Arabian Journal for Science and Engineering (2024) 49:12203–12217
https://doi.org/10.1007/s13369-023-08538-6

RESEARCH ARTICLE-COMPUTER ENGINEERING AND COMPUTER SCIENCE

Hybrid Attention-Aware Learning Network for Facial Expression Recognition in the Wild

Weijun Gong[1] · Zhiyao La[2] · Yurong Qian[1,2,3] · Weihang Zhou[2]

Received: 8 August 2023 / Accepted: 19 November 2023 / Published online: 5 January 2024

Abstract
Facial expression recognition (FER) in the wild is one of the most challenging visual tasks owing to various uncontrolled factors such as occlusion, pose, and subtle variation in real scenes. These factors can directly affect the robust performance of current networks, especially as most single-feature learning space methods lack the extraction of potential discriminative features and fail to provide a deeper understanding of expressions. To address the above issues, we propose a novel hybrid attention-aware learning network (HALNet), which comprises a feature compactness network (FCN), a hybrid attention enhancement network (HAEN), and a joint loss optimization strategy. First, FCN performs basic expression feature extraction and optimizes intra- and inter-class distributions simultaneously. Afterward, HAEN constructs a multi-level feature enhancement space by fusing hybrid attention based on CNN and transformer in parallel to effectively improve the profound understanding of expressions. Finally, the expression classification is performed by supervised optimization with joint loss. Extensive experiments are assessed on some of the widest employed wild expression datasets, and results indicate our method is superior to several present state-of-the-art methods, obtaining accuracies of 90.29%, 90.04%, and 61.75% on RAF-DB, FERPlus, and AffectNet, respectively. The cross-dataset and occlusion and pose variation datasets assessment further substantiate our approach's sound generalization and robustness.

Keywords Facial expression recognition · Hybrid attention learning · Transformer · Occlusion and pose variation

1 Introduction

As a very important biological feature of emotional cognition, facial expression is one of the most natural and direct signal transmission modes of human emotional expression and the most effective manner of understanding and communicating human emotional states. The impact of expression-based emotional intelligence in the advancement of artificial intelligence is garnering increasing attention from scholars. In particular, automatic facial expression recognition (FER) includes extensive applications across many areas, including human–computer interaction, psychological assessment, medical monitoring, and public safety [1–4]. Consequently, the study of facial expression recognition has been receiving increasing interest from researchers, and a series of related works have been continuously carried out.

With the advancement of FER, significant recognition performance has been achieved on some small-scale, single-background, non-occlusion, and non-pose variant expression datasets such as CK+ [5], Oulu-CASIA [6], and MMI [7] in controlled laboratory environments. However, as shown in Fig. 1, many occlusions, poses, subtle expression variations, lighting, and image quality present in real scenes make recognition of expressions in such scenarios significantly more challenging. The performance of recognition on large-scale facial expression datasets in the wild, for instance, RAF-DB [8], FERPlus [9], and AffectNet [10], still has much potential for upgrading. Hence, expression recognition in real environments has become one of the focuses of current research.

During the FER research, conventional machine learning methods are mainly used early on to obtain engineering features from small controlled expression datasets [11–14]

✉ Yurong Qian
qyr@xju.edu.cn

1 School of Information Science and Engineering, Xinjiang University, Urumqi 830046, China

2 School of Software, Xinjiang University, Urumqi 830046, China

3 Key Laboratory of Signal Detection and Processing in Xinjiang Uygur Autonomous Region, Urumqi 830046, China

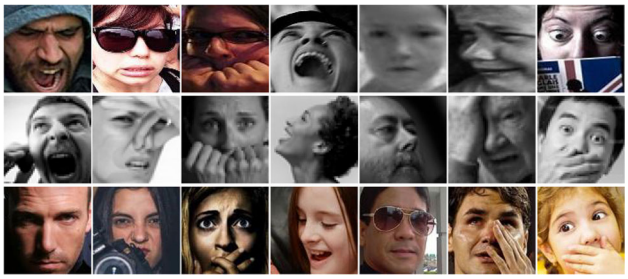


Fig. 1 RAF-DB (top), FERPlus (middle), and AffectNet (bottom) samples of wild expressions. The FER faces significant challenges due to various occlusions, poses, lighting, subtle variations, image resolution, etc.

with limited recognition performance. With the increasing demand for large-scale real expression applications, learning-based feature extraction based on deep learning can more fully capture rich expression information, and end-to-end network models further enhance expression recognition efficiency. A growing number of deep learning models have continued to improve wild FER task performance [15–17]. As attention mechanisms continue to be applied to a variety of computer tasks [18–21], attention focusing on core regions has been further investigated in wild FER [22, 23]. Some scholars have devised center loss [24] and island loss [25] for optimizing the inter- and intra-class distribution of expressions and further guiding the effective learning of the model. However, there are still some issues that need to be further addressed as follows: a large number of occlusions, poses, and small expression spans between classes of real facial expressions present enormous challenges, while most current methods use a single feature extraction and representation space lacking a deep understanding of the overall expression, and insufficient feature extraction capability for regions of interest causing weak robustness problems in recognition; since there are significant inter-class similarity and intra-class variability in wild expression datasets, along with labeling bias arising from subjective and objective reasons, such problems can also lead to degradation of the final recognition performance if they are not optimally guided by model training.

To address the above concerns effectively, we introduce a novel deep learning model for wild FER called the hybrid attention-aware learning network (HALNet). The model mainly involves a feature compactness network (FCN), a hybrid attention enhancement network (HAEN), and a joint optimization strategy component. First, FCN employs the lightweight ResNet-18 as a baseline model for extracting basic expression features along with compactness loss to construct a more sensible intra- and inter-class spatial distribution of features. Afterward, HAEN constructs a multi-level feature enhancement space through parallel hybrid attention fusion to effectively enhance the deep understanding of expressions. Particularly, HAEN consists of a shift enhancement transformer module (SETM), a channel attention enhancement module (CAEM), and a spatial attention enhancement module (SAEM). SETM sequentially feeds the generalized features after channel shifting into a multi-head self-attention learning network and a gated-aware forward network to boost global internal contextual interaction understanding. CAEM captures more meaningful channel information through efficient global channel interaction. SAEM captures the most informative regions using a convolutional codec constructed with a large convolutional kernel. Finally, the label-softened classification loss function is combined with compactness loss to jointly supervise the optimization of the network to further enhance the learning capability of the network. By effectively combining the above network learning modules and under the constant oversight by joint loss, our approach eventually demonstrates significant recognition performance and robustness on multiple wild expression datasets, occlusion and pose variation datasets, and cross-dataset.

In summary, the major contributions of this paper are outlined below:

- We propose a novel HALNet method that can better understand expression variations and capture key discriminative features through a designed hybrid attention model, and combined with optimization in the feature space can effectively address the lack of recognition ability caused by problems such as occlusion and pose, further improving the robust performance of wild FER.
- We design an efficient parallel hybrid attention fusion enhancement network called HAEN, which can better model the global contextual internal information by the self-attention joint unit constructed based on multiple sub-networks designed by transformer, while the channel attention and spatial attention units constructed based on CNNs focus on the most meaningful feature regions and the most representative feature regions. This fusion network can enhance the deep understanding of expression from details to the whole and unfold the final strong discrimination recognition more effectively.
- We assess our HALNet method on three of the most popular wild expression datasets: RAF-DB, FERPlus, and AffectNet. Experimental results show that our approach achieves state-of-the-art performance. Furthermore, the effectiveness of our method is further demonstrated by the excellent performance achieved on the occlusion and pose variation datasets as well as on the cross-dataset.

2 Related Work

In this section, we will describe previous related research work on both FER in the wild and attention-based FER.

2.1 FER in the Wild

The rapid development in emotional intelligence takes the study of facial expression recognition to a new level. While most of the early research focused on laboratory-controlled expression recognition and achieved superior recognition results, facial expression recognition that is more in line with the actual natural environment has received growing focus from researchers as various application requirements are proposed. Moreover, constrained by the various challenging issues mentioned in the previous section, more researchers have shifted from expression analysis based on traditional manual features to further recognition and optimization research using deep learning techniques.

To address the uncertainty of expressions, Wang et al. [26] introduced a self-cure network (SCN) to reduce the degree of overfitting of the network by regularly arranging the learned adaptive weights and re-labeling the samples discriminated as uncertain. Zhang et al. [27] proposed a relative uncertainty learning (RUL) model that uses uncertainty as weights to blend facial features and devise a cumulative loss for better uncertainty learning. Yan et al. [28] presented an efficient label noise robust network (LRN) to further suppress the heteroskedasticity uncertainties arising from inter-class label noise by exploring the inter-class correlations. Some scholars have improved network recognition performance by designing loss functions to optimize network learning. Li et al. [29] proposed the separate loss to strengthen the discriminability of different classes of expressions by normalizing the cosine similarity to optimize the intra- and inter-class distances. Fan et al. [30] presented the RW loss that further optimizes the feature space and integrates a sample weighting strategy to control uncertainty for improving the recognition performance of the model. Farzaneh et al. [31] proposed a discriminant distribution-agnostic loss to deal with the problem of limited learning ability of the network due to the extreme class imbalance phenomenon. Siqueira et al. [32] constructed the ensembles with shared representations (ESRs) network to enhance expression recognition, further reducing redundant data and improving computational efficiency. Liu et al. [33] designed a point adversarial self-mining (PASM) method to enhance the learning ability of the network by gradually generating learning materials and continuously iterating the teacher network to simulate human learning behavior. Ruan et al. [34] designed a feature decomposition reconstruction learning (FDRL) methodology to obtain discriminative facial features by learning latent intra- and inter-feature relationships through joint feature decomposition and feature reconstruction networks under joint loss optimization. Zhao et al. [35] proposed the EfficientFace model to obtain higher wild expression recognition accuracy by establishing global and local feature extractors and the corresponding training optimization strategies. Jiang et al. [36] further designed the identity and pose disentangled facial expression recognition (IPD-FER) network to separate the expression components from head pose and identity for mining more effective discriminative features.

2.2 Attention-Based FER

The visual attention mechanism utilizes feature weight reconstruction to select more discriminative features to perform visual tasks such as image classification more accurately by focusing more on those core regions of interest. Wang et al. [22] built the regional attention network (RAN) that captures critical regional features for pose and occlusion expressions through self-attention and relational attention to further alleviate the degradation of expression recognition performance induced by pose variation and occlusion. Li et al. [37] proposed a slide-patch and whole-face attention model with SE blocks (SPWFA-SE), incorporating local and global features derived through spatial attention to improve expression recognition accuracy. Xia et al. [38] presented ADC-Net, which leverages channel attention to obtain discriminative features generated by scrambled core local subregions of the expression for final recognition. Zhao et al. [39] presented a global multi-scale and local attention network (MA-Net) for obtaining robust global and local features by constructing global multi-scale and local attention modules to further enhance the recognition of pose and occlusion expressions in real scenes. Guo et al. [40] introduced a multi-region attention transformation framework (MATF) that merges local and global details of faces to achieve multi-region expression correlation by fusing local detail information and coarse global features through an attention transformation network. Liu et al. [41] constructed an adapted multilayer perceptual attention network (AMP-Net) according to facial perception mechanisms and facial attributes to enhance the robustness of recognition by adaptively capturing critical information from local, global, and salient facial regions. Wang et al. [42] built a lightweight attentional embedding network (LAENet) based on CNN-based spatial attention to better focus on emotionally relevant locations in images. Ruan et al. [43] designed an adaptive deep disturbance-disentangled learning (ADDL) model, which can adaptively isolate multiple disturbances from facial expression images. It exploits the advantages of multi-task learning and adversarial transfer learning and achieves a good recognition performance with the assistance of multi-level attention. Zhang et al. [44] presented an enhanced global–local feature learning with

priority (EDGL-FLP) method, which further enhances the discriminative ability of various expressions with the support of feature extraction without auxiliary information and priority-based feature attention fusion. With the introduction of the visual transformer, Ma et al. [45] proposed visual transformers with feature fusion (VTFF) method, which fused the extracted CNN features with manual LBP features and used a self-attention-based transformer to further enhance the understanding of various complex expressions. Liang et al. [46] also introduced a convolutional-transformer dual branch network (CT-DBN) to address occlusion and pose variation by fusing local and global features through a parallel CNN and a self-attention-based transformer. Sun et al. [47] proposed an appearance and geometry transformer (AGT) model to further improve the recognition accuracy of wild FER by using two self-attention-based transformers for simultaneous feature extraction and fusion of heterogeneous data consisting of images and graphics.

Most of the current methods mentioned above improve the discriminative power of facial features mainly by constructing different CNN-based channel or spatial attention mechanisms, and some of the lesser methods use a single self-attention-based transformer to learn expression discriminative features. However, these approaches do not pay particular attention to the issue of missing information in subspaces built in a single-attention model. In contrast, our presented hybrid attention mechanism can focus on both the meaningful (channel) and most informative (spatial) expression areas and further enhance the internal relevance learning of expression features in the multi-head (self-attention) mode. The fused hybrid attention network can capture more substantial discriminative features, significantly improving the recognition ability. This approach enhances the deep understanding of various emotions, facilitates learning nuances among facial expressions, better deals with occlusion and pose problems, and can significantly boost the robustness and accuracy of FER under real-world scenarios.

3 Proposed Method

3.1 Overview

Figure 2 illustrates the overall network structure of HALNet. Given a sample of facial expressions, our FCN initially employs the lightweight backbone network ResNet-18 for extracting basic expression features while exploiting the designed compactness loss to build a more rational intra- and inter-class spatial distribution. Subsequently, the focus is on designing HAEN more from a global perspective to construct a multi-level attention space to enhance the focus on key expression regions. In particular, HAEN consists of the SETM, CAEM, and SAEM. SETM feeds channel-shifted features sequentially into a multi-head self-attention learning network and a gated-aware forward network to enhance global intra-contextual interaction understanding. CAEM builds global channel interactions focusing on more meaningful channel information. SAEM devises an efficient codec structure based on large convolutional kernels for group convolution to obtain the most informative areas. The above transformer architecture-based self-attention with global perception capability and the CNN structure-based channel attention and spatial attention with detailed capture advantages are fully integrated to obtain a hybrid enhanced attention map with complementary advantages. Finally, the training and optimization of the whole network are completed under the supervision of the joint loss optimization strategy to further improve the accuracy and robustness of the method.

3.2 Feature Compactness Network (FCN)

Aiming to construct a relatively efficient and lightweight feature extraction model, we adopt the shallow ResNet-18 for the baseline model. In particular, its residual units can control the network degradation and better tackle the problem of gradient disappearance and explosion.

When the ith expression sample x is input, the network feature output of the last layer is obtained through the backbone network $\mathcal{S}$:

$$x_i{}' = \mathcal{S}(x_i, w) \tag{1}$$

where w denotes the weight parameter of the network.

Compactness loss Given that wild facial expressions have remarkable intra-class variability and inter-class similarity, inspired by improved methods such as center loss [48, 49] in solving such problems, we design this compactness loss function for the intra- and inter-class spatial distance optimization of facial expressions for more accurate recognition on this basis. We hereby construct a multi-level automatic codec to rebuild the adaptive feature space weights. This asymmetric encoder first flattens the output x' of the last layer and transforms it into a low-dimensional space with 128 dimensions, then remaps it again into a 1024-dimensional sub-high-dimensional feature, and eventually maps it again to a 512-dimensional output subspace yielding the final spatial feature weights, thereby not only reducing the redundant information, but also further enhancing the adaptive expression capability of the enhanced features. The whole process is shown below:

$$h_l = \delta(W_l^T h_{l-1} + b_l), \quad l = 1, 2, \cdots n \tag{2}$$

$$\omega_i = \psi(\tau(h_{l=3})) \tag{3}$$

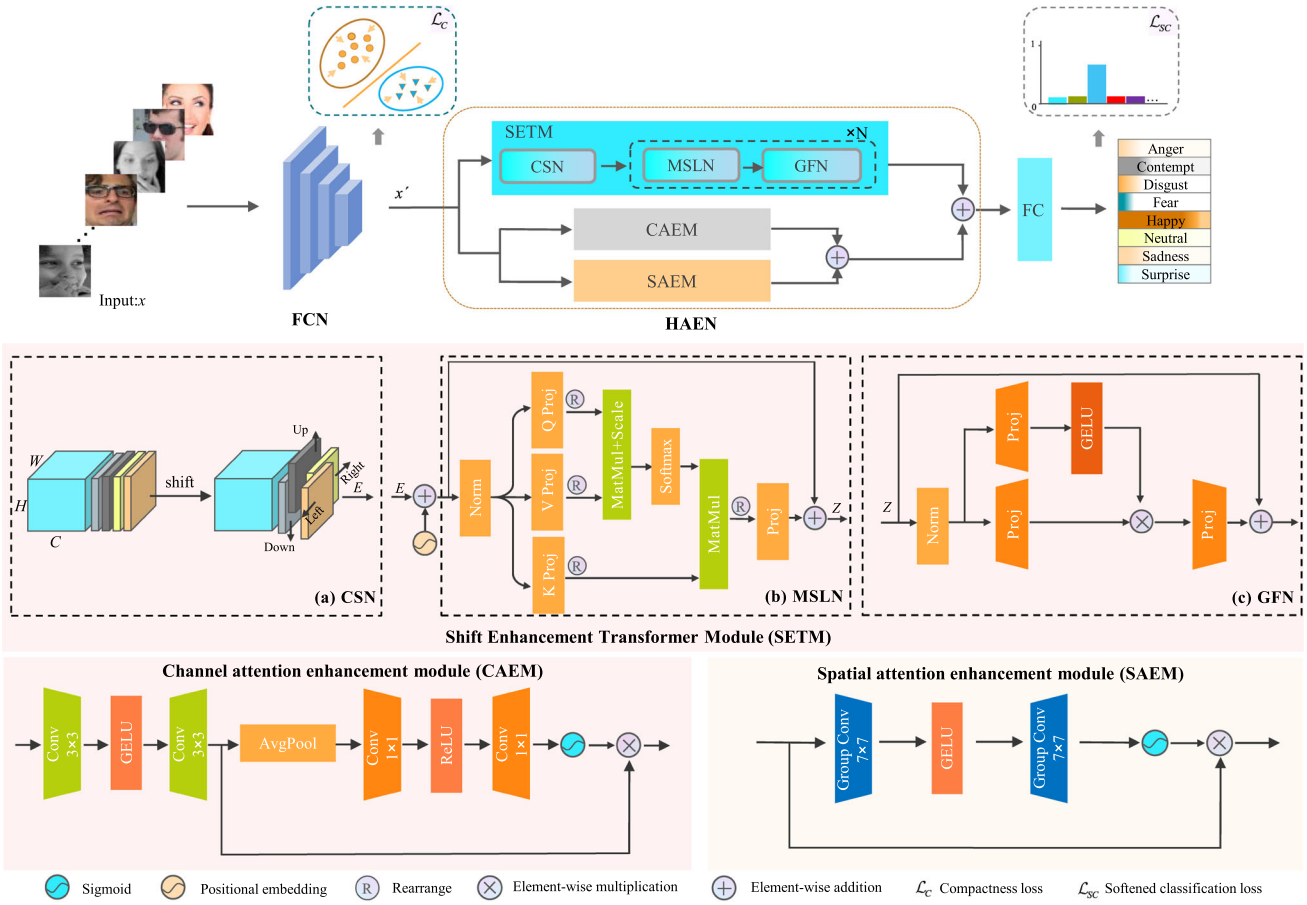


Fig. 2 Overall framework of the proposed HALNet approach. FCN extracts basic facial features using ResNet-18 as a baseline and builds more rational intra- and inter-class distributions by compactness loss. HAEN consists of SETM, CAEM, and SAEM to construct a multilayer attention subspace fusion enhancement module. Final expression classification is eventually performed under the joint supervision of compactness loss and softened classification loss

where h_l denotes the feature output of the lth layer, and when $l = 1$, the network input h_0 is in the form of the representation after x' flattening. b_l is the bias here set to 0, δ indicates the ReLU activation function to enhance the nonlinear capability of the network, and τ refers to the tanh activation function of the last layer. ψ is the softmax to further map the output weight results between 0 and 1.

After obtaining the final adaptive weights ω, the following compactness loss is established:

$$\mathcal{L}_{\mathrm{C}} = \frac{1}{M}\sum_{i=1}^{M} \omega_i \circ \left\| \hat{x}_i - c_{y_i} \right\|_2^2 \tag{4}$$

where $\hat{x}_i$ denotes the feature vector for class y_i after x' averaging pooling, $c_{y_i} \in R^D$ represents the corresponding class center, $y_i \in \{1, 2, \cdots n\}$, and $\| \cdot \|_2$ stands for $L2$ regularization. $\circ$ denotes the dot product by the element, and M denotes the training sample number of the mini-batch. This loss function achieves a better reconstruction of the intra-class feature distribution, making the intra-class features exhibit better compactness and more significant inter-class spacing.

3.3 Hybrid Attention Enhancement Network (HAEN)

This chapter introduces the HAEN in detail, which comprises SETM, CAEM, and SAEM modules, for further obtaining rich and critical expression features of facial regions of interest.

Shift enhancement transformer module (SETM) To address the limitations of local feature extraction in CNN, we propose SETM, which can better model global contextual information and enhance internal correlation understanding. As shown in Fig. 2, SETM acts as a deep improvement module of the ViT [50] encoder, mainly consisting of three parts, namely CSN, MSLN, and GFN.

1) Channel shift network (CSN): To enhance the generalization and robustness of the transformer encoder, inspired by the idea of shift operations [51, 52], as depicted in Fig. 2a,

we introduce a channel shift network before entering the self-attention network. Given an input feature x' of size $C \times H \times W$, some of the channels are selected to be shifted sequentially in pixel units along four spatial directions, such as left, right, up, and down, with the removed pixels no longer used, and the empty pixels are filled with 0. In addition to the above channel operations, the remaining channels remain unchanged. Finally, the feature output E with the same shape after the above transformation is obtained. The specific process is shown as follows:

$$\begin{aligned} e[0:H,\ 1:W]_{0:\alpha C} &\Leftarrow x'[0:H,\ 0:W-1]_{0:\alpha C} \\ e[0:H,\ 0:W-1]_{\alpha C:2\alpha C} &\Leftarrow x'[0:H,\ 1:W]_{\alpha C:2\alpha C} \\ e[0:H-1,\ 0:W]_{2\alpha C:3\alpha C} &\Leftarrow x'[1:H,\ 0:W]_{2\alpha C:3\alpha C} \\ e[1:H,\ 0:W]_{3\alpha C:4\alpha C} &\Leftarrow x'[0:H-1,\ 0:W]_{3\alpha C:4\alpha C} \end{aligned} \tag{5}$$

where α represents the scale factor for the number of selected channels and is set to 1/12. We combine the above-transformed channels with the remaining untransformed channels as the final output E. The implementation mechanism of this module is very simple and efficient and does not contain any parameters.

2) Multi-head self-attention learning network (MSLN): MSLN is used to calculate attention weights to improve the model's representation of inputs by boosting more comprehensive attention through parallel multi-subspace learning, as illustrated in Fig. 2b. Based on the enhanced features E gained from CSN, we first encode the input in terms of learnable positional embeddings and then feed the encoded features into a multi-head self-attention learning unit to compute the query (Q), key (K), and value (V) vectors, individually. Moreover, the following operations are performed to obtain the final attention weights based on obtaining the above vectors and are specifically described as follows:

$$\begin{aligned} Q &= W_d^Q LN(E) \in \mathbb{R}^D \\ K &= W_d^K LN(E) \in \mathbb{R}^D \\ V &= W_d^V LN(E) \in \mathbb{R}^D \end{aligned} \tag{6}$$

$$Att\ (\ Q,\ K,\ V\ ) = Softmax(\frac{QK^T}{\sqrt{D}})V \tag{7}$$

where W represents the corresponding parameter matrix, LN denotes layer normalization, $D = C/H$ means the dimensionality of each attention head, C is the original embedding dimension, and H refers to the number of attention heads. The final attention embedding maps Z obtained by the residual operation is delivered to the subsequent network unit as follows:

$$Z = Att\ (\ Q,\ K,\ V\ ) + E \tag{8}$$

3) Gated-aware feed-forward network (GFN): Conventional feed-forward networks dealing with self-attention mechanisms by channel conversion through only two layers of linear networks may not be sufficient to adapt to complex processes [53]. For better capturing the key expression feature information from the information flow, which enables the effective feature information to enter the subsequent network and improve the features' discriminative power, we design a gated-aware feed-forward network block, as illustrated in Fig. 2c, which can control the flow of network information more precisely and effectively.

For a given input Z, the process of GFN is described as follows:

$$\mathcal{A}_T = \mathcal{F}_2(\delta(\mathcal{F}_1(LN(Z))) \otimes \mathcal{F}_1(LN(Z))) + Z \tag{9}$$

where LN denotes layer normalization, $\mathcal{F}_1$ represents doubling the original number of channels using a linear function, $\mathcal{F}_2$ represents projecting the expanded number of channels to the original ones with a linear function, δ stands for the GELU activation function, and $\otimes$ means element multiplication. $\mathcal{A}_T$ is the attention feature mapping obtained under the whole SETM unit above. GFN guides the flow of information past various layers in the pipeline, ultimately producing more accurate feature information after multilayer stacking.

Channel attention enhancement module (CAEM) To further emphasize which features are the most important for expression recognition, we built a channel attention enhancement module to more effectively mine all channels with high correlation to key features. First, we design two 3×3 convolutions to further enhance the nonlinear feature extraction ability and expand the field of view of the network, based on which global average pooling is performed followed by 1×1 convolution for channel compression, and 1×1-based convolutional channel expansion is performed again under the action of ReLU activation function. Ultimately, the reconstructed channel features are sigmoid-activated and element-wise multiplied with the features before pooling to construct the final attentional feature mapping $\mathcal{A}_C$. The whole process is shown below in detail:

$$\hat{x} = f_{conv}^{3\times3,\,c}(\delta(f_{conv}^{3\times3,\,c/r}(x')) \tag{10}$$

$$\mathcal{A}_C = \sigma(f_{conv}^{1\times1,\,c}(\delta(f_{conv}^{1\times1,\,c/s}(GAP(\hat{x}))))) \otimes \hat{x} \tag{11}$$

where $f_{conv}^{3\times3,\,c/r}$ and $f_{conv}^{1\times1,\,c/s}$ represent the convolution operation with 3×3 and 1×1 convolution kernels, and c is the original number of channels, which is set to 512 here, and r and s are scaling factors with values of 4 and 16, respectively. GAP stands for the global average pooling, while δ and σ are the ReLU and sigmoid activation functions, respectively, and $\otimes$ represents the element-wise multiplication.

Spatial attention enhancement module (SAEM) To better focus on the most meaningful facial expression regions in various faces, we further design a spatial attention enhancement module, which aims to capture important expression region features more precisely by focusing on the critical parts related to the task from the spatial dimension. To enhance the capture of the focused regions under global perception, we first employ a large convolution kernel of 7 × 7 for the output x' of the last layer of the backbone network with a group convolution size of 64 and reduce the channel dimension, followed by a group convolution of 7 × 7 convolution kernels for feature enhancement extraction through channel reconstruction after the nonlinear enhancement of the activation function, and finally multiply with the original x' after sigmoid activation to obtain the final spatial attention feature mapping $\mathcal{A}_S$, which is formulated as follows:

$$\mathcal{A}_S = \sigma(f_{gconv}^{7\times7,c}(\delta(f_{gconv}^{7\times7,c/r}(x')))) \otimes x' \tag{12}$$

where x' represents the input, $f_{gconv}^{7\times7,c/r}$ denotes the group convolution of the 7 × 7 convolution kernel, c refers to the original channel number, and r is the scaling factor, here is set to 4. δ refers to the GELU activation function, σ is the sigmoid function, and $\otimes$ represents the element-wise multiplication.

Ultimately, in order to take full advantage of the above three attentions for more effective expression recognition, we perform a final hybrid attention fusion enhancement based on transformer and CNN, denoted as follows:

$$\mathcal{A} = \mathcal{A}_T + (\mathcal{A}_C + \mathcal{A}_S) \tag{13}$$

3.4 Model Joint Optimization Strategy

For better final expression classification, we adopt a joint network optimization learning strategy in applying hybrid attention features for linear fully connected classification. We first introduce label softening techniques in the commonly used cross-entropy supervised classification loss, using label smoothing to deal with the problem of label bias caused by the presence of a large number of subjective or objective factors in the wild FER dataset and to prevent overfitting caused by mislabeling that leads to paranoid modeling of the network around the wrong answer. Moreover, we combine the compactness loss in the feature compactness network to optimize the intra-class distribution in the feature learning process, further improving the accuracy and robustness of expression recognition. The final joint loss function is shown below:

$$\mathcal{L} = \mathcal{L}_{SC} + \lambda \mathcal{L}_C \tag{14}$$

where $\mathcal{L}_C$ stands for the compactness loss, λ is a control factor to determine the degree of involvement of this loss in the overall loss, and $\mathcal{L}_{SC}$ is the softened classification loss as the basic classification loss is shown in detail below:

$$\mathcal{L}_{SC} = -\sum_{i=1}^{N} \log(p_i)[(1-\delta) * y_i + \delta/N] \tag{15}$$

where p_i represents the predicted probability after softmax, δ denotes the smoothing factor, which is set to 0.1, y_i is 1 when the label is correct and 0 when the label is incorrect, and N refers to the number of categories.

4 Experimental Results

4.1 Datasets

To better reflect the recognition performance of the model, we first perform validation on datasets, including the most commonly used RAF-DB, FERPlus, and AffectNet wild expression datasets, while the occlusion and pose variation datasets are evaluated for targeted robustness. Additionally, we perform a cross-dataset assessment of CK + to assess the model's generalization capability. Details are described below:

RAF-DB [8] consists of a basic or composite dataset including 29,672 real expressions collected using the Internet, which is labeled by 40 skilled professionals. Our experiments are carried out on a basic dataset comprising 12,271 training samples and 3,068 test samples to recognize seven basic expressions.

FERPlus [9] is a real scene expression dataset by relabeling the FER2013 [54] dataset used in the ICML 2013 challenge into ten classes. The dataset contains 28,709, 3,589, and 3,589 training, validation, and test images of size 48 × 48 pixels. Consistent with the preceding research, we perform overall accuracy measures on the eight classes of expressions containing contempt.

AffectNet [10] is the largest dataset of uncontrolled expressions to date, which consists of 450,000 manually annotated images of facial expressions obtained from several Internet search engines. The dataset is extremely challenging and includes a large number of varied ethnicities, poses, occlusions, illuminations, backgrounds, and other complex factors, and the categories are very uneven. Similar to the FERPlus dataset selection, we chose a training set of 287,651 images and a test set of 4,000 images to evaluate the accuracy of the eight basic facial expressions.

CK+ [5] is a laboratory expression dataset extended from Cohn-Kanade (CK) and includes 593 video sequences composed of 123 subjects, where 327 sequences are tagged

as seven basic expressions and contempt. Every sequence contains a range of expressions, starting from a neutral expression in the first frame and progressing toward the peak response of the corresponding expression in the last frame. Consistent with preceding studies, the initial frame of each sequence served as the neutral expression and the final frame as the target expression, leading to 618 images labeled as seven basic emotions and 654 images labeled as eight emotions for cross-dataset validation.

Occlusion and pose variation datasets [22] consist of six dedicated test subsets for occlusion and pose constructed based on the RAF-DB and FERPlus test set and the validation set of AffectNet-8. The occlusion subset consists of occlusion-RAF-DB, occlusion-FERPlus, and occlusion-AffectNet, each containing 735, 605, and 683 samples, while the pose subset consists of samples with angles greater than 30° and angles greater than 45°, where pose-RAF-DB includes 1,247 and 558 expressions with angles greater than 30 and 45, respectively. Pose-FERPlus consists of 1,170 and 633 corresponding expressions, and pose-AffectNet also comprises 1,949 and 985 samples of different angles, respectively.

4.2 Implementation Details

For all acquired facial expression samples, we unify images to a size of 236 × 236. An on-the-fly enhancement strategy is used to prevent overfitting and improve generalization by performing a training image, including random crop, random horizontal flip, normalization, and random erasing operations to obtain an input image of 224 × 224 pixels in size. The test image is also center-cropped and normalized to obtain a test input image size of 224 × 224. We adopt an end-to-end model training approach, using ResNet-18 as the backbone model, where ResNet-18 is pre-trained on the MS-Celeb-1M [55] face dataset. The model is trained by setting the mini-batch size to 64, and using an SGD optimizer with a momentum of 0.9, and a fixed weight decay of 5e-4, while the initial learning rate is set to 0.04. The depth N of the coding layer in SETM is 3, and the hyperparameter λ is fixed at 0.01. The model is trained on both RAF-DB and FERPlus datasets for 60 epochs decaying by a factor of 10 every 20 epochs. We also train on AffectNet for 30 epochs, decaying by a factor of 5 every 5 epochs, with a dynamically balanced sampling tactic to automatically rebalance the classes to address extreme imbalances.

The HALNet model has a parametric count of 30.04M and GFLOPs of 2.56. The training durations for RAF-DB, FERPlus, and AffectNet are 39.94, 67.08, and 305.68 min, respectively. The whole experiment is achieved using the PyTorch platform on the NVIDIA RTX 2080Ti GPU hardware base.

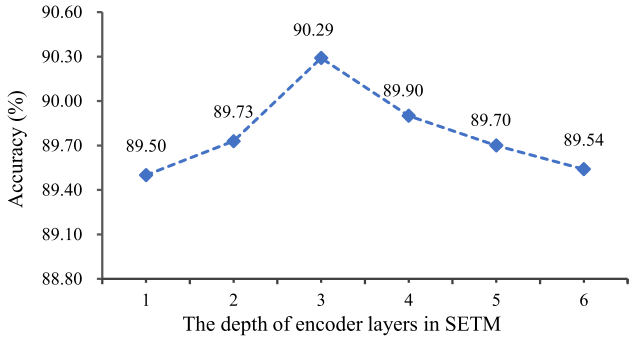


Fig. 3 Ablation study of encoding layer depth in SETM on RAF-DB dataset

4.3 Ablation Studies

We proceed to conduct ablation studies to further assess the impact of model components and some key parameters on model performance.

Evaluation of components for HALNet To analyze the contribution of each component of HALNet to the learning ability of the network, we gradually add shift enhancement transformer module (SETM), channel attention enhancement module (CAEM), and spatial attention enhancement module (SAEM) to the baseline model (ResNet-18) to investigate their effects on the model performance. Table 1 lists the results of our analysis on the wild expression datasets RAF-DB, FERPlus, and AffectNet. When SETM, CAEM, and SAEM are joined to the baseline network alone, it increases by 0.94%, 0.84%, and 0.91% on RAF-DB, 1.05%, 1.15%, and 0.99% on FERPlus in turn, as well as 1.10%, 0.95%, and 1.22% on AffectNet, respectively, which indicates that the above different attention learning modes can contribute to better mining of expression key features from different perspectives. When the above modules are combined pairwise, there is a considerable improvement in the overall performance of the model. The fusion of SETM with other modules is more obvious, which indicates that SETM can better learn potential contextual correlations after model fusion. Moreover, when the three enhancement modules are mix-fused, the model performance shows a significant increase, which is 2.15%, 2.16%, and 2.32% higher than the baseline on RAF-DB, FERPlus, and AffectNet, respectively. The above results further suggest that our model can combine the advantages of channel and spatial attention and enhance the correlation between the contexts within the features, which enables the model to further enhance the understanding of expressions.

Evaluation of layer depth in SETM We conduct the corresponding ablation experiments to validate the effect of the depth of the encoding layers in the SETM on the model performance. It is observed that the accuracy of the RAF-DB demonstrates variation with an increasing number of layers (0–6), as indicated in Fig. 3. Notably, the model achieves

Table 1 Ablation study for each module of HALNet

SETM	CAEM	SAEM	Accuracy (%)		
			RAF-DB	FERPlus	AffectNet
✗	✗	✗	88.14	87.88	59.43
✓	✗	✗	89.08	88.93	60.53
✗	✓	✗	88.98	89.03	60.38
✗	✗	✓	89.05	88.87	60.65
✓	✓	✗	89.63	89.66	61.05
✓	✗	✓	89.70	89.53	61.28
✗	✓	✓	89.34	89.38	60.98
✓	✓	✓	**90.29**	**90.04**	**61.75**

The best outcomes are highlighted in bold

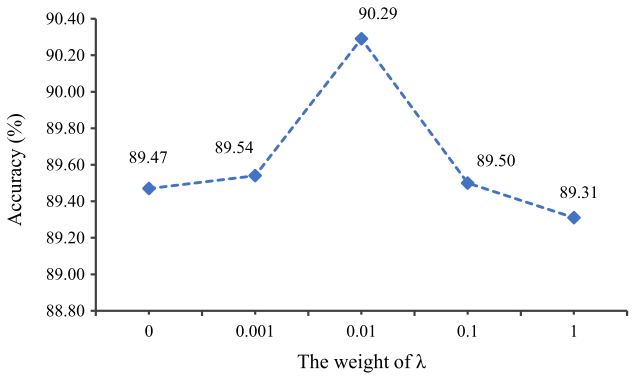


Fig. 4 Ablation study of the parameter λ in the network on the RAF-DB dataset

Table 2 Comparison of ViT [50] and our SETM on model performance

Methods	Accuracy (%)		
	RAF-DB	FERPlus	AffectNet
MSLN + FN (ViT)	89.63	89.53	61.25
CSN + MSLN + FN	90.03	89.82	61.50
MSLN + GFN	89.90	89.79	61.58
CSN + MSLN + GFN (SETM)	**90.29**	**90.04**	**61.75**

The best results are in bold

the highest accuracy (90.29%) when the depth is 3, followed by a decreasing trend of accuracy with further increase in the depth. The ablation results indicate that the SETM with a certain depth of coding layers achieves the best learning capability, while shallower or deeper models have limited learning capability or redundancy making the model performance degraded.

Evaluation of parameter λ in the model We further investigate the effect of the hyperparameter λ in the model loss function on the network performance, as depicted in Fig. 4. We assess the λ values from 0 to 1; in turn, the experimental results show that our approach attains the best performance when $\lambda = 0.01$ and shows a decreasing trend with increasing parameter values. Therefore, the final value of λ is fixed at 0.01.

Performance evaluation between SETM and ViT To better illustrate the effectiveness of our designed modules, we first perform benchmark experiments using the standard multi-head self-attention learning network (MSLN) and feed-forward network (FN) blocks from the ViT [50] encoder and then conduct ablation experiments with our channel shift network (CSN) and gated-aware feed-forward network (GFN), respectively, to verify the effect of different modules on the model performance. It is shown in Table 2 that adding the CSN to the ViT or replacing the FN in the ViT with the GFN block provides further performance improvements over the standard ViT encoder. Moreover, the final SETM constructed by combining all the modules mentioned above leads to an incremental improvement in the overall model performance on all three datasets, further proving the design's effectiveness and good generalization.

4.4 Visualization

We further carry out some data visualizations to illustrate the effectiveness of our network.

Feature attention visualization We adopt Grad-CAM [56] to visualize and compare the baseline model without using attention and the HALNet model with hybrid attention on some wild expression samples, respectively, as illustrated in Fig. 5. The results clearly demonstrate that the feature maps without hybrid attention exhibit a divergent energy distribution, lacking sufficient focus on the core regions of the expressions. Moreover, the feature maps after using our designed HALNet with mixed attention are able to focus more precisely on key regions, especially some occluded and

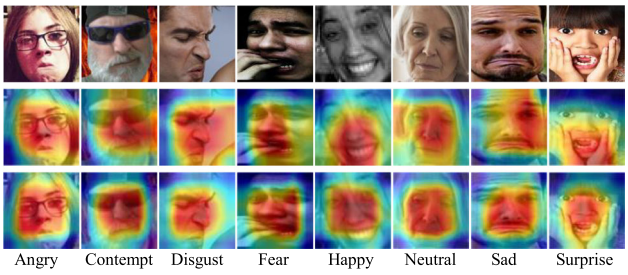


Fig. 5 Feature heatmaps are performed via the Grad-CAM tool, displaying the original samples (top), the feature maps of the baseline model (middle), and the feature maps of HALNet (bottom). Noticeably, our hybrid attention model demonstrates a more precise capture of the expression region of interest

deflected expressions, indicating a significant improvement of our model in fine-grained expression classification.

Feature distribution visualization To further evaluate the effectiveness of our approach, the t-SNE [57] is employed to visualize the distribution of 2D images obtained from the baseline methodology (ResNet-18 with cross-entropy loss) and our HALNet approach. Figure 6 clearly illustrates that the baseline method lacks significant aggregation and differentiation among various categories of expressions. In contrast, our HALNet approach exhibits superior spatial distribution and constraint capabilities. It demonstrates enhanced cohesiveness among intra-class expressions and presents stronger expression discrimination overall.

4.5 Comparison with State-of-the-Art Methods

We first perform a comparison of the proposed HALNet approach with several current state-of-the-art methods on RAF-DB, FERPlus, and AffectNet wild datasets in turn. Moreover, we evaluate the performance of the occlusion and pose variation datasets while conducting a cross-dataset assessment of CK + to verify the generalizability of the methodology. The whole process is evaluated using the overall sample accuracy as the metric.

Performance on RAF-DB Compared with some state-of-the-art approaches from Table 3, our HALNet obtains the highest results, yielding an accuracy rate of 90.29%. When compared with attention-based methods such as RAN [22], AMP-Net [41], ADDL [43], CT-DBN [46], and AGT [47], we outperform the best AGT by 0.77%. For methods SCN [26], RUL [27], and LRN [28], which address label noise and inconsistency, our method surpasses the best RUL method by 1.31%. For the loss optimization-based approaches, separate loss [29], DDA [31], and FDRL [34], our approach exceeds the optimal FDRL by 0.82%. Compared with other network structure models, PASM [33] and EfficientFace [35], our method still surpasses the optimal PASM by 1.61%. Furthermore, as evident from the confusion matrix depicted

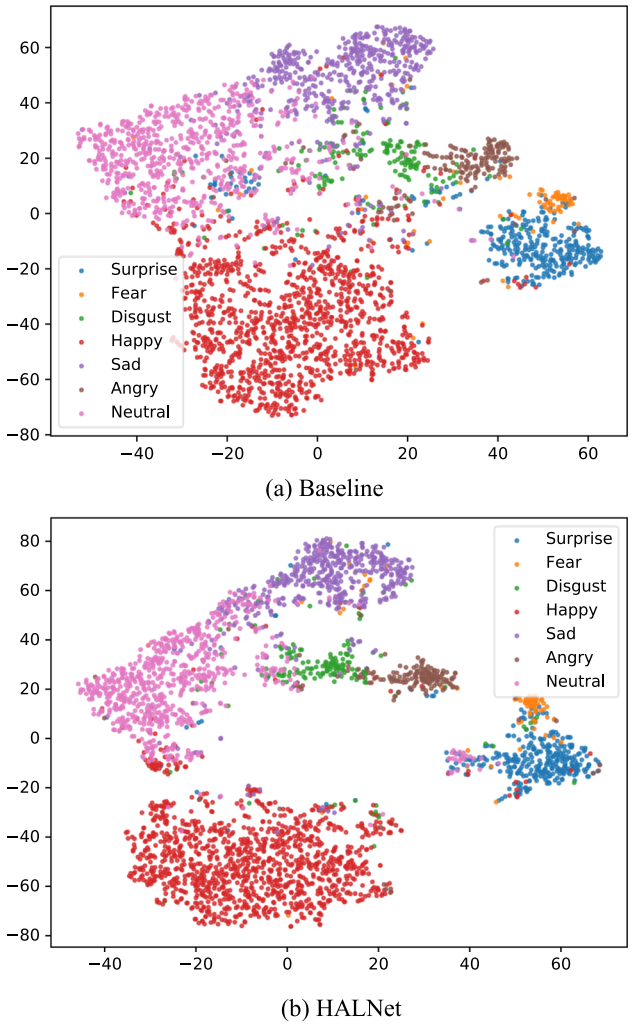
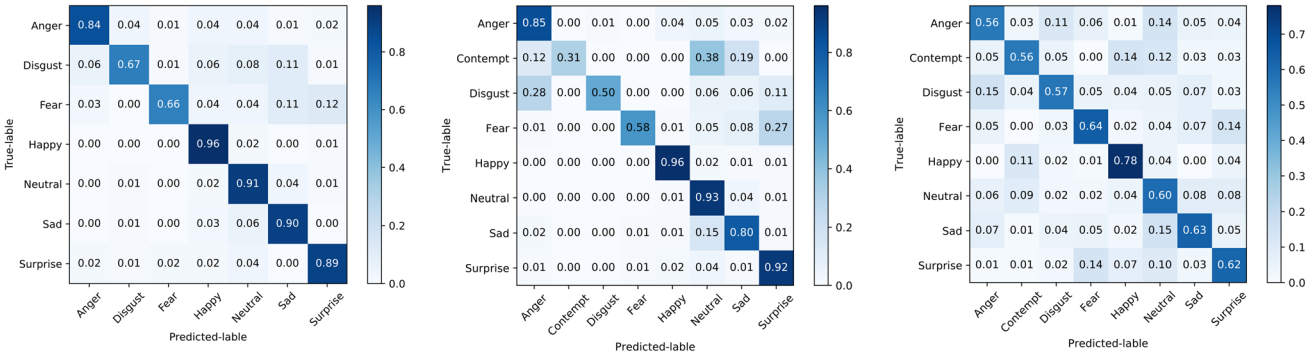


Fig. 6 Feature distribution of the RAF-DB dataset on baseline and HALNet is visualized with *t*-SNE

in Fig. 7a, our approach demonstrates superior recognition performance across most categories except for the disgust and fear categories, which are easily recognized as other expression categories with tiny spans, since very few training samples are available.

Performance on FERPlus The results of comparing various methods on FERPlus are shown in Table 4, where our designed model achieves an optimal accuracy of 90.04%. Compared with the attention-based RAN [22], VTFF [45], ADC-Net [38], MATF [40], PACVT [59], CT-DBN [46], and AGT [47] methods, our method outperforms AGT by 0.64%. When compared with SCN [26] and LRN [28], which are improved for the label problem, our method surpasses the best LRN method by 0.51%, and it also achieves a performance improvement of 1.62% over the optimal model in comparison with the loss-optimized RW loss [30] and IPD-FER [36] methods. Meanwhile, our method improves by 2.28% over the best CNN + BOVW [58] compared to



(a) RAF-DB (b) FERPlus (c) AffectNet

Fig. 7 Confusion matrices with our HALNet on RAF-DB, FERPlus, and AffectNet datasets

Table 3 Comparison to the state-of-the-art methods on RAF-DB

Method	Year	Acc. (%)
Separate loss [29]	2019	86.38
RAN [22]	2020	86.90
DDA [31]	2020	86.90
SCN [26]	2020	87.03
PASM [33]	2021	88.68
EfficientFace [35]	2021	88.36
FDRL [34]	2021	89.47
RUL [27]	2021	88.98
LRN [28]	2022	88.91
AMP-Net [41]	2022	89.25
ADDL [43]	2022	89.34
CT-DBN [46]	2023	88.40
AGT [47]	2023	89.52
HALNet (ours)	2023	**90.29**

The best results are in bold

Table 4 Comparison to the state-of-the-art methods on FERPlus

Method	Year	Acc. (%)
CNN + BOVW [58]	2019	87.76
ESRs [32]	2020	87.25
RAN [22]	2020	88.55
SCN [26]	2020	88.01
RW loss [30]	2020	87.60
VTFF [45]	2021	88.81
ADC-Net [38]	2021	88.90
LRN [28]	2022	89.53
IPD-FER [36]	2022	88.42
MATF [40]	2022	89.34
PACVT [59]	2023	88.72
CT-DBN [46]	2023	89.17
AGT [47]	2023	89.40
HALNet (ours)	2023	**90.04**

The best results are in bold

other feature extraction network models. From Fig. 7b, it can be seen that our method is prone to confusion with other expressions with smaller spans on expressions of contempt, fear, and disgust with very few samples, and the recognition rate is still not high. However, the rest of the categories show better recognition performance.

Performance on AffectNet Table 5 gives the comparative results of the different models. Our method achieves an optimal accuracy of 61.75% on AffectNet with eight classes. Compared to the attention-based RAN [22], SPWFA-SE [37], MA-Net [39], LAENet-SA [42], AMP-Net [41], and EDGL-FLP [44], our approach outperforms the best AMP-Net by 0.36%. Compared with LRN [28] based on label optimization, our method exceeds it by 0.92%. Simultaneously, when compared with other feature learning network CNN + BOVW [58], ESRs [32], and EfficientFace [35], we are 1.86% higher than the best EfficientFace. The confusion

Table 5 Comparison to the state-of-the-art methods on AffectNet-8

Method	Year	Acc. (%)
CNN + BOVW [58]	2019	59.58
ESRs [32]	2020	59.30
RAN [22]	2020	59.50
SPWFA-SE [37]	2020	59.23
SCN [26]	2020	60.23
EfficientFace [35]	2021	59.89
MA-Net [39]	2021	60.29
LRN [28]	2022	60.83
LAENet-SA [42]	2022	61.22
AMP-Net [41]	2022	61.39
EDGL-FLP [44]	2023	61.09
HALNet (ours)	2023	**61.75**

The best results are in bold

Table 6 Performance of cross-dataset evaluation on CK+

Method	Train	Test	Acc. (%)
gACNN [60]	RAF-DB	CK +	81.07
SPWFA-SE [37]	RAF-DB	CK +	81.72
VTFF [45]	RAF-DB	CK +	81.88
CT-DBN [46]	RAF-DB	CK +	82.67
PACVT [59]	RAF-DB	CK +	82.10
VTFF [45]	FERPlus	CK +	83.79
CT-DBN [46]	FERPlus	CK +	81.50
PACVT [59]	FERPlus	CK +	83.88
SPWFA-SE [37]	AffectNet	CK +	85.44
VTFF [45]	AffectNet	CK +	86.24
LAENet-SA [42]	AffectNet	CK +	85.10
PACVT [59]	AffectNet	CK +	85.86
HALNet (ours)	RAF-DB	CK +	**85.92**
HALNet (ours)	FERPlus	CK +	**86.85**
HALNet (ours)	AffectNet	CK +	**91.59**

The best results are in bold

matrix in Fig. 7c further demonstrates the recognition rate of our method in each category. As a whole, the recognition rates of existing methods on AffectNet are generally low, mainly because of the large data size of this dataset, the slight difference between different classes of expressions, the lack of labeling accuracy, etc., causing a more significant challenge of easy misclassification.

Performance on CK+ To further validate the generalization capability of our designed model, a cross-dataset evaluation is performed. We sequentially train our model on RAF-DB, FERPlus, and AffectNet wild datasets and then evaluate them on CK+ . The cross-validation of seven basic emotion categories is performed on RAF-DB, and the other two datasets are cross-validated on the eight basic expressions containing contempt. The comparison results for all methods are presented in Table 6, and our model outperforms the best method by 3.25%, 2.97%, and 5.35% on the three datasets, respectively. These outcomes further illustrate the superior generalization performance of our method, especially on the large-scale AffectNet dataset.

Performance on occlusion and pose variation datasets As illustrated in Table 7, we perform an in-depth assessment of the model on three occlusion and pose variation datasets, respectively. Our method improves over the several methods and indicates some advantages over the present best methodology. First, our model surpasses the optimal AMP-Net [41] by 1.52% on the occlusion datasets occlusion-RAF-DB and outperforms CT-DBN [46] by 0.99% on the occlusion-FERPlus, respectively. In the following comparison on the pose dataset, our model outperforms the current optimal AMP-Net by 0.79% and 0.15% on pose-RAF-DB and pose-AffectNet, respectively, when the pose is larger than 30 degrees. Moreover, our model also reveals an improvement of 1.61%, 0.27%, and 0.16% over the optimal model when compared to the pose-RAF-DB, pose-FERPlus, and pose-AffectNet for poses larger than 45 degrees. Even though very few methods have a slight advantage over ours compared to the best methods, our overall performance still shows an upward trend. Through the above experimental evaluation, our model shows better robustness and recognition performance in coping with the real expressions of pose and occlusion problems.

Table 7 Comparison to state-of-the-art approaches on occlusion and pose variation datasets

(a) Performance of the occlusion-RAF-DB and the pose-RAF-DB datasets

Method	Occlusion	Pose ($\geq 30°$)	Pose ($\geq 45°$)
ResNet-18 [22]	80.19	84.04	83.15
RAN [22]	82.72	86.74	85.20
MA-Net [39]	83.65	87.89	87.99
EfficientFace [35]	83.24	88.13	86.92
VTFF [45]	83.95	87.97	88.35
AMP-Net [41]	85.28	89.75	88.35
CT-DBN [46]	84.90	88.21	86.20
HALNet (ours)	**86.80**	**90.54**	**89.96**

(b) Performance of the occlusion-FERPlus and the pose-FERPlus datasets

Method	Occlusion	Pose ($\geq 30°$)	Pose ($\geq 45°$)
ResNet-18 [22]	73.33	78.11	75.50
RAN [22]	83.63	82.23	80.40
VTFF [45]	84.79	88.29	87.20
AMP-Net [41]	85.44	88.52	87.57
CT-DBN [46]	85.79	**90.60**	87.50
HALNet (ours)	**86.78**	89.32	**87.84**

(c) Performance of the occlusion-AffectNet and the pose-AffectNet datasets

Method	Occlusion	Pose ($\geq 30°$)	Pose ($\geq 45°$)
ResNet-18 [22]	49.48	50.10	48.50
RAN [22]	58.50	53.90	53.19
MA-Net [39]	59.59	57.51	57.78
EfficientFace [35]	59.88	57.36	56.87
VTFF [45]	62.98	60.61	61.00
AMP-Net [41]	**64.27**	61.37	61.16
HALNet (ours)	63.54	**61.52**	**61.32**

The best results are in bold

4.6 Discussion

From the above extensive experiments and results, our approach surpasses some current state-of-the-art methods, which are still mainly from CNN-based channel or spatial attention for local enhancement such as RAN [22], SPWFA-SE [37], AMP-Net [41], or using transformer-based self-attention, like VTFF [45], CT-DBN [46] and AGT [47], to increase the global feature discriminative power. For complex non-rigid structures such as expressions, if the feature information in key regions and the correlation between global feature contexts cannot be captured simultaneously, this will result in limited comprehension, which is not conducive to further recognition improvement. Our hybrid attention is precisely based on the strong capturing ability of CAEM and SAEM attention units constructed by CNN for key features in the focus region and SETM attention units designed based on transformer's self-attention with stronger global context comprehension ability for complementary multi-attention information fusion, which can be seen from the attention structure ablation experiments in Table 1, and the multi-dataset experimental results that include occlusion and pose expressions also attain more competitive results than single attention model. While other methods, such as SCN [26], RUL [27], and LRN [28], only perform loss optimization in terms of label noise or intra-class and inter-class distributions in feature space, our approach further improves the model performance by designing a joint loss optimization strategy for label noise and spatial distribution on top of the mixed attention as mentioned in the above approach as well.

We further provide some samples of correctly classified and misclassified expressions under our model training as shown in Fig. 8 . From the prediction results of the three datasets, it is noted that our model can correctly classify expressions with occlusion, pose, and some expressions with low image quality, but misclassification still exists for some expressions with superimposed poses and occlusions as well as for expressions with tiny expression spans, which are still highly challenging to comprehend when the model relies only on a single image. It can also be observed that some labeling errors caused by subjective factors lead to deviations between the predicted results of the model and the true labels, especially the large-scale dataset AffectNet with low labeling accuracy.

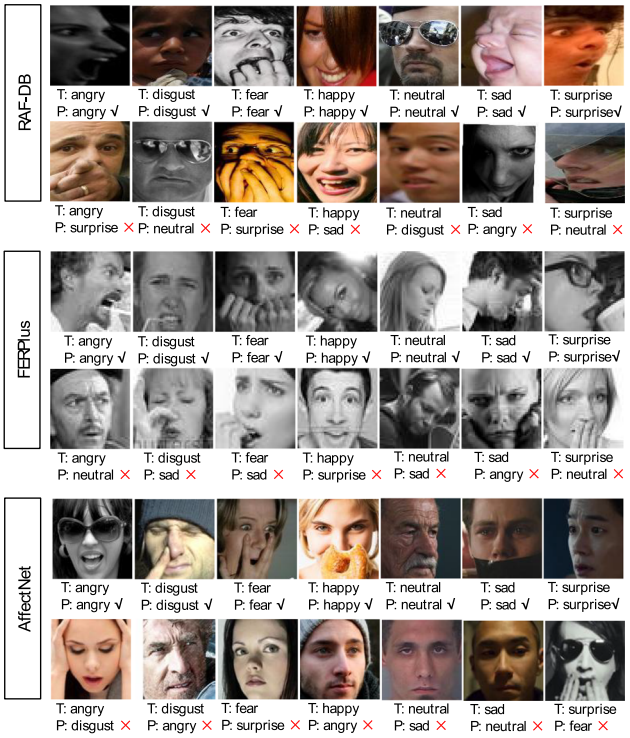


Fig. 8 Some sample examples of correct classification and misclassification on the three datasets. *T* stands for true labels, and *P* stands for predicted labels

5 Conclusion

With the increasing interest in developing innovative applications of biological features in intelligent science based on deep learning methods, it is crucial to improve the recognition rate and robustness of models. This paper proposes a hybrid attention-aware learning network (HALNet) for wild FER, which can more effectively deeply understand and accurately recognize real expressions, including occlusion and pose variations. Initially, a lightweight FCN captures the basic expressive features while simultaneously optimizing the intra-class and inter-class distribution. Then, the hybrid attention enhancement network HAEN is focused on designing a multi-level attention fusion network by SETM, CAEM, and SAEM to more effectively capture discriminative features that facilitate accurate recognition. Finally, the expression classification is performed under joint supervised loss optimization. We perform experiments on three wild expression datasets, demonstrating that our method surpasses some state-of-the-art methods. The estimation of occlusion and pose variation datasets as well as cross-dataset further validates the well-generalization and robustness of our approach.

Since human perception of emotions is a multifactor-triggered process, and our current model only targets single-modal and static expression recognition, the performance will be limited by the singularity of the temporal sequence and the singularity of the modality. Therefore, in the next step, we will further conduct research on dynamic expression and multimodal (such as expression, speech, and body gesture) emotion recognition.

Acknowledgements This work was supported in part by the National Science Foundation of China under Grant 61966035, 62266043 and

U1803261, in part by National Science and Technology Major Project under Grant 95-Y50G37-9001-22/23, and in part by Basic Research Foundation of Universities in the Xinjiang Uygur Autonomous Region of China under Grant 2021D01C083.

Author Contributions All authors were involved in the conceptualization and design of the study. GW and LZ performed material preparation, data collection, and analysis. GW and ZW wrote the first draft of the manuscript, and all authors commented on previous versions of the manuscript. QY read and approved the final manuscript.

Data Availability The datasets used in the paper are all freely available, and the use request is authorized for non-profit purposes.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Liu, Z.; Wu, M.; Cao, W.; Chen, L.; Xu, J.; Zhang, R.; Meng, Z.; Jun, M.: A facial ex pression emotion recognition based human-robot interaction system. IEEE CAA J. Autom. Sin. **4**(4), 668–676 (2017)
2. Corneanu, C.A.; Simón, M.O.; Cohn, J.F.; Guerrero, S.E.: Survey on RGB, 3D, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications. IEEE Trans. Pattern Anal. Mach. Intell. **38**(8), 1458–1568 (2016)
3. Fei, Z.; Erfu, Y.; David, L.; Stephen, B.; Winifred, I.; Xia, L.; Huiyu, Z.: Deep convolution network based emotion analysis towards mental health care. Neurocomputing **388**, 212–227 (2020)
4. Bisogni, C.; Castiglione, A.; Hossain, S.; Narducci, F.; Umer, S.: Impact of deep learning approaches on facial expression recognition in healthcare industries. IEEE Trans. Ind. Inform. **18**(8), 5619–5627 (2022)
5. 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: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 94–101 (2010)
6. Zhao, G.; Huang, X.; Taini, M.; Li, S.Z.; PietikäInen, M.: Facial expression recognition from near-infrared videos. Image Vis. Comput. **29**(9), 607–619 (2011)
7. Pantic, M.; Valstar, M.; Rademaker, R.; Maat, L.: Web-based database for facial expression analysis. In: Proceedings of the IEEE International Conference on Multimedia and Expo (ICME), pp. 5 (2005)
8. Li, S.; Deng, W.; Du, J.P.: Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2852–2861 (2017)
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 ACM International Conference on Multimodal Interaction (ICMI), pp. 279–283 (2016)
10. Mollahosseini, A.; Hasani, B.; Mahoor, M.H.: AffectNet: A database for facial ex pression, valence, and arousal computing in the wild. IEEE Trans. Affect. Comput. **10**(1), 18–31 (2017)
11. Zhao, G.; Pietikainen, M.: Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE Trans. Pattern Anal. Mach. Intell. **29**(6), 915–928 (2007)
12. Aamir, M.; Ali, T.; Shaf, A.; Irfan, M.; Saleem, M.Q.: ML-DCNNet: multi-level deep convolutional neural network for facial expression recognition and intensity estimation. Arab. J. Sci. Eng. **45**(12), 10605–10620 (2020)
13. Happy, S.L.; Routray, A.: Automatic facial expression recognition using features of salient facial patches. IEEE Trans. Affect. Comput. **6**(1), 1–12 (2014)
14. Yan, Y.; Zhang, Z.; Chen, S.; Wang, H.: Low-resolution facial expression recognition: A filter learning perspective. Signal Process. **169**, 107370 (2020)
15. Zhang, Z.; Luo, P.; Loy, C.C.; Tang, X.: From facial expression recognition to inter personal relation prediction. Int. J. Comput. Vis. **126**, 550–569 (2018)
16. Sepas-Moghaddam, A.; Etemad, A.; Pereira, F.; Correia, P.L.: Capsfield: Light field-based face and expression recognition in the wild using capsule routing. IEEE Trans. Image Process. **30**, 2627–2642 (2021)
17. Arnaud, E.; Dapogny, A.; Bailly, K.: Thin: Throwable information networks and application for facial expression recognition in the wild. IEEE Trans. Affect. Comput. (2022)
18. Fan, Q.; Zhuo, W.; Tang, C. K.; Tai, Y. W.: Few-shot object detection with attention-RPN and multi-relation detector. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4013–4022 (2020)
19. Valanarasu, J. M. J.; Oza, P.; Hacihaliloglu, I.; Patel, V. M.: Medical transformer: Gated axial-attention for medical image segmentation. In: Proceedings of the Medical Image Computing and Computer Assisted Intervention (MICCAI), pp. 36–46 (2021)
20. Liu, Z.; Wen, C.; Su, Z.; Liu, S.; Sun, J.; Kong, W.; Yang, Z: Emotion-semantic-aware dual contrastive learning for epistemic emotion identification of learner-generated reviews in MOOCs. IEEE Trans. Neural Netw. Learn. Syst. (2023).
21. Liu, Y.; Li, G.; Lin, L.: Cross-modal causal relational reasoning for event-level visual question answering. IEEE Trans. Pattern Anal. Mach. Intell. **45**(10), 11624–11641 (2023)
22. 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. **29**, 4057–4069 (2020)
23. Cai, J.; Meng, Z.; Khan, A.S.; 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. (2022)
24. Wen, Y.; Zhang, K.; Li, Z.; Qiao, Y.: A discriminative feature learning approach for deep face recognition. In: Proceedings of the European Conference on Computer Vision (ECCV), vol. 14, pp. 499–515 (2016)
25. Cai, J.; Meng, Z.; Khan, A.S.; Li, Z.; O'Reilly, J.; Tong, Y.: Island loss for learning discriminative features in facial expression recognition. In: Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition (FG), pp. 302–309 (2018)
26. Wang, K.; Peng, X.; Yang, J.; Lu, S.; Qiao, Y.: Suppressing uncertainties for large-scale facial expression recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6897–6906 (2020)
27. Zhang, Y.; Wang, C.; Deng, W.: Relative uncertainty learning for facial expression recognition. In: Proceedings of Advanced Neural Information Processing Systems, vol. 34, pp. 17616–17627 (2021)
28. Yan, H.; Gu, Y.; Zhang, X.; Wang, Y.; Ji, Y.; Ren, F.: Mitigating label-noise for facial expression recognition in the wild. In: 2022 IEEE International Conference on Multimedia and Expo (ICME), pp. 1–6 (2022)
29. Li, Y.; Lu, Y.; Li, J.; Lu, G.: Separate loss for basic and compound facial expression recognition in the wild. In: Proceedings of the

Asian Conference on Machine Learning (ACML), pp. 897–911 (2019)
30. Fan, X.; Deng, Z.; Wang, K.; Peng, X.; Qiao, Y.: Learning discriminative representation for facial expression recognition from uncertainties. In: Proceedings of the IEEE International Conference on Image Processing (ICIP), pp. 903–907 (2020)
31. Farzaneh, A.H.; Qi, X.: Discriminant distribution-agnostic loss for facial expression recognition in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 406–407 (2020)
32. Siqueira, H.; Magg, S.; Wermter, S.: Efficient facial feature learning with wide ensemble-based convolutional neural networks. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 5800–5809 (2020)
33. Liu, P.; Lin, Y.; Meng, Z.; Lu, L.; Deng, W.; Zhou, J.T.; Yang, Y.: Point adversarial self-mining: a simple method for facial expression recognition. IEEE T. Cybern. 1–12 (2021)
34. Ruan, D.; Yan, Y.; Lai, S.; Chai, Z.; Shen, C.; Wang, H.: Feature decomposition and reconstruction learning for effective facial expression recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7660–7669 (2021)
35. Zhao, Z.; Liu, Q.; Zhou, F.: Robust lightweight facial expression recognition network with label distribution training. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, no. 4, pp. 3510–3519 (2021)
36. Jiang, J.; Deng, W.: Disentangling identity and pose for facial expression recognition. IEEE Trans. Affect. Comput. **13**(4), 1868–1878 (2022)
37. 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. (2020)
38. Xia, H.Y.; Li, C.; Tan, Y.; Li, L.; Song, S.: Destruction and reconstruction learning for facial expression recognition. IEEE Multimedia **28**(2), 20–28 (2021)
39. 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. **30**, 6544–6556 (2021)
40. Guo, Y.; Huang, J.; Xiong, M.; Wang, Z.; Hu, X.; Wang, J.; Hijji, M.: Facial expressions recognition with multi-region divided attention networks for smart education cloud applications. Neurocomputing **493**, 119–128 (2022)
41. 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. **32**(9), 6253–6266 (2022)
42. Wang, C.; Xue, J.; Lu, K.; Yan, Y.: Light attention embedding for facial expression recognition. IEEE Trans. Circuits Syst. Video Technol. **32**(4), 1834–1847 (2021)
43. 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. Vision **130**(2), 455–477 (2022)
44. Zhang, Z.; Tian, X.; Zhang, Y.; Guo, K.; Xu, X.: Enhanced discriminative global-local feature learning with priority for facial expression recognition. Inf. Sci. **630**, 370–384 (2023)
45. Ma, F.; Sun, B.; Li, S.: Facial expression recognition with visual transformers and attentional selective fusion. IEEE Trans. Affect. Comput. (2021)
46. Liang, X.; Xu, L.; Zhang, W.; Zhang, Y.; Liu, J.; Liu, Z.: A convolution-transformer dual branch network for head-pose and occlusion facial expression recognition. Vis. Comput. 1–14 (2022)
47. Sun, N.; Song, Y.; Liu, J.; Chai, L.; Sun, H.: Appearance and geometry transformer for facial expression recognition in the wild. Comput. Electr. Eng. **107**, 108583 (2023)
48. Wen, Y.; Zhang, K.; Li, Z.; Qiao, Y.: A discriminative feature learning approach for deep face recognition. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 499–515 (2016)
49. Farzaneh, A.H.; Qi, X.: Facial expression recognition in the wild via deep attentive center loss. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 2402–2411 (2021)
50. Dosovitskiy, A.; et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: Proceedings of the International Conference on Learning Representations (ICLR), pp. 1–22 (2020)
51. Jeon, Y.; Kim, J.: Constructing fast network through deconstruction of convolution. In: Proceedings of Advanced Neural Information Processing Systems, vol. 31 (2018)
52. Wang, G.; Zhao, Y.; Tang, C.; Luo, C.; Zeng, W.: When shift operation meets vision transformer: An extremely simple alternative to attention mechanism. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, no. 2, pp. 2423–2430 (2022)
53. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, J.; Gomez, A.N.; Kaiser, L.; Polosukhin, I.: Attention is all you need. In: Proceedings of the Advances in Neural Information Processing Systems, vol. 30 (2017)
54. Goodfellow, I.J.; Erhan, D.; Carrier, P.L.; Courville, A.; Mirza, M.; Hamner, B.; Bengio, Y.: Challenges in representation learning: A report on three machine learning contests. In: Proceedings of the International Conference on Neural Information Processing, pp. 117–124 (2013)
55. Guo, Y.; Zhang, L.; Hu, Y.; He, X.; Gao, J.: Ms-celeb-1m: a dataset and benchmark for large-scale face recognition. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 87–102 (2016)
56. 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: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 618–626 (2017)
57. Van der Maaten, L.; Hinton, G.: Visualizing data using t-SNE. J. mach. Learn. Res. **9**(11), 2579–2605 (2008)
58. Georgescu, M.I.; Ionescu, R.T.; Popescu, M.: Local learning with deep and hand-crafted features for facial expression recognition. IEEE Access **7**, 64827–64836 (2019)
59. Liu, C.; Hirota, K.; Dai, Y.: Patch attention convolutional vision transformer for facial expression recognition with occlusion. Inf. Sci. **619**, 781–794 (2023)
60. Li, Y.; Zeng, J.; Shan, S.; Chen, X.: Occlusion aware facial expression recognition using CNN with attention mechanism. IEEE Trans. Image Process. **28**(5), 2439–2450 (2018)