



# Self-supervised facial expression recognition with fine-grained feature selection

Heng-Yu An<sup>1</sup> · Rui-Sheng Jia<sup>1</sup>

Accepted: 18 February 2024

© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2024

## Abstract

Facial expression recognition (FER) holds significant practical implications in real-world scenarios such as human–computer interaction, fatigue driving detection, and learning engagement analysis. Nonetheless, acquiring large-scale and high-quality annotated facial expression datasets is profoundly challenging due to the inherent ambiguity of facial images and concerns over privacy. Consequently, this paper introduces a self-supervised facial expression recognition method based on mask image modeling. This method can learn multi-level facial feature representations without expensive labels and achieves commendable facial expression recognition performance through further fine-grained feature selection. Specifically, we propose the multi-level feature selector (MFS). The MFS comprises two pivotal components: the multi-level feature combiner and the feature selector. During the pre-training stage, the multi-level feature combiner is employed to integrate multi-level features, effectively addressing the vision transformer’s deficiencies in capturing high-frequency facial semantics. Subsequently, in the fine-tuning stage, the feature selector can automatically differentiate highly discriminative regions, extracting fine-grained features. Subsequently, we use graph convolutional networks to further mine the latent connections among fine-grained features, ultimately deriving an integrated feature with enhanced discriminative capabilities. Through such fine-grained facial feature selection, we can mitigate performance degradation induced by inter-class similarities and intra-class variations. Experimental results on the RAF-DB, AffectNet, and FER + datasets demonstrate that our approach significantly outperforms other self-supervised methods in recognition performance and closely approaches the state-of-the-art methods in supervised learning. The code is available at <https://github.com/Greysahy/MFS>.

**Keywords** Facial expression recognition · Self-supervised learning · Multi-level feature combine · Feature selection

## 1 Introduction

Facial expression recognition is a computer vision technique that facilitates emotion recognition in uncontrolled environments based on facial feature analysis. As one of the most potent signals in humans, facial expressions play a pivotal role for computers to decipher human emotional states and behavioral intentions. Hence, achieving accurate facial expression recognition is paramount in developing intelligent systems (such as smart robots and virtual reality) that can perceive and respond to human emotions.

In recent years, researchers have achieved significant advancements in facial expression recognition thanks to the development of deep learning technologies and the availability of large-scale facial expression datasets. Supervised learning methods based on convolutional neural networks (CNNs) and Vision Transformers have been introduced to facial expression recognition tasks, demonstrating outstanding performance. Some research methodologies have ventured to incorporate intricate attention mechanisms [1, 2] or to utilize prior knowledge (such as facial landmarks) to guide the networks [3–5], subsequently attaining even greater accuracy in facial expression recognition.

While these methods have effectively enhanced network performance, they all face a substantial limitation: they primarily focus on supervised learning. The visual complexity of facial expression images, coupled with their marked inter-class similarities and intra-class variations, means that a significant amount of time and specialized expertise is

---

✉ Rui-Sheng Jia  
jrs716@163.com

<sup>1</sup> College of Computer Science and Engineering, Shandong University of Science and Technology, Qingdao 266590, China

required for annotating facial images. Moreover, considering the privacy-sensitive nature of facial expressions and the subjective annotation biases among different graders, the constructed datasets demand rigorous validation processes. This makes the acquisition of large-scale annotated facial expression data exceedingly challenging, suggesting that future approaches should lean toward reduced label dependency, such as semi-supervised [6] or self-supervised methods.

The Mask Auto encoder [7], in its application to visual representation learning, has successfully transferred BERT-style pre-training strategies to the domain of computer vision. This method realized high-quality unsupervised representation learning by establishing an asymmetric encoder–decoder structure based on the vision transformer. Nevertheless, due to the absence of image-specific inductive biases in the vision transformer, its core multi-head attention mechanism tends to focus more on global information, often overlooking low-level, high-frequency details. This characteristic poses challenges for pixel-level facial unit reconstruction tasks, making it difficult to acquire high-quality facial representations during the pre-training phase.

To address the deficiencies in existing facial expression recognition efforts, this paper introduces a novel training strategy for facial expression recognition models: the multi-level feature selector (MFS). This method can learn multi-level facial representations in unlabeled data and carries out unsupervised fine-grained feature selection during the fine-tuning phase, achieving high-precision facial expression recognition. During the pre-training phase, we designed the multi-level feature combiner. It aims to integrate multiple latent features within masked images, compensating for the vision transformer’s shortfall in high-frequency information, thereby aiding the model in acquiring rich facial representations. In the fine-tuning phase, we devised the feature selector. During the learning process, this module can adaptively filter out non-discriminative features based on the discriminative power of the feature units themselves, consequently highlighting highly discriminative regions. Considering that highly discriminative regions are spatially distributed in a discrete manner, we will obtain a set of sparse data following feature filtration. Merely concatenating these features and employing multi-layer perceptron (MLP) for aggregation would lead to a substantial loss of spatial information. When dealing with sparse data, some studies have employed graph structures to aggregate node information, generating enhanced feature representations [8]. By successfully modeling the intricate high-order feature interactions among sparse data, these methodologies have achieved commendable performance. We conceptualize the filtered features as a graph structure. Employing graph convolutional networks, we have achieved efficient graph feature extraction on the discriminative feature map, thereby delving into the latent connections

among discriminative feature units. With the Feature Selector, we can capture granular facial details, thus overcoming the intrinsic inter-class similarities and intra-class variations of facial expressions. The primary contributions of this paper can be summarized as follows:

- (1) We propose a self-supervised facial expression recognition algorithm named MFS. During the pre-training phase, with the assistance of the multi-level feature combiner, the backbone network can learn multi-level facial feature representations without the need for expensive labeling.
- (2) We designed the Feature Selector, which, through meticulously crafted granular feature selection and feature aggregation strategies, assists the network in learning superior decision boundaries, addressing the inherent ambiguities associated with facial expressions.
- (3) We evaluated the proposed MFS across multiple datasets. Experimental results indicate that MFS significantly outperforms other self-supervised methods and closely approaches the results of state-of-the-art supervised techniques.

The structure of this paper is organized as follows: Section 2 provides an overview of the related work on facial expression recognition; Section 3 delves into the specific implementation details of MFS; Section 4 presents the experimental results and ablation studies of MFS on the RAF-DB, AffectNet, and FER + datasets; In Sect. 5, we discuss various attempts and explorations undertaken during the research process. Finally, Sect. 6 summarizes the primary contributions of this study.

## 2 Related work

From the early methods based on handcrafted features [9, 10] to those based on end-to-end learning [11–14], facial expression recognition has always garnered significant attention. Notably, the majority of research aimed at improving facial expression recognition still focuses on extracting distinctive facial expression features using advanced computational models under supervised settings [15–19]. While these studies have achieved commendable accuracy, they heavily rely on labeled training data. Consequently, these methods might suffer substantial performance degradation when faced with low-quality and noisy labels. To address this issue, some researchers have proposed corresponding methodologies to reduce the network’s reliance on fine-grained labels. For instance, Li et al. [20, 21] reclassified the seven basic facial expressions into four coarse-grained classes and employed

coarse labels to assist in fine-grained label supervised learning and contrastive learning, thereby mitigating the performance degradation caused by the similarity of facial expressions. [22] Designed a training paradigm that employed contrastive learning for self-supervised facial expression recognition in multi-view images. Although effective, the scalability of this method is constrained due to its heavy dependence on specific datasets. Subsequently, Shu et al. [23] applied a contrastive self-supervised learning approach to static single-view facial images, effectively enhancing the performance of self-supervised learning in facial expression recognition tasks. Recently, many studies have started to adopt mask image modeling as a self-supervised framework to learn effective facial representations [24–26]. Ma et al. [24] utilized a Mask Auto Encoder pre-trained on large-scale facial images and achieved state-of-the-art performance in facial action unit analysis tasks.

In addition, to address performance degradation stemming from pose variations, facial occlusions, inherent intra-class variability, and inter-class similarity in facial expressions, some studies have suggested employing fine-grained features for facial recognition [27, 28]. These methods can be broadly categorized into those based on facial landmarks [3–5, 29] and those leveraging attention mechanisms [1, 30, 31]. Zheng et al. [3] utilized a pre-trained facial landmarks detector to locate facial landmarks during the data preprocessing stage. They then inputted the salient regions containing these facial landmarks as prior knowledge into the feature extractor, guiding the feature extraction process. Shi et al. [5] introduce a multi-pose block occlusion face recognition method grounded on feature point location. This method segments the face based on facial landmarks and occlusion regions, thereby proficiently mitigating the influences of pose variations and occlusion on face recognition performance. In [30], an end-to-end network architecture for facial expression recognition based on attention mechanisms was proposed. This design emphasized focusing attention on the face while ignoring background noise. [31] Proposed an encoder–decoder attention operation that can focus more on the regions of muscle movements beneath the facial skin, such as the mouth, eyes, and nose, allowing the network to extract deep facial expression features better. Similarly, Wang et al. [1] introduced a region-based attention network architecture that, by capturing local facial features, displayed robustness against facial occlusions and pose variations.

Our MFS is a self-supervised training approach that does not require expensive labels. Compared to the Mask Auto encoder with a vanilla vision transformer as its backbone, this method can integrate multi-level features during pre-training, achieving superior facial representation learning. Additionally, we have designed an unsupervised feature selection strategy that can adaptively choose highly discriminative fine-grained facial features during the fine-tuning process

while simultaneously disregarding non-salient regions. This differs from previous methods based on facial landmarks or those utilizing complex attention mechanisms.

### 3 Methodology

The overall framework of MFS is illustrated in Fig. 1. The entire training process is divided into two stages: The pre-training stage (a) and the Fine-tuning stage (b). The detailed structure of the Multi-level Feature Combiner is depicted on the right side(c).

In the initial phase, we employed a vision transformer backbone augmented with a multi-level feature combiner for self-supervised pre-training. This approach facilitated the network to acquire multi-level facial representations. Subsequently, in the second phase, we inherited the weights of the encoder from the first phase (without freezing) and performed fine-tuning of the entire network's parameters based on fine-grained features extracted by the Feature Selector, resulting in the ultimate model. Detailed training specifics for each phase will be further elucidated in the following section.

#### 3.1 Multi-level facial feature learning

This paper employs an asymmetric encoder–decoder structure of the Mask Auto encoder as the primary framework for self-supervised learning. For a given facial image  $I^{C \times H \times W}$ , we divide it into  $n$  patches. Among these,  $(n - k)$  patches are masked; while, the remaining  $k$  visible patches are fed into the encoder to encode latent features. Subsequently, these latent features are passed into a lightweight decoder to reconstruct the masked pixels.

As the backbone architecture for the encoder, the Vision Transformer excels at modeling global information with its core multi-head self-attention mechanism. However, due to the low-pass filtering nature of multi-head attention [32], Vision Transformers may lack emphasis on high-frequency features. Considering that the masked facial images exhibit noticeable sparsity at the semantic level, high-frequency texture features hold significant value for pixel-level facial reconstruction tasks.

The structure of the multi-level feature combiner is depicted on the right side of Fig. 1. "When the facial image  $I^{C \times H \times W}$  is fed into the vision transformer encoder, we obtain an intermediate feature set  $F = \{f_0, f_1, f_2, \dots, f_i\}$ , where  $f_i$  represents the output feature of the  $i$ -th transformer block. The multi-level feature combiner selects features from different levels within  $F$  and projects them using affine layers to align them in the feature space with

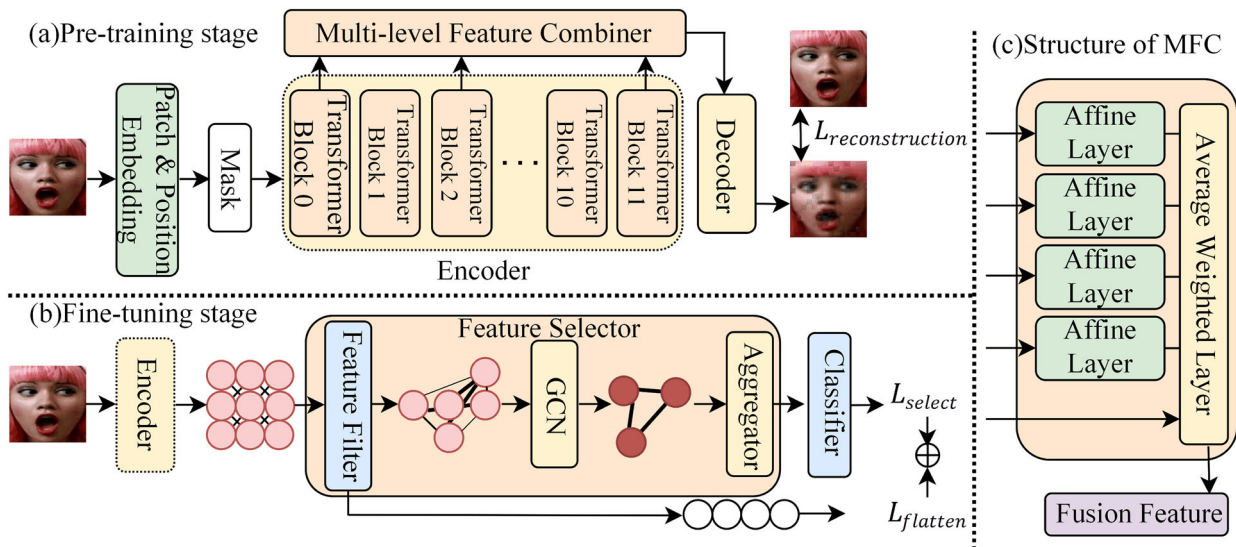


Fig. 1 Overall framework of MFS

the deepest layer’s feature. In the end, we obtain the integrated multi-level concatenated feature  $F_m$ :

$$F_m = \text{stack}(\text{affine}(f_i), f_{n-1}) \tag{1}$$

where  $n$  represents the number of transformer blocks in the encoder. Subsequently, we compute the fused feature  $\hat{F}$  by taking a weighted average of  $F_m$  using an average weighted layer:

$$\hat{F} = \sum_i (F_{m,i} \cdot w_i) \tag{2}$$

where  $w_i$  represents the weight of the  $i$ -th level feature in the fused feature. We initialize all  $w_i$  values to be the same and dynamically optimize them during the subsequent training process.

Finally, the fused feature  $\hat{F}$  is fed into the decoder to reconstruct the masked pixels, resulting in the reconstructed image  $I_r$ . The reconstruction loss  $L_{reconstruction}$  is defined as the pixel-wise mean squared error loss between the reconstructed image and the input image:

$$L_{reconstruction} = \frac{1}{N} \sum_{i=1}^N (\text{pixel}_{r,i} - \text{pixel}_{input,i})^2 \tag{3}$$

where  $N$  represents the total number of pixels in a single image, and  $\text{pixel}_{x,i}$  represents the  $i$ -th pixel value of image  $x$ .

It is important to note that in the multi-level feature combiner, simply fusing features from all levels may lead to information redundancy and introduce noise, resulting in a degradation of the model’s performance. In this paper, for facial expression recognition, we have chosen to select the output features from Block<sub>0</sub>, Block<sub>2</sub>, Block<sub>4</sub>, and Block<sub>6</sub> and

fuse them with the deepest layer features. Detailed experiments regarding this choice will be further discussed in Sect. 5.

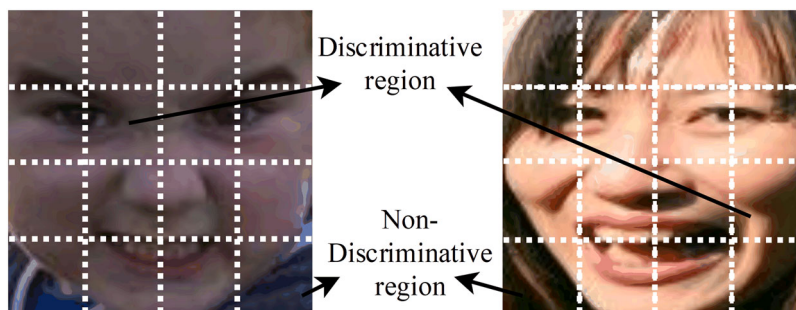
### 3.2 Fine-grained feature fine-tuning

After splitting facial images into  $n$  patches, we can observe that not all regions exhibit significant discriminative information (Fig. 2). Regions that solely contain hair and clothing or are nearly monochromatic are commonly found in facial images across different expression categories. If these areas were used for recognition, their predicted probabilities would likely exhibit a relatively flat distribution. Conversely, selecting regions that contain facial landmarks for recognition yields more discriminative prediction probabilities.

Based on the previous analysis, to make more effective use of the multi-level facial feature representations learned during the pre-training phase, we introduced a feature selector during the fine-tuning stage. The purpose of this design is to adaptively filter out background noise and focus on critical fine-grained facial features, thereby achieving more precise parameter optimization.

The Feature Selector treats each token input to the network as an independent feature unit and performs feature filtering based on the discriminative capacity of each feature unit itself. We use the features extracted by the vision transformer encoder, represented as  $f_i \in R^{L \times D}$ , as the input to the Feature Filter, where  $L$  represents the length of the feature sequence, and  $D$  represents the output dimension of the transformer block. In the feature filter, we first project the input features into a  $C$ -dimensional space (where  $C$  is the total number of predicted categories). Subsequently, we apply the Softmax function to calculate the category prediction scores

**Fig. 2** Discriminative/non-discriminative in facial image



for each feature unit:

$$\mathbf{logits}_i = \text{Projection}(D, C)(f_i), s_i = \text{Softmax}(\mathbf{logits}_i) \tag{4}$$

Among all the feature units, we select the top  $s$  units with the highest confidence as discriminative features; while, the remaining  $L-s$  units are considered non-discriminative features. Since the selected discriminative features exhibit notable local and sparse characteristics, we treat them as a discrete feature map. In the feature fusion stage, to preserve their original spatial scale and spatial structure integrity, we employ a Graph convolutional network (GCN) to process the discriminative feature map, further exploring potential relationships between different features:

$$f'_i = \sigma(A_{\text{adaptive}} W f_i), f_i \in R^{C \times s} \tag{5}$$

Leveraging a graph convolutional network allows us to learn the influence between different feature units and incorporate this influence into the final output features, enabling effective graph feature extraction. Subsequently, by using an aggregator to consolidate the feature map, we ultimately feed the fused features into a classifier to obtain facial expression recognition results. Graph convolutional networks (GCNs) can effectively integrate relationships between multiple facial feature units without disrupting the original feature structure. This enables the model to learn more precise decision boundaries.

Based on the earlier analysis, we aim to achieve more precise recognition results by relying on fine-grained discriminative features. Therefore, we employ cross entropy to compute the classification loss for the logits corresponding to discriminative features:

$$L_{\text{select}} = L_{\text{CE}}(\mathbf{target}_i, \mathbf{logits}_i) \tag{6}$$

Meanwhile, for the features corresponding to non-discriminative regions, we consider them as "background" information. As these features contribute relatively little to classification, we anticipate their prediction probabilities to exhibit a relatively flat distribution. Therefore, we define the

flatten loss as follows:

$$L_{\text{flatten}} = \frac{1}{N} \sum_{i=1}^N (\tanh(\mathbf{logits}_i) + 1)^2 \tag{7}$$

Through the backpropagation of  $L_{\text{flatten}}$ , we aim to drive the logits' values toward negative infinity, thereby obtaining a flat prediction probability distribution.

Based on the above, the overall loss, denoted as  $L$ , can be expressed as:

$$L = \lambda_s L_{\text{select}} + \lambda_f L_{\text{flatten}} \tag{8}$$

where  $\lambda_s$  and  $\lambda_f$  are the weighting parameters for  $L_{\text{select}}$  and  $L_{\text{flatten}}$ , respectively. In our experiments, we set  $\lambda_s = \lambda_f = 1$ .

The objective of designing the feature selector is to enable the model to automatically learn highly discriminative regions within images without relying on pre-extracted facial landmarks or other fine-grained semantic information. Furthermore, it seeks to achieve finer and more accurate recognition by mining the latent relationships between discriminative features. By extracting fine-grained features from facial images, we can more effectively distinguish those facial expressions that exhibit confusion.

## 4 Experiment

### 4.1 Experiment settings

Datasets

**RAF-DB** [11] is one of the most renowned benchmark datasets in facial expression recognition. This dataset comprises 29,672 facial images meticulously annotated by 40 trained annotators. We only utilized 15,339 images that were labeled with six basic emotions and neutral expressions. Out of this subset, 12,271 images were allocated for training purposes; while, the remaining 3068 were designated for testing.

**AffectNet** [33] stands as the largest facial expression recognition dataset to date, offering annotations for both classification and emotional valence-arousal dimensions. This

dataset was assembled by querying facial expression-related keywords in three search engines, resulting in a collection of over one million images, with manual annotations for 450,000 images. It encompasses eight emotion categories, including seven primary facial expressions and the additional category of contempt.

**FER +** [34], derived from the FER2013 dataset, features 28,709 training samples, 3589 validation samples, and 3589 testing samples. All images are in grayscale format and were collected via the Google search engine. The images have a uniform resolution of  $48 \times 48$  pixels. Each image was independently annotated by ten different annotators. Like AffectNet, the facial images in FER + are also annotated for eight different expressions.

#### Implementation details

The facial images used for training were resized to  $224 \times 224$  pixels. During the pre-training phase, we conducted training for 400 epochs with a batch size of 16. Subsequently, we performed fine-tuning for 50 epochs with a batch size 32. The training process employed the AdamW optimizer with an initial learning rate of  $1e-3$ . The initial 10% of epochs were designated as the warm-up stage. Following that, we employed a cosine annealing learning rate scheduler. The proposed method was implemented using the PyTorch framework.

## 4.2 Experimental results

### Comparison with self-supervised learning methods

In this study, we conducted a systematic evaluation of the performance differences between the proposed MFS method and other self-supervised learning methods. Initially, we carried out pre-training on the AffectNet dataset and further fine-tuned the model on multiple diverse datasets (with random sampling) to obtain evaluation results. Additionally, we assessed the effectiveness of MFS's pre-training on the RAF-DB dataset. It is worth noting that due to the limited information contained in the  $48 \times 48$ -pixel grayscale images in the FER + dataset, we did not perform pre-training evaluation on the model using the FER + dataset.

According to Table 1, our MFS not only excels in fine-tuning but also demonstrates outstanding performance in cross-dataset transfer learning tasks. MFS achieved the highest accuracy rates of 63.49%, 60.75%, 91.45%, and 90.16% on four different datasets, significantly surpassing other self-supervised learning methods. These results further confirm the ability of MFS to learn facial representations with stronger robustness and broader generalization capabilities.

Comparison with supervised learning state-of-the-art methods

In this section, we further explored the performance differences between MFS and state-of-the-art supervised learning methods. Specifically, we selected a backbone pre-trained

on AffectNet and fine-tuned it under class-balanced sampling/random sampling conditions. The choice of sampling strategy was relevant to the sample distribution in the dataset. As observed in Table 2, MFS demonstrates competitive performance. On AffectNet7, AffectNet8, RAF-DB, and FER +, MFS exhibits performance differences of only 0.98%, 0.89%, 0.76%, and 0.70%, respectively, compared to state-of-the-art methods. Furthermore, MFS can rapidly adapt the pre-trained backbone to other data domains with minimal computational overhead, a feat that traditional supervised learning methods struggle to achieve.

## 4.3 Ablation study

To investigate the impact of different components of the MFS on the final results, we conducted extensive ablation experiments on the RAF-DB dataset, where MFC denotes multi-level feature combiner, and FS denotes feature selector (Table 3).

When MFC and FS were not utilized, the model achieved an accuracy of 86.27%. When MFC and FS were used individually, the accuracy increased by 1.83% and 0.89%, respectively. However, when MFC and FS were combined, the accuracy improved to 89.41%. This improvement is not simply the mechanical summation of the two components but rather a result of their synergistic enhancement. By integrating these two techniques, the network can perform fine selection among a richer set of facial representations, leading to a significant performance gain.

Furthermore, we explored the number of discriminative features selected (denoted as 's' in Eq. 5). The results indicated that the Vision Transformer backbone with a patch size 16 performed best when selecting 128 features. We speculate that this is related to the characteristics of facial images, where approximately 60% of the information in aligned facial images is crucial. This finding may provide valuable insights for future research.

We also attempted to visualize the regions the model focused on using Grad-CAM [46] to further analyze the effectiveness of the feature selector.

In Fig. 3, we can observe the impact of not using Feature Selector. (b) and using feature selector (c) on the model. Compared to (b), (c) can pinpoint more discriminative and fine-grained facial features within the facial image. This further demonstrates that Feature Selector can significantly enhance the network's ability to extract fine-grained features.

In addition, we investigated the model's confusion matrices (Fig. 4) when utilized both without and with the Feature Selector.

As Fig. 4 shows, noticeable inter-class and intra-class confusions were observed within the Fear and Disgust categories. Upon integrating the feature selector, accuracy was substantially improved by 23% and 7% for Fear and Disgust,

**Table 1** Experimental result of self-supervised learning methods with random sampling

Method	Pre-trained	Fine-tuning/transfer learning			
		AffectNet7	AffectNet8	RAF-DB	FER +
SimCLR[35]	AffectNet	51.80	48.11	83.51	81.48
MoCo[36]	AffectNet	53.69	50.17	80.74	83.58
MoCoV2[37]	AffectNet	56.00	53.40	85.46	86.74
SimSiam[38]	AffectNet	59.97	56.79	85.53	86.58
BYOL[39]	AffectNet	57.09	55.31	87.65	87.63
MFS(Ours)	AffectNet	63.49	60.75	91.45	90.16
	RAF-DB	60.04	57.49	89.41	88.84

**Table 2** Comparison with supervised learning methods. Experiments on AffectNet used class-balanced sampler

Method	Year	AffectNet7	AffectNet8	RAF-DB	FER +
pACNN[16]	ICPR 2018	55.33	–	83.27	–
gACNN[40]	TIP 2018	58.78	–	85.07	–
RAN[1]	TIP 2020	–	59.50	86.90	88.55
SCN[41]	CVPR 2020	–	60.23	88.14	88.01
DAN[17]	2023	65.69	62.09	89.70	–
Efficient-Face[18]	AAAI 2021	63.70	59.89	88.36	–
MA-Net[2]	TIP 2021	64.53	60.29	88.40	–
KT[21]	TIP 2021	63.97	–	88.07	90.49
HO Loss[42]	TVC 2022	64.02	–	89.47	89.03
Meta-Face2Exp[43]	CVPR 2022	64.23	–	88.54	–
APViT[44]	TAC 2022	66.91	–	91.98	<b>90.86</b>
MM-Net[45]	TVC 2023	65.05	–	89.77	89.34
POSTER[3]	ICCV 2023	67.31	63.34	92.05	–
POSTER + + [4]	2023	<b>67.49</b>	<b>63.77</b>	<b>92.21</b>	–
<b>MFS(Ours)</b>	2023	66.51	62.88	91.45	90.16

Bold is employed to denote the optimal performance under different methods/settings

**Table 3** Ablation on each component of MFS

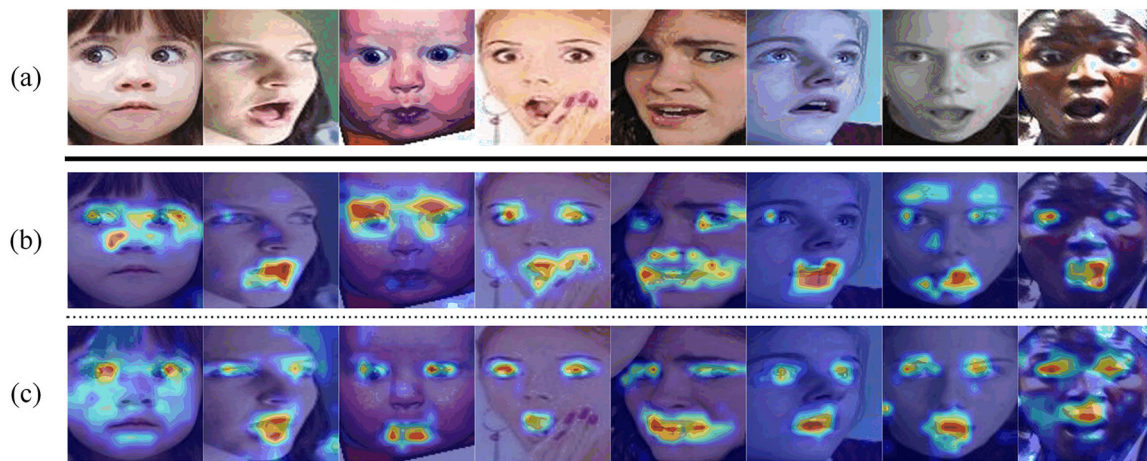
Pre-trained	MFC	FS	Feature selected	RAF-DB
RAF-DB	×	×	196	86.27
	✓	×	196	88.10
	×	✓	128	87.16
	✓	✓	32	88.27
	✓	✓	64	88.49
	✓	✓	128	<b>89.41</b>
	✓	✓	160	88.46

Bold is employed to denote the optimal performance under different methods/settings

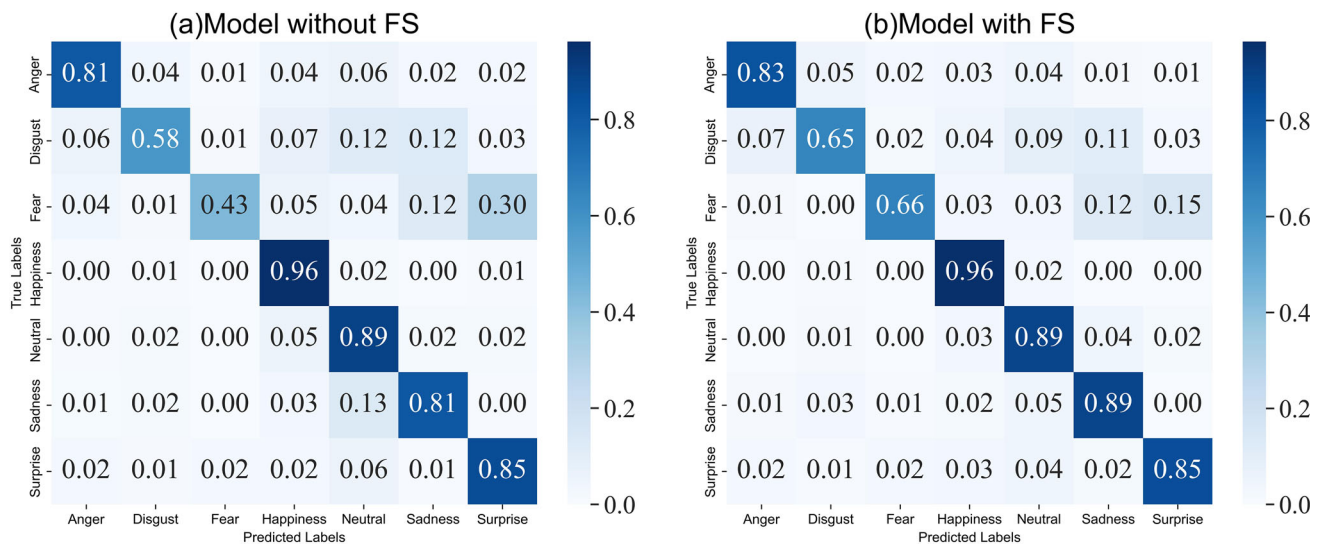
respectively. Concurrently, the confusions between Fear, Disgust, and other categories notably decreased. These outcomes indicate that the feature selector aids the network in capturing subtler distinctions among similar facial expressions, thereby mitigating performance deterioration resulting from inter-class similarities and intra-class variations to a certain extent.

## 5 Discussion

In this section, we will delve into various attempts and explorations made during the research process, laying the foundation for future studies.



**Fig. 3** Attention visualization via Grad-CAM, wherein **a** represents the input, **b** is without the application of Feature Selector, and **c** is with the incorporation of feature selector



**Fig. 4** Confusion matrix of model without (a)/with (b) feature selector

**Table 4** Experimental results of different fusion solutions

Fusion blocks(ViT-base)	Acc. (%)
0, 2, 4, 11	88.95
0, 2, 4, 6, 11	<b>89.41</b>
0, 2, 4, 6, 8, 11	89.05
0, 2, 4, 6, 8, 10, 11	88.69

Bold is employed to denote the optimal performance under different methods/settings

## 5.1 Transformer blocks used in multi-level feature combiner

Regarding facial representation learning, we experimented with the fusion of multiple hierarchical features. We used

features from five different levels as the final fusion scheme based on the results (Table 4).

Furthermore, we examined the weights assigned to each layer in Fig. 5. In this analysis, it was observed that the features from the deepest layer consistently held the highest weight. In most cases, the weights of *Block*<sub>0</sub> closely approximated the weights of the deepest-level features. For shallower features, their weights decreased as the layers became deeper. In the case of the fusion of seven layers of features, the features from *Block*<sub>10</sub> displayed negative weights. These findings suggest that: (1) Shallow features contribute significantly to the final output, particularly the features from *Block*<sub>0</sub>. (2) If too many layers are involved in the fusion, redundant information may be introduced into the network.



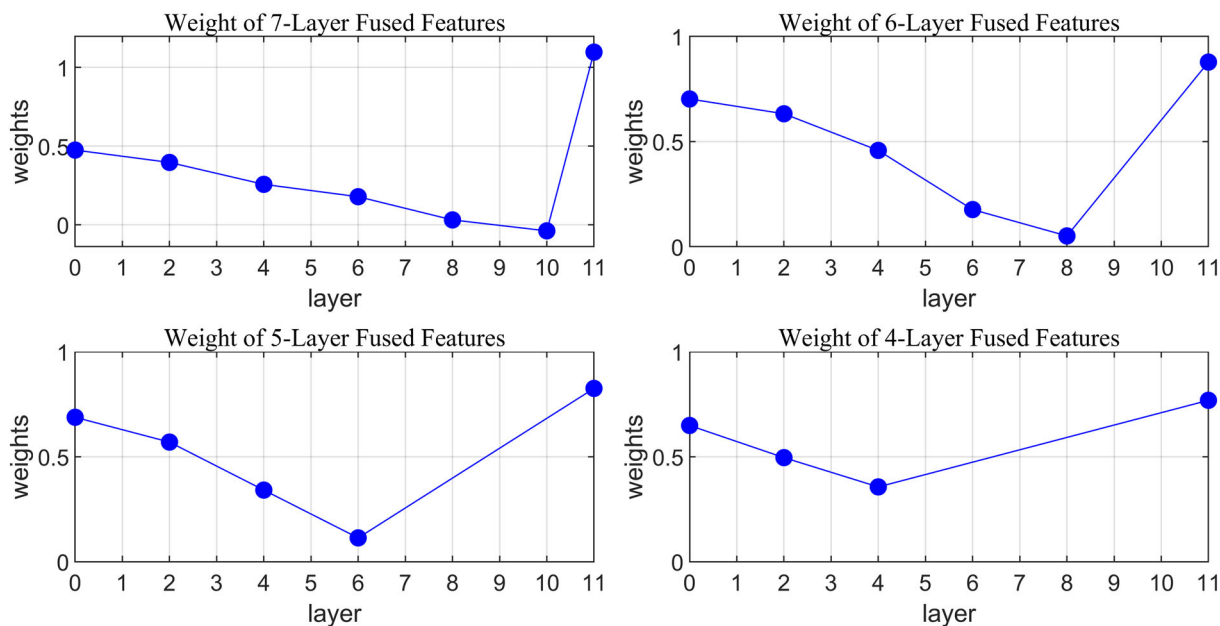


Fig. 5 Weights of features from different transformer blocks

### 5.2 Mask strategy

Students who only practice simple problems will find it challenging to handle complex exams. In order to enable the model to learn a better facial feature representation, we attempt to guide mask generation during the pre-training process based on the attention maps of a well-trained model, which sets a higher challenge for the Mask Image Modeling task.

Given the input features  $f \in R^{B \times L \times D}$  for Block<sub>11</sub>. During the self-attention computation,  $f$  undergoes three independent linear mapping layers, respectively generating the  $Q$ ,  $K$ , and  $V$  matrices. Subsequently, through the scaled dot-product attention computation, we obtain the attention matrix  $A_h \in R^{L \times L}$ . We take the average attention matrices from multiple heads to obtain the averaged attention matrix  $\hat{A}$ .

$$A_i = \text{softmax}\left(\frac{QK^T}{\sqrt{D}}\right)V, \hat{A} = \frac{1}{h} \sum_{i=1}^h A_i \tag{9}$$

In the attention matrix  $\hat{A}$ , each row corresponds to the attention distribution of a token. We extract the attention vector associated with the class token from the first row and resize it into a square shape. This results in the attention map  $M \in R^{(\sqrt{L-1}) \times (\sqrt{L-1})}$  associated with the class token. Based on  $M$ , we propose the following two masking strategies:

- (1) Attention-high: Prioritize masking patches with high attention scores.

- (2) Attention-clue: While masking high-attention-score patches, retain a certain percentage of the patches with the highest attention scores as reference clues for subsequent reconstruction (Fig. 6).

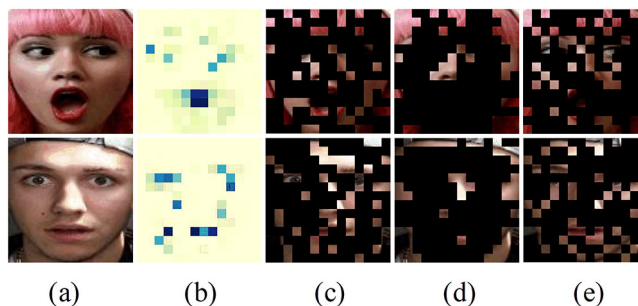
Experimental results indicate that using attention maps to guide mask generation significantly increases the difficulty of the mask image modeling task (as shown in Fig. 7). The loss value for pre-training with the attention-high and attention-clue masking strategies is notably higher than that with the random strategy. Regrettably, the increased task difficulty did not lead to enhanced performance.

Based on Table 5, we can observe that models using random masking significantly outperform those following the attention-high and attention-clue strategies in terms of recognition accuracy. Upon further analysis, it seems unwise to increase task difficulty during the early stages of training. A more ideal approach might be incrementally raising the challenge once the model has accumulated a foundational knowledge base. However, pinpointing the exact moment to escalate this difficulty remains a challenge. A more promising strategy might involve adopting a learnable masking method, allowing the model to autonomously adjust the difficulty of the task. Therefore, designing a learnable masking strategy tailored explicitly for facial representation learning will be the central direction of our future research.

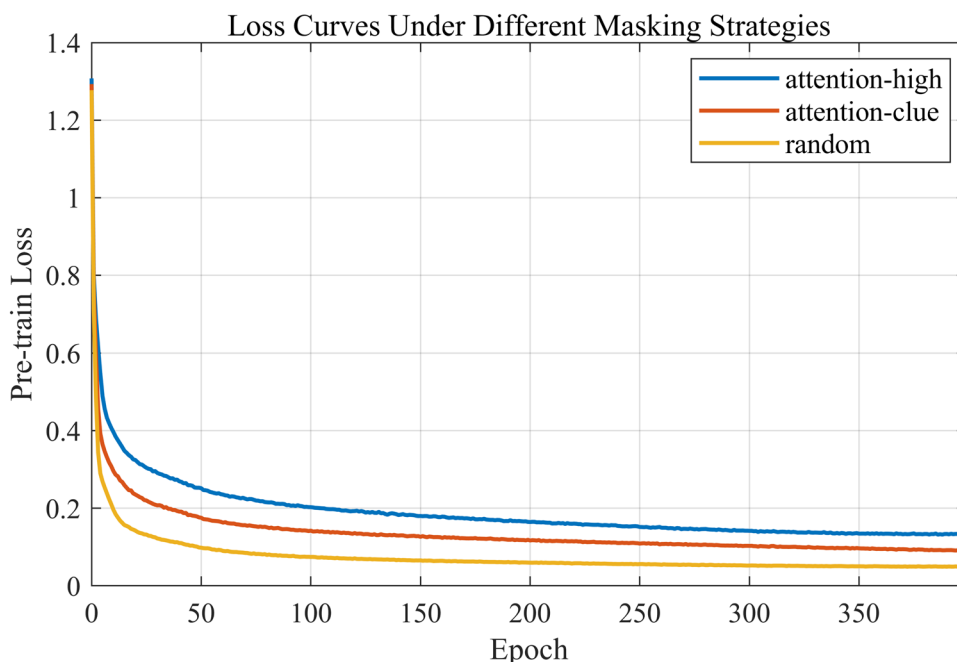
### 5.3 Flatten loss

We explored several strategies to make the prediction probability distribution of non-discriminative features more uniform. Specifically, we applied an activation to the logits and

**Fig. 6** Investigation of masking Strategies: **a** input **b** attention map **c** random **d** attention-high **e** attention-clue



**Fig. 7** Loss curves under various masking strategies



**Table 5** Experimental results of different masking strategies

Mask strategy	Mask ratio(clue ratio)(%)	Acc. (%)
Random	75	<b>89.41</b>
Att. high	75	84.94
Att. clue	75(15)	87.06

Bold is employed to denote the optimal performance under different methods/settings

guided all activation values to converge to a fixed label value gradually:

$$L_{\text{flatten}} = \text{MSE}(\text{Activation}(\mathbf{logits}), \mathbf{label}) \quad (10)$$

**Table 6** Experimental results of different flatten loss

Act./Label	ReLU/0	Variance/0	Tanh/-1	None/(1/num of classes)
Acc. (%)	88.52	87.74	<b>89.41</b>	88.49

Bold is employed to denote the optimal performance under different methods/settings

The experimental results are shown in Table 6. After comprehensive comparison and analysis, we found that among all tested combinations, (tanh/−1) produced the best results. Based on this observation, we chose this combination as the flatten loss.

## 6 Conclusion

In this paper, we propose a self-supervised training strategy for facial expression recognition called MFS (multi-level feature selector). During the pre-training phase, we employ the multi-level feature combiner to achieve multi-level facial representation learning. Subsequently, utilizing a meticulously

designed feature selector, the network can adaptively filter out fine-grained features with discriminative solid power. These features are then fed into a graph convolutional network for graph feature extraction and aggregation. MFS effectively addresses challenges in facial expression recognition related to relying on large-scale annotated data, handling inter-class similarities, and intra-class variations. Experimental results on multiple FER benchmarks demonstrate that the proposed MFS outperforms the supervised learning baseline and other self-supervised methods.

**Author contributions** Heng-Yu An contributed to investigation, conceptualization, methodology, software, and writing—original draft; Rui-Sheng Jia contributed to supervision, methodology, writing—review & editing, and funding acquisition.

**Funding** This work was supported by the Humanities and Social Science Fund of the Ministry of Education of the People's Republic of China (22YJAZH036).

**Data availability** The datasets utilized in this study, namely RAF-DB, FERPlus, and AffectNet, can be accessed at the following respective URLs: RAF-DB: <http://www.whdeng.cn/raf/model1.html>. AffectNet: <http://mohammadmahoor.com/affectnet>. FER + : <https://github.com/Microsoft/FERPlus>

## Declarations

**Conflict of interest** The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## References

- 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). <https://doi.org/10.1109/tip.2019.2956143>
- 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). <https://doi.org/10.1109/tip.2021.3093397>
- Zheng, C., Mendieta, M., Chen, C.: Poster: a pyramid cross-fusion transformer network for facial expression recognition. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3146–3155 (2023) <https://doi.org/10.1109/iccvw60793.2023.00339>
- Mao, J., Xu, R., Yin, X., Chang, Y., Nie, B., Huang, A.: POSTER V2: a simpler and stronger facial expression recognition network. Preprint at [arXiv:2301.12149](https://arxiv.org/abs/2301.12149). (2023) <https://doi.org/10.48550/arXiv.2301.12149>
- Shi, J., Xiu, Y., Tang, G.: Research on occlusion block face recognition based on feature point location. *Comput. Anim. Virtual Worlds* **33**(3–4), e2094 (2022). <https://doi.org/10.1002/cav.2094>
- Li, H., Wang, N., Yang, X., Wang, X., Gao, X.: Towards semi-supervised deep facial expression recognition with an adaptive confidence margin. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4166–4175 (2022) <https://doi.org/10.1109/cvpr52688.2022.00413>
- He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009 (2022) <https://doi.org/10.1109/cvpr52688.2022.01553>
- Xie, Z., Zhang, W., Sheng, B., Li, P., Chen, C.P.: BaGFN: broad attentive graph fusion network for high-order feature interactions. *IEEE Trans. Neural Netw. Learn. Syst.* (2021). <https://doi.org/10.1109/TNNLS.2021.3116209>
- Ekman, P., Friesen, W.V.: *Facial Action Coding Systems*. Consulting Psychologists Press (1978)
- Chen, J., Chen, Z., Chi, Z., Fu, H.: Facial expression recognition based on facial components detection and hog features. In: *International Workshops on Electrical and Computer Engineering Subfields*, pp. 884–888 (2014)
- Li, S., Deng, W., Du, J.: 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*, pp. 2852–2861 (2017) <https://doi.org/10.1109/cvpr.2017.277>
- 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: *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pp. 302–309. IEEE (2018) <https://doi.org/10.1109/fg.2018.00051>
- 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) <https://doi.org/10.1109/wacv48630.2021.00245>
- Zhao, S., Cai, H., Liu, H., Zhang, J., Chen, S.: Feature selection mechanism in CNNs for facial expression recognition. In: *BMVC*, **12**, pp. 317 (2018) <https://doi.org/10.1109/ieeegcc.2009.5734265>
- Hasani, B., Negi, P.S., Mahoor, M.H.: BRReG-NeXt: facial affect computing using adaptive residual networks with bounded gradient. *IEEE Trans. Affect. Comput.* **13**(2), 1023–1036 (2020). <https://doi.org/10.1109/TAFFC.2020.2986440>
- Li, Y., Zeng, J., Shan, S., Chen, X.: Patch-gated CNN for occlusion-aware facial expression recognition. In: *2018 24th International Conference on Pattern Recognition (ICPR)*, pp. 2209–2214. IEEE (2018) <https://doi.org/10.1109/ICPR.2018.8545853>
- Wen, Z., Lin, W., Wang, T., Xu, G.: Distract your attention: multi-head cross attention network for facial expression recognition. *Biomimetics* **8**(2), 199 (2023). <https://doi.org/10.3390/biomimetics8020199>
- 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*, **35** (4), pp. 3510–3519 (2021) <https://doi.org/10.1609/aaai.v35i4.16465>
- Li, H., Wang, N., Yang, X., Wang, X., Gao, X.: Unconstrained facial expression recognition with no-reference de-elements learning. *IEEE Trans. Affect. Comput.* (2023). <https://doi.org/10.1109/tip.2022.3186536>
- Li, H., Wang, N., Yang, X., Gao, X.: CRS-CONT: a well-trained general encoder for facial expression analysis. *IEEE Trans. Image Process.* **31**, 4637–4650 (2022). <https://doi.org/10.1109/tip.2022.3186536>
- Li, H., Wang, N., Ding, X., Yang, X., Gao, X.: Adaptively learning facial expression representation via cf labels and distillation. *IEEE Trans. Image Process.* **30**, 2016–2028 (2021). <https://doi.org/10.1109/tip.2021.3049955>
- Roy, S., Etemad, A.: Self-supervised contrastive learning of multi-view facial expressions. In: *Proceedings of the 2021 International Conference on Multimodal Interaction*, pp. 253–257 (2021) <https://doi.org/10.1145/3462244.3479955>
- Shu, Y., Gu, X., Yang, G.-Z., Lo, B.: Revisiting self-supervised contrastive learning for facial expression recognition. Preprint at [arXiv:2210.03853](https://arxiv.org/abs/2210.03853). (2022) <https://doi.org/10.48550/arXiv.2210.03853>

24. Ma, B., An, R., Zhang, W., Ding, Y., Zhao, Z., Zhang, R., et al.: Facial action unit detection and intensity estimation from self-supervised representation. Preprint at [arXiv:2210.15878](https://arxiv.org/abs/2210.15878). (2022) <https://doi.org/10.48550/arXiv.2210.15878>
25. Cai, Z., Ghosh, S., Stefanov, K., Dhall, A., Cai, J., Rezatofghi, H., et al.: Marlin: masked autoencoder for facial video representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1493–1504 (2023) <https://doi.org/10.1109/cvpr52729.2023.00150>
26. Sun, L., Lian, Z., Liu, B., Tao, J.: Mae-dfer: efficient masked autoencoder for self-supervised dynamic facial expression recognition. In: Proceedings of the 31st ACM International Conference on Multimedia, pp. 6110–6121 (2023) <https://doi.org/10.48550/arXiv.2307.02227>
27. Esmaili, V., Shahdi, S.O.: Automatic micro-expression apex spotting using Cubic-LBP. *Multimedia Tools Appl.* **79**, 20221–20239 (2020). <https://doi.org/10.1007/s11042-020-08737-5>
28. Esmaili, V., Mohassel Feghhi, M., Shahdi, S.O.: Spotting micro-movements in image sequence by introducing intelligent cubic-LBP. *IET Image Proc.* **16**(14), 3814–3830 (2022). <https://doi.org/10.1049/ipr2.12596>
29. Happy, S., Routray, A.: Automatic facial expression recognition using features of salient facial patches. *IEEE Trans. Affect. Comput.* **6**(1), 1–12 (2014). <https://doi.org/10.1109/TAFFC.2014.2386334>
30. Marrero Fernandez, P. D., Guerrero Pena, F. A., Ren, T., Cunha, A.: Feratt: facial expression recognition with attention net. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 0–0 (2019) <https://doi.org/10.1109/cvprw.2019.00112>
31. Li, H., Wang, N., Yu, Y., Yang, X., Gao, X.: LBAN-IL: a novel method of high discriminative representation for facial expression recognition. *Neurocomputing* **432**, 159–169 (2021). <https://doi.org/10.1016/j.neucom.2020.12.076>
32. Park, N., Kim, S.: How do vision transformers work?. Preprint at [arXiv:2202.06709](https://arxiv.org/abs/2202.06709) (2022) <https://doi.org/10.48550/arXiv.2202.06709>
33. Mollahosseini, A., Hasani, B., Mahoor, M.H.: Affectnet: a database for facial expression, valence, and arousal computing in the wild. *IEEE Trans. Affect. Comput.* **10**(1), 18–31 (2017). <https://doi.org/10.1109/TAFFC.2017.2740923>
34. Barsoum, E., Zhang, C., Ferrer, C. C., Zhang, Z.: Training deep networks for facial expression recognition with crowd-sourced label distribution. In: Proceedings of the 18th ACM International Conference on Multimodal Interaction, pp. 279–283 (2016) <https://doi.org/10.1145/2993148.2993165>
35. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning, pp. 1597–1607. PMLR (2020)
36. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9729–9738. (2020) <https://doi.org/10.1109/cvpr42600.2020.00975>
37. Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. Preprint at [arXiv:2003.04297](https://arxiv.org/abs/2003.04297). (2020) <https://doi.org/10.48550/arXiv.2003.04297>
38. Chen, X., He, K.: Exploring simple siamese representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15750–15758 (2021) <https://doi.org/10.1109/cvpr46437.2021.01549>
39. Grill, J.-B., Strub, F., Alché, F., Tallec, C., Richemond, P., Buchatskaya, E., et al.: Bootstrap your own latent—a new approach to self-supervised learning. *Adv. Neural. Inf. Process. Syst.* **33**, 21271–21284 (2020)
40. 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). <https://doi.org/10.1109/TIP.2018.2886767>
41. 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, pp. 6897–6906 (2020) <https://doi.org/10.1109/cvpr42600.2020.00693>
42. Li, H., Xiao, X., Liu, X., Guo, J., Wen, G., Liang, P.: Heuristic objective for facial expression recognition. *Vis. Comput.* **39**(10), 4709–4720 (2023). <https://doi.org/10.1007/s00371-022-02619-7>
43. Zeng, D., Lin, Z., Yan, X., Liu, Y., Wang, F., Tang, B.: Face2exp: combating data biases for facial expression recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 20291–20300 (2022) <https://doi.org/10.1109/cvpr52688.2022.01965>
44. Xue, F., Wang, Q., Tan, Z., Ma, Z., Guo, G.: Vision transformer with attentive pooling for robust facial expression recognition. *IEEE Trans. Affect. Comput.* (2022). <https://doi.org/10.1109/TAFFC.2022.3226473>
45. Xia, H., Lu, L., Song, S.: Feature fusion of multi-granularity and multi-scale for facial expression recognition. *Vis. Comput.* (2023). <https://doi.org/10.1007/s00371-023-02900-3>
46. 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, pp. 618–626 (2017) <https://doi.org/10.1109/iccv.2017.74>

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.



**Heng-Yu An** currently studying Computer Science and Technology at Shandong University of Science and Technology. His main interests involve the application of neural networks in the field of visual affective computing.



**Rui-Sheng Jia** is currently a Full professor at the College of Computer Science and Engineering, Shandong University of Science and Technology, China. He has more than 30 first-author publications and has more than 50 co-author publications. His research interest includes artificial intelligence, big data processing, information fusion, microseismic monitoring and inversion.