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

Yang Wang¹,², Haiyang Mei²,¹, Qirui Bao¹, Ziqi Wei³, Mike Zheng Shou², Haizhou Li⁵,², Bo Dong¹ and Xin Yang¹*

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 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.
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 sufficiently 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 distill (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 distill loss items: (b) hit consistency loss and (c) temporal consistency loss.

The loss function is defined as:

$$L_{\text{total}} = (1 - \alpha) \cdot L_{\text{Cls}} + \alpha \cdot L_{\text{HCKD}} + \alpha \cdot L_{\text{TCKD}},$$

where $L_{\text{Cls}}$ is the classification loss to enforce the output at each timestep close to the true label; $L_{\text{HCKD}}$ and $L_{\text{TCKD}}$ 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 $\alpha$ 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:

$$L_{\text{Cls}} = \frac{1}{T} \sum_{t=1}^{T} L_{CE}(O_{\text{stu}}(t), y),$$

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

\[
L_{\text{HCKD}} = L_{\text{MSE}}(S_{\text{stu}}, S_{\text{tea}}),
\]

\[
S_{\text{stu}} = \frac{1}{C_{\text{stu}}} \sum_{c_{\text{stu}}=1}^{C_{\text{stu}}} O_{\text{stu}}(c_{\text{stu}}),
\]

\[
S_{\text{tea}} = \frac{1}{C_{\text{tea}}} \sum_{c_{\text{tea}}=1}^{C_{\text{tea}}} O_{\text{tea}}(c_{\text{tea}}),
\]

where \( L_{\text{MSE}} \) measures the mean squared error between student correctly-predicted signal \( S_{\text{stu}} \) and teacher correctly-predicted signal \( S_{\text{tea}} \). \( S_{\text{stu}} \) is obtained by averaging \( C_{\text{stu}} \) student/teacher prediction distributions \( O_{\text{stu}}(c_{\text{stu}}) \) at the correctly-predicted timestep \( c_{\text{stu}} \). When \( C_{\text{stu}} \) 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,

\[
L_{\text{TCKD}} = \frac{1}{T} \sum_{t=1}^{T} L_{\text{MSE}}(O_{\text{stu}}(t), O_{\text{tea}}(t)),
\]

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] 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., \( 128 \times 128 \)) 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

![Figure 3: Examples from our DSEE dataset.](image)

3163
Table 1: Comparison among event-based datasets for emotion recognition.

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

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,

\[
UAR = \frac{1}{N_c} \sum_{i=1}^{N_c} \frac{TP_i}{TP_i + FN_i},
\]

\[
WAR = \frac{TP + FN}{TP + TN + FP + FN},
\]

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
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 p.J whereas ANN MAC operations demand 4.6 p.J. 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) × 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 × (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
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 → 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.

<table>
<thead>
<tr>
<th>Methods</th>
<th>input region</th>
<th>Accuracy for different emotions</th>
<th>Accuracy for different lightings</th>
<th>Metrics (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Ha</td>
<td>Sa</td>
<td>An</td>
</tr>
<tr>
<td>Resnet18 + LSTM</td>
<td>Face</td>
<td>72.3</td>
<td>61.9</td>
<td>78.3</td>
</tr>
<tr>
<td>Resnet50 + GRU</td>
<td>Face</td>
<td>66.3</td>
<td>62.2</td>
<td>78.3</td>
</tr>
<tr>
<td>3D Resnet18 [Hara et al., 2018]</td>
<td>Face</td>
<td>62.7</td>
<td>56.5</td>
<td>58.9</td>
</tr>
<tr>
<td>R(2+1)D [Tran et al., 2018]</td>
<td>Face</td>
<td>52.1</td>
<td>53.5</td>
<td>27.1</td>
</tr>
<tr>
<td>Former DFER [Zhao and Liu, 2021]</td>
<td>Face</td>
<td>71.1</td>
<td>54.8</td>
<td>67.2</td>
</tr>
<tr>
<td>Former DFER w/o pre-train</td>
<td>Face</td>
<td>63.1</td>
<td>50.7</td>
<td>49.9</td>
</tr>
<tr>
<td>Eyemotion [Hickson et al., 2019]</td>
<td>Eye</td>
<td>72.3</td>
<td>66.4</td>
<td>76.9</td>
</tr>
<tr>
<td>Eyemotion w/o pre-train</td>
<td>Eye</td>
<td>71.0</td>
<td>68.2</td>
<td>75.0</td>
</tr>
<tr>
<td>EMO [Wu et al., 2020]</td>
<td>Eye</td>
<td>73.7</td>
<td>59.2</td>
<td>70.8</td>
</tr>
<tr>
<td>EMO w/o pre-train</td>
<td>Eye</td>
<td>68.3</td>
<td>64.4</td>
<td>70.7</td>
</tr>
<tr>
<td>SEEN [Zhang et al., 2023]</td>
<td>Eye</td>
<td>72.2</td>
<td>65.3</td>
<td>78.5</td>
</tr>
<tr>
<td>MSKD (Ours) teacher network</td>
<td>Eye</td>
<td><strong>80.1</strong>*</td>
<td><strong>73.7</strong>*</td>
<td><strong>83.8</strong>*</td>
</tr>
<tr>
<td>MSKD (Ours) student network</td>
<td>Eye</td>
<td><strong>73.8</strong>*</td>
<td><strong>71.9</strong>*</td>
<td><strong>80.2</strong>*</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Methods</th>
<th>Eyemotion</th>
<th>EMO</th>
<th>SEEN</th>
<th>MSKD (Ours)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Energy (J)</td>
<td>823.69</td>
<td>46.00</td>
<td>81.64</td>
<td><strong>7.59</strong></td>
</tr>
<tr>
<td>Multiple to Ours</td>
<td></td>
<td>108.52</td>
<td>6.06</td>
<td>10.76</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Training Losses</th>
<th>WAR ↑</th>
<th>UAR ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a$ $L_{cls}$ w/o TET [Deng et al., 2022]</td>
<td>70.77</td>
<td>71.37</td>
</tr>
<tr>
<td>$b$ $L_{cls}$</td>
<td>71.90</td>
<td>72.57</td>
</tr>
<tr>
<td>$c$ $L_{cls} + L_{CE}$</td>
<td>73.83</td>
<td>74.15</td>
</tr>
<tr>
<td>$d$ $L_{cls} + L_{TCKD}$</td>
<td>74.42</td>
<td>74.88</td>
</tr>
<tr>
<td>$e$ $L_{cls} + L_{CE} + L_{TCKD}$</td>
<td><strong>75.27</strong>*</td>
<td><strong>75.66</strong>*</td>
</tr>
</tbody>
</table>

Table 6: Ablation study on synergistic knowledge distillation.

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

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 $L_{cls}$ or temporal consistency knowledge distillation loss $L_{TCKD}$ deteriorates the performance ($G$ and $H$ 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 (2022ZD02010500), 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-FabriICS: 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


