Proceedings of the 5th International Conference on Computing Innovation and Applied Physics
DOI: 10.54254/2753-8818/2026.31464

# *HEC-Net: Hierarchical Event-RGB Cross-Modal Fusion for Single-Eye Emotion Recognition*

**Sichen Shao[1], Bohan Liang[2], Jie Zhang[1], Junjie Cao[1*]**

[1]*School of Mathematical Sciences, Dalian University of Technology, Dalian, China*
[2]*School of Electronic and Information Engineering, Liaoning Technical University, Huludao, China*
**Corresponding Author. Email: jjcao@dlut.edu.cn*

***Abstract.*** Recognizing emotions from a single eye has been a hard work over time, mainly because muscle movements are hard to perceive and event data is susceptible to noise. Most existing methods concentrated on semantic fusion, neglecting the potential benefits of local interaction between different modalities. Considering this problem, we design a new framework called HEC-Net which hierarchically fuses different modalities' information. Our HEC-Net begins with a dual-stream Spatio-Temporal Feature Extraction (STFE) module to encode texture and motion while suppressing noise via Top-k selection, which follows a "perceive and select" design philosophy. Then we form a three-stage local-to-global fusion process to fuse the information in a more layered approach. The whole approach includes a multi-window fusion, a pyramid structure and a global fusion. The multi-window fusion constrains interactions locally, enabling the model to firstly concentrate on details. We employ a pyramid structure to capture the changes in blinking actions over time. Finally, the global fusion process aggregates dispersed cues into a coherent global representation. HEC-Net achieves a state-of-the-art UAR of 92.30% on the SEE dataset, which remains stable and efficient under various lighting conditions.

***Keywords:*** single-eye expression recognition, event-based vision, multimodal fusion, temporal shift module, hierarchical alignment

## 1. Introduction

Facial expression recognition (FER) is an important task for its irreplaceability in VR application. While RGB-based methods [1,2] perform well in ideal environment, their performance gets worse under extreme lighting, forcing us to find a more robust way. Event cameras offer a solution with High Dynamic Range (HDR) and microsecond temporal resolution [3]. However, effectively using different modalities' information is not easy. Existing methods like SEEN [4] and HI-Net [5] depend on straightforward concatenation or heavy recursive blocks that prioritize global integration to get better performance. They neglect the significant potential of local interactions to form deeper relations between asynchronous event transients and synchronous RGB textures. These missing components lead to the loss of critical micro-expression cues getting covered in background noise.

Considering those problems, we propose HEC-Net. Inspired by the recent research on the image recognition [6] and micro-expression analysis [7], we design a Spatio-Temporal Feature Extraction

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-8818/2026.31464

(STFE) module which combines TSM [8] (Temporal Shift Module) and a Top-k selection to extract information from the environment's noise. We secondly form a local-to-global hierarchical fusion strategy which includes three stages: (1) Multi-window fusion which enforces correspondence in a pixel level; (2) Temporal alignments which capture multiscale blinking actions in a pyramid structure; and; (3) Global fusion which synthesizes global representations via transformer at the semantic level. HEC-Net's design embodies a multi-level hierarchical architecture. Our contributions include:

• We design a STFE module that combines TSM and Top-k selection for denoised extraction.

• We propose a hierarchical fusion strategy that progressively resolves spatial misalignment, synchronizes multi-scale temporal dynamics, and fuses global semantics.

• Results on the SEE dataset have achieved SOTA performance (92.30% Unweighted Average Recall, UAR).

## 2. Related work

Early works on single-eye emotion recognition are mainly based on RGB-based methods [1, 2] which largely depended on traditional CNN architectures like ResNet. However, those RGB-based models are unable to maintain accuracy in challenging lighting environments. To cope with it, event cameras [3] have come out as a powerful tool in high-dynamic-range environments which bring out works like SEEN [4] which focus on cross-modal fusion, whereas HI-Net [5] employed recursive fusion that incurs high computational overhead. Unlike previous work, we prioritize efficiency via hierarchical alignment, which is conceptually similar to recent Transformer advancements like Swin [9] and focal self-attention [10].

## 3. Methodology

As we illustrated in Figure 1, HEC-Net is a dual-stream framework which can be divided into two primary parts. The whole process begins with the Spatio-Temporal Feature Extraction (STFE) module, which applies a "perceive-and-select" mechanism to get clean texture and motion signals from noise. The architecture subsequently follows a progressive fusion process. We firstly form spatial fusion to get deep correlations via multi-window cross-attention, which is followed by a pyramid structure model to synchronize cross-frame dynamics. Finally, a Global fusion block consolidates these cues in a semantic level to ensure robust global classification.

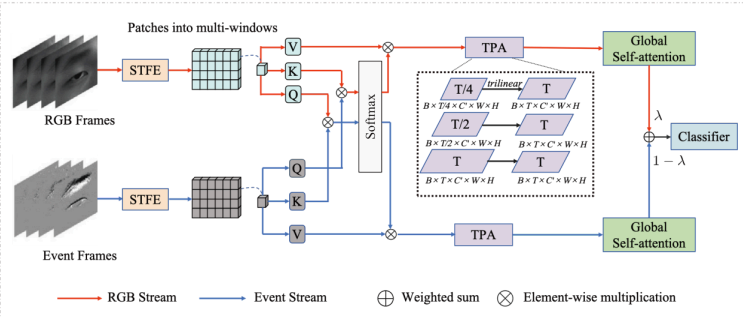


Figure 1. The overall architecture of the HEC-Net

DOI: 10.54254/2753-8818/2026.31464

### 3.1. Spatio-Temporal Feature Extraction (STFE)

The transience of micro-expressions brings a dilemma: while we try to avoid standard 2D CNNs due to rapid changes, full 3D CNNs remain too computationally expensive for edge applications. Furthermore, while event sensors offer high temporal resolution, they meet a distinct challenge that artifacts like hot pixels often simulate real movement, creating false positives that confuse standard motion encoders. To cope with this problem, we introduce the STFE module which emulates the human visual system to filter complex information in daily environment. Our module is able to actively capture salient muscle dynamics from noise in a "perceive-and-select" way.

#### 3.1.1. Temporal modeling and multi-scale perception

Temporal Modeling: To capture rapid micro-movements efficiently, we combine the Temporal Shift Module (TSM) [8] into the 2D backbone. TSM moves a small portion of the channel along the time dimension to facilitate inter-frame information exchange, enabling the encoder to simulate the 3D time receptive domain without additional parameters. This process aggregates temporal contexts from adjacent frames, effectively enriching the static frame-level representations with dynamic evolution cues. Multi-scale Perception: To adapt to varying eye deformation scales, we employ parallel dilated convolutions with different ratesd$\in\{1,2,4\}$. Not only can this exponentially expand the spatial receptive field, but it also generates a comprehensive feature library that serves as the Motion Query (Q) and Context Keys (K) in the subsequent selection stage.

#### 3.1.2. Dual-branch Top-k attentive selection

Standard softmax inevitably assigns weights to background noise. To avoid this, we propose a hard-selection strategy. We firstly compute the relevance matrix M between Motion Query (Q) and Context Keys (K) and retain only the top-k significant values. Then we employ two parallel branches: a Salient Branch ($k_1$) focusing on focal features and a Complementary Branch ($k_2$) preserving broader context,

$$M_k = softmax(Top - K(M, k)) \tag{1}$$

where $M_k$ denotes the attention weights after selecting the top-k values from the similarity matrix M for each row. The final feature is generated by fusing these branches via learnable parameters α and β applied to Value (V):

$$Attn = (a \times M_{K_1} + \beta \times M_{K_2}) \times V \tag{2}$$

where this mechanism effectively suppresses background noise while ensuring feature robustness.

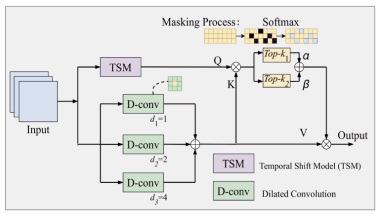


Figure 2. The detailed structure of the Spatio-Temporal Feature Extraction (STFE) module

DOI: 10.54254/2753-8818/2026.31464

### 3.2. Hierarchical spatio-temporal fusion

Due to the fundamental differences between different modalities, misalignment in both spatial resolution and temporal rates often exist between RGB frames (appearance) and event streams (motion).A simple concatenation is insufficient to bridge this gap. Therefore, we propose a local-to-global hierarchical fusion strategy which progressively builds dense correlations, captures cross-frame blinking motions, and aggregates dispersed cues into a coherent representation.

#### 3.2.1. Stage 1: multi-window spatial fusion

To efficiently build deep correlations between two modalities, we design Window-based Cross-Attention. We partition features into patch-level windows (e.g., 8×8). In each window, RGB features serve as the structural anchor (Query Q) to align Event features (Key K, Value V). Symmetrically, Event features serve as the temporal anchor (Query) to guide RGB features, ensuring bi-directional alignment. The calibrated feature is computed as

$$F_x' = softmax(\frac{Q_y K_x^T}{\sqrt{d}} V_x + F_X,\ \forall (x,y) \in (r,e),(e,r) \tag{3}$$

where x denotes the modality of the Query ($x \in \{r,e\}$), and r and e represent RGB and Event, respectively. This anchors sparse motion signals to precise facial landmarks.

#### 3.2.2. Stage 2: temporal alignment and global semantic

After resolving spatial misalignment in Stage 1, the features still exhibit disparate motion frequencies. To bridge this temporal gap and synthesize the final decision, we propose a unified strategy that first synchronizes multi-scale blinking actions and then integrates global semantics. Micro-expressions manifest at varying temporal scales, ranging from rapid, high-frequency flickers (e.g., blinks) to sustained, low-frequency movements (e.g., frowns). To capture this diversity, we devise Temporal Pyramid Alignment (TPA) using a “Split-Align-Fuse” strategy. We first construct a temporal pyramid by down-sampling the input feature $F_{in}$into varying resolutions t= {T, T/2, T/4}. Crucially, to avoid artifacts caused by standard upsampling, we employ trilinear interpolation to strictly align all low-resolution pyramid levels back to the original dimension T. The synchronized feature is obtained via adaptive fusion:

$$F_{tpa} = \sum_{l=0}^{2} \omega_l \cdot Interp_{trilinear}(P_l) \tag{4}$$

where $\omega_l$ denotes the learnable weights for each scale, ensuring the network adaptively emphasizes the most informative motion frequency. With spatio-temporally aligned features, we proceed to synthesize high-level semantics. A transformer block is deployed to model long-range dependencies across the entire sequence. To optimally balance the modalities under varying lighting conditions, a cross-modal gate dynamically fuses the streams using a learnable coefficient λ:

$$F_{global} = \lambda \bigodot F_r^{final} + (1-\lambda) \bigodot F_e^{final} \tag{5}$$

where λ is a learnable parameter balancing the RGB and Event streams. Finally, the fused representation $F_{global}$ is fed into a Classifier consisting of fully connected layers to predict the probability of emotion classes. The entire network is optimized using standard Cross-Entropy Loss.

DOI: 10.54254/2753-8818/2026.31464

# 4. Experiment

## 4.1. Setup and implementation

The network is optimized via standard Cross-Entropy Loss. Performance is reported using Weighted Average Recall (WAR) and Unweighted Average Recall (UAR).We evaluate HEC-Net on the SEE Dataset [4] which is the only authoritative dataset. It comprises 128,712 samples under diverse lighting conditions. Following the standard protocol, we utilize 1,638 sequences for training and 767 for testing. The model is implemented in PyTorch on an NVIDIA RTX 3090 GPU. All inputs are resized to 224×224. We train for 150 epochs with a batch size of 16 using the AdamW optimizer (initial lr =1×10-4, weight decay=1×10-4) and a Cosine Annealing scheduler. The network is optimized via standard Cross-Entropy Loss. Our performance is reported using Weighted Average Recall (WAR) and Unweighted Average Recall (UAR).

## 4.2. Comparative analysis SOTA performance

As shown in Table 1, RGB baselines (e.g., EMO [1]) degrade notably in low-light (61.8%→60.1%) due to texture reliance. Hybrid methods like SEEN [4] rely on simple concatenation, failing to address spatial misalignment. In contrast, HEC-Net establishes a new benchmark with a 92.30% UAR, outperforming the best baseline HI-Net [5] by +4.6%. This validates that our hierarchical fusion aligns features better than simple concatenation. HEC-Net also shows exceptional stability. While achieving peak accuracy in Normal lighting (93.6%), it maintains 92.0% in Low-light, significantly surpassing RGB-based baselines (e.g., EMO [1] at 60.1%). This confirms that the STFE module successfully leverages event data to compensate for texture loss in dark environments.

Table 1. Comparison with SOTA

| Methods | Metrics(%) | | Accuracy under lighting conditions (%) | | | | Accuracy of emotion classification (%) | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | WAR | UAR | Normal | Overexposure | Low-light | HDR | Happy | Sadness | Anger | Disgust | Surprise | Fear | Neutral |
| EMO | 63.1 | 63.3 | 61.8 | 62.8 | 60.1 | 69.6 | 75.0 | 75.1 | 70.2 | 48.1 | 37.5 | 54.1 | 82.8 |
| EMO w/o pre-train | 53.2 | 53.3 | 46.1 | 60.2 | 55.5 | 58.9 | 62.0 | 73.2 | 60.1 | 38.7 | 25.7 | 48.0 | 65.3 |
| Eyemotion | 78.8 | 79.5 | 79.0 | 81.8 | 81.5 | 72.5 | 74.3 | 85.5 | 79.5 | 74.3 | 69.1 | 79.2 | 94.5 |
| Eyemotion w/o pretrain | 75.9 | 77.2 | 77.8 | 75.9 | 79.8 | 69.7 | 79.6 | 85.7 | 81.2 | 71.2 | 54.7 | 87.7 | 85.2 |
| SEEN | 83.6 | 84.1 | 83.3 | 85.6 | 80.8 | 84.8 | 85.0 | 89.9 | 92.2 | 76.7 | 72.1 | 88.0 | 90.3 |
| HI-Net | 86.9 | 87.7 | 84.6 | 90.3 | 87.2 | 85.2 | 93.4 | 95.5 | 87.8 | 85.3 | 79.4 | 91.2 | 89.8 |
| Ours | 91.5 | 92.3 | 93.6 | 92.6 | 90.6 | 89.2 | 93.4 | 97.01 | 98.9 | 91.7 | 94.0 | 84.4 | 89.4 |

## 4.3. Ablation study

We checked how each part contributes to the model using the SEE dataset (Table 2). First, we looked at the STFE module. It is used to catch time-based changes. When we replaced the RGB-STFE with a normal ResNet-18 (Row A), the UAR dropped by 2.7%. This shows that TSM is needed to see small movements. Also, removing the Event-STFE caused a bigger drop of 4.1% (Row B). This proves that the Top-k selection is important. It removes background hot-pixels so we can get real motion signals. Next, we tested the interaction strategies. We found that one-way methods are not good enough. Using only RGB to guide Event (Row C) or Event to guide RGB (Row D) gave lower scores (90.6% and 90.1%). The full model got 92.3%. This means we need two-way interaction to use both spatial and temporal clues well. Finally, looking at TPA (Row E),

DOI: 10.54254/2753-8818/2026.31464

the performance dropped by 1.5% without it. This tells us that a single time scale is not enough. We need to handle different motion speeds, like fast blinks versus slow frowns.

Table 2. Ablation analysis of HEC-Net

| Variant / Components | WAR (%) | UAR (%) |
|---|---|---|
| RGB Stream Only (ResNet-18) | 83.5 | 84.2 |
| Event Stream Only (ResNet-18) | 84.8 | 85.6 |
| w/o STFE on RGB (Standard ResNet) | 88.5 | 89.6 |
| w/o STFE on Event (Standard ResNet) | 87.1 | 88.2 |
| One-way Interaction (from rgb to event) | 89.8 | 90.6 |
| One-way Interaction (from event to rgb) | 89.4 | 90.1 |
| w/o TPA (Single Temporal Scale) | 90.2 | 90.8 |
| w/o Trilinear Interpolation | 90.5 | 91.1 |
| HEC-Net (Ours) | 91.5 | 92.3 |

As shown in Figure 3, we visualize Grad-CAM heatmaps to verify semantic focus. Compared to SEEN (1st row), our HEC-Net (2nd row) precisely locates eye regions (e.g., pupils) even under degraded lighting, whereas SEEN is distracted by noise. In contrast, removing STFE (3rd row) results in highly diffused attention, validating the noise-filtering capability of TSM and Top-k. Similarly, excluding Multi-window or TPA (4th/5th rows) noticeably deteriorates semantic alignment. These results visually reverse-validate the indispensability of each proposed component.

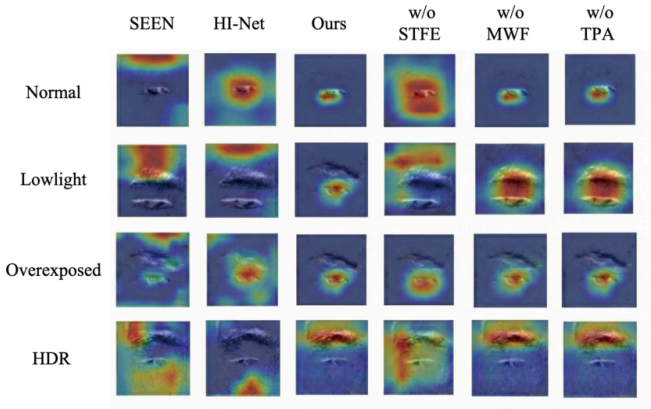


Figure 3. Heat map visualization of SEEN, our HEC-Net, and ablation studies (without (w/o) STFE and without (w/o) multi-window fusion and without TPA)

## 5. Conclusion

In this paper, we introduce a brand new framework “HEC-Net” to deal with the difficulties of spatio-temporal misalignment and noise interference in single-eye emotion recognition. Our method achieves a state-of-the-art UAR of 92.30% by combining a TSM-enhanced STFE module and a three-stage hierarchical fusion strategy. Our experiment’s results show that our method can maintain high stability with minimal performance degradation under low-light and overexposed conditions, which confirms the robustness and effectiveness of our hierarchical alignment paradigm. In future work, we will form further studies to facilitate real-time deployment on edge devices.

DOI: 10.54254/2753-8818/2026.31464

## References

[1] Wu H, Feng J and Tian X, et al. “Emo: real-time emotion recognition from single-eye images for resource-constrained eyewear devices, ” In Proc. ACM MobiSys, 448-461, 2020.
[2] Hickson S, Dufour N and Sud A, et al. “Eyemotion: classifying facial expressions in vr using eye-tracking cameras, ” In Proc. IEEE WACV, pp. 1626-1635, 2019.
[3] Gallego G, Delbrück T and Orchard G, et al. “Event-based vision: a survey, ” IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 1, pp. 154-180, 2020.
[4] Zhang H, Zhang J and Dong B, et al. “In the blink of an eye: event-based emotion recognition, ”ACM SIGGRAPH Conf. Proc., Article 1, pp. 1-11, 2023.
[5] Han R, Liu X and Zhang Y, et al. “Hierarchical event-rgb interaction network for single-eye expression recognition, ” Inf. Sci., vol. 690, Art. no. 121539, 2024.
[6] He K, Zhang X, Ren S, and Sun J. “Deep residual learning for image recognition, ” In Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 770-778, 2016.
[7] Yan W, Li X, Wang S, Zhao G, Liu Y, and Chen Y. “Casme II: an improved spontaneous micro-expression database and the baseline evaluation, ” PLoS ONE, vol. 9, no. 1, p. e86041, 2014.
[8] Lin J, Gan C, and Han S. “TSM: temporal shift module for efficient video understanding, ” In Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), pp. 7083-7093, 2019.
[9] Liu Z, Lin Y and Cao Y, et al. “Swin transformer: hierarchical vision transformer using shifted windows, ” In Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), pp. 10012-10022, 2021.
[10] Yang J, Li C, Zhang P, et al. “Focal self-attention for local-global interactions in vision transformers, ” In Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 34, pp. 30013–30025, 2021.