Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.054644

ARTICLE

A Facial Expression Recognition Method Integrating Uncertainty Estimation and Active Learning

Yujian Wang[1], Jianxun Zhang[1,*] and Renhao Sun[2]

[1]School of Computer Science and Engineering, Chongqing University of Technology, Chongqing, 400054, China

[2]Data Space Research Institute of Hefei Comprehensive National Science Center, Hefei, 230088, China

*Corresponding Author: Jianxun Zhang. Email: zjx@cqut.edu.cn

Received: 03 June 2024 Accepted: 23 August 2024 Published: 15 October 2024

ABSTRACT

The effectiveness of facial expression recognition (FER) algorithms hinges on the model's quality and the availability of a substantial amount of labeled expression data. However, labeling large datasets demands significant human, time, and financial resources. Although active learning methods have mitigated the dependency on extensive labeled data, a cold-start problem persists in small to medium-sized expression recognition datasets. This issue arises because the initial labeled data often fails to represent the full spectrum of facial expression characteristics. This paper introduces an active learning approach that integrates uncertainty estimation, aiming to improve the precision of facial expression recognition regardless of dataset scale variations. The method is divided into two primary phases. First, the model undergoes self-supervised pre-training using contrastive learning and uncertainty estimation to bolster its feature extraction capabilities. Second, the model is fine-tuned using the prior knowledge obtained from the pre-training phase to significantly improve recognition accuracy. In the pre-training phase, the model employs contrastive learning to extract fundamental feature representations from the complete unlabeled dataset. These features are then weighted through a self-attention mechanism with rank regularization. Subsequently, data from the low-weighted set is relabeled to further refine the model's feature extraction ability. The pre-trained model is then utilized in active learning to select and label information-rich samples more efficiently. Experimental results demonstrate that the proposed method significantly outperforms existing approaches, achieving an improvement in recognition accuracy of 5.09% and 3.82% over the best existing active learning methods, Margin, and Least Confidence methods, respectively, and a 1.61% improvement compared to the conventional segmented active learning method.

KEYWORDS

Expression recognition; active learning; self-supervised learning; uncertainty estimation

1 Introduction

Facial Expression Recognition (FER) is a crucial research area within computer vision, with extensive applications across various fields such as human-computer interaction [1], autism detection [2], and safe driving [3]. The continuous advancements in deep learning techniques, particularly the widespread adoption of the Transformer model [4], have increased demand for large-scale datasets.

However, the primary bottleneck hindering further progress in FER is the acquisition of sufficiently large labeled datasets.

To address this challenge, current research in FER increasingly focuses on methods to efficiently extract and learn feature representations from limited labeled data, thereby mitigating the constraints imposed by dataset size on model performance [5–7]. In FER research, constructing labeled datasets is time-consuming and labor-intensive despite the vast availability of unlabeled image resources on the Internet, as illustrated in Fig. 1. In contemporary research, one of the primary challenges is identifying valuable samples for labeling within the vast expanse of unlabeled data, especially when labeling resources are limited. Active learning offers a promising solution to this issue. It employs a strategic approach to select the most informative unlabeled samples for labeling. These newly annotated data points subsequently enhance the model, improving its capability to recognize essential features. Studies have applied active learning methods to the FER domain in recent years, achieving specific results [8]. However, the quality of labeling remains a critical factor limiting performance. High-quality labeling is particularly challenging due to the subjectivity of labelers and the ambiguity of face images in the wild. This uncertainty often results in labeling inconsistencies and errors, exacerbating the "cold-start" problem. In this context, the model faces difficulties in learning effective feature representations from the initially labeled dataset, resulting in the suboptimal selection of valuable samples in subsequent active learning iterations. When this issue is pronounced, the model's performance may deteriorate compared to scenarios without the application of an active learning method. Furthermore, the performance of different active learning methods on the same dataset can be inconsistent. Some simple methods may not perform as well as more complex ones on the same dataset, challenging the assumption that more complex methods are always better.

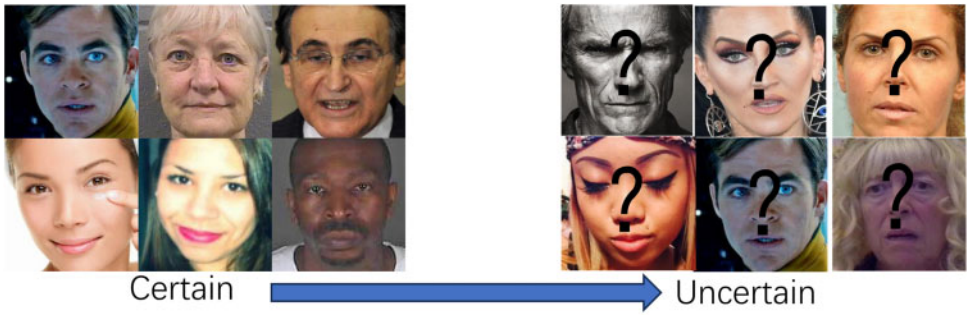


Figure 1: Description of uncertainty in real-face images of RAF-DB. Sample quality has an excellent influence on active learning

In response to the challenges above, this paper proposes a novel end-to-end recognition network, FUCLE (Fusion of Uncertainty Control and Active Learning) for FER. This network excels in image feature extraction and integrates the re-labeling of uncertain labels and an active learning mechanism for unlabeled data. Specifically, the network structure is divided into three modules: First, we introduce a supervised contrastive learning pre-training strategy, which guides the network to learn shared and distinctive information, thereby extracting underlying representations from the entire unlabeled dataset and establishing a solid foundation for the subsequent active learning phase. Secondly, a dynamic re-labeling module (DRM) is developed during the pre-training stage. This module employs a self-attention importance weighting mechanism to assign weights to the images involved in pre-training. It uses a ranking regularization module to group and regularize these weights. This dynamically re-labels unreliable samples and handles ambiguous annotations, reducing

the negative impact of labeling errors on model training and enhancing the model's performance in the initial training stages. This step significantly enhances the dependability of the pre-trained model. Ultimately, fine-tuning ensures that the initial model, which has undergone active learning, encounters a diverse set of representative samples right from the beginning. This exposure allows the model to acquire a more extensive understanding of facial features during the early stages of training. Our contributions to this work can be summarized as follows:

(1) We propose an innovative active learning method for FER, which combines the pre-training and fine-tuning stages and introduces a dynamic relabeling module in the pre-training phase, effectively solving the cold-start problem in active learning. Experiments show that this method significantly improves model performance.

(2) A dynamic ranking regularization mechanism is devised to supervise the learning process of the self-attention weighting network. This mechanism ensures that the Facial Expression Uncertainty-based Contrastive Learning FUCLE (Facial Expression Uncertainty-based Contrastive Learning) framework can learn meaningful and important weights, providing a robust foundation for subsequent steps, particularly the relabeling module.

(3) Extensive validation is conducted on three widely used benchmark datasets (FER13, RAFDB, and KDEF). Experimental results showcase the method's superiority over several existing state-of-the-art approaches in terms of performance. Notably, it achieves state-of-the-art (SOTA) results, with accuracy rates of 66.81%, 80.70%, and 94.17% on FER13, RAFDB, and KDEF, respectively.

2 Related Work

2.1 Active Learning

Active learning focuses on selectively annotating the most informative samples from unlabeled datasets through a well-designed selection mechanism. This process aims to enhance the model's performance most efficiently [9].

Deep learning models typically require large datasets to support their complex training processes. However, obtaining these labeled datasets often presents numerous challenges and limitations in practical applications. Consequently, active learning, which intelligently selects the most valuable samples for labeling, plays a crucial role in enhancing the efficiency of data labeling, reducing labeling costs, and optimizing model performance [10–12]. For example, GLISTER [13] employs a two-stage optimization strategy that applies various sampling methods to the entire dataset. This ensures that the selected samples accurately represent the overall distribution and characteristics of the data. This requirement increases computational complexity and cost. Additionally, adversarial attacks have been introduced into active learning to estimate decision boundaries for different classes. Using this approach, researchers can select samples near the decision boundary for labeling [14].

2.2 Uncertainty Estimate

In tackling the problem of labeling ambiguity, current research focuses on uncertainty estimation with noisy labels, aiming to mitigate the detrimental effects of ambiguous labeling on recognition effectiveness. Specifically, significant progress has been made in the area of uncertainty estimation. MentorNet, proposed by Jiang et al. [15], guides the underlying deep network through curriculum learning, enabling the network to prioritize samples likely to be labeled correctly. The confident learning approach proposed by Northcutt et al. [16] handles labeling errors by estimating the joint distribution of noisy and clean labels. The method introduced by Shu et al. [17] overcomes the

over-segmentation problem present in biased training data by adaptively learning explicit weighting functions. The symmetric learning approach explored by Wang et al. [18] solves the learning complex problem by introducing a symmetric loss function and effectively copes with the overfitting of the cross-entropy loss function under noisy labels. Li et al. [19] proposed a complexity-aware center loss method that distinguishes between easy and hard samples to generate more representative center vectors, thereby reducing the interference of complex samples and improving facial expression recognition accuracy. However, while avoiding the interference of hard-to-recognize complex samples on the class centers, some useful class information from these complex samples is discarded.

2.3 Self-Supervised Learning

Self-supervised Learning (SSL) has emerged as a prominent approach in unsupervised representation learning, demonstrating significant performance improvements across various fields. The core idea of SSL is to define an auxiliary task, known as the pre-text task, which enables the model to learn the data's intrinsic structure and key features without any human annotation [20–22]. Recent advances in SSL have focused on comparative learning as a framework. Wang et al. [23] proposed a self-supervised learning-based occlusion facial expression recognition method that generates diverse occluded facial images by randomly adding occluders and defines occlusion prediction as a pre-training task. This enables the model to learn occlusion-invariant facial features. Morais et al. [24] allows models to learn valuable representations from large amounts of unlabeled data and then use these representations to fine-tune downstream tasks such as emotion recognition. Torpey proposed that many of the SimCLR methods [25] play a remarkable role in training high-quality representations by maximizing the consistency of two different augmented versions of an unlabeled image in the embedding space. Li et al. [26] proposed a novel Structural Self-Contrast Learning (SSCL) method and its enhanced version, SSCL+, for effectively learning features for FER. This approach considers the relationships between feature map positions, reducing the interference of irrelevant features by encouraging two different representations of the same sample to have similar structures. The method extracts facial expression features in a self-supervised manner without relying on specifically labeled augmented samples, thus mitigating the impact of noisy labels and addressing the problem of insufficient samples by increasing the number of training samples. Li et al. [27] also introduced a method that integrates two cognitive feature modules into deep neural networks to enhance FER. The Adaptive Feature Relative Transformation (AFRT) module creates explicit features through relative transformation. In contrast, the Adaptive Feature Graph Convolutional Network (AFGCN) module considers the interaction between expression classes to generate complementary features, significantly improving FER performance. Chen et al. [28] proposed an automatic analysis method based on computer vision to predict student engagement in collaborative learning environments using a multi-modal deep neural network (MDNN). Umer et al. [29] developed a deep fusion model-based FER method, which extracts facial regions through preprocessing and integrates features from multiple Convolutional Neural Network (CNN) models for controlling a music player. Shen et al. [30] introduced a FER algorithm based on the ResNet18 network structure and a multi-channel attention mechanism. Integrating Multi-Channel Attention (MCA) into ResNet18, the approach harmonizes self-attention and channel attention mechanisms to extract richer facial expression features.

3 Method

To effectively leverage active learning to improve recognition accuracy in expression recognition, we propose a simple and efficient network called FUCLE (Fusion of Uncertainty Control and Active Learning for facial expression recognition). This section provides a brief overview of the

FUCLE network structure, followed by an introduction to its three modules: the contrastive learning pre-training module, the DRM module, and the active learning module. Finally, the experimental procedure for the overall network is outlined.

3.1 Overview

To address the cold start challenge in active learning, a dual-phase training approach is introduced. This approach combines pre-training, integrating a dynamic relabeling module with active learning. In the initial phase, we utilize the SimCLR framework for contrastive learning, employing data augmentation and contrastive loss to learn meaningful representations. This pre-training phase aims to capture the fundamental structure of the data, providing a robust initialization for the subsequent active learning phase. Notably, during the pre-training phase, we leverage the re-labeling mechanism of the DRM module to suppress labels of highly uncertain samples, thereby achieving better model performance during pre-training. In the subsequent step, active learning is implemented by iteratively selecting and labeling the most informative samples, utilizing the representations gained during the pre-training stage. This approach ensures a strong initialization for the active learning phase, mitigating the cold start problem. As illustrated in Fig. 2, given an expression recognition model $E(\cdot)$ with a set of labeled training samples $\boldsymbol{I}$, and a subset of small labeled training samples $\boldsymbol{I}_s$ ($\boldsymbol{I}_s \in \boldsymbol{I}$), the pre-training phase proceeds as follows: First, labeled data $x \in \boldsymbol{I}$ is selected. Using a variable image enhancement method, the original image x is generated. x and an enhanced image x'. A deep neural network (e.g., ResNet50) is then employed to extract feature mappings x and x' from these two augmented images, utilizing the principles of SimCLR. Next, the similarity between the two augmented images derived from the same original image is computed. A contrastive loss function is subsequently calculated. *RRLOSS* (see Eq. (6)), where the objective is to reduce the differences between projections of the same image and increase the differences between projections of distinct images. Specifically, The integration of a DRM module takes place, which weighs the images in the batch according to their characteristics. The images are dynamically ranked according to these weights, with those exhibiting high uncertainty being relabeled. The model is pre-trained in a self-supervised manner using the entire labeled set, enabling it to learn robust feature representations from the data.

The traditional active learning training process is followed in the second step. The prior knowledge gained from the first step aids the model in learning discriminative representations without overfitting. Initially, the frame uses the data x_s from the small labeled subset $x_s \in \boldsymbol{I}_s$. Then select batches of quantitative data and incrementally add data with predictive values $\overline{p}$. more significant than the threshold σ to the training dataset. This approach facilitates more efficient identification of representative samples during subsequent cycles of the active learning process, thus effectively addressing the cold-start issue.

3.2 A Framework for Supervised Comparative Learning

Supervised Contrastive Learning (SCL) is a contrastive learning method incorporating supervised information to enhance a model's feature representation capabilities. Unlike traditional self-supervised contrastive learning, SCL leverages category labels to construct positive and negative sample pairs, improving performance in supervised tasks. When applied to pre-trained models for expression recognition, the SCL framework can significantly enhance the model's ability to distinguish between different expression features.

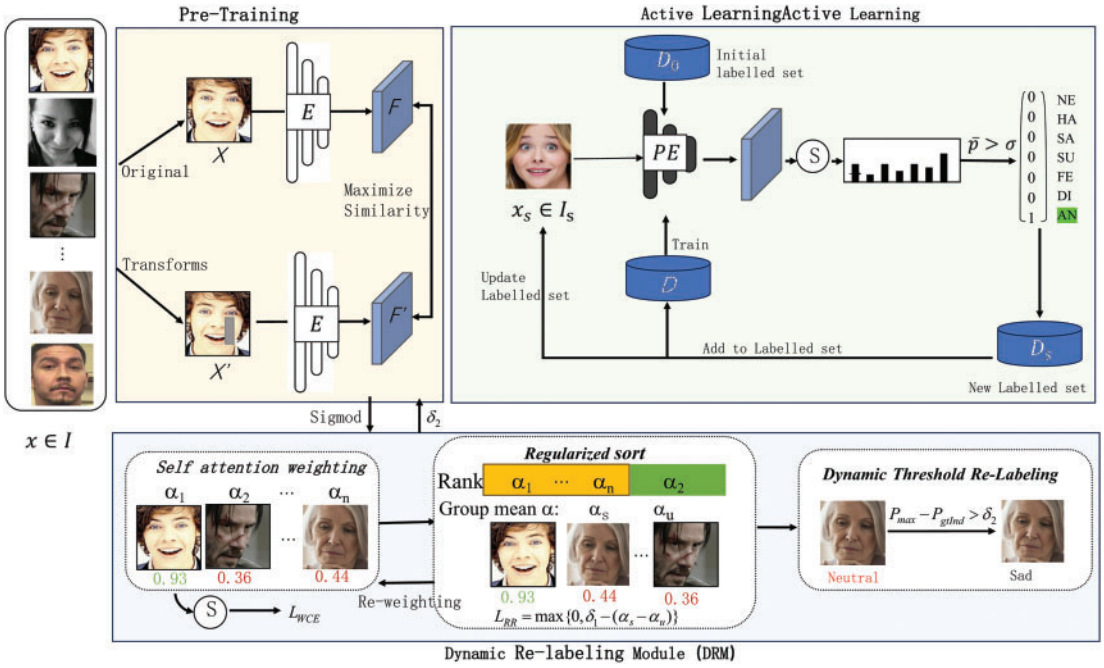


Figure 2: An overview of the two-step training approach: initially, the pre-training model employs comparative self-supervised learning objectives. Simultaneously, it utilizes the DRM module to relabel data with high uncertainty, thereby enabling the model to learn the fundamental data representations. Finally, Further fine-tuning of the initialized model is conducted with active learning to enhance its performance

SimCLR stands out as a prominent contrastive self-supervision technique that has significantly contributed to advancing contrastive learning within the realm of computer vision [31,32]. The core idea of this method lies in learning from both positive and negative samples. Positive samples refer to variations or modifications of the input samples, commonly created via data augmentation, while all other samples are considered negative in relation to the input sample. Contrastive learning effectively captures the underlying data representation by clustering positive samples together and pushing them apart from negative samples within the embedding space.

In expression recognition, supervised contrastive learning harnesses label information to construct more comprehensive sets of positive and negative sample pairs, enhancing the model's ability to discern different expressions. Specifically, the augmented version and additional samples from the same expression category are considered positive samples for each sample. To elaborate, for a given small batch of data$\{(x_k, y_k)\}_{k=1}^{N}$, where x_k represents a sample and y_k denotes the corresponding expression category label, the positive sample set is defined as $P(i) = \{p \in \{1, \ldots, N\} \mid y_p = y_i, p \neq i\}$, the loss function for supervised contrastive learning is defined as follows:

$$L_{i,j} = \frac{1}{|P(i)|} \sum_{p \in P(i)} -log \frac{\exp\left(\frac{\sin(z_i, z_j)}{\tau}\right)}{\sum_{a=1}^{N} \mathbf{1}_{[a \neq i]} \exp\left(\frac{\sin(z_i, z_a)}{\tau}\right)} \tag{1}$$

where $\mathbf{z}_i$ and $\mathbf{z}_j$ are the embedding representations of sample "i" and positive sample p, respectively, τ is a temperature parameter, and $\mathbf{1}_{[a\neq i]}$ is an indicator function that is 1 when $k \neq i$ and 0 otherwise.

In the model, supervised contrastive learning efficiently leverages expression category labeling information to acquire a more discriminative feature representation. This enhances the model's capacity to differentiate between various expressions and bolsters its generalization performance in expression recognition tasks. Compared to traditional unsupervised contrastive learning methods, supervised contrastive learning demonstrates superior recognition accuracy and robustness in expression recognition tasks.

3.3 Uncertainty Estimation Based Relabeling Module

3.3.1 Self-Attention Weighting

The DRM (Dynamic Re-labeling Module), based on standard CNN structures, includes three main elements: self-attentive importance weighting, dynamic rank regularization, and relabeling of low-weighted groups. The self-attentive module evaluates each sample's training value. Some samples may be deemed more significant, while uncertain ones may be assigned lower importance. Let $T = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N] \in R^{D\times N}$, where D and N represent the feature dimensions and the number of samples, respectively. The self-attentive importance weighting module processes this input and assigns an importance weight to each feature. In this study, we employ the log-weighted cross-entropy loss, as described in [33]. This module consists of a fully connected linear layer coupled with a sigmoid-type activation function. The mathematical representation of this module is as follows:

$$\alpha_i = \sigma\left(\mathbf{W}_a^{\mathrm{T}}\mathbf{x}_i\right) \tag{2}$$

where α_i is the important weight of the ith sample, $\mathbf{W}_a^T$ is the parameter used for attention in the FC layer, and σ is a sigmod-type function.

3.3.2 Log-Weighted Cross-Entropy Loss

During the loss weighting process utilizing attention weights, a straightforward method involves multiplying the weight of each individual sample by its corresponding sample loss. However, during training, it is observed that a practical phenomenon occurs: the attention weights tend to converge to zero. This occurs because when the weights approach zero, the loss function also diminishes, resulting in a trivial solution. There is a need for a strategy to constrain or regularize these weights to ensure accurate reflection of the multi-class cross-entropy loss. this approach is termed logit-weighted cross-entropy loss (WCE-Loss), characterized by the formula:

$$L_{WCE} = -\frac{1}{N}\sum_{i=1}^{N} log\frac{e^{\alpha_i \mathbf{W}_{y_i}^{\mathrm{T}}\mathbf{x}_i}}{\sum_{j=1}^{C} e^{\alpha_i \mathbf{W}_j^{\mathrm{T}}\mathbf{x}_i}} \tag{3}$$

where W_j is the jth classifier. Liu et al. [34] showed that L_{WCE} is positively correlated with α.

3.3.3 Regularized Sort

In the above module, the self-attention weights can vary arbitrarily within the range of (0, 1). To better handle uncertain samples, a rank regularization module is integrated to adjust the attention weights. In this module, the attention weights are first sorted in descending order and then divided into two groups based on a predefined ratio. The purpose of rank regularization is to ensure that the

average attention weight of the high-importance group remains marginally higher than that of the low-importance group. To accomplish this, the rank regularization loss (RR-Loss) is formally defined as:

$$L_{RR} = \max\ \{0,\ \delta_1 - (\alpha_s - \alpha_u)\} \tag{4}$$

where δ_1 is a marginal value that controls the proportion of the overall number of low-ranking groups, and $\alpha_s - \alpha_u$ is the value of the standard deviation of a batch sample of groups having high importance groups. During training, the comprehensive loss function is defined as:

$$L_{all} = \gamma L_{RR} + (1 - \gamma)\, L_{WCE} \tag{5}$$

where γ is the trade-off ratio.

3.3.4 Dynamic Threshold Re-Labeling

Uncertain samples typically exhibit lower importance weights, suggesting a need for a strategy to reassign labels to these samples. The central difficulty in adjusting annotations lies in determining their accuracy. The relabeling module employs a method to assess samples with lower importance weights based on Softmax probabilities. For each sample, a comparison is made between the highest predicted probability and the probability assigned to the original label. When the former exceeds the latter, the sample is reassigned a new pseudo-label. Formally, the dynamic thresholding relabeling process operates as follows:

$$y' = \begin{cases} l_{max} & \text{if } P_{max} - P_{gtInd} > \delta_2 \\ l_{org} & \text{otherwise} \end{cases} \tag{6}$$

δ_2 represents the threshold value, P_{max} denotes the highest probability among all predictions, P_{gtInd} signifies the probability assigned specifically to the given label. l_{org} and l_{max} are the indexes of the original given label and the maximum prediction, respectively. In particular, to further enhance the effectiveness of the ranking process, we propose the dynamic relabeling boundary method, which can affect the robustness and error correction behavior of the model through the increase and decrease of margin. Increasing δ_2 reduces the frequency of relabeling but may miss some subtle correction opportunities, while decreasing δ_2 increases the frequency of reliability but also increases the risk of the rough model making wrong corrections. The specific dynamic adjustment method is as follows:

$$\delta_2 = \begin{cases} \delta_2 + increase & if\ bestval > Td \\ \max\,(0,\ \delta_2 - decrease) & otherwise \end{cases} \tag{7}$$

Increases and decreases represent the magnitude of the value for each adjustment, respectively.

4 Experimentation

In this section, the experimental configuration and outcomes are detailed. Initially, implementation specifics and dataset characteristics are provided. Subsequently, the learning outcomes of various active learning techniques are examined, followed by confirmation of our module's efficacy through ablation tests and sensitivity analyses to justify each parameter choice. Conclusively, the findings of our proposed approach are showcased, with an in-depth exploration of the method's multiple facets.

4.1 Datasets

The RAF-DB is a natural FER dataset comprising 30,000 labeled facial images. For this study, images representing seven basic expressions have been selected: neutral, happy, sad, surprised, fear, disgust, and anger. The training set comprises 12,271 images, while the test set comprises 3068 images.

The FER13 is a large-scale FER dataset expanded from FER2013, containing 28,709 training images and 3589 test images. In addition to the seven primary expressions present in RAF-DB, this dataset includes the expression of contempt. Each image in FER13 has been manually labeled with one of eight expressions by ten human annotators.

The KDEF dataset has been deliberately illuminated with relatively even, soft lighting and the subjects wear uniform T-shirt colors. The portraits in the test set are free of beards, earrings, glasses, and visible makeup. The dataset consists of 70 individuals (35 males and 35 females) aged between 20 and 30. Each person displays seven different expressions, each captured from five different angles, resulting in a total of 4900 color images, each sized at $562 * 762$ pixels.

4.2 Comparison Experiments

In this section, a comprehensive evaluation of the performance of various active learning strategies on three expression recognition datasets—FER13, RAF-DB, and KDEF—is conducted. The comparative analysis covers a range of active learning methods, such as random sampling, and highlights several significant patterns. Firstly, while some active learning methods demonstrate superior performance compared to random sampling on specific datasets, overall, the efficacy of existing methods in the FER task remains to be enhanced. Our experimental results indicate that although certain active learning methods (Entropy, Margin, and Least Confidence) show improvements over random sampling on some datasets, the overall performance gains for FER tasks are not as significant as expected. This suggests that existing methods may not fully optimize facial expression datasets' inherent complexity and variability. In Table 1, on the KDEF dataset, except for the Least Confidence and GLISTER methods, all other methods exhibit lower performance than Random Sampling. This suggests the necessity for more specialized active learning methods to enhance performance in the FER task. Secondly, by comparing the performance of different methods on the same dataset, Observations indicate that comparatively straightforward methods, such as Least Confidence, can achieve superior results under specific conditions. The Least Confidence method consistently exhibits performance enhancements across the three datasets—FER13, RAF-DB, and KDEF. Particularly noteworthy is its performance on the KDEF dataset, where it achieves an accuracy of 87.03%, surpassing the accuracy of random sampling at 85.57%. This underscores the effectiveness of simpler methods in certain scenarios.

Table 1: Performance of different active learning methods on FER13, RAF-DB and KDEF datasets

Methods	FER13	RAF-DB	KDEF
Random sampling (baseline)	63.16	73.82	83.77
Entropy [35]	65..88	76.99	81.78
Margin [23]	65.97	76.88	83.55
Least confidence [23]	65.83	77.33	87.03
BADGE [31]	64.22	76.88	84.22

(Continued)

Table 1 (continued)

Methods	FER13	RAF-DB	KDEF
GLISTER [13]	64.13	76.88	84.71
CoreSet [32]	64.90	75.11	81.22
Bald [36]	65.22	75.32	79.88
Adv.deepfool [37]	64.21	78.01	80.11
Pre-training + DRM + Least confidence (ours)	**66.81**	**80.70**	**94.17**

However, it is also noted that while most active learning methods realize improvements on more extensive datasets (e.g., FER13 and RAF-DB), their performance is relatively suboptimal on smaller datasets (e.g., KDEF). This phenomenon may be attributed to the "cold-start" problem, wherein the model struggles to learn a generalized representation due to the limited number of initial labeled samples. An active learning approach is proposed that integrates the DRM during pre-training to mitigate this issue and enhance model performance across all datasets. As depicted in the table, this approach significantly enhances model performance on the KDEF dataset, achieving an accuracy of 94.17%, markedly higher than other methods. Moreover, the method demonstrates notable performance improvements on the FER13 and RAF-DB datasets, with accuracies reaching 66.81% and 80.70%, respectively. In summary, the comparative data presented in the table underscores the performance disparities among various active learning methods in the FER task. Our experimental findings demonstrate that active learning methods combined with DRM pre-training effectively enhance performance in expression recognition tasks, particularly on smaller datasets. This outcome underscores the potential of active learning methods in FER tasks and provides robust empirical support for our subsequent research endeavors.

4.3 Ablation Experiments

In this section, the influence of SSL (self-supervised learning) pre-training on the effectiveness of various active learning approaches is investigated through a rigorous set of ablation experiments.

The primary aim is to not only to assess the improvement in model performance solely resulting from pre-training but also to explore how pre-training, when integrated with active learning strategies, effectively addresses the cold-start challenge inherent in active learning. The analysis commences by examining the data presented in Table 2. Compared against the best-existing benchmark method (Least confidence), Experiments were conducted on the pre-training-based active learning method (Pre-training + Least confidence) across three datasets (FER13, RAFDB, KDEF). Results indicate that the combination of pre-training and active learning approach elevates the model's recognition accuracy from 65.83% to 66.17% on the FER13 dataset. Notably, Significant performance enhancements are observed on the RAFDB and KDEF datasets, reaching 78.36% and 92.71%, respectively. This marked improvement suggests that the pre-training process endows the model with a more nuanced representation of initial features, enabling the model to converge to superior solutions more swiftly in subsequent active learning iterations. Particularly, it excels in the KDEF dataset, which is characterized by its smaller data size. Subsequently, an active learning approach incorporating the uncertainty estimation module DRM (Pre-training-DRM + Least confidence) is introduced. This approach not only leverages the initial feature representation from pre-training but also guides the

model to downplay uncertain data by computing uncertainty estimates of feature representations. As depicted in Table 2, when combining DRM and Least confidence active learning methods atop pre-training, performance across all three datasets is further bolstered to 66.81%, 80.70%, and 94.17%, respectively. This notable performance enhancement underscores that the active learning approach, integrating feature representation uncertainty and pre-training, adeptly identifies and utilizes critical information within the data, thereby enhancing model performance. Through pre-training, the model acquires generalized features and representations from the data, facilitating rapid adaptation and convergence within the active learning process. Concurrently, the active learning method integrates feature representation uncertainty and accurately discerns and leverages key data information, further refining model performance.

Table 2: Module validity experiments

Method	Baseline	Pre-training	Uncertainty	FER13	RAFDB	KDEF
Least confidence	√			65.83	77.33	87.03
Pre-training + Least confidence	√	√		66.17	78.36	92.71
Pre-training + DRM + Least confidence	√	√	√	**66.81**	**80.70**	**94.17**

4.4 Computational Cost Analysis

A detailed analysis is conducted on the computational costs associated with different methods. The assessment includes an objective examination of GPU memory usage, CPU memory consumption, runtime efficiency, and accuracy achieved. This approach ensures an unbiased evaluation of the performance characteristics and inherent trade-offs of each methodology. As shown in Table 3, while the pre-training and uncertainty estimation steps increase computational costs, they also significantly enhance model accuracy. To enhance the analysis's credibility, computational cost comparisons based on ablation experiments have been included. Firstly, regarding GPU memory usage, the Least Confidence method uses 1.2 GB of memory. After incorporating pre-training, the memory usage increases to 23.4 GB, and with the addition of the DRM module, it slightly rises to 23.5 GB. This indicates that pre-training and uncertainty estimation steps require substantial GPU memory. Secondly, CPU memory usage exhibits a similar trend. The Least Confidence method uses 0.63 GB of CPU memory. After adding pre-training, this increases to 2.1 GB, and further inclusion of the DRM module raises it to 3.0 GB. This shows that pre-training and uncertainty estimation demand significant CPU resources. In terms of runtime, the Least Confidence method takes 427 s. With pre-training, the runtime increases to 446.56 s; with the DRM module, it is 449.05 s. Although there is an increase in runtime, it is acceptable considering the improvement in accuracy. While the complexity increases, it is justified by the significant improvement in accuracy. Finally, regarding accuracy, the Least Confidence method achieves 87.03%. With the introduction of pre-training, accuracy rises to 92.71%, and with the DRM module, it further increases to 94.17%. This demonstrates that, although pre-training and uncertainty estimation increase computational costs, they significantly enhance model accuracy.

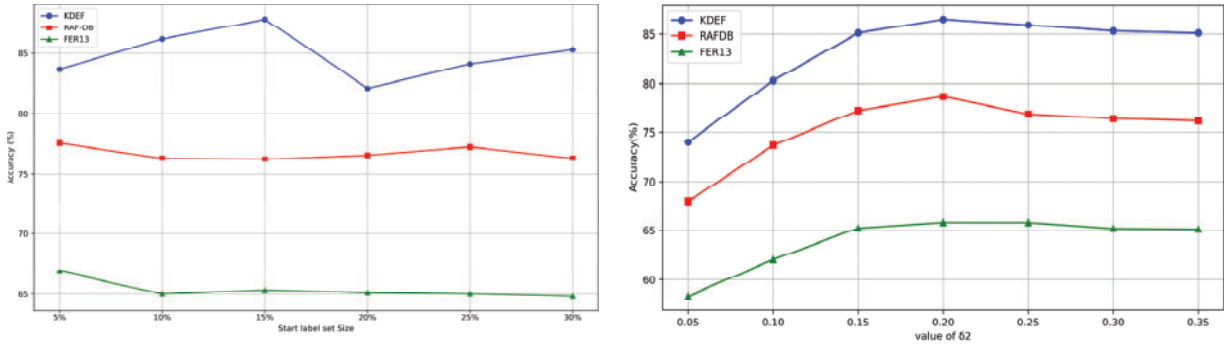
Table 3: The computational cost of each method on the KDEF dataset for one epoch

Cost	Method		
	Least confidence	Pre-training + Least confidence	Pre-training + DRM + Least confidence
Graphics card memory (G)	1.2	23.4	23.5
CPU memory (G)	0.63	2.1	3.0
Run time (s)	427	446.56	449.05
Accuracy (%)	87.03	92.71	94.17

4.5 Sensitivity Experiment

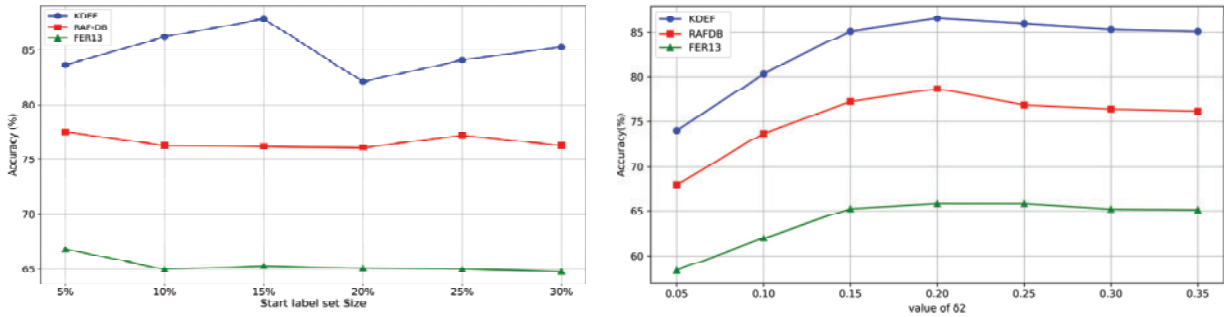
A comprehensive sensitivity study was conducted to thoroughly understand the roles of pre-training and active learning methods in the overall model training process, focusing on several key hyperparameters. These include: (a) The size of the initial labeled dataset for active learning. (b) The number of active learning cycles. (c) Two boundary threshold hyperparameters.

The focus is on the active learning method demonstrating the best performance in the FER task. Experiments were conducted across three datasets: FER13, RAF-DB, and KDEF. In Fig. 3, Initial Labeled Dataset Size: The size of the initial labeled dataset is a crucial hyperparameter for active learning. Choosing a larger value reduces the number of samples available for selection in subsequent cycles, while selecting too small a value varying initial labeled dataset sizes on the FER task was investigated using a two-step training scheme. (a) The results show that selecting smaller initial values leads to superior performance. For example, FER13 and RAF-DB achieve their peak performance with an initial labeling set of just 5%. However, for KDEF, optimal accuracy is reached with a 15% initial labeled dataset size. This pattern can be traced back to the latent representations learned through self-supervised pre-training, which facilitates efficient learning from a small labeled dataset initially and enables the selection of the most informative samples, thereby potentially addressing cold-start issues. Furthermore, as the initial labeled dataset size increases, there is a noticeable rise in standard deviation, especially evident in the KDEF dataset. (b) Regarding the number of training cycles, it is another pivotal factor in active learning. More cycles permit the model to identify more representative samples in later iterations but come with increased training expenses. The sensitivity to different numbers of training cycles was evaluated and it was found that the optimal performance for the FER13, RAF-DB, and KDEF datasets is achieved with 7 training cycles. This indicates that a judicious choice of the number of training cycles significantly enhances model performance while avoiding unnecessary computational overhead. (c) Fig. 4 shows threshold hyperparameters δ_1 and δ_2 the effects of different values δ_1 (ranging from 0 to 0.35) on the KDEF, RAF-DB, and FER13 datasets. Experimental results demonstrate that all three datasets achieve high accuracy when δ_1 values are close to 0.15. Similarly, for δ_2 ranging from 0 to 0.35, optimal accuracy is attained when δ_2 values are close to 0.2. Smaller δ_2 values lead to more incorrect relabeling operations, whereas larger $\delta_2$values result in fewer or no relabeling operations.



(a)Initial labelled set size. (b) Number of cycles.

Figure 3: Sensitivity studies of the parameters of the two-step training scheme on three datasets



(a) Impact of δ_1 (b) Impact of δ_2

Figure 4: Evaluation of the margin δ_1 and δ_2 on the KDEF/RAF-DB/FER13 dataset

5 Conclusion

In this paper, we introduced a novel self-supervised active learning approach integrating uncertainty estimation to address the prevalent cold-start problem encountered in small and medium-sized datasets within FER. While deep learning models have shown success in FER tasks, their performance heavily relies on large-scale models and extensive labeled data. However, acquiring labeled data is resource-intensive, particularly for smaller datasets. Therefore, efficiently utilizing limited labeled data has become a pivotal research focus. Firstly, through a combination of comparative self-supervised pre-training and uncertainty estimation, our model learns the underlying feature representation of the overall unlabeled dataset without requiring an abundance of labeled data.

Furthermore, the model's feature extraction capability is enhanced by incorporating the Dynamic Rank Regularization (DRM) module, enabling a more precise capture of intrinsic regularities and variations in the data during feature learning. In the second stage, the pre-trained model is leveraged for active learning, strategically selecting and labeling information-rich samples. The efficacy of this phase hinges on the effective utilization of the pre-trained models' feature extraction capability. The proposed method demonstrates significant improvements in recognition performance on limited labeled data. Experimental results showcase substantial enhancements in FER accuracy compared to random sampling and learning loss-based active learning methods. Overall, our proposed method validates its effectiveness and introduces novel insights for research in the realm of facial expression recognition.

In future investigations, we intend to examine model pruning methods and lightweight approaches to refine the algorithm. Our objective is to preserve detection precision while minimizing computational demands, thereby enabling smoother integration into real-world scenarios.

Acknowledgement: Thanks are extended to the editors and reviewers.

Funding Statement: This project is supported by National Science Foundation of China (61971078); Chongqing Municipal Education Commission Science and Technology Major Project (KJZD-M202301901).

Author Contributions: Yujian Wang conducted the experiment. Renhao Sun and Jianxun Zhang analyzed the data. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data used to support the findings of this study are available from the corresponding author upon request.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] M. Derball, M. Jarrah, and P. Randhawa, "Autism spectrum disorder detection: Video games based facial expression diagnosis using deep learning," *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 1, pp. 111–119, 2023. doi: 10.14569/issn.2156-5570.

[2] A. Perosanz, O. Martínez, P. Espinosa-Blanco, I. García, M. AlRashaida and J. F. López-Paz, "Comparative analysis of emotional facial expression recognition and empathy in children with prader-willi syndrome and autism spectrum disorder," *BMC Psychol.*, vol. 12, pp. 94–108, 2024. doi: 10.1186/s40359-024-01590-3.

[3] L. Sharara *et al.*, "A real-time automotive safety system based on advanced AI facial detection algorithms," *IEEE Trans. on Intell. Veh.*, vol. 9, no. 6, pp. 5080–5100, 2024. doi: 10.1109/TIV.2023.3272304.

[4] K. Han *et al.*, "A survey on vision transformer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 87–110, 2023. doi: 10.1109/TPAMI.2022.3152247.

[5] B. Fang, X. Li, G. X. Han, and J. H. He, "Rethinking pseudo-labeling for semi-supervised facial expression recognition with contrastive self-supervised learning," *IEEE Access*, vol. 11, pp. 45547–45558, 2023. doi: 10.1109/ACCESS.2023.3274193.

[6] J. R. Zhong, T. X. Chen, and L. H. Yi, "Face expression recognition based on NGO-BILSTM model," *Front. Neurorobot.*, vol. 17, pp. 1150038–1150048, 2023. doi: 10.3389/fnbot.2023.1155038.

[7] X. Zhang, D. B. Huang, H. Y. Li, Y. J. Zhang, Y. Xia and J. Z. Liu, "Self-training maximum classifier discrepancy for EEG emotion recognition," *CAAI Trans. on Intel. Tech.*, vol. 8, no. 4, pp. 1480–1491, 2023. doi: 10.1049/cit2.12174.

[8] M. U. Ahmed, K. J. Woo, K. Y. Hyeon, M. R. Bashar, and P. K. Rhee, "Wild facial expression recognition based on incremental active learning," *Cogn. Syst. Res.*, vol. 52, pp. 212–222, 2018. doi: 10.1016/j.cogsys.2018.06.017.

[9] K. Konyushkova, R. Sznitman, and P. Fua, "Learning active learning from data," *Adv. Neural Inf. Process. Syst.*, vol. 30, pp. 4228–5238, 2017. doi: 10.5555/3294996.3295177.

[10] J. K. Yuan *et al.*, "Bi3D: Bi-domain active learning for cross-domain 3D object detection," in *Proc. 5th Conf. Comput. Vis. Pattern Recognit.*, Oxford, England, Jun. 2023, pp. 15599–15608.

[11] D. Yuan *et al.*, "Active learning for deep visual tracking," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 2, pp. 1–13, 2023. doi: 10.1109/TNNLS.2023.3266837.

[12] V. L. Nguyen, M. H. Shaker, and E. Hüllermeier, "How to measure uncertainty in uncertainty sampling for active learning," *Mach. Learn.*, vol. 111, pp. 89–122, 2022. doi: 10.1007/s10994-021-06003-9.
[13] K. Killamsetty, D. Sivasubramanian, G. Ramakrishnan, and R. Iyer, "GLISTER: Generalization based data subset selection for efficient and robust learning," in *Proc. 35th AAAI Conf. Artif. Intell.*, Vancouver, Canada, Feb. 2021, pp. 8110–8118.
[14] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. 37th Int. Conf. Mach. Learn.*, New York, NY, USA, Nov. 2020, pp. 1597–1607.
[15] L. Jiang, Z. Y. Zhou, T. Leung, L. J. Li, and F. F. Li, "MentorNet: Learning data-driven curriculum for very deep neural networks on corrupted labels," in *Proc. 35th Int. Conf. Mach. Learn.*, Stockholm, Sweden, Jul. 2018, pp. 2304–2313.
[16] C. Northcutt, L. Jiang, and I. Chuang, "Confident learning: Estimating uncertainty in dataset labels," *J. Artif. Intell. Res.*, vol. 70, pp. 1373–1411, 2021. doi: 10.1613/jair.1.12125.
[17] J. Shu *et al.*, "Meta-Weight-Net: Learning an explicit mapping for sample weighting," *Adv Neural Inf. Process. Syst.*, vol. 32, pp. 1373–1411, 2019. doi: 10.48550/arXiv.1902.07379.
[18] J. H. Wang, H. Ding, and S. F. Wang, "Occluded facial expression recognition using self-supervised learning," in *Proc. 16th Asian Conf. Comput. Vis.*, Macau, China, Dec. 2022, pp. 1077–1092.
[19] H. H. Li, X. Yuan, C. L. Xu, R. Zhang, X. Y. Liu and L. Q. Liu, "Complexity aware center loss for facial expression recognition," *Vis. Comput.*, vol. 9, pp. 1–10, 2024. doi: 10.1007/s00371-023-03221-1.
[20] G. D. Xu, Z. W. Liu, X. X. Li, and C. C. Loy, "Knowledge distillation meets self-supervision," in *Proc. 2th Eur. Conf. Comput. Vis.*, Glasgow, UK, Aug. 2020, pp. 588–604.
[21] S. Z. Ravid and L. C. Yann, "To compress or not to compress—self-supervised learning and information theory: A review," *Spec. Issue Inf. Theoretic Methods Deep Learn. Theory Appl.*, vol. 26, no. 3, pp. 252–270, 2024. doi: 10.3390/e26030252.
[22] V. Rani, S. T. Nabi, M. Kumar, A. Mittal, and K. Kumar, "Self-supervised learning: A succinct review," *Arch. Comput. Methods Eng.*, vol. 30, pp. 2761–2775, 2023. doi: 10.1007/s11831-023-09884-2.
[23] D. Wang and Y. Shang, "A new active labeling method for deep learning," in *Proc. 24th Int. Joint Conf. Neural Netw.*, Beijing, China, Jul. 2014, pp. 112–119.
[24] E. Morais *et al.*, "Speech emotion recognition using self-supervised features," in *Proc. 47th Int. Conf. Acoust. Speech Signal Process*, Singapore, May 2022, pp. 6922–6926.
[25] D. Torpey and R. Klein, "DeepSet SimCLR: Self-supervised deep sets for improved pathology representation learning," 2024, *arXiv:2402.15598*.
[26] H. H. Li, J. H. Zhu, G. H. Wen, and H. Y. Zhong, "Structural self-contrast learning based on adaptive weighted negative samples for facial expression recognition," *Vis. Comput.*, vol. 15, pp. 1–12, 2024. doi: 10.1007/s00371-024-03349-8.
[27] H. H. Li, X. L. Xiao, X. Y. Liu, G. H. Wen, and L. Q. Liu, "Learning cognitive features as complementary for facial expression recognition," *Int. J. Intell. Syst.*, vol. 1, pp. 1–15, 2024. doi: 10.1155/2024/7321175.
[28] Y. Chen, J. Zhou, Q. Gao, J. Gao, and W. Zhang, "MDNN: Predicting student engagement via gaze direction and facial expression in collaborative learning," *Comput. Model. Eng. Sci.*, vol. 136, no. 1, pp. 381–401, 2023. doi: 10.32604/cmes.2023.023234.
[29] S. Umer, R. K. Rout, S. Tiwari, A. A. AlZubi, J. M. Alanazi and K. Yurii, "Human-computer interaction using deep fusion model-based facial expression recognition system," *Comput. Model. Eng. Sci.*, vol. 135, no. 2, pp. 1165–1185, 2023. doi: 10.32604/cmes.2022.023312.
[30] T. Shen and H. Xu, "Facial expression recognition based on multi-channel attention residual network," *Comput. Model. Eng. Sci.*, vol. 135, no. 1, pp. 539–560, 2023. doi: 10.32604/cmes.2022.022312.
[31] J. T. Ashs, C. C. Zhang, A. Krishnamurthy, J. Langford, and A. Agarwal, "Deep batch active learning by diverse, uncertain gradient lower bounds," 2019. Accessed: Feb. 24, 2020. [Online]. Available: https://arxiv.org/abs/1906.03671
[32] O. Sener and S. Savarese, "Active learning for convolutional neural networks: A core-set approach," 2017. Accessed: Jan. 1, 2018. [Online]. Available: https://arxiv.org/abs/1708.00489

[33] W. Hu, Y. Y. Huang, F. Zhang, and R. R. Li, "Noise-tolerant paradigm for training face recognition CNNs," in *Proc. 2th Conf. Comput. Vis. Pattern Recognit.*, Los Angeles, CA, USA, Oct. 2019, pp. 11887–11896.
[34] W. Y. Liu, Y. D. Wen, Z. D. Yu, M. Li, B. S. Raj and L. Song, "SphereFace: Deep hypersphere embedding for face recognition," in *Proc. 1th Conf. Comput. Vis. Pattern Recognit.*, Honolulu, Hawaii, Jul. 2017, pp. 212–220.
[35] J. X. Wu, J. X. Chen, and D. Huang, "Entropy-based active learning for object detection with progressive diversity constraint," in *Proc. 4th Conf. Comput. Vis. Pattern Recognit.*, Tel Aviv, Israel, Dec. 2022, pp. 9397–9406.
[36] Y. Gal, R. Islam, and Z. Ghahramani, "Deep bayesian active learning with image data," in *Proc. 34th Int. Conf. Mach. Learn.*, Amsterdam, Netherlands, Aug. 2017, pp. 1183–1192.
[37] M. Ducoffe and F. Precioso, "Adversarial active learn-ing for deep networks: A margin based approach," 2018, *arXiv:1802.09841*.