Engineering Applications of Artificial Intelligence 129 (2024) 107623

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

Adaptive graph-based feature normalization for facial expression recognition

Yu-Jie Xiong [a,b,*], Qingqing Wang [b,**], Yangtao Du [b], Yue Lu [b]

[a] School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai, 201620, China
[b] Shanghai Key Laboratory of Multidimensional Information Processing, East China Normal University, Shanghai, 200241, China

ARTICLE INFO

Keywords:
Facial expression recognition
Data uncertainties
Feature normalization
Poisson graph generator

ABSTRACT

Facial Expression Recognition (FER) suffers from data uncertainties caused by ambiguous facial images and annotators' subjectiveness, resulting in excursive semantic and feature covariate shifting problems. Existing works usually correct mislabeled data by estimating noise distribution, or guide network training with knowledge domain that learned from clean data, neglecting the associative relations of expression samples. In this work, we propose an **A**daptive **G**raph-based **F**eature **N**ormalization (AGFN) to protect FER models from data uncertainties by normalizing feature distributions with the association of expressions. Specifically, we propose a Poisson graph generator to adaptively construct topological graphs for samples in each mini-batches via a sampling process, and correspondingly design a coordinate descent strategy to optimize proposed model. Our method outperforms state-of-the-art works with accuracies of 91.84%, 91.11% and 61.38% on three benchmark datasets, i.e., FERPlus, RAF-DB and AffectNet. Especially, when the percentage of mislabeled data significantly increases (e.g., to 20%), our method surpasses existing works by 14.09%, 21.12% and 13.67% on above datasets. Our code is publicly available in https://github.com/X-Lab-CN/AGBFN.

1. Introduction

Facial expression is a natural signal to convey emotions and intentions of human beings. Therefore, facial expression recognition (FER) is essential for machines to understand human behaviors and interact with humans. Though great efforts have been made in last decades and promising progress has been achieved, FER still suffers from data uncertainty problem, i.e., data is frequently mislabeled because of the subjectiveness of annotators and the ambiguities of facial images. Existing works on FER rarely focus on this problem, except IPA2LT (Zeng et al., 2018) and Self-Cure Network (SCN) (Wang et al., 2020a), which suppress data uncertainties by discovering the latent truth from inconsistent pseudo labels or relabeling mislabeled data by weighting and ranking samples in a mini-batch. Meanwhile, the facial shadow caused by insufficient lighting leads to a sharp decrease in recognition rate, making it difficult for the system to meet practical requirements. Facial similarity, inability to recognize facial features, blurred facial images caused by motion, or incorrect camera focus all lead to inaccurate facial information received. The above are the main reasons for inaccurate identification (Schlett et al., 2022). With the growth of training samples gathered from internet, data uncertainties have introduced significant challenges to FER, manifesting as disruptive semantic and feature covariate shifts, where distributions of individual classes are with serious overlaps because of mislabeled data.

To further indicate the data uncertainty problem in FER, a more intuitive sample distribution analyzation is given in Fig. 1, where samples from a widely used benchmark FER dataset, i.e., FERPlus, is statistically analyzed according to their labels. In FERPlus, a sample is labeled by 10 annotators, and for each sample, we denote labels with the most and the second most votes as 1st label and 2nd label respectively. Then, three statistics have been made for each classes: the total number of samples (#Total), the number of samples whose 1st label is with a vote number less than 5 (#1st<5), and the number of samples whose 2nd label only has 0, 1, or 2 fewer votes than their 1st label (#2nd-1st≤2). Overall, there are 35 487 samples in FERPlus, of which 3729 (10.5%) satisfy #1st<5 and 4951 (13.95%) satisfy #2nd-1st≤2. That means more than 10% samples are labeled with small confidence and a large portion of samples suffer from data uncertainty problem.

Though learning with noisy labels has been studied extensively in the community of computer vision, existing works (Zheng et al., 2021; Karim et al., 2022) mainly focus on correcting mislabeled data by estimating label quality and noise distribution, or guiding network training with knowledge learned from clean data, neglecting the associative relations of samples. As pointed in Anderson and Bower (2014), when people encounter a vague facial image with fuzzy expression, they often associate it with other images sharing similar expressions, instead of

* Corresponding author at: School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai, 201620, China.
** Corresponding author.
E-mail addresses: xiong@sues.edu.cn (Y.-J. Xiong), qqwang0723@hotmail.com (Q. Wang).

https://doi.org/10.1016/j.engappai.2023.107623
Received 29 May 2023; Received in revised form 20 October 2023; Accepted 25 November 2023
Available online 5 December 2023

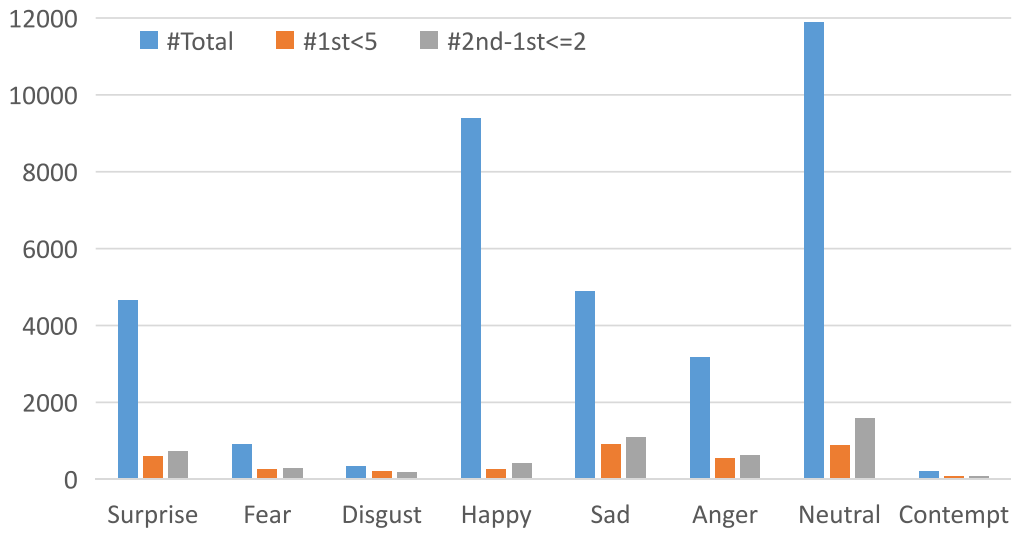

Fig. 1. Sample distribution analyzation in FERPlus dataset regarding to data uncertainty problem.

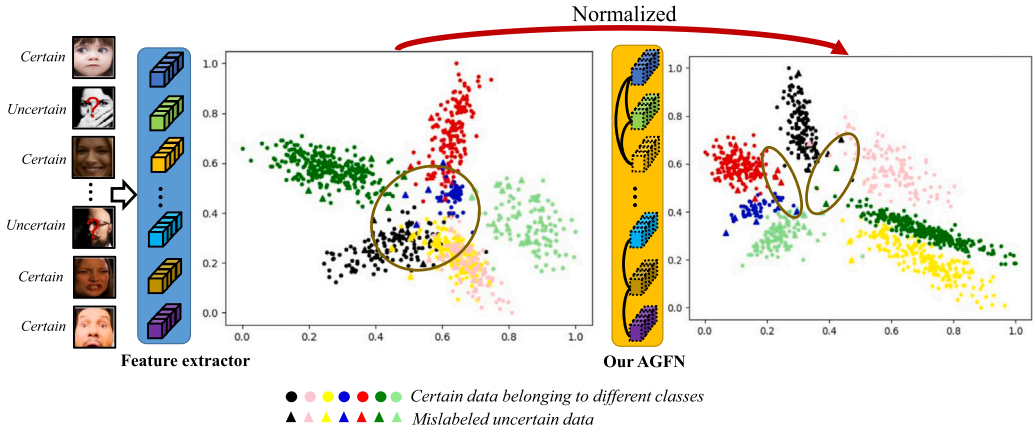

Fig. 2. The excursive semantic and feature covariate shifting problems caused by uncertain data. After being normalized by our AGFN, feature distribution is with much clearer boundaries although mislabeled samples are still there.

starting at its parts for research, which is called human adaptively associative learning process. In other words, humans tend to make associative comparisons when discerning subtle expressions.

In this work, we propose an efficient normalization method called Adaptive Graph-based Feature Normalization (AGFN) to tackle the data uncertainty problem by normalizing feature distributions with the association of expressions. Specifically, given feature maps extracted from facial images, AGFN firstly projects them into emotional feature vectors. Then, under the assumption that the probability of sample connections satisfies Poisson distribution and corresponding parameters are closely related to samples' similarities, a Poisson graph generator is designed to adaptively construct topological graphs for samples in each mini-batches. Afterwards, Graph Convolutional Network (GCN) is exploited to convey the semantic information of associated samples because expressions present in facial images can be reflected by other images sharing similar features in a proper feature space. The combination of convolutional topology and Poisson generator solve the uncertainty of facial expression association and similarity in data. In addition, since the calculation of adjacent matrices used for graph generation involves a sampling process, parameters of our network cannot be optimized by the widely used gradient descent method. Therefore, we design a coordinate descent strategy to jointly optimize parameters of neural networks and the sampling process.

As shown in Fig. 2, after normalizing features with proposed AGFN, individual classes can be split with much clearer boundaries though mislabeled data still exists. According to our experiments, the recognition performance is improved by large margins from 85.37% to 91.84% and from 85.89% to 91.11% on the benchmark datasets FERPlus and RAF-DB when a naive FER model is equipped with proposed AGFN, implying the importance of tackling the data uncertainty problem in FER as well as the effectiveness of our AGFN. Moreover, we also conduct experiments on synthetic datasets where a large portion of samples (e.g., 20%) are mislabeled, finding that our AGFN-equipped network surpasses existing state-of-the-art works significantly by 14.09%, 21.12% and 13.67% on FERPlus, RAF-DB and AffectNet datasets, respectively. The main contributions of this paper are as follows:

(1) We propose leveraging the associative relationships between expressions to address the challenges posed by data uncertainties, namely the excursive semantic and feature covariate shifting problems. Additionally, we introduce a highly effective normalization method called AGFN to enhance the performance of FER models.
(2) To construct topological graphs for samples in each mini-batch, we have developed a Poisson graph generator that utilizes a sampling process. Additionally, we leverage GCN to normalize feature distributions. Furthermore, we employ a coordinate descent strategy to jointly optimize the parameters involved in both the neural networks and the sampling process.

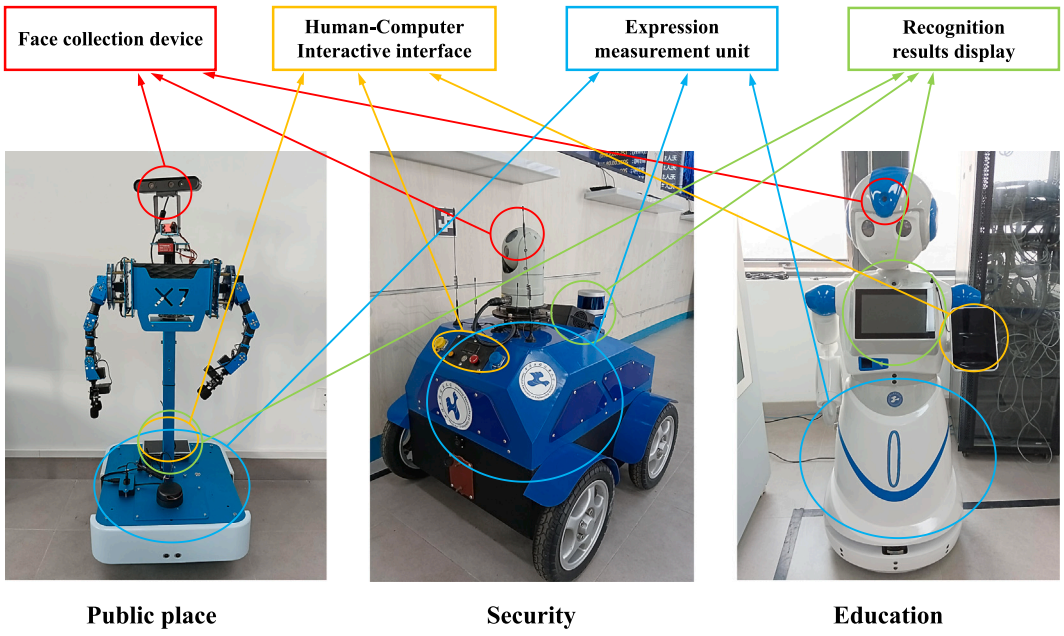

Fig. 3. Three types of facial expression recognition measurement devices.

(3) In comparison to state-of-the-art methods, our AGFN achieves superior or comparable performance on three widely-used benchmark datasets. Particularly, when a significant portion (e.g., 20%) of samples are mislabeled, our AGFN-equipped network exhibits significantly enhanced robustness and effectiveness.

2. Related works

2.1. Facial expression recognition

Feature extraction and expression classification are the two basic modules of typical FER pipeline, and in past years, most of related works have focused on designing more effective and robust feature extractors. For example, works (Ding et al., 2017; Wang et al., 2017; Li and Lima, 2021; Wu et al., 2021) explored various deep neural networks to extract more powerful features, including VGG network, Inception network, Residual network and Capsule network etc. However, Li et al. (2021) pointed out that deeper and wider network structures tended to increase storage and computing cost. Therefore, they proposed to learn more discriminative representations from another perspective by designing a more adaptive supervised objective named AdaReg loss, which re-weighted the importance coefficients of categories to tackle with the class imbalance problem and obtain more powerful representations. Acharya et al. (2018) held the view that the widely used convolutional layers and average pooling layers only captured first-order statistics, so they proposed to extract the second order statistic features with covariance pooling. Moreover, in practice, pose variations, occlusions and uneven illuminations always resulted in low quality facial images, on which FER models usually failed to extract discriminative features. Therefore, Wang et al. (2020b) designed a regional attention network to improve the performance of FER. Inspired by the psychological theory that expressions could be decomposed into multiple facial action units, Liu et al. (2015) constructed a deep network called AU-inspired Deep Networks (AUDN) to combine the informative local appearance variation and high-level representation. Generally, objective functions of FER networks considered each sample independently, while Zhao et al. (2016) designed a peak-piloted deep network (PPDN) to supervise the intermediate feature responses of hard samples, i.e., ones with non-peak expression, with easy samples, i.e., ones with peak expression.

It is a challenging task to accurately extract all the correlated handcrafted features due to the effect of variations caused by emotional state (Heidari et al., 2023). Researchers (Mohan et al., 2021b) proposed FER-NET: a convolution neural network to distinguish FEs efficiently with the help of the softmax classifier. FLEPNet (Karnati et al., 2022a), a texture-based feature-level ensemble parallel network, was also proposed to solve the FER problem. The parallel network FLEPNet uses multi-scale convolutional and multi-scale residual block-based DCNN as building blocks. It is pointed out that, the modern FER systems based on deep neural networks mainly suffer from two problems: overfitting due to the inadequate availability of training data and complications unassociated with the expressions, such as occlusion, posture, illumination, and identity bias (Karnati et al., 2023). Karnati et al. (2022b) presented a deep convolution neural network (DCNN) named LieNet to precisely detect the multiscale variations of deception automatically. Mohan et al. (2021a) proposed a two-stage approach for FER. The former one finds out local features from face images using a local gravitational force descriptor, while, in the latter part, the descriptor is fed into a deep convolution neural network.

Recently, Transformer (Vaswani et al., 2017) showed its power in various neural language processing and computer vision tasks, including FER. For instance, Ma et al. (2021) proposed visual Transformers with feature fusion to translate facial images into sequences of visual words and perform expression recognition from a global perspective. Huang et al. (2021) also utilized two attention mechanisms to conduct low-level feature learning and obtain high-level semantic representation. Specifically, a grid-wise attention mechanism was used to capture the dependencies of different facial image regions and a visual transformer attention mechanism was exploited to learn global representation from a sequence of visual semantic tokens.

In summary, in the era of deep learning, FER performance has been significantly improved by learning more discriminative representations with more powerful neural networks, i.e., networks with attention mechanisms or equipped with Transformers. However, existing works omitted the data uncertainty problem, and as shown in Fig. 2, classifiers failed to distinguish samples well with features learned from noisy data.

2.2. GCN-based FER models

Geometric information was essential to FER representation learning, and Graph Convolutional Network (GCN) was effective to capture

geometric dependencies. Therefore, GCN was widely used in FER to model the geometrical relations of key landmarks on facial images and learn more representative features. For example, Zhao et al. (2021) proposed GA-FER to encode facial landmarks and explore the structural information of facial components. The geometric and appearance knowledge were combined by a GCN equipped with multiple blocks and attention mechanisms so that more comprehensive high-semantic representation and global characteristics of expressions were obtained. Ruan et al. (2021) viewed the expression information as the combination of the shared information and the unique information, so they designed a Feature Decomposition and Reconstruction (FDRL) FER model, where topological graphs were built upon a set of latent features that decomposed from basic features, and GCN was utilized to learn discriminative expression features from graphs. Liu et al. (2020a) detected facial action units with GCN that took latent representation vectors learned by an auto-encoder as input. Nian et al. (2019) performed unified facial attribute recognition by decoupling features. In this work, label dependencies were captured by GCN to handle the correlations between facial attributes. Liu et al. (2020b) utilized an improved GCN to handle facial images with low resolution or partial occlusion. Liu et al. (2020c) utilized GCN to learn significant facial expression features that concentrated on certain regions after extracting features with CNN. Lo et al. (2020) applied GCN to discover the dependency of action unit nodes for micro-expression categorization. Xie et al. (2020a) designed an AU-assisted Graph Attention Convolutional Network (AU-GACN) to extract discriminative features for subtle micro-expressions by fully exploiting the relation between action units.

As we can see, GCN variants were extensively explored in FER. However, existing FER models mainly utilized GCN to capture geometric information of different regions or landmarks on facial images so that more powerful representations were obtained. Moreover, topological graphs used in existing works were usually constructed with staircase functions. By contrast, in our work, GCN is employed to normalize features adaptively, and our graph is built upon samples from each mini-batches with a dynamic graph generator, who models the relation between sample connection probabilities and feature similarities with Poisson distribution.

2.3. Measurements and applications

This chapter is used to introduce the measurement devices and application scenarios of FER.

Typically, FER systems employ mobile robots for data collection. This approach leverages the mobility of robots to efficiently gather and transmit data, providing reliable support for the system's operation. As shown in Fig. 3, three different types of robots used in various fields are displayed, with each functional component labeled accordingly. The facial acquisition device as the visual input component of the system, capturing and recording face image data in real time. Users can interact with the robots through the human–computer interaction interface, while this interface also provides feedback from the robots. The robots collect, analyze and feedback facial expression data in real time, thereby providing valuable applications in various fields to enhance human–computer interaction and deepen the understanding of emotion recognition.

FER has many applications in practical scenarios. Four different scenarios including education, public crowds, healthcare and indoor security (Wang et al., 2022; Zhong, 2023) are shown in Fig. 4. In the field of education, FER is used to monitor students' reactions during online classes or training sessions, helping educators make adjustments to their teaching methods. In dense crowd scenarios, FER help in identifying the emotional states of individuals in a crowd. This is useful for understanding the mood of the crowd, identifying potential threats, and ensuring public safety in large gatherings, protests, or events. In the field of healthcare, FER is valuable for both mental health assessment, aiding in the diagnosis and treatment planning of conditions like depression and anxiety, and for pain assessment in clinical settings. This technology is especially beneficial for patients who may have difficulty communicating their pain levels, such as infants or the elderly, as it allows healthcare providers to effectively evaluate pain levels and adjust treatment plans accordingly. In the field of security and surveillance, detect threats and prevent crime by assessing emotions and abnormal behavior, identifying individuals of interest in crowded spaces, ensuring access control, and assisting in missing person searches. Additionally, it plays a vital role in border control, financial transactions, and search and rescue operations.

2.4. Uncertainties in facial expression recognition

Uncertainties result in mislabeled data, which seriously affects the performance of FER models. Though learning with noisy labels has attracted extensive attentions in the community of computer vision, it is rarely studied in the field of FER. On the other hand, existing works usually address this issue by pre-training networks on weak data and then fine-tuning them with true labels, guiding the training of networks with knowledge learned from clean data, or relabeling mislabeled data together with learning more powerful representations and classifiers (Veit et al., 2017; Xie et al., 2020b; Wang et al., 2022),. For example, to alleviate the harm from noisy data, Li et al. (2017b) designed a unified distillation framework, where the distillation process was guided with a knowledge graph, to 'hedge the risk' of learning from noisy labels. Apparently, above works tried to estimate label quality or noisy distribution with a small set of clean data, while works without utilizing clean data usually introduced additional constrains or distributions on noisy data. For example, Mnih and Hinton (2012) proposed more robust loss functions to deal with omission noise and registration noise on aerial image datasets. For the task of FER, Zeng et al. (2018) was the first to improve FER performance by addressing the data uncertainty issue. They assigned more than one labels to each samples and discovered the latent truth from the inconsistent pseudo labels with an end-to-end LTNet. Afterwards, Wang et al. (2020a) proposed to suppress data uncertainties by weighting and ranking samples in a mini-batch with a self-attention mechanism, followed by modifying samples' labels in the lowest-ranked group. Also others provided a novel database for natural facial expression to construct leveraging the social images and then trained a deep model based on the naturalistic dataset (Peng et al., 2016). An amount of social labeled images are obtained from the image search engines by using specific keywords. And some devote their researches how to leverage noisy data in the web to boost the FER performance. They proposed model is implemented in an end-to-end weakly supervised manner and enjoys several merits (Zhang et al., 2021a). It utilizes massive noisy labeled data to boost the performance of the FER classifier trained on a small set of clean labels.

Though previous works have explored how to discover the truth of mislabeled data and prevent networks from the harm of noisy data, they neglect the associative relations of expressions. Therefore, in this work, we propose to protect FER models from data uncertainties by tackling the excursive semantic and feature covariate shifting problems with the association of expressions. Our experiments demonstrate the necessity of handling data uncertainties in FER and the effectiveness of our proposed strategy.

3. Our method

Distinguishing face expressions is challenging because of the ambiguity of facial images and subjectiveness of annotators, which result in mislabeled data, and further cause excursive semantic and feature covariate shifting. In this work, we propose to learning from noisy FER data from the perspective of associative learning. Intuitively, facial images containing similar subtle expressions are most likely to share the same labels and according to human associative learning

Fig. 4. Application of facial expression recognition in different scenarios.

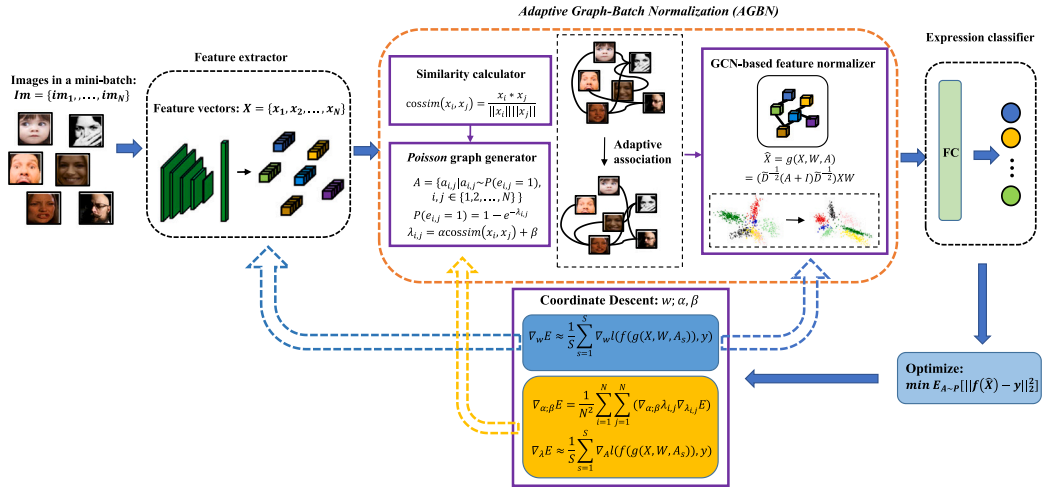

Fig. 5. Architecture of proposed AGFN-equipped network. The AGFN module is composed of similarity calculator, Poisson Graph generator and GCN-based feature normalizer, and can be conveniently inserted into FER models between their feature extractors and expression classifiers.

mechanism (Anderson and Bower, 2014), humans tend to correlate objects with similar abstract features. Therefore, exchanging semantic information among samples with high similarity can help to normalize features of individual samples, leading to more discriminative feature representations. Toward this end, we design a feature normalization method named Adaptive Graph-based Feature Normalization (AGFN). In general, a FER model consists of two based components, i.e., a feature extractor that extracts features from facial images and a expression classifier that distinguishes corresponding expressions. Our AGFN can be conveniently inserted into any FER models between these two components, as shown in Fig. 5. Specifically, AGFN exploits a novel graph generator to dynamically and adaptively construct topological graphs for samples within each mini-batches according to their similarities. In this generator, adjacent matrices of topological graphs are determined by a sampling process. Then, GCN is used to convey semantic information among associated samples. Since gradient calculation rules of parameters from neural networks and above sampling process are different, traditional gradient descent method is not applicable anymore. Therefore, we propose a new coordinate descent strategy to optimize our network in an end-to-end way. More details are given below.

3.1. Poisson graph generator

Traditional graph-based methods usually connect samples with high similarities with the widely used threshold-based staircase function. However, samples with very similar features may not belong to the same classes and high similarities only imply high probabilities of sharing the same class labels. Especially in the task of FER, expressions have serious ambiguities and samples belong to different emotion categories may seem very similar. Therefore, if the topological graph is built with the staircase function, hard samples with higher similarities but belonging to different classes will always been connected, misleading models to learn bad representations.

To address this issue, we propose to model the relations between feature similarities and sample connection probabilities with Poisson distribution. In statistics, a Poisson distribution is a discrete probability distribution that is used to describe how many times an event is likely to

occur over a specified period, as shown in Eq. (1), where k and λ denote times and average times that an event happens per unit time. In human associative learning, two samples are usually compared for multiple times to confirm whether they belong to the same classes, and different regions of interest are usually looked every time. Obviously, the higher similarity two samples have, the more times they are compared, and the higher probability they are connected in the topological graph. Therefore, we assume that probabilities of sample connections satisfy Poisson distribution and corresponding parameters are closely related to samples' similarities. Subsequently, a novel Poisson graph generator is proposed to calculate the adjacent matrices of topological graphs with a sampling process.

$$Po(k;\lambda) = \frac{\lambda^k}{k!}e^{-\lambda}, k = 0, 1, \dots, K \tag{1}$$

For simplicity, we build topological graph with samples from the same mini-batch, so for a batch size setting of N, there are N input images $Im = \{im_1, im_2, \dots, im_N\}$, which are also N nodes of our target graph. Assuming corresponding feature maps extracted by CNN-based feature extractors are $M = \{m_1, m_2, \dots, m_N\}$, we further project them into emotional feature vectors $X = \{x_1, x_2, \dots, x_N\}$ with a Multi-Layer Perceptron (MLP). Thus, the representation of ith node in the graph is x_i.

The key of GCN-based models is the construction of adjacent matrix. To achieve this, we firstly calculate similarities of samples with cosine similarity coefficient, as formulated in Eq. (2), where $cossim(x_i, x_j)$ represents the similarity between sample im_i and im_j.

$$cossim(x_i, x_j) = \frac{x_i * x_j}{\|x_i\|\|x_j\|} \tag{2}$$

Intuitively, we associate an object to different ones for multiple times to capture more detailed information. Therefore, the connection probability $p(e_{i,j} = 1)$ of sample im_i and sample im_j can be computed with Eq. (3), meaning that samples should be connected if they need to be compared for $k \geq 1$ times.

$$P(e_{i,j} = 1) = \sum_{k=1}^{\infty} Po(k;\lambda) \tag{3}$$

In other words, samples should not be connected if they do not need to be compared, so we further rewrite Eq. (3) to Eq. (4), where the Poisson parameter $\lambda_{i,j}$ is computed with the similarity $cossim(x_i, x_j)$ via a linear function described in Eq. (5). Here, parameters α and β are introduced to scale the probability distribution, and are learned during the training procedure. The learnable scale parameter α and β in matrix A that represent the generalization features of facial expressions can deal with the uncertainty preferably.

$$\begin{aligned} P(e_{i,j} = 1) &= \sum_{k=1}^{\infty} Po(k;\lambda) = 1 - Po(0;\lambda_{i,j}) \\ &= 1 - \frac{e^{-\lambda_{i,j}}\lambda_{i,j}^0}{0!} = 1 - e^{-\lambda_{i,j}} \end{aligned} \tag{4}$$

$$\lambda_{i,j} = \alpha cossim(x_i, x_j) + \beta \tag{5}$$

Afterwards, we sample the adjacent matrix A according to $A \sim P$ (see Eq. (6)) for samples in current mini-batch. Expectations of the sampling process will be optimized as introduced in Section 3.3.

$$A = \{a_{i,j} | a_{i,j} \sim P(e_{i,j} = 1), i, j \in \{1, 2, \dots, N\}\} \tag{6}$$

Another advantage of constructing topological graphs with stochastic mechanism is that different contrastive objects can be seen in different iterations, which benefits models' robustness. The mechanism is similar to the well-known DropConnect strategy (Wan et al., 2013), who randomly drops neurons' connections of neural networks to alleviate the over-fitting problem.

3.2. Feature normalization with GCN

To tackle with the excursive semantic and feature covariate shifting problems, we imitate human associative learning procedure by conveying semantic information among associated samples with GCN, which is built upon samples' topological graphs generated by our Poisson graph generator. The employed GCN is in the second order Chebyshev expansion, as formulated in Eq. (7), where A is the adjacent matrix obtained by above sampling strategy, W is trainable parameters, I is an identity matrix, $\widetilde{D}$ is a diagonal matrix, X denotes samples' emotional feature vectors and $\hat{X}$ represents the expected normalized features. Here, $A + I$ means a self-loop edge is added to the graph and $\widetilde{D}^{\frac{-1}{2}}$ is a degree matrix used to weight information from associated samples. For ith sample, more associated samples result in greater value of $\widetilde{D}_{ii}$, which means less information from associated samples will be passed to current sample.

$$\begin{aligned} \hat{X} &= g(X, W, A) = (\widetilde{D}^{\frac{-1}{2}}(A + I)\widetilde{D}^{\frac{-1}{2}})XW \\ \widetilde{D}_{ii} &= 1 + \sum_j A_{i,j} \end{aligned} \tag{7}$$

Note that, GCN has been widely used in existing FER models, but it is usually employed to capture geometric information, which is different from ours. Here we discuss the time complexity of the GCN. We designed a pairwise matrix element to calculate cosine similarity, due to the use of one multiplication and division for each two elements, the calculation amount is 2. Through α and β scaling, the calculation amount for one multiplication and addition is 2. Then, through one Poisson distribution sampling: one subtraction and exponential operation, the calculation amount is 2. So the calculation amount for this process is 6. Therefore, the amount of calculation required to construct matrix A is $6N \cdot 6N = 36N^2$. Through normalization, diagonal matrix D is calculated, and the calculation complexity is $N^2 + N$; The computational complexity of $A + I$ is N^2; When calculating the square root of the D matrix, the computational complexity is N, and when calculating X, there are four times of matrix multiplication. So we get the final computational complexity is $O(N^3)$.

3.3. Optimization with coordinate descent

Suppose our loss function is defined as Eq. (8), where $f(.)$ denotes the expression classifier, then our final goal is twofold: (1) optimizing parameters W involved in neural networks, including feature extractor, GCN and expression classifier; and (2) learning parameters α and β used to find the best adjacent matrix $A \in \mathcal{H}_N$ in Eqs. (4) and (5). Here, $\mathcal{H}_N$ is the convex hull of the set of all possible adjacency matrices under the Poisson distribution. Furthermore, the objective of our network is to minimize the expectation formulated in Eq. (9).

$$\ell(f(\hat{X}), y) = \|f(\hat{X}) - y\|_2^2 \tag{8}$$

$$J = \min_{W,\alpha,\beta} E_{A\sim P}[\ell(f(\hat{X}), y)] \tag{9}$$

Since the gradient calculation of α and β is different from that of W, and the function used to calculate $\nabla_\alpha E$ and $\nabla_\beta E$ are not differentiable, traditional gradient descent strategy is not applicable for our network optimization. Therefore, under the assumption that parameters to be optimized are independent from each other, we design a new coordinate descent strategy to train our network in an end-to-end way. Concretely, we update W with the tractable approximate learning dynamics shown in Eq. (10), and obtain the approximate gradient $\nabla_W E$ with Eq. (11), where $P(A)$ is the probability of sampling A from distribution P with Eq. (6), S represents the pre-defined sampling times and A_s denotes the result of the sth sampling.

$$\hat{W} = W - \gamma_1 \nabla_W E \tag{10}$$

$$\begin{aligned}\nabla_W E &= \nabla_W E_{A\sim P}[\ell(f(\hat{X}), y)] \\ &= \sum P(A)\nabla_W \ell(f(g(X, W, A)), y) \\ &\approx \frac{1}{S}\sum_{s=1}^{S} \nabla_W \ell(f(g(X, W, A_s)), y)\end{aligned} \tag{11}$$

On the other hand, we update α and β with Eq. (12), where $\nabla_\alpha \lambda_{i,j} = cossim(x_i, x_j)$, $\nabla_\beta \lambda_{i,j} = 1$ and $\nabla_\lambda E$ is obtained with an estimator (see Eqs. (14)~(16)).

$$\begin{aligned}\hat{\alpha} &= \alpha - \gamma_2 \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\nabla_\alpha \lambda_{i,j}\nabla_{\lambda_{i,j}} E) \\ \hat{\beta} &= \beta - \gamma_2 \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\nabla_\beta \lambda_{i,j}\nabla_{\lambda_{i,j}} E)\end{aligned} \tag{12}$$

According to Mohamed et al. (2020), continuous distributions have a simulation property that samples can be drawn from them in both direct and indirect ways, and for the general case $\hat{x} \sim p_{(x;\theta)}$, we can draw a sample $\hat{x}$ in an indirect way by firstly sampling $\hat{\epsilon}$ from a simple base distribution $p_{(\epsilon)}$, which is independent of the parameters θ, and then transforming $\hat{\epsilon}$ to $\hat{x}$ through a sampling path $sp(\epsilon;\theta)$. This procedure (or sampling process) can be formulated as Eq. (13).

$$\hat{x} \sim p_{(x;\theta)} \quad \equiv \quad \hat{x} = sp(\hat{\epsilon};\theta), \quad \hat{\epsilon} \sim p_{(\epsilon)} \tag{13}$$

Afterwards, according to the *Law of the Unconscious Statistician* (LOTUS) (Grimmett and Stirzaker, 2001), even though the distribution of a function with variable x is unknown, we still can calculate this function's expectation with corresponding sampling path and base distribution, as shown in Eq. (14). This is a widely used way to reparametrize a probabilistic system.

$$E_{p(x;\theta)}[f(x)] = E_{p(\epsilon)}[f(sp(\epsilon;\theta))], \tag{14}$$

Therefore, a path-wise estimator for gradient $\nabla_\theta E_{p(x;\theta)}[f(x)]$ can be calculated with Eq. (15).

$$\begin{aligned}\nabla_\theta E_{p(x;\theta)}[f(x)] &= \nabla_\theta \int p(\epsilon) f(sp(\epsilon;\theta))\mathrm{d}\epsilon \\ &= \int p(\epsilon)\nabla_x f(x)|_{x=sp(\epsilon;\theta)} \nabla_\theta sp(\epsilon;\theta)\mathrm{d}\epsilon \\ &= E_{p(x;\theta)}[\nabla_x f(x)\nabla_\theta x]\end{aligned} \tag{15}$$

However, in our case, the distribution $A \sim P$ is discontinuous because elements of A are binarized, so we approximately estimate $\nabla_\lambda E$ with an inexact but smooth reparameterization of $A \sim P$ (see Eq. (16)). Specifically, we employ the identity mapping $A = sp(\epsilon;\lambda) = 1 - Po(0;\lambda)$ of straight-through estimators (STE) (Bengio et al., 2013), and accordingly, get $|\nabla_\lambda A| = |\nabla_\lambda Po(0;\lambda)| \approx I$.

$$\begin{aligned}\nabla_\lambda E &= \nabla_\lambda E_{A\sim P}[\ell(f(\hat{X}), y)] \\ &= E_{A\sim P}[\nabla_\lambda A \nabla_A \ell(f(g(X, W, A)), y)] \\ &\approx \frac{1}{S}\sum_{s=1}^{S} \nabla_A \ell(f(g(X, W, A_s)), y)\end{aligned} \tag{16}$$

The proposed algorithm is summarized in Algorithm 1. When calculating the gradient descent algorithm, the matrix subtraction complexity is N^2, the coefficient multiplication complexity is $2 * N^2$, where $\nabla_\alpha \lambda_{i,j}$ and $\nabla_\beta \lambda_{i,j}$ is N^2, and $\nabla_W E$ complexity is $S * N * N = SN^2$. Therefore, the final calculation complexity for α, β is $O\left(N^4\right)$. Similarly, the computational complexity of W includes matrix subtraction N^2, coefficient multiplication N^2 and $\nabla_W E$ computation SN^2. So the final complexity of the proposed algorithm is $O\left(N^4\right)$.

4. Experiments

In this section, we provide details of our implementation and conduct extensive experiments to prove the effectiveness and robustness of our AGFN on datasets with uncertainties.

Algorithm 1 The AGFN gradient descent algorithm.

Input: γ_1, γ_2Learning rate for W, α, β learnable parameters for gradient updating;
$f(g(X, W, A_s)), y$: The objective optimization function and labeled values;
l : The loss function for the AGFN;
Output: W, α, β
while W, α, β not converged **do**
$W \leftarrow W - \gamma_1 \nabla_W E$
$\alpha \leftarrow \alpha - \gamma_2 \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\nabla_\alpha \lambda_{i,j}\nabla_{\lambda_{i,j}} E\right)$
$\beta \leftarrow \beta - \gamma_2 \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\nabla_\beta \lambda_{i,j}\nabla_{\lambda_{i,j}} E\right)$
The parameter $\nabla_W E, \nabla_\alpha \lambda_{i,j}, \nabla_\beta \lambda_{i,j}, \nabla_{\lambda_{i,j}}$ is updated by following equations:
$\nabla_W E \approx \frac{1}{S}\sum_{s=1}^{S} \nabla_W l\left(f\left(g\left(X, W, A_s\right)\right), y\right)$
$\nabla_\alpha \lambda_{i,j} = \mathrm{cossim}\left(x_i, x_j\right), \nabla_\beta \lambda_{i,j} = 1$
$\nabla_{\lambda_{i,j}} E \approx \frac{1}{S}\sum_{s=1}^{S} \nabla_A l\left(f\left(g\left(X, W, A_s\right)\right), y\right)$
end while

4.1. Datasets and implementation

RAF-DB (Li et al., 2017a) contains 12,271 training images and 3068 test images collected from thousands of individuals. In our experiments, only images belonging to the 7 basic expressions (i.e., neutral, happiness, surprise, sadness, anger, disgust and fear) are used.

FERPlus (Barsoum et al., 2016) consists of 28,000 training images and 3000 test images collected by Google search engine. Compared with RAF-DB, FERPlus includes an extra expression, i.e., contempt, resulting in 8 expression classes. In addition, samples in FERPlus are labeled by 10 annotators, and in this work, we select labels with the most votes as the ground truth of each samples.

AffectNet (Mollahosseini et al., 2017) is a large-scale dataset created from the internet with three search engines and 1250 emotion related tags in six languages. It consists of more than 1,000,000 facial images in the wild and the same as FERPlus, samples in AffectNet are classified into 8 expression classes.

In our implementation, we embed the proposed AGFN into a naive FER baseline model, who employs ResNet-18 as its feature extractor and a fully-connected layer as its expression classifier. The projected emotional feature vectors are with a dimension of 512 and the batch size is set to 256. Moreover, we set the learning rates γ_1 and γ_2 (used in Eqs. (10) and (12)) to 0.01 and 0.001, respectively, and in order to speed up the training procedure, we pre-train the baseline model for about 10 epochs.

4.2. Comparison with state-of-the-art FER models

In this part, we compare the proposed AGFN-equipped network with state-of-the-art FER models to demonstrate the effectiveness of our AGFN and the importance of dealing with data uncertainties in FER. Generally, existing FER models can be grouped into three categories, i.e., traditional CNN-based ones, typical GCN-based ones and the most advanced Transformer-based ones. Especially, Transformer (Vaswani et al., 2017) has largely fueled the performance of various computer vision tasks including FER, and the Transformer-based solutions achieve the most promising performance at present. For example, FER-VT (Huang et al., 2021) exploits grid-wise attention and visual Transformer to learn long-range inductive biases between different facial regions, and TransFER (Xue et al., 2021) learns rich relation-aware local representations with Transformer. From Table 1,

Table 1
Comparison with existing works on RAF-DB and FERPlus datasets.

Types	Method	FERPlus	RAF-DB	AffectNet$_8$	AffectNet$_7$
CNN-based	IPA2LT (Zeng et al., 2018)	–	86.77	–	55.71
	SCN (Wang et al., 2020a)	89.35	88.14	60.23	–
	KTN (Li et al., 2021)	90.49	88.07	–	63.97
	RAN (Wang et al., 2020b)	88.55	86.90	52.97	–
	DLP-CNN (Li et al., 2017a)	–	84.22	–	–
	SeNet (Albanie et al., 2018)	88.8	–	–	–
	gACNN (Li et al., 2019)	–	85.07	–	58.78
	DACL (Farzaneh and Qi, 2021)	–	87.78	–	65.20
	FLEPNet (Mohan et al., 2021b)	–	87.56	–	
	FER-NET (Karnati et al., 2022a)	–	82	–	
	FG-AGR (Li et al., 2023)	–	90.81	–	
Transformer-based	FER-VT (Huang et al., 2021)	90.04	88.26	–	–
	TransFER (Xue et al., 2021)	90.83	90.91	–	**66.23**
GCN-based	GA-FER (Zhao et al., 2021)	–	87.52	–	–
	FDRL (Ruan et al., 2021)	–	89.47	–	–
Ours	**Baseline**	85.37	85.89	–	–
	Baseline-GCN	88.8	88.95	–	–
	Baseline-AGFN	**91.84**	**91.11**	**61.38**	66.09

both FER-VT and TransFER surpass other existing works significantly. Even though, our AGFN-based network still outperforms FER-VT by 1.8% and 2.85% on FERPlus and RAF-DB datasets respectively, and the accuracies are 1% and 0.2% higher than that of TransFER on both datasets, indicating the superiority of proposed AGFN-equipped network. Moreover, from the comparison with our baseline model we can see that when integrating the proposed AGFN module, recognition accuracies on FERPlus and RAD-DB are significantly elevated from 85.37% to 91.84% and from 85.89% to 91.11%, respectively, which further demonstrates the effectiveness of our graph-based feature normalization strategy.

Among in-the-wild FER datasets, AffectNet is the most affected by the uncertainty samples, and the scale of AffectNet is much larger than that of the widely used FERPlus and RAF-DB datasets. Therefore, we also conducted experiments on AffectNet. Note that, AffectNet has 8 emotion categories as in FERPlus, but works like TransFER (Xue et al., 2021) and KTN (Li et al., 2021) test their models on only 7 of them, where the class of "contempt" is excluded. In contrast, works like SCN (Wang et al., 2020a) and RAN (Wang et al., 2020b) conduct experiments on all of the 8 emotion categories. From our statistical analysis in Fig. 1, the percentage of low-confidence samples in "Contempt" class is much higher than that in other classes like "Happy", "Surprise" and "Sad", which means the "Contempt" class introduces more data uncertainties and makes AffectNet dataset more challenging. Therefore, for a fair comparison on AffectNet, we conduct two groups of experiments, denoted as AffectNet$_8$ and AffectNet$_7$ in Table 1, to compare the performance of our network with that of existing works separately. As we can see, our AGFN-equipped network achieves comparable performance with TransFER in the case of AffectNet$_7$ and outperforms all of other existing works in both AffectNet$_8$ and AffectNet$_7$ cases, proving that our AGFN also works well on more challenging datasets with larger scales.

On the other hand, IPA2LT (Zeng et al., 2018) and SCN (Wang et al., 2020a) also improves FER performance by tackling with noisy labels and the data uncertainty problem. Specifically, IPA2LT trains a FER model from multiple inconsistently labeled datasets and large scale unlabeled data with a scheme that discovers latent truth from inconsistent pseudo labels, while SCN relabels mislabeled data by weighting and ranking samples in a mini-batch with a self-attention mechanism. In contrast, our AGFN protects FER models from data uncertainties by alleviating the excursive semantic and feature covariate shifting problems with associative relations of expressions. From Table 1, our network achievers better performance than both PA2LT and SCN. Especially, it surpasses SCN (Wang et al., 2020a) by 2.49%, 2.91% and 1.15% on FERPlus, RAF-DB and AddectNet datasets, respectively, indicating that AGFN is with more advantages than SCN and IPA2LT when dealing with the data uncertainty problem.

4.3. Comparison with GCN-based FER models

GCN is widely used in FER to capture geometric information, which is further combined with appearance information to learn more powerful and discriminative representations for facial images. Our network also utilizes GCN, but there is significant difference between our work and existing GCN-based methods. Firstly, GCN is used to convey semantic information among samples from the same mini-batch, rather than capturing geometric information from single facial images. Secondly, topological graphs of existing GCN-based models are built upon different regions, landmarks or features of single facial images, while nodes of our graph represent different samples from the same mini-batch. Moreover, we design a new sampling strategy named Poisson graph generator to build topological graphs dynamically. Here, Poisson distribution is utilized to model the relation of sample connection probabilities and feature similarities. By contrast, existing GCN-based models build their graphs with the widely used staircase function or according to spatial distances. Additionally, to optimize parameters from the sampling process and neural networks in an end-to-end way, our network is trained with a new coordinate descent strategy, instead of gradient descent strategy that widely used in other GCN-based methods.

Among existing GCN-based FER models, GA-FER (Zhao et al., 2021) and FDRL (Ruan et al., 2021) achieve the most promising performances. Specifically, GA-FER (Zhao et al., 2021) builds graphs upon landmarks of single face images, and FDRL (Ruan et al., 2021) constructs graphs with latent features obtained by decomposing basic features of face images. In addition, GA-FER (Zhao et al., 2021) utilizes a GCN variant equipped with complicated attention mechanism, while FDRL (Ruan et al., 2021) employs the naive version of GCN. We also design a GCN-based baseline model named Baseline-GCN by integrating GCN into our baseline model with the traditional staircase function, i.e., connecting samples with similarities greater than 0.5 directly to build topological graphs. As shown in Table 1, our network outperforms GA-FER, FDRL and Baseline-GCN significantly by 3.95%, 1.64% and 2.61%, respectively on RAF-DB dataset, indicating the superiority of network over other existing GCN-based FER models.

4.4. Performance of AGFN on datasets with serious uncertainties

With the growth in scale of training samples gathered from internet, the problem of data uncertainty is getting more and more severe. To further explore the effectiveness of AGFN on datasets with serious uncertainties, we conduct extra experiments on our synthetic datasets. Specifically, we randomly select 10% or 20% samples from FERPlus, RAF-DB and AffectNet datasets, and assign wrong labels to them.

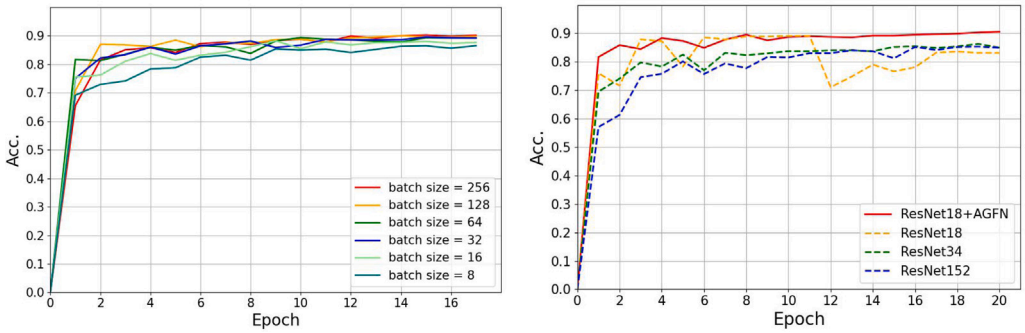

Fig. 6. Performance of our network with different batch size (left) and comparison of networks with AGFN and deeper backbones (right) on RAF-DB dataset.

Table 2
Comparison results on synthetic datasets with noise ratios of 10% and 20%. Models are all trained from scratch and 8 emotion categories are evaluated on AffectNet dataset.

Method	FERPlus		RAF-DB		AffectNet	
	10%	20%	10%	20%	10%	20%
CurriculumNet (Guo et al., 2018)	–	–	68.50	61.23	–	–
MetaCleaner (Guo et al., 2018)	–	–	68.45	61.35	–	–
SCN (Wang et al., 2020a)	78.53	72.46	70.26	63.50	45.23	41.63
Ours	**87.03**	**86.55**	**87.02**	**84.62**	**59.12**	**55.30**

Most of the image noise conforms to normal distribution, so we have added examples of normal distribution image noise generation on the basis of the above. The comparison results are shown in Table 2, where the proposed network is compared with SCN (Wang et al., 2020a) and other two state-of-the-art noise-tolerant methods, i.e., CurriculumNet (Guo et al., 2018) and MetaCleaner (Zhang et al., 2019). CurriculumNet (Guo et al., 2018) handles massive amounts of noisy labels and data imbalance on large-scale web images by leveraging curriculum learning, which measures and ranks the complexity of data in an unsupervised manner, while MetaCleaner (Zhang et al., 2019) learns to a clean representation for an object category according to a small noisy subset from the same category.

From Table 2, the recognition accuracy of our network is obviously higher than that of other three methods. Especially, when the noise ratio is 10%, our network outperforms SCN by 8.5%, 16.76% and 13.89% on FERPlus, RAF-DB and AffectNet, respectively, while the improvements are 14.09%, 21.12% and 13.67% on above datasets for the noise ratio of 20%. Therefore, our AGFN-equipped network is with better effectiveness and robustness than other works when the data becomes more noisy. Note that, among all of the three datasets, our method achieves the most improvements on RAF-DB. This may be explained by the fact that RAF-DB is annotated by 40 people with crowdsourcing, while FERPlus and AffectNet are labeled by experts. Therefore, noise introduced to RAF-DB is much more than that introduced to FERPlus and AffectNet. This further demonstrates the effectiveness of our method in dealing with noisy data.

4.5. Performance on datasets with occlusion

In practice, occlusion frequently happens and results in data uncertainties. Therefore, following RAN (Wang et al., 2020b), we also conduct additional experiments on Occlusion-FERPlus and Occlusion-RAF-DB, which are generated from FERPlus and RAF-DB by Wang et al. (2020b), to evaluate the performance of our network. The experimental results are listed in Table 3. Here, RAN (Wang et al., 2020b) adaptively captures the importance of facial regions for occlusion FER, and CVT (Ma et al., 2021) translates facial images into sequences of visual words and performs expression recognition from a global perspective with convolutional visual Transformers. As shown in Table 3, our network outperforms all of the listed methods on datasets with occlusion. We owe the success to gathering semantic information from samples with high similarities, which alleviates information missing caused by occlusion.

Table 3
Performance on datasets with occlusion.

Method	Occlusion-FERPlus	Occlusion-RAF-DB
RAN (Wang et al., 2020b)	83.63	82.72
CVT (Ma et al., 2021)	84.79	83.95
FER-VT (Huang et al., 2021)	85.24	84.32
Ours	**85.95**	**86.53**

4.6. Parameter sensitivity analysis

In this part, we analyze the parameter sensitivity of our AGFN-equipped network in terms of batch size and backbone depth. Firstly, since our Poisson graph generator constructs topological graphs for samples within mini-batches, the setting of batch size is important. Therefore, as shown in Fig. 6, we evaluate the effect of different batch size settings to the performance of proposed network. As we can see, when the batch size is set to greater than 16, the performance is barely changed, and for batch size less than 16, the drop of performance is also in a reasonable range, so our proposed network is with strong robustness.

On the other hand, our AGFN module introduces additional parameters to the baseline network. It is a common sense that increasing the scale of deep networks is an effective way to enhance model's ability. Therefore, to prove that the performance improvement of our network is contributed by the strategy of utilizing associative relations of expressions rather than increasing model's complexity, we conduct extra experiments in Fig. 6, where accuracies of baseline models with ResNet18, ResNet34 and ResNet152 backbones are 85.89%, 85.95% and 86.02%, respectively. However, when equipped with proposed AGFN, the accuracy of baseline model with ResNet18 is elevated from 85.89% to 91.11%, which is 5.16% and 5.09% higher than that of

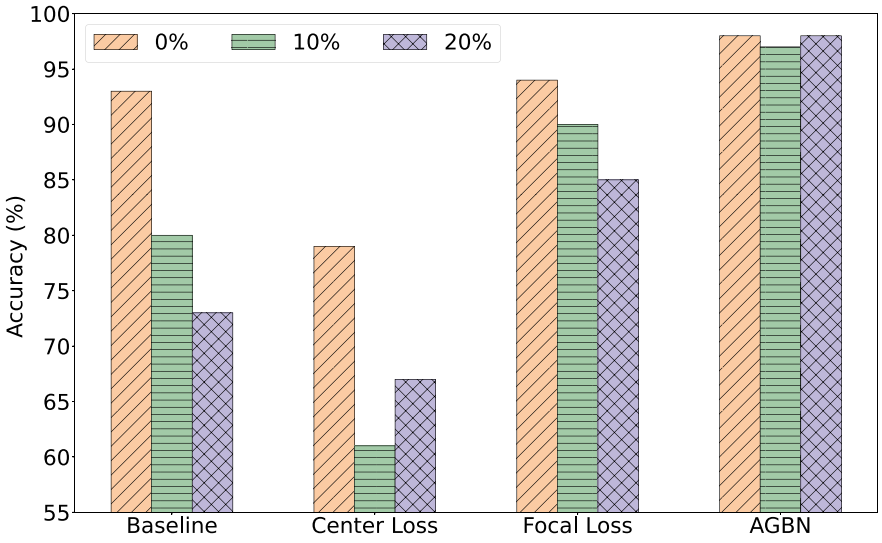

Fig. 7. Comparison of different methods with mislabeled samples.

baseline models with ResNet34 and ResNet152. Therefore, we can conclude that integrating AGFN to baseline model is more effective than simply increasing network's depth.

4.7. Generalizability of AGFN

To evaluate the generalizability of proposed AGFN on dealing with data uncertainties, we conduct experiments on the well-known MNIST dataset. Here, we randomly select 10,000 samples from the training set of MNIST and assign wrong labels to (0%, 10%, 20%) of them. These samples form our training set and the original test set of MNIST is kept as our test set. As shown in Fig. 7, we can see that our AGFN is able to effectively decrease the intra-class variance and, meanwhile, increase the inter-class variance, resulting in a better recognition accuracy, which is 20% higher than that of the baseline. Nearly all models are affected by mislabeled data, with baseline and Focal Loss exhibiting linear changes in response. However, AGFN pays greater attention to the interconnections among the data itself, thus not being influenced by mislabeled individual samples. As evident from the results, our approach maintains a stable performance, consistently achieving around 97% recognition rate. On the other hand, Center Loss (Wen et al., 2016) also exhibits similar characteristics. It also utilizes relations of samples to enhance the discriminative power of extracted features. However, different from our AGFN, Center Loss is a supervision signal who builds connections among samples according to their labels, while our AGFN constructs topological graphs in an unsupervised way by utilizing samples' similarities. Center Loss simultaneously learns a center for deep features of each class and penalizes distances between deep features and their corresponding class centers. In contrast, our AGFN conveys semantic information among samples with GCN to normalize the distribution of features. Therefore, mislabeled data would affect the learning of class centers in Center Loss, resulting in performance degradation, while our AGFN is able to protect FER models from mislabeled data. To provide stronger evidence and assist readers in understanding the distinct feature learning strategies among different models, we also provide the high-dimensional feature visualizations based on t-SNE in Fig. 8.

4.8. Analysis for samples with ambiguity

To intuitively understand the superiority of our AGFN-equipped network, we present a analysis for samples with ambiguity in Fig. 9. From samples in the first three columns, the baseline model usually assigns similar scores to its top-2 predictions and fails to generate correct predictions. In contrast, our network not only generates correct predictions but also predicts top-2 scores with relatively larger distance. Moreover, from the fourth sample in the first row, the baseline model fails to distinguish the 'Fear' expression from 'Surprise', 'Sad' and 'Anger'. It assigns the ground truth label 'Fear' a score of 0.19, which is the same as that of 'Sad' and 'Anger', and lower than that of 'Surprise'. In contrast, our network predicts the expression as 'Fear' with a score of 0.25, which is obviously higher than that of 'Sad' and 'Anger'.

5. Conclusion

This paper proposes to utilize the associative relations of expressions to tackle the excursive semantic and feature covariate shifting problems caused by data uncertainties in FER. It presents an effective feature normalization method named AGFN, who exploits a Poisson graph generator to dynamically and adaptively construct topological graphs for samples in each mini-batches, and employs GCN to convey semantic information among samples. Additionally, to jointly optimize parameters involved in neural networks and the sampling process, a novel coordinate descent strategy is designed. Extensive experiments demonstrate the effectiveness of proposed AGFN and the importance of addressing the data uncertainty problem. Specifically, boundaries of different classes in the feature space become much clearer when extracted features are normalized with proposed AGFN. Moreover, our AGFN-equipped network not only outperforms existing works on the benchmark datasets FERPlus, RAF-DB and AffectNet but also surpasses state-of-the-art works by 14.09%, 21.12% and 13.67% on above datasets when the percentage of mislabeled data significantly increases (i.e., to 20%). Facial Expression Recognition finds widespread application across various domains. It serves as a pivotal tool for emotion analysis, enabling the detection and classification of facial expressions to assess emotions, thereby benefitting market research and user experience enhancement in human–computer interaction (Zhang et al., 2021b). Our method can also be applied to early warning of unexpected events in multiple dense crowd scenarios, such as detecting individual expressions and emotional tendencies in train station dense crowds to provide early warning of unexpected events. This paper mainly discusses the influence of Poisson generators on facial expressions recognition, but the generation of noise is also very important. Our future work mainly focuses on how to study the robustness of noise distribution to sample prediction.

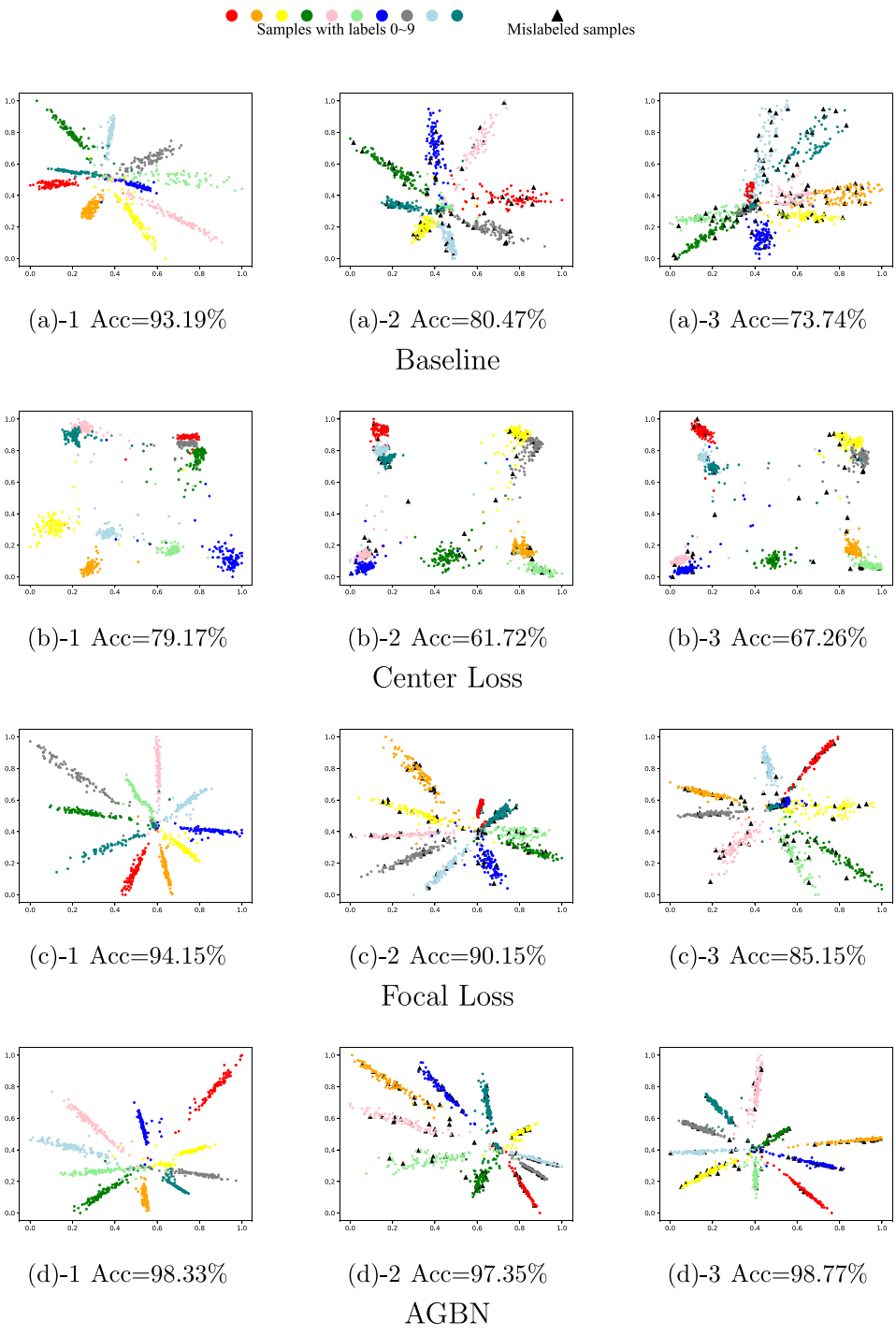

Fig. 8. Comparison of feature visualizations on the MNIST. The first/second/last column of subfigures represents 0%/10%/20% mislabeled samples.

CRediT authorship contribution statement

Yu-Jie Xiong: Writing – review & editing, Validation, Software, Funding acquisition. **Qingqing Wang:** Writing – review & editing, Investigation, Validation. **Yangtao Du:** Writing – original draft, Conceptualization, Software, Methodology. **Yue Lu:** Supervision, Resources.

Declaration of competing 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.

Data availability

Data will be made available on request.

Acknowledgments

This work is supported by National Natural Science Foundation of China (62006150), Science and Technology Commission of Shanghai Municipality (21DZ2203100), and Shanghai Key Laboratory of Multidimensional Information Processing, East China Normal University (No. MIP20225, Fundamental Research Funds for the Central Universities).

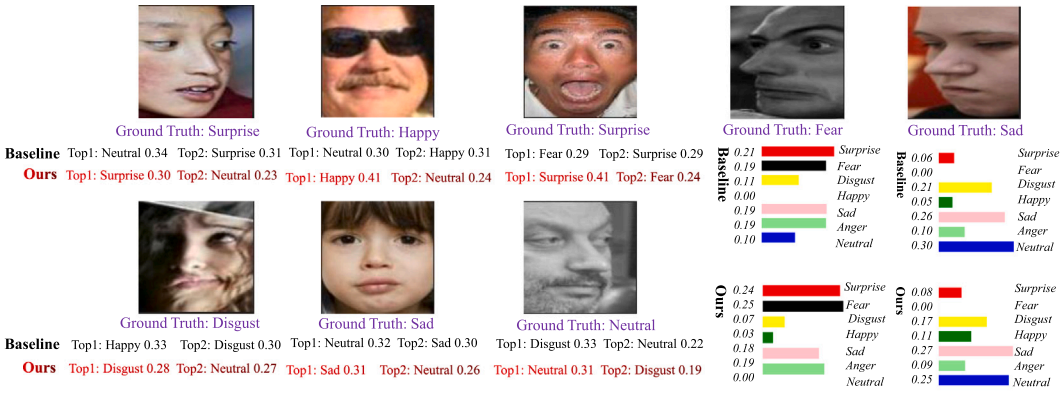

Fig. 9. Results of samples with ambiguity.

References

Acharya, D., Huang, Z., Paudel, D., et al., 2018. Covariance pooling for facial expression recognition. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 367–374.

Albanie, S., Nagrani, A., Vedaldi, A., et al., 2018. Emotion recognition in speech using cross-modal transfer in the wild. In: ACM International Conference on Multimedia. pp. 292–301.

Anderson, J., Bower, G., 2014. Human Associative Memory. Psychology Press.

Barsoum, E., Zhang, C., Ferrer, C.C., Zhang, Z., 2016. Training deep networks for facial expression recognition with crowd-sourced label distribution. In: ACM International Conference on Multimodal Interaction. pp. 279–283.

Bengio, Y., Léonard, N., Courville, A.C., 2013. Estimating or propagating gradients through stochastic neurons for conditional computation. arXiv abs/1308.3432.

Ding, H., Zhou, S., Chellappa, R., 2017. Facenet2expnet: Regularizing a deep face recognition net for expression recognition. In: IEEE International Conference on Automatic Face & Gesture Recognition. pp. 118–126.

Farzaneh, A., Qi, X., 2021. Facial expression recognition in the wild via deep attentive center loss. In: IEEE Winter Conference on Applications of Computer Vision. pp. 2402–2411.

Grimmett, G., Stirzaker, D., 2001. Probability and Random Processes. Oxford University Press.

Guo, S., Huang, W., Zhang, H., et al., 2018. Curriculumnet: Weakly supervised learning from large-scale web images. In: European Conference on Computer Vision. pp. 135–150.

Heidari, A., Javaheri, D., Toumaj, S., Navimipour, N.J., Rezaei, M., Unal, M., 2023. A new lung cancer detection method based on the chest CT images using federated learning and blockchain systems. Artif. Intell. Med. 141, 102572.

Huang, Q., Huang, C., Wang, X., Jiang, F., 2021. Facial expression recognition with grid-wise attention and visual transformer. Inform. Sci. 580, 35–54.

Karim, N., Khalid, U., Esmaeili, A., Rahnavard, N., 2022. CNLL: A semi-supervised approach for continual noisy label learning. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 3878–3888.

Karnati, M., Seal, A., Bhattacharjee, D., Yazidi, A., Krejcar, O., 2023. Understanding deep learning techniques for recognition of human emotions using facial expressions: A comprehensive survey. IEEE Trans. Instrum. Meas. 72, 1–31.

Karnati, M., Seal, A., Yazidi, A., Krejcar, O., 2022a. FLEPNet: Feature level ensemble parallel network for facial expression recognition. IEEE Trans. Affect. Comput. 13 (4), 2058–2070.

Karnati, M., Seal, A., Yazidi, A., Krejcar, O., 2022b. LieNet: A deep convolution neural network framework for detecting deception. IEEE Trans. Cogn. Dev. Syst. 14 (3), 971–984.

Li, S., Deng, W., Du, J., 2017a. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 528527–2861.

Li, C., Li, X., Wang, X., Huang, D., Liu, Z., Liao, L., 2023. FG-AGR: Fine-grained associative graph representation for facial expression recognition in the wild. IEEE Trans. Circuits Syst. Video Technol. 1. http://dx.doi.org/10.1109/TCSVT.2023.3237006.

Li, B., Lima, D., 2021. Facial expression recognition via ResNet-50. Int. J. Cogn. Comput. Eng. 2, 57–64.

Li, H., Wang, N., Ding, X., Yang, X., Gao, X., 2021. Adaptively learning facial expression representation via C-F labels and distillation. IEEE Trans. Image Process. 30, 2016–2028.

Li, Y., Yang, J., Song, Y., et al., 2017b. Learning from noisy labels with distillation. In: International Conference on Computer Vision. pp. 1910–1918.

Li, Y., Zeng, J., Shan, S., Chen, X., 2019. Occlusion aware facial expression recognition using CNN with attention mechanism. IEEE Trans. Image Process. 28, 2439–2450.

Liu, Z., Dong, J., Zhang, C., Wang, L., Dang, J., 2020a. Relation modeling with graph convolutional networks for facial action unit detection. In: International Conference on Multimedia Modeling. pp. 489–501.

Liu, M., Li, S., Shan, S., Chen, X., 2015. Au-inspired deep networks for facial expression feature learning. Neurocomputing 159, 126–136.

Liu, Z., Li, L., Wu, Y., Zhang, C., 2020b. Facial expression relation based on improved graph convolutional networks. In: International Conference on Multimedia Modeling. pp. 527–539.

Liu, D., Zhang, H., Zhou, P., 2020c. Video-based facial expression recognition using graph convolutional networks. In: International Conference on Pattern Recognition. pp. 607–614.

Lo, L., Xie, H., Shuai, H., Cheng, W., 2020. Mer-gcn: Micro-expression recognition based on relation modeling with graph convolutional networks. In: IEEE Conference on Multimedia Information Processing and Retrieval. pp. 79–84.

Ma, F., Sun, B., Li, S., 2021. Facial expression recognition with visual transformers and attention selective fusion. IEEE Trans. Affect. Comput..

Mnih, V., Hinton, G., 2012. Learning to label aerial images from noisy data. In: International Conference on Machine Learning. pp. 567–574.

Mohamed, S., Rosca, M., Figurnov, M., Mnih, A., 2020. Monte Carlo gradient estimation in machine learning. J. Mach. Learn. Res. 21 (132), 1–62.

Mohan, K., Seal, A., Krejcar, O., Yazidi, A., 2021a. Facial expression recognition using local gravitational force descriptor-based deep convolution neural networks. IEEE Trans. Instrum. Meas. 70, 1–12.

Mohan, K., Seal, A., Krejcar, O., Yazidi, A., 2021b. FER-net: facial expression recognition using deep neural net. Neural Comput. Appl. 33, 9125–9136.

Mollahosseini, A., Hasani, B., Mahoor, M., 2017. AffectNet: A dataset for facial expression, valence, and arousal computing in the wild. IEEE Trans. Affect. Comput..

Nian, F., Chen, X., Yang, S., Lv, G., 2019. Facial attribute recognition with feature decoupling and graph convolutional networks. IEEE Access 7, 85500–85512.

Peng, X., Xia, Z., Li, L., Feng, X., 2016. Towards facial expression recognition in the wild: A new database and deep recognition system. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 93–99.

Ruan, D., Yan, Y., Lai, S., et al., 2021. Feature decomposition and reconstruction learning for effective facial expression recognition. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 7660–7669.

Schlett, T., Rathgeb, C., Henniger, O., Galbally, J., Fierrez, J., 2022. Face image quality assessment: A literature survey. ACM Comput. Surv. 54 (10s), 1–49.

Vaswani, A., Shazeer, N., Parmar, N., et al., 2017. Attention is all you need. Adv. Neural Inf. Process. Syst. (NIPS).

Veit, A., Alldrin, N., Chechil, G., Krasin, I., Gupta, A., Belongie, S., 2017. Learning from noisy large-scale datasets with minimal supervision. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 839–847.

Wan, L., Zeiler, M., Zhang, S., Cun, Y., Fergus, R., 2013. Regularization of neural networks using DropConnect. In: International Conference on Machine Learning. pp. 1058–1066.

Wang, K., Peng, X., Yang, J., Lu, S., Qiao, Y., 2020a. Suppresing uncertainties for large-scale facial expression recognition. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 6897–6906.

Wang, K., Peng, X., Yang, J., et al., 2020b. Region attention networks for pose and occlusion robust facial expression recognition. IEEE Trans. Image Process. 29, 4057–4069.

Wang, F., Xiang, X., Liu, C., et al., 2017. Regularizing face verification nets for pain intensity regression. In: IEEE International Conference on Image Processing. pp. 1087–1091.

Wang, Y., Zhou, S., Liu, Y., et al., 2022. ConGNN: Context-consistent cross-graph neural network for group emotion recognition in the wild. Inform. Sci. 610, 707–724.

Wen, Y., Zhang, K., Li, Z., Qiao, Y., 2016. A discriminative feature learning approach for deep face recognition. In: European Conference on Computer Vision. pp. 499–515.

Wu, F., Pang, C., Zhang, B., 2021. FaceCaps for facial expression recognition. Comput. Anim. Virt. Worlds 32, 3–4.

Xie, H., Lo, L., Shuai, H., Cheng, W., 2020a. Au-assisted graph attention convolutional network for micro-expression recognition. In: ACM International Conference on Multimedia. pp. 2871–2880.

Xie, Q., Luong, M., Hovy, E., Le, Q., 2020b. Self-training with noisy student improves ImageNet classification. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 10687–10698.

Xue, F., Wang, Q., Guo, G., 2021. Transfer: Learning relation-aware facial expression representations with transformers. In: International Conference on Computer Vision. pp. 3601–3610.

Zeng, J., Shan, S., Chen, X., 2018. Facial expression recognition with inconsistently annotated datasets. In: European Conference on Computer Vision. pp. 222–237.

Zhang, W., Wang, Y., Qiao, Y., 2019. Metacleaner: Learning to hallucinate clean representations for noisy-labeled visual recognition. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 7373–7382.

Zhang, F., Xu, M., Xu, C., 2021a. Weakly-supervised facial expression recognition in the wild with noisy data. IEEE Trans. Multimed. 24, 1800–1814.

Zhang, X., Yang, X., Zhang, W., Li, G., Yu, H., 2021b. Crowd emotion evaluation based on fuzzy inference of arousal and valence. Neurocomputing 445 (20), 194–205.

Zhao, X., Liang, X., Liu, L., Li, T., et al., 2016. Peak-piloted deep network for facial expression recognition. In: European Conference on Computer Vision. pp. 425–442.

Zhao, R., Liu, T., Huang, Z., et al., 2021. Geometry-aware facial expression recognition via attentive graph convolutional networks. IEEE Trans. Affect. Comput..

Zheng, G., Awadallah, A., Dumais, S., 2021. Meta label correction for noisy label learning. Proc. AAAI Conf. Artif. Intell. 35 (12), 11053–11061.

Zhong, H., 2023. Research And Application of Lightweight Face Expression Recognition Model for Embedded Devices (Master Thesis). University of Electronic Science and Technology of China, Chengdu.