ISSN: 2472-3703 DOI: 10.6911/WSRJ.202505_11(5).0014

MACA-Net: A Multi-head Attention and Clustering-guided Network for Facial Expression Recognition

Dongmei Ma[1, a], Zhitao Zheng[1, b, *]

[1]School of Physics and Electronic Engineering Northwest Normal University, Gansu Lanzhou, China

[a]madongmei@nwnu.edu.cn, [b]1921769825@qq.com

[*] Corresponding author

Abstract

Facial expression recognition (FER) in uncontrolled environments poses significantly greater challenges compared to controlled settings. On the one hand, various real-world interferences—such as illumination variations, pose deviations, and occlusions—severely hinder the extraction of robust and discriminative features. On the other hand, field-collected expression datasets often exhibit high intra-class variability and low inter-class separability, further increasing the difficulty of accurate expression classification and degrading overall recognition performance. To address these challenges, this paper proposes a novel FER framework, MACA-Net, which integrates an enhanced loss function with a multi-head attention-based feature extraction strategy. Specifically, a Multi-head Cross Attention (MCHA) module is introduced to capture diverse and complementary local features, thereby enriching the representational capacity of regional expressions. In addition, an Adaptive Feature Clustering Loss (AFC-Loss) is designed to promote intra-class compactness and inter-class dispersion in the learned feature space, effectively improving the model's discriminative power. Extensive experiments conducted on two challenging FER benchmarks—RAF-DB and FerPlus—demonstrate that MACA-Net achieves recognition accuracies of 89.01% and 89.85%, respectively, outperforming several state-of-the-art methods and validating the effectiveness of the proposed approach.

Keywords

Facial Expression Recognition; Multi-head Cross Attention; Adaptive Feature Clustering Loss.

1. INTRODUCTION

Facial expressions serve as indispensable nonverbal cues in human communication, playing a crucial role in conveying emotional states and social intentions. Compared to other communication modalities such as speech, text, and body language, facial expressions offer a more intuitive and efficient means of expressing emotions. According to Mehrabian's seminal study [1], up to 55% of communicative information is transmitted through facial expressions, while vocal tone and verbal content contribute only 38% and 7%, respectively—underscoring the central role of facial expressions in nonverbal communication.

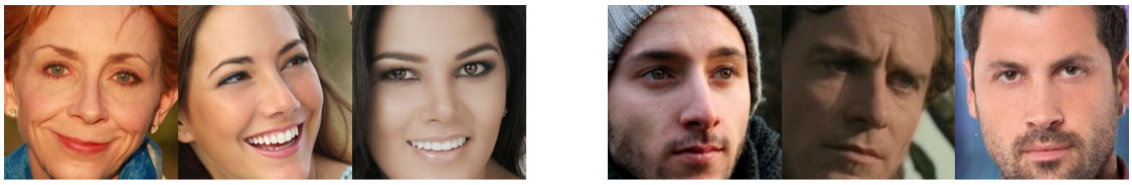
With the rapid development of artificial intelligence and computer vision, facial expression recognition (FER) has emerged as a significant research focus. The primary objective of FER is to automatically identify emotional categories (e.g., happiness, sadness, anger, surprise, fear) by analyzing the dynamic or static variations in key facial regions such as the eyebrows, eyes, and

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202505_11(5).0014

mouth corners. FER holds immense potential across various real-world applications, including human–computer interaction [2], medical diagnosis assistance [3], intelligent driving systems [4], and public safety surveillance [5].

Despite the remarkable progress brought by deep learning in FER, models still face considerable challenges in real-world, unconstrained environments. In-the-wild conditions introduce unpredictable variations in image acquisition, which severely impact model performance. Unlike laboratory-controlled datasets such as CK+ [6], MMI [7], and JAFFE [8], real-world datasets like RAF-DB [9], FerPlus [10] and AffectNet [11] exhibit diverse expression intensities, significant pose variations, illumination inconsistencies, and occlusions. These factors greatly compromise the robustness and generalization capability of FER models.

Furthermore, FER is intrinsically challenged by semantic ambiguity. On one hand, a single emotional category may correspond to multiple distinct facial expressions, leading to large intra-class variability. For example, a "happy" expression may manifest as a subtle smile or an exuberant grin. On the other hand, different emotional categories may share similar facial features—resulting in high inter-class similarity. For instance, “happy” and “surprised,” or “sad” and “angry,” can appear visually alike in localized facial regions. These challenges—broad intra-class diversity and ambiguous inter-class boundaries—significantly increase the difficulty of accurate classification. Figure 1 illustrates typical examples of such intra-class variations and inter-class similarities in expression images.



(a) Intra-Class Variability (b) Inter-Class Similarity

Figure 1. Examples of Intra-class Variation and Inter-class Similarity in Facial Expressions

To address the aforementioned challenges in facial expression recognition (FER), researchers have developed various strategies to optimize the distribution of learned features. Among them, Center Loss and its variants are widely employed to simultaneously minimize intra-class variance and maximize inter-class separability, thereby enhancing the clustering compactness and discriminative power of feature representations.

Building upon this foundation, we propose an improved loss function termed Adaptive Feature Clustering Loss (AFC-Loss). As an extension of Center Loss, AFC-Loss dynamically adjusts the compactness strength across different classes according to their distributional characteristics. This design enhances the expressiveness of learned features and facilitates more structured feature space organization.

In parallel, effectively balancing global structural understanding and local detail modeling remains a core issue in FER. Recent advances have demonstrated that attention mechanisms are well-suited for dynamically emphasizing informative facial regions. To this end, we design a novel Multi-Head Cross-Attention (MHCA) module that enables the model to adaptively capture discriminative local features. Each attention head in MHCA is capable of focusing on different facial regions based on content-aware context, thereby improving local modeling flexibility and robustness.

To further enhance the complementarity among attention heads, we introduce an additional Mutual Information-based Loss (MI-Loss). This constraint encourages different attention

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202505_11(5).0014

channels to attend to semantically diverse and complementary regions of the face, promoting greater feature diversity and reducing redundancy.

In summary, we propose a novel FER framework named MACA-Net (Multi-head Attention and Clustering-Aware Network), built upon a ResNet-34 backbone. The main contributions of this work are as follows:

1.AFC-Loss: An adaptive feature clustering loss that improves feature discriminability by dynamically compressing intra-class distances and enhancing inter-class separability.

2.MHCA Module: A parallel multi-head cross-attention mechanism that adaptively attends to key facial regions, thereby strengthening the model's sensitivity to subtle expression variations.

3.MI-Loss: A mutual information-based loss function that enforces diversity among attention heads, guiding the network to learn complementary and semantically distinct facial features.

2. PROPERTIES

2.1. Center-Loss

To effectively distinguish between different classes in facial expression recognition, a widely adopted strategy is to increase inter-class separability and reduce intra-class variability through feature clustering loss functions. Among them, the Center Loss is one of the most commonly used approaches.

The primary objective of Center Loss is to enhance intra-class compactness by minimizing the distance between sample features and their corresponding class centers in the embedding space. This is achieved by computing a representative center for each class and encouraging features of the same class to be as close as possible to this center, thereby reducing intra-class variance and promoting more discriminative feature representations.

The formulation of the Center Loss is defined as follows:

$$L_c = \frac{1}{2}\sum_{i=1}^{m} \| x_i - c_{y_i} \|^2 \tag{1}$$

Here, x_i denotes the feature representation of a sample, and c_{yi} represents the center of the class y_i to which the sample belongs. During the network training process, the feature representations are updated through backpropagation, driving the sample features to gradually converge towards their respective class centers. This process effectively minimizes intra-class variance and contributes to a more compact and discriminative feature space.

2.2. Mutual Information

Mutual information (MI) is a fundamental concept in information theory that quantifies the amount of information shared between two random variables. Specifically, it measures how much knowing one variable reduces the uncertainty about the other. Unlike traditional linear correlation measures such as covariance or Pearson's correlation coefficient, mutual information is capable of capturing both linear and non-linear dependencies, making it a powerful and versatile tool. As a result, it has been widely applied in various domains, including feature selection, representation learning, and deep neural networks.

Given two discrete random variables X and Y, with joint distribution PXY(x, y) and marginal distributions PX(x) and PY(y), the mutual information between X and Y is defined as:

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202505_11(5).0014

$$I(X;Y)=\sum_{x\in X}\sum_{y\in Y}P_{XY}(x,y)\log\left(\frac{P_{XY}(x,y)}{P_X(x)P_Y(y)}\right) \quad (2)$$

This formulation captures how much the joint distribution PXY(x,y) diverges from the product of the marginals, thereby reflecting the dependence between the two variables. When the mutual information is zero, it indicates that the two variables are completely independent and share no information. Conversely, a higher mutual information value signifies a stronger statistical dependence, implying that knowing one variable significantly reduces the uncertainty about the other.

2.3. Self-Attention Mechanism

In contrast to traditional attention mechanisms, self-attention enables adaptive modeling of relationships between arbitrary positions within a given input, demonstrating a robust capability for capturing global dependencies. Initially, this mechanism was predominantly applied in the field of natural language processing [12], and has since been extended to computer vision tasks [13] to enhance the model's capacity to perceive and integrate global contextual information within images.

The core principle of self-attention lies in the global, weighted modeling of features, achieved by constructing three distinct sets of vectors—Query, Key, and Value—and computing the attention weights between arbitrary positions in the input. Specifically, the attention output is formulated as a weighted sum of the value vectors, with the weights being determined by the similarity between the query and the key. This approach effectively strengthens long-range dependencies between spatial locations, thereby facilitating the model's understanding of the overall image structure. The mechanism is mathematically expressed as follows:

$$\text{Attention(Q,K,V)=softmax}(\frac{QK^T}{\sqrt{d_k}})V \quad (3)$$

Where, Q, K, V denote the query, key, and value, respectively, and d_k are the feature dimensions, which are used to scale the control gradient stability.

3. FACIAL EXPRESSION RECOGNITION ALGORITHM WITH MULTI-HEAD CROSS-ATTENTION AND ADAPTIVE FEATURE CLUSTERING LOSS

3.1. Generation Network

To improve the differentiation between categories and effectively extract features from multiple aspects of expression data, thereby enhancing the expression recognition accuracy, this paper proposes MACA-Net—an expression recognition model that integrates adaptive feature loss with a multi-head attention mechanism. The model utilizes ResNet34 as the backbone network. The overall architecture of the proposed model is depicted in Figure 2.

As illustrated in Figure 2, the model's overall process can be summarized as follows: First, the input facial image is passed through the feature extraction network to extract the fundamental facial expression features. In this study, we employ ResNet34 as the feature extraction network. The extracted features are then optimized using the feature clustering function, AFC-Loss, which consolidates the features of the same class and disperses those of different classes. Following this, the Multi-Head Cross Attention (MHCA) mechanism is applied to learn attention maps for multiple local facial expression regions, adaptively assigning attention weights to

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202505_11(5).0014

these regions. Finally, the neural network combines the features extracted by all attention heads to predict the expression category of the input image.

With these enhancements, the proposed model can effectively capture subtle features of facial expressions, strengthen the discriminative power between categories, and significantly improve the overall accuracy of expression recognition.

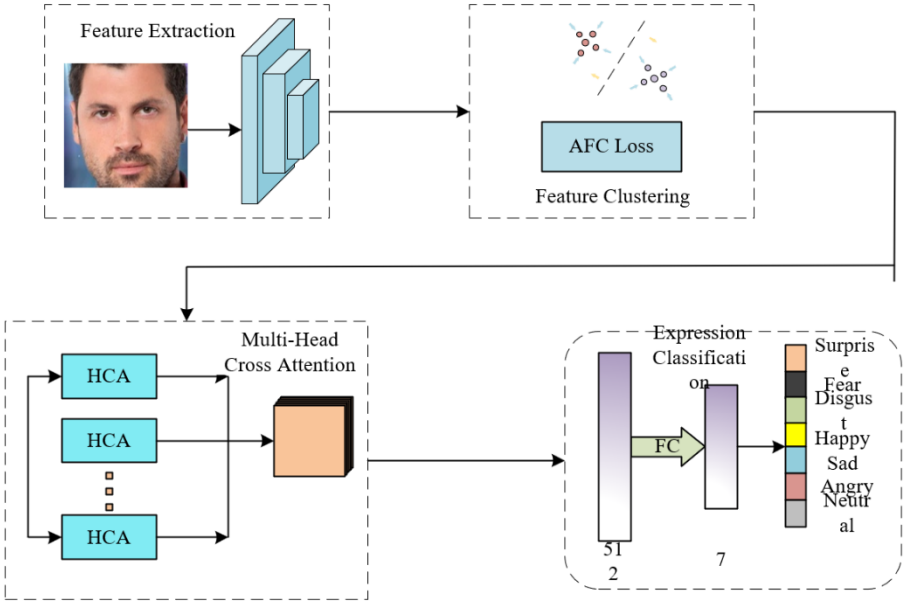


Figure 2. General Block Diagram of the MACA-Net Model

3.2. Adaptive feature clustering loss

To effectively distinguish between different classes, one common approach is to improve the separation between classes by utilizing a feature clustering loss function. A widely used method is the Center Loss, which enhances feature clustering by optimizing intra-class distances. However, this method primarily uses Euclidean distance as the metric, which measures the absolute distance in the feature space. This approach is sensitive to the magnitude of the features and does not account for their directionality. In real-world scenarios, facial expressions are often captured under non-frontal angles, varying lighting conditions, and diverse poses, which makes relying on Euclidean distance problematic.

In contrast, Cosine Similarity is frequently used to measure the relative similarity between high-dimensional features, as it captures the similarity of vector directions rather than magnitudes. This makes cosine similarity invariant to rotation and scale changes, which is particularly beneficial for facial expression recognition under natural conditions. The formula for cosine similarity is as follows:

$$L_{cos} = 1 - \frac{x_i \cdot c_{y_i}}{\| x_i \| \ \| c_{y_i} \|} \tag{4}$$

Here, x_i represents the feature vector of the i-th sample, and c_{yi} is the feature center of the class y_i to which the sample belongs.

Cosine similarity is widely used to measure the relative similarity between high-dimensional features. However, when used alone, it may cause samples from the same class to be distributed

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202505_11(5).0014

in different directions, which negatively impacts clustering performance. On the other hand, Euclidean distance is heavily influenced by feature scales and ignores the directional aspect of features. To address this, this paper proposes a hybrid metric strategy that combines the advantages of both approaches. Therefore, the key challenge for improving expression recognition accuracy lies in effectively combining Euclidean distance and cosine similarity for feature clustering optimization.

To enhance inter-class discriminability and optimize intra-class compactness, this paper introduces a novel Adaptive Feature Clustering Loss (AFC-Loss). The core idea is to adaptively adjust the weights between Euclidean distance and cosine similarity to achieve a more optimal feature measurement. Specifically, a learnable parameter a is introduced to dynamically adjust the contribution of each metric. The formula is as follows:

$$L = \alpha L_{euclidean} + (1-\alpha) L_{\cos ine} \tag{5}$$

The Euclidean distance is used to minimize the L2 distance between the sample features and their corresponding class centers, thereby clustering samples of the same class together. On the other hand, cosine similarity is employed to maximize the cosine similarity between the sample and class center, ensuring that the feature directions remain aligned.

To effectively combine Euclidean distance and cosine similarity, this paper introduces a learnable parameter a, which dynamically adjusts the weight between the two metrics. The adaptive strategy for a is as follows:

$$\alpha = \frac{1}{1+e^{-\theta}} \tag{6}$$

Where a is a learnable parameter, and its range is constrained within [0,1] through a sigmoid function. This allows a to dynamically adjust the weight between Euclidean and cosine losses during the training process, based on the data characteristics.

3.3. Multi-Cross Attention

In uncontrolled facial expression recognition, local details play a crucial role. To capture the expression features from different regions, this paper proposes a multi-head cross-attention mechanism to focus on various facial regions. This attention mechanism consists of several Cross-Attention (HCA) modules. While classical attention mechanisms like CBAM are effective in feature extraction, they still have limitations in expressing spatial relationships and channel dependencies, especially in complex scenarios where they may overlook the interaction between local details and global context. Therefore, this chapter introduces a cross-attention mechanism that integrates spatial attention, self-attention, and channel attention information. By adopting a parallel structure and feature fusion approach, this design achieves richer and more robust feature enhancement.

The cross-attention mechanism employs a parallel structure that merges the outputs of the spatial and channel attention sub-modules, while incorporating a self-attention mechanism to further enhance the model's ability to capture long-range dependencies and complex feature interactions. The design is based on the following three core ideas:

1. Spatial + Channel Parallel Modeling: Spatial attention focuses on "where," and channel attention focuses on "what." Their parallel integration complements each other, enabling full-dimensional feature enhancement.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202505_11(5).0014

2. Self-Attention to Enhance Spatial Expression Power: By incorporating self-attention into spatial attention, the model can better understand long-range dependencies between different regions.

3. Concatenation Fusion for Robustness: Through feature concatenation and 1×1 convolutions, the spatial and channel attention results are fused, improving expression consistency and feature usability.

The structure of the cross-attention mechanism is shown in Figure 3 below:

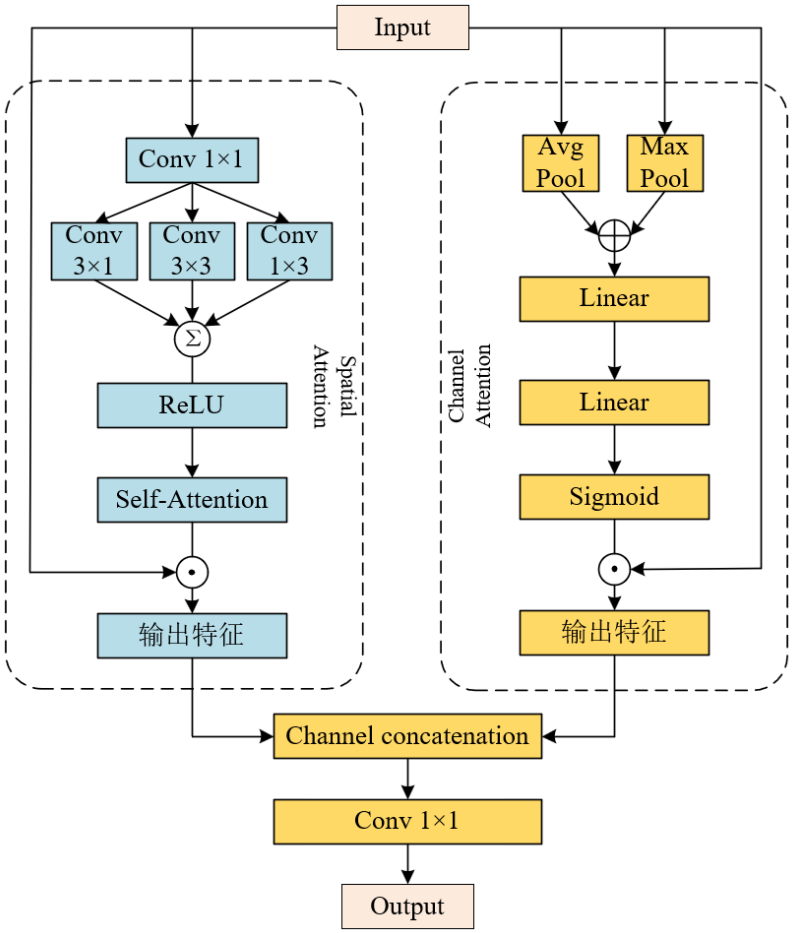


Figure 3. The Structure of Cross-Attention

As illustrated in Figure 3, the cross-attention head is composed of two main branches: spatial attention and channel attention, which are combined side by side. Let the output feature of the feature extraction network be denoted as X. This feature is input into both the spatial attention and channel attention branches separately. In each branch, the features undergo weighted attention, and the results are subsequently fused through concatenation and a 1×1 convolution. The detailed process is as follows:

1. Spatial Attention Branch: This branch is designed to capture long-range dependencies across spatial locations. Initially, the input features undergo dimensionality reduction to expand the local receptive field. Spatial information along both horizontal and vertical axes is extracted using 3×3 convolution, 1×3 convolution, and 3×1 convolution, respectively. This allows for effective modeling of spatial dependencies within the input feature map. The formula is as follows:

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202505_11(5).0014

$$\mathrm{Y} = \mathrm{Re}\,LU(Conv_{3\times3}(Conv_{1\times1}(X)) + Conv_{1\times3}(\bullet) + Conv_{3\times1}(\bullet)) \tag{7}$$

The self-attention mechanism is subsequently constructed and the input features are weighted for output:

$$X_{sa} = X \odot Z_{self}(Y) \tag{8}$$

Where Z_{self} represents the self-attention constructed by 1×1 convolution and $\odot$ represents the element-by-element product.

2. Channel Attention Branch. This branch mainly models the dependencies between different channels. Firstly, the input features are subjected to global maximum pooling and global average pooling, and the two channel statistics are summed up to enhance the importance of different channels:

$$Y = GAP(X) + GMP(X) \tag{9}$$

The channel attention weight y is subsequently obtained through a shared two-layer fully connected network:

$$y = \sigma(W_2 \cdot \mathrm{Re}\,LU(BN(W_1(Y)))) \tag{10}$$

The final channel attention weighted output is as follows:

$$X_{out} = Conv_{1\times1}([\mathrm{X}_{sa}, \mathrm{X}_{ca}]) \tag{11}$$

In order to effectively fuse the features extracted from multiple attention heads the following process is designed. Firstly, the output weighted features X_i of each attention head are downscaled by global average pooling to vector form f_i:

$$\mathrm{f}_i = GAP(\mathrm{X}_i) = \frac{1}{H \times W}\sum_{h=1}^{H}\sum_{w=1}^{W}\mathrm{X}_i(h,w) \tag{12}$$

Subsequently all the output vectors are stacked into a 3D tensor:

$$\mathrm{F} = \mathrm{Stack}(\mathrm{f}_1, \cdots, \mathrm{f}_H) \tag{13}$$

Subsequently, in order to assign dynamic weights to the features learnt by the different attentional heads, the importance of each head was calculated using Log-Softmax normalised on the second dimension (Head dimension):

$$\hat{\mathrm{F}} = \mathrm{LogSoftmax}(\mathrm{F}, \mathrm{dim}=1) \tag{14}$$

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202505_11(5).0014

Finally, all attention head outputs are fused to obtain a unified feature representation using a weighted summation strategy:

$$f_{\text{fused}} = \sum_{i=1}^{H} \hat{F}_i \tag{15}$$

The above method can effectively fuse the features extracted from multiple attention heads for subsequent expression classification.

3.4. Mutual Information Loss Functions

In a multi-head cross-attention mechanism, each attention head is expected to capture distinct local or channel-specific features from the input image, thereby improving the model's capacity to recognize complex expression patterns. However, in practice, the attention maps generated by different heads often exhibit substantial overlap, leading to redundant feature representations and reduced diversity among attention outputs. To address this issue and promote more diverse and complementary feature extraction, we introduce a regularized loss function inspired by mutual information theory. This loss serves to suppress redundancy by penalizing overlapping responses among different attention heads, thus encouraging each head to focus on unique regions or feature dimensions. As a result, the overall discriminative power of the model is significantly enhanced.

To encourage each attention head to focus on distinct feature regions, the proposed method minimizes the mutual information between attention head outputs. A high mutual information value between two attention heads suggests significant overlap in the features they extract, indicating redundancy. In contrast, a low mutual information value implies that the attention heads have learned complementary representations by attending to different spatial regions or semantic attributes.

However, directly computing mutual information is mathematically intractable and computationally intensive in deep learning settings. To overcome this challenge, we adopt a simplified surrogate measure based on the diagonal elements of the covariance matrix, which serves as an efficient approximation of mutual information.

Concretely, the outputs of multiple attention heads for each sample are first flattened and concatenated into a single feature vector. The covariance matrix of these concatenated vectors is then computed, and its diagonal elements are interpreted as indicators of the degree of self-correlation within each head. Lower values along the diagonal suggest reduced redundancy and greater diversity among the attention heads. Based on this insight, the following mutual information loss function is formulated to regularize the training process:

$$L_{MI} = \frac{1}{H} \sum_{i=1}^{H} \log(1 + \frac{H}{\sigma_i^2 + \varepsilon}) \tag{16}$$

Where σ_i^2 denotes the i-th term on the diagonal of the covariance matrix, and ε is a small constant introduced to avoid numerical instability, which is set to $1e-5$ in this paper.

3.5. Overall Loss Function

The model proposed in this chapter is composed of three primary components: a feature extraction network, an adaptive feature clustering loss function, and a multi-head cross-attention mechanism. To enable effective end-to-end training, the overall loss function must incorporate the contributions from all these components—namely, the adaptive feature

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202505_11(5).0014

clustering loss, the mutual information loss associated with the multi-head cross-attention mechanism, and the cross-entropy loss for image classification.

In line with common practices in deep learning, the final objective is formulated by aggregating these individual losses into a unified loss function, which guides the joint optimization of the entire model during training:

$$L = \mathrm{a} \times L_{AFC} + u \times L_{MI} + L_{cls} \tag{17}$$

Where a and u denote the coefficients of the adaptive feature clustering loss function and the mutual information loss function, respectively. After experiments, this paper sets and to 1.0, see chapter 4.4.

4. EXPERIMENTS

4.1. Experimental environment setup

The experimental environment of this paper is as follows: the CPU uses i7 13700F; the GPU is NVIDIA RTX3090 with 24GB of video memory and 32GB of RAM. All the experiments are conducted in Ubuntu22.04 environment, and the deep learning frameworks are PyTorch2.0 and CUDA12.0.

4.2. Datasets

We conducted experiments on two widely used in-the-wild facial expression datasets: RAF-DB and FerPlus.

RAF-DB: The Real-world Affective Faces Database (RAF-DB) consists of approximately 30,000 facial images collected from the Internet (e.g., Flickr), each annotated with either basic or compound expressions by 40 trained annotators. In our experiments, only images labeled with basic emotions were used. The dataset was divided into 12,271 images for training and 3,068 images for testing.

FerPlus: FerPlus is an enhanced version of the original FER2013 dataset, offering more accurate emotion annotations through a crowdsourcing approach. Each image is labeled by 10 different annotators, which provides a more reliable ground truth compared to the original single-label annotations of FER2013. The dataset covers eight emotion categories: the six basic emotions (happiness, surprise, sadness, anger, disgust, fear), along with *contempt* and *neutral*. It contains 28,709 training images, 3,589 validation images, and 3,589 test images.

In this study, ResNet34—pretrained on the MS-Celeb-1M [24] face recognition dataset—was adopted as the backbone network. All input images were resized to 224 × 224 pixels prior to training. To improve the generalization ability of the model and mitigate overfitting, a series of data augmentation techniques were applied, including horizontal flipping, random rotation, random cropping, and random erasing.

Different training configurations were employed for the two in-the-wild facial expression datasets. Specifically, for the RAF-DB dataset, the model was trained for 60 epochs using the stochastic gradient descent (SGD) optimizer with a momentum of 0.9 and a batch size of 32. The initial learning rate was set to 0.01 and reduced by a factor of 10 every 15 epochs. For the FerPlus dataset, the training also lasted for 60 epochs with the same optimizer settings, but the batch size was increased to 64. The learning rate schedule remained consistent, with decay applied every 15 epochs by a factor of 0.1.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202505_11(5).0014

4.3. Ablation analysis

To verify the feasibility and effectiveness of each proposed module, ablation experiments were conducted on both the RAF-DB and FerPlus datasets. The baseline architecture used throughout all experiments is ResNet-34, and all models were trained from scratch without any pre-trained weights.

The detailed experimental results are presented in Table 1. In this table, MHCA refers to the *Multi-Head Cross Attention* module, AFC-Loss denotes the *Adaptive Feature Clustering Loss*, and MHCA* represents the Multi-Head Cross Attention module *without* the Mutual Information Loss.

The first row corresponds to the baseline ResNet-34 model without any of the proposed modules, while the final row presents the full model proposed in this chapter, integrating all enhancements.

Table 1. Ablation experiments for each component on the RAF-DB dataset as well as the FerPlus dataset

MHCA	AFC-Loss	MHCA*	RAF-DB	FerPlus
			84.17	84.51
✓			86.11	86.53
	✓		85.33	85.47
		✓	85.75	85.96
✓	✓		87.19	87.79

As shown in Table 1, on the RAF-DB dataset, incorporating the *Adaptive Feature Clustering Loss* (AFC-Loss) improves the accuracy of the baseline ResNet-34 model by 1.16%. The addition of the *Multi-Head Cross Attention* (MHCA) mechanism leads to a more significant improvement of 1.94%. When both modules are combined—i.e., in the proposed model MACA-Net—the accuracy increases by a total of 3.02% over the baseline. Furthermore, introducing the *Mutual Information Loss* enhances the performance of the MHCA module, contributing an additional 0.36% improvement in recognition accuracy.

Similarly, on the FerPlus dataset, the addition of AFC-Loss yields a 0.96% accuracy gain, while MHCA improves the baseline by 2.02%. When both modules are used together, the model achieves a 3.28% improvement over the baseline. The incorporation of *Mutual Information Loss* further enhances the effectiveness of the MHCA module, resulting in an additional 0.57% gain in accuracy on the FerPlus dataset.

These ablation results clearly demonstrate the effectiveness of the proposed model architecture and validate the individual contributions of each component.

4.4. Experiments on hyperparameter values

Since the model proposed in this chapter involves two hyperparameters, to ensure that their effects do not interfere with each other, this study first conducts experiments to determine the optimal setting a for the Adaptive Feature Clustering Loss on the RAF-DB dataset. During this stage, the Multi-Head Cross-Attention mechanism is not incorporated into the model. The corresponding experimental results are presented in Figure 4.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202505_11(5).0014

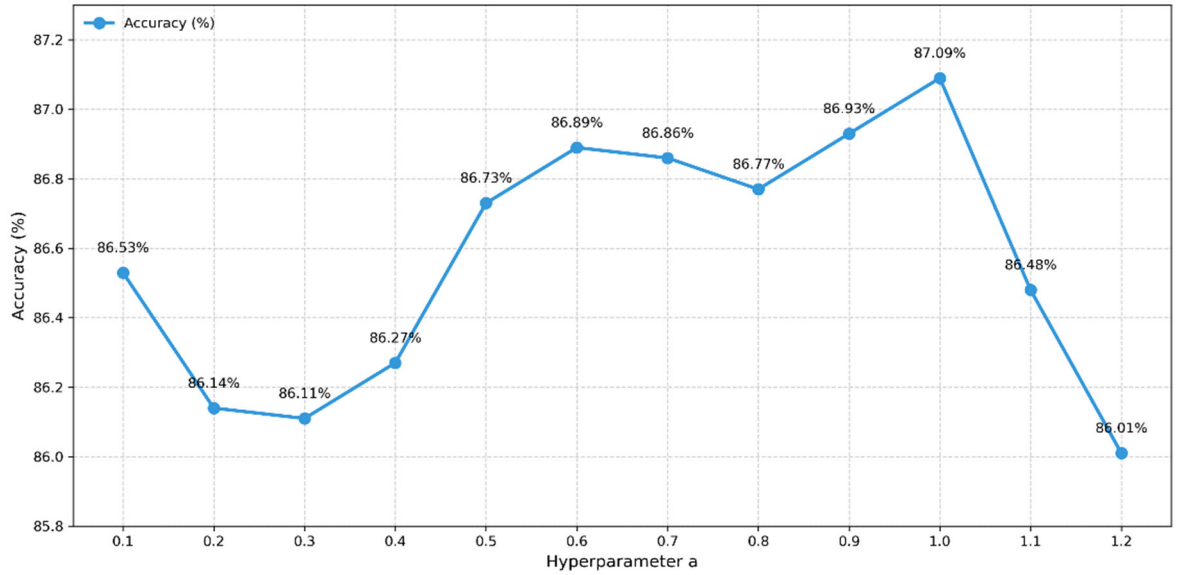


Figure 4. Effect of different values of a on accuracy

As shown in Figure 4, the model achieves the highest accuracy when the parameter a of the Adaptive Feature Clustering Loss is set to 1. Overall, as a increases, the model performance initially fluctuates slightly, reaches a peak near a = 1, and then begins to decline. This indicates that the parameter is sensitive to the regularization effect of the loss function, and proper weight settings can effectively balance intra-class compactness and inter-class separation, thereby improving recognition performance.

Specifically, with a fixed at 1, the Multi-Head Cross-Attention network and the Mutual Information Loss function are introduced, and the coefficient u is varied. The experimental results are shown in Figure 5.

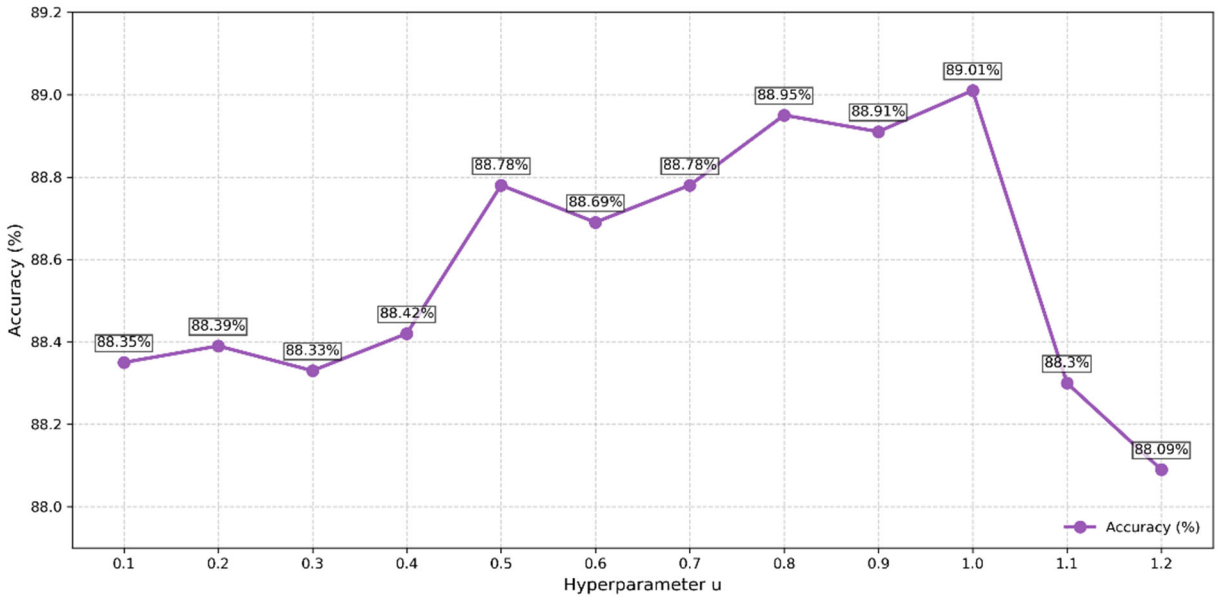


Figure 5. Effect of different values of u on accuracy

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202505_11(5).0014

As shown in Figure 5, the model achieves the highest accuracy when the parameter u of the Mutual Information Loss function is set to 1. Generally, as u increases, the model accuracy shows an initial gradual increase, eventually reaching a point of saturation. The highest accuracy of 89.01% is obtained when u equals 1, after which performance significantly declines. This suggests that excessively large weights for the mutual information loss may diminish the positive influence of other loss terms on model training.

4.5. Experiments on the number of attentions of multiple cross-attention mechanisms

In order to be able to focus on several different regions of the facial expression, a multi-head cross-attention mechanism is designed in this chapter, and in order to determine exactly how many heads of attention work best, experiments on this are done in this chapter, and the results are shown in Figure 6 below.

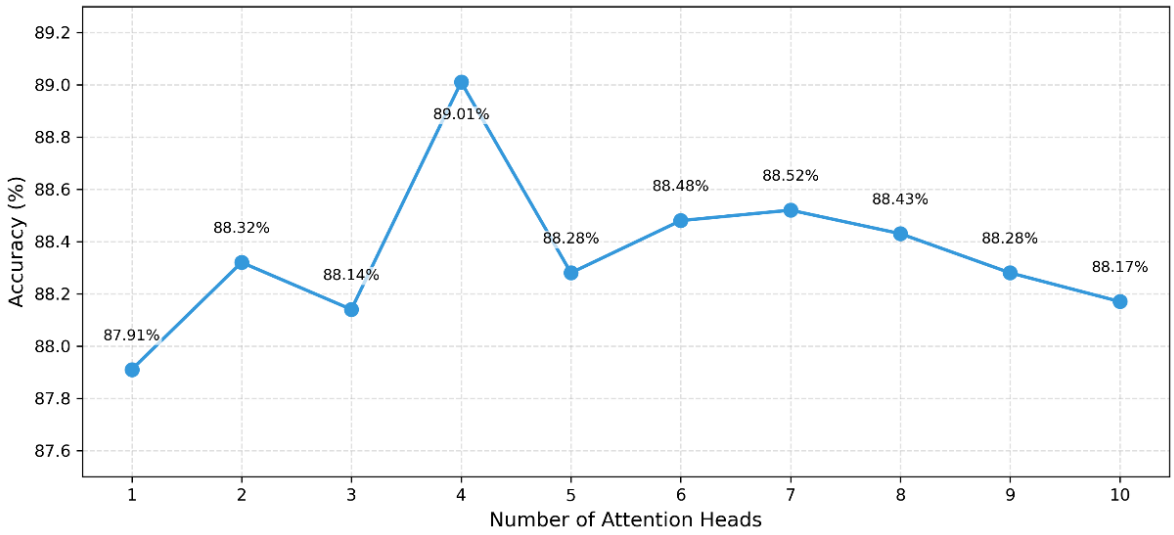


Figure 6. Effect of different number of attention heads on accuracy

As shown in Fig. 6, the multi-head cross-attention structure proposed in this paper is superior to a single attention module, furthermore, the use of four attention heads maximises the performance gain.

4.6. Feature clustering visualization experiments

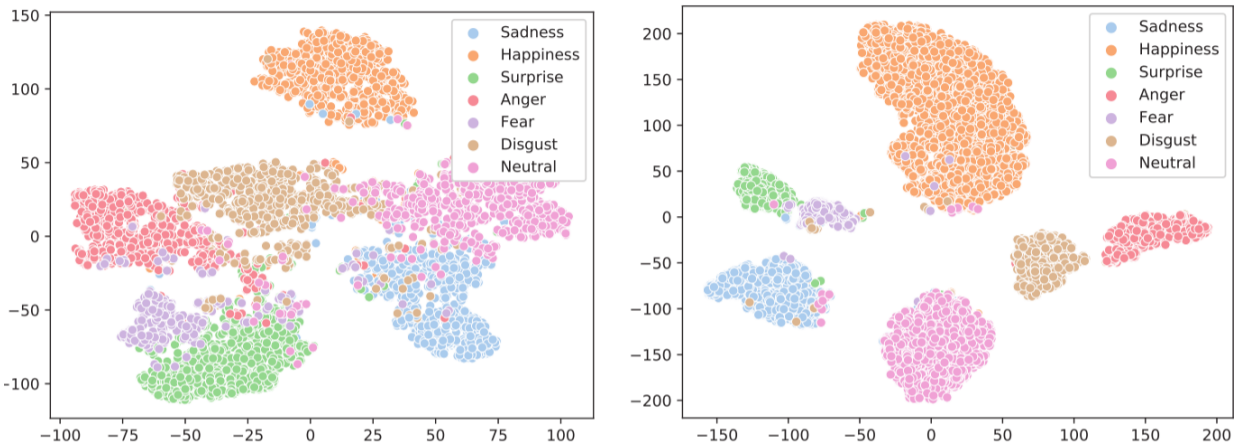
In this paper, an Adaptive Feature Clustering Loss (AFC-Loss) is designed by combining Euclidean distance and cosine similarity to address the challenges of expression images with small inter-class and large intra-class differences. The effectiveness of this loss function is quantitatively evaluated through ablation experiments, as shown in Table 1. To further demonstrate its impact on feature clustering, t-SNE [14] (t-Distributed Stochastic Neighbor Embedding) is employed for feature visualization. t-SNE is a nonlinear dimensionality reduction technique specifically designed for visualizing high-dimensional data. Its core principle is to preserve the similarity (probability distribution) of samples in high-dimensional space when projected to lower dimensions (usually 2D or 3D), ensuring that samples which are similar in the original space remain close to each other in the lower-dimensional projection.

To isolate the influence of the AFC-Loss component, ResNet-18 is used as the base model for the t-SNE visualization, with the comparison model incorporating only AFC-Loss. The RAF-DB dataset is used for this experiment, and the results of the feature visualization are presented in Figure 7.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202505_11(5).0014

As shown in Figure 7, (a) represents the feature visualization of the baseline model on the RAF-DB dataset, while (b) shows the feature visualization after the addition of the AFC-Loss. The categories 0, 1, 2, 3, 4, 5, and 6 in the figure correspond to the seven expression categories: Surprise, Fear, Disgust, Happiness, Sadness, Anger, and Neutral, respectively.

From the figure, it is evident that, compared to the baseline, the feature visualization with the inclusion of AFC-Loss demonstrates an increased distance between different categories, with a clear boundary separating them. Additionally, the distance between features of the same category decreases, resulting in more compact and well-defined clusters. These observations indicate that the adaptive feature clustering function proposed in this chapter effectively enhances inter-class separation and reduces the intra-class variability.



(a) Baseline (b) AFC-Loss

Figure 7. Visualisation of t-SNE features

4.7. Quantitative performance

In order to verify the effectiveness of the proposed method in this paper, quantitative performance comparisons are made with excellent models in recent years on two commonly used field datasets, RAF-DB and FerPlus. The comparison results are shown in Table 2 below.

Table 2. Performance comparison on RAF-DB as well as FerPlus dataset

Method	RAF-DB	FER_Plus
RAN[15]	86.90	88.55
IPA2LT[16]	86.77	/
SCN[17]	87.03	88.01
DMUE[18]	89.42	89.51
PACVT[19]	88.21	88.72
EAC[20]	89.99	89.64
GAAVE[21]	89.29	89.83
Ours	89.01	89.85

As shown in Table 2, the model proposed in this paper achieves the highest classification accuracy on the FerPlus dataset. Although it does not attain the best performance on the RAF-DB dataset, the accuracy gap between the proposed model and the top-performing method is marginal. These results collectively demonstrate the effectiveness and competitiveness of the proposed model.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202505_11(5).0014

5. CONCULUSION

To more effectively capture subtle changes in facial expressions and improve recognition accuracy, this paper proposes a robust and efficient facial expression recognition method based on residual networks. The method introduces innovative improvements in both feature clustering and the attention mechanism, as outlined below:

First, with respect to feature clustering, an *Adaptive Feature Clustering Loss* (AFC-Loss) is proposed, which dynamically adjusts the distribution of features both between and within classes, thereby enhancing the discriminative power of the feature space. Second, to improve the model's ability to capture local facial features, this paper designs a *Multiple Cross-Attention Mechanism* (MCHA). This mechanism enables the model to simultaneously focus on multiple key regions of the face, extracting fine-grained features with complementary information from these regions.

Experiments conducted on two field expression recognition datasets, RAF-DB and FerPlus, demonstrate the effectiveness and robustness of the proposed method.

REFERENCES

[1] Lee J, Kim S, Kim S, et al. Multi-Modal Recurrent Attention Networks for Facial Expression Recognition[J]. IEEE Transactions on Image Processing, 2020, 29: 6977-6991.

[2] Samara A, Galway L, Bond R, et al. Affective state detection via facial expression analysis within a human-computer interaction context[J]. Journal of Ambient Intelligence and Humanized Computing. 2019,10(6):2175-2184.

[3] He L, Guo CG, Tiwari P, Dang W, et al. Intelligent system for depression scale estimation with facial expressions and case study in industrial intelligence[J]. International journal of intelligent systems. 2022, 37(12): 10140-10156.

[4] Jeong M, Ko BC. Driver's Facial Expression Recognition in Real-Time for Safe Driving[J]. Sensors,2018,18(12):4270-4288.

[5] Xie Z H, Cheng S J. Micro-Expression spotting based on a short-duration prior and multi-stage feature extraction[J]. Electronics, 2023,12(2):434-434.

[6] Lucey P, Cohn J F, Kanade T, et al. The extended Cohn-Kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression[C]//Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition, San Francisco, Jun 13-18, 2010. Washington: IEEE Computer Society, 2010: 94-101.

[7] Pantic M, Valstar M F, Rademaker R, et al. Webbased database for facial expression analysis[C]//Proceedings of the 2005 IEEE International Conference on Multimedia and Expo, Amsterdam, Jul 6- 9, 2005. Washington: IEEE Computer Society, 2005: 317-321.

[8] Lyons M J, Akamatsu S, Kamachi M, et al. Coding facial expressions with gabor wavelets[C]//Proceedings of the 3rd International Conference on Face & Gesture Recognition, Apr 14-16, 1998. Washington: IEEE Computer Society, 1998: 200-205.

[9] Li S, Deng W, Du J P. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2852-2861.

[10] Barsoum E, Zhang C, Ferrer C C, et al. Training deep networks for facial expression recognition with crowd-sourced label distribution[C]//Proceedings of the 18th ACM international conference on multimodal interaction. 2016: 279-283.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202505_11(5).0014

[11] Mollahosseini A, Hasani B, Mahoor M H. AffectNet: a database for facial expression, valence, and arousal computing in the wild[J]. IEEE Transactions on Affective Computing, 2019, 10(1): 18-31.

[12] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems. 2017, 30.

[13] Chen M, Radford A, Child R, et al. Generative pretraining from pixels[C]// International conference on machine learning. PMLR, 2020: 1691-1703.

[14] Van der Maaten L, Hinton G. Visualizing data using t-sne.[J]. Journal of Machine Learning Research, 2008, 9(11).

[15] Wang K, Peng X, Yang J, et al. Region attention networks for pose and occlusion robust facial expression recognition[J]. IEEE Transactions on Image Processing, 2020, 29: 4057-4069..

[16] Zeng J, Shan S, Chen X. Facial expression recognition with inconsistently annotated datasets[C]. Proceedings of the European Conference on Computer Vision (ECCV). 2018: 222-237.

[17] Wang K, Peng X, YANG J, et al. Suppressing uncertainties for large-scale facial expression recognition[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 6897-6906.

[18] She J, Hu Y, Shi H, et al. Dive into ambiguity: Latent distribution mining and pairwise uncertainty estimation for facial expression recognition[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 6248-6257.

[19] Liu C, Hirota K, Dai Y. Patch attention convolutional vision transformer for facial expression recognition with occlusion[J]. Information Sciences, 2023, 619: 781-794.

[20] Zhang Y, Wang C, Ling X, et al. Learn from all: Erasing attention consistency for noisy label facial expression recognition[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 418-434.

[21] Zheng J, Li B, Zhang S, et al. Attack can benefit: An adversarial approach to recognizing facial expressions under noisy annotations[C]. Proceedings of the AAAI Conference on Artificial Intelligence: Vol. 37. 2023: 3660-3668.