

Apprenticeship-Inspired Elegance: Synergistic Knowledge Distillation Empowers Spiking Neural Networks for Efficient Single-Eye Emotion Recognition

Yang Wang^{1,2}, Haiyang Mei^{2,1}, Qirui Bao¹, Ziqi Wei³, Mike Zheng Shou², Haizhou Li^{5,2}, Bo Dong⁴ and Xin Yang^{1*}

¹Key Laboratory of Social Computing and Cognitive Intelligence, Dalian University of Technology

²National University of Singapore

³Institute of Automation, Chinese Academy of Sciences

⁴Independent Researcher

⁵The Chinese University of Hong Kong, Shenzhen (CUHK-Shenzhen)

{yangwang06,1612000589}@mail.dlut.edu.cn, haiyang.mei@outlook.com, ziqi.wei@ia.ac.cn, {mike.zheng.shou,dongshuhao12}@gmail.com, haizhouli@cuhk.edu.cn, xinyang@dlut.edu.cn

Abstract

We introduce a novel multimodality synergistic knowledge distillation scheme tailored for efficient single-eye motion recognition tasks. This method allows a lightweight, unimodal student spiking neural network (SNN) to extract rich knowledge from an event-frame multimodal teacher network. The core strength of this approach is its ability to utilize the ample, coarser temporal cues found in conventional frames for effective emotion recognition. Consequently, our method adeptly interprets both temporal and spatial information from the conventional frame domain, eliminating the need for specialized sensing devices, *e.g.*, event-based camera. The effectiveness of our approach is thoroughly demonstrated using both existing and our compiled single-eye emotion recognition datasets, achieving unparalleled performance in accuracy and efficiency over existing state-of-the-art methods.

1 Introduction

Real-time emotion recognition is pivotal in enhancing human-centered interactive experiences, such as virtual reality (VR) and augmented reality (AR) applications [Picard, 2003; Zhang *et al.*, 2023]. This technology’s proficiency in precisely decoding users’ emotional states can greatly elevate the VR/AR experience. More significantly, it facilitates the personalization of these experiences to cater to the distinct emotional requirements of each user, thereby offering uniquely immersive experiences and boosting engagement.

In the context of VR/AR, the devices are typically affixed to a user’s face, which inherently accommodates the variances in performance that may arise from different head positions. While beneficial for head pose accommodation, this placement presents a significant challenge: the majority of the facial area is obscured by the device, diminishing the effectiveness of traditional emotion recognition methods that

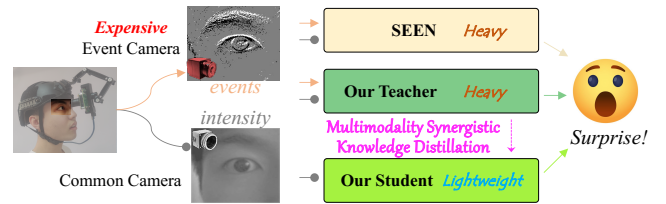


Figure 1: Surpassing the state-of-the-art SEEN in real-time single-eye emotion recognition (SER), our method achieves lightweight inference by requiring solely intensity frames, obviating the need for data from expensive event cameras. This is facilitated by our novel synergistic knowledge distillation strategy, enabling real-time and accurate SER on resource-constrained devices for the first time.

rely on facial action units. To counteract this limitation, the focus is shifting towards eye-based emotion recognition techniques [Hickson *et al.*, 2019; Wu *et al.*, 2020]. Yet, these methods often depend on personalized initialization or necessitate capturing the peak phase of an emotion [Hickson *et al.*, 2019], which can limit their practicality.

Recently, an event-based single-eye-based approach [Zhang *et al.*, 2023] capitalizes on both temporal and spatial cues, enhancing accuracy in emotion recognition without needing personalization or capturing peak emotional phases. It proves effective under various real-world lighting conditions, including low-light and high-dynamic-range environments. The success of this approach is underpinned by the advantages of event-based cameras, which offer a higher dynamic range (140 dB vs. 80 dB in traditional cameras) and a significantly finer temporal resolution (about 0.001 ms, in contrast to 10 ms in conventional frame-based cameras). However, despite these advancements, event-based cameras are still nascent, especially when compared to the more established conventional frame-based camera technologies. This relative immaturity translates to higher costs, making the widespread adoption of event-based cameras in VR/AR devices not currently a cost-effective solution.

In our study, we developed an *apprenticeship-learning-based* approach to address the aforementioned limitation.

*Corresponding author.

Our method efficiently learns from both event and frame domains but uses only conventional frames for inference; see Figure 1. In particular, our approach employs a knowledge distillation process, where a teacher network, trained on multimodal data (event and frame), transfers its insights to a student network. This student network is then adept at harnessing spatial and temporal cues from conventional frames. The key to this method’s success lies in leveraging the sufficient coarser temporal cues present in conventional frames for effective emotion recognition, eliminating the need for more expensive event-based cameras. The efficacy of our distillation scheme is reinforced by two novel consistency losses we developed: hit consistency and temporal consistency. Hit consistency ensures a match in the correct prediction distribution between the teacher and student networks. Temporal consistency, on the other hand, encompasses all temporal predictions, considering both correct and incorrect ones. We extensively validate the efficacy of our approach on both SEE dataset [Zhang *et al.*, 2023] and our Diverse Single-eye Event-based Emotion (DSEE) datasets, demonstrating the best performance among all competing state-of-the-art methods. In summary, our contributions include:

- Developing a novel framework for an unimodal student network to distill knowledge from an event-enhanced multimodal teacher network;
- Creating a new large-scale dataset with both frame and event for advancing eye-based emotion recognition;
- Achieving significant improvements in accuracy and efficiency over competing state-of-the-art methods.

2 Related Work

Facial-based emotion recognition has received significant attention in recent years given its diverse and practical applications in the field of security, health, communication, etc. A number of facial emotion recognition datasets have been developed to facilitate the research and development of this field, such as CK+ [Lucey *et al.*, 2010], MUG [Aifanti *et al.*, 2010], MMI [Pantic *et al.*, 2005], Oulu-CASIA [Zhao *et al.*, 2011], and ADFES [Van Der Schalk *et al.*, 2011]. The majority of previous research focuses on analyzing the entire face, with various techniques introduced for effective facial feature learning [Xue *et al.*, 2021; Ruan *et al.*, 2021], addressing uncertainties in facial expression data [Zhang *et al.*, 2021b], handling partial occlusions [Georgescu and Ionescu, 2019; Houshmand and Khan, 2020], and utilizing temporal cues [Sanchez *et al.*, 2021; Deng *et al.*, 2020]. However, in many practical scenarios, it is not always feasible to observe the entire face, which triggers growing interest in identifying emotions based solely on information from the eye area.

Eye-based emotion recognition is a branch of occluded facial emotion recognition. Years of dedicated investigation have yielded substantial progress, demonstrating the potential of this avenue for enhancing the accuracy and robustness of emotion detection. [Hickson *et al.*, 2019] utilized images of both eyes captured with an infrared gaze-tracking camera within a virtual reality headset to infer emotional expressions while [Wu *et al.*, 2020] relied on infrared single-eye observations to address camera synchronization and data bandwidth

issues when monitoring both eyes. Both constructed systems necessitate a personalized initialization procedure: the former requires a personalized neutral image while the latter needs a reference feature vector of each emotion. The requirement for a personalized setup renders these systems intrusive and non-transparent to the user, potentially raising privacy concerns. Additionally, neither system incorporates temporal cues, which are crucial for robust emotion recognition [Sanchez *et al.*, 2021]. Most recently, [Zhang *et al.*, 2023] proposed a new Single-eye Event-based Emotion dataset (SEE) and a real-time emotion recognition method SEEN that integrates event and intensity cues and achieves enhanced emotion recognition. Our method distinguishes itself from SEEN by distilling enriched knowledge from both event and intensity modalities to a lightweight SNN model during training. This eliminates the requirement for the expensive event camera during the inference stage while enabling high-speed and low-energy-cost emotion recognition.

SNN-based knowledge distillation has emerged as a promising approach to address the challenges of training deep SNNs directly with a loss function, which is hindered by the non-differentiable nature of spiking signals [Wei *et al.*, 2024; Zhang *et al.*, 2021a; Liu *et al.*, 2020a; Liu *et al.*, 2020b]. This technique unlocks the potential of deep SNNs for efficient yet accurate inference. Related works can be broadly categorized into two categories: those distilling knowledge [Ji *et al.*, 2023] from differently-structured and pre-trained artificial neural networks (ANNs) [Xu *et al.*, 2023b; Takuya *et al.*, 2021] or SNNs [Xu *et al.*, 2023a; Kushawaha *et al.*, 2021], and those distilling knowledge from itself [Dong *et al.*, 2023; Deng *et al.*, 2022]. Unlike the prior studies, our proposed multimodality knowledge distillation strategy achieves knowledge transfer from a multimodal input network to an unimodal input network.

3 Methodology

While existing research has demonstrated the effectiveness of multimodal networks for real-time emotion recognition (*e.g.*, [Zhang *et al.*, 2023]), their inherent complexity and need for multimodal data raise the question: *can a lightweight unimodal network achieve comparable performance?* In this work, we make the first investigation into this question by proposing a novel method leveraging knowledge distillation.

As illustrated in Figure 2(a), we first train a cumbersome teacher network that takes as input the events data and intensity frames. The teacher network uses an SNN-based SEW-Resnet-18 [Fang *et al.*, 2021] and a CNN-based Resnet-18 [He *et al.*, 2016] to extract event features and intensity features, respectively, based on which to perform emotion recognition via features fusion and fully connected classifier. We then construct a lightweight SNN-based student network that operates solely on intensity frames and consists of five consecutive feature extraction layers and a classifier, optimizing it using a classification loss (subsection 3.1) alongside two synergistic knowledge distillation losses that ensure prediction distribution harmony with the teacher network at both granular (hit consistency, subsection 3.2) and comprehensive (temporal consistency, subsection 3.3) levels. Formally, the

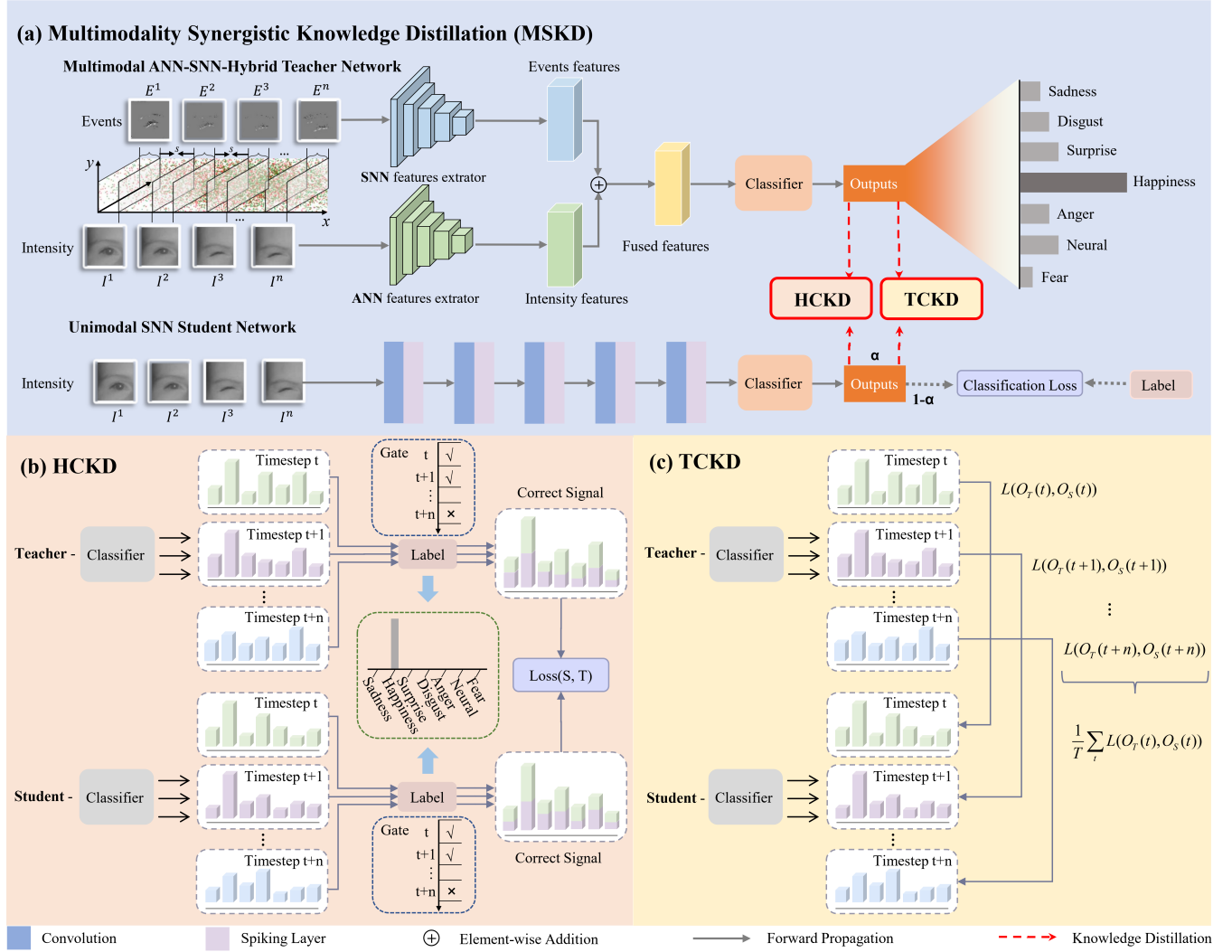


Figure 2: Overview of our proposed multimodality synergistic knowledge distillation (MSKD) framework (a) which consists of a multimodal input ANN-SNN-hybrid teacher network (top) and an unimodal input SNN student network (bottom), as well as two synergistic knowledge distillation loss items: (b) hit consistency loss and (c) temporal consistency loss.

loss function is defined as:

$$\mathcal{L}_{\text{total}} = (1 - \alpha) * \mathcal{L}_{\text{Cls}} + \alpha * \mathcal{L}_{\text{HCKD}} + \alpha * \mathcal{L}_{\text{TCKD}}, \quad (1)$$

where $\mathcal{L}_{\text{Cls}}$ is the classification loss to enforce the output at each timestep close to the true label; $\mathcal{L}_{\text{HCKD}}$ and $\mathcal{L}_{\text{TCKD}}$ are the hit consistency and temporal consistency knowledge distillation losses to improve the prediction distribution matching with the teacher network at the individual correctly predicted timestamp and each of all timestamps, respectively; and α is the weighting parameter to balance the classification loss and distillation losses, which is initialized as 0.5 and increased by 0.1 after every 30 training epochs.

3.1 Classification Loss

Typically, it's not easy to efficiently train deep SNNs due to the non-differentiability of its activation function [Ding *et al.*, 2022; Zhang *et al.*, 2022; Wang *et al.*, 2023], which disables

the widely used gradient descent approaches for traditional ANNs. Although the adoption of surrogate gradient (SG) formally allows for the back-propagation of losses, the discrete spiking mechanism differentiates the loss landscape of SNNs from that of ANNs, failing the SG methods to achieve desirable accuracy. To alleviate this, we follow [Deng *et al.*, 2022] to adopt the temporal efficient training (TET) approach to compensate for the loss of momentum in the gradient descent with SG so that the training process can converge into flatter minima with better generalizability. Equipped with TET, our classification cross-entropy loss can be defined as:

$$\mathcal{L}_{\text{Cls}} = \frac{1}{T} \sum_{t=1}^T \mathcal{L}_{\text{CE}}(O_{\text{stu}}(t), y), \quad (2)$$

where $O_{\text{stu}}(t)$ represents the pre-synaptic input of the student classifier at the t -th timestep; $\mathcal{L}_{\text{CE}}$ denotes the cross-entropy loss; T is the total timesteps; and y indicates the ground truth

one-hot vector. Different from [Zhang *et al.*, 2023] that directly optimizes the integrated potential, our classification loss optimizes every moment’s pre-synaptic inputs, which helps the network have more robust time scalability.

3.2 Hit Consistency Knowledge Distillation

Despite with lightweight architecture, the student network is expected to recognize emotions as correctly as the cumbersome teacher network. This inspires our hit consistency knowledge distillation (HCKD) loss (Figure 2(b)) which penalizes the distribution difference between teacher and student networks *at the correctly-predicted timestep*. Formally,

$$\mathcal{L}_{\text{HCKD}} = \mathcal{L}_{\text{MSE}}(S_{stu}, S_{tea}), \quad (3)$$

$$S_{stu} = \frac{1}{C_{stu}} \sum_{c_{stu}=1}^{C_{stu}} O_{stu}(c_{stu}), \quad (4)$$

$$S_{tea} = \frac{1}{C_{tea}} \sum_{c_{tea}=1}^{C_{tea}} O_{tea}(c_{tea}), \quad (5)$$

where $\mathcal{L}_{\text{MSE}}$ measures the mean squared error between student correctly-predicted signal S_{stu} and teacher correctly-predicted signal S_{tea} . S_{stu}/S_{tea} is obtained by averaging C_{stu}/C_{tea} student/teacher prediction distributions $O_{stu}(c_{stu})/O_{tea}(c_{tea})$ at the correctly-predicted timestep c_{stu}/c_{tea} . When C_{stu}/C_{tea} equals to zero, we assign $1/N_c$ as its value (N_c is the total number of emotion categories).

3.3 Temporal Consistency Knowledge Distillation

The HCKD emphasizes the harmony of averaged distributions for all correctly-predict timesteps between student and teacher networks, which could help the student network approach the teacher network in terms of the *overall* correct predictions. However, HCKD does not consider the temporal consistency between student and teacher networks at each timestep, ignoring the distillation of rich knowledge embedded in the temporal patterns and dynamics. To address this limitation, we introduce a temporal consistency knowledge distillation (TCKD) loss (Figure 2(c)), which enforces the student network to learn temporal patterns and dynamics from the teacher by quantifying the discrepancy in temporal dynamics, via computing the mean squared error between their prediction distributions *at each timestep*. Formally,

$$\mathcal{L}_{\text{TCKD}} = \frac{1}{T} \sum_{t=1}^T \mathcal{L}_{\text{MSE}}(O_{stu}(t), O_{tea}(t)), \quad (6)$$

where T is the number of all timesteps. Finally, by combining HCKD loss and TCKD loss, we form a new and powerful synergistic knowledge distillation strategy to empower a lightweight unimodal student network for efficient SER.

4 Dataset

The scarcity of publicly available datasets is a major challenge in eye-based emotion recognition research. Early related datasets include the active infrared lighting/camera datasets Eyemotion [Hickson *et al.*, 2019] (both eyes) and EMO [Wu *et al.*, 2020] (single eye). Recently, [Zhang *et al.*, 2023]

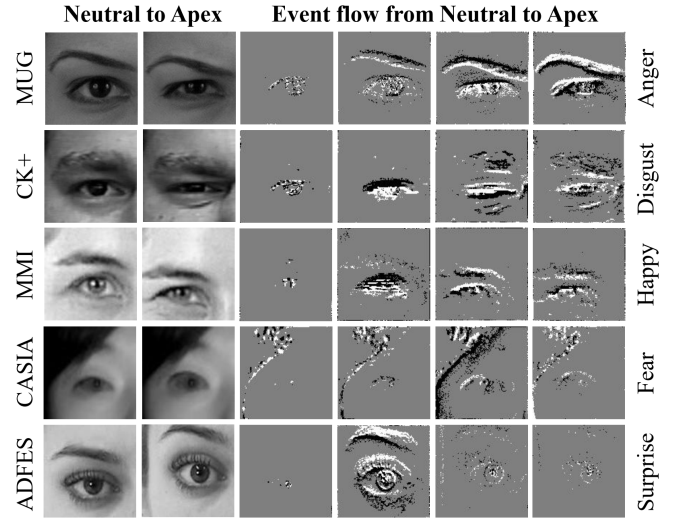


Figure 3: Examples from our DSEE dataset.

et al., 2023] collected a Single-eye Event-based Emotion (SEE) dataset for event-enhanced emotion recognition. 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. Despite the pioneering effort to introduce events modality for improving recognition accuracy, SEE is highly imbalanced, with a majority of samples from challenging scenes and a minority of samples from normal scenes. In addition, SEE has limited example diversity in terms of participants’ age, gender, race, etc. These two shortcomings significantly impair the generalization ability of models trained on this dataset.

To address the above limitations, we introduce a Diverse Single-eye Event-based Emotion (DSEE) dataset. DSEE contains intensity video frames and corresponding real/synthetic events as well as a ground truth emotion label. To the best of our knowledge, DSEE is currently the largest single-eye event-based emotion benchmark (kindly see Table 1 for a summary and Figure 3 for representative examples).

4.1 Protocols for Data Acquisition

Besides including SEE [Zhang *et al.*, 2023] as a subset, we ensure a wide diversity and broad coverage of our DSEE by inheriting and further processing existing facial emotion recognition video datasets including CK+ [Lucey *et al.*, 2010], MUG [Aifanti *et al.*, 2010], MMI [Pantic *et al.*, 2005], Oulu-CASIA [Zhao *et al.*, 2011], and ADFES [Van Der Schalk *et al.*, 2011]. As illustrate in Figure 4, we first use a multi-task cascaded convolutional network (MTCNN) [Zhang *et al.*, 2016] to locate and crop the right eye regions from the given facial sequence. Then we resize the cropped region to a fixed resolution (*i.e.*, 128×128) to accommodate different instances. Next, we feed the resized crop sequence into v2e [Hu *et al.*, 2021], a video-to-event converter for realistic events simulation, to obtain the corresponding raw events. Finally, we follow prior works [Rebecq *et al.*, 2019; Mei *et al.*, 2023; Delbruck *et al.*, 2023] to convert the raw events into event frames. By the above steps, we can obtain

Datasets	Sequences	Frames	Subjects	Emotion	Age range	Race number	Intensity	Events	Real	Synthetic
MUG	983	70596	52	7	20-35	1	✓	✓	×	✓
CK+	327	5876	118	7	18-50	3	✓	✓	×	✓
MMI	126	11099	21	6	20-32	3	✓	✓	×	✓
Oulu-CASIA	1440	31995	80	6	23-58	1	✓	✓	×	✓
Adfaces	216	32350	22	10	18-25	2	✓	✓	×	✓
SEE	3224	175185	113	7	19-28	1	✓	✓	✓	×
DSEE (Ours)	6235	317447	394	7	18-58	6	✓	✓	✓	✓

Table 1: Comparison among event-based datasets for emotion recognition.

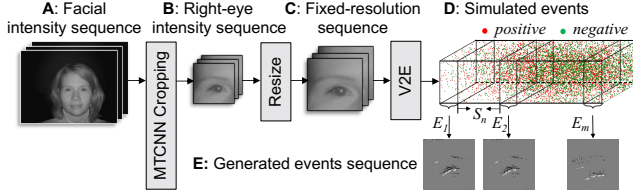


Figure 4: Illustration of the single-eye events data synthesis process.

the intensity sequence and corresponding events sequence as well as emotion label triplets, covering diverse examples.

4.2 Dataset Statistical Analysis

We further present dataset statistics in in Table 1. It can be seen that (i) our DSEE dataset provides a substantial edge with its abundance of sequences, frames, and subjects, exceeding the capacity of current alternatives and enabling robust and generalizable research endeavors; (ii) our DSEE also covers participants with a wider range of age and more race numbers, providing more diversity for training and evaluating the model; and (iii) both real and synthetic events data are included in our DSEE, facilitating both the synthetic-data-based and real-data-based research and experiments, as well as the further exploration of synthetic-to-real transfer.

5 Experiments

5.1 Experimental Settings

Implementation Details. We implement our method in PyTorch [Paszke *et al.*, 2019] and perform the experiments on a server with Intel(R) Xeon(R) Gold 6240R @ 2.40GHz CPU and NVIDIA GeForce RTX 4090 GPU. The training of our method can be divided into two stages. First, we pretrain the ANN-SNN-hybrid teacher network with both intensity frame and events data as input, for 180 epochs with a batch size of 32. The training is optimized via stochastic gradient descent (SGD) with a momentum of 0.9 and a weight decay of 0.001. The learning rate is initialized as 0.015 and decayed by a factor of 0.94 after each epoch. Second, our multimodality synergistic knowledge distillation strategy is adopted for the training of unimodal SNN student network. Except for the loss function, other training configurations for the student network are identical to those of the teacher network.

Evaluation Dataset and Metrics. We evaluate the effectiveness of method on both the existing event-based single-eye emotion recognition dataset SEE [Zhang *et al.*, 2023]

as well as our newly constructed DSEE dataset. We adopt two widely used metrics for quantitatively assessing the emotion classification performance: Unweighted Average Recall (UAR) and Weighted Average Recall (WAR). UAR reflects the average accuracy of different emotion classes without considering instances per class, while WAR indicates the accuracy of overall emotions [Schuller *et al.*, 2011]. Formally,

$$UAR = \frac{1}{N_c} \sum_{i=1}^{N_c} \frac{TP_i}{TP_i + FN_i}, \quad (7)$$

$$WAR = \frac{TP + FN}{TP + TN + FP + FN}, \quad (8)$$

where N_c is the total number of emotion classes; TP and FP are true positive and false positive, respectively; TN and FN are true negative and false negative, respectively. For UAR and WAR, a higher value indicates a better performance.

5.2 Comparison to State-of-the-arts

Table 2 reports the comparison results of our proposed MSKD approach against state-of-the-art methods on SEE dataset [Zhang *et al.*, 2023]. From the results, we observe that: (i) eye-based emotion recognition methods are generally superior to facial-based ones. This reflects that the eyes serve as a crucial region for emotional expression as it contains rich information about subtle nuances in facial expressions. Focusing on this specific region enables more precise feature extraction, enhancing the discriminative power of the algorithm. Furthermore, eye-based methods are more robust to different light conditions. The reason behind this is that unlike the entire face, which may exhibit variations in shadowing and contrast under different lighting conditions, the eyes are less susceptible to these variations. The relatively small size of the eyes and their position in the middle of the face makes them less prone to extreme illumination changes, resulting in increased algorithmic robustness; (ii) our teacher network achieves the best overall performance. For example, it outperforms the existing state-of-the-art method SEEN [Zhang *et al.*, 2023] by 2.6% and 2.5% in terms of WAR and UAR, respectively. This demonstrates the superior capability of the teacher network, laying a strong foundation for the subsequent training of the student network; and (iii) the proposed multimodality synergistic knowledge distillation strategy enables the student network to achieve comparable results to SEEN [Zhang *et al.*, 2023] despite the absence of event data. This demonstrates the effectiveness of our strategy in knowledge transfer and opens doors for efficient model training in

Methods	input region	Accuracy for different emotions							Accuracy for different lightings				Metrics (%)	
		Ha	Sa	An	Di	Su	Fe	Ne	Nor	Over	Low	HDR	WAR $\uparrow$	UAR $\uparrow$
Resnet18 + LSTM	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
Resnet50 + GRU	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
3D Resnet18 [Hara <i>et al.</i> , 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
R(2+1)D [Tran <i>et al.</i> , 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
Former DFER [Zhao and Liu, 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
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
Eyemotion [Hickson <i>et al.</i> , 2019]	Eye	74.3	85.5	79.5	74.3	69.1	79.2	<u>94.5</u>	79.0	81.8	81.5	72.5	78.8	79.5
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
EMO [Wu <i>et al.</i> , 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
EMO 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
SEEN [Zhang <i>et al.</i> , 2023]	Eye	85.0	89.9	<u>92.2</u>	76.7	72.1	87.7	85.2	<u>83.3</u>	<u>85.6</u>	80.8	84.8	83.6	84.1
MSKD (Ours) <i>teacher network</i>	Eye	85.6	91.7	92.3	<u>79.0</u>	79.4	<u>88.0</u>	90.3	84.4	89.1	88.3	82.7	86.2	86.6
MSKD (Ours) <i>student network</i>	Eye	<u>83.2</u>	<u>90.6</u>	89.8	79.3	<u>72.2</u>	89.4	86.3	<u>83.3</u>	<u>85.6</u>	<u>83.9</u>	<u>83.0</u>	<u>84.0</u>	<u>84.4</u>

Table 2: Quantitative comparison against state-of-the-arts. All methods are retrained and tested on the SEE dataset. The abbreviations are defined as Ha $\rightarrow$ Happiness; Sa $\rightarrow$ Sadness; An $\rightarrow$ Anger; Di $\rightarrow$ Disgust; Su $\rightarrow$ Surprise; Fe $\rightarrow$ Fear; Ne $\rightarrow$ Neutrality; Nor $\rightarrow$ Normal; Over $\rightarrow$ Overexposure; Low $\rightarrow$ Low-Light. The first and second best results are highlighted in **bold** and underline, respectively.

modality-scarce scenarios. We also validate the effectiveness of our method on our newly constructed DSEE dataset. Our proposed method also demonstrates competitive performance on our DSEE dataset, as evidenced by the top two rankings of our teacher and student networks in Table 3. Notably, our student network surpasses SEEN [Zhang *et al.*, 2023] by a large margin in both WAR (3.4%) and UAR (3.1%). This superiority remains consistent across diverse emotion categories and varied lighting conditions, demonstrating the robust capability of our approach for emotion recognition.

5.3 Efficiency Analysis

The objective of this work is to develop an effective yet lightweight single-eye emotion recognition method to enhance its practical usability. Having established the effectiveness of our single-eye emotion recognition method, this section delves into a meticulous analysis of its efficacy along three key dimensions: computational complexity, inference speed, and energy cost. Table 4 shows that among the existing eye-based emotion recognition methods, Eyemotion [Hickson *et al.*, 2019] outperforms EMO [Wu *et al.*, 2020] and SEEN [Zhang *et al.*, 2023] in emotion recognition accuracy but lags in efficiency. Benefiting from the utilization of SNNs and weight-copy scheme, ANN-SNN-hybrid SEEN [Zhang *et al.*, 2023] achieves comparable recognition accuracy as ANN-based Eyemotion [Hickson *et al.*, 2019] in a more efficient way. Our method takes advantage of multimodality synergistic knowledge distillation, achieving the best recognition performance and better computational efficiency than SEEN [Zhang *et al.*, 2023]. In addition, we conduct the energy cost comparison in Table 5. Energy estimations are predicated on [Horowitz, 2014]’s examination of 45 nm CMOS technology, as adopted in [Rathi and Roy, 2021; Li *et al.*, 2021], in which SNN addition operations cost 0.9 *pJ* whereas ANN MAC operations demand 4.6 *pJ*. From the results we can draw the conclusion that our method is much more energy efficient than existing methods, *e.g.*, 108.52 and 10.76 times more energy efficient than Eyemotion [Hickson

et al., 2019] and SEEN [Zhang *et al.*, 2023], respectively.

5.4 Ablation Study

Effectiveness of Hit / Temporal Consistency Knowledge Distillation. Table 6 reports the results on the DSEE testing set when using different loss functions to train the student network. It can be seen that: (i) the temporal efficient training (TET) approach [Deng *et al.*, 2022] can benefit the training of SNN-based student network, *i.e.*, *b* is better than *a*; (ii) both hit and temporal consistency knowledge distillation can help improve the performance (*c* and *d* perform better than *b*); and (iii) the complementary nature of hit and temporal consistency knowledge distillation is evident in their synergistic interaction, which leads to demonstrably superior model performance (*e* is better than *c* and *d*).

Effects of different event data configurations. We explored multiple ways to utilize event data as input for training our teacher network, aiming to determine the most effective approach. To enable recognition during any phase of the emotion, we adopted [Zhang *et al.*, 2023]’s approach of using a uniformly distributed random starting point and corresponding sequence length for testing. Specifically, a start point is selected to ensure that the rest sequence is longer than the testing length which is defined as the total accumulation time of all included event frames, *x*, and a skip time between two adjacent event frames where all events are ignored, *y*, denoted as *ExSy*. We express both accumulation time and skip time as multiples of 1/30 s. This yields a testing length of *ExSy* corresponding to $(x + (x - 1) \times y)/30$ s. From Table 7(A-F, M), we observe that: (i) with four input event frames, longer skip time yields better performance (*i.e.*, *A*, *B*, *C*, and *M* gets better in order). This shows that a longer testing length is beneficial as it contains more temporal information; and (ii) benefited from more temporal cues, with the same $3 \times (1/30)$ s skip time, more event frames generate improved performance (*D*, *E*, *F*, and *M* gets better in order).

Influence of loss function in MSKD. There are multiple choices when measuring the distance between distributions

Methods	input region	Accuracy for different emotions							Accuracy for different lightings				Metrics (%)	
		Ha	Sa	An	Di	Su	Fe	Ne	Nor	Over	Low	HDR	WAR $\uparrow$	UAR $\uparrow$
Resnet18 + LSTM	Face	72.3	61.9	78.3	76.8	69.9	66.9	84.8	71.3	73.1	69.7	79.6	72.2	73.0
Resnet50 + GRU	Face	66.3	62.2	78.3	75.8	68.5	65.7	84.7	69.9	70.5	69.3	78.4	71.6	66.3
3D Resnet18 [Hara <i>et al.</i> , 2018]	Face	62.7	56.5	58.9	50.9	51.5	33.5	62.6	54.9	55.6	44.2	56.5	53.3	53.8
R(2+1)D [Tran <i>et al.</i> , 2018]	Face	52.1	53.5	27.1	62.0	54.7	29.8	35.6	49.1	45.9	36.6	44.4	45.8	45.0
Former DFER [Zhao and Liu, 2021]	Face	71.1	54.8	67.2	64.3	45.2	42.7	82.6	58.1	64.2	60.7	58.8	59.7	61.1
Former DFER w/o pre-train	Face	63.1	50.7	49.9	48.1	40.8	42.0	62.8	50.3	50.7	47.3	53.5	50.2	51.1
Eyemotion [Hickson <i>et al.</i> , 2019]	Eye	72.3	66.4	76.9	74.9	70.4	64.9	85.8	71.3	72.1	69.3	82.0	72.3	73.1
Eyemotion w/o pre-train	Eye	71.0	68.2	75.0	74.8	66.8	67.2	86.3	70.4	70.4	70.1	83.8	71.8	72.7
EMO [Wu <i>et al.</i> , 2020]	Eye	73.7	59.2	70.8	74.2	61.6	61.7	80.2	67.0	68.2	65.5	76.5	68.0	68.8
EMO w/o pre-train	Eye	68.3	64.4	70.7	73.7	62.9	58.6	82.5	67.4	67.0	63.2	78.6	67.8	68.7
SEEN [Zhang <i>et al.</i> , 2023]	Eye	72.2	65.3	78.5	74.9	72.0	61.0	84.2	70.9	74.8	69.8	75.5	71.9	72.6
MSKD (Ours) <i>teacher network</i>	Eye	80.1	73.7	83.8	79.3	75.4	66.8	86.3	76.1	78.6	75.0	86.2	77.4	77.9
MSKD (Ours) <i>student network</i>	Eye	<u>73.8</u>	<u>71.9</u>	<u>80.2</u>	80.3	<u>73.7</u>	66.9	82.8	<u>73.7</u>	<u>76.2</u>	<u>72.4</u>	<u>85.9</u>	<u>75.3</u>	<u>75.7</u>

Table 3: Quantitative comparison against state-of-the-arts. All methods are retrained and tested on the **DSEE** dataset. The abbreviations are defined as Ha $\rightarrow$ Happiness; Sa $\rightarrow$ Sadness; An $\rightarrow$ Anger; Di $\rightarrow$ Disgust; Su $\rightarrow$ Surprise; Fe $\rightarrow$ Fear; Ne $\rightarrow$ Neutrality; Nor $\rightarrow$ Normal; Over $\rightarrow$ Overexposure; Low $\rightarrow$ Low-Light. The first and second best results are highlighted in **bold** and underline, respectively.

Methods	WAR	FLOPs (G)	Params (M)	Time (ms)
Eyemotion	72.3	5.73	25.13	17.5
EMO	68.0	0.32	1.68	7.1
SEEN	71.9	0.95	6.08	7.2
Ours	75.3	0.27	4.04	6.1

Table 4: Computational efficiency comparison of different eye-based emotion recognition methods.

Methods	Eyemotion	EMO	SEEN	MSKD (Ours)
Energy (<i>mJ</i>)	823.69	46.00	81.64	7.59
Multiple to Ours	108.52	6.06	10.76	1.00

Table 5: Estimated energy efficiency comparison of different eye-based emotion recognition methods.

Training Losses		WAR $\uparrow$	UAR $\uparrow$
<i>a</i>	$\mathcal{L}_{\text{Cls}}$ w/o TET [Deng <i>et al.</i> , 2022]	70.77	71.37
<i>b</i>	$\mathcal{L}_{\text{Cls}}$	71.90	72.57
<i>c</i>	$\mathcal{L}_{\text{Cls}} + \mathcal{L}_{\text{HCKD}}$	73.83	74.15
<i>d</i>	$\mathcal{L}_{\text{Cls}} + \mathcal{L}_{\text{TCKD}}$	74.42	74.88
<i>e</i>	$\mathcal{L}_{\text{Cls}} + \mathcal{L}_{\text{HCKD}} + \mathcal{L}_{\text{TCKD}}$	75.27	75.66

Table 6: Ablation study on synergistic knowledge distillation.

in the knowledge distillation process, such as cross-entropy (CE), mean squared error (MSE), and Kullback-Leibler divergence (KLD). As can be seen from Table 7 (*I-M*), replacing MSE with CE or KLD in HCKD or TCKD leads to performance degradation (*i.e.*, *I-L* are lower than *M*). This indicates that MSE is a better choice for our synergistic knowledge distillation. Besides, removing the TET [Deng *et al.*, 2022] in classification loss $\mathcal{L}_{\text{Cls}}$ or temporal consistency knowledge distillation loss $\mathcal{L}_{\text{TCKD}}$ deteriorates the performance (*G* and *H*

Networks		WAR $\uparrow$	UAR $\uparrow$
<i>A</i>	<i>E4S0</i>	67.00	68.05
<i>B</i>	<i>E4S1</i>	72.21	72.85
<i>C</i>	<i>E4S2</i>	73.62	74.17
<i>D</i>	<i>E1S3</i>	62.83	63.58
<i>E</i>	<i>E2S3</i>	68.92	69.67
<i>F</i>	<i>E3S3</i>	72.22	72.77
<i>G</i>	No TET for $\mathcal{L}_{\text{Cls}}$	74.79	75.23
<i>H</i>	No TET for $\mathcal{L}_{\text{TCKD}}$	74.32	74.77
<i>I</i>	$\mathcal{L}_{\text{HCKD}}$: MSE $\rightarrow$ CE	74.91	75.35
<i>J</i>	$\mathcal{L}_{\text{HCKD}}$: MSE $\rightarrow$ KLD	73.87	74.34
<i>K</i>	$\mathcal{L}_{\text{TCKD}}$: MSE $\rightarrow$ CE	74.31	74.71
<i>L</i>	$\mathcal{L}_{\text{TCKD}}$: MSE $\rightarrow$ KLD	73.94	74.42
<i>M</i>	Ours (<i>E4S3</i> ; w/ TET; MSE)	75.27	75.66

Table 7: Quantitative ablation results indicate that each key component in our MSKD contributes to the overall performance.

are worse than *M*), showing the effectiveness of the temporal efficient training strategy.

6 Conclusion

In summary, our research introduces a pioneering approach to enhance single-eye emotion recognition in source-limited wearable devices. Drawing inspiration from apprenticeship learning, our multimodality synergistic knowledge distillation mechanism empowers a lightweight spiking neural network for more effective yet efficient recognition. Extensive validation demonstrates its significant improvement over state-of-the-art methods in both accuracy and efficiency. Furthermore, the establishment of a diverse single-eye emotion benchmark not only validates our method but also lays the foundation for continued exploration and innovation in the realm of single-eye-based methodologies. This work signifies a notable advancement in practical usability and user experience in the context of emotion recognition, opening new avenues for innovation and exploration in this dynamic field.

Acknowledgements

This work was supported in part by National Key Research and Development Program of China (2022ZD0210500), the National Natural Science Foundation of China under Grants 62332019/U21A20491, and the Distinguished Young Scholars Funding of Dalian (No. 2022RJ01). Yang Wang was supported by a Chinese Scholarship Council (CSC) grant. This work is also supported by IAF, A*STAR, SOITEC, NXP, National University of Singapore under FD-fAbrICS: Joint Lab for FD-SOI Always-on Intelligent & Connected Systems (Award I2001E0053). Haiyang Mei and Mike Zheng Shou did not receive any funding for this work.

References

- [Aifanti *et al.*, 2010] Niki Aifanti, Christos Papachristou, and Anastasios Delopoulos. The mug facial expression database. In *WIAMIS*, pages 1–4, 2010.
- [Delbruck *et al.*, 2023] Tobi Delbruck, Zuowen Wang, Haiyang Mei, Germain Haessig, Damien Joubert, Justin Haque, Yingkai Chen, Moritz B. Milde, and Viktor Gruev. Live demo: E2p-events to polarization reconstruction from pdavis events. In *CVPRW*, pages 3973–3975, 2023.
- [Deng *et al.*, 2020] Didan Deng, Zhaokang Chen, Yuqian Zhou, and Bertram Shi. Mimamo net: Integrating micro- and macro-motion for video emotion recognition. In *AAAI*, volume 34, pages 2621–2628, 2020.
- [Deng *et al.*, 2022] Shikuang Deng, Yuhang Li, Shanghang Zhang, and Shi Gu. Temporal efficient training of spiking neural network via gradient re-weighting. *arXiv:2202.11946*, 2022.
- [Ding *et al.*, 2022] Jianchuan Ding, Bo Dong, Felix Heide, Yufei Ding, Yunduo Zhou, Baocai Yin, and Xin Yang. Biologically inspired dynamic thresholds for spiking neural networks. In *NeurIPS*, pages 6090–6103, 2022.
- [Dong *et al.*, 2023] Yiting Dong, Dongcheng Zhao, and Yi Zeng. Temporal knowledge sharing enable spiking neural network learning from past and future. *arXiv:2304.06540*, 2023.
- [Fang *et al.*, 2021] Wei Fang, Zhaofei Yu, Yanqi Chen, Tiejun Huang, Timothée Masquelier, and Yonghong Tian. Deep residual learning in spiking neural networks. *NeurIPS*, 34:21056–21069, 2021.
- [Georgescu and Ionescu, 2019] Mariana-Iuliana Georgescu and Radu Tudor Ionescu. Recognizing facial expressions of occluded faces using convolutional neural networks. In *ICONIP*, volume 1142, pages 645–653, 2019.
- [Hara *et al.*, 2018] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *CVPR*, pages 6546–6555, 2018.
- [He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.
- [Hickson *et al.*, 2019] Steven Hickson, Nick Dufour, Avneesh Sud, Vivek Kwatra, and Irfan Essa. Eyemotion: Classifying facial expressions in vr using eye-tracking cameras. In *WACV*, pages 1626–1635, 2019.
- [Horowitz, 2014] Mark Horowitz. 1.1 computing’s energy problem (and what we can do about it). In *ISSCC*, pages 10–14, 2014.
- [Houshmand and Khan, 2020] Bitan Houshmand and Naimul Mefraz Khan. Facial expression recognition under partial occlusion from virtual reality headsets based on transfer learning. In *IEEE Sixth International Conference on Multimedia Big Data*, 2020.
- [Hu *et al.*, 2021] Yuhuang Hu, Shih-Chii Liu, and Tobi Delbruck. v2e: From video frames to realistic dvs events. In *CVPR*, pages 1312–1321, 2021.
- [Ji *et al.*, 2023] Zhong Ji, Jingwei Ni, Xiyao Liu, and Yanwei Pang. Teachers cooperation: team-knowledge distillation for multiple cross-domain few-shot learning. *Frontiers of Computer Science*, 17(2):172312, 2023.
- [Kushawaha *et al.*, 2021] Ravi Kumar Kushawaha, Saurabh Kumar, Biplab Banerjee, and Rajbabu Velmurugan. Distilling spikes: Knowledge distillation in spiking neural networks. In *ICPR*, pages 4536–4543, 2021.
- [Li *et al.*, 2021] Yuhang Li, Yufei Guo, Shanghang Zhang, Shikuang Deng, Yongqing Hai, and Shi Gu. Differentiable spike: Rethinking gradient-descent for training spiking neural networks. In *NeurIPS*, volume 34, pages 23426–23439, 2021.
- [Liu *et al.*, 2020a] Qianhui Liu, Gang Pan, Haibo Ruan, Dong Xing, Qi Xu, and Huajin Tang. Unsupervised aer object recognition based on multiscale spatio-temporal features and spiking neurons. *IEEE TNNLS*, 31(12):5300–5311, 2020.
- [Liu *et al.*, 2020b] Qianhui Liu, Haibo Ruan, Dong Xing, Huajin Tang, and Gang Pan. Effective aer object classification using segmented probability-maximization learning in spiking neural networks. In *AAAI*, volume 34, pages 1308–1315, 2020.
- [Lucey *et al.*, 2010] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *CVPRW*, pages 94–101, 2010.
- [Mei *et al.*, 2023] Haiyang Mei, Zuowen Wang, Xin Yang, Xiaopeng Wei, and Tobi Delbruck. Deep polarization reconstruction with pdavis events. In *CVPR*, pages 22149–22158, 2023.
- [Pantic *et al.*, 2005] Maja Pantic, Michel Valstar, Ron Rade-maker, and Ludo Maat. Web-based database for facial expression analysis. In *ICME*, pages 5–pp, 2005.
- [Paszke *et al.*, 2019] 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. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, pages 8024–8035, 2019.
- [Picard, 2003] Rosalind W Picard. Affective computing: challenges. *International Journal of Human-Computer Studies*, 59(1-2):55–64, 2003.
- [Rathi and Roy, 2021] Nitin Rathi and Kaushik Roy. Diet-snn: A low-latency spiking neural network with direct input encoding and leakage and threshold optimization. *IEEE TNNLS*, 2021.
- [Rebecq et al., 2019] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. High speed and high dynamic range video with an event camera. *IEEE TPAMI*, 43(6):1964–1980, 2019.
- [Ruan et al., 2021] Delian Ruan, Yan Yan, Shenqi Lai, Zhenhua Chai, Chunhua Shen, and Hanzhi Wang. Feature decomposition and reconstruction learning for effective facial expression recognition. In *CVPR*, pages 7660–7669, 2021.
- [Sanchez et al., 2021] Enrique Sanchez, Mani Kumar Telamekala, Michel Valstar, and Georgios Tzimiropoulos. Affective processes: stochastic modelling of temporal context for emotion and facial expression recognition. In *CVPR*, pages 9074–9084, 2021.
- [Schuller et al., 2011] B. Schuller, B. Vlasenko, F. Eyben, M. Woßlmer, A. Stuhlsatz, A. Wendemuth, and G. Rigoll. Cross-corpus acoustic emotion recognition: Variances and strategies. *IEEE Transactions on Affective Computing*, 1(2):119–131, 2011.
- [Takuya et al., 2021] Sugahara Takuya, Renyuan Zhang, and Yasuhiko Nakashima. Training low-latency spiking neural network through knowledge distillation. In *IEEE Symposium in Low-Power and High-Speed Chips*, pages 1–3, 2021.
- [Tran et al., 2018] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *CVPR*, pages 6450–6459, 2018.
- [Van Der Schalk et al., 2011] Job Van Der Schalk, Skyler T Hawk, Agneta H Fischer, and Bertjan Doosje. Moving faces, looking places: validation of the amsterdam dynamic facial expression set (adfs). *Emotion*, 11(4):907, 2011.
- [Wang et al., 2023] Yang Wang, Bo Dong, Yuji Zhang, Yunduo Zhou, Haiyang Mei, Ziqi Wei, and Xin Yang. Event-enhanced multi-modal spiking neural network for dynamic obstacle avoidance. In *ACM MM*, pages 3138–3148, 2023.
- [Wei et al., 2024] Wenjie Wei, Malu Zhang, Jilin Zhang, Ammar Belatreche, Jibin Wu, Zijong Xu, Xuerui Qiu, Hong Chen, Yang Yang, and Haizhou Li. Event-driven learning for spiking neural networks. *arXiv preprint arXiv:2403.00270*, 2024.
- [Wu et al., 2020] Hao Wu, Jinghao Feng, Xuejin Tian, Edward Sun, Yunxin Liu, Bo Dong, Fengyuan Xu, and Sheng Zhong. 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*, pages 448–461, 2020.
- [Xu et al., 2023a] Qi Xu, Yaxin Li, Xuanye Fang, Jianrong Shen, Jian K Liu, Huajin Tang, and Gang Pan. Biologically inspired structure learning with reverse knowledge distillation for spiking neural networks. *arXiv:2304.09500*, 2023.
- [Xu et al., 2023b] Qi Xu, Yaxin Li, Jianrong Shen, Jian K Liu, Huajin Tang, and Gang Pan. Constructing deep spiking neural networks from artificial neural networks with knowledge distillation. In *CVPR*, pages 7886–7895, 2023.
- [Xue et al., 2021] Fanglei Xue, Qiangchang Wang, and Guodong Guo. Transfer: Learning relation-aware facial expression representations with transformers. In *CVPR*, pages 3601–3610, 2021.
- [Zhang et al., 2016] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10):1499–1503, 2016.
- [Zhang et al., 2021a] Malu Zhang, Jiadong Wang, Jibin Wu, Ammar Belatreche, Burin Amornpaisannon, Zhixuan Zhang, Venkata Pavan Kumar Miriyala, Hong Qu, Yansong Chua, Trevor E Carlson, et al. Rectified linear post-synaptic potential function for backpropagation in deep spiking neural networks. *IEEE transactions on neural networks and learning systems*, 33(5):1947–1958, 2021.
- [Zhang et al., 2021b] Yuhang Zhang, Chengrui Wang, and Weihong Deng. Relative uncertainty learning for facial expression recognition. *NeurIPS*, 34:17616–17627, 2021.
- [Zhang et al., 2022] Jiqing Zhang, Bo Dong, Haiwei Zhang, Jianchuan Ding, Felix Heide, Baocai Yin, and Xin Yang. Spiking transformers for event-based single object tracking. In *CVPR*, pages 8801–8810, 2022.
- [Zhang et al., 2023] Haiwei Zhang, Jiqing Zhang, Bo Dong, Pieter Peers, Wenwei Wu, Xiaopeng Wei, Felix Heide, and Xin Yang. In the blink of an eye: Event-based emotion recognition. In *ACM SIGGRAPH*, pages 1–11, 2023.
- [Zhao and Liu, 2021] Zengqun Zhao and Qingshan Liu. Former-dfer: Dynamic facial expression recognition transformer. In *ACM MM*, pages 1553–1561, 10 2021.
- [Zhao et al., 2011] Guoying Zhao, Xiaohua Huang, Matti Taini, Stan Z Li, and Matti Pietikäinen. Facial expression recognition from near-infrared videos. *Image and vision computing*, 29(9):607–619, 2011.