Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.062621

ARTICLE

A Semi-Lightweight Multi-Feature Integration Architecture for Micro-Expression Recognition

Mengqi Li, Xiaodong Huang* and Lifeng Wu

Information Engineering College, Capital Normal University, Beijing, 100048, China

*Corresponding Author: Xiaodong Huang. Email: hxd@cnu.edu.cn

Received: 23 December 2024; Accepted: 02 April 2025; Published: 09 June 2025

ABSTRACT: Micro-expressions, fleeting involuntary facial cues lasting under half a second, reveal genuine emotions and are valuable in clinical diagnosis and psychotherapy. Real-time recognition on resource-constrained embedded devices remains challenging, as current methods struggle to balance performance and efficiency. This study introduces a semi-lightweight multifunctional network that enhances real-time deployment and accuracy. Unlike prior simplistic feature fusion techniques, our novel multi-feature fusion strategy leverages temporal, spatial, and differential features to better capture dynamic changes. Enhanced by Residual Network (ResNet) architecture with channel and spatial attention mechanisms, the model improves feature representation while maintaining a lightweight design. Evaluations on SMIC, CASME II, SAMM, and their composite dataset show superior performance in Unweighted F1 Score (UF1) and Unweighted Average Recall (UAR), alongside faster detection speeds compared to existing algorithms.

KEYWORDS: Micro-expressions; DynamicFusionResNet (DFR-Net); feature fusion; attention mechanism

1 Introduction

Facial micro-expressions are involuntary, spontaneous facial muscle movements that typically last no more than half a second [1], occurring when individuals try to suppress or mask emotions, revealing underlying feelings. Accurate recognition of micro-expressions has significant applications in fields like various domains criminal investigations [2], clinical diagnostics [3], and social interactions [4], improving emotion recognition and decision-making. Micro-expression recognition can be utilized to detect subtle emotional cues that patients may conceal during psychological assessments, particularly in the early diagnosis of emotional disorders. By identifying these hidden negative emotions, clinicians can make more accurate assessments and provide timely interventions. Additionally, micro-expression recognition enables therapists to monitor patients' emotional responses in real-time during therapy sessions. This technology can help evaluate the impact of different treatment methods on patients' emotions, allowing therapists to adjust their strategies accordingly for better therapeutic outcomes.

In recent years, automatic micro-expression recognition has emerged as a prominent research area. Early studies on automatic micro-expression recognition employed methods based on Local Binary Patterns (LBP) [5] or optical flow [6] for feature extraction, followed by classification using traditional machine learning techniques such as Support Vector Machines (SVM) [7] or Random Forests (RF) [8].

With the advent and development of deep learning, researchers have increasingly turned to Convolutional Neural Networks (CNNs) for micro-expression recognition [9], achieving significant improvements

in system performance. More recently, large-scale Transformer-based models [10] have also been applied to micro-expression recognition, effectively capturing long-range dependencies and global contextual information in facial expressions.

However, most models excessively prioritize recognition accuracy, resulting in a large number of parameters and, consequently, slower inference speeds. To address this, we propose a lightweight, high-performance model suitable for real-time applications, such as psychological diagnostics. Our approach features a novel multi-feature fusion strategy that dynamically combines temporal, spatial, and differential features, enhancing recognition performance.

As depicted in Fig. 1, we present a novel tri-branch model for micro-expression identification, aiming to utilize features from three different modalities. This paradigm is concretely realized through a network architecture termed DynamicFusionResNet (DFR-Net). Initially, shallow features are extracted independently by each branch. The temporal branch is responsible for capturing dynamic variations over time, while the spatial branch focuses on extracting static information from individual frames. In parallel, the differential branch aims to detect subtle changes, which are essential for discerning micro-expressions.

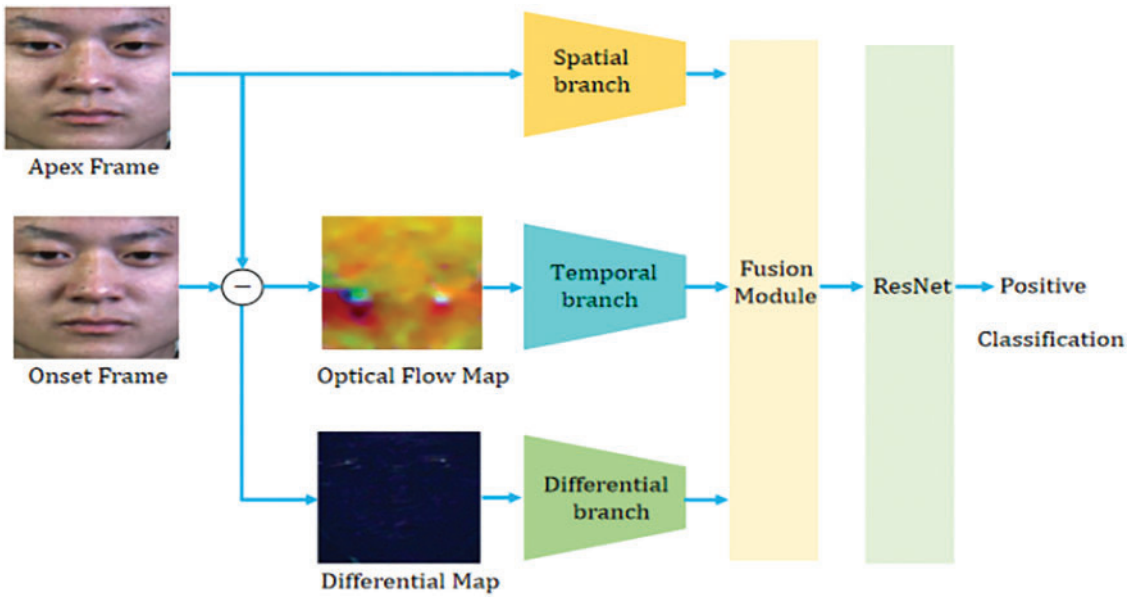

Figure 1: The proposed three-branch micro-expressions recognition paradigm

Following the feature extraction stage, we employ a carefully designed multi-feature fusion strategy. This fusion allows the model to prioritize critical expression variations across different spatial regions and adjust its attention to various feature modalities. Subsequently, the fused features are passed through a modified ResNet [11], which incorporates channel-spatial attention mechanisms. This augmentation is specifically intended to enhance the recognition accuracy by enabling the model to focus more effectively on informative regions and features within the micro-expression data.

The contributions of our work are summarized as follows:

- We propose a novel tri-branch paradigm for micro-expression recognition that leverages three distinct modalities: temporal, spatial, and differential features. These features are fused early in the network to enable joint learning and interaction across modalities.

- We design a tailored multi-feature fusion strategy that adaptively integrates the different modalities based on their inherent characteristics. This approach enables the model to effectively capture the importance of expression variations across different spatial regions, while also adjusting the model's attention to specific feature modalities.
- We evaluated various modules, assessing their parameter count and inference speed, and ultimately selected our proposed channel-spatial attention ResNet to process the fused feature maps. This module is specifically designed to identify and focus on the most informative regions within the micro-expression feature maps, striking a balance between efficiency and accuracy in capturing subtle micro-expressions.
- Our DFR-Net demonstrates competitive performance on three widely-used micro-expression datasets (CASME II, SAMM, and SMIC), as well as on their combined dataset. Our model significantly reduces the number of parameters and achieves faster convergence compared to state-of-the-art models.

2 Related Works

Traditional amicro-expression recognition algorithms typically rely on handcrafted feature extraction followed by classification using machine learning models. These techniques can be generally classified into texture-based and motion-based strategies. Texture-based methods, such as Local Binary Pattern (LBP) [6], which generates binary sequences by comparing pixel values with their surrounding neighborhood, thereby capturing texture information within images. In contrast, motion-based methods depend on optical flow techniques, which obtain motion information by describing the displacement of pixels between adjacent frames. Notable texture-based methods include LBP-TOP [12], LBP-SIP [13], STLBP-IP [14], LBP-MOP [15], and DiSTLBP-RIP [16], whereas optical flow-based algorithms comprise MDMO [17], FDM [18], ALSTP [19], BI-WOOF [20], and FHOFO [21].

As deep learning continues to evolve, Convolutional Neural Networks (CNNs) have become the go-to approach for micro-expression recognition, showcasing remarkable performance gains [9]. For example, Gan et al. [22] leveraged optical flow data between the starting and peak frames, feeding it into a custom CNN model to extract and classify features. Quang et al. [23] took a different route by employing Capsule Networks (CapsuleNet), where features were first pulled from a pre-trained Residual Network (ResNet) [11] and then classified using a primary capsule layer and dynamic routing. Liong et al. [24] crafted a three-path shallow neural network that taps into horizontal optical flow, vertical optical flow, and optical strain features to boost both accuracy and model expressiveness. Xia et al. [25] introduced a method based on Recursive Convolutional Networks (RCN), employing a shallower network structure and low-resolution input data to reduce model complexity, while incorporating modules that do not require additional parameters to enhance feature representation. Zhou et al. [26] proposed a micro-expression prediction method based on attention mechanisms, which integrates expression-specific feature learning and attention extraction modules, demonstrating excellent performance in micro-expression recognition tasks. Lei et al. [27] upped the ante by integrating facial graph representation learning and graph convolutional networks (GCN) with action unit information, ultimately proposing a fusion model to sharpen recognition performance. Zhao et al. [28] proposed an apex frame detection method based on Unimodal Pattern Constrained (UPC), combined with a local attention module, and introduced an end-to-end training approach for the ME-PLAN framework. Verma et al. [29] introduced a refined hybrid module that integrates mixed spatiotemporal operations with an optimal path exploration network to design a lightweight architecture. Similarly, Wei et al. [30] combined both elementary and advanced geometric movement data to discern distinctive characteristics, while employing a self-learning approach to dynamically model inter-node relationships and strengthen the correspondence between facial landmarks in micro-expressions.

In the field of micro-expression recognition, the application of Transformer models is also increasing. For example, Zhao et al. [31] introduced a dual-path neural network that utilizes a pre-trained Swin Transformer as a feature extractor, enhancing computational efficiency through local window-based self-attention mechanisms. Zhai et al. [32] designed a framework combining self-supervised learning, proposing three Transformer-based feature fusion modules for extracting multi-level information features. Wang et al. [33] developed a Hierarchical Transformer Network (HTNet) that partitions facial regions into distinct parts and utilizes local self-attention mechanisms to capture subtle movements in each region. Wang et al. [34] addressed the issue of unilateral local movements by devising a strategy that partitions facial information into global and local regions, thereby achieving high-precision micro-expression recognition. Bao et al. [35] developed a framework that combines self-expression reconstruction with memory contrastive learning; they introduced a supervised prototype-based memory contrastive learning module to mine discriminative features. Zhang et al. [36] proposed a hierarchical feature aggregation network that employs a multi-scale attention module to capture subtle local variations while also establishing global dependencies.

In addition to above-mentioned approaches, other methodologies in the micro-expression domain feature a temporal augmentation technique proposed by Wang et al. [37] to mitigate the challenge of limited data through pre-training, and a dual-branch meta-auxiliary learning framework. Wang et al. [38] utilized a primary task branch to learn micro-expression features alongside an auxiliary task to enhance the extraction of discriminative features.

Inspired by previous research, this study aims to design a relatively lightweight network by improving feature selection and fusion modules, while fully leveraging the characteristics of different modalities. We propose a three-dimensional feature set consisting of temporal features extracted from optical flow, spatial features derived from the apex frame, and differential features computed between the onset and apex frames. Additionally, we introduce a simple yet efficient feature fusion strategy that combines convolutional layers with attention mechanisms to effectively exploit the complementarity of different features, thereby enhancing feature richness and model robustness. Finally, the fused features are input into a channel-space attention-optimized ResNet for classification. Experimental results demonstrate the effectiveness of this module.

3 Method

As illustrated in Fig. 2, our proposed DynamicFusion-ResNet (DFR-Net) architecture comprises three primary branches: a Spatial branch, a Temporal branch, and a Differential branch. The Spatial branch extracts static information from high-resolution apex frame images, capturing detailed spatial cues. The Temporal branch is designed to extract dynamic features by analyzing motion variations between the apex and onset frames, leveraging optical flow to capture temporal changes. The Differential branch focuses on extracting subtle features by analyzing static changes between the apex and onset frames. Section 3.1 elaborates on the data preprocessing steps, including apex frame acquisition, optical flow computation, and the extraction of differential features. In Section 3.2, we introduce a novel fusion strategy that integrates the three types of features. Finally, Section 3.3 details the learning process of the fused features using a ResNet architecture enhanced with spatiotemporal attention mechanisms, aiming to obtain a comprehensive representation for micro-expression recognition.

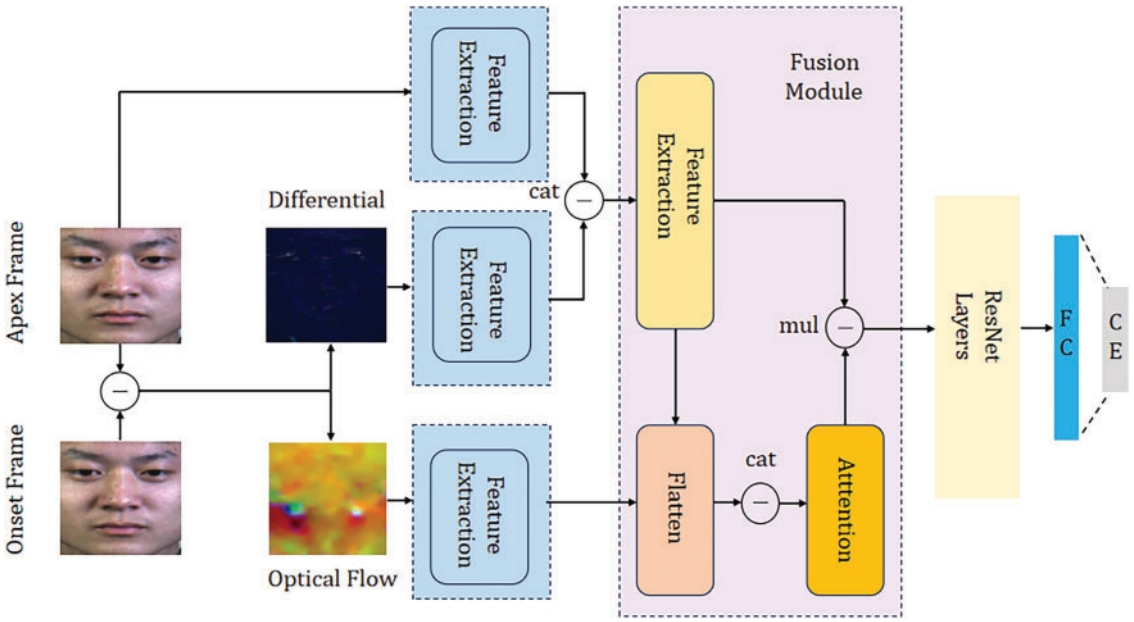

Figure 2: The overall pipeline of our proposed DynamicFusionResNet (DFR-Net) for micro-expression recognition. The network architecture comprises three main branches: the spatial branch, the temporal branch, and the differential branch. Each branch performs initial feature extraction at a shallow level before the features are fused. Subsequently, the fused features are processed through a ResNet framework for further learning. CE denotes the cross-entropy loss function

3.1 Data Preprocessing

3.1.1 Apex Frame Acquisition

Pinpointing the apex frame with precision is a cornerstone of micro-expression analysis, as it enables the extraction of optical flow features that capture the most pronounced facial muscle activity. Since the SMIC database [39] lacks ground-truth labels for apex frames, an automated detection system becomes indispensable. In this research, we employ the Divide and Conquer based on Regions of Interest (D&C-RoIs) technique [40] to autonomously determine the apex frame index in video sequences. This method has gained traction in recent micro-expression studies for its consistent reliability in apex frame identification, which in turn bolsters the accuracy of micro-expression detection.

The D&C-RoIs method begins by computing Local Binary Pattern (LBP) features [41] for each frame within three key facial sub-regions, namely the left eye and eyebrow, right eye and eyebrow, and the mouth. Following this, a correlation coefficient is calculated to quantify the variation in LBP features between the onset frame and subsequent frames. This variation is mathematically expressed as:

$$d = \frac{\sum_{i=1}^{B} h_{1i} \times h_{2i}}{\sqrt{\sum_{i=1}^{B} h_{1i}^2 \times \sum_{i=1}^{B} h_{2i}^2}}, \tag{1}$$

where B represents the number of bins in the LBP histograms, h_1 denotes the histogram of the onset frame, and h_2 corresponds to the histograms of the other frames. The frame that exhibits the highest variation in the Regions of Interest (RoIs) is selected as the apex frame, corresponding to the peak facial muscle activity.

To refine the apex frame identification, the D&C strategy is applied to the rate of feature variation, effectively searching for the frame index that corresponds to the local maximum. This systematic process guarantees the dependable identification of the peak frame, essential for obtaining the optical flow data pivotal in micro-expression analysis.

Let $S = [s_1, s_2, \ldots, s_n]$ denote a set of n micro-expression video clips, where the ith sample video clip s_i is represented as:

$$s_i = \left\{ f_{i,j} \mid i = 1, \ldots, n; j = 1, \ldots, F_i \right\}, \tag{2}$$

here, F_i is the total number of image frames in the ith sequence. Each sequence has a solitary apex frame, labeled as $f_{i,a}$, which can be located at any frame index between the onset (first frame) $f_{i,1}$ and offset (last frame) $f_{i,n}$. Thus, the apex frame can be formally expressed as:

$$f_{i,\alpha} \in \{ f_{i,1}, \ldots, f_{i,F_i} \}, \tag{3}$$

the D&C-RoIs approach is employed to predict the apex frame $f_{i,a}$ for each video sequence within the SMIC database, thereby facilitating the subsequent tasks involved in micro-expression recognition.

3.1.2 Feature Extraction

To calculate the horizontal and vertical optical flow components, vectors are derived between the onset and apex frames. These optical flow vectors serve as an essential tool for describing motion displacements within facial regions and have shown significant promise in micro-expression recognition tasks. The assumption that image brightness remains constant between consecutive frames gives rise to the following equation:

$$I(x, y, t) = I(x + \delta x, y + \delta y, t + \Delta t), \tag{4}$$

where x, y represent pixel coordinates and t denotes time. By applying a Taylor series expansion, the optical flow constraint equation can be expressed as:

$$\frac{\partial I}{\partial x} u(x, y) + \frac{\partial I}{\partial y} v(x, y) + \frac{\partial I}{\partial t} = 0, \tag{5}$$

in this equation, $u(x, y)$ and $v(x, y)$ represent the horizontal and vertical components of the optical flow feature map, respectively. The optical flow feature map can be formulated as follows:

$$V = \{ (u(x, y), v(x, y)) \mid x = 1, 2, \ldots, X; y = 1, 2, \ldots, Y \}, \tag{6}$$

where X and Y represent the frame's width W and height H, respectively. The optical flow feature map $V = \left[V_x, V_y \right]$, and $V \in \mathbb{R}^{W \times H \times 2}$.

Our method entails calculating the first-order derivatives of the optical flow field, a process termed optical strain, for assessing variations. Optical strain essentially gives us a measure of how much the face is moving, which is super useful for understanding the nuanced muscle movements that define micro-expressions. By crunching the optical strain numbers, we can get a grip on the fine details of facial motion, a key factor in pinpointing micro-expressions accurately. The specific formula is as follows:

$$V_z = \sqrt{\left(\frac{\partial V_x}{\partial x} \right)^2 + \left(\frac{\partial V_y}{\partial y} \right)^2 + \frac{1}{2} \left(\left(\frac{\partial V_x}{\partial y} \right)^2 + \left(\frac{\partial V_y}{\partial x} \right)^2 \right)}, \tag{7}$$

finally, three-dimensional optical flow feature maps are formed and represented as $V_m = \left[V_x, V_y, V_z\right]$ and $V_m \in \mathbb{R}^{W \times H \times 3}$.

In addition, we compute the difference between the apex frame and the onset frame to capture subtle variations that occur between these two key frames. This difference reflects the nuanced changes in facial expressions and is mathematically expressed as follows:

$$V_z = D(x, y) = |I_{\text{apex}}(x, y) - I_{\text{onset}}(x, y)|, \tag{8}$$

where $D(x, y)$ represents the difference map, $I_{\text{apex}}(x, y)$ is the pixel intensity at coordinates (x, y) in the apex frame, and $I_{\text{onset}}(x, y)$ is the corresponding pixel intensity in the onset frame.

We employ three distinct feature inputs to our network architecture: optical flow maps derived from the temporal branch, the apex frame as the spatial branch, and the absolute difference map as the differential branch. The optical flow map, generated between the onset and apex frames, effectively encodes motion patterns that occur over time, reflecting the dynamic changes in facial muscle movement. The apex frame serves as the spatial representation, capturing fine-grained details of the facial expression at its peak intensity. Finally, the absolute difference map highlights the localized changes by emphasizing the differences between the onset and apex frames, thereby providing a focused input that captures the subtle variations critical for micro-expression recognition.

3.2 Multi-Feature Fusion Strategy

Inspired by Gan et al. [22], who proposed an innovative descriptor merging the context derived from optical flow with convolutional neural networks (CNNs) and drawing insights from Liong et al. [24], whose design can extract discriminative high level features from three optical flow features, we propose a multi-branch fusion strategy to carry out the task of micro-expression recognition. It aimed at effectively leveraging information from different feature maps, including optical flow images, static images, and difference images. Since these three feature maps capture various dimensions of expression changes, it is necessary to conduct appropriate pre-processing so as to ensure effective information integration.

As illustrated in Fig. 3, we first perform preliminary shallow feature extraction on the spatial feature map (static images) and the difference feature map. The spatial feature map captures static details of expressions, such as facial texture and geometry, while the difference feature map reflects subtle changes between the apex and onset frames, particularly those localized changes inherent in micro-expressions. For the shallow feature extraction, we apply convolutional layers with 3 × 3 kernels and a stride of 1, followed by ReLU activation, to capture low-level details such as edges and textures in both feature maps.

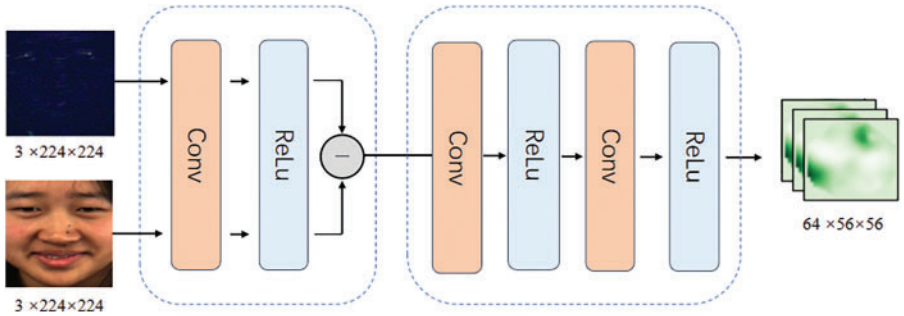

Figure 3: Spatial and differential features, after undergoing shallow feature extraction, are concatenated along the channel dimension, followed by the extraction of deep-level features

To achieve effective fusion between these two feature maps, we concatenate the spatial feature map where F_s and the difference feature map F_d along the channel dimension. This concatenation ensures that both the static details and subtle local changes are preserved, providing a rich feature representation for subsequent layers. The concatenated feature map is further processed through a fusion network, consisting of two convolutional layers and ReLU activation functions, to extract deep fused features, forming the fused feature map F_{fused}. The formula is as follows:

$$F_{fused} = \sigma\left(f_2\left(\sigma\left(f_1\left([F_s, F_d]\right)\right)\right)\right), \tag{9}$$

where f_1 and f_2 denote the convolution kernels, $[F_s, F_d]$ represents concatenation along the channel dimension, and σ is the ReLU activation function.

As shown in Fig. 4, after fusing the spatial and difference feature maps, we introduce the temporal feature map (optical flow images) for further fusion. The optical flow feature map captures motion patterns between the apex and onset frames, effectively reflecting the dynamic changes in facial muscles. First, we perform shallow feature extraction using the same approach. Then, we perform adaptive average pooling on the fused and temporal feature maps to reduce them to 1×1 feature maps, effectively aggregating global information. Next, these reduced feature maps are concatenated along the channel dimension and processed through two convolutional layers and ReLU activation functions to obtain an attention weight map. The formula is as follows:

$$F_{final} = F_{fused} \cdot A, \tag{10}$$

where $\cdot$ denotes element-wise multiplication.

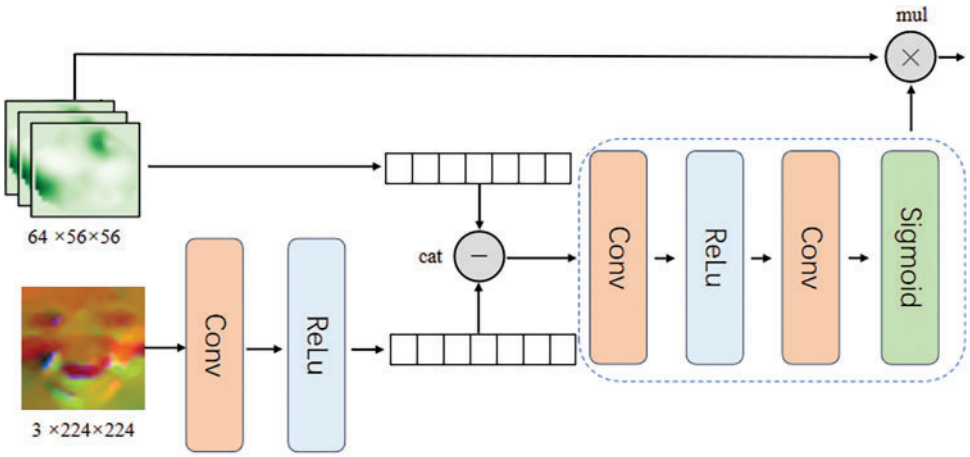

Figure 4: The optical flow images first undergo shallow feature extraction. Subsequently, the optical flow images are flattened and merged with the spatial-differential feature maps to obtain attention weights, which are then multiplied with the original spatial-differential feature maps to produce the fused image

Through the above multi-branch fusion strategy, the network can effectively integrate useful information from different feature maps. The spatial feature map provides static details of the face, the difference feature map highlights subtle changes, and the temporal feature map captures dynamic motion information. By rationally integrating these features, the information contained in these feature maps is effectively

consolidated, providing a more comprehensive and accurate feature representation for the task of micro-expression recognition. Moreover, the attention mechanism-based fusion strategy enhances the network's sensitivity to temporal information, thereby improving recognition accuracy.

3.3 Channel-Spatial Attention ResNet

Upon the completion of the fusion of spatial feature maps with differential feature maps, as well as their integration with temporal feature maps, the resulting fused feature maps are subjected to further feature extraction and classification through a standard ResNet [11] model. The architecture proposed in this paper leverages a ResNet-based network, augmented with an attention mechanism, to enhance the learning of salient features. ResNet is a critically important neural network architecture in the domain to deep learning. By introducing residual blocks, it effectively mitigates the issues of vanishing and exploding gradients that typically occur during the training of deep neural networks. In tasks involving multimodal feature fusion, solely relying on residual connections may be insufficient to capture the intricate details of the image features. To address this, we have incorporated Channel Attention (CA) and Spatial Attention (SA) mechanisms into the residual blocks of ResNet, aiming to heighten the model's focus on crucial features. These attention mechanisms facilitate the model in capturing significant information across both spatial and channel dimensions, thereby enhancing the model's recognition capabilities.

The improved ResNet architecture is still composed of multiple stacked residual layers, with each layer containing several residual blocks. The primary function of these blocks is to extract and propagate features from the input image, progressively refining them into more abstract high-level representations through multiple layers of stacking. A typical residual block can be mathematically represented as follows:

$$F_{\text{out}} = \sigma\left(x + f_2\left(\sigma\left(f_1\left(x\right)\right)\right)\right), \tag{11}$$

here, x denotes the input feature map, f_1 and f_2 represent the convolutional weights of the two layers, and σ is the ReLU activation function.

This equation illustrates that the input feature map undergoes two successive convolutions, batch normalization, and activations, after which it is added to the original input to produce the final output. To further optimize the feature extraction process, as illustrated in Fig. 5, we have introduced the Channel Attention (CA) and Spatial Attention (SA) mechanisms within the residual block. The revised residual block computation is as follows:

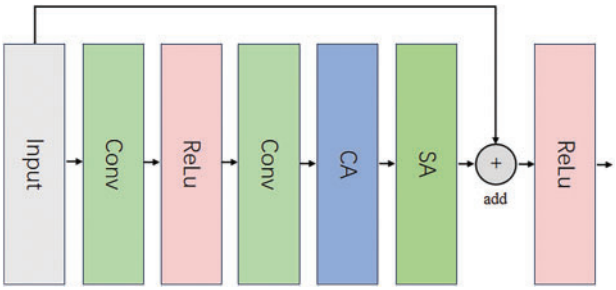

Figure 5: The input feature map sequentially passes through a residual block that incorporates both channel-wise and spatial attention mechanisms, followed by a residual connection with the original input feature map. Finally, it is processed through a ReLU activation function before outputting

$$F_{\text{out}} = \sigma\left(x + SA\left(CA\left(f_2\left(\sigma\left(f_1\left(x\right)\right)\right)\right)\right)\right), \tag{12}$$

The Channel Attention (CA) mechanism dynamically adjusts the weights of different channels. Its calculation is expressed as:

$$CA\left(x\right) = \sigma_2\left(f_2\left(\sigma_1\left(f_1\left(AvgPool\left(x\right) + MaxPool\left(x\right)\right)\right)\right)\right), \tag{13}$$

where σ_1 denotes the ReLU activation function and σ_2 is the Sigmoid activation function, f_1 and f_2 are the convolutional weights for channel attention, and AvgPool and MaxPool represent adaptive average pooling and max pooling operations, respectively.

The Spatial Attention (SA) mechanism, on the other hand, modulates the spatial distribution of the feature map, and is computed as follows:

$$SA\left(x\right) = \sigma\left(f_1\left(\left[Avg\left(x\right), Max\left(x\right)\right]\right)\right), \tag{14}$$

in this expression, f_1 is the convolutional weight for spatial attention, $\left[Avg\left(x\right), Max\left(x\right)\right]$ denotes the concatenation operation along the channel dimension, and *Avg* and *Max* represent the average and maximum operations along the channel dimension, respectively.

By integrating the Channel Attention and Spatial Attention mechanisms into ResNet, the model becomes more adept at dynamically adjusting the importance of information within the fused feature maps, significantly improving the capture of crucial features.

The Channel Attention mechanism enables the network to automatically assign different levels of importance to each feature channel, ensuring that more discriminative channels—those that are critical for recognizing micro-expressions—are emphasized. This attention mechanism helps the network focus on the most informative feature channels, effectively filtering out less relevant channels, which enhances the overall performance of the model.

Similarly, the Spatial Attention mechanism enables the model to focus on important regions of the image, such as areas where the face undergoes subtle movements. It achieves this by assigning higher attention weights to spatial locations that contain significant micro-expression features, thereby enhancing the model's capacity to identify subtle, confined alterations in facial expressions. The fusion of both attention mechanisms allows the model to simultaneously refine feature channels and focus on critical spatial regions, making the network more sensitive to both temporal and spatial variations.

This enhancement not only increases the model's adaptability to various modalities, such as static images, dynamic motion patterns, and differences between frames, but also improves its overall performance while maintaining the network's depth. By using attention mechanisms, the network can selectively focus on the most relevant information in both the channel and spatial dimensions, which leads to better recognition accuracy for micro-expressions.

4 Experiments and Discussion

4.1 Datasets

This study employed experiments across three key databases: SMIC [39], CASME II [42], and SAMM [43,44]. The SMIC database, a creation of the University of Oulu in Finland, boasts 164 natural micro-expression clips gathered from 16 individuals. These clips are categorized into three distinct types: positive, negative, and surprise. The CASME II database, issued by the Chinese Academy of Sciences' Institute

of Psychology, contains 247 micro-expressions cherry-picked from around 3000 facial expressions. This collection, sourced from 26 people, is divided into seven categories: happiness, surprise, disgust, suppression, sadness, fear, and more. The SAMM database, crafted by Manchester Metropolitan University in the UK, houses 159 micro-expression videos featuring 29 participants. The samples here are categorized into eight groups: happiness, surprise, disdain, anger, disgust, fear, sadness, and additional categories.

In this study, we adopted the same approach as used in the MEGC2019 Challenge [45] to standardize categories across different datasets, resulting in a unified dataset. Specifically, for the CASME II dataset, we classified 'happiness' under the 'positive' category, while 'disgust', 'repression', 'sadness', and 'fear' were grouped under 'negative'. The 'surprise' category was retained as is, and the remaining categories were excluded. Similarly, for the SAMM dataset, 'happiness' was assigned to the 'positive' category, while 'contempt', 'anger', 'disgust', 'fear', and 'sadness' were categorized as 'negative'. The 'surprise' category remained unchanged, and other categories were omitted. The detailed composition of the combined dataset is provided in Table 1.

Table 1: The experiments are implemented on SMIC, CASME II and SAMM databases

	SMIC	CASME II	SAMM
Samples	164	145	133
Subject	16	24	28
AUs	✗	✓	✓
Negative	70	88	92
Positive	51	32	26
Surprise	43	25	15

4.2 Setup

For our experiments, we utilized an NVIDIA GeForce RTX 4090 GPU equipped with 24 GB of memory. The model was fine-tuned with a learning rate set at 0.00001, leveraging the Adam optimizer alongside the cross-entropy loss function to effectively minimize errors. Training spanned 100 epochs, incorporating an early stopping protocol [46] that kicked in if the validation loss plateaued for 10 consecutive epochs. We kept a close eye on the training loss to ensure it was on the right track. To strike a balance between memory efficiency and convergence speed, we maintained a batch size of 32 throughout the training process. To enhance the robustness of our training data and mitigate overfitting, we employed data augmentation techniques like horizontal flipping and random cropping. Furthermore, we normalized the input images using the dataset's mean and standard deviation, ensuring the pixel values were centered around zero with a unit variance.

4.3 Evaluation Metrics

For this study, we employed the leave-one-subject-out (LOSO) cross-validation approach to perform experiments on both the unified dataset and the three original datasets. The performance of the models was evaluated using Unweighted F1 Score (UF1) and Unweighted Average Recall (UAR). The LOSO method ensures a thorough assessment of the models generalization ability by systematically leaving one subject out for testing while training on the others. This approach provides a robust estimate of model performance across diverse subject-specific characteristics, enhancing the credibility and reliability of the results.

The Unweighted F1 Score (UF1) is the mean of all class-wise $F1$ scores, where each $F1$ score is the harmonic mean of precision and recall, expressed as:

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}, \tag{15}$$

where *Precision* is the proportion of accurately predicted positive cases to the overall predicted positives, while *Recall* is the ratio of correctly predicted positive cases to all actual positive cases. The $UF1$ is then computed as:

$$UF1 = \frac{1}{N} \sum_{i=1}^{N} F1_i, \tag{16}$$

and Unweighted Average Recall is the average of recall values across all classes and is computed as:

$$UAR = \frac{1}{N} \sum_{i=1}^{N} Recall_i. \tag{17}$$

Using $UF1$ and UAR as performance metrics provides a more comprehensive evaluation of model performance, especially in the context of imbalanced datasets.

To achieve a balance between model performance and lightweight design, it is essential to compute the parameter count, FLOPs (Floating Point Operations), and inference time associated with various modules. FLOPs serve as a metric to assess the computational load during model inference. Given that the duration of micro-expressions typically lasts less than half a second [1], real-time micro-expression recognition necessitates that the inference time is constrained to approximately 200 ms, which is half the duration of the micro-expression. In our deployment experiments, we utilized an Intel i7-1165G7 CPU (2.8 GHz, 8 cores) with 16 GB of RAM, ensuring that the selected model achieves an inference time of around 50 ms, with a parameter count exceeding merely 3 million. This facilitates deployment in other environments, such as embedded hardware, while maintaining commendable performance.

4.4 Comparison with State-of-the-Arts

In Table 2, we present a comparison between our proposed method and both traditional and deep learning techniques across the composite (Full), SMIC, CASME II, and SAMM datasets. The evaluation metrics used include UF1 and UAR.

Table 2: Comparison of micro-expression recognition performance in terms of Unweighted F1-score (UF1) and Unweighted Average Recall (UAR) on the composite (Full), CASME II, SMIC and SAMM databases

	Full		SMIC		CASME II		SAMM	
Method	**UF1**	**UAR**	**UF1**	**UAR**	**UF1**	**UAR**	**UF1**	**UAR**
LBP-TOP [12]	0.5882	0.5785	0.2000	0.5280	0.7026	0.7429	0.3954	0.4102
BI-WOOF [20]	0.6296	0.6227	0.5727	0.5829	0.7805	0.8026	0.5211	0.5139
CapsuleNet [23]	0.6520	0.6506	0.5820	0.5877	0.7068	0.7018	0.6209	0.5989
ApexNet [22]	0.7196	0.7096	0.6817	0.6695	0.8764	0.8681	0.5409	0.5392
STSTNet [24]	0.7353	0.7605	0.6801	0.7013	0.8382	0.8686	0.6588	0.6810
RCN [25]	0.7432	0.7190	0.6326	0.6441	0.8512	0.8123	0.7601	0.6715
ME-PLAN [28]	0.7715	0.7864	0.7127	0.7256	0.8632	0.8778	0.7164	0.7418

(Continued)

Table 2 (continued)

	Full		SMIC		CASME II		SAMM	
Method	**UF1**	**UAR**	**UF1**	**UAR**	**UF1**	**UAR**	**UF1**	**UAR**
FeatRef [26]	0.7838	0.7832	0.7011	0.7083	0.8915	0.8873	0.7372	0.7155
FGRL-AUF [27]	0.7914	0.7933	0.7192	0.7215	0.8798	0.8710	0.7751	0.7890
Ours	0.7657	0.7295	0.6579	0.6422	0.8669	0.8491	0.7889	0.7325

Our proposed model, which incorporates three types of feature fusion (static, optical flow, and differential features) and leverages channel and spatial attention mechanisms within a semi-lightweight ResNet architecture, achieves a favorable balance between recognition performance and inference speed. Specifically, while our model outperforms most of the other models in terms of UF1 and UAR, it also maintains competitive inference speed, achieving around 60 ms per inference. Meanwhile, our model was evaluated on three independent datasets as well as on their combined dataset, and the results demonstrate its robust generalization capability.

In contrast, methods that achieve faster inference times tend to have lower UF1 and UAR scores, indicating that a trade-off between accuracy and speed is often necessary. On the other hand, methods that show higher UF1 and UAR scores than ours generally suffer from slower inference speeds, suggesting a potential sacrifice in real-time performance.

This highlights the primary advantage of our approach, which effectively addresses the challenge of real-time micro-expression recognition by focusing on both accuracy and inference efficiency. Utilizing feature fusion along with attention mechanisms enriches the model's grasp of input data, thereby enhancing recognition performance.

Due to the inherent complexity of nonconvex optimization [47], it is challenging to rigorously prove the convergence of the model. Therefore, this study employs extensive experimental validation to ascertain model convergence, specifically by monitoring the loss function curve and utilizing early stopping techniques [46] to ensure convergence. To further guarantee the model's generalizability, a Leave-One-Subject-Out (LOSO) cross-validation scheme is adopted. Finally, evaluations on the Full, SMIC, CASME II, and SAMM datasets demonstrate that the model exhibits excellent performance and confirm its robust generalization capability.

To analyze the model performance more rigorously, we conducted t-tests for both UF1 and UAR to assess their differences with other models. The significance level was set at $p < 0.05$. The results show:

For UF1, the differences were significant ($p < 0.05$) with the following models: LBP-TOP, BI-WOOF, CapsuleNet, and RCN, while the differences were not significant ($p \geq 0.05$) with the following models: ApexNet, STSTNet, ME-PLAN, FeatRef, and FGRL-AUF.

For UAR, the differences were significant ($p < 0.05$) with the following models: LBP-TOP, BI-WOOF, and CapsuleNet, while the differences were not significant ($p \geq 0.05$) with the following models: ApexNet, STSTNet, RCN, ME-PLAN, FeatRef, and FGRL-AUF.

Figs. 6 and 7 illustrate the confusion matrix for the best experimental outcomes on the composite dataset, SMIC, CASME II, and SAMM. Due to the greater number of negative samples, the model demonstrates higher accuracy in classifying the negative class. Additionally, the high quality of the CASME II dataset contributes to the model's strong performance, especially in distinguishing the negative class. However,

the variation in lighting conditions within the SAMM dataset impacts the model's ability to effectively differentiate between positive and negative samples.

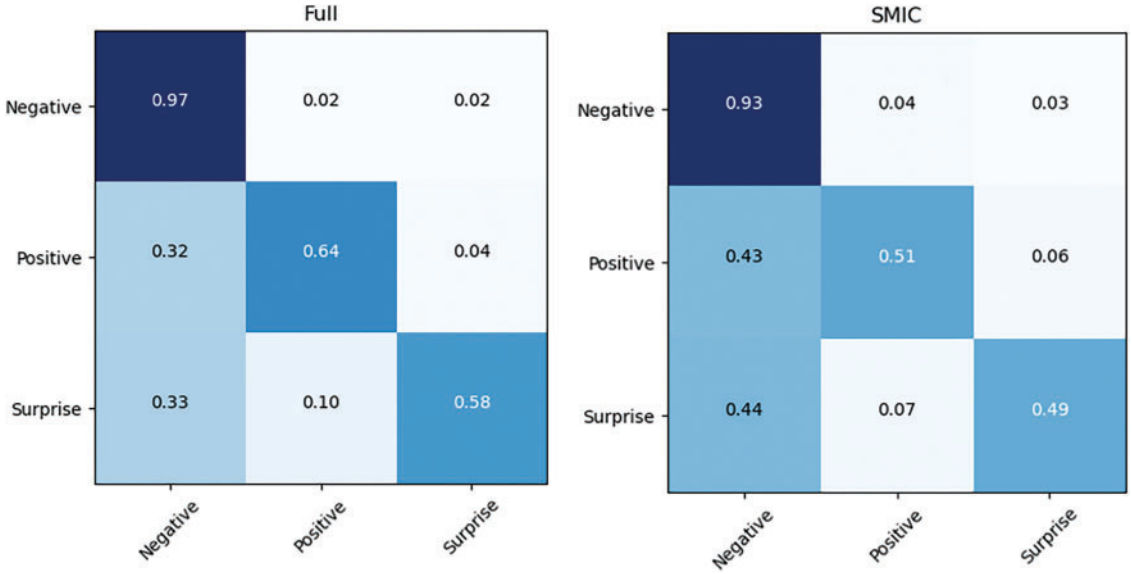

Figure 6: Confusion matrices of our proposed model on the composite database and SMIC with 3 classes

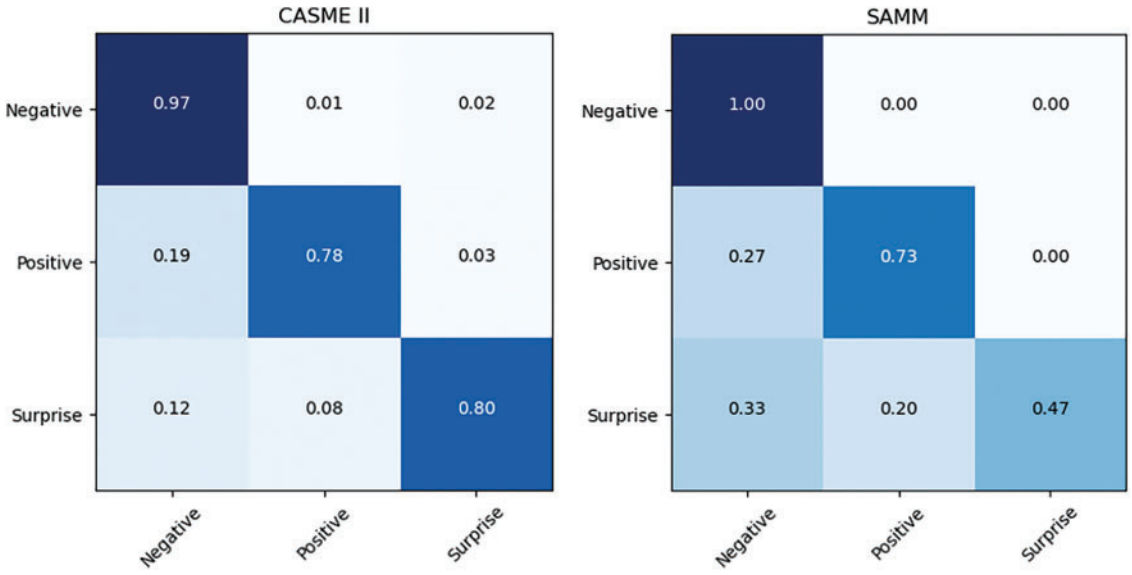

Figure 7: Confusion matrices of our proposed model on CASME II and SAMM datasets with 3 classes

It is noteworthy that prior research has produced models that outperform the results presented in our experiments, with the majority being based on transformer architectures [33]. Our estimations indicate that these models typically possess a parameter count exceeding 100 million and FLOPs surpassing 10 billion, with inference times greater than 300 ms. Consequently, such models were not considered in our analysis.

4.5 Ablation Studies

In this paper, we explore the impact of various modules on the model from different perspectives. Section 4.5.1 investigates the effects of different features on the model's performance. Section 4.5.2 analyzes the impact of different fusion methods on the model. Section 4.5.3 investigates the parameter count and inference speed of the modules used after feature fusion. Section 4.5.4 evaluates the effect of attention mechanisms in ResNet on the model's performance. Section 4.5.5 presents experiments and discussions on the impact of varying the number of layers in the ResNet architecture.

4.5.1 Impact of Feature Selection

Table 3 demonstrates the impact of various features on the model's performance. An ablation study was performed by combining each of the three input features in pairs, utilizing the fusion method proposed in this paper. If a feature was excluded, the corresponding fusion step was omitted. Experimental results reveal that the model primarily relies on optical flow features, which capture the motion characteristics during micro-expression transitions. Additionally, while spatial and differential features also contribute to the model's performance, the impact of optical flow features is more pronounced.

Table 3: The impact of different features on the model performance. Here, T represents temporal features, S denotes spatial features, and D refers to differential features

	Full		SMIC		CASME II		SAMM	
Feature	**UF1**	**UAR**	**UF1**	**UAR**	**UF1**	**UAR**	**UF1**	**UAR**
S + D	0.6474	0.6332	0.5547	0.5896	0.7217	0.7155	0.6317	0.5782
T + D	0.7067	0.6733	0.6340	0.6208	0.7588	0.7323	0.7436	0.7107
T + S	0.7368	0.7051	0.6409	0.6273	0.8339	0.8176	0.7531	0.7236
T + S + D	0.7657	0.7295	0.6579	0.6422	0.8669	0.8491	0.7889	0.7325

4.5.2 Impact of Feature Selection

Table 4 showcases the outcomes of various fusion techniques applied to the three distinct feature types. Initially, we standardize the dimensions of the three input feature maps, followed by the implementation of conventional fusion methods like element-wise addition, multiplication, and concatenation. These are then juxtaposed with the innovative fusion strategy introduced in this study. The findings unequivocally highlight the superior performance of our proposed fusion method.

Table 4: Evaluation of Feature Fusion Techniques. The input feature maps are standardized to identical dimensions, and conventional approaches like element-wise addition, multiplication, and concatenation are benchmarked against the innovative fusion strategy introduced in this study

	Full		SMIC		CASME II		SAMM	
Fusion	**UF1**	**UAR**	**UF1**	**UAR**	**UF1**	**UAR**	**UF1**	**UAR**
Add	0.6856	0.6563	0.6039	0.5946	0.7669	0.7609	0.6874	0.6353
Mul	0.6758	0.6422	0.6323	0.6181	0.7053	0.6803	0.6708	0.6407
Cat	0.6841	0.6500	0.6269	0.6142	0.7167	0.6907	0.6963	0.6663
Ours	0.7657	0.7295	0.6579	0.6422	0.8669	0.8491	0.7889	0.7325

4.5.3 Analysis of Module Parameter Count and Inference Speed

Table 5 presents the parameters and performance metrics of various modules following feature fusion, evaluated based on UF1 and UAR on a mixed dataset, as well as the parameter count and FLOPs. Additionally, it includes the average inference time computed over 100 inference runs. Among the models, AlexNet has the smallest parameter count but demonstrates suboptimal performance. In contrast, Vision Transformer (ViT) achieves better performance; however, it requires significantly greater computational resources and exhibits prolonged inference times. Consequently, we ultimately opted for ResNet, which provides a more balanced compromise between performance and computational efficiency. ResNet strikes an effective balance between high performance and inference speed due to its deeper architecture and residual connections. The residual connections allow for efficient training of deeper models, improving the model's ability to capture complex features without significant increases in computational burden.

Table 5: Comparative analysis of parameter count, performance metrics (UF1 and UAR), FLOPs, and average inference time of various modules following feature fusion

Module	UF1	UAR	Parameter	Speed	FLOPs
AlexNet	0.6269	0.6142	2.384 M	39.872	1.093 G
VGG	0.6741	0.6399	7.104 M	59.137	2.672 G
Inception	0.6811	0.6469	3.218 M	45.134	1.123 G
ResNet	0.7333	0.6991	3.273 M	54.713	2.229 G
ViT	0.7877	0.7418	85.892 M	308.997	17.801 G

4.5.4 Impact of Attention Mechanisms

Table 6 investigates the impact of incorporating attention mechanisms into ResNet. The study compares the performance of the model without any attention mechanism, with spatial attention, and with channel attention. The results indicate that channel attention is a key factor in improving the model's performance, showing the most significant gains on the SAMM dataset. Furthermore, the simultaneous application of both spatial and channel attention yields improvements across all evaluation metrics on each dataset.

Table 6: The effect of incorporating attention mechanisms into ResNet. This table compares the performance of the model without attention, with Spatial Attention (SA), and with Channel Attention (CA)

	Full		SMIC		CASME II		SAMM	
Attention	UF1	UAR	UF1	UAR	UF1	UAR	UF1	UAR
None	0.7333	0.6991	0.6323	0.6214	0.8198	0.7941	0.7583	0.7103
SA	0.7436	0.7097	0.6402	0.6270	0.8461	0.8282	0.7570	0.7068
CA	0.7596	0.7248	0.6470	0.6334	0.8586	0.8386	0.7949	0.7453
All	0.7657	0.7295	0.6579	0.6422	0.8669	0.8491	0.7889	0.7325

Moreover, Table 7 presents a comparative analysis of the parameter count, FLOPs, and inference time associated with the inclusion of the attention module. It is observed that the increase in parameter count and FLOPs remains within acceptable limits, while the inference time satisfies real-time requirements. Furthermore, the incorporation of the attention module significantly enhances the model's recognition accuracy.

Table 7: Comparison of parameter count, FLOPs, and inference time before and after the inclusion of the attention module

Attention	UF1	UAR	Parameter	Speed	FLOPs
None	0.7333	0.6991	3.273 M	54.713	2.229 G
All	0.7657	0.7295	3.296 M	65.944	2.231 G

4.5.5 Impact of ResNet Layers

Table 8 presents the experimental results for ResNet with varying numbers of layers. This study examines ResNet configurations with layer counts ranging from 1 to 4. It is observed that as the number of layers increases, the network acquires sufficient feature representation capabilities. However, the transition from 3 to 4 layers yields only marginal improvements in these representations, while simultaneously increasing both the model's parameter count and inference time.

Table 8: Performance comparison of ResNet with varying numbers of layers. The table compares the performance of ResNet models with different layer configurations (1 to 4 layers)

	Full		SMIC		CASME II		SAMM	
Layers	UF1	UAR	UF1	UAR	UF1	UAR	UF1	UAR
1	0.6851	0.6504	0.6319	0.6177	0.7203	0.6907	0.6899	0.6628
2	0.7470	0.7107	0.6428	0.6285	0.8374	0.8178	0.7788	0.7197
3	0.7657	0.7295	0.6579	0.6422	0.8669	0.8491	0.7889	0.7325
4	0.7739	0.7387	0.6566	0.6410	0.8750	0.8595	0.8201	0.7675

Table 9 illustrates the impact of each layer on the model's parameters and inference speed. When the layer count is equal to 4, there is an observable improvement in recognition capability compared to the scenario with 3 layers; however, this comes at the cost of a several-fold increase in parameter count and a corresponding impact on inference speed. Therefore, this study selects the configuration with 3 layers as the optimal choice. It is worth noting that if deployment on more lightweight devices is desired, a trade-off in recognition accuracy may be warranted, allowing for the selection of a configuration with 2 layers.

Table 9: Analysis of the impact of varying layer counts on model parameters and inference speed. The table compares the recognition capabilities, parameter counts, and inference speeds for ResNet configurations with 2, 3, and 4 layers

Layers	UF1	UAR	Parameter	Speed	FLOPs
1	0.6851	0.6504	0.456 M	40.108	1.404 G
2	0.7470	0.7107	1.046 M	47.282	1.817 G
3	0.7657	0.7295	3.296 M	65.944	2.231 G
4	0.7739	0.7387	11.412 M	89.513	2.640 G

4.6 Application Prospects

Micro-expression recognition technology holds significant potential in clinical diagnosis and psychotherapy [3], with the real-time recognition capability enhancing its practical value. First, in mental health monitoring, real-time micro-expression recognition can serve as an auxiliary tool for clinicians, enabling more accurate assessments of patients' emotional states. By analyzing micro-expression changes in real-time, healthcare professionals can quickly detect negative emotions such as anxiety, depression, or excessive stress that patients may not explicitly express. Compared to traditional emotional assessment methods, real-time recognition not only improves the timeliness and accuracy of diagnosis but also provides immediate feedback on emotional fluctuations, helping to adjust and optimize psychological treatment in real-time.

Additionally, real-time micro-expression recognition has unique advantages in the early diagnosis of emotional disorders. Many patients with emotional disorders find it difficult to express their true emotions verbally, and real-time micro-expression detection can sensitively capture emotional fluctuations, offering faster and more accurate emotional assessments. This allows clinicians to tailor more personalized treatment plans, enhancing the overall effectiveness of therapy.

5 Conclusion

This paper introduces a novel semi-lightweight deep learning architecture, DynamicFusionResNet (DFR-Net), which strikes a balance between lightweight design and recognition performance. The model incorporates dynamic, static, and optical flow features as input feature maps. The DFR-Net architecture is specifically designed to accommodate these diverse features and introduces an innovative fusion strategy to integrate them effectively. Additionally, we employ spatial-channel attention mechanisms within ResNet to enhance feature representation and improve the model's discriminative power. Experimental results demonstrate that the DFR-Net model performs exceptionally well across multiple benchmarks, proving competitive among semi-lightweight networks. Its semi-lightweight nature also facilitates real-time inference, making it suitable for various practical applications.

However, for micro-expression recognition tasks, the limited availability of training data constrains the model's learning effectiveness. Additionally, some datasets suffer from low quality due to varying lighting conditions, which further compromises the reliability of facial feature extraction. Future research may focus on generating realistic micro-expressions to alleviate the issue of limited training data and increase the model's practical relevance in practical contexts.

Acknowledgement: Not applicable.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm contribution to the paper as follows: Study conception: Mengqi Li; Draft preparation: Mengqi Li; Writing, review, and editing: Xiaodong Huang, Lifeng Wu. All authors reviewed the results 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, Xiaodong Huang, upon reasonable request.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Porter S, Ten Brinke L. Reading between the lies: identifying concealed and falsified emotions in universal facial expressions. Psychol Sci. 2008;19(5):508–14. doi:10.1111/j.1467-9280.2008.02116.x.
2. O'Sullivan M, Frank MG, Hurley CM, Tiwana J. Police lie detection accuracy: the effect of lie scenario. Law Hum Behav. 2009;33(6):530–8. doi:10.1007/s10979-008-9166-4.
3. Mikhailova ES, Vladimirova TV, Iznak AF, Tsusulkovskaya EJ, Sushko NV. Abnormal recognition of facial expression of emotions in depressed patients with major depression disorder and schizotypal personality disorder. Biol Psychiatry. 1996;40(8):697–705. doi:10.1016/0006-3223(96)00032-7.
4. Frank M, Herbasz M, Sinuk K, Keller A, Nolan C. I see how you feel: training laypeople and professionals to recognize fleeting emotions. In: Annual Meeting of the International Communication Association; 2009 May 21–25; Chicago, IL, USA. p. 1–35. doi:10.1007/s12144-019-00359-x.
5. Ojala T, Pietikäinen M, Harwood D. A comparative study of texture measures with classification based on featured distributions. Pattern Recognit. 1996;29(1):51–9. doi:10.1016/0031-3203(95)00067-4.
6. Horn BKP, Schunck BG. Determining optical flow. Artif Intell. 1981;17(1–3):185–203. doi:10.1016/0004-3702(81)90024-2.
7. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20(3):273–97. doi:10.1007/BF00994018.
8. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32. doi:10.1023/A:1010933404324.
9. Patel D, Hong X, Zhao G. Selective deep features for micro-expression recognition. In: Proceeding of 23rd International Conference on Pattern Recognition (ICPR); 2016 Dec 4–8; Cancun, Mexico. p. 2258–63. doi:10.1109/icpr.2016.7899972.
10. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16 × 16 words: transformers for image recognition at scale. arXiv:2010.11929. 2020.
11. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceeding of IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016 Jun 27–30; Las Vegas, NV, USA. p. 770–8. doi:10.1109/cvpr.2016.90.
12. Pfister T, Li X, Zhao G, Pietikäinen M. Recognising spontaneous facial micro-expressions. In: Proceeding of IEEE International Conference on Computer Vision Workshops (ICCV Workshops); 2011 Nov 6–13; Barcelona, Spain. p. 1449–56. doi:10.1109/iccv.2011.6126401.
13. Wang Y, See J, Phan RCW, Oh YH. LBP with six intersection points: reducing redundant information in LBP-TOP for micro-expression recognition. In: Proceeding of 12th Asian Conference on Computer Vision (ACCV); 2014 Nov 1–5; Singapore. p. 525–37. doi:10.1007/978-3-319-16865-4_34.
14. Huang X, Wang SJ, Zhao G, Pietikäinen M. Facial micro-expression recognition using spatiotemporal local binary pattern with integral projection. In: Proceeding of IEEE International Conference on Computer Vision Workshop (ICCVW); 2015 Dec 7–13; Santiago, Chile. p. 1–9. doi:10.1109/iccvw.2015.10.
15. Wang Y, See J, Phan RCW, Oh YH. Efficient spatiotemporal local binary patterns for spontaneous facial micro-expression recognition. PLoS One. 2015;10(5):e0124674. doi:10.1371/journal.pone.0124674.
16. Huang X, Wang SJ, Liu X, Zhao G, Feng X, Pietikäinen M. Discriminative spatiotemporal local binary pattern with revisited integral projection for spontaneous facial micro-expression recognition. IEEE Trans Affect Comput. 2017;10(1):32–47. doi:10.1109/TAFFC.2017.2713359.
17. Liu YJ, Zhang JK, Yan WJ, Wang SJ, Zhao G, Fu X. A main directional mean optical flow feature for spontaneous micro-expression recognition. IEEE Trans Affect Comput. 2015;7(4):299–310. doi:10.1109/TAFFC.2015.2485205.
18. Xu F, Zhang J, Wang JZ. Microexpression identification and categorization using a facial dynamics map. IEEE Trans Affect Comput. 2017;8(2):254–67. doi:10.1109/TAFFC.2016.2518162.
19. Zhang S, Feng B, Chen Z, Huang X. Micro-expression recognition by aggregating local spatio-temporal patterns. In: Proceeding of 23rd International Conference Multimedia Modeling (MMM); 2017 Jan 4–6; Reykjavik, Iceland. p. 638–48. doi:10.1007/978-3-319-51811-4_52.
20. Liong ST, See J, Wong K, Phan RCW. Less is more: micro-expression recognition from video using apex frame. Signal Process Image Commun. 2018;62(10):82–92. doi:10.1016/j.image.2017.11.006.

21. Happy SL, Routray A. Fuzzy histogram of optical flow orientations for micro-expression recognition. IEEE Trans Affect Comput. 2017;10(3):394–406. doi:10.1109/TAFFC.2017.2723386.
22. Gan YS, Liong ST, Yau WC, Huang YC, Tan LK. Off-apexnet on micro-expression recognition system. Signal Process Image Commun. 2019;74(2):129–39. doi:10.1016/j.image.2019.02.005.
23. Quang NV, Chun J, Tokuyama T. Capsulenet for micro-expression recognition. In: Proceedings of 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019); 2019 May 14–18; Lille, France. p. 1–7. doi:10.1109/fg.2019.8756544.
24. Liong ST, Gan YS, See J, Khor HQ, Huang YC. Shallow triple stream three-dimensional CNN (STSTNet) for micro-expression recognition. In: Proceedings of 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019); 2019 May 14–18; Lille, France. p. 1–5. doi:10.1109/fg.2019.8756567.
25. Xia Z, Peng W, Khor HQ, Feng X, Zhao G. Revealing the invisible with model and data shrinking for composite-database micro-expression recognition. IEEE Trans Image Process. 2020;29:8590–605. doi:10.1109/TIP.2020.3018222.
26. 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.
27. Lei L, Chen T, Li S, Li J. Micro-expression recognition based on facial graph representation learning and facial action unit fusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); 2021 Jun 19–25; Nashville, TN, USA. p. 5–6. doi:10.1109/cvprw53098.2021.00173.
28. Zhao S, Tang H, Liu S, Zhang Y, Wang H, Xu T, et al. ME-PLAN: a deep prototypical learning with local attention network for dynamic micro-expression recognition. Neural Netw. 2022;153(8):427–43. doi:10.1016/j.neunet.2022.06.024.
29. Verma M, Lubal P, Vipparthi SK, Abdel-Mottaleb M. RNAS-MER: a refined neural architecture search with hybrid spatiotemporal operations for micro-expression recognition. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision; 2023 Jan 2–7; Waikoloa, HI, USA. p. 4770–9. doi:10.1109/wacv56688.2023.00475.
30. Wei J, Peng W, Lu G, Li Y, Yan J, Zhao G. Geometric graph representation with learnable graph structure and adaptive au constraint for micro-expression recognition. IEEE Trans Affect Comput. 2023;15(3):1343–57. doi:10.1109/TAFFC.2023.3340016.
31. Zhao X, Lv Y, Huang Z. Multimodal fusion-based Swin transformer for facial recognition micro-expression recognition. In: Proceedings of the IEEE International Conference on Mechatronics and Automation (ICMA); 2022 Aug 7–10; Guilin, China. p. 780–5. doi:10.1109/icma54519.2022.9856162.
32. Zhai Z, Zhao J, Long C, Xu W, He S, Zhao H. Feature representation learning with adaptive displacement generation and transformer fusion for micro-expression recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023 Jun 17–24; Vancouver, BC, Canada. p. 22086–95. doi:10.1109/cvpr52729.2023.02115.
33. 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.
34. Wang F, Li J, Qi C, Wang L, Wang P. JGULF: joint global and unilateral local feature network for micro-expression recognition. Image Vis Comput. 2024;147(1):105091. doi:10.1016/j.imavis.2024.105091.
35. Bao Y, Wu C, Zhang P, Shan C, Qi Y, Ben X. Boosting micro-expression recognition via self-expression reconstruction and memory contrastive learning. IEEE Trans Affect Comput. 2024;15(4):2083–96. doi:10.1109/TAFFC.2024.3397701.
36. Zhang M, Yang W, Wang L, Wu Z, Chen D. HFA-Net: hierarchical feature aggregation network for micro-expression recognition. Complex Intell Syst. 2025;11(3):1–20. doi:10.1007/s40747-024-01660-4.
37. Wang T, Shang L. Temporal augmented contrastive learning for micro-expression recognition. Pattern Recognit Lett. 2023;167(9):122–31. doi:10.1016/j.patrec.2023.02.003.
38. Wang J, Tian Y, Yang Y, Chen X, Zheng C, Qiang W. Meta-auxiliary learning for micro-expression recognition. arXiv:2024.12024. 2024.

39. Li X, Pfister T, Huang X, Zhao G, Pietikäinen M. A spontaneous micro-expression database: inducement, collection and baseline. In: Proceedings of the 10th IEEE International Conference on Automatic Face and Gesture Recognition (FG); 2013 Apr 22–26; Shanghai, China. p. 1–6. doi:10.1109/fg.2013.6553717.
40. Liong ST, See J, Wong K, Ngo ACL, Oh YH, Phan R. Automatic apex frame spotting in micro-expression database. In: Proceedings of the 3rd IAPR Asian Conference on Pattern Recognition (ACPR); 2015 Nov 3–6; Kuala Lumpur, Malaysia. p. 665–9. doi:10.1109/acpr.2015.7486586.
41. Ahonen T, Hadid A, Pietikäinen M. Face recognition with local binary patterns. In: Proceedings of the European Conference on Computer Vision (ECCV); 2004 May 11–14; Prague, Czech Republic. p. 469–81. doi:10.1007/978-3-540-24670-1_36.
42. Yan WJ, Li X, Wang SJ, Zhao G, Liu YJ, Chen YH, et al. CASME II: an improved spontaneous micro-expression database and the baseline evaluation. PLoS One. 2014;9(1):e86041. doi:10.1371/journal.pone.0086041.
43. Davison AK, Lansley C, Costen N, Tan K, Yap MH. SAMM: a spontaneous micro-facial movement dataset. IEEE Trans Affect Comput. 2016;9(1):116–29. doi:10.1109/TAFFC.2016.2573832.
44. Yap CH, Kendrick C, Yap MH. SAMM long videos: a spontaneous facial micro-and macro-expressions dataset. In: Proceedings of the 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG); 2020 Nov 16–20; Buenos Aires, Argentina. p. 771–6. doi:10.1109/fg47880.2020.00029.
45. See J, Yap MH, Li J, Hong X, Wang SJ. MEGC 2019—the second facial micro-expressions grand challenge. In: Proceedings of the 14th IEEE International Conference on Automatic Face and Gesture Recognition (FG); 2019 May 14–18; Lille, France. p. 1–5. doi:10.1109/fg.2019.8756611.
46. Prechelt L. Early stopping—but when? In: Montavon G, Orr G, Muller KR, editors. Neural networks: tricks of the trade. Berlin/ Heidelberg: Springer Berlin Heidelberg; 2002. p. 55–69. doi:10.1007/978-3-642-35289-8_5.
47. Du S, Lee J, Li H, Wang L, Zhai X. Gradient descent finds global minima of deep neural networks. In: Proceedings of the 36th International Conference on Machine Learning; 2019 Jun 9–15; Long Beach, CA, USA. p. 1675–85. doi:10.1109/allerton.2019.8919696.