	An	Di	Su	Fe	Ne	Nor	Over	Low	HDR	WAR ↑	UAR ↑	FLOPS (G)	Time (ms)
Resnet18 + LSTM [2016; 1997]	Face	57.8	86.0	64.9	46.5	9.2	81.6	59.8	57.9	60.4	53.9	52.5	56.3	58.0	7.9	5.0
Resnet50 + GRU [2020a; 2016]	Face	27.9	38.0	49.7	44.5	6.9	70.0	5.6	43.0	35.7	28.9	32.8	35.2	34.7	17.3	10.3
3D Resnet18 [2018]	Face	54.8	45.4	67.7	23.8	37.2	42.8	81.6	51.9	51.4	44.8	47.8	49.1	50.5	8.3	21.2
R(2+1)D [2018]	Face	63.6	45.5	65.7	27.8	33.3	37.9	86.6	54.3	50.3	44.4	49.3	49.7	51.5	42.4	47.3
Former DFER [2021]	Face	81.5	75.2	85.8	59.4	39.3	50.8	78.6	70.1	65.4	66.2	61.1	65.8	67.2	8.3	7.7
Former DFER w/o pre-train	Face	44.1	65.2	46.0	66.5	28.0	50.3	36.1	47.0	51.9	45.6	47.2	48.0	48.0	8.3	7.7
Eyemotion [2019]	Eye	74.3	85.5	79.5	74.3	69.1	79.2	94.5	79.0	81.8	**81.5**	72.5	78.8	79.5	5.7	17.5
Eyemotion w/o pre-train	Eye	79.6	85.7	81.2	71.2	54.7	71.6	**96.4**	77.8	75.9	79.8	69.7	75.9	77.2	5.7	17.5
EMO [2020]	Eye	75.0	75.1	70.2	48.1	37.5	54.1	82.8	61.8	62.8	60.1	69.6	63.1	63.3	0.3	7.1
EMO w/o pre-train	Eye	62.0	73.2	60.1	38.7	25.7	48.0	65.3	46.1	60.2	55.5	58.9	53.2	53.3	0.3	7.1
Ours(E4-S0)	Eye	76.0	85.0	85.8	74.8	66.8	79.9	85.3	78.0	80.0	78.1	78.3	78.6	79.1	0.9	7.2
Ours(E4-S1)	Eye	76.9	89.2	88.9	76.3	69.0	82.3	86.6	78.5	83.4	80.5	81.0	80.9	81.3	0.9	7.2
Ours(E7-S0)	Eye	76.7	86.8	87.6	74.2	66.2	82.4	86.7	78.1	80.9	77.3	82.1	79.6	80.1	1.5	10.7
Ours(E4-S3)	Eye	**85.0**	89.9	**92.2**	76.7	**72.1**	**87.7**	85.2	**83.3**	85.6	80.8	**84.8**	**83.6**	**84.1**	0.9	7.2
Ours(E7-S1)	Eye	79.0	**90.9**	91.1	77.2	71.7	85.0	84.4	82.4	**86.7**	79.8	80.3	82.4	82.7	1.5	10.7
Ours(E13-S0)	Eye	77.9	88.7	90.2	**79.2**	69.7	87.6	84.6	81.1	86.5	79.4	81.8	82.3	82.5	2.6	19.0

conditions than SEEN with E4-S3. We posit that Eymotion benefits from the Imagenet[Deng et al. 2009] pre-training process; without this pre-training step, Eyemotion's accuracy is 1% less than the one offered by SEEN with E4-S3 setting. Moreover, we note that Eyemotion requires a personalization preprocessing step, which requires subtracting a mean neutral image for each person. Personalization dramatically increases the accuracy of neutral emotion estimation regardless of whether Eyemotion is pre-trained on ImageNet or not.

We compare SEEN with three different sequence lengths: 4/30 s, *i.e.,* E4-S0; 7/30 s, *i.e.,* E4-S1 group; 13/30 s, *i.e.,* E4-S3 group. The experimental results show that the accuracy of SEEN improves with longer sequence length under all lighting conditions, especially under HDR conditions. Note, all other prior video-based approaches require the full video sequences; consequently, their delay time is the length of an input sequence. In contrast, our method can flexibly adjust the delay time by changing input settings. Figures 4 and 5 qualitatively demonstrate the benefits of our method compared to prior eye-based emotion recognition methods. In Table 1, the complexity and processing speed of each competing approach are also provided. As the temporal feature extractor processes event frames iteratively, the complexity and processing time increase with the number of event frames. Nevertheless, with the E4-S3 setting, our method offers the second fastest processing speed, but it is more than 20% more accurate than the fastest method, EMO.

6.3 Ablation Study

To gain better insight into the abilities of SEEN, we perform a series of ablation studies that investigate a) the impacts of input, b) the influence of each component of SEEN, and c) the impact of outputs. Table 2 summarizes the experimental results.

Table 2: Quantitative ablation comparisons show that: a) both the first and last intensity frames are essential for providing discriminative features; b) all components of SEEN contribute to the overall performance (except experiment E under the E4-S0 setting); and c) potential averaging is necessary results in a more accurate performance.

		E4-S0		E4-S1		E4-S3	
	Networks	WAR	UAR	WAR	UAR	WAR	UAR
A	w/o I^n	77.1	77.6	79.9	80.2	81.3	81.8
B	$I^n \rightarrow I^2$	76.4	76.9	80.1	80.6	81.8	82.2
C	$[I^1, ..., I^n]$	78.0	78.4	79.9	80.2	82.9	83.3
D	No weight copy	77.5	78.0	79.6	80.0	82.1	82.6
E	No Att. weight copy	78.7	79.2	80.7	81.1	83.0	83.2
F	SNN → CNN	50.2	50.2	53.2	53.2	55.7	55.6
G	SNN → LSTM	52.9	53.0	55.3	55.2	55.8	55.7
H	SNN → Transformer	69.2	69.8	73.6	74.2	77.1	77.3
I	SNN → 3D CNN	54.3	54.3	57.7	57.7	59.9	59.9
J	Last potential	76.6	77.2	78.8	79.2	81.1	81.7
K	Last spike	55.7	54.8	59.5	58.9	63.2	62.8
L	Mean spike	63.5	63.2	64.1	63.6	69.7	69.5
M	Ours	78.6	79.1	80.9	81.3	83.6	84.1

Impacts of Input. SEEN leverages the first and last intensity frames. Experiments (A), (B) and (C) gauge the impact of the intensity frames: experiment (A) only uses the first intensity frame, experiment (B) replaces the last intensity frame with the second frame, and experiment (C) uses all the intensity frames corresponding to the included event frames. The results of (A) and (B) demonstrate spatial differences are critical for T to extract descriptive temporal cues. Compared to experiments (A) and (B), the results of

experiment (C) show that using more intensity frames slightly increases performance. However, compared to our method, the setup dramatically increases data bandwidth.

Influence of SEEN components. We investigate the effectiveness of the different components that comprise SEEN: 1) the effectiveness of the weight-copy scheme (experiments (D) and (E)) and 2) the benefits of SNNs (experiments (F) to (I)). These two experiment groups show that SEEN with all components offers the best performance, except experiment E under the E4-S0 setting. Experiments (F) to (I) show that replacing the CNN-SNN with a 3-layer CNN, LSTM, Transformer, or 3D CNN significantly degrades performance. A CNN fails to extract useful temporal cues, so the performance degradation further justifies the inclusion of temporal cues. Although LSTM, Transformer, and 3D CNN can extract temporal cues, they are less effective than SNNs. Notably, an SNN neuron's spiking mechanism acts as temporal memory and a natural noise filter, which is beneficial for robust emotion recognition.

Impact of outputs. SEEN estimates emotions based on the average of n membrane potentials; see Equation 8 and Equation 14. To better understand the impact of this design decision, we conduct three ablation experiments: instead of using the average of n membrane potentials, we define the prediction score based on the potential generated by the last event frame only (experiment (J)); similar to the previous but using output spikes instead of potential (experiment (K)); and finally using the average of n output spikes instead of the n membrane potentials for emotion classification, *i.e.,* remove the $\mathcal{M}$ operator in Equation 8 (experiment (L)). These results show that membrane potentials are more effective signals than spikes. We posit that the higher precision of membrane potentials (float vs. binary for spikes) offers more discriminative features for emotion classification. When a membrane potential triggers a spike, the potential is reset to 0. However, it becomes a problem if we leverage the potential as an output signal since the rest operation breaks the temporal cues. To address the problem, we design to use the average of the output potentials as the output signal. Experiment (J) validates the effectiveness of this design.

7 CONCLUSION

In this work, we introduce a novel wearable single-eye-based emotion recognition prototype that can effectively estimate emotions under challenging lighting conditions. To this end, we investigate event-based camera inputs for emotion recognition. Due to the high dynamic range and temporal resolution of event-based cameras, the captured events can robustly encode temporal information under different lighting conditions. However, the captured events are asynchronous, noisy, and lack texture cues. We introduce SEEN, a novel learning-based solution to extract informative temporal cues for emotion recognition. SEEN introduces two novel design components: a weight-copy scheme and a CNN-SNN-based temporal feature extractor. The former injects spatial attention into temporal feature extraction during the training and inference phases. The latter exploits both spatial awareness and the spiking mechanism of SNNs to provide discriminative features for emotion classification effectively. Our extensive experimental results show that SEEN can effectively estimate an emotion from any phase of the emotion. To the best of our knowledge, SEEN is the first attempt at leveraging event-based cameras and SNNs for emotion recognition tasks.

ACKNOWLEDGMENTS

This work was supported in part by the National Key Research and Development Program of China (2022ZD0210500), the National Natural Science Foundation of China under Grants 61972067/U21A2049-1, and the Distinguished Young Scholars Funding of Dalian (No. 2022RJ01). Pieter Peers was supported in part by NSF grant IIS-1909028. Felix Heide was supported by an NSF CAREER Award (2047359), a Packard Foundation Fellowship, a Sloan Research Fellowship, a Sony Young Faculty Award, a Project X Innovation Award, and an Amazon Science Research Award.

REFERENCES

Bradley M. Appelhans and Linda J. Luecken. 2006. Heart Rate Variability as an Index of Regulated Emotional Responding. *Review of General Psychology* 10, 3 (2006), 229–240. https://doi.org/10.1037/1089-2680.10.3.229

Xavier P. Burgos-Artizzu, Julien Fleureau, Olivier Dumas, Thierry Tapie, François LeClerc, and Nicolas Mollet. 2015. Real-Time Expression-Sensitive HMD Face Reconstruction. In *SIGGRAPH Asia 2015 Technical Briefs* (Kobe, Japan) *(SA '15)*. Association for Computing Machinery, New York, NY, USA, Article 9, 4 pages. https://doi.org/10.1145/2820903.2820910

Jean Costa, François Guimbretière, Malte F. Jung, and Tanzeem Choudhury. 2019. BoostMeUp: Improving Cognitive Performance in the Moment by Unobtrusively Regulating Emotions with a Smartwatch. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 2, Article 40 (jun 2019), 23 pages. https://doi.org/10.1145/3328911

David Couret, Pierre Simeone, Sébastien Freppel, and Lionel J Velly. 2019. The effect of ambient-light conditions on quantitative pupillometry: a history of rubber cup. *Neurocritical Care* 30 (2019), 492–493.

Didan Deng, Zhaokang Chen, and Bertram E Shi. 2020a. Multitask emotion recognition with incomplete labels. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)* (Buenos Aires, Argentina). IEEE, 592–599. https://doi.org/10.1109/FG47880.2020.00131

Didan Deng, Zhaokang Chen, Yuqian Zhou, and Bertram Shi. 2020b. Mimamo net: Integrating micro-and macro-motion for video emotion recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. Assoc Advancement Artificial Intelligence, 2621–2628.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 248–255. https://doi.org/10.1109/CVPR.2009.5206848

Jianchuan Ding, Bo Dong, Felix Heide, Yufei Ding, Yunduo Zhou, Baocai Yin, and Xin Yang. 2022. Biologically Inspired Dynamic Thresholds for Spiking Neural Networks. In *Advances in Neural Information Processing Systems*. https://doi.org/10.48550/arXiv.2206.04426

Paul Ekman and Wallace V Friesen. 1978. *Facial action coding systems*. Consulting Psychologists Press.

Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J. Davison, Jörg Conradt, Kostas Daniilidis, and Davide Scaramuzza. 2022. Event-Based Vision: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 1 (2022), 154–180. https://doi.org/10.1109/TPAMI.2020.3008413

Daniel Gehrig, Antonio Loquercio, Konstantinos G Derpanis, and Davide Scaramuzza. 2019. End-to-end learning of representations for asynchronous event-based data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5633–5643. https://doi.org/10.1109/ICCV.2019.00573

Daniel Gehrig, Michelle Rüegg, Mathias Gehrig, Javier Hidalgo-Carrió, and Davide Scaramuzza. 2021. Combining Events and Frames Using Recurrent Asynchronous Multimodal Networks for Monocular Depth Prediction. *IEEE Robotics and Automation Letters* 6, 2 (2021), 2822–2829. https://doi.org/10.1109/LRA.2021.3060707

Mariana-Iuliana Georgescu and Radu Tudor Ionescu. 2019. Recognizing facial expressions of occluded faces using convolutional neural networks. In *International Conference on Neural Information Processing*, Vol. 1142. Springer, 645–653. https://doi.org/10.1007/978-3-030-36808-1_70

Wulfram Gerstner and Werner M. Kistler. 2002. *Spiking Neuron Models: Single Neurons, Populations, Plasticity*.

Anna Gruebler and Kenji Suzuki. 2014. Design of a Wearable Device for Reading Positive Expressions from Facial EMG Signals. *IEEE Transactions on Affective Computing* 5, 3 (2014), 227–237. https://doi.org/10.1109/TAFFC.2014.2313557

Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. 2018. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?. In *Proceedings of the IEEE conference*

on Computer Vision and Pattern Recognition. 6546–6555. https://doi.org/10.1109/CVPR.2018.00685

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778. https://doi.org/10.1109/CVPR.2016.90

Steven Hickson, Nick Dufour, Avneesh Sud, Vivek Kwatra, and Irfan Essa. 2019. Eyemotion: Classifying facial expressions in VR using eye-tracking cameras. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 1626–1635. https://doi.org/10.1109/WACV.2019.00178

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Bita Houshmand and Naimul Mefraz Khan. 2020. Facial expression recognition under partial occlusion from virtual reality headsets based on transfer learning. In *2020 IEEE Sixth International Conference on Multimedia Big Data (BigMM)*. IEEE, 70–75. https://doi.org/10.1109/BigMM50055.2020.00020

Xinya Ji, Hang Zhou, Kaisiyuan Wang, Qianyi Wu, Wayne Wu, Feng Xu, and Xun Cao. 2022. EAMM: One-Shot Emotional Talking Face via Audio-Based Emotion-Aware Motion Model. In *ACM SIGGRAPH 2022 Conference Proceedings (SIGGRAPH '22)*. 1–10. https://doi.org/10.1145/3528233.3530745

Anil Kag and Venkatesh Saligrama. 2021. Time adaptive recurrent neural network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 15149–15158. https://doi.org/10.1109/CVPR46437.2021.01490

Xavier Lagorce, Garrick Orchard, Francesco Galluppi, Bertram E Shi, and Ryad B Benosman. 2017. Hots: a hierarchy of event-based time-surfaces for pattern recognition. *IEEE transactions on pattern analysis and machine intelligence* 39, 7 (2017), 1346–1359. https://doi.org/10.1109/TPAMI.2016.2574707

Jiyoung Lee, Seungryong Kim, Sunok Kim, Jungin Park, and Kwanghoon Sohn. 2019. Context-aware emotion recognition networks. In *Proceedings of the IEEE/CVF international conference on computer vision*. 10143–10152. https://doi.org/10.1109/ICCV.2019.01024

Jiyoung Lee, Sunok Kim, Seungryong Kim, and Kwanghoon Sohn. 2020. Multi-modal recurrent attention networks for facial expression recognition. *IEEE Transactions on Image Processing* 29 (2020), 6977–6991. https://doi.org/10.1109/TIP.2020.2996086

Hao Li, Laura Trutoiu, Kyle Olszewski, Lingyu Wei, Tristan Trutna, Pei-Lun Hsieh, Aaron Nicholls, and Chongyang Ma. 2015. Facial Performance Sensing Head-Mounted Display. *ACM Trans. Graph.* 34, 4, Article 47 (jul 2015), 9 pages. https://doi.org/10.1145/2766939

Mi Li, Hongpei Xu, Xingwang Liu, and Shengfu Lu. 2018. Emotion recognition from multichannel EEG signals using K-nearest neighbor classification. *Technology and Health Care* 26 (04 2018), 509–519. https://doi.org/10.3233/THC-174836

Junxiu Liu, Guopei Wu, Yuling Luo, Senhui Qiu, Su Yang, Wei Li, and Yifei Bi. 2020. EEG-Based Emotion Classification Using a Deep Neural Network and Sparse Autoencoder. *Frontiers in Systems Neuroscience* 14 (2020). https://doi.org/10.3389/fnsys.2020.00043

Jorge C. Lucero and Kevin G. Munhall. 1999. A model of facial biomechanics for speech production. *The Journal of the Acoustical Society of America* 106 5 (1999), 2834–2842. https://doi.org/10.1121/1.428108

Ana I Maqueda, Antonio Loquercio, Guillermo Gallego, Narciso García, and Davide Scaramuzza. 2018. Event-based vision meets deep learning on steering prediction for self-driving cars. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5419–5427. https://doi.org/10.1109/CVPR.2018.00568

Sebastiaan Mathôt. 2018. Pupillometry: Psychology, Physiology, and Function. *Journal of Cognition* 1 (02 2018). https://doi.org/10.5334/joc.18

Seungjun Nah, Sanghyun Son, and Kyoung Mu Lee. 2019. Recurrent neural networks with intra-frame iterations for video deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8094–8103. https://doi.org/10.1109/CVPR.2019.00829

Jingping Nie, Yigong Hu, Yuanyuting Wang, Stephen Xia, and Xiaofan Jiang. 2020. SPIDERS: Low-Cost Wireless Glasses for Continuous In-Situ Bio-Signal Acquisition and Emotion Recognition. In *2020 IEEE/ACM Fifth International Conference on Internet-of-Things Design and Implementation (IoTDI)*. 27–39. https://doi.org/10.1109/IoTDI49375.2020.00011

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 8024–8035. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

Delian Ruan, Yan Yan, Shenqi Lai, Zhenhua Chai, Chunhua Shen, and Hanzi Wang. 2021. Feature decomposition and reconstruction learning for effective facial expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7660–7669. https://doi.org/10.1109/CVPR46437.2021.00757

Enrique Sanchez, Mani Kumar Tellamekala, Michel Valstar, and Georgios Tzimiropoulos. 2021. Affective Processes: stochastic modelling of temporal context for emotion and facial expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9074–9084. https://doi.org/10.1109/CVPR46437.2021.00896

B. Schuller, B. Vlasenko, F. Eyben, M. Wo?Llmer, A. Stuhlsatz, A. Wendemuth, and G. Rigoll. 2011. Cross-Corpus Acoustic Emotion Recognition: Variances and Strategies. *IEEE Transactions on Affective Computing* 1, 2 (2011), 119–131. https://doi.org/10.1109/T-AFFC.2010.8

Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. 2018. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 6450–6459. https://doi.org/10.1109/CVPR.2018.00675

Yanxiang Wang, Xian Zhang, Yiran Shen, Bowen Du, Guangrong Zhao, Lizhen Cui Cui Lizhen, and Hongkai Wen. 2022. Event-Stream Representation for Human Gaits Identification Using Deep Neural Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 7 (2022), 3436–3449. https://doi.org/10.1109/TPAMI.2021.3054886

Hao Wu, Jinghao Feng, Xuejin Tian, Edward Sun, Yunxin Liu, Bo Dong, Fengyuan Xu, and Sheng Zhong. 2020. EMO: Real-time emotion recognition from single-eye images for resource-constrained eyewear devices. In *Proceedings of the 18th International Conference on Mobile Systems, Applications, and Services*. 448–461. https://doi.org/10.1145/3386901.3388917

Yujie Wu, Lei Deng, Guoqi Li, Jun Zhu, and Luping Shi. 2018. Spatio-temporal backpropagation for training high-performance spiking neural networks. *Frontiers in neuroscience* 12 (2018), 331. https://doi.org/10.3389/fnins.2018.00331

Fanglei Xue, Qiangchang Wang, and Guodong Guo. 2021. Transfer: Learning relation-aware facial expression representations with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3601–3610. https://doi.org/10.1109/ICCV48922.2021.00358

Jiqing Zhang, Xin Yang, Yingkai Fu, Xiaopeng Wei, Baocai Yin, and Bo Dong. 2021b. Object tracking by jointly exploiting frame and event domain. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 13043–13052. https://doi.org/10.1109/ICCV48922.2021.01280

Yuhang Zhang, Chengrui Wang, and Weihong Deng. 2021a. Relative Uncertainty Learning for Facial Expression Recognition. *Advances in Neural Information Processing Systems* 34 (2021), 17616–17627.

Zengqun Zhao and Qingshan Liu. 2021. Former-DFER: Dynamic Facial Expression Recognition Transformer. In *Proceedings of the 29th ACM International Conference on Multimedia*. 1553–1561. https://doi.org/10.1145/3474085.3475292

Figure 4: We show four examples across four different emotions, Fear, Anger, Disgust, and Happiness, under overexposure and normal lighting conditions. The frames marked with red boxes are the inputs for EMO [Wu et al. 2020] and Eyemotion [Hickson et al. 2019], which is also the first input frame of our approach. Our approach offers the most accurate emotion predictions under all test settings.

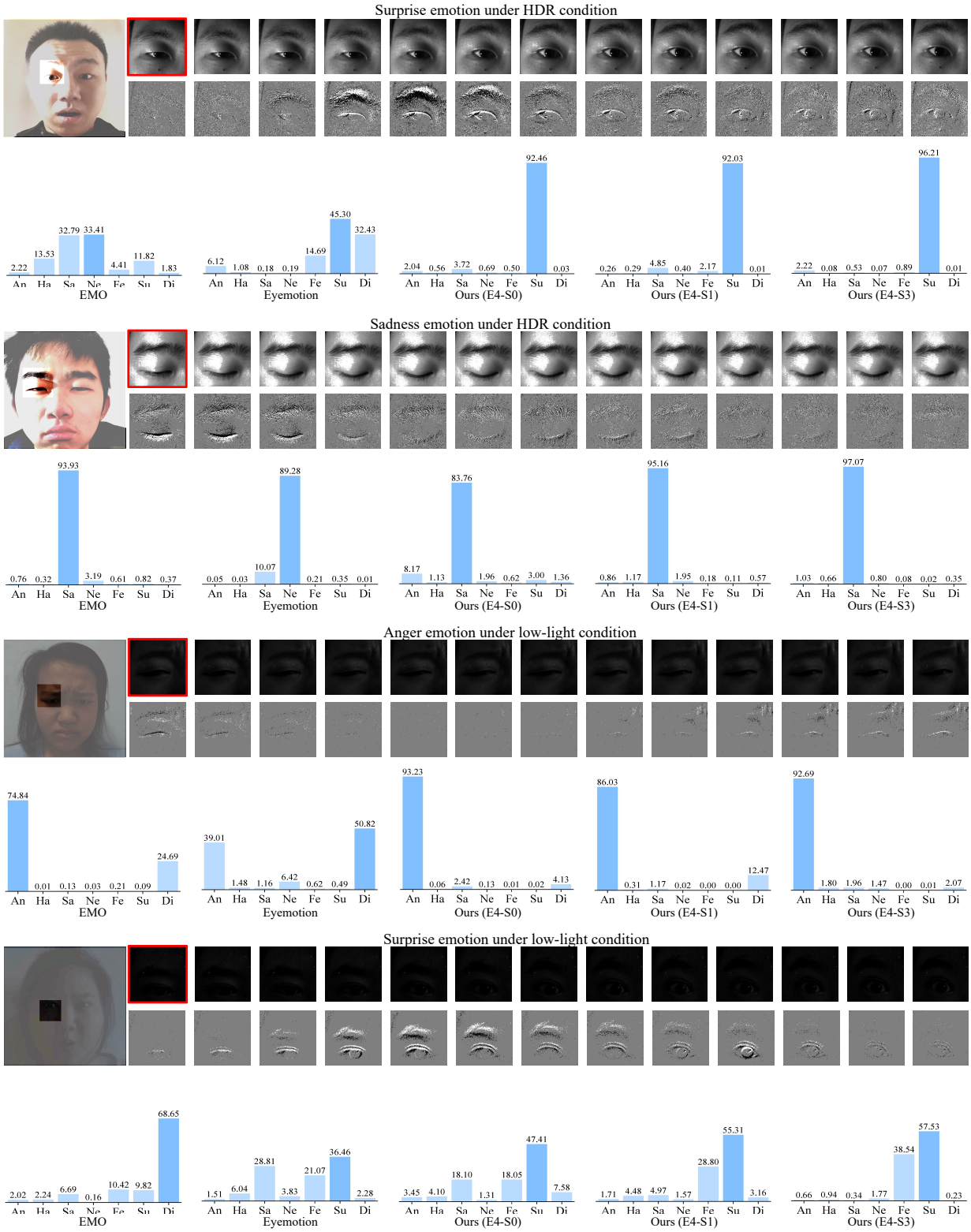


Figure 5: We show four additional examples across another four different emotions, Surprise, Sadness, Anger, and Surprise, under HDR and low-light conditions. The frames marked with red boxes are the inputs for EMO [Wu et al. 2020] and Eyemotion [Hickson et al. 2019], which is also the first input frame of our approach. Our approach offers the most accurate emotion predictions under all test settings.