Adaptively Learning Facial Expression Representation via C-F Labels and Distillation

Hangyu Li, Nannan Wang, *Member, IEEE*, Xinpeng Ding, Xi Yang, *Member, IEEE*, and Xinbo Gao, *Senior Member, IEEE*

***Abstract*— Facial expression recognition is of significant importance in criminal investigation and digital entertainment. Under unconstrained conditions, existing expression datasets are highly class-imbalanced, and the similarity between expressions is high. Previous methods tend to improve the performance of facial expression recognition through deeper or wider network structures, resulting in increased storage and computing costs. In this paper, we propose a new adaptive supervised objective named AdaReg loss, re-weighting category importance coefficients to address this class imbalance and increasing the discrimination power of expression representations. Inspired by human beings' cognitive mode, an innovative coarse-fine (C-F) labels strategy is designed to guide the model from easy to difficult to classify highly similar representations. On this basis, we propose a novel training framework named the emotional education mechanism (EEM) to transfer knowledge, composed of a knowledgeable teacher network (KTN) and a self-taught student network (STSN). Specifically, KTN integrates the outputs of coarse and fine streams, learning expression representations from easy to difficult. Under the supervision of the pre-trained KTN and existing learning experience, STSN can maximize the potential performance and compress the original KTN. Extensive experiments on public benchmarks demonstrate that the proposed method achieves superior performance compared to current state-of-the-art frameworks with *88.07%* on RAF-DB, *63.97%* on AffectNet and *90.49%* on FERPlus.**

***Index Terms*— Facial expression recognition, emotional education mechanism, adaptive regular loss, coarse-fine labels, knowledge distillation.**

I. Introduction

FACIAL expression is one of the most important ways to convey emotion in interpersonal communication. Facial expression recognition (FER) refers to the use of computers to extract features of detected faces so that computers can understand facial expressions resulting from human thinking and respond to people's needs. This research is the frontier of multidisciplinary research in image processing, pattern recognition, psychology, affective computing and computer vision. It is quite vital to automatic emotion analysis systems, especially for achieving strong robustness in real situations such as illumination and nonlinear view variations. Numerous FER research endeavours have been conducted and have achieved great performance; FER thus offers a wide range of applications, such as in the fields of medical diagnosis, public safety and so on. In this paper, we review recent works on FER. According to our analysis, the reason why the development of FER is restricted is that large FER models fail to consider the human experience under the influence of class-imbalanced data in solving highly similar expression problems.

In recent years, unconstrained large-scale datasets, such as RAF-DB [1], AffectNet [2], FER2013 [3], and FERPlus [4], have greatly facilitated the development of automated facial expression analysis. There is no doubt that large amounts of facial expression images provide data support for algorithms. However, for existing FER datasets collected from the Internet, it is extremely difficult to obtain sufficient negative expressions (such as disgust, fear and anger) for the major reason of the subtle difference and difficult labeling. To be specific, negative expressions often share the same muscle changes and then are hard to determine labels by annotators with high confidence, leading directly to the decline in quantity and making the label distribution of in-the-wild FER datasets highly imbalanced. As shown in Fig. 1 (left), the number of happy expressions is far greater than other types of expressions: happiness is 16 times more common than fear. Even though all negative emotions can be conveyed by four expressions (fear, disgust, sadness and anger), there are still nearly 1,000 fewer negative expressions than there are expressions of happiness. Existing

Manuscript received June 8, 2020; revised October 6, 2020; accepted January 2, 2021. Date of publication January 13, 2021; date of current version January 22, 2021. This work was supported in part by the National Key Research and Development Program of China under Grant 2018AAA0103202; in part by the National Natural Science Foundation of China under Grant 61922066, Grant 61876142, Grant 61772402, Grant 61976166, and Grant 62036007; in part by the National High-Level Talents Special Support Program of China under Grant CS31117200001; in part by the Xidian University Intellifusion Joint Innovation Laboratory of Artificial Intelligence; in part by the Fundamental Research Funds for the Central Universities; and in part by the Innovation Fund of Xidian University. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Heng Tao Shen. *(Corresponding author: Nannan Wang.)*

Hangyu Li, Nannan Wang, and Xi Yang are with the State Key Laboratory of Integrated Services Networks, School of Telecommunications Engineering, Xidian University, Xi'an 710071, China (e-mail: hangyuli.xidian@gmail.com; nnwang@xidian.edu.cn; yangx@xidian.edu.cn).

Xinpeng Ding is with the State Key Laboratory of Integrated Services Networks, School of Electronic Engineering, Xidian University, Xi'an 710071, China (e-mail: xpding.xidian@gmail.com).

Xinbo Gao is with the Chongqing Key Laboratory of Image Cognition, Chongqing University of Posts and Telecommunications, Chongqing 400065, China (e-mail: gaoxb@cqupt.edu.cn).

Digital Object Identifier 10.1109/TIP.2021.3049955

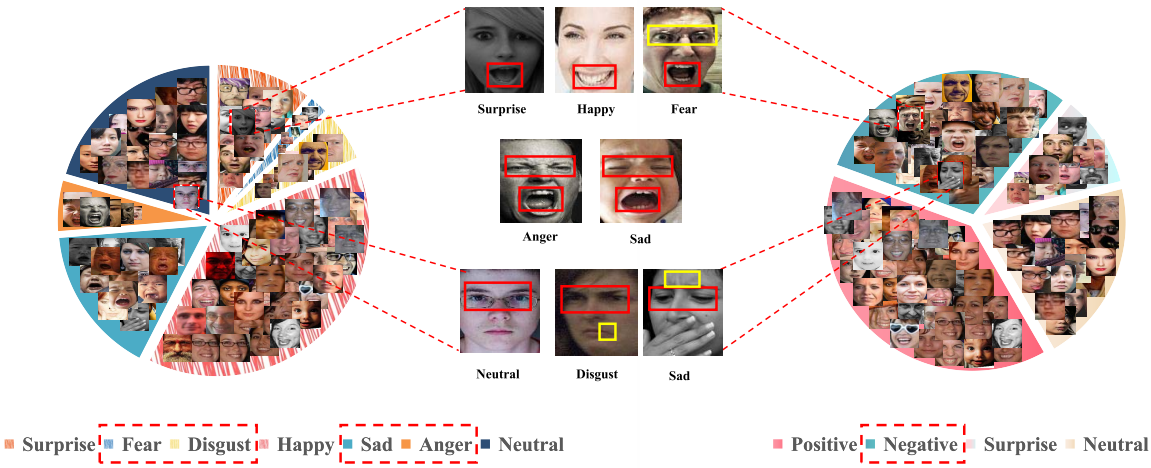

Fig. 1. An example of the data distribution and expression similarity from RAF-DB. The red frames represent similar changes in facial areas. The yellow frames represent different changes in facial areas. It is worth noting that angry and sad expressions often appear to be highly similar, with the two examples above showing the same changes in regions such as the eyes, mouth and eyebrows.

studies on FER [1], [5] were aimed at supervision signals to further improve the discrimination power of expression representation. However, the generalization of CNNs obtained for this purpose is limited by the imbalance of data. Therefore, it is necessary to emphasize the difference between small-scale categories and other (well-classified) examples; i.e., ***how to solve the class-imbalanced FER problem through the adaptive re-weighting of feature differences is worth exploring.***

In addition to the above problem, many expressions require similar muscle movements in the facial area [6], [7], causing small muscle contractions without obvious features. As shown in Fig. 1, sadness and anger produce highly similar changes in the eyes and mouth, while disgust often changes the corner of the mouth slightly, which is vital to determine its category. More importantly, the representation of similar expressions obtained by CNNs is weak, which leads to learning scarcely any important representation information. Specifically, recently developed deep attention methods [8], [9] are affected by high similarity and hardly notice subtle changes in the face (such as the corners of the mouth being slightly raised), which are important for expression features. In addition, the adaptive search for important local expression regions imposes higher requirements for model inference; i.e., cropping regions and patches is a time-consuming task. In contrast, humans do not focus on local information when dealing with the high similarity problem. The cognitive mode of human beings usually considers the whole, preliminarily judges a general range and then determines its category more precisely. Hence, ***motivated by this cognitive mode, an easy-to-difficult strategy should be considered in the FER model.***

Most existing deep CNNs have received extensive attention and have been widely applied in different application fields [10]–[15]. These efforts depend heavily on deep networks with millions or even billions of parameters, and the availability of GPUs with high computing power is the key to their success. However, existing FER models based on deep CNNs are computationally expensive and memory intensive, which impedes their deployment in limited devices, such as handheld devices, smartphones and sensors. It is crucial for real-world FER applications to reduce the storage and computing costs of models. In addition, although the lightweight FER model can achieve comparable results, its latent representation ability remains to be developed. ***It is a natural development to compress the FER model while taking an in-depth look at the representation capability.***

To achieve these goals, we design a complete FER technique in three steps. First, we propose a novel adaptive regular (AdaReg) loss, which is good for monitoring CNNs learning class-imbalanced expression representations. Second, the C-F labels strategy provides a reference for the FER model, which can distinguish expressions ranging from easy to difficult. Specifically, based on existing labels (happiness, surprise, sadness, anger, disgust, fear, and neutral, namely, fine labels), we redefine coarse labels (positive, negative, neutral and surprise) as new auxiliary supervision signals to solve the high similarity FER problem, avoiding the cost of searching the expression region. Third, given the above two aspects, we can achieve a large-scale model with superior performance. To deploy the FER model efficiently and unlock the potential of the lightweight model, we propose a new emotional education mechanism (EEM). Specifically, the pre-trained teacher model (KTN) monitors a student network (STSN) to achieve better performance and also self-learns facial features based on label information. We argue that this framework provides a novel method for the implementation of an intelligent expression recognition system. The main contributions of this paper are as follows.

- We propose an adaptive regular loss function, AdaReg, which can adaptively obtain the importance coefficients of expression representations and effectively combine the class imbalance with the distribution of high-dimensional features.
- Different from the original labels in datasets, we divide expressions into four categories to provide a new idea for

the FER network and reduce the influence of similarity. The coarse-grained result plays an important role in early warning in intelligent video surveillance.

- We propose a new FER scheme named the *emotional education mechanism* to achieve a compression model with better performance. To the best of our knowledge, this is the first attempt to utilize knowledge distillation to help supervise the learning of expression features. With the joint supervision of KTN and C-F labels, highly discriminative features can be obtained for robust FER, as supported by our experimental results.
- We significantly advance the state-of-the-art performance on public FER datasets with both intra-dataset and cross-dataset testing, providing a promising direction for the community and building powerful FER solutions in practice.

The rest of this paper is organized as follows. Section II presents a brief review of FER and related techniques in three respects. The proposed method is introduced in Section III. Section IV provides extensive experimental results and analyses, and the conclusion is given in Section V.

II. Related Work

In this section, we review the previous works that are related to ours in three respects, i.e., facial expression recognition and related techniques, namely, class-imbalanced classification and knowledge distillation.

A. Facial Expression Recognition

In general, a facial expression recognition system mainly includes three modules: face detection (or face alignment), feature extraction and a classifier. Specifically, given a facial image to be detected, the face detector (such as MTCNN [16] and FAN [17]) locates the face in the complex background. Then, the feature extractor captures facial expression features and determines expression categories through the classifier. In recent years, conventional hand-crafted feature-based [18]–[22] and deep learning-based methods have been proposed for facial expression analysis. On one hand, texture and shape are important clues to FER and have been used accordingly. Some hand-crafted features have been proposed in previous works, for example, LBP [23], Gabor [24], HOG [25] and NMF [26]. Most of these features are based on images collected under constrained conditions, such as CK+ [27], MMI [28], Oulu-CASIA [29] and other laboratory-collected datasets. The most common of these features is the LBP, which is an operator used to describe the local texture characteristics of an image. Combined with support vector machines (SVMs) [19], LBP features can be applied to FER. On the other hand, FER via deep learning has achieved a series of breakthroughs in recent years [1], [5], [8], [10], [30]. Zeng *et al.* [30] proposed the deep sparse autoencoders (DSAE) to learn robust and discriminative facial expression features. Next, large-scale facial expression datasets, such as RAF-DB [1], AffectNet [2] and EmotioNet [31], also greatly facilitate deep FER research. Very recently, Li *et al.* [1] and Cai *et al.* [5] supervised the learning process in CNNs by a loss function, which offers more discriminative expression information. With the attention mechanism, Li *et al.* [8] first proposed an ACNN to distinguish facial expressions from partially occluded faces. Wang *et al.* [9] proposed a region-based attention network to capture the importance of facial regions, which were cropped from facial landmarks. In addition, Zeng *et al.* [32] first considered uncertainties and the inconsistent annotation problem and improved the FER performance. Wang *et al.* [33] suppressed uncertainties in FER and achieved the best results to date.

B. Learning With Imbalance

The imbalance in the FER task mainly comes from minority classes (negative emotions). Despite the abundance of image resources on the Internet, it is still difficult to obtain negative expressions in real scenes, resulting in highly class-imbalanced FER training datasets. To resolve this sample imbalance, intuitive ideas include re-weighting [34]–[36] and re-sampling [37]–[39]. Lin *et al.* [36] reshaped the standard cross-entropy (CE) loss to down-weight well-classified samples. Cao *et al.* [40] proposed the combination of re-weighting and re-sampling, which are more effective for the later stage of CNN training, and designed a regularization method to realize good generalization for less frequent classes. *In contrast, our work focuses on the adaptive estimation of the importance coefficients in a mini-batch and the classification of imbalanced samples by increasing the cross-class feature difference in Euclidean space.*

C. Knowledge Distillation

Especially for FER in the wild, excessively deep or wide CNNs are often required to achieve excellent performance, which has greatly limited model deployment. Most recently, the concept of knowledge distillation [41]–[43] has been proposed to achieve better compression ratios while retaining the performance of the original model as much as possible and has achieved great successes in various vision problems including face detection and recognition [44]–[47], object detection and tracking [48], [49], saliency estimation [50], etc. Hinton *et al.* [41] first proposed knowledge distillation and introduced soft targets to induce the training of student networks. On the basis of the former, Romero *et al.* [42] considered middle feature maps in a teacher network and guided student hidden layers. Similarly, Ge *et al.* [46] firstly proposed a two-stream learning approach to recognize low-resolution faces via selective knowledge distillation. Next, Ge *et al.* [47] proposed the bridge distillation to turn a complex face model to a lightweight one. *In contrast, our work focuses on the transfer of fully connected features, i.e., highly discriminant facial expression feature vectors that supervise the training of the self-taught student network. While ensuring that the original teacher network is compressed in a proper proportion, the student network can fulfil its potential well.*

III. Methodology

In this section, we elaborate our proposed approach. We first visually display the distribution of expression data. In view

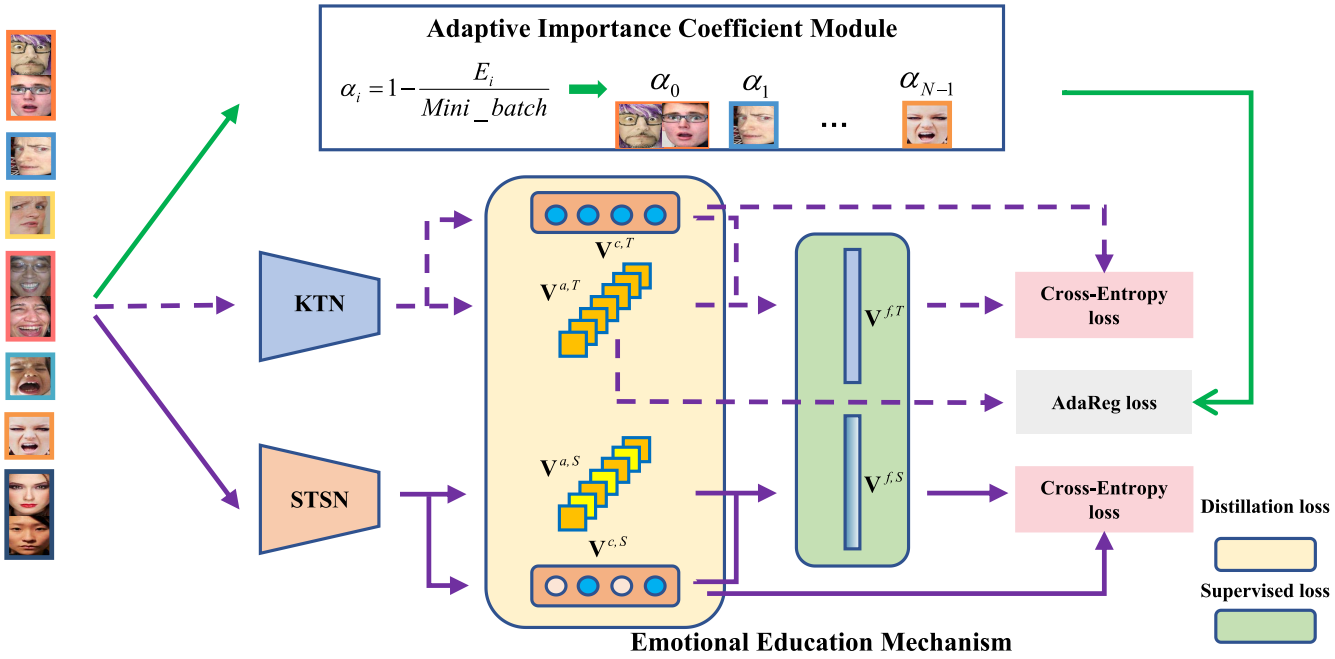

Fig. 2. An overview of our method. A batch of facial expression images is fed into two modules: adaptive importance coefficient module and emotional education mechanism (EEM) module. The former obtains the importance coefficient (weight) of each category and participates in the calculation of AdaReg loss. EEM consists of the knowledgeable teacher network (KTN), self-taught student network (STSN) and two knowledge transfer modules. In the process of learning, the highly discriminant representation, coarse vector and fine vector obtained by KTN are transferred into STSN. Finally, the classification of expressions is realized through CE loss. *It is worth noting that AdaReg loss supervises only the learning of KTN.*

of the extreme imbalance of expression labels, we propose a novel adaptive importance coefficient module to re-weight the expression features. Inspired by human beings' cognitive mode, we design the C-F labels strategy to monitor the model training. Considering the limited model storage, we propose a novel emotional education mechanism (EEM), followed by some discussions.

A. Adaptive Importance Coefficient Module

We introduce an adaptive importance coefficient module to re-weight the contribution of the tail expression representation. Compared with the number of positive samples, the number of samples pertaining to the four negative expressions is greatly lacking. As shown in Fig. 1, the combined number of happy expressions and neutral faces represents more than half of the total. In particular, expressions of fear constitute only two percent of the total. This uneven data distribution directly prohibits the deep features learned by CNNs from reaching the best performance. To resolve this drawback, small expression features should be given a high importance coefficient, while large ones should be given a low importance coefficient. Specifically, let $\mathbf{Y} = [y_1, y_2, \ldots, y_M]$ denote the original label in a mini-batch, where the adaptive importance coefficient module takes $\mathbf{Y}$ as the input and outputs the importance coefficients of each type of expression, formulated as,

$$\alpha_i = 1 - \frac{E_i}{M}, \quad i = 0, 1 \cdots, N-1, \tag{1}$$

where E_i denotes the number of the i^{th} expression in a mini-batch, N represents the maximum label index and M is the mini-batch size. Instead of considering the importance of neutral faces, we focus on other expressions, which are important to FER. During the experiment, we found that superior performance can be achieved without large batches, and a comparison of the settings will be shown in an ablation study IV-D.

1) Adaptive Regular (AdaReg) Loss: Like charges repel, while opposite charges attract; this is one of the fundamental laws in physics. However, the representation of facial expressions is just the opposite. Our goal is to maximize the distances between different emotions as much as possible to supervise the CNN training. Inspired by [10], [51], to maximize the differences between facial expressions in high-dimensional space, we design a regular term as shown in Eq. 2. Specifically, we first initialize every expression category representation $\{\mathbf{e}_i\}_{i=0}^{N}$. With the network training, $\mathbf{e}_i$ is updated iteratively following [10], and the L2 norm between different category representations is calculated to measure the difference in high-dimensional space. We set the reciprocal of the sum of all differences as the final measurement signal. However, this measure is highly susceptible to an imbalanced label distribution. At the same weight ratio, negative emotions are unfairly measured.

$$\mathcal{L}_{Reg} = \frac{1}{\lambda \sum_{i=0}^{N-1} \sum_{j=i+1}^{N} \left\| \mathbf{e}_i - \mathbf{e}_j \right\|_2}, \tag{2}$$

where λ is an expansion parameter to expand the space scale. We will prove its effects in IV-D.

To alleviate unfair measurements, we introduce an adaptive importance coefficient and propose the AdaReg loss,

as shown in Eq. 3. Motivated by [36], we adaptively re-weight the importance of the samples from the perspective of the feature distribution, avoiding manual regulation. Specifically, the importance coefficient α is utilized to weight the representation differences between different categories, i.e., to devote more attention to the differences between tailed expression features and other features. Finally, the objective function is minimized to adaptively monitor the sample distribution. The specific effect will be shown in the following visualization result IV-F.

$$\mathcal{L}_{AdaReg} = \frac{1}{\lambda \sum_{i=0}^{N-1} \sum_{j=i+1}^{N} \alpha_i \left\| \mathbf{e}_i - \mathbf{e}_j \right\|_2}. \tag{3}$$

2) Relation to Prior Work: Here, we discuss the relation between the AdaReg loss and center loss [10], focal loss [36] and uniform loss [51], which share a similar design concept but different emphases. The comparison results in Section IV-C show the superiority of the AdaReg loss for the FER task.

Relation to Center Loss [10]. The center loss (CL) is devoted to narrowing the difference from the same class in identifying similarity measurements of significant value. The lack of inter-class consideration is not suitable for the complex FER task, which is essentially intended to distinguish different-class expressions.

Relation to Focal Loss [36]. Focal loss (FL) focuses on reshaping the standard cross-entropy loss, which has limited constraints on the expression distribution. AdaReg loss focuses on enhancing the expression representation power in high-dimensional space. On the other hand, FL adopts fixed weights to describe sample importance, while AdaReg can determine the importance coefficient automatically.

Relation to Uniform Loss [51]. Uniform loss (UL) prevents excessive repulsion by adding one to each distance and calculates the reciprocal sum of each distance separately but ignores the overall properties of the distance set. However, AdaReg loss aggregates differences across the representation space to maximize the overall distance.

B. Coarse-Fine Labels Strategy

According to the facial action coding system (FACS) [6], [7], a facial expression is the combination of several facial action units. Existing large-scale facial expression datasets are divided into six or seven facial expressions and neutral faces. However, fine-grained expression labelling poses a similarity challenge to FER, especially in the wild. Take the fear and sad expressions as examples. The both expressions change in the same two action units (AU), i.e., inner brow raiser (AU_1) and brow lowerer (AU_4). The high similarity therefore makes an automatic FER system inefficiently learn similar representations and incorrectly classify some expressions.

In recent years, the coarse-to-fine strategy has been widely used in computer vision tasks [52]–[55], which makes features with different grains influence each other. To distinguish highly similar facial expressions, we propose the coarse-fine (C-F) labels strategy to re-cluster four categories and predict categories from easy (coarse) to difficult (fine) to improve the fine-grained FER performance. Specifically, four emotions (fear, disgust, sadness and anger) are uniformly classified as negative expressions, the happy expression is defined as a positive expression, and the surprised expression and neutral face remain the same, thus constituting four coarse labels (positive, negative, surprise and neutral). There are two purposes of this strategy. One is to simulate the human cognitive mode to predict the approximate range (coarse label) of an expression easily and then to determine the specific state (fine label). Second, as new supervision information, the coarse labels, which can assist the learning of original labels, can reduce the impact of similarity and further improve the FER performance. This strategy provides a reference for the following network (KTN/STSN). We will evaluate this strategy in IV-D.

TABLE I

THE NETWORK STRUCTURE OF KTN AND STSN. FOR EXAMPLE, THE INPUT IS THE EXPRESSION IMAGE WITH A SIZE OF 224 × 224 × 3. IT IS WORTH NOTING THAT ONLY IN STSN, TWO RESNET-18 NETWORKS REALIZE WEIGHT SHARING. IN TWO STREAMS, TWO VECTORS (256 DIM) OBTAINED BY THE FIRST FC ARE AVERAGED TO CALCULATE ADAREG. IN ADDITION, FINE OUTPUT IS OBTAINED THROUGH TWO FC LAYERS IN THE FINE STREAM, WHILE THE COARSE STREAM ADOPTS MULTI-LEVEL FC MODE; I.E., 256 IS MAPPED TO 4 AND THEN TO 7

Output	KTN/STSN (Fine Stream)	KTN/STSN (Coarse Stream)
224×224×3	**Input Image**	**Input Image**
7×7×2048/512	ResNet50$_1$/18	ResNet50$_2$/18
2048/512	Average Pool	Average Pool
256	FC1-1	FC2-1
4	-	FC2-2
7	FC1-2	FC2-3

C. Emotional Education Mechanism

To explore the deep potential and practicability of lightweight models, we develop a neural network called the emotional education mechanism (EEM) to learn more discriminative expression information to solve the FER task. As shown in Fig. 2, the EEM model is composed of several modules: knowledgeable teacher network (KTN), self-taught student network (STSN) and two knowledge transfer modules. In this section, we first introduce the structure details of KTN and STSN, respectively. We finally present the detailed implementation of knowledge transfer.

Let $\{(\mathbf{X}_m, y_m, y_m^c)\}_{m=1}^{M}$ denote the set of training data, where $\mathbf{X}_m$ represents a training image, $\mathbf{y}_m$ denotes its original fine label and $\mathbf{y}_m^c$ denotes its coarse label. To complete the personified cognitive mode and make full use of C-F labels, we design a simple yet efficient network. The details of the network structure are shown in Table I.

1) Knowledgeable Teacher Network: KTN, a high-complexity model with superior performance, is an important module of EEM and directly affects the performance of the student network. KTN is a Conv-Streams structure composed of two CNNs to extract expression

representation. The former (fine-stream) focuses on directly learning the original fine label information of facial expressions, while the latter (coarse-stream) obtains the coarse label information by multiple FC mapping and further improves the representation power of the former as auxiliary information. Specifically, we use two 50-layer ResNets as feature extractors, followed by the coarse vector $\mathbf{V}_m^{c,T}$ and expression vector $\mathbf{V}_m^{a,T}$ (obtained through equalization and used to update the category representation). $\mathbf{V}_m^{c,T}$ directly uses the softmax loss to obtain the coarse classification results. Under the supervision of AdaReg loss, differences between different-class representations are adaptively maximized in high-dimensional space. Finally, we combine both coarse and fine stream outputs to yield the final fine vector $\mathbf{V}_m^{f,T}$.

In summary, in the training process of KTN, the whole loss function is defined in Eq. 4, where $\mathcal{L}_{CE}$ denotes the cross-entropy loss. In addition, λ_f, λ_c and λ_A are used to balance three parts. We will evaluate their impact in the ablation study IV-D.

$$\mathcal{L}_{KTN} = \lambda_f \mathcal{L}_{CE}(\mathbf{V}_m^{f,T}, y_m) + \lambda_A \mathcal{L}_{AdaReg} + \lambda_c \mathcal{L}_{CE}(\mathbf{V}_m^{c,T}, y_m^c). \quad (4)$$

2) Self-Taught Student Network: On the premise of independent learning of C-F labels information, STSN utilizes the rich knowledge in KTN as supervision signals, compresses the original KTN and fully demonstrates its potentiality. Moreover, under the guidance of KTN, STSN fully learns the rich expression information in KTN. The details of the internal structure are similar to those in KTN. The only difference is that we use two ResNet-18 networks as backbones and share the convolutional layers of both. The specific STSN training process is actually the transfer of knowledge from KTN to STSN, which will be detailed below.

3) Overall Knowledge Transfer: Given an expression image $\mathbf{X}_m$, it is simultaneously used as input for both KTN and STSN to obtain the coarse feature vector $\mathbf{V}_m^c$ and high-dimensional feature vector $\mathbf{V}_m^a$, respectively. These vectors are mapped and the equalization operation is adopted by fully connected layers to obtain the fine vector $\mathbf{V}_m^f$. $\mathbf{V}_m^c$ and $\mathbf{V}_m^f$ are calculated by softmax loss to obtain four-class and multi-class results. It is worth noting that it is necessary to transfer the rich knowledge in KTN to STSN to realize knowledge transfer. Therefore, we design the distillation loss and supervised loss to achieve a high degree of consistency among all the feature vectors above.

Inspired by [41], [42], we focus on transferring highly discriminant expression representations and C-F label prediction experiences. Specifically, the purpose of the distillation loss $\mathcal{L}_D$ is to reduce the difference in internal knowledge ($\mathbf{V}_m^c$ and $\mathbf{V}_m^a$) in KTN and STSN. To ensure that STSN learns highly discriminant expression representations, we directly calculate the mean square error between $\mathbf{V}_m^{a,T}$ and $\mathbf{V}_m^{a,S}$, as shown in Eq. 5. However, because of the coarse classification result $\mathbf{V}_m^c$ learned by STSN, it directly reflects the probability distribution of coarse labels, and thus, we reduce the difference by calculating the KL divergence. Specifically, the softmax and logarithmic operations are carried out for $\mathbf{V}_m^{c,S}$, and $\mathbf{V}_m^{'c,S} = log(softmax(\mathbf{V}_m^{c,S}))$ is obtained. Then, only the softmax calculation is performed on $\mathbf{V}_m^{c,T}$, and $\mathbf{V}_m^{'c,T}$ is obtained to calculate the difference between the two, as shown in Eq. 6. By calculating $\mathcal{L}_D$, the differences in coarse identification and expression representation in KTN and STSN can be minimized.

$$\mathcal{L}_D = \frac{1}{M}\sum_{m=1}\left(\mathbf{V}_m^{a,S} - \mathbf{V}_m^{a,T}\right)^2 \quad (5)$$

$$+\frac{1}{M}\sum_{m=1}(\mathbf{V}_m^{'c,T} * (log(\mathbf{V}_m^{'c,T}) - \mathbf{V}_m^{'c,S})). \quad (6)$$

Moreover, the $\mathbf{V}_m^f$ obtained through FC mapping are subjected to the supervision of softmax loss to obtain the probability of seven expressions and determine the expression attributes of the input facial image. To make the outputs of the two networks as consistent as possible, we design the supervised loss $\mathcal{L}_S$ in Eq. 7; i.e., we calculate KL divergence between the predicted fine result $\mathbf{V}_m^{f,S}$ and conditional ground-truth distribution $\mathbf{V}_m^{f,T}$.

$$\mathcal{L}_S = \frac{1}{M}\sum_{m=1}(\mathbf{V}_m^{'f,T} * (log(\mathbf{V}_m^{'f,T}) - \mathbf{V}_m^{'f,S})), \quad (7)$$

where $\mathbf{V}_m^{'f,T}$ is calculated by $\mathbf{V}_m^{f,T}$ through the softmax operation and $\mathbf{V}_m^{'f,S} = log(softmax(\mathbf{V}_m^{f,S}))$.

In summary, in the process of knowledge transfer, that is, under the supervision of KTN, the training process of STSN is defined in Eq. 8, where $\mathcal{L}_{CE}$ denotes the cross-entropy loss. The hyperparameters λ_1 and λ_2 are used to balance the loss function. The corresponding ablation study is conducted in Sec. IV-D.

$$\mathcal{L}_{STSN} = \mathcal{L}_D + \mathcal{L}_S + \lambda_1 \mathcal{L}_{CE}(\mathbf{V}_m^{f,S}, y_m) + \lambda_2 \mathcal{L}_{CE}(\mathbf{V}_m^{c,S}, y_m^c). \quad (8)$$

IV. Experiments

In this section, we conduct extensive experiments on the FER task to verify the effectiveness of our method. In the following, we describe the evaluation datasets and metrics (Sec. IV-A), implementation details (Sec. IV-B), results (Sec. IV-D, IV-E) and visualization (Sec. IV-F). Furthermore, we explore the impact of class-imbalanced FER in Sec. IV-C.

A. Datasets & Metrics

1) Datasets: We evaluate the proposed approach on three public unconstrained facial expression datasets, RAF-DB [1], AffectNet [2] and FERPlus [4]. **RAF-DB** includes nearly 30k facial images with two different subsets processed by 40 annotators. In our experiment, a single-label subset with seven classes of basic emotions was used, including 12,271 images as training data and 3,068 images as testing data. **AffectNet** is by far the largest facial expression database in the wild. It consists of approximately 400k manually annotated images. Following the settings in [8], we also used facial images with six basic emotions and neutral faces in our experiment, resulting in 280,000 training images and 3,500 testing

TABLE II
THE EVALUATION OF CLASS-IMBALANCE IMPACTS ON RAF-DB

Method	Fear	Disgust	Sadness	Anger	RAF-DB (Ave.)
Center Loss [10]	45.95	50.00	85.77	78.40	74.94
Uniform Loss [51]	47.30	53.13	78.24	80.25	74.56
Reg Term	58.11	55.63	87.66	80.86	77.97
Focal Loss [36]	**59.46**	55.63	82.85	76.54	76.86
LDAM [40]	**59.46**	52.50	85.98	78.40	77.06
AdaReg Loss	58.11	**60.63**	**91.21**	**81.48**	**79.15**

images. **FERPlus** is extended from FER2013 [3], providing a set of new labels created by 10 crowd-sourced annotators. It consists of 28,709 training images, 3,589 validation images and 3,589 testing images. Unlike the first two, *Contempt* is introduced, resulting in 8 classes in FERPlus.

2) Performance Metrics: The performance metrics we employed here are the average rate and the accuracy. The average rate denotes the average of all expressions (basic expressions plus neutral face), i.e., $Average = \frac{1}{N+1}\sum_{i=0}^{N} ACC_i$, where ACC_i denotes the accuracy for each category. The accuracy is the overall accuracy for all classes.

B. Implementation Details

1) Data Processing: In our experiments, all images are detected by a face alignment algorithm [17] and resized to 224×224 pixels. Our proposed approach is implemented with the PyTorch framework. By default, the ResNet-50/18 in EEM is pre-trained on the ImageNet dataset, and facial representations are extracted from the last convolutional layer. To prevent over-fitting in the training process, we adopt the same data augmentation strategy in all experiments. Specifically, a 200×200 facial image is randomly cropped from the original image and randomly flipped horizontally as the input.

2) Training Setting: We conduct all experiments on a workstation with two NVIDIA TITAN Xp GPUs. We use Adam to optimize our proposed network with initial learning rate of 1e-4 in all training phases. The weight decay parameter is 5e-4 during the whole training process, and the mini-batch size is fixed to 32 by default, whose influence will be discussed in the ablation study IV-D. The learning rate (lr) is updated every four epochs, and the multiplier factor (gamma) of decreasing lr is 0.2. In practice, during the training process, the parameters λ_f, λ_c and λ_A are set as 5, 3 and 0.01, respectively. Furthermore, in the process of STSN training, λ_1 and λ_2 are set as 3 and 1.

C. Exploring the Impact of Class Imbalance on FER

The class imbalance of FER originates mainly from having too few negative emotions (i.e., fear, disgust, sadness and anger). Considering the quantitative analysis of the performance on negative expressions, we verify the robustness of AdaReg loss. In Table II, for a fair comparison, we use ResNet-18 as the backbone and compare AdaReg loss with two state-of-the-art imbalanced-classification methods on RAF-DB, namely, focal loss [36] and LDAM [40]. In addition, to highlight the efficiency of AdaReg compared with similar work, we compare it with two baselines [10], [51] (without considering an imbalance).

TABLE III
EVALUATION OF THE C-F LABELS AND ADAREG LOSS IN KTN IN TERMS OF ACCURACY

C-F Labels	Reg	AdaReg	RAF-DB	FERPlus
×	×	×	85.69	82.82
×	✓	×	86.44	85.81
×	×	✓	87.22	86.54
✓	×	×	87.00	86.35
✓	✓	×	87.48	88.83
✓	×	✓	88.07	90.49

TABLE IV
EVALUATION OF TWO MODULES IN KNOWLEDGE TRANSFER IN TERMS OF AVERAGE/ACCURACY ON RAF-DB AND ACCURACY ON FERPLUS

$\mathcal{L}_D$	$\mathcal{L}_S$	RAF-DB	FERPlus
×	×	77.68/86.25	88.71
✓	×	80.04/86.73	89.21
×	✓	79.26/86.57	89.06
✓	✓	80.32/87.52	89.66

As shown in Table II, our AdaReg is significantly superior to other methods. First, the Reg term achieves better results than UL, indicating the importance of the overall distance. Second, compared to the baseline, FL and LDAM have limited improvement and outstanding performance for only fear and sadness. This may be explained by the fact that they only re-weight CE loss and do not affect the feature distribution. Third, our AdaReg achieves the best results for disgust, sadness and anger and the second-best for fear. Specifically, our AdaReg outperforms the three second-best results by 5%, 5.23% and 1.23%. This shows that AdaReg is effective in class-imbalanced FER and improves performance for small negative emotions.

D. Ablation Study

1) Evaluation of C-F Labels and AdaReg in KTN: To evaluate the effect of modules in KTN, we design the ablation study to examine the influences of C-F labels strategy, Reg term and AdaReg loss on RAF-DB and FERPlus. When the C-F labels strategy is not used, the fine stream can achieve FER alone. As auxiliary supervision information, the coarse label guides the model to simulate the human cognitive mode, i.e., adopting an easy-to-difficult strategy to explore important expression information. In addition, the purpose of AdaReg loss is to adaptively learn highly discriminative expression representations and eliminate the impact of class imbalance. Therefore, the effects of feature discrimination and imbalance should be respectively considered.

As shown in Table III, we offer the following four observations. First, compared with the baseline (first row), each module added improves performance. The reason for this is that the performance of the fine stream is limited under the supervision of the fine labels and CE loss, and the force on the representation is not sufficient. Second, although the Reg term improves the feature discrimination performance to some

TABLE V

THE INTRA-TESTING RESULTS ON RAF-DB IN TERMS OF THE AVERAGE RATE. FOR THE AVERAGE, LARGER VALUES INDICATE BETTER PERFORMANCE. THE BEST RESULT IS SHOWN IN BOLD, AND THE SECOND-BEST RESULT IS UNDERLINED. THIS ALSO APPLIES TO THE FOLLOWING TABLES

Method	Surprise	Fear	Disgust	Happiness	Sadness	Anger	Neutral	Average
VGG	67.48	14.87	33.75	92.41	77.41	53.09	82.35	60.19
ResNet	72.04	22.97	18.13	90.04	59.21	62.58	82.50	58.35
CL-CNN [10]	86.63	9.46	14.38	94.09	83.26	71.61	83.68	63.30
DLP-CNN [1]	81.16	62.16	52.15	92.83	80.13	71.60	80.29	74.20
IL-CNN [5]	78.42	32.43	45.63	91.48	76.78	70.99	75.74	67.35
DeepEmo [56]	-	-	-	-	-	-	-	68.20
Boosting-POOF [57]	80	64	57	89	74	73	76	73.19
STSN	**86.63**	59.46	**66.25**	94.35	**87.87**	**82.72**	85.00	80.32
KTN	83.28	**68.92**	65.62	**94.60**	87.24	81.48	**88.53**	**81.38**

extent, it is still affected by the imbalanced classes, and thus, the performance is further improved by adaptively adjusting the importance coefficient. Third, the gains of C-F labels and AdaReg are close (third and fourth rows), but that of AdaReg is slightly higher. Through our analysis, it is hypothesized that KTN lacks the constraint on category representation, while the features in the fine stream are highly discriminative owing to AdaReg. Finally, when adding a new module, we obtain the best improvement through the C-F labels strategy; i.e., the baseline is improved from 82.82% to 86.35% on FERPlus. We believe that the reason for this is that the coarse stream learns important representations from easy to difficult, which contributes to fine stream learning and effectively reduces the effects of highly similar expression.

2) Effectiveness of Knowledge Transfer in EEM: In this part, we explore the impacts of modules on knowledge transfer, namely, the advantages of the emotional education mechanism. Specifically, we compare the effects of distillation loss and supervised loss on RAF-DB and FERPlus. As shown in Table IV, we can intuitively obtain some observations in the following. First, both modules (losses) improve the initial performance compared to the baseline (first row). Due to the self-study of STSN under the supervision of CE loss, the representation ability is not fully effective. Thus, with the rich knowledge transfer of KTN, all the potential within the lightweight model is explored to further improve the performance of STSN. Second, the maximum gain is obtained by distillation loss; i.e., the baseline is improved from 77.68% to 80.04% in terms of the average rate. This indicates that compared with the fine results, internal knowledge in KTN plays an important role in FER. In addition, the feature visualization results of KTN and STSN highlight the importance of knowledge transfer in Sec. IV-F. Besides evaluating the effects of two supervision signals in EEM, we compare several lightweight backbones in STSN on RAF-DB, as shown in Table VI. It shows that several backbones with EEM achieve better performance compared to them without EEM, indicating that the EEM can explore latent learning ability of lightweight models.

3) Impact of the Mini-Batch Size M: M is the number of facial expressions in a mini-batch, and is directly related to the effects of importance coefficient and AdaReg loss. We study the impacts of different M values from 8 to 128 on RAF-DB. The results are shown in Fig. 3 (left). M defaults to 32 and

TABLE VI

EVALUATION OF DIFFERENT LIGHTWEIGHT BACKBONES IN STSN ON RAF-DB IN TERMS OF ACCURACY

Backbone in STSN	Strategy	RAF-DB
SqueezeNet [58]	w/o EEM	77.44
	w/ EEM	78.13
ShuffleNet [59]	w/o EEM	79.21
	w/ EEM	80.48
MobileNet [60]	w/o EEM	85.79
	w/ EEM	86.64

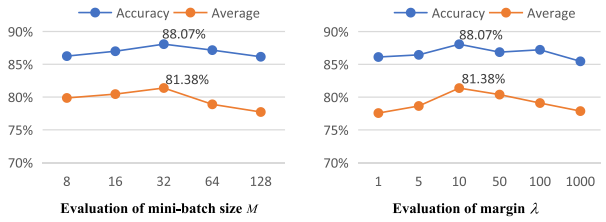

Fig. 3. Evaluation of the mini-batch size M and expansion parameter λ on RAF-DB. Larger values correspond to better performance.

achieves the best classification performance. The smaller M is, the smaller the number of each category, especially negative expressions, which is bad for the generalization of the algorithm and leads to performance degradation. In addition, the higher the value of M, the more the performance is decreased since it excessively considers negative emotions. More importantly, compared with the batch size (1,024) in [33], our method is approachable for researchers with limited equipment.

4) Impact of the Expansion Parameter λ: It is generally known that high-dimensional representation in CNNs is limited between 0 and 1 through the BN layer and that the reciprocal of the difference is too large, which is not conductive to optimization. The parameter λ is used to control the representation difference and enlarge the scale of the representation space. We explore the effects of different expansion parameters λ from 1 to 1,000 on RAF-DB. As shown in Fig. 3 (right), λ is set to 10 by default. Small λ values reduce the model performance since the AdaReg calculation value is too large and the optimization is affected. Large λ excessively considers

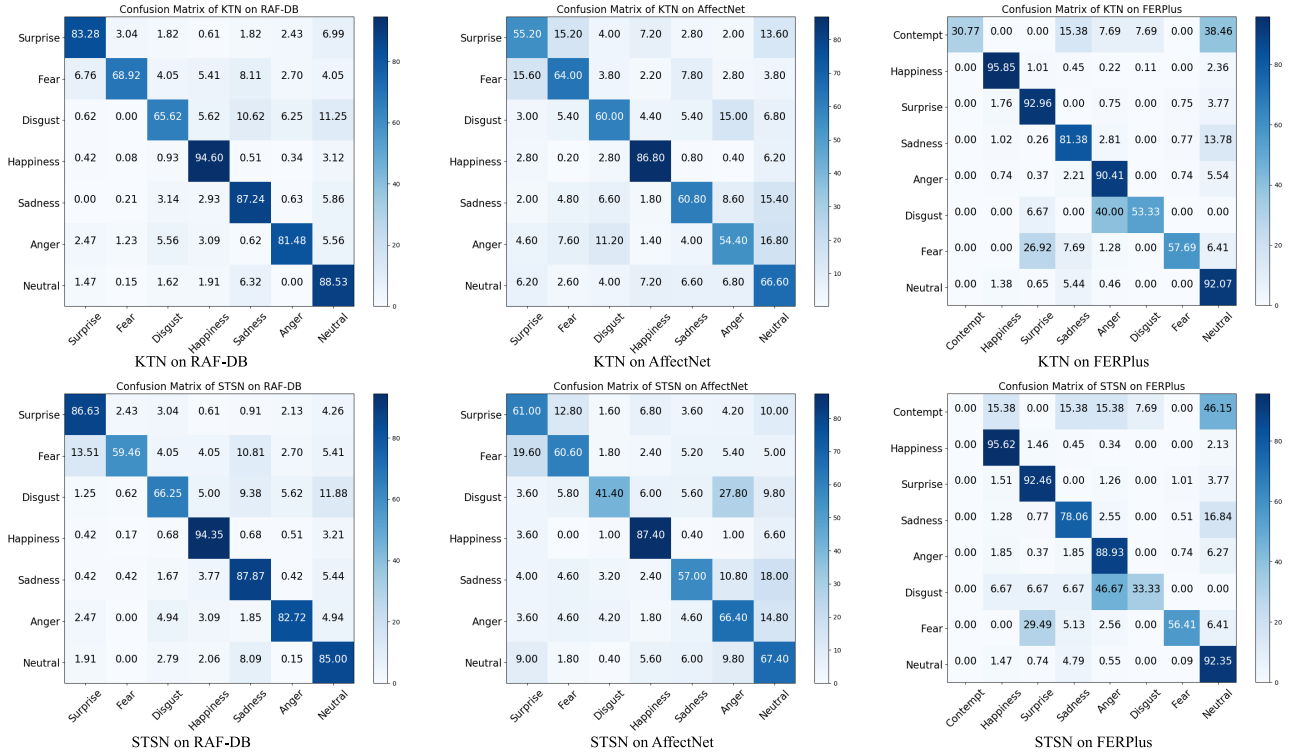

Fig. 4. The confusion matrices of KTN and STSN on the RAF-DB, AffectNet and FERPlus test sets.

the representation difference and unreasonably decreases the overall loss value.

E. Comparison With State-of-the-Art Methods

Our proposed method is compared with the state-of-the-art methods on three datasets. We conduct cross-dataset testing to verify the generalization ability of the proposed algorithm for unknown data distributions. It is worth noting that to better highlight the superiority of each method, we utilize VGG [61] and ResNet [62] under the supervision of CE loss as the baseline. We report the performance of two parts in EEM, STSN and KTN as our results.

1) Results on RAF-DB: To better highlight the advantages of the proposed method, the results obtained by VGG under the supervision of the island or center loss are used as the performance of IL-CNN or CL-CNN on RAF-DB. A comparison of the different approaches in terms of the average rate is shown in Table V. We also report a relevant comparison of the accuracy results in Table VII (second column).

Table V reveals the following. 1) Compared with the state-of-the-art methods, our method achieves the best performance for all categories. In particular, KTN achieves the three best values and four second-best values. STSN realizes the four best values and two second-best values, especially the superior performance with respect to negative emotions. 2) For the average rate, KTN achieves the best result, outperforming the existing best result by 7%. Although the compression ratio is improved by more than four times, the performance loss of STSN is slightly reduced; however, the result is still 6% higher than with DLP-CNN implementation, which satisfies the realistic needs. 3) According to the performance with respect to negative expression, the results of four negative emotions are far beyond those of the existing methods. Especially for the fear emotion, as the least sampled expression, our results exceed the best result by almost 5%. This indicates that our method is of great significance to the recognition of negative expressions.

As shown in Table VII, our method continues to outperform the state-of-the-art techniques. The performance improvement is significant. In particular, both STSN and SCN [33] utilize ResNet-18 as the backbone, but the results of STSN are still 0.5% higher than those of SCN without an additional data pre-training backbone. To highlight the advantages of coarse labels, we report the results of four categories in KTN: **93.84%** (positive), **89.13%** (negative), **83.38%** (neutral) and **79.03%** (surprise). From these results, we can intuitively find that the recognition performance in all four categories is excellent. The results can serve an important role for early warning and the preliminary screening of the major attributes of expressions in criminal investigation and teaching.

2) Results on AffectNet: AffectNet, as the largest dataset of facial expressions, is a very challenging dataset. The testing set has not been released yet, and thus, we evaluate our method only on the validation set. In this part, we provide the comparison of the accuracy rate results obtained by existing approaches based on experiments. For the sake of fairness, we use the same settings as in [9], [33]. As shown in Table VII (third column), compared with the state-of-the-art methods, our KTN achieves the best result at 63.97%. In particular, STSN, a lightweight model, achieves the second-best performance, which is close to that of the original model; compared to gACNN, our STSN reduces model storage by more than

TABLE VII

COMPARISONS OF DIFFERENT METHODS ON THE RAF-DB AND AFFECTNET DATASETS IN TERMS OF THE ACCURACY RATE

Method	RAF-DB	AffectNet
VGG	78.16	50.46
ResNet	74.87	54.37
DLP-CNN [1]	80.89	54.47
GAN-Inpainting [63]	81.87	52.97
pACNN [8]	83.27	55.33
CL-CNN [10]	81.91	54.11
IL-CNN [5]	82.30	56.06
gACNN [8]	85.07	58.78
IPA2LT [32]	86.77	55.71
RAN [9]	86.90	52.97
CovPool [64]	87.00	-
SCN [33]	87.03	60.23
STSN	87.52	63.03
KTN	**88.07**	**63.97**

TABLE VIII

COMPARISONS OF DIFFERENT METHODS ON FERPLUS IN TERMS OF THE ACCURACY RATE

Method	FERPlus
VGG	83.87
ResNet	83.52
PLD [4]	85.1
RAN [9]	88.55
SeNet50 [65]	88.8
RAN-VGG16 [9]	89.16
SCN [33]	88.01/89.35
STSN	89.66
KTN	**90.49**

30M and improves performance by 4%. Moreover, the results of four categories in KTN, **87.40%** (positive), **87.75%** (negative), **55.00%** (neutral) and **43.60%** (surprise), are also competitive.

3) Results on FERPlus: Table VIII compares the performance of our method with the state-of-the-art results on the FERPlus dataset. It can be seen that our method performs the best. Specifically, the result obtained by KTN exceeds that obtained by SCN by 1.14%. It should be noted that SCN utilizes IR50 [66] as the backbone to achieve its performance (89.35%), while our STSN uses a small model to achieve and exceed it by 0.31%. With the same backbone (ResNet-18), our result is 1.65% better than the SCN result.

4) Cross-Dataset Evaluation: The purpose of cross-testing is to verify the generalization ability of the pre-trained model. Specifically, we set up two cross-validation schemes. The first is the trained model on RAF-DB with testing on AffectNet. The second is to exchange the order of the two datasets and complete the relevant experiments again. In [32], there is a certain degree of annotation bias between different datasets; thus, the results of cross-validation are not ideal at present but are used to reliably verify the generalization capacity of different methods.

A comparison of the results is shown in Table IX. The results reveal the following. 1) Our method achieves better performance than the other methods with no exceptions. Although the focal loss and LDAM improve the generalization result to some extent, the effect is not obvious. However, our KTN achieves extremely competitive results, indicating that our model has better generalization power. 2) Interestingly, for the second validation scheme, STSN achieves results superior to KTN. This may be is due to the large scale and internal diversity of AffectNet.

TABLE IX

CROSS-DATASET EVALUATION COMPARISON ON THE RAF-DB AND AFFECTNET DATASETS IN TERMS OF THE ACCURACY RATES. THE RESULTS FOR AFFCTNET AND RAF-DB ARE PROVIDED IN TWO PARTS: ACCURACY AND AVERAGE

Method	Train	Test	Train	Test
	RAF-DB	AffectNet	AffectNet	RAF-DB
VGG	36.74		72.33/53.04	
ResNet	28.77		63.79/39.12	
DLP-CNN [1]	38.37		72.43/53.52	
IL-CNN [5]	39.31		73.99/55.80	
Focal Loss [36]	43.03		75.09/56.90	
LDAM [40]	46.51		74.09/57.13	
STSN	48.49		**76.99/60.92**	
KTN	**49.60**		76.53/59.63	

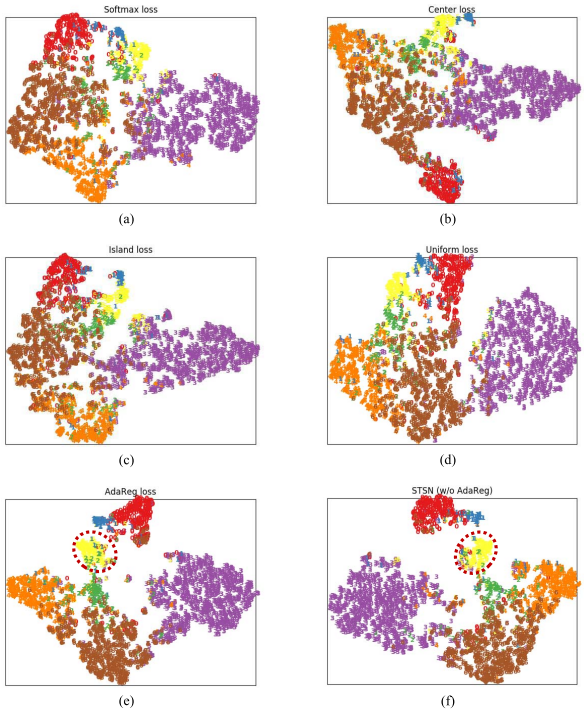

Fig. 5. A visualization study of the feature distribution obtained by KTN using different values of the loss, softmax loss (a), center loss (b), island loss (c), uniform loss (d) and AdaReg loss (e) on RAF-DB. STSN (without AdaReg) (f) is provided to observe the effect of knowledge transfer.

F. Visualization

Our proposed AdaReg loss constrains the distribution of expression representations in high-dimensional space. To verify the effect, we visualize the feature distribution and analyse why AdaReg performs well in this subsection.

We adopt the t-SNE [67] method to visualize representation distribution. As shown in Fig. 5, we compare the feature

distribution of KTN under the supervision of different supervision signals. It is clear that the feature with AdaReg loss (Fig. 5(e)) presents superior performance. Although CL considers the intra-class distance, the inter-class distance cannot be increased. Inspired by CL, UL considers the inter-class similarity but is affected by the label imbalance, and the feature distribution is not obvious. What is most impressive is that our AdaReg loss can achieve a clear boundary between different categories with a large blank space inside. In addition, for the visualization of KTN and STSN (Fig. 5(f)), it is obvious that they are highly similar, indicating that STSN has learned rich knowledge in KTN, especially without AdaReg.

V. Conclusion

This paper proposed the AdaReg loss and C-F labels strategy to solve a previously unexplored problem: how to solve the FER problem of distinguishing highly similar expressions from easy to difficult under the influence of the class imbalance. To the best of our knowledge, this is the first work to address class-imbalanced FER. We first proposed the emotional education mechanism (EEM) to resolve the inadequate representation of the lightweight model, providing a solution for the practical application of the FER system. Specifically, we utilized the knowledgeable teacher network (KTN) to achieve a self-taught student network (STSN) with excellent performance. Experimental results on three large-scale public FER datasets demonstrated the superiority and effectiveness of the proposed approach.

References

[1] S. Li, W. Deng, and J. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2852–2861.

[2] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Trans. Affect. Comput.*, vol. 10, no. 1, pp. 18–31, Jan. 2019.

[3] I. J. Goodfellow *et al.*, "Challenges in representation learning: A report on three machine learning contests," in *Proc. Int. Conf. Neural Inf. Process.* Berlin, Germany: Springer, 2013, pp. 117–124.

[4] E. Barsoum, C. Zhang, C. C. Ferrer, and Z. Zhang, "Training deep networks for facial expression recognition with crowd-sourced label distribution," in *Proc. 18th ACM Int. Conf. Multimodal Interact. (ICMI)*, Oct. 2016, pp. 279–283.

[5] J. Cai, Z. Meng, A. S. Khan, Z. Li, J. OReilly, and Y. Tong, "Island loss for learning discriminative features in facial expression recognition," in *Proc. 13th IEEE Int. Conf. Automat. Face Gesture Recognit. (FG)*, May 2018, pp. 302–309.

[6] E. Friesen and P. Ekman, *Facial Action Coding System: A Technique for the Measurement of Facial Movement*, vol. 3. Palo Alto, CA, USA: Consulting Psychologists Press, 1978.

[7] W. V. Friesen *et al.*, "EMFACS-7: Emotional facial action coding system," Univ. California San Francisco, San Francisco, CA, USA, 1983, vol. 2, no. 36, p. 1.

[8] Y. Li, J. Zeng, S. Shan, and X. Chen, "Occlusion aware facial expression recognition using CNN with attention mechanism," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2439–2450, May 2019.

[9] K. Wang, X. Peng, J. Yang, D. Meng, and Y. Qiao, "Region attention networks for pose and occlusion robust facial expression recognition," *IEEE Trans. Image Process.*, vol. 29, pp. 4057–4069, 2020.

[10] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 499–515.

[11] H. Wang *et al.*, "CosFace: Large margin cosine loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5265–5274.

[12] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[13] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 568–576.

[14] N. Zeng, Z. Wang, H. Zhang, K.-E. Kim, Y. Li, and X. Liu, "An improved particle filter with a novel hybrid proposal distribution for quantitative analysis of gold immunochromatographic strips," *IEEE Trans. Nanotechnol.*, vol. 18, pp. 819–829, 2019.

[15] N. Zeng *et al.*, "Deep-reinforcement-learning-based images segmentation for quantitative analysis of gold immunochromatographic strip," *Neurocomputing*, to be published, doi: 10.1016/j.neucom.2020.04.001.

[16] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.

[17] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2D & 3D face alignment problem? (And a dataset of 230,000 3D facial landmarks)," in *Proc. Int. Conf. Comput. Vis.*, 2017, pp. 1021–1030.

[18] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with Gabor wavelets," in *Proc. 3rd IEEE Int. Conf. Automat. Face Gesture Recognit.*, Apr. 1998, pp. 200–205.

[19] Y. Luo, C.-M. Wu, and Y. Zhang, "Facial expression recognition based on fusion feature of PCA and LBP with SVM," *Optik-Int. J. Light Electron Opt.*, vol. 124, no. 17, pp. 2767–2770, Sep. 2013.

[20] C. Shan, S. Gong, and P. W. McOwan, "Robust facial expression recognition using local binary patterns," in *Proc. IEEE Int. Conf. Image Process.*, vol. 2, Sep. 2005, p. II-370.

[21] Y. Hu, Z. Zeng, L. Yin, X. Wei, X. Zhou, and T. S. Huang, "Multi-view facial expression recognition," in *Proc. 8th IEEE Int. Conf. Automat. Face Gesture Recognit.*, Sep. 2008, pp. 1–6.

[22] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Image Vis. Comput.*, vol. 27, no. 6, pp. 803–816, May 2009.

[23] M. Pietikäinen, A. Hadid, G. Zhao, and T. Ahonen, *Computer Vision Using Local Binary Patterns* (Computational Imaging and Vision), vol. 40. London, U.K.: Springer, 2011, doi: 10.1007/978-0-85729-748-8.

[24] C. Liu and H. Wechsler, "Gabor feature based classification using the enhanced Fisher linear discriminant model for face recognition," *IEEE Trans. Image Process.*, vol. 11, no. 4, pp. 467–476, Apr. 2002.

[25] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2005, pp. 886–893.

[26] R. Zhi, M. Flierl, Q. Ruan, and W. B. Kleijn, "Graph-preserving sparse nonnegative matrix factorization with application to facial expression recognition," *IEEE Trans. Syst. Man, Cybern. B, Cybern.*, vol. 41, no. 1, pp. 38–52, Feb. 2011.

[27] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.-Workshops*, Jun. 2010, pp. 94–101.

[28] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jun. 2005, p. 5.

[29] G. Zhao, X. Huang, M. Taini, S. Z. Li, and M. Pietikäinen, "Facial expression recognition from near-infrared videos," *Image Vis. Comput.*, vol. 29, no. 9, pp. 607–619, Aug. 2011.

[30] N. Zeng, H. Zhang, B. Song, W. Liu, Y. Li, and A. M. Dobaie, "Facial expression recognition via learning deep sparse autoencoders," *Neurocomputing*, vol. 273, pp. 643–649, Jan. 2018.

[31] C. F. Benitez-Quiroz, R. Srinivasan, and A. M. Martinez, "EmotioNet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5562–5570.

[32] J. Zeng, S. Shan, and X. Chen, "Facial expression recognition with inconsistently annotated datasets," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 222–237.

[33] K. Wang, X. Peng, J. Yang, S. Lu, and Y. Qiao, "Suppressing uncertainties for large-scale facial expression recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6897–6906.

[34] C. Huang, Y. Li, C. C. Loy, and X. Tang, "Learning deep representation for imbalanced classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5375–5384.

[35] S. H. Khan, M. Hayat, M. Bennamoun, F. A. Sohel, and R. Togneri, "Cost-sensitive learning of deep feature representations from imbalanced data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 8, pp. 3573–3587, Aug. 2018.
[36] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.
[37] N. Japkowicz, "The class imbalance problem: Significance and strategies," in *Proc. Int. Conf. Artif. Intell.*, 2000, pp. 1–7.
[38] Y. Cui, Y. Song, C. Sun, A. Howard, and S. Belongie, "Large scale fine-grained categorization and domain-specific transfer learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4109–4118.
[39] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009.
[40] K. Cao, C. Wei, A. Gaidon, N. Arechiga, and T. Ma, "Learning imbalanced datasets with label-distribution-aware margin loss," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 1565–1576.
[41] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*. [Online]. Available: http://arxiv.org/abs/1503.02531
[42] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "FitNets: Hints for thin deep nets," 2014, *arXiv:1412.6550*. [Online]. Available: http://arxiv.org/abs/1412.6550
[43] T. Fukuda, M. Suzuki, G. Kurata, S. Thomas, J. Cui, and B. Ramabhadran, "Efficient knowledge distillation from an ensemble of teachers," in *Proc. Interspeech*, 2017, pp. 3697–3701.
[44] P. Luo, Z. Zhu, Z. Liu, X. Wang, and X. Tang, "Face model compression by distilling knowledge from neurons," in *Proc. AAAI Conf. Artif. Intell.*, 2016, vol. 30, no. 1, pp. 1–7.
[45] X. Dong and Y. Yang, "Teacher supervises students how to learn from partially labeled images for facial landmark detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 783–792.
[46] S. Ge, S. Zhao, C. Li, and J. Li, "Low-resolution face recognition in the wild via selective knowledge distillation," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 2051–2062, Apr. 2019.
[47] S. Ge, S. Zhao, C. Li, Y. Zhang, and J. Li, "Efficient low-resolution face recognition via bridge distillation," *IEEE Trans. Image Process.*, vol. 29, pp. 6898–6908, 2020.
[48] G. Chen, W. Choi, X. Yu, T. Han, and M. Chandraker, "Learning efficient object detection models with knowledge distillation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 742–751.
[49] S. Ge, Z. Luo, C. Zhang, Y. Hua, and D. Tao, "Distilling channels for efficient deep tracking," *IEEE Trans. Image Process.*, vol. 29, pp. 2610–2621, 2020.
[50] J. Li, K. Fu, S. Zhao, and S. Ge, "Spatiotemporal knowledge distillation for efficient estimation of aerial video saliency," *IEEE Trans. Image Process.*, vol. 29, pp. 1902–1914, 2020.
[51] Y. Duan, J. Lu, and J. Zhou, "UniformFace: Learning deep equidistributed representation for face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3415–3424.
[52] J. Zhang, S. Shan, M. Kan, and X. Chen, "Coarse-to-fine auto-encoder networks (CFAN) for real-time face alignment," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 1–16.
[53] S. Zhu, C. Li, C. C. Loy, and X. Tang, "Face alignment by coarse-to-fine shape searching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4998–5006.
[54] M. Zhang, N. Wang, Y. Li, R. Wang, and X. Gao, "Face sketch synthesis from coarse to fine," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 1–8.
[55] P. Gao, K. Lu, J. Xue, L. Shao, and J. Lyu, "A coarse-to-fine facial landmark detection method based on self-attention mechanism," *IEEE Trans. Multimedia*, early access, Apr. 30, 2020, doi: 10.1109/TMM.2020.2991507.
[56] W. Deng, J. Hu, S. Zhang, and J. Guo, "DeepEmo: Real-world facial expression analysis via deep learning," in *Proc. Vis. Commun. Image Process. (VCIP)*, Dec. 2015, pp. 1–4.
[57] Z. Liu, S. Li, and W. Deng, "Boosting-POOF: Boosting part based one vs one feature for facial expression recognition in the wild," in *Proc. 12th IEEE Int. Conf. Automat. Face Gesture Recognit. (FG)*, May 2017, pp. 967–972.
[58] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size," 2016, *arXiv:1602.07360*. [Online]. Available: http://arxiv.org/abs/1602.07360
[59] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6848–6856.
[60] A. G. Howard *et al.*, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*. [Online]. Available: http://arxiv.org/abs/1704.04861
[61] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556
[62] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[63] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1125–1134.
[64] D. Acharya, Z. Huang, D. P. Paudel, and L. Van Gool, "Covariance pooling for facial expression recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 367–374.
[65] S. Albanie, A. Nagrani, A. Vedaldi, and A. Zisserman, "Emotion recognition in speech using cross-modal transfer in the wild," in *Proc. 26th ACM Int. Conf. Multimedia*, Oct. 2018, pp. 292–301.
[66] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4690–4699.
[67] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

Hangyu Li received the B.Eng. degree in electronic and information engineering from Shandong University, Jinan, China, in 2017. He is currently pursuing the Ph.D. degree with the School of Telecommunications Engineering, Xidian University, Xi'an, China. His current research interests include computer vision, machine learning, and affective computing.

Nannan Wang (Member, IEEE) received the B.Sc. degree in information and computation science from the Xi'an University of Posts and Telecommunications, Xi'an, China, in 2009, and the Ph.D. degree in information and telecommunications engineering from Xidian University, Xi'an, in 2015. From September 2011 to September 2013, he has been a Visiting Ph.D. Student with the University of Technology, Sydney, NSW, Australia. He is currently a Professor with the State Key Laboratory of Integrated Services Networks, Xidian University. He has published more than 100 articles in refereed journals and proceedings, including the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE (TPAMI), *International Journal of Computer Vision* (IJCV), NeurIPS, and ECCV. His current research interests include computer vision, pattern recognition, and machine learning.

Xinpeng Ding received the B.Eng. degree in software engineering from Xidian University, Xi'an, China, in 2018, where he is currently pursuing the M.S. degree with the School of Electronic Engineering. His current research interests include computer vision, pattern recognition, and machine learning.

Xi Yang (Member, IEEE) received the B.Eng. degree in electronic information engineering and the Ph.D. degree in pattern recognition and intelligence systems from Xidian University, Xi'an, China, in 2010 and 2015, respectively. From 2013 to 2014, she was a Visiting Ph.D. Student with the Department of Computer Science, The University of Texas at San Antonio, San Antonio, TX, USA. In 2015, she joined the State Key Laboratory of Integrated Services Networks, School of Telecommunications Engineering, Xidian University, where she is currently an Associate Professor in communications and information systems. Her current research interests include image/video processing, computer vision, and multimedia information retrieval.

Xinbo Gao (Senior Member, IEEE) received the B.Eng., M.Sc., and Ph.D. degrees in electronic engineering, signal and information processing from Xidian University, Xi'an, China, in 1994, 1997, and 1999, respectively. From 1997 to 1998, he was a Research Fellow with the Department of Computer Science, Shizuoka University, Shizuoka, Japan. From 2000 to 2001, he was a Post-Doctoral Research Fellow with the Department of Information Engineering, The Chinese University of Hong Kong, Hong Kong. Since 2001, he has been with the School of Electronic Engineering, Xidian University. He is currently a Cheung Kong Professor of Ministry of Education, China, a Professor of pattern recognition and intelligent system, Xidian University, and a Professor of computer science and technology, Chongqing University of Posts and Telecommunications. He has published six books and around 300 technical articles in refereed journals and proceedings. His current research interests include image processing, computer vision, multimedia analysis, machine learning, and pattern recognition. He is a fellow of the Institute of Engineering and Technology and the Chinese Institute of Electronics. He served as the General Chair/Co-Chair, the Program Committee Chair/Co-Chair, or a PC Member for around 30 major international conferences. He is on the Editorial Boards of several journals, including *Signal Processing* (Elsevier) and *Neurocomputing* (Elsevier).