Scilight

*Journal of Machine Learning and Information Security*
https://www.sciltp.com/journals/jmlis

*Article*

# A Directional Attention Fusion and Multi-Head Spatial-Channel Attention Network for Facial Expression Recognition

Yukun Shao, Yang Li and Baiqiang Wu *

School of Automation Science and Electrical Engineering, Beihang University, Beijing 100191, China
* Correspondence: wubaiqiang@buaa.edu.cn

**How To Cite**: Shao, Y.; Li, Y.; Wu, B. A Directional Attention Fusion and Multi-Head Spatial-Channel Attention Network for Facial Expression Recognition. *Journal of Machine Learning and Information Security* **2025**, *1*(1), 7.

Received: 15 September 2025
Revised: 21 October 2025
Accepted: 22 October 2025
Published: 5 November 2025

**Abstract:** Facial Expression Recognition (FER), as a cutting-edge affective computing technology, holds significant application value in the field of human–computer interaction. However, due to intra-class variations, subtle inter-class differences, and environmental interference, FER in unconstrained scenarios remains challenging. To address these limitations, this paper proposes a Directional Attention Fusion and Multi-Head Spatial-Channel Attention Network (DAF-MHSCA). Firstly, coarse-grained facial features are extracted through a ResNet18 backbone network, followed by the capture of detailed expression features via an adaptive feature calibration (AFC) mechanism. Subsequently, we introduce a directional attention fusion (DAF) module, which generates spatial attention maps through both self-attention and cross-attention mechanisms along the width and height directions. Finally, a multi-head spatial-channel attention (MHSCA) module is incorporated, which integrates the spatial attention maps to perform channel-wise and spatial-wise attention on the features, ultimately enabling emotion recognition through a classifier. The competitive experimental results on five datasets have shown that our proposed method achieves notable improvements over state-of-the-art methods.

**Keywords:** facial expression recognition; directional attention fusion; multi-head spatial-channel attention

## 1. Introduction

Facial expressions serve as critical signals for conveying emotional states, leveraging their intuitiveness and information richness to effectively communicate feelings and thoughts [1]. With advancements in computer vision, Facial Expression Recognition (FER) has emerged as a significant research direction [2], demonstrating extensive application value and research potential in fields such as medical diagnosis, driver safety, and virtual reality [3].

However, FER performance in real-world scenarios remains suboptimal [4]. The primary limitation stems from the fact that most existing methods are developed using high-quality images captured in controlled laboratory settings, which are characterized by minimal interference from illumination variations, viewing angles, or complex backgrounds. In contrast, real-world data is susceptible to numerous challenges: variations in subjects, uncontrolled environmental illumination, complex background clutter, and the inherent difficulty in obtaining large-scale, accurately annotated datasets due to the time-consuming, subjective, and ambiguous nature of expression labeling itself [5]. Furthermore, existing techniques exhibit limited capability in discriminating subtle expression differences. Significant intra-class variations arise from physiological differences, environmental interference, and labeling subjectivity. Simultaneously, small inter-class differences result from physiological similarities between certain expressions and the insufficient sensitivity of algorithms to subtle distinguishing features [6].

To address these challenges, researchers have proposed various improvements. Deng et al. [7] enhanced feature discriminative power with their Deep Locality Preserving Neural Network (DLP) but struggled to focus on crucial regions. Wen et al. [8] improved attention to local regions via their Distract your Attention Network (DAN) but lacked effective modeling of inter-region relationships. Their subsequent work introducing bi-directional attention

heads generated attention maps in two directions but failed to incorporate interaction between these directions. Zhang et al. [9] proposed a Dual-Directional Attention Mixed Feature Network (DDAMFN), utilizing bi-directional attention heads and an attention loss mechanism to effectively capture long-range dependencies and focus on important regions; however, it still faced difficulties in distinguishing easily confusable expressions. Josep et al. [10] enhanced the multi-task performance of DDAMFN through optimized loss functions and data balancing but did not thoroughly validate its effectiveness for unconstrained scenarios using mainstream large-scale datasets.

To overcome these limitations, this paper proposes a novel Directional Attention Fusion and Multi-Head Spatial-Channel Attention Network (DAF-MHSCA). The method employs a modular pipeline: (1) A ResNet18 backbone for extracting primary facial features and an Adaptive Feature Calibration (AFC) mechanism leveraging multi-scale dilated convolutions to capture detailed facial features; (2) A Directional Attention Fusion (DAF) module utilizing self-attention and cross-attention to reinforce directional features and generate attention maps; (3) A Multi-Head Spatial-Channel Attention (MHSCA) module jointly optimizing attention maps with the original features to produce classification results. The directional convolutional structure of DAF-MHSCA effectively models directional feature correlations, while its hybrid attention mechanism precisely localizes key expression regions, significantly enhancing the model's discriminative capability for subtle inter-class differences and substantial intra-class variations.

To validate the effectiveness of the proposed method, comprehensive experiments are conducted on five public datasets. On RAF-DB, DAF-MHSCA achieves a recognition accuracy of 92.49%. For AffectNet-7 and AffectNet-8 [11], our approach achieves accuracies of 67.72% and 65.20%, respectively. On FER2013 and FERPlus, our approach attains an accuracy of 74.33% and 91.38%, respectively, demonstrating significant superiority over existing methods. These results comprehensively validate the advancement and effectiveness of the proposed framework.

## 2. Materials and Methods

In this section, we propose a novel facial expression recognition approach based on Directional Attention Fusion and Multi-Head Spatial-Channel Attention (DAF-MHSCA). The proposed architecture comprises three components: an adaptive feature calibration (AFC) mechanism, a directional attention fusion (DAF) module, and a multi-head spatial-channel attention (MHSCA) module. Some acronyms are summarized in Table 1.

**Table 1.** Glossary of Key Acronyms.

| Acronym | Full Name |
|---|---|
| AFC | Adaptive Feature Calibration |
| CA | Channel Attention |
| MHSCA | Multi-Head Spatial-Channel Attention |
| DAF | Directional Attention Fusion |
| FER | Facial Expression Recognition |
| GAP | Global Average Pooling |
| BN | Batch Normalization |
| ReLU | Rectified Linear Unit Activation Function |

### *2.1. Overall Architecture of DAF-MHSCA*

The overall architecture of the proposed DAF-MHSCA is shown in Figure 1. Firstly, the preprocessed input image is fed into the ResNet18 backbone to extract facial features, and the AFC mechanism employs multi-scale dilated convolutions to capture expression-related details at varying receptive fields, particularly within micro-expression regions. Subsequently, through dedicated and hybrid attention mechanisms operating along the height and width dimensions, DAF generates three attention-weighted representations: Height-direction self-attention features, Width-direction self-attention features and Cross-direction attention features. These representations are synthesized to compute a spatial attention map encoding discriminative expression information. Finally, facial features extracted earlier in the network are routed to the MHSCA. MHSCA independently derives spatial attention features and channel attention features. The spatial attention map generated by DAF is then applied to further refine both the spatial and channel features within MHSCA, enhancing their expression-related characteristics.

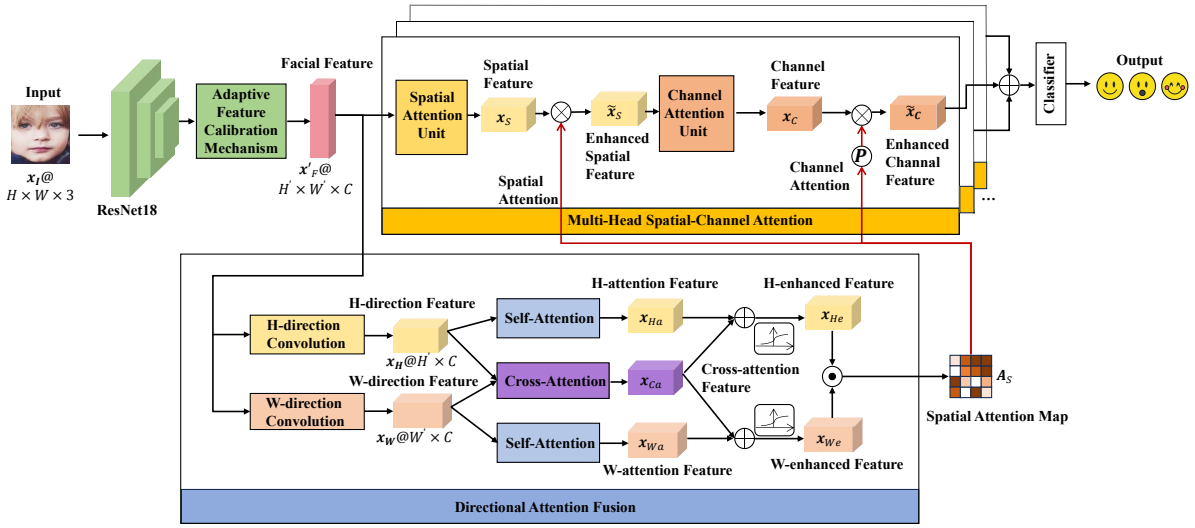


**Figure 1.** The architecture of the DAF-MHSCA.

### *2.2. Backbone Feature Extraction and Adaptive Feature Calibration Mechanism*

We adopts the ResNet18 [12] residual network as the backbone feature extraction network. The ResNet18 network, pre-trained on large-scale FER datasets, has convolutional kernels in its initial layers that have learned universal facial texture and edge features. We employ transfer learning, using a ResNet18 network pre-trained on the MS-Celeb-1M [13] face recognition dataset and adapting it for expression recognition tasks on the RAF-DB, AffectNet-7, AffectNet-8, FER2013 and FERPlus datasets.

The initial layer of ResNet18 employs a $7 \times 7$ convolutional kernel with stride 2 for feature extraction. Subsequent to the initial layer, batch normalization (BatchNorm), ReLU activation, and max pooling (MaxPool) operations are applied. The input image, denoted as $x_I \in \mathbb{R}^{H \times W \times 3}$, undergoes progressive downsampling through multiple stages to yield the final facial feature $x_F \in \mathbb{R}^{H_1 \times W_1 \times C}$, where $H_1 = H/32$, $W_1 = W/32$, and $C = 512$.

Although features extracted by the ResNet18 demonstrate reasonable accuracy, the limited depth and complexity of the network constrain its ability to capture subtle expression variations. The Adaptive Feature Calibration (AFC) mechanism addresses this limitation through multi-scale dilated convolutions that capture expression details at varying receptive fields, particularly in micro-expression regions such as eyebrows and eye corners. The AFC mechanism employs spatial attention to automatically filter irrelevant information while utilizing channel reweighting to suppress expression-irrelevant feature channels. Finally, a residual connection structure amplifies inter-class feature differences, enhancing recognition precision. The architectural of the AFC is illustrated in Figure 2, and the pseudocode of the AFC is shown in Algorithm 1.

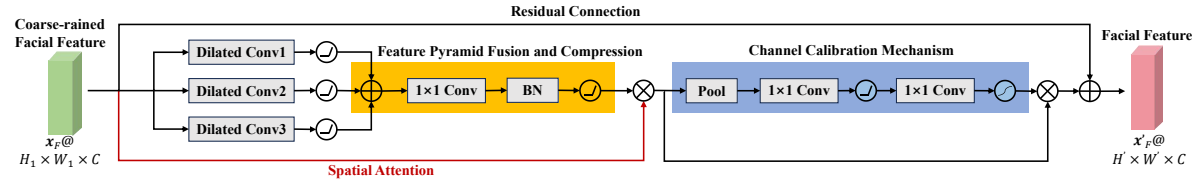


**Figure 2.** The architecture of the AFC.

The AFC mechanism processes $x_F$ using three dilated convolutions with different dilation rates to capture multi-scale expression details $x_{F_1}$, $x_{F_2}$, and $x_{F_3}$, thereby enhancing facial micro-expression feature extraction.

$$\begin{cases} x_{F_1} = \text{ReLU}\left(\text{Conv2d}_{3\times3}^{p_1}(x_F)\right) \\ x_{F_2} = \text{ReLU}\left(\text{Conv2d}_{3\times3}^{p_2,d_2}(x_F)\right) \\ x_{F_3} = \text{ReLU}\left(\text{Conv2d}_{3\times3}^{p_4,d_4}(x_F)\right) \\ x_{F_{\text{sum}}} = x_{F_1} + x_{F_2} + x_{F_3} \end{cases} \tag{1}$$

where $\text{Conv2d}_{3\times3}^{p_1}(\cdot)$ represents a standard dilated convolution with $3 \times 3$ kernel and padding = 1; $\text{Conv2d}_{3\times3}^{p_2,d_2}(\cdot)$ denotes a medium-receptive-field dilated convolution with $3 \times 3$ kernel, padding = 2, and dilation rate = 2;

$\mathrm{Conv2d}_{3\times3}^{p_4,d_4}(\cdot)$ signifies a large-receptive-field dilated convolution with $3 \times 3$ kernel, padding = 4, and dilation rate = 4.

**Algorithm 1:** Adaptive Feature Calibration (AFC) Module

**Require:** Coarse-grained facial feature $x_F \in \mathbb{R}^{H_1\times W_1\times C}$ (output from ResNet18)
**Ensure:** Calibrated feature $x'_F \in \mathbb{R}^{H_1\times W_1\times C}$ (same dimension as $x_F$)
**Step 1: Multi-scale Dilated Convolution**
Apply 3 × 3 dilated conv (dilation = 1, padding = 1) + ReLU to $x_F \rightarrow x_{F1}$
Apply 3 × 3 dilated conv (dilation = 2, padding = 2) + ReLU to $x_F \rightarrow x_{F2}$
Apply 3 × 3 dilated conv (dilation = 4, padding = 4) + ReLU to $x_F \rightarrow x_{F3}$
Sum $x_{F1}, x_{F2}, x_{F3} \rightarrow x_{Fsum}$
**Step 2: Feature Fusion & Spatial Calibration**
Apply 1 × 1 conv + BN + ReLU to $x_{Fsum} \rightarrow x_{Fint}$
Apply 1 × 1 conv + sigmoid to $x_F$ (get spatial attention weights) $\rightarrow x_{FcalS\_weight}$
Element-wise multiply $x_{Fint}$ with $x_{FcalS\_weight} \rightarrow x_{FcalS}$
**Step 3: Channel Calibration**
Apply global average pooling (GAP) to $x_{FcalS} \rightarrow x_{Fgap}$
Apply 1 × 1 conv to $x_{Fgap}$ (dimension reduction) $\rightarrow x_{Fdre}$
Apply ReLU + 1 × 1 conv to $x_{Fdre}$ (dimension recovery) $\rightarrow x_{Fdra}$
Apply sigmoid to $x_{Fdra}$ (get channel attention weights) $\rightarrow x_{FcalC\_weight}$
Element-wise multiply $x_{FcalS}$ with $x_{FcalC\_weight} \rightarrow x_{FcalC}$
**Step 4: Residual Connection**
Sum original $x_F$ with $x_{FcalC} \rightarrow x'_F$ **return** $x'_F$

The multi-scale features $x_{F_{\text{sum}}}$ are then processed through the feature pyramid fusion and compression module to generate the integrated feature $x_{F_{\text{int}}}$. This module concatenates the three branch outputs $x_{F_1}$, $x_{F_2}$, and $x_{F_3}$ along the channel dimension, applies $1 \times 1$ convolution to halve the channel count, and calibrates features by matrix multiplication with spatial attention weights. The integrated and spatially calibrated features are computed as:

$$\begin{cases} x_{F_{\text{int}}} = \mathrm{ReLU}\left(\mathrm{BN}\left(\mathrm{Conv2d}_{1\times1}(x_{F_{\text{sum}}})\right)\right) \\ x_{F_{\text{calS}}} = \sigma\left(\mathrm{Conv2d}_{1\times1}(x_F)\right) \odot x_{F_{\text{int}}} \end{cases} \tag{2}$$

where $\sigma(\cdot)$ represents the sigmoid activation function, and $\odot$ indicates element-wise multiplication.

Following spatial calibration, channel calibration is performed through two $1 \times 1$ convolutions that learn feature channel weights. The first convolution reduces dimensionality of the pooled features to produce $x_{F_{\text{Dre}}}$, while the second convolution upscales the activated reduced-dimension feature to generate $x_{F_{\text{Dra}}}$. The channel-calibrated feature $x_{F_{\text{calC}}}$ is obtained by weighting the spatial attention output with channel weights to enhance critical channel responses. Finally, residual connection preserves original input information:

$$\begin{cases} x_{F_{\text{Dre}}} = \mathrm{Conv2d}_{1\times1}\left(\mathrm{GAP}(x_{F_{\text{calS}}})\right) \\ x_{F_{\text{Dra}}} = \mathrm{Conv2d}_{1\times1}\left(\mathrm{ReLU}\left(x_{F_{\text{Dre}}}\right)\right) \\ x_{F_{\text{calC}}} = \sigma\left(x_{F_{\text{Dra}}}\right) \odot x_{F_{\text{calS}}} \\ x'_F = x_{F_{\text{calC}}} + x_F \end{cases} \tag{3}$$

The output $x'_F \in \mathbb{R}^{H'\times W'\times C}$ maintains identical dimensionality as the input feature $x_F \in \mathbb{R}^{H_1\times W_1\times C}$.

### *2.3. Directional Attention Fusion*

In facial expression recognition tasks, fine-grained associations of local features and dynamic interactions of cross-directional features are critical to enhancing recognition accuracy. Traditional feature fusion methods struggle to effectively model the complementarity between horizontal and vertical features while introducing redundant noise. To address this limitation, we propose a Directional Attention Fusion (DAF) Module that achieves dynamic fusion of bidirectional features through a synergistic mechanism combining self-attention and cross-directional attention. This design significantly enhances the model's perception of subtle expression variations. The architecture of the DAF is illustrated in Figure 3.

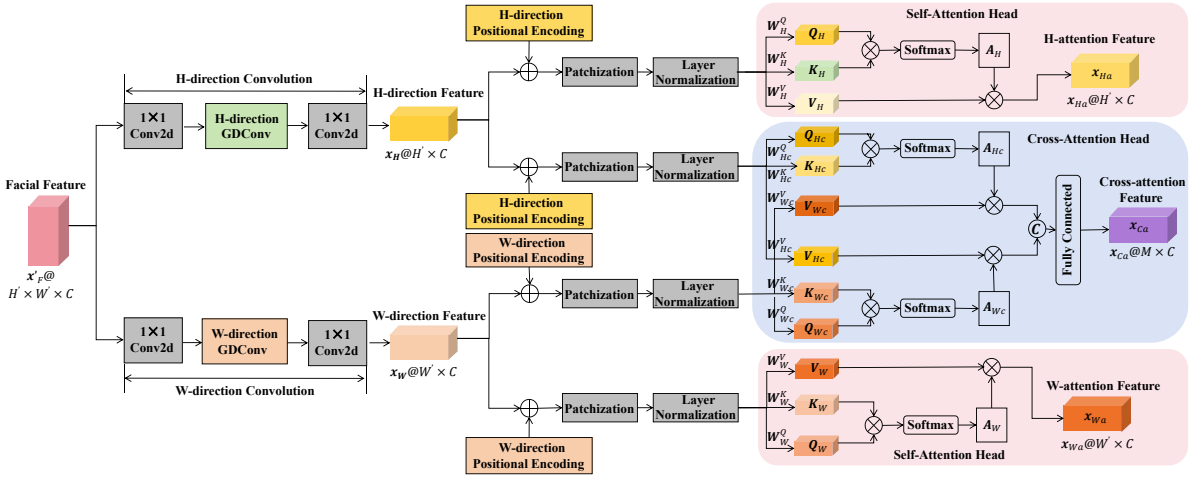


**Figure 3.** The architecture of the DAF.

Firstly, a dual-direction convolution captures fine-grained facial texture information along orthogonal spatial directions. We employ asymmetric convolutional kernels to model transverse stretching (e.g., mouth corners) and longitudinal contractions (e.g., eyebrow movements) through separate processing paths. This directional decomposition enables more effective modeling of expression-related deformations than isotropic convolution operations.

Formally, $x'_F \in \mathbb{R}^{H' \times W' \times C}$ denotes the feature map extracted from the AFC module, where $H'$, $W'$, and $C$ represent height, width, and channel dimensions respectively. The dual-direction convolution processes $x'_F$ through two parallel convolution branches: height-direction processing and width-direction processing.

$$\begin{cases} x_H = \text{Conv2D}_{1\times1}\left(\text{GDConv}_H\left(\text{Conv2D}_{1\times1}(x'_F)\right)\right) \\ x_W = \text{Conv2D}_{1\times1}\left(\text{GDConv}_W\left(\text{Conv2D}_{1\times1}(x'_F)\right)\right) \end{cases} \tag{4}$$

where $\text{Conv2D}_{1\times1}(\cdot)$ denotes pointwise convolution; $\text{GDConv}_H(\cdot)$ represents global dilated convolution with vertical orientation; $\text{GDConv}_W(\cdot)$ denotes global dilated convolution with horizontal orientation; $x_H \in \mathbb{R}^{H' \times C}$ donates the height-direction feature; $x_W \in \mathbb{R}^{W' \times C}$ donates the width-direction feature. $\text{GDConv}_H$ and $\text{GDConv}_W$ are the global convolution functions in the $H$ and $W$ directions, respectively.

Subsequently, height-direction features $x_H \in \mathbb{R}^{H' \times C}$ and width-direction features $x_W \in \mathbb{R}^{W' \times C}$ are partitioned into patches and projected into four embedded representations through linear transformations. To preserve spatial relationships, positional encodings are incorporated, followed by layer normalization to enhance stability, resulting in four normalized features: $x_H\prime$, $x_H\prime\prime$, $x_W\prime$, $x_W\prime\prime$.

The self-attention heads process directional features independently. For height-direction features, the query $Q_H$, key $K_H$, and value $V_H$ are computed as:

$$\begin{bmatrix} Q_H & K_H & V_H \\ Q_W & K_W & V_W \end{bmatrix} = \begin{bmatrix} x'_H & 0 \\ 0 & x'_W \end{bmatrix} \begin{bmatrix} W_H^Q & W_H^K & W_H^V \\ W_W^Q & W_W^K & W_W^V \end{bmatrix} \tag{5}$$

The self-attention features are then generated through scaled dot-product attention.

$$\begin{cases} x_{Ha} = \text{softmax}\left(\frac{Q_H K_H^T}{\sqrt{C}}\right) V_H \\ x_{Wa} = \text{softmax}\left(\frac{Q_W K_W^T}{\sqrt{C}}\right) V_W \end{cases} \tag{6}$$

The cross-attention head enables simultaneous focus on both directions. The attention outputs are computed as:

$$x_{Ca} = \text{concat}\left\{\text{softmax}\left(\frac{Q_{Hc} K_{Hc}^\top}{\sqrt{C}}\right) V_{Wc}, \text{softmax}\left(\frac{Q_{Wc} K_{Wc}^\top}{\sqrt{C}}\right) V_{Hc}\right\} W_{Ca} \tag{7}$$

where $\text{concat}\{\cdot\}$ denotes concatenation and $W_{Ca}$ is a learnable weight matrix.

The cross-attention features $x_{Ca}$ are combined with self-attention features through residual connections and sigmoid activation. The spatial attention map $A_s$ is generated through element-wise multiplication of enhanced features.

$$A_s = \sigma(x_{Ha} + x_{Ca}) \odot \sigma(x_{Wa} + x_{Ca}) \tag{8}$$

This attention map dynamically weights expression-relevant regions while suppressing irrelevant background information.

The pseudocode of the DAF is shown in Algorithm 2, which provides three key advantages: (1) Directional feature interaction: Models complementary relationships between orthogonal facial movements; (2) Multi-scale context: Captures both local details and global dependencies; (3) Adaptive weighting: Dynamically emphasizes expression-critical regions.

**Algorithm 2:** Pseudocode for Directional Attention Fusion (DAF) Module

**Require:** $x'_F$: Feature map from AFC
**Ensure:** $A_s$: Spatial attention map
**Step 1: Extract directional features**
Compute height-direction features ($x_H$) via vertical global dilated convolution + 1 × 1 convolution
Compute width-direction features ($x_W$) via horizontal global dilated convolution + 1 × 1 convolution
**Step 2: Add positional encoding & normalize**
Add height-specific positional encoding to $x_H$, apply layer normalization
Add width-specific positional encoding to $x_W$, apply layer normalization
**Step 3: Compute self-attention**
Generate Q/K/V from normalized $x_H$, compute height self-attention ($x_{Ha}$)
Generate Q/K/V from normalized $x_W$, compute width self-attention ($x_{Wa}$)
**Step 4: Compute cross-attention**
Cross-attend $x_H$ with $x_W$ and vice versa, concatenate results to get $x_{Ca}$
**Step 5: Generate spatial attention map**
Sum $x_{Ha}$ + $x_{Ca}$ and $x_{Wa}$ + $x_{Ca}$, apply sigmoid to both
Element-wise multiply the two results to get $A_s$
**return** $A_s$

### *2.4. Multi-Head Spatial-Channel Attention Module*

We propose the Multi-Head Spatial-Channel Attention (MHSCA) module, which performs refined weighting and enhancement on facial features. This process focuses on spatial regions and channel information crucial for expression recognition, thereby enhancing the model's discriminative power. The architectural of the MHSCA is illustrated in Figure 4.

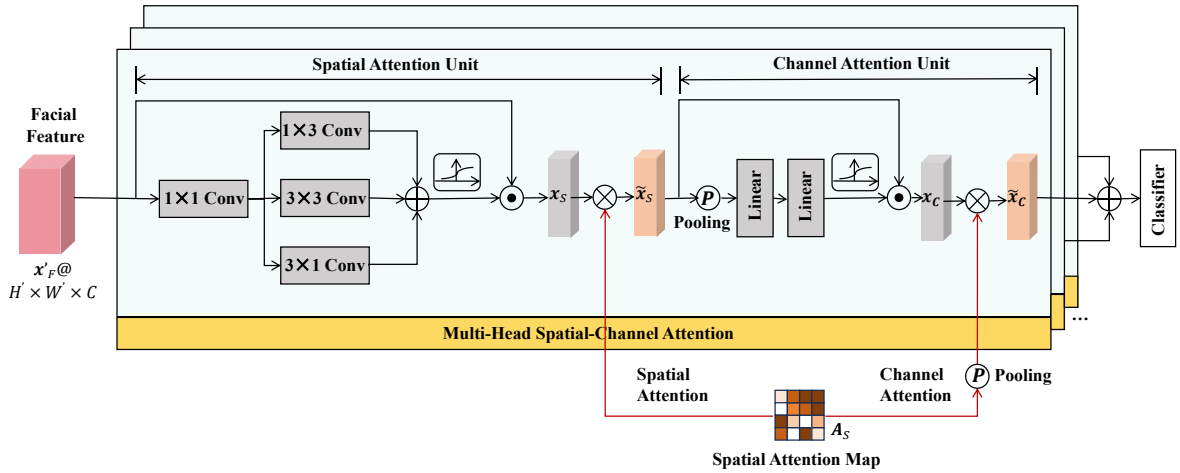


**Figure 4.** The architecture of the MHSCA.

Channel attention focuses on semantically significant feature content by emphasizing important channels while suppressing redundant ones. Spatial attention prioritizes discriminative spatial locations by weighting critical regions. By integrating both mechanisms with a multi-head design, spatial attention heads concentrate on local facial regions through dynamic weight allocation, suppressing background interference [14]. Concurrently, channel attention heads enhance expression-relevant feature channels while suppressing irrelevant ones [15]. The multi-head mechanism attends to different facial regions, significantly improving recognition accuracy in key areas.

The proposed Multi-Head Spatial-Channel Attention (MHSCA) module processes features through coordinated

spatial and channel attention mechanisms. Taking $x'_F \in \mathbb{R}^{H' \times W' \times C}$ as input, MHSCA employs multi-scale convolutional spatial attention to capture local spatial relationships. To model long-range dependencies, it weights spatial features using the attention map $A_s$ from the bidirectional feature fusion module. Channel relationships are enhanced through channel attention map $A_c$, derived by pooling $A_s$, which weights channel features to amplify discriminative characteristics.

The spatial attention head processes $x'_F$ to generate spatial attention features $x_s$. These are enhanced using spatial attention map $A_s$ to produce $\widetilde{x}_s$, which is fed to the channel attention head. The channel attention head outputs channel attention features $x_c$, which are refined using channel attention map $A_c$ (obtained by pooling $A_s$) to produce $\widetilde{x}_c$. Finally, multi-head enhanced channel attention features are summed and fed to the classifier for expression recognition. The pseudocode of the DAF is shown in Algorithm 3.

Firstly, the spatial attention head processes $x_F$ through parallel convolutional branches.

$$x_s = \delta\Big(\text{Conv}_{1\times3}\left(\text{Conv}_{1\times1}(x_F)\right) + \text{Conv}_{3\times3}\left(\text{Conv}_{1\times1}(x_F)\right) + \text{Conv}_{3\times1}\left(\text{Conv}_{1\times1}(x_F)\right)\Big) \odot x_F \tag{9}$$

where $\delta(\cdot)$ denotes the Sigmoid activation function and $\odot$ represents element-wise multiplication. The enhanced spatial features are obtained via matrix multiplication:

$$\widetilde{x}_s = A_s \cdot x_s \tag{10}$$

Then, the channel attention head processes $\widetilde{x}_s$ through pooling and linear layers. The linear layers are primarily employed for feature transformation and dimensionality adjustment, aiming to enhance the expressive capability of channel attention. The first linear layer reduces the feature dimensionality to decrease computational complexity, while the second linear layer restores the original dimension and generates channel-wise weights, emphasizing key feature.

$$x_c = \delta\left(\text{Linear}\left(\text{Linear}\left(\text{GAP}(\widetilde{x}_s)\right)\right)\right) \odot \widetilde{x}_s \tag{11}$$

Enhanced channel features are computed as:

$$\widetilde{x}_c = A_c \cdot x_c = \text{GAP}(A_s) \cdot x_c \tag{12}$$

The final output aggregates features from $N_h$ attention heads:

$$x_{\text{out}} = \sum_{i=1}^{N_h} \widetilde{x}_c^{(i)} \tag{13}$$

where $\widetilde{x}_c^{(i)}$ represents enhanced channel features from the $i$-th attention head. This aggregated feature vector is fed to the classifier for expression recognition.

**Algorithm 3:** Pseudocode for Multi-Head Spatial-Channel Attention (MHSCA) Module

**Require:** $x'_F$: Feature map from AFC; $A_s$: Spatial attention map from DAF; $N_h$: Number of attention heads
**Ensure:** $x_{out}$: Final enhanced feature for classification
**Step 1: Compute spatial attention features**
Apply parallel 1 × 3, 3 × 3, 3 × 1 convolutions to $x'_F$, sum + sigmoid, multiply with $x'_F$ to get $x_s$
Enhance $x_s$ by element-wise multiplying with $A_s$ to get $\tilde{x}_s$
**Step 2: Compute channel attention features**
Apply global average pooling to $\tilde{x}_s$, pass through linear layer + sigmoid, multiply with $\tilde{x}_s$ to get $x_c$
Compute channel attention map $A_c$ via global average pooling on $A_s$
Enhance $x_c$ by element-wise multiplying with $A_c$ to get $\tilde{x}_c$
**Step 3: Aggregate multi-head features**
Repeat Steps 1–2 for $N_h$ heads, collect all $\tilde{x}_c^{(i)}$ (i = 1 to $N_h$)
Sum all $\tilde{x}_c^{(i)}$ to get $x_{out}$
**return** $x_{out}$

### *2.5. Loss Function*

To effectively train the proposed DAF-MHSCA model, we design a multi-task joint optimization loss function system that synergistically combines cross-entropy loss, feature affinity loss, and attention diversity loss.

2.5.1. Weighted Cross-Entropy Loss

To address significant class imbalance in FER datasets, we introduce adaptive class weights into the standard cross-entropy loss. For a training set with $C$ classes where class $c$ contains $N_c$ samples, the weighted cross-entropy loss $L_{wce}$ and class weight $w_c$ are defined as:

$$\begin{cases} L_{wce} = -\frac{1}{N}\sum_{i=1}^{N} w_{y_i} \log\left(\frac{e^{z_{i,y_i}}}{\sum_{c=1}^{C} e^{z_{i,c}}}\right) \\ w_c = \left(\frac{N_{\max}}{N_c+\epsilon}\right)^{1/2} \end{cases} \tag{14}$$

where $N$ is the total number of samples, $y_i$ is the true label of sample $i$, $z_{i,c} \in \mathbb{R}$ is the logit output for class $c$, $N_{\max} = \max\{N_1, \cdots, N_C\}$, and $\epsilon = 10^{-6}$ prevents division by zero. The square root operation ensures minority classes receive higher weights without excessive compensation.

2.5.2. Feature Affinity Loss

The feature affinity loss $L_{dna}$ enhances intra-class compactness through dynamic normalization in feature space, addressing limitations of traditional center-based losses that directly constrain Euclidean distances to class centroids. This loss computes the normalized distance between sample features and their class centers:

$$L_{dna} = \frac{1}{N}\sum_{i=1}^{N} w_{y_i} \frac{\|\mathbf{f}_i - \mathbf{m}_{y_i}\|_2^2}{\sigma_M^2 + \epsilon} \tag{15}$$

where $\mathbf{f}_i \in \mathbb{R}^d$ is the feature vector of sample $i$, $\mathbf{m}_{y_i} \in \mathbb{R}^d$ is the centroid of class $y_i$, and $\sigma_M^2$ represents the variance of class centroids in feature space. The denominator dynamically normalizes feature distances based on inter-class dispersion, forcing the network to learn tighter clusters for closely spaced classes. From a mathematical perspective, this normalization effectively increases the margin between different class centroids in the normalized feature space. By penalizing large intra-class distances relative to inter-class separation, $L_{dna}$ directly addresses the challenge of significant intra-class variation in FER, leading to more discriminative features that are crucial for distinguishing subtle expression differences.

2.5.3. Attention Diversity Loss

Specifically designed for multi-head attention mechanisms, $L_{mhv}$ enhances orthogonality across attention heads to prevent redundancy and improve localization of discriminative facial regions:

$$L_{mhv} = -\frac{\alpha}{BHW}\sum_{b=1}^{B}\sum_{h=1}^{H}\sum_{w=1}^{W} \text{Var}(\mathbf{A}_{b,h,w}) + \frac{2\beta}{N(N-1)}\sum_{i\neq j} \cos(\mathbf{A}_i, \mathbf{A}_j) \tag{16}$$

where $\mathbf{A} \in \mathbb{R}^{B\times H\times W\times N}$ denotes attention weights, $\text{Var}(\cdot)$ computes head-wise variance, and $\cos(\cdot)$ measures cosine similarity between attention head matrices. Hyperparameters $\alpha = 0.7$ and $\beta = 0.3$ balance the variance maximization and similarity minimization terms. This formulation ensures different attention heads focus on complementary facial regions. Mathematically, maximizing the variance of each attention head encourages them to produce sharp, high-contrast attention maps, which helps in pinpointing critical local features. Simultaneously, minimizing the cosine similarity between heads promotes feature diversity, ensuring comprehensive coverage of various facial regions. This cooperative mechanism enables the model to fuse multi-region cues effectively, which is fundamental for accurate recognition under complex real-world conditions.

2.5.4. Multi-Loss Integration

The total loss $L_{total}$ integrates all components through a gradient-gated weighting scheme:

$$L_{total} = L_{wce} + L_{dna} + \theta L_{mhv} \tag{17}$$

where $\theta = 0.3$ scales the attention diversity term. The gradient gating mechanism automatically adjusts the effective weights during backpropagation by monitoring the relative magnitudes of each loss component's gradients.

## 3. Experiments and Discussion

### *3.1. Datasets*

To evaluate the effectiveness of DAF-MHSCA, this section presents its performance assessment on five public facial expression datasets: RAF-DB, AffectNet-7, AffectNet-8, FER2013 and FERPlus.

RAF-DB [7] is a facial expression dataset comprising more than 29,670 images collected from the Internet. The dataset provides two subsets, both annotated by 40 trained human coders. The subset used for expression classification consists of 12,271 training images and 3068 test images, all aligned and cropped to a size of 100 × 100 pixels. In the experiments, only the single-label subset containing seven basic emotions (happiness, anger, fear, surprise, neutral, sadness, disgust) is utilized.

AffectNet [11] is a large-scale facial expression dataset containing over 1,000,000 facial images gathered from the Internet. It includes two benchmark branches: AffectNet-7 and AffectNet-8. AffectNet-7 contains 287,401 images across seven expression categories, with 283,901 used for training and 3500 for testing. AffectNet-8 extends to 287,651 training samples and 4000 test samples, with an additional "contempt" category accounting for approximately 1.30% of the total.

The FER2013 facial expression dataset comprises 35,886 facial expression images. The dataset is divided into 28,708 training images, with 3589 images each allocated to a public test set and a private test set. Each image is a grayscale image of fixed size, 48 × 48 pixels, and is categorized into one of seven expression classes: anger, disgust, fear, happy, sad, surprised, and neutral.

The FERPlus dataset is an extension of the FER2013 dataset, comprising 28,709 training images and 3589 validation images. It provides annotations for each image with multiple emotional labels, covering eight emotion categories: neutral, happy, sad, surprise, fear, disgust, and contempt, thereby offering higher-quality and more fine-grained label information.

### *3.2. Implementation Details*

This experiment is conducted based on the PyTorch deep learning framework and executed on eight NVIDIA L40S GPUs (each with 48 GB VRAM). The backbone network employed is a ResNet18 model pre-trained on the MS-Celeb-1M face recognition dataset. All images are resized to 224 × 224 pixels. To ensure fair comparisons and mitigate overfitting, data augmentation techniques are applied, including random rotation, erasure, and horizontal flipping. The number of attention heads ($N_h$) in the MHSCA is set to 4. The model is trained for 40 epochs on each dataset using the Adam optimizer, with different initial learning rates set for each dataset. The learning rate for the RAF-DB dataset is set to 0.1, while for the AffectNet, FER2013, and FERPlus datasets, the learning rate is set to 0.0001. And the batch size for the FERPlus is set to 64, while for the RAF-DB, AffectNet, and FER2013 datasets, the batch size is set to 256. The specific implementation details are shown in the Table 2.

**Table 2.** Implementation Details.

| Dataset | Learning Rate | Number of Epochs | Batch Size | Optimizer |
|---|---|---|---|---|
| RAF-DB | 0.1 | 40 | 256 | Adam |
| AffectNet-7 | 0.0001 | 40 | 256 | Adam |
| AffectNet-8 | 0.0001 | 40 | 256 | Adam |
| FER2013 | 0.0001 | 40 | 256 | Adam |
| FERPlus | 0.0001 | 40 | 64 | Adam |

### *3.3. Data Augmentation*

In facial expression recognition research, image enhancement serves as a core technique for data preprocessing, addressing the "distribution gap" between training datasets and real-world scenarios through simulation of environmental complexity and diversity. This study incorporates an image enhancement module prior to input data processing. The module performs preprocessing through cropping and scaling operations to adjust image dimensions. Data augmentation techniques including grayscale conversion, contrast enhancement, and Gaussian noise addition are employed to eliminate irrelevant background regions, expand the dataset, and reduce overfitting.

### *3.4. Overall Performance*

We conduct a comprehensive evaluation of our proposed DAF-MHSCA approach against several state-of-the-art facial expression recognition methods on the RAF-DB, AffectNet-7, AffectNet-8, FER2013 and FERPlus

datasets, as shown in Table 3. The experimental results reveal that the proposed DAF-MHSCA model achieves state-of-the-art results, outperforming contemporary methods on all evaluated benchmarks. Specifically, DAF-MHSCA attains accuracies of 92.49% on the RAF-DB, 67.72% on the AffectNet-7, 65.20% on the AffectNet-8, 74.33% on the FER2013, and 91.38% on the FERPlus, surpassing the next-best approaches like MFER and DDAMFN, which highlights its enhanced capability in mitigating intra-class variations and inter-class ambiguities under real-world conditions.

Additionally, a comparative analysis is conducted between DAF-MHSCA and state-of-the-art transformer-based FER models, namely VTFF, GAAVE, and MMATrans. As shown in the Table 3, the proposed DAF-MHSCA surpasses the transformer-based models in performance across all datasets. This advantage stems from the Directional Attention Fusion (DAF) Module in DAF-MHSCA, which generates spatial attention maps along the width-direction and height-direction through self-attention and cross-attention mechanisms, effectively capturing subtle expression variations. In contrast, VTFF relies on global self-attention and may overlook local directional features. GAAVE focuses on handling noisy labels but lacks discriminative power on clean data. And MMATrans overemphasizes modeling muscle relationships while neglecting the importance of directional feature interactions and efficient attention mechanisms.

**Table 3.** Performance comparison for RAF-DB, AffectNet-7, AffectNet-8, FER2013 and FERPlus datasets.

| Method | Accuracy (%) | | | | | |
|---|---|---|---|---|---|---|
| | Year | RAF-DB | AffectNet-7 | AffectNet-8 | FER2013 | FERPlus |
| MVT [16] | 2021 | 88.62 | 64.57 | 61.40 | – | 89.22 |
| DACL [17] | 2021 | 87.78 | 65.20 | – | 68.54 | – |
| VTFF [18] | 2023 | 88.14 | – | 61.85 | – | 88.81 |
| DAN [8] | 2023 | 89.70 | 65.69 | 62.02 | 71.19 | 89.05 |
| DDAMFN [9] | 2023 | 91.35 | 67.03 | 64.25 | – | 90.74 |
| FEDA [19] | 2023 | 89.03 | 64.53 | – | 69.72 | 85.26 |
| GAAVE [20] | 2023 | 89.29 | 65.60 | 62.78 | – | 89.83 |
| MMATrans [21] | 2024 | 89.67 | 64.89 | – | – | 90.32 |
| MFER [22] | 2024 | 92.08 | 67.06 | 63.15 | 73.45 | 91.09 |
| DAF-MHSCA | 2025 | 92.49 | 67.72 | 65.20 | 74.33 | 91.38 |

### *3.5. Visualization Results*

Figure 5 presents the confusion matrices for three methods-DAN, DDAMFN, and our DAF-MHSCA on the RAF-DB dataset. These matrices illustrate the correspondence between predicted labels and ground-truth labels across the seven emotion categories. In comparison to DAN and DDAMFN, the DAF-MHSCA approach exhibits enhanced performance in multiple emotion classes, particularly demonstrating superior prediction accuracy for emotions such as “disgust” and “fear”. This outcome further substantiates the efficacy and competitive advantage of the DAF-MHSCA in emotion recognition tasks.

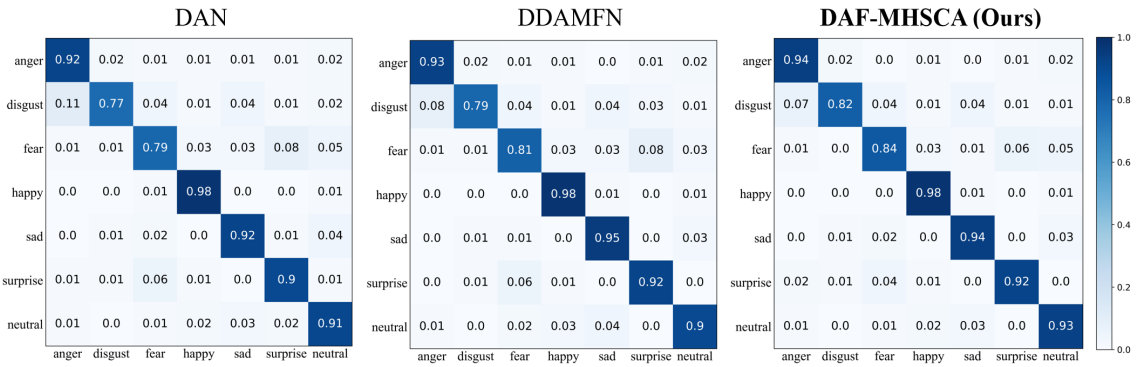


**Figure 5.** The confusion matrixes of our DAF-MHSCA and baseline methods on the RAF-DB dataset.

Figure 6 presents the t-SNE visualization results of DAN, DDAMFN, and the proposed DAF-MHSCA method on the RAF-DB dataset. In this visualization, the seven expression categories within the RAF-DB dataset are distinctly represented by seven unique colors. Compared to DAN and DDAMFN, the DAF-MHSCA approach

achieves more pronounced separation between different emotion categories, as evidenced by clearer inter-class boundaries. This enhanced clustering performance demonstrates the superior capability of DAF-MHSCA to discriminate nuanced emotional states, particularly when processing complex expression data.

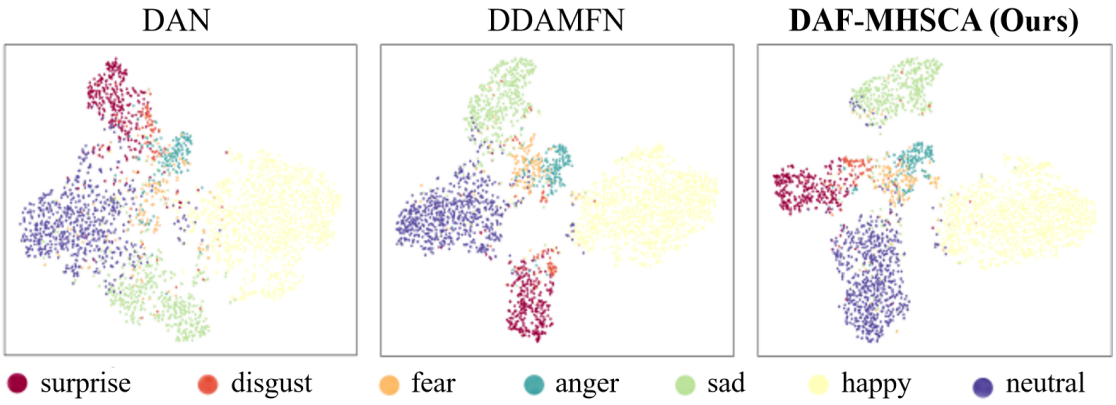


**Figure 6.** Our DAF-MHSCA and baseline methods for t-SNE visualization on the RAF-DB dataset.

Furthermore, Figure 7 illustrates the attention maps of DAN, DDAMFN, and the proposed DAF-MHSCA method for seven distinct expressions on the RAF-DB dataset. In these visualizations, TL denotes True Labeland while PL indicates Predicted Label. Predictions inconsistent with ground-truth labels are highlighted in red, while correct predictions are marked in green. The colored regions across attention maps represent salient areas prioritized by each model. Quantitative analysis reveals that both DAN and DDAMFN exhibit misclassifications for certain emotions, whereas DAF-MHSCA demonstrates significantly higher accuracy. This performance advantage substantiates the efficacy of its hybrid attention mechanism in emotion recognition tasks. Critically, the attention maps of DAF-MHSCA consistently concentrate on key facial action units—such as periocular regions and oral commissures—which aligns with psychological studies of expression encoding. This targeted focus validates the architectural importance of integrating mixed-attention mechanisms for discriminative feature localization.

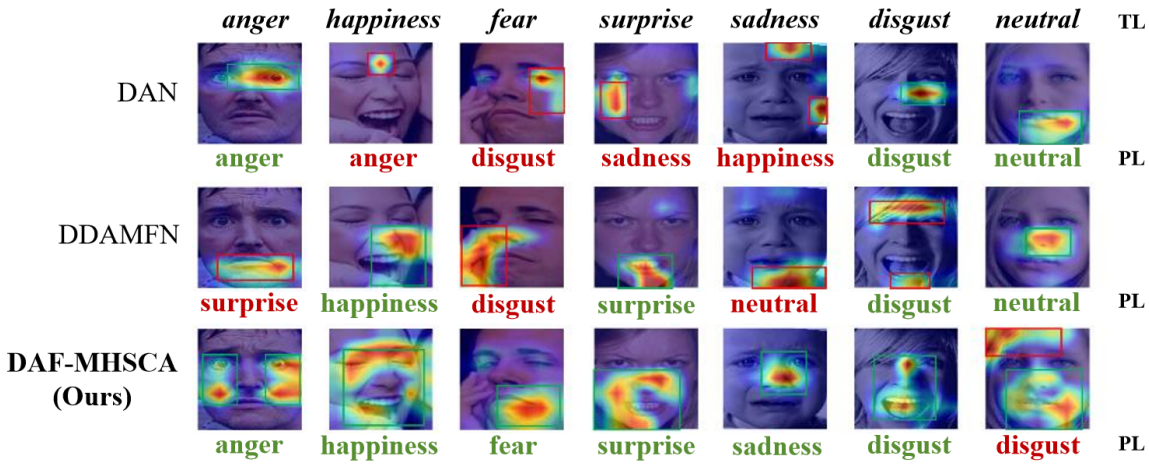


**Figure 7.** The attention maps of Our DAF-MHSCA and baseline methods on the RAF-DB dataset.

### *3.6. Ablation Experiments*

Table 4 presents a comprehensive ablation study of the DAF-MHSCA model across five benchmark datasets (RAF-DB, AffectNet-7, AffectNet-8, FER2013, and FERPlus), systematically evaluating the contribution of each core component to the overall recognition performance. The results demonstrate that the removal of the AFC, DAF, CA (Channel Attention), and MHSCA components consistently degrades accuracy, underscoring their synergistic importance. For instance, on the challenging RAF-DB dataset, excluding AFC reduces accuracy from 92.49% to 91.69%, while removing DAF results in a more significant decline to 90.96%, indicating the critical role of DAF in modeling directional feature interactions. Similarly, on AffectNet-7, ablation of CA lowers accuracy from 67.72% to 67.37%, and on FERPlus, omitting MHSCA reduces performance from 91.38% to 91.15%, highlighting how the multi-head mechanism of MHSCA enhances discriminative power.

These performance degradations can be attributed to the distinct functional roles of each component: the multi-scale dilated convolutions of AFC are essential for resolving micro-expression details and mitigating intra-class variations, as evidenced by the consistent drops across datasets; the hybrid self-attention and cross-attention mechanisms are pivotal for modeling long-range dependencies and suppressing environmental noise, particularly in unconstrained scenarios; CA optimizes channel-wise feature recalibration to amplify emotion-relevant signals; and MHSCA integrates spatial and channel attentions to localize key facial regions. The full DAF-MHSCA model achieves state-of-the-art results by synergistically combining these elements, ensuring robustness against subtle inter-class differences and intra-class variations, as validated by the superior accuracy metrics. This ablation analysis confirms that the integrated architecture is not merely additive but multiplicative, with each component providing complementary enhancements to feature representation and generalization.

**Table 4.** Ablation experiments for RAF-DB, AffectNet-7, AffectNet-8, FER2013, FERPlus datasets.

| Method | Accuracy (%) | | | | |
|---|---|---|---|---|---|
| | **RAF-DB** | **AffectNet-7** | **AffectNet-8** | **FER2013** | **FERPlus** |
| w/o AFC | 91.69 | 67.54 | 64.88 | 72.85 | 90.64 |
| w/o DAF | 90.96 | 67.01 | 64.20 | 71.19 | 90.07 |
| w/o CA | 91.52 | 67.37 | 64.72 | 72.78 | 90.94 |
| w/o MHSCA | 91.89 | 67.50 | 65.01 | 73.46 | 91.15 |
| DAF-MHSCA | 92.49 | 67.72 | 65.20 | 74.33 | 91.38 |

### *3.7. Sensitivity Analysis*

Table 5 demonstrates the impact of varying the number of attention heads ($N_h$) on recognition accuracy across multiple datasets. The results indicate that the model achieves optimal performance when $N_h = 4$, with accuracies of 92.49% on RAF-DB, 67.72% on AffectNet-7, 65.20% on AffectNet-8, and 91.38% on FERPlus. This suggests that a moderate number of heads allows the model to effectively capture diverse and complementary spatial-channel features without introducing excessive redundancy or overfitting. Further increasing $N_h$ to 8 or 16 leads to a decline in performance, likely due to increased computational complexity and potential over-specialization of attention mechanisms, reducing generalizability. These findings underscore the importance of carefully balancing model capacity and representational power in multi-head attention architectures for facial expression recognition.

**Table 5.** Sensitivity to number of attention heads ($N_h$).

| $N_h$ | Accuracy (%) | | | |
|---|---|---|---|---|
| | **RAF-DB** | **AffectNet-7** | **AffectNet-8** | **FERPlus** |
| 1 | 90.82 | 65.60 | 64.25 | 89.91 |
| 2 | 91.74 | 66.57 | 64.83 | 90.64 |
| 4 | 92.49 | 67.72 | 65.20 | 91.38 |
| 8 | 92.07 | 65.89 | 65.02 | 90.97 |
| 16 | 91.33 | 65.42 | 64.58 | 90.12 |

Figure 8 presents the sensitivity analysis of $\theta$, which controls the attention diversity loss term in the multi-loss integration. This analysis aims to evaluate the impact of $\theta$ on classification accuracy to ensure an optimal balance among loss components. $\theta$ is searched over the range of 0.1 to 1.0 with a step size of 0.1, and the model performance is measured on the RAF-DB dataset. As shown in the Figure 8, the model achieves peak performance of 92.49% when $\theta = 0.3$. When $\theta$ is too small ($\theta < 0.3$), insufficient emphasis on diversity may lead to redundant attention heads, reducing the model's discriminative power. Conversely, higher values of $\theta$ ($\theta > 0.3$) prioritize diversity excessively, overshadowing other loss terms and impairing feature learning. In contrast, $\theta = 0.3$ strikes an effective trade-off.

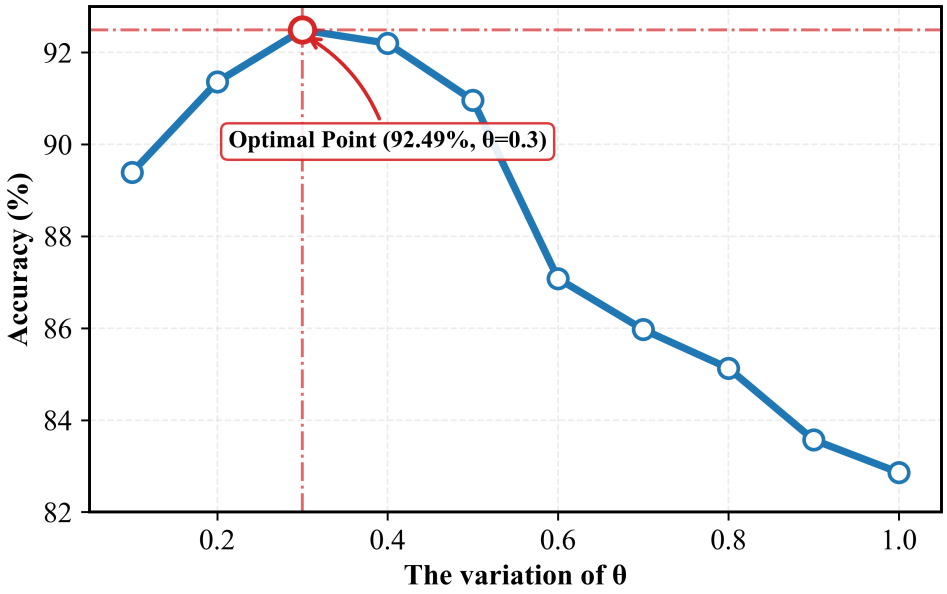


**Figure 8.** Hyperparameter Sensitivity Analysis ($\theta$) on the RAF-DB dataset.

### *3.8. Computational Efficiency Analysis*

To comprehensively evaluate the computational efficiency of the proposed DAF-MHSCA model, we conduct a comparative analysis with three state-of-the-art transformer-based methods: VTFF, GAAVE, and MMATrans. The evaluation metrics include memory usage (MB) and training time (ms/iter), which are critical for real-world deployment scenarios. As illustrated in Figure 9, a scatter plot visualizes the trade-off between these metrics, emphasizing the overall efficiency.

The results demonstrate that DAF-MHSCA achieves superior efficiency, with a memory usage of only 16.5 MB and a training time of 43 ms/iter, significantly outperforming VTFF, GAAVE, and MMATrans. This efficiency stems from the streamlined architecture of DAF-MHSCA, which integrates directional attention fusion and multi-head spatial-channel attention without excessive parameters. In contrast, the higher computational cost of VTFF aligns with its use of visual transformers and feature fusion mechanisms, while MMATrans relies on muscle relationship mining that increases memory overhead. GAAVE, though faster than VTFF, still exhibits higher memory usage due to its adversarial vulnerability estimation modules. The efficiency of DAF-MHSCA enhances its suitability for resource-constrained environments, such as embedded systems or real-time applications, without compromising accuracy.

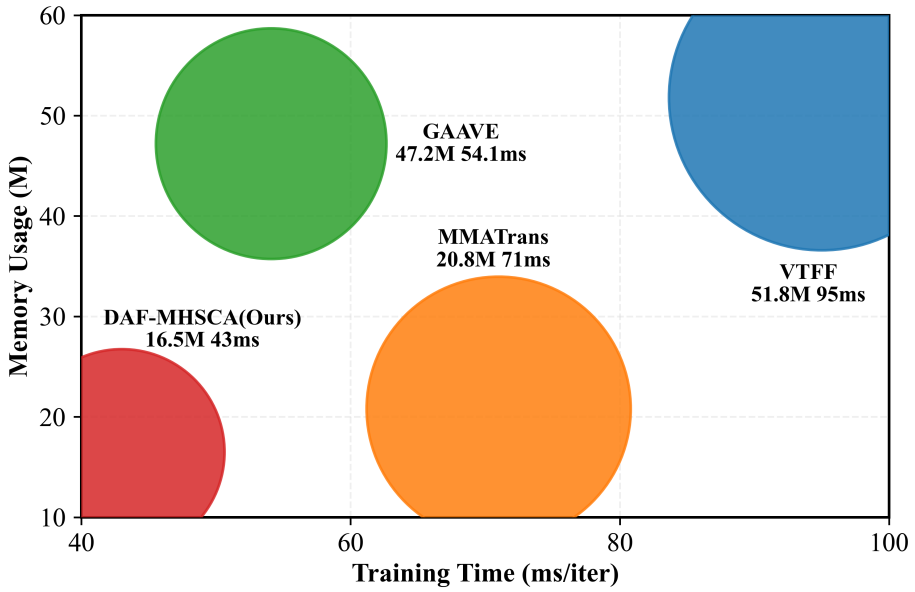


**Figure 9.** The computational efficiency comparison between our model and other transformer-based models.

## 4. Conclusions

In this paper, we propose a novel Directional Attention Fusion and Multi-Head Spatial-Channel Attention Network (DAF-MHSCA) to address the challenges of facial expression recognition in unconstrained scenarios, such as intra-class variations, subtle inter-class differences, and environmental interference. The DAF-MHSCA framework effectively captures both coarse-grained and detailed expression features through a ResNet18 backbone and an Adaptive Feature Calibration (AFC) mechanism. The introduced Directional Attention Fusion (DAF) module generates spatial attention maps through both self-attention and cross-attention mechanisms along the width and height directions. The Multi-Head Spatial-Channel Attention (MHSCA) module refines spatial and channel features for improved recognition. Extensive experiments on RAF-DB, AffectNet, FER2013, and FERPlus datasets demonstrate that DAF-MHSCA achieves state-of-the-art performance, outperforming existing methods across all benchmarks. The ablation studies further validate the contribution of each component to the overall framework. This work provides an effective solution for real-world facial expression recognition and highlights the potential of integrating directional feature extraction with hybrid attention mechanisms for fine-grained visual tasks.

While the proposed DAF-MHSCA achieves competitive performance, we acknowledge its limitations. Firstly, the incorporation of multi-head attention and directional fusion mechanisms introduces additional computational complexity compared to CNN-based architectures. However, it is noteworthy that our model is inherently more efficient than other transformer-based models, as it avoids the quadratic complexity of global self-attention by leveraging directional convolutions and localized attention mechanisms. Future work will explore lightweight attention designs for deployment on resource-constrained devices. Secondly, our model is evaluated on static images. Extending it to dynamic video sequences is a promising direction, as the directional attention mechanism could capture temporal dependencies across frames. Finally, our model relies on pre-trained weights from face recognition datasets (MS-Celeb-1M). Future work will investigate self-supervised pre-training strategies tailored to expression features to reduce this dependency.

### Author Contributions

Y.S.: data processing, result analysis, investigation, visualization; Y.L.: conceptualization, supervision, resources, writing—review and editing; B.W.: methodology, project administration, validation, writing—original draft preparation. All authors have read and agreed to the published version of the manuscript.

### Funding

This research was funded by the National Natural Science Foundation of China under Grant 623B2011; the National Natural Science Foundation of China under Grant 62325301; the National Natural Science Foundation of China under Grant U24B20186.

### Institutional Review Board Statement

Not applicable.

### Informed Consent Statement

Not applicable.

### Data Availability Statement

All data included in this study may be made available upon request via contacting the corresponding author.

### Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper. All facial images used in the study, including those in Figure 1 and Figure 7, are sourced from publicly available datasets, ensuring compliance with ethical standards and copyright regulations.

### Use of AI and AI-assisted Technologies

During the preparation of this work, the authors used Deepseek-V3 to polish the language of the manuscript. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

## References


1. Schmidt, K.L.; Cohn, J.F. Human facial expressions as adaptations: Evolutionary questions in facial expression research. *Am. J. Phys. Anthropol. Off. Publ. Am. Assoc. Phys. Anthropol.* **2001**, *116*, 3–24.
2. Li, S.; Deng, W. Deep facial expression recognition: A survey. *IEEE Trans. Affect. Comput.* **2020**, *13*, 1195–1215.
3. Li, Y.; Yang, G.; Su, Z.; et al. Human activity recognition based on multienvironment sensor data. *Inf. Fusion* **2023**, *91*, 47–63.
4. Ekundayo O; Viriri S. Facial expression recognition: a review of methods, performances and limitations. In Proceedings of the 2019 Conference on Information Communications Technology and Society (ICTAS), Durban, South Africa, 6–8 March 2019.
5. Kopalidis, T.; Solachidis, V.; Vretos, N.; et al. Advances in facial expression recognition: a survey of methods, benchmarks, models, and datasets. *Information* **2024**, *15*, 135.
6. Pham, T.D.; Duong, M.T.; Ho, Q.T.; et al. CNN-based facial expression recognition with simultaneous consideration of inter-class and intra-class variations. *Sensors* **2023**, *23*, 9658.
7. Li, S.; Deng, W. Reliable Crowdsourcing and Deep Locality-Preserving Learning for Unconstrained Facial Expression Recognition. *IEEE Trans. Image Process.* **2019**, *28*, 356–370.
8. Wen, Z.; Lin, W.; Wang, T.; et al. Distract your attention: Multihead cross attention network for facial expression recognition. *Biomimetics* **2023**, *8*, 199.
9. Zhang, S.; Zhang, Y.; Zhang, Y.; et al. A dual direction attention mixed feature network for facial expression recognition. *Electronics* **2023**, *12*, 3595.
10. Cabacas-Maso, J.; Ortega-Beltrán, E.; Benito-Altamirano, I.; et al. Enhancing facial expression recognition through dual-direction attention mixed feature networks: Application to 7th ABAW challenge. In Proceedings of the European Conference on Computer Vision (ECCV), Milan, Italy, 29 September–4 October 2024.
11. Mollahosseini, A.; Hasani, B.; Mahoor, M.H. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Trans. Affect. Comput.* **2017**, *10*, 18–31.
12. He, K.; Zhang, X.; Ren, S.; et al. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016.
13. Guo, Y.; Zhang, L.; Hu, Y.; et al. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 11–14 October 2016.
14. Uniyal, S.; Agarwal, R.. Analyzing Facial Emotion Patterns in AffectNet with Deep Neural Networks. In Proceedings of the 2024 1st International Conference on Advances in Computing, Communication and Networking (ICAC2N), Greater Noida, India, 16–17 December 2024.
15. Song, C.H.; Han, H.J.; Avrithis, Y. All the attention you need: Global-local, spatial-channel attention for image retrieval. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 3–8 January 2022.
16. Li, H.; Sui, M.; Zhao, F.; et al. MVT: Mask vision transformer for facial expression recognition in the wild. *arXiv* **2021**, arXiv:2106.04520.
17. Farzaneh, A.H.; Qi, X. Facial expression recognition in the wild via deep attentive center loss. In Proceedings of the IEEE/CVF winter conference on applications of computer vision (WACV), Virtual, 5–9 January 2021.
18. Ma, F.; Sun, B.; Li, S.. Facial expression recognition with visual transformers and attentional selective fusion. *IEEE Trans. Affect. Comput.* **2021**, *14*, 1236–1248.
19. Liu, H.; Cai, H.; Lin, Q.; et al. FEDA: Fine-grained emotion difference analysis for facial expression recognition. *Biomed. Signal Process. Control* **2023**, *79*, 104209.
20. Zheng, J.; Li, B.; Zhang, S.; et al. Attack can benefit: An adversarial approach to recognizing facial expressions under noisy annotations. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), Washington, DC, USA, 7–14 February 2023.
21. Liu, H.; Zhou, Q.; Zhang, C.; et al. MMATrans: Muscle movement aware representation learning for facial expression recognition via transformers. *IEEE Trans. Ind. Inform.* **2024**, *20*, 13753–13764.
22. Xu, J.; Li, Y.; Yang, G.; et al. Multiscale facial expression recognition based on dynamic global and static local attention. *IEEE Trans. Affect. Comput.* **2025**, *16*, 683–696.