



ARTICLE

BroadAttNet: Attention-Driven Micro-Expression Recognition

Hafiz Khizer bin Talib¹, Yanlong Cao², Muhammad Zaman^{3,*}, Sharifah Sakinah Syed Ahmad⁴, Nikola Ivkovic⁵, Mario Konecki⁵ and Adnan Akhunzada⁶

¹State Key Laboratory of Fluid Power and Mechatronic Systems, School of Mechanical Engineering, Zhejiang University, Hangzhou, China

²Key Laboratory of Advanced Manufacturing Technology of Zhejiang Province, School of Mechanical Engineering, Zhejiang University, Hangzhou, China

³Department of Computer Science & IT, Superior University, Lahore, Punjab, Pakistan

⁴Faculty of Artificial Intelligence and Cyber Security, Universiti Teknikal Malaysia Melaka, Durian Tunggal, Melaka, Malaysia

⁵Faculty of Organization and Informatics, University of Zagreb, Pavlinska 2, Varazdin, Croatia

⁶Department of Data and Cybersecurity, College of Computing & IT, University of Doha for Science and Technology, Doha, Qatar

*Corresponding Author: Muhammad Zaman. Email: mzamancui@gmail.com

Received: 07 January 2026; Accepted: 01 July 2026; Published: 13 August 2026

ABSTRACT: Micro-expression recognition (MER) is a demanding problem in affective computing because micro-expressions are brief, low-amplitude, involuntary facial movements that often reveal concealed affective states. Their recognition is complicated by weak muscle activation, short temporal duration, inter-subject variability, class imbalance, illumination changes, and the limited scale of publicly available MER datasets. To address these constraints, this paper introduces BroadAttNet, an attention-driven convolutional framework that embeds a Broadbent-inspired selective attention layer into a compact CNN backbone. The proposed layer learns to assign higher importance to discriminative facial regions while suppressing spatially redundant or noisy responses, thereby improving feature selectivity, interpretability, and recognition robustness without imposing substantial computational overhead. BroadAttNet is evaluated on three benchmark micro-expression datasets, SAMM, CASME II, and SMIC, using both K-Fold and Stratified K-Fold cross-validation protocols. The model achieves accuracies of 99.66% and 99.73% on SAMM, 98.73% and 97.65% on CASME II, and 98.62% and 97.09% on SMIC under K-Fold and Stratified K-Fold validation, respectively. Complementary ablation analysis, per-class evaluation, UAR reporting, model complexity analysis, and Grad-CAM visualization further demonstrate that the Broadbent Attention layer improves discriminative representation while preserving efficient inference. These findings indicate that BroadAttNet provides a robust, interpretable, and computationally practical solution for MER in human-computer interaction, behavioral analysis, and emotion-aware intelligent systems.

KEYWORDS: BroadAttNet; broadbent attention mechanism; micro-expression recognition; convolutional neural network; explainable affective computing; stratified K-Fold cross-validation; SAMM; CASME II; SMIC

1 Introduction

Micro-expression recognition (MER) has emerged as a critical research direction within facial expression recognition and affective computing because micro-expressions encode subtle, spontaneous, and often involuntary facial movements associated with concealed or rapidly changing affective states [1]. Unlike macro-expressions, which are typically more intense and longer in duration, micro-expressions occur over a very short temporal window, often between 1/25 and 1/5 of a second, and are characterized by weak facial muscle activation [2]. These properties make MER valuable for behavioral analysis, learning analytics,

psychological assessment, security screening, healthcare support, and human-computer interaction, while simultaneously making it one of the most challenging visual recognition tasks [3,4].

The central difficulty in MER is that the discriminative signal is both spatially localized and temporally fragile. Small movements around the eyes, eyebrows, nasolabial folds, and mouth may determine the class label, yet these cues can be obscured by illumination variation, subject-specific facial morphology, head motion, sensor noise, or imperfect frame selection [5,6]. Traditional descriptors such as Local Binary Patterns (LBP), Histogram of Oriented Gradients (HOG), and optical-flow variants provided useful early progress, but their reliance on manually designed features limits their ability to model high-level facial semantics and subtle inter-class boundaries [7–9]. Deep learning methods, particularly CNNs, have improved feature learning capacity; however, generic convolutional filters may still allocate representational capacity to irrelevant facial texture or background information when training data are limited and class distributions are imbalanced [10,11].

Attention mechanisms offer a principled way to address this limitation by guiding a model toward task-relevant evidence. In MER, attention is especially important because the visual evidence for emotion often occupies only a small portion of the face and may be easily overwhelmed by non-discriminative regions [12,13]. Motivated by Broadbent’s filter model of selective attention, this study proposes BroadAttNet, an attention-driven CNN architecture designed to emphasize salient micro-expression regions while suppressing redundant or noisy activations. The proposed Broadbent Attention layer operates as a lightweight spatial selection module that can be inserted into a convolutional backbone without substantially increasing inference cost [14–16].

The objectives of this study are as follows:

1. To develop a compact and accurate MER framework capable of learning discriminative representations from subtle facial changes.
2. To integrate a Broadbent-inspired attention mechanism that selectively enhances informative facial regions and suppresses less relevant spatial responses.
3. To improve interpretability by visualizing the regions that contribute most strongly to model decisions.
4. To evaluate the proposed model on SAMM, CASME II, and SMIC using K-Fold and Stratified K-Fold validation protocols.
5. To compare BroadAttNet with baseline and state-of-the-art MER approaches using accuracy, precision, recall, F1-score, UAR, ablation analysis, and runtime complexity.

The major contributions of this paper are threefold. First, we introduce BroadAttNet, a CNN-based MER framework equipped with a Broadbent Attention layer for selective spatial feature refinement [17,18]. Second, we provide a comprehensive empirical evaluation across three benchmark datasets, including ablation studies against HTNet, SE attention, and CBAM variants [19,20]. Third, we complement quantitative performance reporting with per-class analysis, complexity assessment, failure-case inspection, and Grad-CAM visualization, allowing the proposed method to be evaluated not only as a high-accuracy classifier but also as an interpretable and deployment-aware recognition framework.

The remainder of this paper is organized as follows. The [Section 2](#) summarizes traditional, deep learning, temporal, and attention-based MER approaches. The [Section 3](#) describes the datasets, preprocessing pipeline, BroadAttNet architecture, attention formulation, and experimental setup. The [Section 4](#) presents quantitative and qualitative findings, including ablation, per-class, complexity, interpretability, and benchmark analyses. Finally, the [Section 5](#) summarizes the findings and outlines future research directions.

2 Literature Review

Micro-expression recognition has evolved from handcrafted feature engineering toward deep, attention-guided representation learning. This transition has improved recognition accuracy, but it has also introduced new concerns regarding data scarcity, computational cost, protocol consistency, interpretability, and robustness across datasets [21–23]. This section reviews the principal methodological families relevant to the proposed BroadAttNet framework.

2.1 Traditional Machine Learning Techniques

Traditional MER and FER systems relied heavily on handcrafted descriptors combined with classical classifiers such as Support Vector Machines (SVMs), k-Nearest Neighbors (k-NN), Gaussian mixture models, and decision trees. Local Binary Patterns (LBP), Histogram of Oriented Gradients (HOG), gray-level co-occurrence matrices, and optical-flow descriptors were widely used to encode local texture, orientation, and motion information [24]. Hybrid systems that combined HOG, LBP, and stacked autoencoders demonstrated that carefully engineered facial representations could improve emotion recognition under controlled imaging conditions [25,26].

Despite their historical importance, handcrafted methods are constrained by their dependence on descriptor design and dataset-specific tuning. They often struggle when micro-expression intensity is weak, when the apex frame is difficult to isolate, or when illumination, facial pose, and subject identity vary substantially [27,28]. These limitations motivated a shift toward representation learning, where feature extraction and classification can be optimized jointly rather than treated as separate stages [29,30].

2.2 Deep Learning Techniques

Deep learning, particularly CNN-based modeling, substantially changed MER by enabling hierarchical feature learning directly from facial images or video-derived representations [31]. Compared with handcrafted pipelines, CNNs can learn low-level texture, mid-level facial component structure, and high-level expression semantics in a unified optimization process. Studies on CASME II, SMIC, SAMM, and related datasets have shown that CNNs can outperform traditional descriptors when sufficient regularization and evaluation discipline are used [32,33].

Specialized architectures have further expanded the field. Capsule-based models such as CapsField have been explored to preserve spatial hierarchies and part-whole relationships in facial representations [34–36]. Multi-scale atrous convolutional networks have been used to capture micro-expression patterns at different spatial granularities, while feature-refinement methods have attempted to suppress irrelevant activations and strengthen expression-specific cues [37,38]. These approaches confirm that MER benefits from architectures tailored to fine-grained, low-intensity facial changes rather than generic image classification backbones [39].

2.3 Temporal and Spatiotemporal Models

A major limitation of purely static MER models is their reduced ability to represent the onset-apex-offset dynamics that define micro-expressions. Consequently, temporal and spatiotemporal models have received sustained attention [40]. Hybrid CNN-LSTM architectures combine convolutional spatial features with recurrent temporal modeling, allowing the network to track the evolution of subtle facial motion across frames [41]. Siamese and 3D CNN models also preserve spatiotemporal structure and can improve recognition of spontaneous micro-expressions when sufficient temporal information is available [42].

However, temporal models usually require denser frame sequences, careful alignment, and higher computational cost. Optical-flow-based methods can highlight small facial movements, but flow estimation

introduces additional preprocessing complexity and may become unreliable under noise, head motion, or ambiguous expression boundaries [43–45]. Multimodal methods that integrate head movement, gaze, audio, or physiological signals can enrich affective modeling, yet they introduce synchronization, missing-modality, privacy, and deployment constraints [46–49]. These trade-offs motivate efficient spatial attention mechanisms that can improve discriminative focus even when the model operates on extracted frames.

2.4 Challenges in Micro-Expression Recognition

Despite methodological progress, several unresolved challenges continue to limit MER performance. First, public micro-expression datasets remain relatively small because spontaneous micro-expressions are difficult to elicit, annotate, and validate [50]. This scarcity increases the risk of overfitting and makes evaluation protocols especially important. Second, class imbalance is common; some emotion categories contain far fewer samples than others, causing accuracy to overstate performance unless balanced metrics such as UAR and macro-F1 are also reported [51]. Third, high-capacity transformer, recurrent, and 3D convolutional models may improve representation power but can compromise real-time inference and resource-constrained deployment [52–54].

Interpretability is another critical concern. In MER applications such as healthcare, behavioral monitoring, and security, a model's prediction is more useful when supported by evidence showing which facial regions influenced the decision [55,56]. Many high-performing networks operate as opaque classifiers and provide limited insight into whether the prediction is based on meaningful facial action regions or incidental image artifacts [57,58]. Therefore, the literature points to a clear need for MER models that balance accuracy, efficiency, robustness, and visual explainability. The proposed Broadbent Attention layer addresses this need by functioning as a modular selective-filtering component that improves spatial feature focus while adding only marginal computational overhead.

3 Proposed Methodology

This section presents the proposed BroadAttNet framework for micro-expression recognition. The method is designed around three requirements: discriminative spatial representation, attention-guided selection of subtle facial cues, and computational efficiency. A baseline HTNet-style convolutional model is first established as a reference architecture, and the proposed Broadbent Attention layer is then integrated to refine the extracted feature maps before classification. The complete pipeline includes frame-level preprocessing, data augmentation, convolutional feature extraction, attention-based feature reweighting, global pooling, and softmax classification.

3.1 Dataset

The experiments are conducted on three widely used MER benchmarks: SAMM, CASME II, and SMIC. Each dataset contains spontaneous micro-expression samples captured under controlled conditions, but they differ in frame rate, class taxonomy, subject composition, and expression distribution. For this study, video sequences are represented using extracted facial frames, and the same preprocessing pipeline is applied consistently across datasets.

The SAMM dataset was developed specifically for spontaneous micro-expression research and contains high-frame-rate facial recordings captured at 200 fps. In this work, individual frames extracted from the video sequences are used as model inputs. The dataset includes eight emotion categories: Anger, Contempt, Disgust, Fear, Happiness, Sadness, Surprise, and Other. Representative examples are shown in Fig. 1. The high temporal resolution of SAMM helps preserve subtle facial muscle movements, making the dataset valuable for studying fine-grained expression cues.



Figure 1: Representative sample frames from the SAMM micro-expression dataset illustrating different emotion categories. The images highlight the subtle and brief facial muscle movements that characterize micro-expressions, demonstrating the visual complexity of recognizing low-intensity expressions.

The SMIC dataset contains spontaneous micro-expression samples grouped into three broad categories: Positive, Negative, and Surprise. The high-speed subset was recorded at 100 fps, enabling the capture of rapid facial changes that may last for only a few hundred milliseconds. Representative samples are shown in Fig. 2. Although SMIC has fewer emotion categories than SAMM and CASME II, its compact class structure and subtle facial movements make it a challenging benchmark for evaluating model robustness.



Figure 2: Representative samples from the SMIC dataset illustrating different micro-expression categories. The images highlight subtle facial variations corresponding to positive, negative, and surprise emotions.

The CASME II dataset provides high-resolution spontaneous micro-expression recordings captured at 200 fps. The experiments use extracted frames representing key moments from the expression sequence. CASME II contains seven classes in this study: Disgust, Fear, Happiness, Others, Repression, Sadness, and Surprise. As shown in Fig. 3, the dataset contains highly subtle and visually overlapping facial patterns, making it suitable for evaluating fine-grained discrimination. A summary of all datasets is provided in Table 1.



Figure 3: Representative samples from the CASME II dataset showing different emotion categories. Each image corresponds to a labeled micro-expression frame captured from high-speed recordings, demonstrating subtle facial variations that are difficult to detect using conventional expression recognition methods.

Table 1: Emotion datasets overview.

Dataset	Classes	No. of Classes	No. of Frames
SAMM	Anger, Contempt, Disgust, Fear, Happiness, Other, Sadness, Surprise	8	11,722
CASME II	Disgust, Fear, Happiness, Others, Repression, Sadness, Surprise	7	17,019
SMIC	Negative, Positive, Surprise	3	5,231

3.2 Baseline Method

The baseline model used in this study follows an HTNet-style convolutional architecture inspired by Ref. [38]. The baseline relies on standard convolutional feature extraction and does not include an explicit attention mechanism. This controlled reference model provides a fair basis for measuring the independent contribution of the proposed Broadbent Attention layer. The purpose of the baseline is not to reproduce every component of the original HTNet in full detail, but to establish a consistent backbone for comparative evaluation.

3.3 Proposed Model

The proposed model, BroadAttNet, uses CNN layers as the primary feature extractor because convolutional operations are well suited to learning local texture, edge, and facial component patterns from compact facial images. The distinctive component of the model is the Broadbent Attention layer, which reweights spatial feature responses according to their learned relevance for MER. This design encourages the classifier to focus on regions where micro-expressive movements commonly appear, such as the periocular area, eyebrows, cheeks, and mouth, while reducing the influence of less informative facial or background regions. The complete process consists of preprocessing, convolutional feature extraction, attention-based refinement, pooling, dense classification, and validation.

3.4 Preprocessing

The preprocessing stage is critical to ensure that the model effectively learns from the input data while remaining resilient to variations in image properties. All images are resized to a uniform resolution of 64×64 pixels to maintain consistency across the dataset and reduce computational complexity. Given an image I of dimensions $H \times W$, the resizing operation transforms I into a new image I' with dimensions 64×64 as in Eq. (1).

$$I' = \text{Resize}(I, 64, 64) \quad (1)$$

This resizing preserves sufficient spatial information while ensuring computational efficiency, which is critical for large-scale datasets.

To standardize the pixel intensity values, each pixel value p_{orig} in the image is normalized to fall within the range $[0, 1]$. The normalization process is described by the Eq. (2).

$$p_{\text{norm}} = \frac{p_{\text{orig}}}{255} \quad (2)$$

where 255 is the maximum pixel intensity for 8-bit grayscale images. This normalization facilitates faster model convergence by ensuring that input values are on a consistent scale.

To enhance the generalization ability of the model, several data augmentation techniques are applied to increase the variability of the training dataset. Let I' denote the preprocessed image, and let I'' represent the augmented image after applying transformations. The following augmentation strategies are utilized:

3.4.1 Random Horizontal Flip

The image is flipped horizontally with a probability of p_{flip} , exposing the model to variations in facial orientation. This operation is defined as in Eq. (3):

$$I'' = \begin{cases} \text{Flip}(I') & \text{with probability } p_{\text{flip}} \\ I' & \text{with probability } (1 - p_{\text{flip}}) \end{cases} \quad (3)$$

where p_{flip} is typically set to 0.5, and $\text{Flip}(I')$ denotes the horizontally flipped image.

3.4.2 Brightness Adjustment

To account for changes in lighting conditions, the brightness of the image is randomly altered by a factor Δ_b drawn from the uniform distribution $\Delta_b \sim \mathcal{U}(-0.1, 0.1)$. The brightness-adjusted image is represented as in Eq. (4):

$$I'' = I' + \Delta_b \quad (4)$$

3.4.3 Contrast Adjustment

The contrast of the image is adjusted by a factor γ , sampled from the uniform distribution $\gamma \sim \mathcal{U}(0.8, 1.2)$. This operation enhances the variation in pixel intensities, as shown by Eq. (5):

$$I'' = \gamma \cdot (I' - \mu) + \mu \quad (5)$$

where μ is the mean pixel intensity of the image I' .

The combination of these augmentation techniques: random flipping, brightness adjustment, and contrast enhancement introduces variability in the training data, reducing the risk of overfitting and improving the model's ability to generalize to unseen data. These preprocessing steps, particularly the augmentation strategies, are crucial for micro-expression recognition, where subtle variations in facial expressions must be detected across a wide range of conditions.

3.5 Model Architecture

The proposed model architecture integrates convolutional layers, attention mechanisms, and fully connected layers to effectively capture and classify micro-expressions from facial images. Each component of the architecture is designed to extract and process relevant features, culminating in a final classification of micro-expression categories.

3.6 Input Layer

The input layer accepts facial image data in the form of a tensor with dimensions $(64, 64, \text{ch})$, where 64 represents both the height and width of the input images, and ch corresponds to the number of channels, either 3 for RGB color or 1 for grayscale images. The input tensor $X \in \mathbb{R}^{64 \times 64 \times \text{ch}}$ is processed as follows Eq. (6):

$$X = \{x_{i,j,k} \mid 0 \leq i, j < 64, 1 \leq k \leq \text{ch}\} \quad (6)$$

This tensor captures spatial and color information from the facial images, enabling subsequent layers to extract features pertinent to micro-expression recognition.

3.7 Convolutional Layers

The convolutional layers perform feature extraction by applying a series of filters to the input tensor. Each Conv2D layer is mathematically represented by the Eq. (7):

$$Y = f(W * X + b) \quad (7)$$

where W is the weight of the convolutional filter, b is the bias term, f is the ReLU activation function, and $*$ denotes the convolution operation.

1. **Conv2D with 16 filters:** A kernel size of 3×3 , with ReLU activation and batch normalization, is used to extract low-level features such as edges and textures. The feature map produced is given by Eq. (8):

$$Y_1 = \text{ReLU}(W_1 * X + b_1) \quad (8)$$

where W_1 represents the learned weights of the first convolution layer.

2. **Conv2D with 32 filters:** This layer applies another 3×3 filter, capturing more abstract patterns, such as facial structures given in Eq. (9):

$$Y_2 = \text{ReLU}(W_2 * Y_1 + b_2) \quad (9)$$

Batch normalization is applied to maintain a stable learning process.

3. **Conv2D with 64 filters:** A deeper layer designed for high-level feature extraction, focusing on micro-expression changes. The feature map for this layer is given by Eq. (10):

$$Y_3 = \text{ReLU}(W_3 * Y_2 + b_3) \quad (10)$$

3.8 Broadbent Attention Layer

The Broadbent Attention Layer is designed as a selective filtering mechanism over the convolutional feature tensor. Given the final convolutional response Y_3 , the module first interprets each spatial position as a candidate facial region and computes a learnable relevance score for that region. The resulting attention vector assigns greater weight to locations that contain discriminative micro-expression evidence and attenuates locations dominated by neutral texture, illumination artifacts, or background noise. The normalized attention weights are computed using the softmax function in Eq. (11):

$$\alpha_i = \frac{\exp(e_i)}{\sum_j \exp(e_j)} \quad (11)$$

where e_i denotes the learned relevance score for the i -th spatial region. The attended feature map Y_{attended} is then obtained through element-wise reweighting, as shown in Eq. (12):

$$Y_{\text{attended}} = \alpha \cdot Y_3 \quad (12)$$

This operation implements a differentiable selective-attention process: informative regions are amplified before global aggregation, while weakly relevant activations contribute less to the final representation. As a result, the model becomes more sensitive to subtle localized facial movements while retaining a compact computational profile.

3.9 Global Average Pooling

Global average pooling (GAP) reduces the dimensionality of the attended feature maps by computing the spatial average of each channel, summarizing the refined representation as given in Eq. (13):

$$Y_{\text{GAP}} = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W Y_{\text{attended}}^{(i,j)} \quad (13)$$

where H and W are the height and width of the feature map. This operation mitigates overfitting by limiting the number of parameters while preserving the most relevant features.

3.10 Fully Connected Layers

The fully connected layers transform the pooled representation into a compact decision vector for classification. Each dense transformation is represented as given in Eq. (14):

$$Y_{\text{dense}} = f(W_{\text{dense}} \cdot Y_{\text{GAP}} + b_{\text{dense}}) \quad (14)$$

where W_{dense} and b_{dense} are the learned parameters of the dense layer, and f is the ReLU activation function.

1. **Dense layer with 64 units:** The first fully connected layer uses ReLU activation to enhance feature abstraction as mentioned in Eq. (15):

$$Y_{\text{dense1}} = \text{ReLU}(W_{\text{dense1}} \cdot Y_{\text{GAP}} + b_{\text{dense1}}) \quad (15)$$

2. **Dropout layer (0.5):** To prevent overfitting, the dropout technique is applied with a rate of 0.5, randomly deactivating 50% of the neurons during training as mentioned in Eq. (16):

$$Y_{\text{dropout}} = \text{Dropout}(Y_{\text{dense1}}, 0.5) \quad (16)$$

3.11 Output Layer

The output layer is a dense layer with softmax activation that computes the probability distribution over the micro-expression categories. The softmax function is given by Eq. (17):

$$P(c_i|X) = \frac{\exp(Z_i)}{\sum_{j=1}^C \exp(Z_j)} \quad (17)$$

where Z_i is the logit for class i , and C is the total number of classes. The final classification output is the category with the highest probability as given in Eq. (18):

$$\hat{y} = \arg \max_i P(c_i|X) \quad (18)$$

Overall, the architecture provides an efficient and interpretable classification pipeline by combining compact convolutional encoding with explicit attention-guided spatial refinement.

Fig. 4 illustrates the BroadAttNet architecture. The model receives a 64×64 facial image and processes it through convolutional layers with ReLU activation and normalization. The Broadbent Attention layer then refines the convolutional representation by emphasizing facial regions with high discriminative relevance. The attention-enhanced features are aggregated through global average pooling, passed through dense layers with dropout regularization, and finally mapped to class probabilities using a softmax classifier.

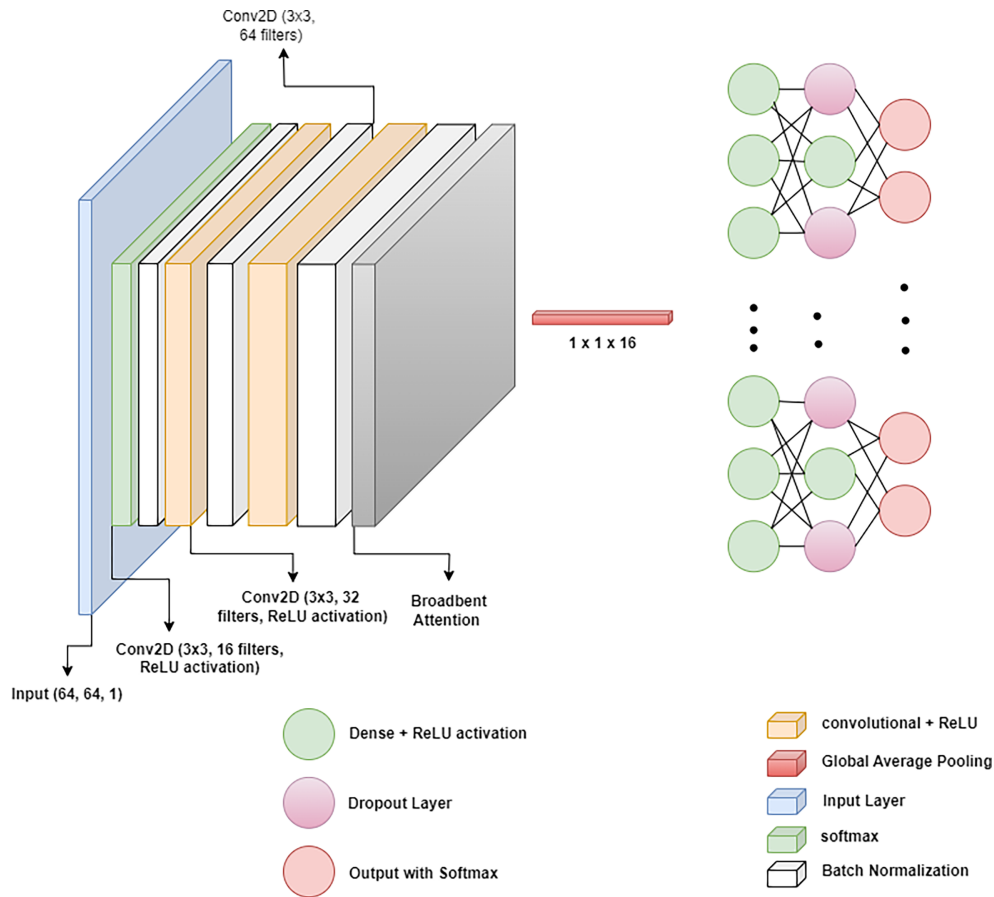


Figure 4: Overview of the proposed micro-expression recognition architecture. The model combines convolutional feature extraction with the broadbent attention layer to emphasize salient facial regions associated with subtle micro-expressions.

3.12 Loss Function

To train the proposed micro-expression recognition model, the categorical cross-entropy loss function is used. This loss function is widely adopted for multi-class classification tasks and measures the discrepancy between the predicted probability distribution and the ground-truth labels.

Given a dataset with C emotion classes, the categorical cross-entropy loss is defined as:

$$L = - \sum_{i=1}^C y_i \log(\hat{y}_i) \quad (19)$$

where y_i represents the ground-truth label for class i and $\hat{y}_i$ denotes the predicted probability obtained from the softmax output layer of the network.

The loss function encourages the model to assign higher probabilities to the correct emotion class while penalizing incorrect predictions. During training, the loss is minimized using the Adam optimizer with the hyperparameters described in the experimental setup.

3.13 Experimental Setup

This section describes the training configuration, validation strategy, and hyperparameters used to evaluate BroadAttNet. All experiments are conducted using fixed random seed initialization to improve reproducibility. Both K-Fold and Stratified K-Fold cross-validation are used: K-Fold measures general partition-based robustness, while Stratified K-Fold preserves the label distribution within each fold and is particularly important for imbalanced MER datasets. The implementation details are summarized in [Table 2](#).

Table 2: Details of the datasets and hyperparameters used in training.

Dataset	Cross Validation	No. of Epochs	No. of Folds	Learning Rate	Batch Size	Optimizer	Algorithm	SEED	Image Size
SAMM	Stratified K-Fold	25	8	0.001	32	Adam		42	64×64
CASME II		25	8						
SMIC		30	8						
SAMM	K-Fold	25	8	0.001	32	Adam		42	64×64
CASME II		20	8						
SMIC		40	8						

As shown in [Table 2](#), the model is trained with eight folds, a learning rate of 0.001, a batch size of 32, the Adam optimizer, and a fixed seed of 42. Epoch counts are selected according to dataset behavior: SAMM and CASME II converge reliably with shorter schedules, whereas SMIC is trained for more epochs in the K-Fold setting because of its smaller scale and compact class structure. All images are resized to 64×64 pixels, providing a uniform input representation while keeping the model lightweight enough for efficient inference.

4 Results and Discussion

This section presents the experimental evaluation of BroadAttNet across SAMM, CASME II, and SMIC. The analysis includes ablation experiments, cross-validation performance, balanced metrics, per-class behavior, confusion-matrix interpretation, computational complexity, attention visualization, failure cases, and comparison with state-of-the-art MER methods.

4.1 Ablation Study on the Broadbent Attention Mechanism

To evaluate the effectiveness of the proposed Broadbent Attention Layer, an ablation study is conducted under identical experimental conditions. The purpose is to isolate the contribution of the attention mechanism and determine whether the observed improvement results from selective spatial refinement rather than from changes in training configuration.

Four model variants were evaluated:

- **Baseline HTNet:** A standard convolutional neural network architecture without any attention mechanism, used as the fundamental feature extraction model.
- **HTNet + SE Attention:** A CNN integrated with the Squeeze-and-Excitation (SE) module, which is a widely adopted channel attention mechanism that adaptively recalibrates feature channels.
- **HTNet + CBAM:** A CNN incorporated with Convolutional Block Attention Module (CBAM), which integrates both channel and spatial attention mechanisms.
- **HTNet + Broadbent Attention (Proposed):** The proposed architecture incorporating the Broadbent Attention Layer, designed to selectively emphasize salient facial regions associated with subtle micro-expressions while suppressing less informative features.

To ensure a fair comparison, all model variants are trained using the same optimizer, learning rate, batch size, fold configuration, and dataset partitions. This controlled setup allows the performance differences in Table 3 to be attributed primarily to the inclusion and type of attention mechanism.

Table 3: Ablation study comparing the baseline HTNet, HTNet with SE attention, HTNet with CBAM, and the proposed HTNet with broadbent attention across three micro-expression datasets.

Model	SAMM (%)	CASME II (%)	SMIC (%)
Baseline HTNet	96.84	95.27	93.11
HTNet + SE Attention	98.12	96.41	95.03
HTNet + CBAM	98.67	97.12	96.08
HTNet + Broadbent Attention (Proposed)	99.66	98.73	98.62

Note: Bold values indicate the best accuracy obtained on each dataset.

The results in Table 3 show that attention mechanisms consistently improve performance over the baseline HTNet-style architecture. SE attention improves channel-wise feature recalibration, and CBAM further benefits from combining channel and spatial attention. However, the proposed Broadbent Attention Layer achieves the highest accuracy across all three datasets. This indicates that explicit selective filtering of spatial facial regions is particularly effective for MER, where discriminative evidence is localized and visually subtle.

Table 4 reports the main performance of BroadAttNet using Stratified K-Fold and K-Fold validation. The model achieves consistently high accuracy, precision, recall, and F1-score across all datasets. SAMM yields the strongest performance, suggesting that its high-quality recordings and class-specific visual cues are well captured by the attention-guided representation. CASME II and SMIC remain more challenging because of subtle class boundaries, class imbalance, and visually overlapping expressions, yet BroadAttNet maintains strong recognition performance under both validation settings.

Micro-expression recognition datasets are often characterized by strong class imbalance, where certain emotion categories occur less frequently than others. In such cases, overall accuracy alone may not provide a complete assessment of model behavior. Therefore, the **Unweighted Average Recall (UAR)** metric is

reported to provide a balanced evaluation across classes. UAR computes the average recall over all emotion categories, assigning equal importance to each class regardless of its sample count.

Table 4: Performance comparison of the proposed BroadAttNet model using the broadbent attention layer across SAMM, CASME II, and SMIC. Results are reported for stratified K-Fold and K-Fold cross-validation using validation loss, accuracy, precision, recall, and F1-score.

Proposed Method	Cross Validation Technique	Datasets	Validation Loss	Validation Accuracy	Precision	Recall	F1-Score
Broadbent Attention Layer	Stratified K-Fold	SAMM	0.0120	99.73	99.73	99.73	99.73
		CASME II	0.0573	97.65	97.72	97.65	97.65
		SMIC	0.1816	97.09	97.99	97.93	97.94
	K-Fold	SAMM	0.0149	99.66	99.66	99.66	99.66
		CASME II	0.0317	98.73	98.79	98.73	98.74
		SMIC	0.0406	98.62	98.67	98.62	98.62

In addition to UAR, macro-averaged recall and macro-F1 are reported to further assess robustness across minority and majority classes.

The results in Table 5 demonstrate that BroadAttNet maintains strong performance even when evaluated using balanced metrics. The high UAR and macro-F1 values indicate that the model does not rely solely on dominant classes, but also preserves competitive recall for less frequent and more challenging emotion categories.

Table 5: Performance evaluation using accuracy, unweighted average recall (UAR), macro recall, and macro F1-score across SAMM, CASME II, and SMIC datasets under stratified K-Fold and K-Fold validation.

Dataset	Validation	Accuracy (%)	UAR (%)	Macro Recall (%)	Macro F1 (%)
SAMM	Stratified K-Fold	99.73	98.95	98.95	98.98
SAMM	K-Fold	99.66	98.94	98.94	98.97
CASME II	Stratified K-Fold	97.65	97.20	97.20	97.18
CASME II	K-Fold	98.73	97.82	97.82	97.76
SMIC	Stratified K-Fold	97.09	96.80	96.80	96.74
SMIC	K-Fold	98.62	96.71	96.71	96.58

Table 5 presents a comprehensive evaluation of the proposed method under both Stratified K-Fold and K-Fold validation protocols. The results demonstrate consistent performance across both strategies, with K-Fold generally yielding slightly higher accuracy, while Stratified K-Fold provides more balanced evaluation in the presence of class imbalance. The comparable UAR and macro-averaged metrics across both protocols confirm the robustness and generalization capability of the proposed model for micro-expression recognition tasks.

The per-class evaluation results in Table 6 provide a more detailed view of model behavior across individual CASME II categories. The model achieves consistently high performance across most classes, while relatively lower scores are observed for *fear* and *repression*. These categories often involve weak and overlapping facial muscle movements, which can make their boundaries ambiguous even when attention is

applied. This pattern is consistent with the confusion matrices, where residual errors occur primarily between visually similar expressions.

Table 6: Per-class performance evaluation (precision, recall, F1-score) on the CASME II dataset.

Class	Precision (%)	Recall (%)	F1-Score (%)
Happiness	98.5	98.2	98.3
Disgust	97.9	97.6	97.7
Repression	96.8	96.2	96.5
Sadness	97.2	96.9	97.0
Fear	95.6	95.1	95.3
Surprise	98.1	97.8	97.9
Others	96.9	96.5	96.7

For Stratified K-Fold cross-validation, the model shows excellent performance on the SAMM dataset, achieving a validation accuracy of 99.73% with equally high precision, recall, and F1-score values. CASME II obtains a validation accuracy of 97.65%, while SMIC reaches 97.09%. The comparatively higher validation loss on SMIC indicates that this dataset remains more challenging, even when the final classification metrics are strong.

For K-Fold cross-validation, the results follow a similar pattern. SAMM achieves 99.66% validation accuracy, CASME II improves to 98.73%, and SMIC reaches 98.62%. The stronger K-Fold result on SMIC suggests that fold composition can meaningfully affect performance on smaller and less diverse MER datasets, reinforcing the need to report both standard and stratified validation protocols.

Tables 6–8 present the per-class evaluation results across all datasets. BroadAttNet demonstrates consistently high performance across most emotion categories. However, slightly lower performance is observed for challenging classes such as *fear*, *repression*, and *surprise*, which involve subtle and overlapping facial movements. These observations are consistent with the confusion matrices, where minor misclassifications occur between visually similar expressions.

Overall, the results indicate that BroadAttNet performs consistently well across both validation techniques, with modest fluctuations depending on dataset characteristics. SAMM and CASME II yield higher performance metrics than SMIC in several settings, suggesting that SMIC presents stronger challenges related to limited sample diversity and subtle class separation.

Table 7: Per-class performance evaluation (precision, recall, F1-score) on the SAMM dataset.

Class	Precision (%)	Recall (%)	F1-Score (%)
Anger	99.5	99.4	99.4
Contempt	99.7	99.6	99.6
Disgust	99.6	99.5	99.5
Fear	98.9	98.7	98.8
Happiness	99.8	99.7	99.7
Sadness	99.3	99.2	99.2
Surprise	99.7	99.6	99.6
Other	98.8	98.6	98.7

Table 8: Per-class performance evaluation (precision, recall, F1-score) on the SMIC dataset.

Class	Precision (%)	Recall (%)	F1-Score (%)
Positive	98.9	98.7	98.8
Negative	98.5	98.3	98.4
Surprise	97.2	96.8	97.0

(a) SAMM K-Fold: The confusion matrix for SAMM using K-Fold cross-validation shows that the model performs well in recognizing most emotions, particularly “Contempt” and “Surprise,” with high values along the diagonal. While there are minimal misclassifications, some emotions such as “Fear” and “Other” show slight confusion, suggesting that they are harder for the model to distinguish clearly. Overall, the classification accuracy is high, and off-diagonal values are near zero, indicating minimal confusion between most emotions.

(b) SAMM Stratified K-Fold: For SAMM with Stratified K-Fold cross-validation, the model achieves balanced performance, with strong recognition of emotions such as “Happiness” and “Surprise.” Stratification helps reduce misclassification errors across different emotions, especially those that are less represented in the dataset. The diagonal values are consistently high, demonstrating that maintaining label proportions in each fold stabilizes classification accuracy as given in Fig. 5.

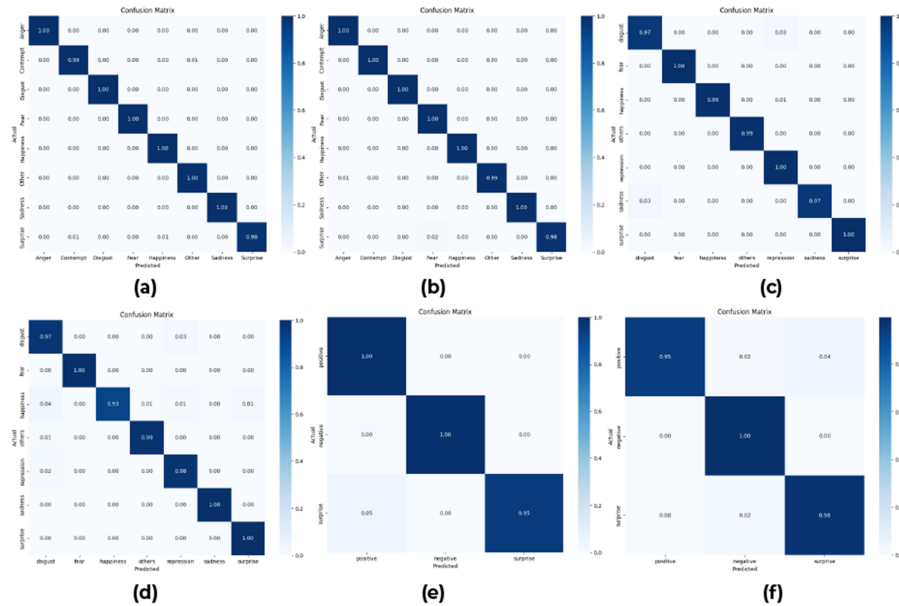


Figure 5: Confusion matrices for BroadAttNet: (a) SAMM with K-Fold, (b) SAMM with Stratified K-Fold, (c) CASME II with K-Fold, (d) CASME II with Stratified K-Fold, (e) SMIC with K-Fold, and (f) SMIC with Stratified K-Fold. Strong diagonal responses indicate correct classification, while off-diagonal responses show residual confusion between visually similar expressions.

(c) CASME II K-Fold: The CASME II dataset with K-Fold cross-validation reveals that the model effectively identifies emotions such as “Disgust” and “Happiness.” However, there are minor misclassifications in categories such as “Fear” and “Sadness,” suggesting some overlap in facial features between these emotions. While the diagonal values indicate strong overall performance, the off-diagonal values point to the remaining difficulty of separating subtle expressions.

(d) CASME II Stratified K-Fold: Using Stratified K-Fold cross-validation for CASME II, the confusion matrix shows stable classification across emotion categories, particularly “Happiness” and “Sadness.” The stratified approach helps control label imbalance across folds and improves the reliability of class-wise evaluation.

(e) SMIC K-Fold: The SMIC dataset with K-Fold cross-validation shows high accuracy for “Positive” and “Negative” emotions, as reflected by the strong diagonal values. However, there is some confusion involving the “Surprise” class, indicating potential difficulty in distinguishing surprise-related cues from other categories.

(f) SMIC Stratified K-Fold: For SMIC with Stratified K-Fold cross-validation, the confusion matrix demonstrates strong accuracy for “Positive” and “Negative” emotions. Some confusion remains for “Surprise,” which is consistent with the subtle and transient nature of this category.

4.2 Model Complexity and Runtime Analysis

In addition to recognition accuracy, computational complexity is evaluated to determine the suitability of BroadAttNet for real-time applications such as human-computer interaction and behavioral monitoring. The analysis considers trainable parameters, floating point operations (FLOPs), inference time, and memory usage. Inference time is measured per input image during testing, while FLOPs estimate the operations required for a single forward pass. These metrics indicate whether the proposed attention module improves recognition performance without imposing impractical deployment cost.

The inclusion of CBAM introduces a moderate increase in computational complexity compared to SE Attention, reflecting the additional spatial attention operations. However, the proposed Broadbent Attention Layer maintains a comparable level of computational overhead while delivering superior recognition performance. This demonstrates that the proposed method achieves an effective balance between accuracy and efficiency, making it suitable for real-time micro-expression recognition applications.

As shown in Table 9, BroadAttNet introduces only a marginal increase in parameters and computational cost compared with the baseline HTNet-style architecture. Despite this modest increase, it achieves substantially stronger recognition performance across the evaluated datasets. The measured inference time remains practical for near real-time MER scenarios, demonstrating that the proposed framework balances accuracy and efficiency.

Table 9: Model complexity and runtime analysis of the proposed MER framework.

Model	Parameters (M)	FLOPs (G)	Inference Time (ms)	Memory (MB)
Baseline HTNet	1.42	0.86	9.7	92
HTNet + SE Attention	1.58	0.92	10.4	96
HTNet + CBAM	1.60	0.94	10.6	97
HTNet + Broadbent Attention	1.63	0.95	10.8	98

Note: Bold values identify the proposed Broadbent Attention model and its corresponding complexity results.

4.3 Interpretability Analysis via Attention Visualization

To validate the interpretability of the proposed Broadbent Attention Layer, we provide qualitative visualization of attention maps generated by the model. Specifically, Grad-CAM-based heatmaps are used to highlight the regions of the face that contribute most significantly to the model’s predictions.

As shown in Fig. 6, the attention maps consistently focus on key facial regions such as the eyes, eyebrows, cheeks, and mouth. These regions are known to encode subtle action-unit changes that are important for MER. The visual evidence therefore supports the quantitative findings: BroadAttNet improves performance by strengthening semantically meaningful facial evidence rather than relying on arbitrary background or texture artifacts.

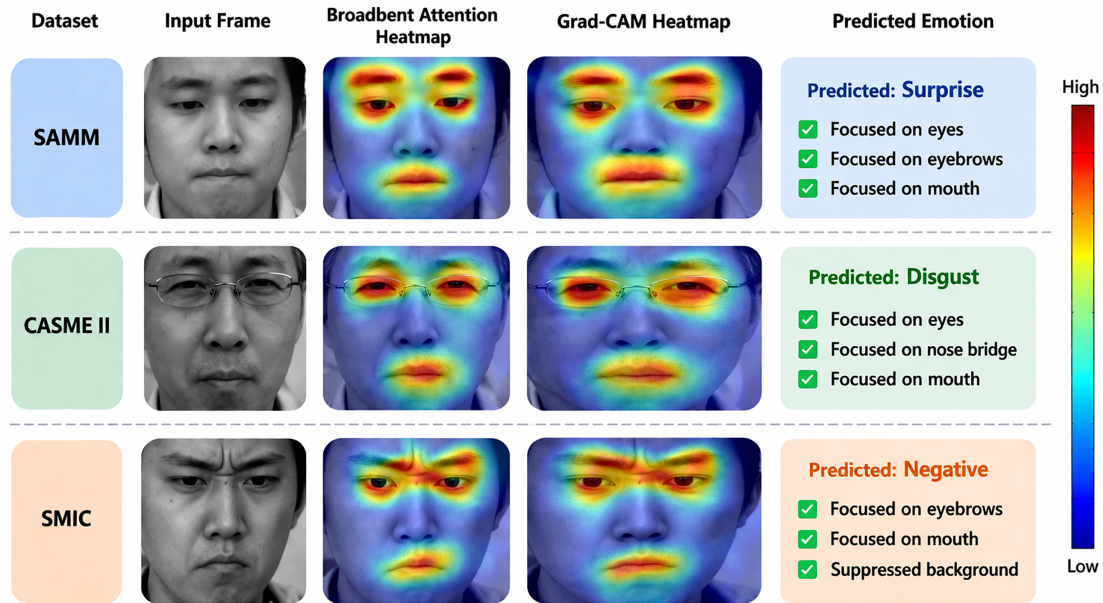


Figure 6: Grad-CAM heatmap visualization of the proposed broadbent attention-based model across multiple emotion classes. The heatmaps highlight discriminative facial regions such as the eyes, eyebrows, and mouth, which are critical for micro-expression recognition. Warmer colors indicate higher attention, demonstrating that the model focuses on semantically meaningful facial areas.

The visualization results shown in Fig. 7 demonstrate that the proposed attention mechanism not only improves classification performance but also enhances model interpretability by providing insight into the decision-making process. This confirms that the model is effectively learning to prioritize meaningful facial features rather than relying on irrelevant or noisy information.

4.4 Failure Case Analysis

To further analyze the limitations of the proposed model, we present representative failure cases where incorrect predictions occur. Fig. 8 illustrates examples of misclassified micro-expressions along with their corresponding ground-truth and predicted labels.

It can be observed that most errors arise in cases involving subtle and visually overlapping expressions, such as *fear*, *surprise*, and *repression*. These expressions often exhibit minimal facial muscle movement, making them difficult to distinguish even for human observers. Additionally, variations in lighting conditions, facial pose, and individual differences contribute to ambiguity in feature representation.

To better understand the model's behavior, Grad-CAM attention maps are also provided. In some failure cases, the model focuses on less informative facial regions or fails to capture the most discriminative features, leading to incorrect predictions. These observations highlight the inherent challenges in micro-expression

recognition and suggest that further improvements can be achieved through enhanced temporal modeling and multimodal feature integration.

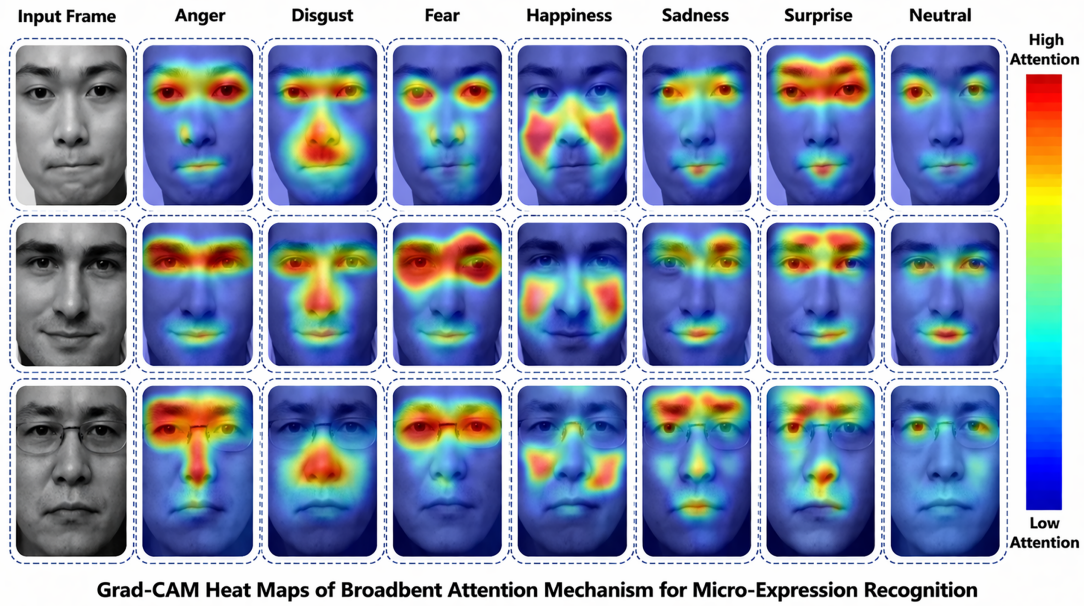


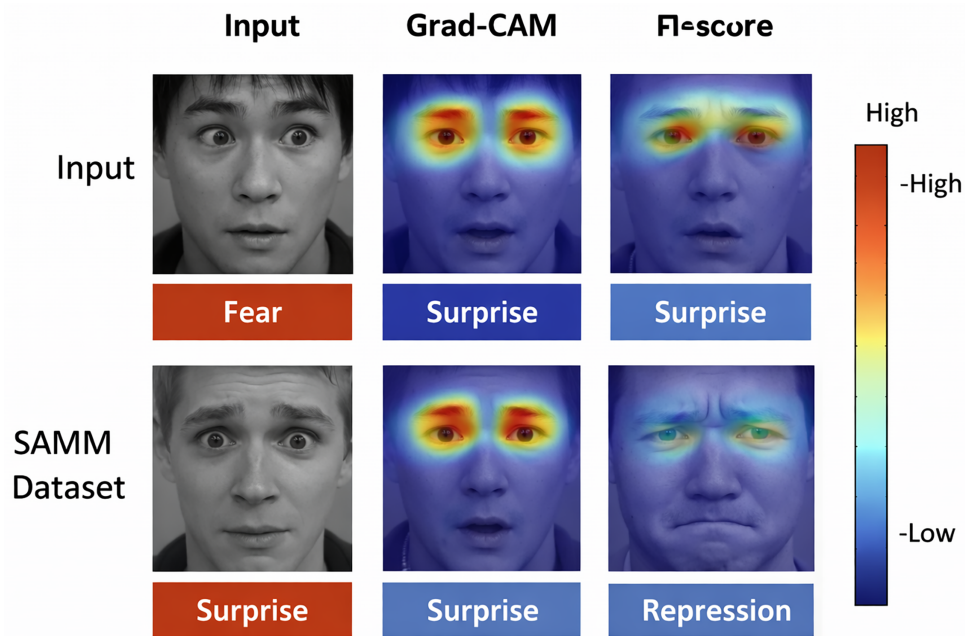
Figure 7: Attention map visualization using Grad-CAM showing regions of importance for micro-expression recognition. The top row shows input facial images, while the bottom row presents corresponding heatmaps, highlighting areas where the model concentrates for decision making.

4.5 Discussion

Although BroadAttNet achieves strong performance across SAMM, CASME II, and SMIC, several limitations remain. MER is inherently difficult because the relevant facial movements are weak, brief, and often visually overlapping. Classes such as *fear*, *surprise*, and *repression* may share similar facial muscle activations, which can lead to false positives or false negatives when the discriminative signal is not sufficiently pronounced.

Subject variability also remains a source of error. Differences in facial structure, lighting, head pose, and expression intensity can introduce noise into the feature extraction process. In addition, because this study uses extracted frames, temporal cues from onset, apex, and offset transitions are not fully modeled. Rare classes with limited samples can also reduce recall for underrepresented emotions. The Broadbent Attention mechanism mitigates these issues by improving spatial selectivity, but future work should incorporate temporal modeling, multimodal signals, improved data balancing, and cross-database validation to further strengthen robustness.

Performance differences across datasets can be explained by their distinct characteristics. SAMM provides high-quality recordings and comparatively clear class cues, contributing to the strongest performance. CASME II includes more subtle inter-class variation, while SMIC has fewer samples and broader class labels, both of which can make generalization more difficult. These observations reinforce the value of attention-based feature refinement for low-intensity facial expression analysis.



Per-class performance evaluation on the SAMM dataset.

Figure 8: Representative failure cases showing misclassified micro-expressions. Each example includes the input image, ground-truth label, predicted label, and corresponding Grad-CAM heatmap. The results highlight challenges in distinguishing subtle and overlapping expressions such as fear, surprise, and repression.

4.6 Benchmark SOTA Analysis

Table 10 compares BroadAttNet with representative MER methods across multiple datasets, including CASME II, SAMM, SMIC, and MEGC2019. The comparison includes conventional feature-based models, graph-based methods, capsule networks, transformer-inspired approaches, and recent attention-based methods. Because published results may differ in preprocessing, class grouping, and validation protocol, the table should be interpreted as a benchmark-level comparison rather than a strictly identical-protocol comparison. Within this context, BroadAttNet demonstrates highly competitive performance across all three evaluated datasets, indicating that selective attention is an effective strategy for enhancing micro-expression representation.

Table 10 shows that attention-aware approaches generally outperform models that do not explicitly focus on discriminative facial regions. BroadAttNet further strengthens this advantage through a lightweight Broadbent-inspired selective filtering mechanism. Its performance gains are supported by the ablation study, per-class results, and Grad-CAM visualizations, all of which indicate improved discrimination of subtle facial regions associated with micro-expressions.

Table 10: State-of-the-art comparison of micro-expression recognition methods. The reported accuracy for HTNet corresponds to results from the original study under its specific evaluation protocol and is not a direct average across datasets. Methods are categorized based on whether they incorporate attention mechanisms.

Model Architecture	Datasets	Accuracy	Precision	Recall	F1-Score
GTS-GN [20]	CASME II	87.21	87.5	86.2	84.8
	SAMM	72.06	72.9	71.4	70.9
MACNN [24]	CASME	73	74	75	73
	CASME II				
PCA, GMM, GLCM, SVM [31]	Emotion Dataset	93	94	92	93
ResNet50 + CapsField [36]	LFFW	89.1	89.3	89.7	88.9
	LFFC				
MER-GCN [38]	CASME II	58.82	57.3	57.5	58.2
	MEGC2019	74.6	74.8	75.3	75.3
IncepTR [42]	CASME II	89.6	89.9	88.4	91.1
	SMIC-HS	65.0	65.3	66.4	65.5
	SAMM	69.4	70.5	70.1	69.1
SRMCL [44]	SMIC	78.98	78.6	77.8	78.87
	CASME II	83.20	83.7	83.9	82.86
	SAMM	74.63	70.3	68.9	65.99
SSRLTS-ViT [48]	CASME II	76.3	76.7	75.4	74.6
	SAMM	71.6	72.3	69.9	71.6
MEMix [50]	CASME II	96.79	95.4	95.1	95.78
	SAMM	82.04	81.6	80.2	81.57
SDGSA [52]	SMIC-HS	76.9	76.3	75.9	75.7
	SAMM	73.4	73.9	72.7	74.2
	CASME II	75.6	76.8	75.1	76.9
	CASME II	81.0	81.22	81.24	81.31
	CASME III	83.40	84.42	82.64	84.71
HTNet (Baseline Method) [38]	SMIC	85.09	85.82	84.34	83.71
	SAMM	86.10	84.72	85.64	83.51
Broadbent Attention Layer	SAMM	99.66	99.73	99.66	99.73
	CASME II	98.73	97.65	98.73	97.65
	SMIC	98.62	97.09	98.62	97.09

Note: Bold values identify results from the proposed Broadbent Attention Layer; italics distinguish its validation protocols and dataset labels.

5 Conclusion

This study introduced BroadAttNet, an attention-driven framework for micro-expression recognition that integrates a Broadbent-inspired selective attention layer within a compact CNN architecture. The proposed attention mechanism improves the model's ability to focus on salient facial regions, including the eyes, eyebrows, cheeks, and mouth, where subtle micro-expression cues are most likely to occur. Experiments on SAMM, CASME II, and SMIC demonstrate strong performance under both K-Fold and Stratified K-Fold validation, with accuracies of 99.66% and 99.73% on SAMM, 98.73% and 97.65% on CASME II, and 98.62% and 97.09% on SMIC, respectively.

Beyond overall accuracy, the study provides a broader evaluation through ablation analysis, UAR and macro-F1 reporting, per-class performance, confusion matrices, runtime complexity, Grad-CAM visualization, and failure-case analysis. The findings show that BroadAttNet improves recognition performance while retaining efficient inference and interpretable attention behavior. Remaining challenges include visually overlapping emotion categories, class imbalance, subject variability, and the limited use of temporal information in frame-level modeling. Overall, BroadAttNet offers a robust and practical contribution to MER, with potential applications in human-computer interaction, behavioral analysis, learning analytics, emotion-aware systems, and psychological research.

Future Directions

Future work will extend BroadAttNet in several directions. First, temporal modeling can be incorporated to capture onset-apex-offset dynamics rather than relying only on extracted frames. Second, multimodal signals such as gaze, audio, or physiological measurements may improve recognition in complex real-world environments. Third, transfer learning, domain adaptation, and cross-database evaluation should be explored to improve robustness when the model is applied to unseen subjects, sensors, and cultural contexts. Finally, the Broadbent Attention layer should be evaluated across additional backbones, including ResNet, MobileNet, and transformer-based architectures, to determine its scalability and general applicability for real-time emotion-aware systems.

Acknowledgement: The authors would like to express their sincere gratitude to the Department of Computer Science & IT, Superior University, Lahore, Pakistan, for providing academic guidance, research facilities, and continuous support throughout this study.

Funding Statement: This work was supported by University of Zagreb, Faculty of Organization and Informatics, institutional project Digital Twins and Smart Solutions and institutional project The Role of Large Language Models in Enhancing Teaching and Learning Methods in Programming and Communication Disciplines, funded by the European Union—NextGenerationEU.

Author Contributions: Muhammad Zaman made the principal and sustained contribution to this research, including conceptualization, study design, methodology development, data acquisition, formal analysis, interpretation of results, implementation of the proposed framework, drafting of the original manuscript, and critical revision of the work for significant intellectual content. Hafiz Khizer bin Talib and Yanlong Cao provided research supervision, conceptual guidance, and critical review of the manuscript. Adnan Akhunzada contributed to data analysis, validation of experimental results, and manuscript review. Sharifah Sakinah Syed Ahmad provided methodological consultation and technical review. Nikola Ivkovic and Mario Konecki contributed through scientific evaluation, conceptualisation, validation of results, and critical revision of the manuscript. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are available from the corresponding author upon reasonable request.

Ethics Approval: The authors declare that this manuscript is an original work and has not been previously published or submitted to any other journal or conference for publication consideration.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Khizer Bin Talib H, Xu K, Cao Y, Xu Y, Xu Z, Zaman M, et al. Micro-expression recognition using convolutional variational attention transformer (ConVAT) with multihead attention mechanism. *IEEE Access*. 2025;13:20054–70. doi:10.1109/access.2025.3530114.
2. Zeng X, Zhao X, Zhong X, Liu G. A survey of micro-expression recognition methods based on LBP, optical flow and deep learning. *Neural Process Lett*. 2023;55(5):5995–6026. doi:10.1007/s11063-022-11123-x.
3. Pahwa R, Gupta V. Detection of microexpressions and subtle emotions. In: *Emotion and facial recognition in artificial intelligence: sustainable multidisciplinary perspectives and applications*. Cham, Switzerland: Springer; 2026. p. 175–205.
4. Gilanie G, Cheema S, Latif A, Saher A, Ahsan M, Ullah H, et al. A robust method of bipolar mental illness detection from facial micro expressions using machine learning methods. *Intell Autom Soft Comput*. 2024;39(1):57–71. doi:10.32604/iasc.2024.041535.
5. Khare SK, Blanes-Vidal V, Nadimi ES, Acharya UR. Emotion recognition and artificial intelligence: a systematic review (2014–2023) and research recommendations. *Inf Fusion*. 2024;102(3):102019. doi:10.1016/j.inffus.2023.102019.
6. Wang Y, Lin Z, Teng Y, Cheng Y, Jiang H, Yang Y. SIMMA: multimodal automatic depression detection via spatiotemporal ensemble and cross-modal alignment. *IEEE Trans Comput Soc Syst*. 2025;12(5):3548–64. doi:10.1109/tcss.2025.3542986.
7. Esmaeili V, Mohassel Feghhi M, Shahdi SO. A comprehensive survey on facial micro-expression: approaches and databases. *Multimed Tools Appl*. 2022;81(28):40089–134. doi:10.1007/s11042-022-13133-2.
8. Liu KH, Jin QS, Xu HC, Gan YS, Liong ST. Micro-expression recognition using advanced genetic algorithm. *Signal Process Image Commun*. 2021;93(3):116153. doi:10.1016/j.image.2021.116153.
9. Dou J, Song Y, Yu H. Hierarchical oversampling based on Cohen's criterion for imbalanced data with missing information. *IEEE Trans Comput Soc Syst*. 2025;12(5):3143–55. doi:10.1109/tcss.2025.3548874.
10. Wang S, Zhao X, Zeng X, Xie J, Luo Y, Chen J, et al. Micro-expression recognition based on EEG signals. *Biomed Signal Process Control*. 2023;86(5):105037. doi:10.1016/j.bspc.2023.105037.
11. Xu K, Cao Y, Zaman M, Toqeer MH, Akhunzada A. Error-corrected quantum CNN for gender-based micro-expression analysis in affective computing. In: *Proceedings of the 2025 7th International Congress on Human-Computer Interaction, Optimization and Robotic Applications (ICHORA)*; 2025 May 23–24; Ankara, Turkiye. New York, NY, USA: IEEE; 2025. p. 1–7.
12. Zhao S, Tao H, Zhang Y, Xu T, Zhang K, Hao Z, et al. A two-stage 3D CNN based learning method for spontaneous micro-expression recognition. *Neurocomputing*. 2021;448:276–89. doi:10.1016/j.neucom.2021.03.058.
13. Liu Y, Li Y, Yi X, Hu Z, Zhang H, Liu Y. Lightweight ViT model for micro-expression recognition enhanced by transfer learning. *Front Neurorobot*. 2022;16:922761. doi:10.3389/fnbot.2022.922761.
14. Geetha A, Mala T, Priyanka D, Uma E. Multimodal emotion recognition with deep learning: advancements, challenges, and future directions. *Inf Fusion*. 2024;105(2):102218. doi:10.1016/j.inffus.2023.102218.
15. Yu Z, Chen X, Qu C. SDGSA: a lightweight shallow dual-group symmetric attention network for micro-expression recognition. *Complex Intell Syst*. 2024;10(6):8143–62. doi:10.1007/s40747-024-01594-x.
16. Talib HKB, Cao Y, Zaman M, Xu K, Akhunzada A. SepViT: a dual-path transformer-convolution framework for apex frame-based microexpression recognition. *IEEE Trans Comput Soc Syst*. 2026;13(3):3912–27. doi:10.1109/tcss.2026.3653634.
17. Zhou L, Mao Q, Huang X, Zhang F, Zhang Z. Feature refinement: an expression-specific feature learning and fusion method for micro-expression recognition. *Pattern Recognit*. 2022;122(1):108275. doi:10.1016/j.patcog.2021.108275.

18. Zhang L, Zhang Y, Sun X, Tang W, Wang X, Li Z. Micro-expression recognition based on direct learning of graph structure. *Neurocomputing*. 2025;619(11):129135. doi:10.1016/j.neucom.2024.129135.
19. Yee NL, Zulkifley MA, Saputro AH, Abdani SR. Apex frame spotting using attention networks for micro-expression recognition system. *Comput Mater Contin*. 2022;73(3):5331–48. doi:10.32604/cmc.2022.028801.
20. Wang X, Shen X. Micro-expression recognition based on apex frame using deep learning. In: *Proceedings of the 2023 19th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*; 2023 Jul 29–31; Harbin, China. New York, NY, USA: IEEE; 2023. p. 1–6.
21. Vice J, Khan MM, Tan T, Murray I, Yanushkevich S. A hierarchical separation and classification network for dynamic microexpression classification. *IEEE Trans Comput Soc Syst*. 2024;11(3):3766–79. doi:10.1109/tcss.2023.3334823.
22. Gupta P. MERASTC: micro-expression recognition using effective feature encodings and 2D convolutional neural network. *IEEE Trans Affective Comput*. 2023;14(2):1431–41. doi:10.1109/taffc.2021.3061967.
23. Yang H, Sun S, Chen J. Deep learning-based micro-expression recognition algorithm research. *Int J Comput Sci Inf Technol*. 2024;2(1):59–70.
24. Yang J, Wu Z, Wu R. Micro-expression recognition based on contextual transformer networks. *Visual Comput*. 2025;41(3):1527–41. doi:10.1007/s00371-024-03443-x.
25. Krause FC, Linardatos E, Fresco DM, Moore MT. Facial emotion recognition in major depressive disorder: a meta-analytic review. *J Affect Disord*. 2021;293(5):320–8. doi:10.1016/j.jad.2021.06.053.
26. Döllinger L, Laukka P, Högman LB, Bänziger T, Makower I, Fischer H, et al. Training emotion recognition accuracy: results for multimodal expressions and facial micro expressions. *Front Psychol*. 2021;12:708867. doi:10.3389/fpsyg.2021.708867.
27. Karnati M, Seal A, Bhattacharjee D, Yazidi A, Krejcar O. Understanding deep learning techniques for recognition of human emotions using facial expressions: a comprehensive survey. *IEEE Trans Instrum Meas*. 2023;72(1):1–31. doi:10.1109/tim.2023.3243661.
28. Agung ES, Rifai AP, Wijayanto T. Image-based facial emotion recognition using convolutional neural network on emognition dataset. *Sci Rep*. 2024;14(1):14429. doi:10.1038/s41598-024-65276-x.
29. Hossain S, Umer S, Rout RK, Tanveer M. Fine-grained image analysis for facial expression recognition using deep convolutional neural networks with bilinear pooling. *Appl Soft Comput*. 2023;134(5):109997. doi:10.1016/j.asoc.2023.109997.
30. Wang R, Song P, Zheng W. Graph-diffusion-based domain-invariant representation learning for cross-domain facial expression recognition. *IEEE Trans Comput Soc Syst*. 2024;11(3):4163–74. doi:10.1109/tcss.2024.3355113.
31. Ashraf I, Kehkashan T, Zaman M, Akhunzada A, Talib HK, Cao Y. Enhancing micro-expression recognition with broadbent attention mechanism: a high-performance approach for emotion detection. In: *Proceedings of the 2024 IEEE 21st International Conference on Smart Communities: Improving Quality of Life using AI, Robotics and IoT (HONET)*; 2024 Dec 3–5; Doha, Qatar. New York, NY, USA: IEEE; 2024. p. 189–94.
32. Matharaarachchi N, Fermi Pasha M. A dual stream spatio-temporal deep network for micro-expression recognition using upper facial features. *Neural Comput Appl*. 2025;37(3):1271–87. doi:10.1007/s00521-024-10374-0.
33. Zhang H, Yin L, Zhang H, Wu X. Facial micro-expression recognition using three-stream vision transformer network with sparse sampling and relabeling. *Signal Image Video Process*. 2024;18(4):3761–71. doi:10.1007/s11760-024-03039-x.
34. Zheng Y, Blasch E. Facial micro-expression recognition enhanced by score fusion and a hybrid model from convolutional LSTM and vision transformer. *Sensors*. 2023;23(12):5650. doi:10.3390/s23125650.
35. Tang Y, Yi J, Tan F. Facial micro-expression recognition method based on CNN and transformer mixed model. *Int J Biom*. 2024;16(5):463–77. doi:10.1504/ijbm.2024.140771.
36. He E, Chen Q, Zhong Q. SL-swin: a transformer-based deep learning approach for macro- and micro-expression spotting on small-size expression datasets. *Electronics*. 2023;12(12):2656. doi:10.3390/electronics12122656.
37. Wang R, Song P, Li S, Ji L, Zheng W. Common Latent embedding space for cross-domain facial expression recognition. *IEEE Trans Comput Soc Syst*. 2024;11(2):2046–56. doi:10.1109/tcss.2023.3276990.

38. Xie M, Chen D, Chen W, Zhang K, Liu W, Liu J, et al. Fast HTNet: leveraging subtle movements for efficient micro-expression recognition. In: Proceedings of the 2024 5th International Conference on Computers and Artificial Intelligence Technology (CAIT); 2024 Dec 20–22; Hangzhou, China. New York, NY, USA: IEEE; 2024. p. 88–93.
39. Kumar RAJ, Bhanu B. Micro-expression classification based on landmark relations with graph attention convolutional network. In: Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); 2021 Jun 19–25; Nashville, TN, USA. p. 1511–20.
40. Shuai T, Beng S, Khalid FB, Rahmat RWBO. Advances in facial micro-expression detection and recognition: a comprehensive review. *Information*. 2025;16(10):876. doi:10.3390/info16100876.
41. Wang Q. Micro-expression recognition method based on CNN-LSTM hybrid network. *Int J Wirel Mob Comput*. 2022;23(1):67. doi:10.1504/ijwmc.2022.125537.
42. Xie T, Sun G, Sun H, Lin Q, Ben X. Decoupling facial motion features and identity features for micro-expression recognition. *PeerJ Comput Sci*. 2022;8:e1140.
43. Pan J, Fang W, Zhang Z, Chen B, Zhang Z, Wang S. Multimodal emotion recognition based on facial expressions, speech, and EEG. *IEEE Open J Eng Med Biol*. 2024;5(10):396–403. doi:10.1109/ojemb.2023.3240280.
44. Saisanthiya D, Supraja P. Neuro-facial fusion for emotion AI: improved federated learning GAN for collaborative multimodal emotion recognition. *IEIE Trans Smart Process Comput*. 2024;13(1):61–8. doi:10.5573/ieiespc.2024.13.1.61.
45. Leng W, Zhao S, Zhang Y, Liu S, Mao X, Wang H, et al. ABPN: apex and boundary perception network for micro- and macro-expression spotting. In: Proceedings of the 30th ACM International Conference on Multimedia; 2022 Oct 10–14; Lisbon, Portugal. p. 7160–4.
46. Wang Z, Yang M, Jiao Q, Xu L, Han B, Li Y, et al. Two-level spatio-temporal feature fused two-stream network for micro-expression recognition. *Sensors*. 2024;24(5):1574. doi:10.3390/s24051574.
47. Zhu X, Ye S, Zhao L, Dai Z. Hybrid attention cascade network for facial expression recognition. *Sensors*. 2021;21(6):2003. doi:10.3390/s21062003.
48. Wang Z, Zhang K, Luo W, Sankaranarayana R. HTNet for micro-expression recognition. *Neurocomputing*. 2024;602(4):128196. doi:10.1016/j.neucom.2024.128196.
49. Li J, Nie J, Guo D, Hong R, Wang M. Emotion separation and recognition from a facial expression by generating the poker face with vision transformers. *IEEE Trans Comput Soc Syst*. 2025;12(4):1548–62. doi:10.1109/tcss.2024.3478839.
50. Zhou J, Liu X, Wang H, Zhang Z, Chen T, Fu X, et al. Seeing through the mask: recognition of genuine emotion through masked facial expression. *IEEE Trans Comput Soc Syst*. 2024;11(6):7159–72. doi:10.1109/tcss.2024.3404611.
51. Wang SJ, Miao YH, Li J, Zhou L, Dong Z, Sun M, et al. Micro-expression key frame inference. *IEEE Trans Affect Comput*. 2025;16(3):1955–69. doi:10.1109/taffc.2025.3548284.
52. Li M, Wang Y, Yang C, Lu Z, Chen J. Automatic diagnosis of depression based on facial expression information and deep convolutional neural network. *IEEE Trans Comput Soc Syst*. 2024;11(5):5728–39. doi:10.1109/tcss.2024.3393247.
53. Zhang L, Hong X, Arandjelovic O, Zhao G. Short and long range relation based spatio-temporal transformer for micro-expression recognition. *IEEE Trans Affective Comput*. 2022;13(4):1973–85. doi:10.1109/taffc.2022.3213509.
54. Shang Z, Wang P, Li X. Micro-expression recognition based on differential feature fusion. *Multimed Tools Appl*. 2024;83(4):11111–26.
55. Gong W, Zhang Y, Wang W, Cheng P, González J. Meta-MMFNet: meta-learning-based multi-model fusion network for micro-expression recognition. *ACM Trans Multimedia Comput Commun Appl*. 2024;20(2):1–20. doi:10.1145/3539576.
56. Li Y, Wei J, Liu Y, Kauttonen J, Zhao G. Deep learning for micro-expression recognition: a survey. *IEEE Trans Affective Comput*. 2022;13(4):2028–46. doi:10.1109/taffc.2022.3205170.
57. Choirina P, Rosiani UD, Fitriani IM, Baqi R. Facial micro expression recognition for feature point tracking using apex frames on CASME II database. *Sinkron*. 2023;8(1):81–9. doi:10.33395/sinkron.v8i1.11946.
58. Tan S, Wu Y. Cross-database micro expression recognition based on apex frame. In: *Parallel Architectures, Algorithms and Programming: 11th International Symposium*; 2020 Dec 28–30; Shenzhen, China.