

Research Article

Open Access



MSAFNet: a novel approach to facial expression recognition in embodied AI systems

Huifang He¹, Runbin Liao², Yating Li³

¹School of Information Engineering, Guangdong Engineering Polytechnic, Guangzhou 510520, Guangdong, China.

²School of Physics and Telecommunication Engineering, South China Normal University, Guangzhou 510006, Guangdong, China.

³School of Data Science and Engineering, and Xingzhi College, South China Normal University, Shanwei 516600, Guangdong, China.

Correspondence to: Prof. Huifang He, School of Information Engineering, Guangdong Engineering Polytechnic, No. 123 Keji Road, Tianhe District, Guangzhou 510520, Guangdong, China. E-mail: hehuifang@gdep.edu.cn; ORCID: 0000-0002-7204-0736

How to cite this article: He, H.; Liao, R.; Li, Y. MSAFNet: a novel approach to facial expression recognition in embodied AI systems. *Intell. Robot.* 2025, 5(2), 313-32. <http://dx.doi.org/10.20517/ir.2025.16>

Received: 24 Oct 2024 **First Decision:** 13 Feb 2025 **Revised:** 11 Mar 2025 **Accepted:** 13 Mar 2025 **Published:** 11 Apr 2025

Academic Editor: Simon Yang **Copy Editor:** Pei-Yun Wang **Production Editor:** Pei-Yun Wang

Abstract

In embodied artificial intelligence (EAI), accurately recognizing human facial expressions is crucial for intuitive and effective human-robot interactions. We introduce multi-scale attention and convolution-transformer fusion network, a deep learning framework tailored for EAI, designed to dynamically detect and process facial expressions, facilitating adaptive interactions based on the user's emotional state. The proposed network comprises three distinct components: a local feature extraction module that utilizes attention mechanisms to focus on key facial regions, a global feature extraction module that employs Transformer-based architectures to capture comprehensive global information, and a global-local feature fusion module that integrates these insights to enhance facial expression recognition accuracy. Our experimental results on prominent datasets such as FER2013 and RAF-DB indicate that our data-driven approach consistently outperforms existing state-of-the-art methods.

Keywords: Facial expression recognition, multi-scale attention, feature fusion, data-driven

1. INTRODUCTION

Facial expression, serving as one of the most direct and natural social signals in human communication^[1], holds a critical role in interpersonal interactions and is a vital conduit for emotional exchange^[2]. The ability to interpret these expressions accurately is fundamental to the paradigm of embodied artificial intelligence (EAI),



© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, sharing, adaptation, distribution and reproduction in any medium or format, for any purpose, even commercially, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.



where systems interact with their environment in a meaningful way. In EAI, recognizing and responding to human emotions through facial expressions can significantly enhance the adaptability and functionality of these systems in various applications such as human-computer interaction, smart healthcare, and safety monitoring in driving scenarios.

The complexity of facial expressions incorporates dynamic changes in facial muscle movements and subtle variations in facial features, which reflect a person's emotional state and intentions. As shown in Figure 1, Ekman *et al.* categorized facial expressions into seven basic emotions: anger, disgust, fear, happiness, sadness, surprise, and neutral, which have been extensively used as a foundation in the development of facial expression recognition (FER) technologies^[3]. As computer vision technology has advanced, data-driven approaches to FER have progressively become more sophisticated, adapting to the diverse and spontaneous nature of human expressions captured in both controlled laboratory environments and in more challenging in-the-wild settings.

Despite the advancements in FER algorithms, the performance of these systems still requires enhancements to cope with the diversity and complexity of real-world human expressions. This ongoing development is emblematic of EAI's core challenge: to create systems that not only perceive but also understand and appropriately respond to human cues in a manner that mirrors human cognitive abilities. Datasets for FER, such as CK+^[4], JAFEE^[5], Oulu-CASIA^[6], RAF-DB^[7], SFEW^[8], and FERPlus^[9], play a crucial role in training these intelligent systems, offering a spectrum of expressions from controlled to naturalistic environments. These datasets help in refining the algorithms to achieve higher accuracy and reliability in expression recognition, thus enabling EAI systems to engage more naturally and effectively with humans.

Convolution neural network (CNN) has achieved a significant performance in FER in other fields. Mollahosseini *et al.* designed a deep neural network consisting of two convolution layers, two pooling layers, and four inception layers, with single-component architecture^[10]. It achieved satisfactory results on different public datasets. Shao *et al.* proposed three different kinds of convolutional neural networks: Light-CNN, dual-branch CNN, and pre-trained CNN, which achieved robust results for facial expression in the wild^[11]. Gurseli *et al.* designed a lightweight CNN for facial emotion recognition, called custom lightweight CNN-based model (CLCM), based on MobileNetV2 architecture^[12]. It achieved performance comparable to or better than MobileNetV2 and ShuffleNetV2.

The advantage of CNN is that it performs local information exchange in a region through convolution operation, which focuses on modeling local relationships. Each convolutional filter is made for a small region. Although CNN can extract more abstract features at a deep level, it still cannot extract enough global features for FER. In contrast, the visual transformer can capture long-distance dependencies between pixels through a self-attention mechanism, which have advantages for global feature extraction and can compensate for the shortcomings of CNN.

Attention mechanisms have also been extensively used to solve the problems of occlusion and pose variation in FER. In FER tasks, the useful features for recognition mainly focus on key areas such as the eyes, nose, and mouth. The attention mechanism improves expression recognition by increasing the weights of these key features. Sun *et al.* proposed AR-TE-CATFFNet integrating three core components: attention-rectified convolution for feature selection, local binary pattern (LBP)/GLCM-based texture enhancement, and cross-attention transformers to fuse RGB-texture features globally, achieving enhanced accuracy and cross-domain generalization^[13]. Tao *et al.* introduced a hierarchical attention network with progressive fusion of local-global contexts and illumination-robust gradients through hierarchical attention modules (HAM), adaptively amplifying discriminative facial cues while suppressing irrelevant regions for improved robustness^[14]. A multilayer perceptual attention network was presented by Liu *et al.* and is capable of learning the potential diversity and essential details of various expressions^[15]. Furthermore, the perceptual attention network can



Figure 1. The samples of 7 basic emotions from RAF-DB, FERPlus and FER2013.

adaptively focus on the local regions with robustness to different datasets. In addition, Zhao *et al.* designed a geometry-guided framework integrating GCN-transformers, constructing spatial-temporal graphs from facial landmarks to model local/non-local dependencies, and employing spatiotemporal attention to prioritize critical regions/frames for video emotion recognition^[16].

To overcome the above shortcomings, in this paper, we propose an end-to-end multi-scale attention (MSA) and convolution-transformer fusion network (MSAFNet) for FER tasks, which can learn local and global features and adaptively model the relationship between them. Our proposed network has three components: the local feature extraction module (LFEM), the global feature extraction module (GFEM), and the global-local feature fusion module (GLFM). A MSA block is embedded into the LFEM, which can adaptively capture the importance of relevant regions of FER, effectively overcoming the inherent limitations of traditional single-scale feature modeling. The proposed MSA block can capture key facial information from different perspectives and improve the performance in occlusion and pose variation conditions. The GFEM can compensate for the shortcomings of the LFEM by capturing long-distance relationships from global images. We designed the GLFM to model the relationship between local and global features. The synergistic operation of these modules significantly enhances micro-expression sensitivity and cross-domain generalization capabilities, with the fusion mechanism dynamically recalibrating feature importance to optimize recognition performance under real-world complexities.

In summary, the main contributions of our work are as follows:

1. A MSAFNet for FER tasks has been proposed, which can capture key information from local and global features and adaptively model the relationship between them.
2. A LFEM and a GFEM have been proposed. Furthermore, a MSA block is designed to embed in the LFEM, which can combine the attention information of spatial dimension with channel dimension without cropping strategies and facial landmark detectors.
3. A GLFM to model the relationship between local and global features has been designed, which can effectively improve recognition performance.
4. Experimental results on three different FER datasets show that MSAFNet obtains competitive results compared with other state-of-the-art methods, proving our model's validity.

2. RELATED WORK

In recent years, many effective FER methods have been proposed. In this part, we mainly introduce previous related methods.

2.1. FER based on traditional approaches

Early FER works mainly on hand-crafted features and traditional machine-learning classification methods. The hand-crafted features can be divided into appearance-based features and geometric features. The commonly used appearance-based features include LBP^[17], Gabor wavelets^[5], and histogram of oriented gradients (HOG)^[18]. Geometric features are obtained by measuring the relative position of significant features, such as eyes, nose, and mouth^[19,20]. Moreover, support vector machine (SVM)^[21] is the most common and effective machine learning classification algorithm. Ghimire *et al.* proposed a FER method using a combination of appearance and geometric features with SVM classification^[21]. Although the traditional methods have a good performance on in-the-lab FER datasets, the performance on in-the-wild FER datasets is significantly degraded. It can be ascribed that lighting, noise, and other factors can easily affect hand-crafted features. Moreover, the method provides better results in facial expression datasets.

2.2. FER based on CNN models

Compared with traditional machine learning methods, deep neural networks, especially CNN, can learn directly from the input reducing the dependence on pre-processing. With the rapid development of deep learning, many deep neural networks such as AlexNet^[22], VGG^[23], and ResNet^[24] are widely used in FER tasks and have shown good performance. Wu *et al.* proposed FER-CHC with cross-hierarchy contrast to enhance CNN-based models by critical feature exploitation^[25]. Teng *et al.* designed typical facial expression network (TFEN) combining dual 2D/3D CNNs for robust video FER across four benchmarks^[26]. Zhao *et al.* developed a cross-modality attention CNN (CM-CNN) that fused grayscale, LBP, and depth features via hierarchical attention mechanisms, effectively addressing illumination/pose variations and enhancing recognition of subtle expressions^[27]. Cai *et al.* introduced probabilistic attribute tree CNN (PAT-CNN) addressing identity-induced intra-class variations through probabilistic attribute modeling^[28]. Liu *et al.* combined CNN-extracted facial features with GCN-modeled high-aggregation subgraphs (HASs) to boost recognition robustness^[29].

2.3. FER based on attention mechanism

Attention mechanism has been widely applied in FER tasks out of their effectiveness in focusing the network on useful regions relevant to expression recognition. Zhang *et al.* proposed a cross-fusion dual-attention network with three innovations: grouped dual-attention for multi-scale refinement, adaptive C2 activation mitigating computational bottlenecks, and distillation-residual closed-loop framework enhancing feature purity^[30]. Li *et al.* developed SPWFA-SE combining Slide-Patch/Whole-Face attention with SE blocks to jointly capture local details and global contexts, improving FER accuracy^[31]. Zhang *et al.* designed lightweight GSDNet using gradual self-distillation for inter-layer knowledge transfer and ACAM with learnable coefficients for adaptive enhancement^[32]. Tao *et al.* introduced a hierarchical attention network integrating local-global gradient features via multi-context aggregation, employing attention gates to amplify discriminative regions^[14]. Chen *et al.* introduced a hierarchical attention network integrating local-global gradient features via multi-context aggregation, employing attention gates to amplify discriminative regions^[33].

2.4. FER based on visual transformer

Transformers^[34] have been widely used in natural language processing (NLP) tasks and have shown significant performance. They are good at capturing the long-distance relation between words by their self-attention mechanism. Inspired by the success of transformers, Dosovitskiy *et al.* proposed ViT^[35], a pure transformer, applied to image patches on classification tasks and has shown significant performance in the field of computing vision, such as object detection^[36], object tracking^[37], and instance segmentation^[38]. Visual transformers are also applied to FER by some researchers. Ma *et al.* introduced a transformer-augmented network (TAN)

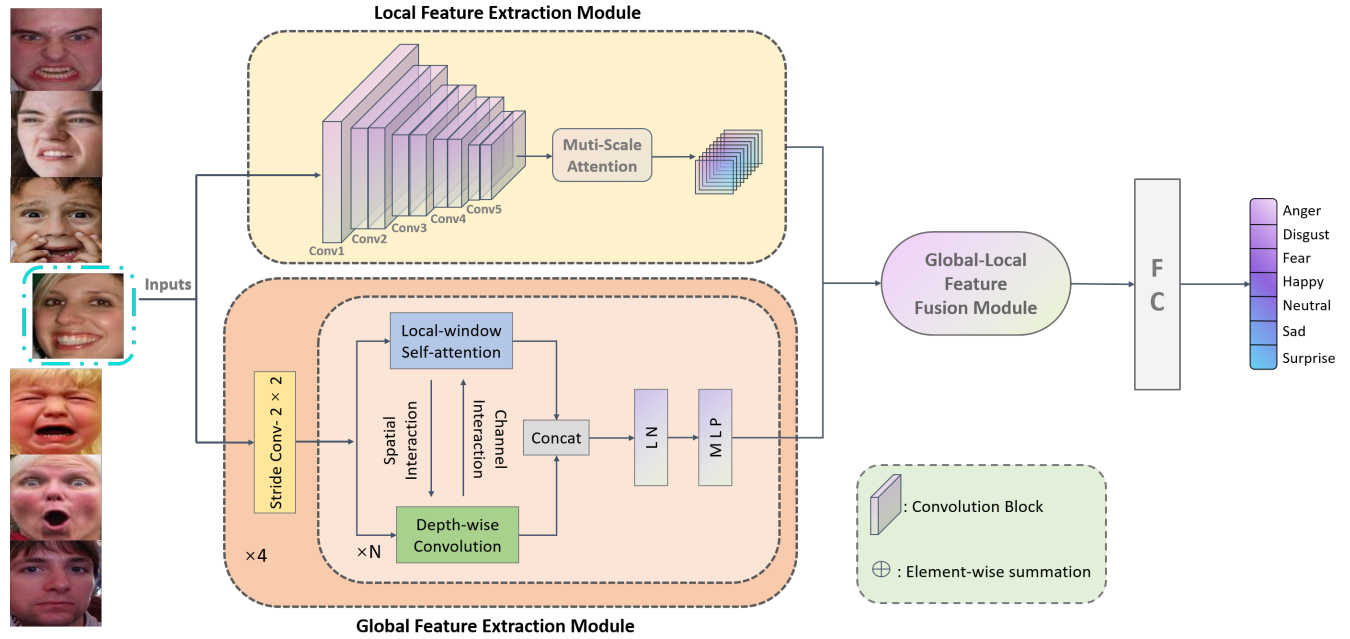


Figure 2. The overview of our proposed method for FER. The proposed method consists of three components, including the LFEM, the GFEM and GLFM. The images and labels are from RAF-DB. FER: Facial expression recognition; LFEM: local feature extraction module; GFEM: global feature extraction module; GLFM: global-local feature fusion module.

combining intra-patch transformers (position-disentangled attention) and inter-patch transformers to capture local features and cross-region dependencies, integrated with online label correction for noise reduction^[39]. Zhang *et al.* developed transformer-based multimodal emotional perception (T-MEP) with triple transformers for audio/image/text features, aligning multimodal semantics through fused self/cross-attention in visual latent space^[40]. Liu *et al.* proposed patch attention convolutional vision transformer (PACVT) that also extracts local and global features, but uses simple add operation to combine different features^[41]. Different from them, our method can obtain rich emotional information from local and global features without a cropping strategy, and employs a learnable way to integrate the relation of local and global features. It is effective to recognize facial expression images.

3. PROPOSED METHOD

3.1. Overview

As shown in Figure 2, our proposed MSAFNet consists of three components, including the LFEM, the GFEM, and the GLFM. The LFEM introduces a CNN (ResNet) as the backbone to extract local features. Specifically, an original facial expression image is fed into the LFEM and the GFEM as input. Concurrently, a MSA block is embedded into the local module to direct the model to focus more on regions that are crucial for expression recognition. The GFEM also converts the original facial image to different tokens with positional information by a linear embedding layer. Then the tokens are fed to the transformer encoder, which can extract global features of the original image. Finally, the GLFM compatibly models the relationship between local and global features, and the output features are utilized for FER.

3.2. LFEM

Inspired by Res2Net^[42], we design a MSA block. In this module, we manipulate ResNet18^[24] as the backbone from given facial images for the LFEM, and the average pool and fully connected layer are removed after the Conv5 block. To optimize the salient information in local regions, this module designs a MSA block which is

embedded after the Conv5 block.

As shown in Figure 3, the feature maps extracted after the Conv5 block are fed into the attention block. After a 1×1 convolution, the feature maps $X1 \in R^{C1 \times H1 \times W1}$ are obtained, where $C1$, $H1$, and $W1$ represent the number of channels, width, and height of the feature map after the convolution operation, respectively. The module splits the feature maps $X1$ into s groups feature map subsets, denoted by x_i , where $i \in \{0, 1, 2, \dots, s-1\}$. The spatial size of each group feature map subset is the same as the input feature maps X , while the number of channels is $C1' = C1/s$. The i -th group feature map subset import $x_i \in R^{C1' \times H1 \times W1}$, $i \in \{0, 1, 2, \dots, s-1\}$. Each x_i is processed by a corresponding 3×3 convolution designated by $f_i(\cdot)$, and the output is denoted by $y_i \in R^{C1' \times H1 \times W1}$. According to the module, the input and output have the same dimension. When $i \geq 1$, the i -th group feature subset is computed with the output of y_{i-1} and then fed as the input of $f_i(\cdot)$. Thus, y_i can be written as:

$$y_i = \begin{cases} f_i(x_i) & i = 0 \\ f_i(x_i + y_{i-1}) & 1 \leq i \leq s-1 \end{cases} \quad (1)$$

When each group feature subset $\{x_i, 0 \leq i \leq s-1\}$ goes through 3×3 convolution, the output $\{y_i, 1 \leq i \leq s-1\}$ can acquire a larger receptive field than $\{y_i, j \leq i\}$. After that, each y_i can contain characteristic information of feature subsets with different receptive field scales and different scales, thus obtaining multi-scale spatial information. Different sizes of s can learn different information, and larger s may get richer scale information. This module sets the size of s as 4, which was carefully chosen through comprehensive ablation studies to balance computational complexity and model performance. This systematic analysis justifies our design choice of $s = 4$ as the optimal configuration that achieves superior accuracy while maintaining reasonable computational demands.

Following the multi-scale spatial attention information, we subsequently compute attention weights along channel dimensions. By using global average pooling, each output y_i from a convolution operation for each group feature subset x_i is condensed into a vector. Then, we employ two fully connected layers to model the channel correlations. In neural networks, activation functions are primarily used to introduce non-linearity, enabling the network to learn complex patterns. Therefore, we use a sigmoid activation function to normalize the output, which can obtain the channel attention weight of each group feature subset $F_i \in R^{C1' \times 1 \times 1}$. It can be defined as:

$$F_i = \sigma(W_2 \rho(W_1 y_i)) \quad (2)$$

where σ denotes the sigmoid function, which normalizes the output as a range of $[0, 1]$, effectively transforming the output into a probability value. Additionally, ρ denotes the ReLU activation function, a typical non-linear function, defined as $f(x) = \max(0, x)$, which maps the input signal to the feature space; W_1 and W_2 denote the FC operation; F_i denotes the channel attention weight of different group feature subsets. Furthermore, it splices the attention weights to acquire the final MSA weights $F \in R^{C1 \times 1 \times 1}$:

$$F = \text{Concat}([F_0, F_2, F_3, \dots, F_{s-1}]) \quad (3)$$

$$X_{local} = X1 \otimes F \quad (4)$$

Finally, we acquire the output X_{local} by multiplying feature maps X with the MSA weights F .

Different from Res2Net, our MSA not only captures the multi-scale information from feature map subsets, but also calculates channel information and aggregates all information from feature map subsets, which can make the attention information richer. It considers both spatial semantic information and channel semantic information and effectively combines information from both spatial and channel dimensions. This design, leveraging self-attention, reduces redundancy, accelerates training, and improves convergence. By emphasizing comprehensive feature fusion and diverse interactions, our MSA ensures more efficient information flow, better addressing gradient vanishing and enhancing training stability in deep architectures.

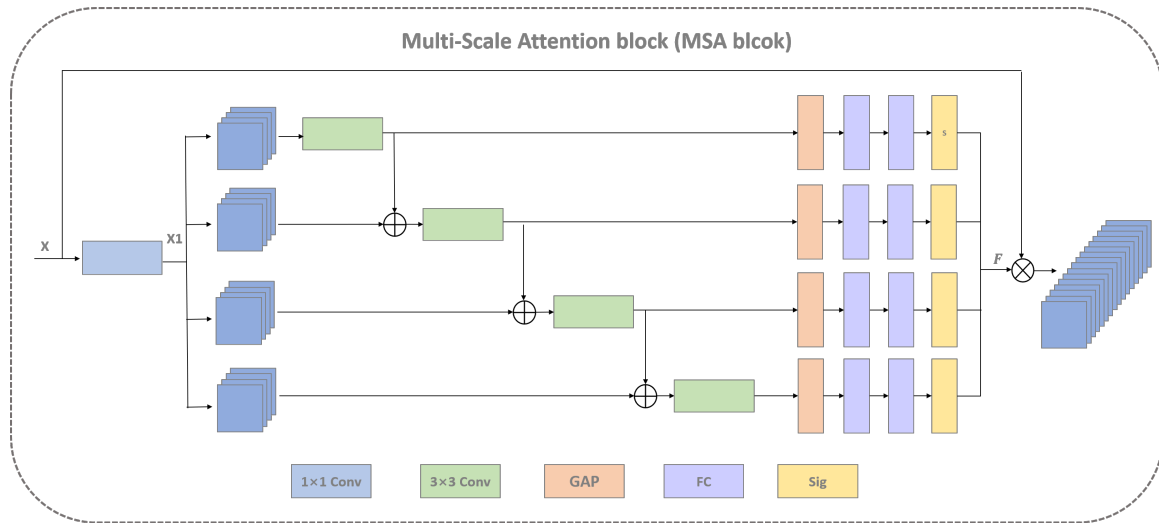


Figure 3. The details of MSA block. MSA: Multi-scale attention; GAP: global average pooling; FC: fully connected layer; Sig: sigmoid function.

3.3. GFEM

Given the significant performance of vision transformers (ViT)^[35] in numerous computer vision tasks, several transformer architectures have been widely adopted. Transformer architecture is good at capturing the long-distance dependencies between pixels. Therefore, the GFEM utilizes a transformer architecture following Mixformer^[43]. As shown in Figure 2, the GFEM combines local-window self-attention (W-MSA) with depth-wise convolution in a parallel design, providing complementary information for each branch. The ViT branch employs W-MSA to model global facial semantics: non-overlapping patches undergo dynamic correlation analysis. Channel interaction (CI_{out}) enhances critical regions via element-wise recalibration. The CNN branch extracts localized details via depthwise convolution, with spatial interaction (SI_{out}) amplifying discriminative regions. Fused features combine both streams channel-wise, processed by the MIX function for robust classification. This dual-stream design enables complementary modeling: W-MSA captures spatial dependencies, while CNN optimizes channel-wise features, achieving multi-scale representation through hierarchical fusion. The channel interaction contains a global average pooling layer and two 1×1 convolution layers with BN layer and GELU activation function. And a sigmoid function is used to generate attention. At last, the channel interaction is applied to the value in W-MSA. The spatial interaction involves two 1×1 convolution layers with BN and GELU, followed by a sigmoid function used to generate attention.

$$CI_{out} = \sigma(\text{Conv}_{1 \times 1}(\text{GELU}(\text{Conv}_{1 \times 1}(\text{GAP}(CI_{in})))))) \quad (5)$$

$$SI_{out} = \sigma(\text{Conv}_{1 \times 1}(\text{GELU}(\text{Conv}_{1 \times 1}(SI_{in})))) \quad (6)$$

Where CI_{in} , CI_{out} represent the input and output of channel interaction; SI_{in} , SI_{out} represent the input and output of spatial interaction; GELU denotes activation function; σ denotes the sigmoid function.

Based on the parallel design, the mix transformer block can be formulated as follows:

$$\hat{z}^{l+1} = \text{CONCAT}(W - \text{MSA}(z^l), \text{Conv}(z^l)) + z^l \quad (7)$$

$$z^{l+1} = \text{MLP}(\text{LN}(\hat{z}^{l+1})) + \hat{z}^{l+1} \quad (8)$$

where CONCAT represents a function that mixes the feature between W-MSA and depth-wise convolution. Conv denotes depth-wise convolution; LN represents layer normalization; \hat{z}^{l+1} and z^{l+1} denote the output features of the CONCAT and the MLP , respectively.

3.4. GLFM

In this field, GLFM effectively integrates features from the LFEM and the GFEM. The primary objective is to combine local and global features to capture richer spatial and channel information, thereby enhancing FER performance. Compared with traditional neural networks that separately process local and global features, our model enables the complementarity between local and global features. As shown in Figure 4, this approach allows for a more accurate description of both the details and overall characteristics of facial expressions. Given local feature maps $X_{local} \in R^{C1 \times H1 \times W1}$ extraction from the LFEM, where $C1, H1, W1$ are the channel dimension, height and width, and global feature maps $X_{global} \in R^{C2 \times L2}$ extracted from the GFEM, where $C2, D2$ are channel dimensions and token numbers. The local feature maps $X_{local} \in R^{C1 \times H1 \times W1}$ are rearrange to $X_{local} \in R^{C1 \times L1}$ where $L1 = H1 \times W1$. This rearrangement of local feature maps is an essential step in the process. By reshaping the feature maps, local and global information can be effectively fused within the same dimensional space, thereby optimizing subsequent operations. This paper presents the first method to fuse local feature maps X_{local} and global feature maps X_{global} via an element-wise summation for information interaction:

$$X_{fusion} = X_{local} \oplus X_{global} \quad (9)$$

Where $X_{fusion} \in R^{C3 \times L3}$ is the interactive features, and \oplus denotes element-wise summation. For purpose of performing information interaction better, the interaction features X_{fusion} are transposed into $x_{fusion}^T \in R^{C3 \times L3}$ and fed into MLP block including two fully connected layers and a non-linearity layer that can c the information on the token level. Employing this method can acquire the feature $x_{fusion1}^T \in R^{C3 \times L3}$ and transpose it into $X_{fusion1} \in R^{C3 \times L3}$. And then the features $X_{fusion1}$ are fed into two different MLP blocks to interact the information on the channel level. The details can be defined as follows:

$$X_{fusion1}^T = W_4 \text{GELU} \left(W_3 \text{LN} \left(x_{fusion}^T \right) \right) \quad (10)$$

$$X_{fusion2} = W_6 \text{GELU} \left(W_5 \text{LN} \left(X_{fusion1} \right) \right) \quad (11)$$

$$X_{fusion3} = W_8 \text{GELU} \left(W_7 \text{LN} \left(X_{fusion1} \right) \right) \quad (12)$$

where LN denotes layer normalization, which is the process of normalizing the output of a specific layer in a neural network; it helps maintain stability during training by preventing difficulties caused by large differences in the outputs of different layers. This can reduce training time, improve stability, and enhance convergence. Additionally, GELU denotes the GELU activation function, which is similar to ReLU but with smoother output for small negative values. It helps mitigate issues such as gradient explosion or vanishing gradients during training and is commonly used in deep learning to transform the model's output. W_3, W_4, W_5, W_6, W_7 and W_8 denote the fully connected operation. And we can get the last fusion features as:

$$X_{Fin} = (X_{local} \otimes \sigma(X_{fusion2})) \oplus (X_{global} \otimes \sigma(X_{fusion3})) \quad (13)$$

where \otimes denotes element-wise multiplication, and \oplus denotes element-wise summation.

4. EXPERIMENT RESULTS

4.1. Datasets

RAF-DB^[7] is a real-world FER dataset containing facial images downloaded from the Internet. The facial images are labeled with seven classes of basic expressions or 12 classes of compound expressions by 40 trained human coders. In our experiment, the proposed method only utilizes seven basic expressions, including anger, disgust, fear, happiness, sadness, surprise, and neutral. It involves 12,271 images for training and 3,069 images for testing.

FER2013^[44] is collected from the Internet that was first for ICML 2013 Challenges in representation learning. It contains 385,887 facial images collected by Google search engine with 28,709 images for training, 3,589

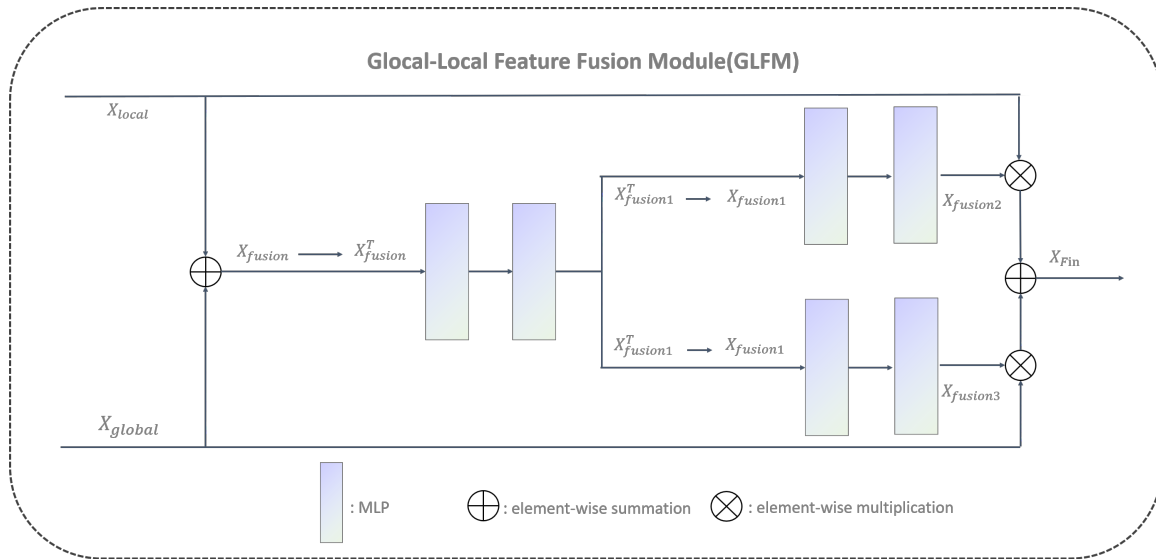


Figure 4. The details of GLFM. MLP denotes multilayer perceptron, \oplus denotes element-wise summation, and \otimes denotes element-wise multiplication. GLFM: Global-local feature fusion module.

images for validating, and 3,589 images for testing. All images are grayscale with the size of 48×48 pixels. The facial images are also labeled with seven basic expression labels, including anger, disgust, fear, happiness, sadness, surprise, and neutral.

FERPlus^[9] is extended from FER2013 dataset which was for ICML 2013 Challenges. It also contains 28,709 training, 3,589 validation, and 3,589 testing images. All images are 48×48 pixels. Different from FER2013 dataset, each image in FERPlus is relabeled by ten annotators and FERPlus adds three categories, including contempt, unknown, and not a face. Therefore, FERPlus dataset has better quality labels than FER2013. The proposed algorithm reports overall sample accuracy on the test set with eight categories.

Occlusion and pose variant datasets are test subsets of RAF-DB and FERPlus collected by the work^[45]. These datasets are Occlusion-RAF-DB, Pose-RAF-DB, Occlusion-FERPlus, and Pose-FERPlus, and the facial expression images are manually annotated with various occlusion and variant poses. The pose datasets can be divided into two types with poses larger than 30 degrees and larger than 45 degrees.

4.2. Implementation details

In the experiments, the multitask cascaded convolutional network (MTCNN)^[46] is employed for face detection in images. The detected faces were subsequently cropped and further downsized to 256×256 pixels. MTCNN was chosen due to its superior accuracy and speed. Its cascaded network structure enables high-precision detection and effectively handles challenging scenarios, including varying illumination, pose variations, and occlusions. Furthermore, the multitask learning framework of MTCNN allows for simultaneous face localization and keypoint detection, thereby enhancing the efficiency of subsequent processing tasks. To prevent over-fitting, we randomly cropped the facial photos to 224×224 pixels during training. We also randomly flip the data during data augmentation. The ResNet18 is pre-trained on MS-Celeb-1M face recognition dataset^[47] similarly to region attention networks (RAN)^[45] and self-cure networks (SCN)^[48]. In the experiments, the model is trained with stochastic gradient descent (SGD) optimizer algorithm. The momentum is 0.9, the weight decay is 0.0005 and the batch size is set to 128. The learning rate is initialized as 0.01 and it decays to 0.9 every five epochs after 30 epochs. And we train the model for 150 epochs in total. Our method is implemented with Pytorch toolbox on GeForce RTX 2080Ti GPU platform.

Table 1. Comparison with other methods on RAF-DB dataset

Methods	Year	Accuracy (%)
gACNN ^[49]	2018	85.07
RAN ^[45]	2020	86.90
SCN ^[48]	2020	87.01
OADN ^[50]	2020	87.16
DACL ^[51]	2021	87.78
MA-Net ^[52]	2021	88.40
FDRL ^[53]	2021	89.47
VTFF ^[54]	2021	88.14
AMP-Net ^[15]	2022	89.25
ADDL ^[55]	2022	89.34
PACVT ^[41]	2023	88.21
GSDNet ^[32]	2024	90.91
DBFN ^[56]	2024	87.65
MSAFNet(ours)	2025	90.06

The bold format is used to indicate the best (highest) accuracy. gACNN: Region attention mechanism; RAN: region attention networks; SCN: self-cure networks; OADN: occlusion-adaptive deep network; DACL: deep attentive center loss; MA-Net: multi-scale and local attention network; FDRL: feature decomposition and reconstruction learning; VTFF: visual transformers with feature fusion; AMP-Net: adaptive multilayer perceptual attention network; ADDL: adaptive deep disturbance-disentangled learning; PACVT: patch attention convolutional vision transformer; DBFN: dual-branch fusion network; MSAFNet: multi-scale attention and convolution-transformer fusion network.

4.3. Comparison with state-of-the-arts

This section compares the proposed approach MSAFNet with several state-of-the-art methods on RAF-DB, FERPlus, FER2013, Occlusion-RAF-DB, Pose-RAF-DB, Occlusion-FERPlus, and Pose-FERPlus. MSAFNet consistently achieves high accuracy and demonstrates stable performance across these benchmarks. Notably, it exhibits strong generalization capabilities, particularly in complex scenarios involving diverse facial expressions and emotion categories.

4.3.1 Results on RAF-DB

Comparison results with other state-of-the-art methods on RAF-DB in recent years with seven emotion categories are shown in Table 1. Multi-scale and local attention network (MA-Net)^[52] utilized global and local features to address the issues both occlusion and pose variation and got an accuracy of 88.40%. Adaptive multilayer perceptual attention network (AMP-Net)^[15] uses different fine-grained features to extract global, local and salient features and obtained recognition accuracy of 89.25% on RAF-DB dataset. As shown in Table 1, our proposed method MSAFNet obtains the recognition accuracy of 89.77% on RAF-DB and achieves 1.66% and 0.81% improvement compared with the MA-Net^[52] and AMP-Net^[15], respectively. Compared to the visual transformers with feature fusion (VTFF)^[54] which used transformers and attention selective fusion, our method has 1.92% improvement. PACVT^[41] can also extract local and global features with attention weights

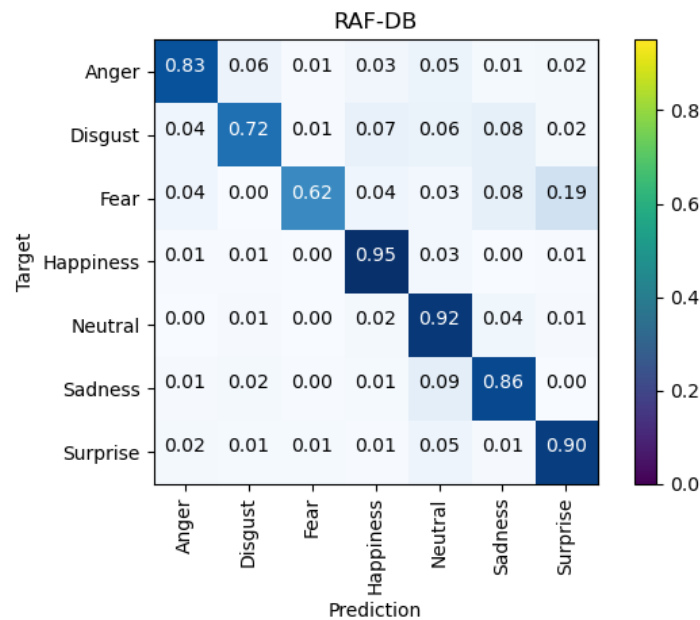


Figure 5. The confusion matrices of MSAFNet on the RAF-DB dataset.

Table 2. Comparison with other methods on FER2013 dataset

Methods	Year	Accuracy (%)
SHCNN [57]	2019	69.10
Pre-trained CNN [11]	2019	71.14
AWHFL [58]	2019	72.67
FreNet [59]	2020	64.41
LBAN-IL [60]	2021	73.11
MSAFNet(ours)	2025	73.25

The bold format is used to indicate the best (highest) accuracy. SHCNN: Shallow convolutional neural network; CNN: convolution neural network; AWHFL: adaptive weighting of handcrafted feature losses; LBAN-IL: local binary attention network with instance loss; MSAFNet: multi-scale attention and convolution-transformer fusion network.

and ViT. Compared with PACVT, our method achieves a higher accuracy by about 1.85%. In the confusion matrix shown in Figure 5, happiness expression has the highest recognition accuracy, while fear and disgust expression has poor performance. This disparity primarily arises from three factors: data scarcity, inter-class similarity and feature subtlety.

4.3.2. Results on FER2013

Table 2 shows the results that our method compares with state-of-the-art methods on FER2013 dataset. Our MSAFNet obtains an accuracy of 73.25% on FER2013 dataset, which is competitive with other advanced methods. The confusion matrix in Figure 6 shows that fear expression is the poorest to recognize and happiness has the highest recognition rate.

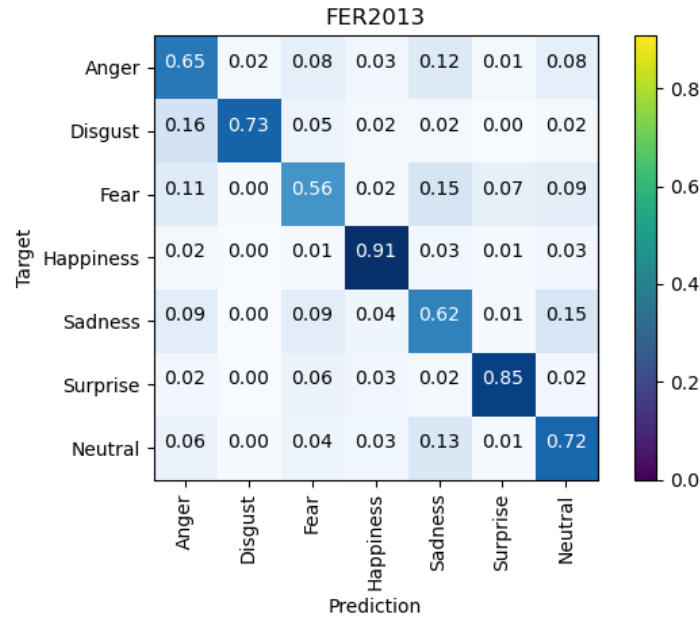


Figure 6. The confusion matrices of MSAFNet on the FER2013 dataset.

4.3.3 Results on FERPlus

Comparison results with other state-of-the-art methods on FERPlus are shown in Table 3. The recognition accuracy on FERPlus has been considerably improved when compared to the FER2013 dataset since FERPlus has been relabeled and non-face images have been removed. As shown in Table 3, our MSAFNet obtains the recognition accuracy of 89.82%. Compared to VTFF^[54] and PACVT^[41] which also utilized transformer architecture, our method achieves 1.01% and 1.1% improvement. The results of the confusion matrix on FERPlus are shown in Figure 7. The confusion matrix shows that happiness, neutral, and surprise have better performance than other expressions, and contempt, disgust, and fear have poor performance. The reason for these results may be that contempt, disgust, and fear lack enough data compared to other expressions.

4.3.4 Results on occlusion and pose variant datasets

To verify the robustness of our method under occlusion and variant pose in real-world scenarios, we conduct experiments and compare the best results with occlusion and pose variant datasets, including Occlusion-RAF-DB, Pose-RAF-DB, Occlusion-FERPlus, and Pose-FERPlus. Tables 4 and 5 show the results compared to the state-of-the-art methods on the RAF-DB and FERPlus datasets for facial occlusion and pose variants. Our MSAFNet obtains competitive performance compared to other methods. For facial occlusion datasets, it achieves superior recognition performance (86.38% and 85.62%) on the RAF-DB and FERPlus datasets. Specifically, our method outperforms the AMP-Net^[15] method by 1.1% and 0.18%, which can demonstrate the robustness of our method under facial occlusion. For pose variant datasets, our MSAFNet is significantly superior to VTFF^[54], MA-Net^[52], and AMP-Net^[15] on RAF-DB dataset with pose larger than 30 degrees and 45 degrees. On FERPlus dataset with pose larger than 30 degrees and pose larger than 45 degrees, our method also achieves higher accuracy compared with other methods. The results on occlusion and pose variant datasets demonstrate the effectiveness of our method.

4.4. Ablation analysis

To evaluate the effectiveness of our method, we perform a series of ablation studies on RAF-DB dataset. In the experiments, we evaluate the impact of the proposed components, the impact of different fusion methods, and the impact of different attention methods, respectively.

Table 3. Comparison with other methods on FERPlus dataset

Methods	Year	Accuracy (%)
CSLD [9]	2016	83.85
ResNet+VGG [61]	2017	87.40
SHCNN [57]	2019	86.54
RAN [45]	2020	88.55
RAN-VGG [45]	2021	89.16
SCN [48]	2020	88.01
VTFF [54]	2021	88.81
PACVT [41]	2023	88.72
GSDNet [32]	2024	90.32
CBAM-4CNN [62]	2024	87.75
MSAFNet(ours)	2025	89.82

The bold format is used to indicate the best (highest) accuracy. CSLD: Crowd-sourced label distribution; VGG: visual geometry group networks; SHCNN: shallow convolutional neural network; RAN: region attention networks; SCN: self-cure networks; VTFF: visual transformers with feature fusion; PACVT: patch attention convolutional vision transformer; GSDNet: gradual self distillation network; CBAM-4CNN: convolutional block attention module with convolutional neural network; MSAFNet: multi-scale attention and convolution-transformer fusion network.

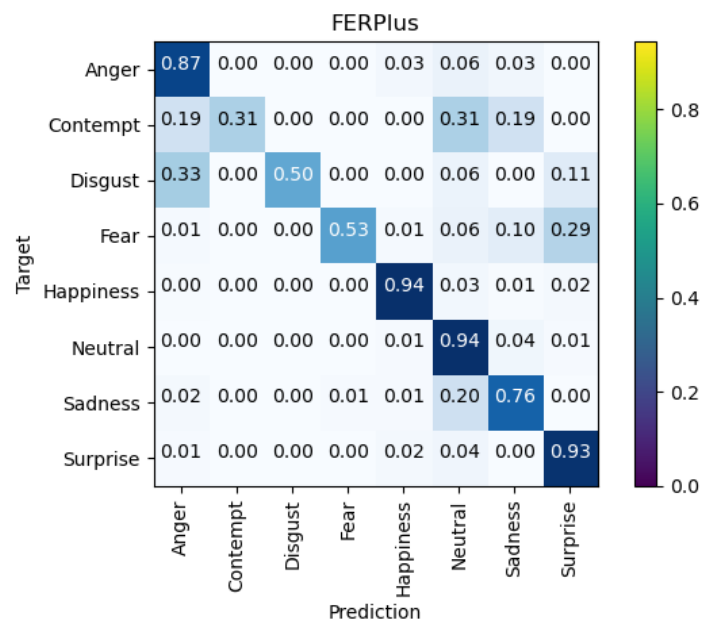
**Figure 7.** The confusion matrices of MSAFNet on the FERPlus dataset.

Table 4. Comparison with other methods on Occlusion-RAF-DB, Pose-RAF-DB

Methods	Occlusion	Pose(30)	Pose(45)
Baseline ^[45]	80.19	84.04	83.15
RAN ^[45]	82.72	86.74	85.20
MA-Net ^[52]	83.65	87.89	87.99
VTFF ^[54]	83.95	87.97	88.35
AMP-Net ^[15]	85.28	89.75	89.25
MSAFNet(ours)	86.38	90.14	89.60

The bold format is used to indicate the best (highest) accuracy. RAN: Region attention networks; MA-Net: multi-scale and local attention network; VTFF: visual transform-ers with feature fusion; AMP-Net: adap-tive multilayer perceptual attention net-work; MSAFNet: multi-scale attention and convolution-transformer fusion network.

Table 5. Comparison with other methods on Occlusion-FERPlus, Pose-FERPlus

Methods	Occlusion	Pose(30)	Pose(45)
Baseline ^[45]	73.33	78.11	75.50
RAN ^[45]	83.63	82.23	80.40
VTFF ^[54]	84.79	88.29	87.20
AMP-Net ^[15]	85.44	88.52	87.57
MSAFNet(ours)	85.62	88.63	88.78

The bold format is used to indicate the best (highest) accuracy. RAN: Region at-tention networks; VTFF: visual transform-ers with feature fusion; AMP-Net: adap-tive multilayer perceptual attention net-work; MSAFNet: multi-scale attention and convolution-transformer fusion network.

4.4.1 Impact of the proposed components

We first conduct the experiments to evaluate the impact of the proposed components, including LFEM, MSA, GFEM, and GLFM, as shown in Table 6. As we can see from the results of the first three rows in Table 6, only employing the GFEM achieves better performance compared to just utilizing the LFEM, due to the global features from GFEM capture holistic facial configurations critical for expression semantics. After adding the MSA block, the performance is 3.1% higher compared to only using LFEM. While combining GFEM, the accuracy achieves 89.18%. This hierarchical interaction ensures global-local feature complementarity GFEM suppresses LFEM's background noise, while LFEM rectifies GFEM's over-smoothing of subtle textures. With the help of the adaptively GLFM we have suggested, our approach achieves the greatest results and improves by 0.88%. The results clearly illustrate that the proposed components of our method can improve performance significantly.

4.4.2 Impact of different fusion methods

To evaluate the impact of the GLFM, we study the effects of different feature fusion strategies. As shown in Table 7, our proposed fusion method GLFM achieves 90.06% which are better result than other feature fusion strategies. The results show that our GLFM can improve the performance for FER. Compared to other fusion strategies, our GLFM employs a learnable way that can integrate local features and global features at the token level and channel level. Thus, our GLFM achieves a better performance.

Table 6. Impact of LFEM, MSA, GFEM, and GLFM on RAF-DB dataset

LFEM	MSA	GFEM	GLFM	Accuracy (%)
√				85.72
		√		86.83
√	√			88.82
√	√	√		89.18
√	√	√	√	90.06

LFEM: Local feature extraction module; GFEM: global feature extraction module; GLFM: global-local feature fusion module; MSA: multi-scale attention.

Table 7. Impact of different fusion methods

Methods	Accuracy (%)
Add	89.18
Concat	88.95
Maximum	89.02
GLFM	90.06

The bold format is used to indicate the best (highest) accuracy. GLFM: Global-local feature fusion module.

Table 8. Impact of different attention mechanisms

Methods	Accuracy (%)
SE [63]	88.20
CBAM [64]	88.23
ECA [65]	88.07
MSA	88.82

The bold format is used to indicate the best (highest) accuracy. SE: Squeeze-and-Excitation; CBAM: convolutional block attention module; ECA: efficient channel attention; MSA: multi-scale attention.

4.4.3 Impact of different attention methods

To evaluate the impact of MSA, we study the effects of different attention mechanisms, including “Squeeze-and-Excitation” (SE) [63], convolutional block attention module (CBAM) [64], and efficient channel attention (ECA) [65]. As shown in Table 8, our proposed MSA outperforms SE, CBAM, and ECA by 0.62%, 0.59%, and 0.75%, respectively. Compared to other attention mechanisms, our MSA achieves the best results and improves performance well.

4.5. Complexity analysis

We compare the number of parameters (params) and floating point operations (FLOPs) of our method with other methods, as shown in Table 9. We can see that the params and FLOPs of our method are only 23.42

Table 9. The number of parameters, FLOPs and accuracy on RAF-DB dataset

Method	Params	FLOPs	Accuracy (%)
MA-Net ^[52]	50.54 M	3.65 G	88.40
AMP-Net ^[15]	105.67 M	4.73 G	89.25
our MSAFNet	23.42 M	3.60 G	90.06

FLOPs: Floating point operations; MA-Net: multi-scale and local attention network; AMP-Net: adaptive multilayer perceptual attention network; MSAFNet: multi-scale attention and convolution-transformer fusion network.

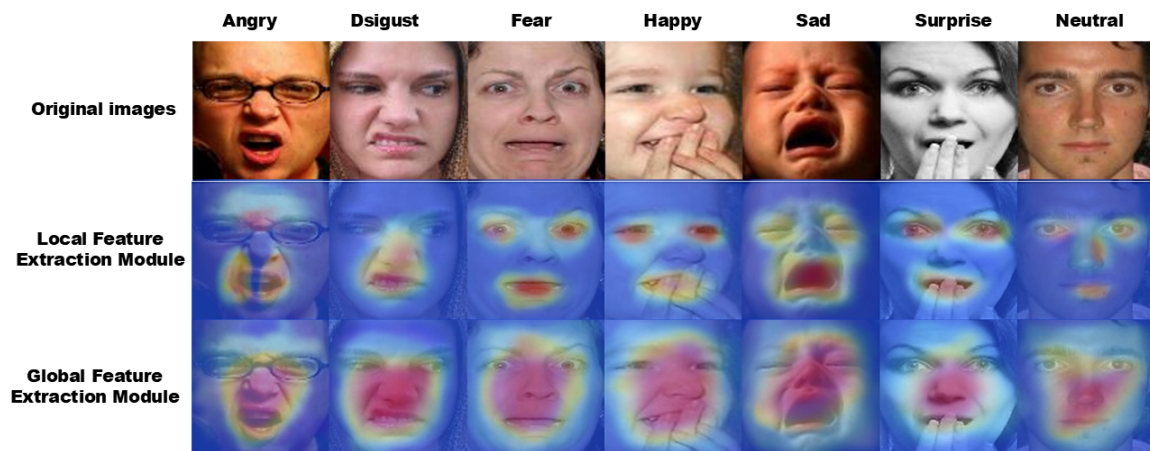


Figure 8. The CAM of LFEM and GFEM. The images and labels are from FER2013 and RAF-DB. CAM: Class activation mapping; LFEM: local feature extraction module; GFEM: global feature extraction module.

M and 3.60 G. The parameters and FLOPs of our method are significantly lower than those of MA-Net^[52] and AMP-Net^[15]. These results demonstrate that our MSAFNet has lower complexity and achieves better performance than other methods.

4.6. Visualization

In this section, in order to better validate the performance of MSA, we utilize gradient-weighted class activation mapping (Grad-CAM)^[66] to visualize SE, CBAM, ECA, and our MSA respectively. As shown in Figure 8, LFEM generates highly localized activations focusing on fine-grained facial components, while GFEM produces broader activation patterns capturing holistic facial structure. This contrast validates the complementary roles of LFEM in micro-feature extraction and GFEM in macro-context modeling. As shown in Figure 9, our MSA enables the network to better focus on the key areas, such as the eyes, nose, and mouth. For facial occlusion or variant poses, MSA can still focus on eyes, nose, and mouth regions, and other attention methods only pay attention to eyes, nose, or mouth regions. The results can further illustrate that our MSA can capture the important information of the regions related to FER, verifying the effectiveness of our method.

5. CONCLUSION

In this paper, we propose an end-to-end MSAFNet for FER tasks that can learn local and global features and adaptively model the relationship between them. Our network contains three modules that can obtain different facial information and are robust to real-world facial expression datasets, including the LFEM, the GFEM, and the GLFM. And a MSA block is designed to adaptively capture the importance of relevant regions of FER. The

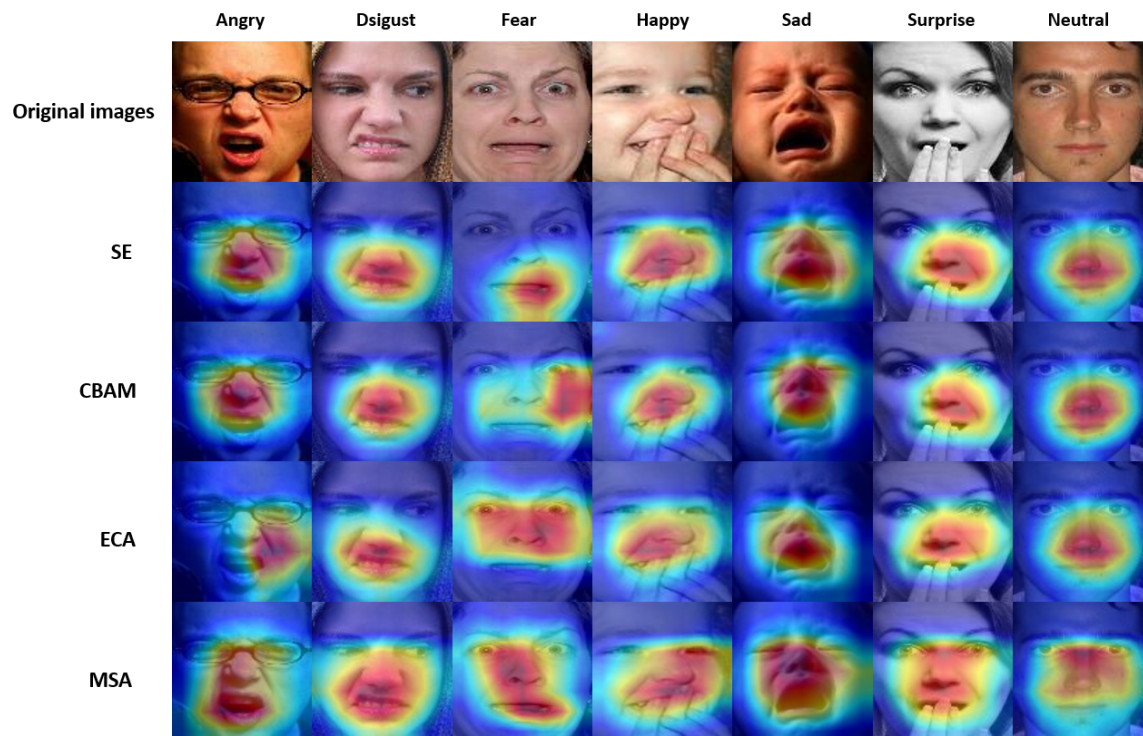


Figure 9. Visualization results. The CAM of MSA is compared with other attention methods. The images and labels are from FER2013 and RAF-DB. CAM: Class activation mapping; MSA: multi-scale attention.

results compared with the existing methods and the ablation experiment show that the proposed method can achieve better performance and have high robustness on real-world facial expression datasets. In future work, we will focus on designing datasets to quantify expressions and establishing evaluation metrics. We will explore how to integrate cognition and deep learning with minimal discrepancies to maximize information extraction. This research will extend to diverse populations, where varied emotional expressions may be present.

DECLARATIONS

Authors' contributions

Made substantial contributions to conception and design of the study and performed data analysis and interpretation: He, H.; Liao, R.; Li, Y.

Performed data acquisition and provided administrative, technical, and material support: He, H.

Availability of data and materials

The datasets used in this study are sourced from publicly available datasets, including RAF-DB, FER2013, and FERPlus. These datasets can be accessed at: RAF-DB: <http://www.whdeng.cn/RAF/model1.html>; FER2013: <https://www.kaggle.com/datasets/msambare/fer2013>; FERPlus: <https://www.kaggle.com/datasets/debanga/facial-expression-recognition-ferplus>. For proprietary or additional datasets used in this study, access requests can be made by contacting the corresponding author at hehuifang@gdep.edu.cn. The code used in this research is available at <https://github.com/SCNU-RISLAB/MSAFNet> or can be obtained upon request.

Financial support and sponsorship

This work was supported in part by the Guangdong Association of Higher Education under the “14th Five-Year” Plan for Higher Education Research Projects, grant number 24GYB148.

Conflicts of interest

All authors declared that there are no conflicts of interest.

Ethical approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Copyright

© The Author(s) 2025.

REFERENCES

1. Bah, I.; Xue, Y. Facial expression recognition using adapted residual based deep neural network. *Intell. Robot.* **2022**, 2, 72–88. [DOI](#)
2. Suguitan, M.; Depalma, N.; Hoffman, G.; Hodgins, J. Face2Gesture: translating facial expressions into robot movements through shared latent space neural networks. *ACM Trans. Hum. Robot Interact* **2024**, 13, 1–18. [DOI](#)
3. Ekman, P.; Friesen, W. V. Constants across cultures in the face and emotion. *J. Pers. Soc. Psychol.* **1971**, 17, 124. [DOI](#)
4. Lucey, P.; Cohn, J. F.; Kanade, T.; Saragih, J.; Ambadar, Z.; Matthews, I. The extended Cohn-Kanade Dataset (CK+): a complete dataset for action unit and emotion-specified expression. In: *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, San Francisco, USA. Jun 13–18, 2010. IEEE, 2010; pp. 94–101. [DOI](#)
5. Lyons, M.; Akamatsu, S.; Kamachi, M.; Gyoba, J. Coding facial expressions with Gabor wavelets. In: *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*, Nara, Japan. Apr 14–16, 1998. IEEE, 1998; pp. 200–5. [DOI](#)
6. Zhao, G.; Huang, X.; Taini, M.; Li, S. Z.; Pietikäinen, M. Facial expression recognition from near-infrared videos. *Image Vis. Comput.* **2011**, 29, 607–19. [DOI](#)
7. Li, S.; Deng, W.; Du, J. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, USA. Jul 21–26, 2017. IEEE, 2017; pp. 2584–93. [DOI](#)
8. Dhall, A.; Goecke, R.; Lucey, S.; Gedeon, T. Collecting large, richly annotated facial-expression databases from movies. *IEEE MultiMedia* **2012**, 19, 34–41. [DOI](#)
9. Barsoum, E.; Zhang, C.; Ferrer, C. C.; Zhang, Z. Training deep networks for facial expression recognition with crowd-sourced label distribution. In: *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. Association for Computing Machinery, 2016; pp. 279–83. [DOI](#)
10. Mollahosseini, A.; Chan, D.; Mahoor, M. H. Going deeper in facial expression recognition using deep neural networks. In: *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Lake Placid, USA. Mar 07–10, 2016. IEEE, 2016; pp. 1–10. [DOI](#)
11. Shao, J.; Qian, Y. Three convolutional neural network models for facial expression recognition in the wild. *Neurocomputing* **2019**, 355, 82–92. [DOI](#)
12. Gursesli, M. C.; Lombardi, S.; Duradoni, M.; Bocchi, L.; Guazzini, A.; Lanata, A. Facial emotion recognition (FER) through custom lightweight CNN model: performance evaluation in public datasets. *IEEE Access* **2024**, 12, 45543–59. [DOI](#)
13. Sun, M.; Cui, W.; Zhang, Y.; Yu, S.; Liao, X.; Hu, B. Attention-rectified and texture-enhanced cross-attention transformer feature fusion network for facial expression recognition. *IEEE Trans. Ind. Informat.* **2023**, 19, 11823–32. [DOI](#)
14. Tao, H.; Duan, Q. Hierarchical attention network with progressive feature fusion for facial expression recognition. *Neural Netw.* **2024**, 170, 337–48. [DOI](#)
15. Liu, H.; Cai, H.; Lin, Q.; Li, X.; Xiao, H. Adaptive multilayer perceptual attention network for facial expression recognition. *IEEE Trans. Circuits Syst. Video Technol* **2022**, 32, 6253–66. [DOI](#)
16. Zhao, R.; Liu, T.; Huang, Z.; Lun, D. P.; Lam, K. M. Spatial-temporal graphs plus transformers for geometry-guided facial expression recognition. *IEEE Trans. Affect. Comput.* **2023**, 14, 2751–67. [DOI](#)
17. Shan, C.; Gong, S.; McOwan, P. W. Facial expression recognition based on local binary patterns: a comprehensive study. *Image Vis. Comput.* **2009**, 27, 803–16. [DOI](#)
18. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, San Diego, USA. Jun 20–25, 2005. IEEE, 2005; pp. 886–93. [DOI](#)
19. Cordea, M. D.; Petriu, E. M.; Petriu, D. C. Three-dimensional head tracking and facial expression recovery using an anthropometric muscle-based active appearance model. *IEEE Trans. Instrum. Meas.* **2008**, 57, 1578–88. [DOI](#)
20. Xu, Z.; Wu, H. R.; Yu, X.; Horadam, K.; Qiu, B. Robust shape-feature-vector-based face recognition system. *IEEE Trans. Instrum. Meas.* **2011**, 60, 3781–91. [DOI](#)
21. Ghimire, D.; Jeong, S.; Lee, J.; Park, S. H. Facial expression recognition based on local region specific features and support vector machines. *Multimed. Tools Appl.* **2017**, 76, 7803–21. [DOI](#)
22. Krizhevsky, A.; Sutskever, I.; Hinton, G. E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **2017**, 60, 84–90. [DOI](#)

23. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556. Available online: <http://arxiv.org/abs/1409.1556>. (accessed on 24 Mar 2025)
24. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, USA. Jun 27–30, 2016. IEEE, 2016; pp. 770–8. DOI
25. Wu, X.; He, J.; Huang, Q.; et al. FER-CHC: facial expression recognition with cross-hierarchy contrast. *Appl. Soft Comput.* **2023**, *145*, 110530. DOI
26. Teng, J.; Zhang, D.; Zou, W.; Li, M.; Lee, D. Typical facial expression network using a facial feature decoupler and spatial-temporal learning. *IEEE Trans. Affect. Comput.* **2023**, *14*, 1125–37. DOI
27. Zhao, R.; Liu, T.; Huang, Z.; Lun, D. P.; Lam, K. M. Geometry-aware facial expression recognition via attentive graph convolutional networks. *IEEE Trans. Affect. Comput.* **2023**, *14*, 1159–74. DOI
28. Cai, J.; Meng, Z.; Khan, A.; Li, Z.; O'Reilly, J.; Tong, Y. Probabilistic attribute tree structured convolutional neural networks for facial expression recognition in the wild. *IEEE Trans. Affect. Comput.* **2023**, *14*, 1927–41. DOI
29. Liu, T.; Li, J.; Wu, J.; Du, B.; Chang, J.; Liu, Y. Facial expression recognition on the high aggregation subgraphs. *IEEE Trans. Image Process.* **2023**, *32*, 3732–45. DOI
30. Zhang, F.; Chen, G.; Wang, H.; Zhang, C. CF-DAN: facial-expression recognition based on cross-fusion dual-attention network. *Comput. Vis. Media* **2024**, *10*, 593–608. DOI
31. Li, Y.; Lu, G.; Li, J.; Zhang, Z.; Zhang, D. Facial expression recognition in the wild using multi-level features and attention mechanisms. *IEEE Trans. Affect. Comput.* **2023**, *14*, 451–62. DOI
32. Zhang, X.; Zhu, J.; Wang, D.; et al. A gradual self distillation network with adaptive channel attention for facial expression recognition. *Appl. Soft Comput.* **2024**, *161*, 111762. DOI
33. Chen, D.; Wen, G.; Li, H.; Chen, R.; Li, C. Multi-relations aware network for in-the-wild facial expression recognition. *IEEE Trans. Circuits Syst. Video Technol.* **2023**, *33*, 3848–59. DOI
34. Vaswani, A.; Shazeer, N.; Parmar, N.; et al. Attention is all you need. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems. NIPS'17*, Red Hook, USA. Curran Associates Inc., 2017; pp. 6000–10. <https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html>. (accessed 2025-03-24)
35. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; et al. An image is worth 16x16 words: transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929. Available online: <https://arxiv.org/abs/2010.11929>. (accessed on 24 Mar 2025)
36. Li, Y.; Miao, N.; Ma, L.; Shuang, F.; Huang, X. Transformer for object detection: review and benchmark. *Eng. Appl. Artif. Intell.* **2023**, *126*, 107021. DOI
37. Chen, X.; Yan, B.; Zhu, J.; Wang, D.; Yang, X.; Lu, H. Transformer tracking. In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, USA. Jun 20–25, 2021. IEEE, 2021; pp. 8122–31. DOI
38. Wang, Y.; Xu, Z.; Wang, X.; et al. End-to-end video instance segmentation with transformers. In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, USA. Jun 20–25, 2021. IEEE, 2021; pp. 8737–46. DOI
39. Ma, F.; Sun, B.; Li, S. Transformer-augmented network with online label correction for facial expression recognition. *IEEE Trans. Affect. Comput.* **2024**, *15*, 593–605. DOI
40. Zhang, X.; Li, M.; Lin, S.; Xu, H.; Xiao, G. Transformer-based multimodal emotional perception for dynamic facial expression recognition in the wild. *IEEE Trans. Circuits Syst. Video Technol.* **2024**, *34*, 3192–203. DOI
41. Liu, C.; Hirota, K.; Dai, Y. Patch attention convolutional vision transformer for facial expression recognition with occlusion. *Inf. Sci.* **2023**, *619*, 781–94. DOI
42. Gao, S. H.; Cheng, M. M.; Zhao, K.; Zhang, X. Y.; Yang, M. H.; Torr, P. Res2Net: a new multi-scale backbone architecture. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 652–62. DOI
43. Chen, Q.; Wu, Q.; Wang, J.; Hu, Q.; Hu, T.; Ding, E. MixFormer: mixing features across windows and dimensions. In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, USA. Jun 18–24, 2022. IEEE, 2022; pp. 5249–59. DOI
44. Goodfellow, I. J.; Erhan, D.; Carrier, P. L.; et al. Challenges in representation learning: a report on three machine learning contests. In: *Neural Information Processing. ICONIP 2013*, Berlin, Heidelberg. Springer Berlin Heidelberg. 2013; pp. 117–24. DOI
45. Wang, K.; Peng, X.; Yang, J.; Meng, D.; Qiao, Y. Region attention networks for pose and occlusion robust facial expression recognition. *IEEE Trans. Image Process.* **2020**, *29*, 4057–69. DOI
46. Zhang, K.; Zhang, Z.; Li, Z.; Qiao, Y. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Process. Lett.* **2016**, *23*, 1499–503. DOI
47. Guo, Y.; Zhang, L.; Hu, Y.; He, X.; Gao, J. MS-Celeb-1M: a dataset and benchmark for large-scale face recognition. In: Leibe B, Matas J, Sebe N, Welling M, editors. *Computer Vision ECCV*. Cham: Springer International Publishing, 2016; pp. 87–102. DOI
48. Wang, K.; Peng, X.; Yang, J.; Lu, S.; Qiao, Y. Suppressing uncertainties for large-scale facial expression recognition. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, USA. Jun 13–19, 2020. IEEE, 2020; pp. 6897–906. DOI
49. Li, Y.; Zeng, J.; Shan, S.; Chen, X. Occlusion aware facial expression recognition using CNN with attention mechanism. *IEEE Trans. Image Process.* **2019**, *28*, 2439–50. DOI
50. Ding, H.; Zhou, P.; Chellappa, R. Occlusion-adaptive deep network for robust facial expression recognition. In: *2020 IEEE International Joint Conference on Biometrics (IJCB)*, Houston, USA. Sep 28 - Oct 01, 2020. IEEE, 2020; p. 1–9. DOI
51. Farzaneh, A. H.; Qi, X. Facial expression recognition in the wild via deep attentive center loss. In: *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Waikoloa, USA. Jan 03–08, 2021. IEEE, 2021; pp. 2401–10. DOI
52. Zhao, Z.; Liu, Q.; Wang, S. Learning deep global multi-scale and local attention features for facial expression recognition in the wild.

- IEEE Trans. Image Process.* **2021**, *30*, 6544–56. DOI
53. Ruan, D.; Yan, Y.; Lai, S.; Chai, Z.; Shen, C.; Wang, H. Feature decomposition and reconstruction learning for effective facial expression recognition. *arXiv* **2021**, arXiv:2104.05160. Available online: <https://doi.org/10.48550/arXiv.2104.05160>. (accessed on 24 Mar 2025)
 54. Ma, F.; Sun, B.; Li, S. Facial expression recognition with visual transformers and attentional selective fusion. *IEEE Trans. Affect. Comput.* **2023**, *14*, 1236–48. DOI
 55. Ruan, D.; Mo, R.; Yan, Y.; Chen, S.; Xue, J. H.; Wang, H. Adaptive deep disturbance-disentangled learning for facial expression recognition. *Int. J. Comput. Vis.* **2022**, *130*, 455–77. DOI
 56. Lv, Z. Facial expression recognition method based on dual-branch fusion network with noisy labels. In: *2024 IEEE 7th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, Chongqing, China. Mar 15–17, 2024. IEEE, 2024; pp. 1608–12. DOI
 57. Miao, S.; Xu, H.; Han, Z.; Zhu, Y. Recognizing facial expressions using a shallow convolutional neural network. *IEEE Access* **2019**, *7*, 78000–11. DOI
 58. Xie, W.; Shen, L.; Duan, J. Adaptive weighting of handcrafted feature losses for facial expression recognition. *IEEE Trans. Cybern.* **2021**, *51*, 2787–800. DOI
 59. Tang, Y.; Zhang, X.; Hu, X.; Wang, S.; Wang, H. Facial expression recognition using frequency neural network. *IEEE Trans. Image Process.* **2021**, *30*, 444–57. DOI
 60. Li, H.; Wang, N.; Yu, Y.; Yang, X.; Gao, X. LBAN-IL: a novel method of high discriminative representation for facial expression recognition. *Neurocomputing* **2021**, *432*, 159–69. DOI
 61. Huang, C. Combining convolutional neural networks for emotion recognition. In: *2017 IEEE MIT Undergraduate Research Technology Conference (URTC)*, Cambridge, USA. Nov 03–05, 2017. IEEE, 2017; p. 1–4. DOI
 62. Yalçın, N.; Alisawi, M. Introducing a novel dataset for facial emotion recognition and demonstrating significant enhancements in deep learning performance through pre-processing techniques. *Heliyon* **2024**, *10*, e38913. DOI
 63. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, USA. Jun 18–23, 2018. IEEE, 2018; pp. 7132–41. DOI
 64. Woo, S.; Park, J.; Lee, J. Y.; Kweon, I. S. CBAM: convolutional block attention module. In: *Computer Vision - ECCV 2018*. Springer, Cham; pp. 3–19. DOI
 65. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: efficient channel attention for deep convolutional neural networks. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, USA. Jun 13–19, 2020. IEEE, 2020; pp. 11531–9. DOI
 66. Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: visual explanations from deep networks via gradient-based localization. In: *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy. Oct 22–29, 2017. IEEE, 2017; pp. 618–26. DOI