Semantic-Rich Facial Emotional Expression Recognition

Keyu Chen, Xu Yang, Changjie Fan, Wei Zhang, and Yu Ding

Abstract—The ability to perceive human facial emotions is an essential feature of various multi-modal applications, especially in the intelligent human-computer interaction (HCI) area. In recent decades, considerable efforts have been put into researching automatic facial emotion recognition (FER). However, most of the existing FER methods only focus on either basic emotions such as the seven/eight categories (e.g., *happiness, anger* and *surprise*) or abstract dimensions (*valence, arousal, etc.*), while neglecting the fruitful nature of emotion statements. In real-world scenarios, there is definitely a larger vocabulary for describing human's inner feelings as well as their reflection on facial expressions. In this work, we propose to address the semantic richness issue in the FER problem, with an emphasis on the granularity of the emotion concepts. Particularly, we take inspiration from former psycho-linguistic research, which conducted a prototypicality rating study and chose 135 emotion names from hundreds of English emotion terms. Based on the 135 emotion categories, we investigate the corresponding facial expressions by collecting a large-scale 135-class FER image dataset and propose a consequent facial emotion recognition framework. To demonstrate the accessibility of prompting FER research to a fine-grained level, we conduct extensive evaluations on the dataset credibility and the accompanying baseline classification model. The qualitative and quantitative results prove that the problem is meaningful and our solution is effective. To the best of our knowledge, this is the first work aimed at exploiting such a large semantic space for emotion representation in the FER problem.

Index Terms—Facial emotion recognition, affective computing, image analysis

1 Introduction

Understanding and recognizing human facial emotional expressions has been an attractive research topic for decades, lying in the intersection area of affective science and human-computer interaction. Despite the natural perception ability that humans obtained from evolution [1], it is never straightforward for computer-based systems to sense and interpret emotions from human facial performances automatically. On one side, the challenge of facial emotion recognition (FER) problem partially comes from the sophisticated facial muscle system, leading to complicated facial behaviors w.r.t. individual's emotional statements, especially under the in-the-wild uncontrolled conditions. On the other side, most of the current FER researches only focus on the abstract level of emotion concepts, but are struggling to cover the entire emotion space [2] sufficiently.

Typically, the categorical model is one of the most popular representations in the FER area, composed of several basic emotion classes, e.g., *happiness*, *anger* and *surprise*. Depending on the psychological conceptualization on specific natural emotions, multiple emotion theorists suggest a variety of category lists individually [3], [4], [5], [6]. However, due to the highly abstract manner of such definitions, there are some arguable ambiguities. For example, given two individual emotion terms, *amazement* and *astonishment*, which are both subject to the *surprise* class [7], their triggered facial expressions are obviously different as *amazement* is rather positive and close to *happiness* while *astonishment* is more negative and associated with *fear* (Fig. 1). Therefore, simply categorizing the various facial expressions into several abstract classes is incapable of representing the numerous and fine-grained emotional statements.

To tackle this issue, several annotated FER datasets are proposed by mixing the basic expressions into compound ones [8], [9], [10], replacing the discrete representations with multi-label distributions [11], [12], or enlarging the emotion sets with a few more classes [13], [14]. Besides, another category of methods follows the circumplex emotion modeling idea [15], whose dimensions are represented by the principle emotion factors, i.e., *valence*, *arousal*, *dominance*, etc. The shortcoming of the dimensional model comes from its difficulty of annotating accurate continuous labels, such as [16], [17]. Nevertheless, the semantic richness issue of recognizable emotion concepts still remains an open problem, which is really challenging to the whole FER community.

In this paper, we aim at studying the FER problem on a semantic-rich level. Different from the previous methods that simply blend or add more emotion classes to enhance the FER quality, we thoroughly exploit the linguistic space and leverage a reasonable lexicon to describe the emotion

- *Keyu Chen, Changjie Fan, Wei Zhang, and Yu Ding are with Netease Fuxi AI Lab, Beijing 100084, China. E-mail: chern9511@gmail.com, {fanchangjie, zhangwei05, dingyu01}@corp.netease.com.*
- *Xu Yang is with PALM Lab, Department of Computer Science, Southeast University (SEU), Nanjing, Jiangsu 211189, China. E-mail: xuyangseu@ieee.org.*

Manuscript received 22 February 2022; revised 19 July 2022; accepted 17 August 2022. Date of publication 24 August 2022; date of current version 15 November 2022.
This work was supported in part by the Key Research and Development Program of Zhejiang Province under Grant 2022C01011, and in part by the Hangzhou Science and Technology Office through the 2022 Key Artificial Intelligence Science and Technology Innovation Project.
(Corresponding author: Keyu Chen.)
Recommended for acceptance by S. Wang.
This article has supplementary downloadable material available at https://doi.org/10.1109/TAFFC.2022.3201290, provided by the authors.
Digital Object Identifier no. 10.1109/TAFFC.2022.3201290

Fig. 1. Facial image samples belong to the same category (*surprise*) defined by the basic emotion model but with contrastive emotional expressions (*amazement* and *astonishment*). The obviously different facial performances indicate the necessity of proposing more fine-grained representation model to handle the abundant emotion semantics.

concepts. Inspired by previous psychological research [7], we extend the recognizable emotions to an exhaustive set, covering 135 English words which can semantically describe most of all distinctive emotional feelings or inner statements of humankind. From the perspective of psychological and linguistic research, the 135 words expand an *almost* complete semantic atlas of the emotion domain [18], [19]. Accordingly, we argue that the 135-class emotion model is desirable for semantic-rich FER research.

Based on the 135-class emotion model, we construct a large-scale FER dataset in a labor-free manner. First, we use the 135 emotion terms as class labels, collect more than one million web images from the internet. Then, we design an automatic data cleaning process by efficiently evaluating the expression consistency of the collected images. To evaluate the label credibility of our categorical dataset, we set up a manual verification test in which multiple participants are required to give their judgments on given images and different emotion labels. In this way, we successfully build up the *Emo135* dataset, which contains 135 emotion categories and 728,946 facial images in total.

Next, we propose a baseline method to validate the feasibility of conducting FER on the semantic-rich representation. Considering the number of emotions to be recognized, there inevitably exist synonyms among the 135 emotion concepts/terms, making it neither reasonable nor possible to regard these categories as individual sets. Our corresponding solution is to evaluate the cross-label correlations via two metrics, i.e., computing the word embedding and facial expression embedding similarity distances. The similarity scores are then transformed into two weight matrices for storing the correlations among 135 emotion classes. Finally, we make the weight matrices as prior knowledge and inject them into the recognition network training softly.

To the best of our knowledge, this is the first work aimed at handling the FER problem with such a large number of emotion categories. The psychological backing of the utilized 135 emotion concepts makes adequate support on our claimed *semantic richness* of the FER problem. In sum, the contributions of this research are three-fold:

- We propose the first semantic-rich facial emotional expression recognition work, with an exhaustive emotion set including 135 concepts comprehensively described the entire emotion domain.
- We automatically construct a large-scale FER dataset *Emo135*, containing 135 fine-grained emotion categories and 728,946 facial images. We believe the open-released dataset could benefit the other research works in the FER community.
- We carefully design a correlation-guided method for fine-grained facial emotional expression recognition. The quantitative and qualitative experiment results indicate that our method can well handle the complicated nature of so many emotions and generate reliable FER predictions with rich semantics.

2 Related Work

This section briefly reviews some related literature to our work, including facial emotion expression representations, datasets, and automatic recognition methods.

2.1 FER Representation and Dataset

Facial emotional expression embodies non-verbal communication of our daily life. In order to technically model the inner emotion statements that are conveyed by facial expressions, there are three common used emotion representations being proposed, including the categorical model [6], the action unit model [20]), and the circumplex model [15]. Among these models, the categorical one consisting of several basic emotion terms is most popular. Typically, it is defined by seven or eight universal recognizable emotions, namely *neutral*, *anger*, *disgust*, *fear*, *happiness*, *sadness*, *surprise*, *contempt*, etc. As a matter of fact, most current FER datasets are built upon these discrete categories, varying on the specific definition of emotion concepts, such as *JAFFE* [21], *CK+* [22], *KDEF* [23],*SFEW* [24], *FER2013* [25], *FER-Wild* [26], *AffectNet* [16], and *Aff-Wild2* [27].

However, until recent years, the basic categorical model has been challenged for its incapability of modeling fine-grained emotion variances. The following researches suggest improving the representation capacity of the emotion model, for example, introducing compound emotion classes [8] and transferring the discrete emotion labels into continuous distributions [11]. Based on these idea, some novel FER datasets are proposed, like *RAF-DB* [10] and *EmotioNet* [9] which includes 18 and 23 basic/compound emotion classes respectively, and *RAF-ML* [12] with continuous label distribution annotations. Furthermore, the latest research work tries to extend the emotion concepts to 54 classes and proposes a corresponding dataset F^2ED [14].

2.2 Facial Expression Features and Classifiers

Image-based facial emotion recognition has been extensively studied for decades. In general, a complete FER method is composed of two algorithm modules, i.e., feature extractor and classifier. Traditional FER approaches usually apply hand-crafted features, such as Gabor Wavelets [28], Local Binary Patterns [29], and Histogram of Oriented Gradients [30]. With the rapid development of deep learning techniques, some pre-trained backbones like ResNet [31] are adopted for extracting high-level features. In terms of the specificity of FER tasks, there are also some expression

embedding models [32], [33] which can eliminate the invariant attributes like pose, identity, and image background from the captured features.

On the other hand, the classifiers integrated into the FER systems have also achieved promising performances in recent years. To solve the occlusion issue caused by large poses, the region-based network [34] is proposed with an attention mechanism. Besides, there are some FER methods considering the facial priors, such as the muscle moving masks [35] and the geodesic distance on 3D shapes [36]. Except for the extensive methods [37], [38], [39], [40], [41] which focus on solving the FER problem independently, there is also another category of methods trying to explore the benefits from multi-task settings [42], [43].

3 Methodology

In this section, we first review the background of our leveraged emotion model, which contains 135 lexicon terms representing the semantic atlas of the emotion domain. Then we introduce the data acquisition and processing details that help us construct a large-scale facial image dataset. Finally, we propose a baseline approach for fine-grained facial emotion recognition by considering the cross-label relationships among the multiple emotions.

3.1 Semantic-Rich Emotion Categories

Emotion knowledge plays an important role in social interaction. Without too much training, even infants can naturally perceive and express emotional feelings at a basic level, e.g., *happiness*, *fear*, and *anger*. Although numerous empirical cognitive studies have demonstrated that ordinary people can reliably name the emotions being expressed from facial images [44], it has been a struggle for psychologists to agree on a formal semantic structure of the human emotion space. To efficiently associate the cognitive emotions with linguistic descriptions, some emotion theorists suggest applying the prototype approach [45] to determine the emotion concepts with a finite set of words [3].

Depending on the definition of emotion varieties, there are different kinds of emotion taxonomies, i.e., emotion representations consisting of different sets of semantic terms (words with specific emotion meanings). Some research works focus on the abstract level of emotion episodes, claiming several universal categories forming the overall structure of the emotion space, such as the seven or eight basic emotions [6].

Another branch of methods digs into the language space to search for every distinctive word representing a particular emotion concept. It is first proposed by Averill who collects 558 English words conveying emotion connotations [46]. Then the 558 words are further cleaned up by grammar roots and evaluated with the *emotion-sorting study* [7]: one hundred twelve participants make their judgments on each word, with a prototypical rating from *"I definitely would call this an emotion."* to *"I definitely would not call this an emotion."* [7]. Finally, there are 135 words left with high enough ratings, and it is responsible for saying that they extensively form the semantic space of the emotion domain. Therefore, in this paper, we refer to the 135 emotion words/categories as the semantic-rich emotion representation. A full list of the 135 words given by Shaver et al. is transcribed in the appendix section.

Fig. 2. Facial image samples of six categories within the *Emo135* dataset. All images are collected from four photo stock websites[1].

We would like to point out that, given the human language is a living entity, some recent studies [47], [48], [49] propose a variety of emotion classes that defend/revise the Shaver's model [7]. Nevertheless, as language research goes on, the ideal emotion semantic atlas shall be updated as well.

3.2 *Emo135* Dataset

With the semantic-rich emotion representation, we establish a facial image dataset *Emo135* according to the 135 emotion words. Technically, our dataset construction process involves two automatic steps, data collection and cleaning.

Data Acquisition. We use the 135 emotion category names as keywords, accompanying several other suffix words such as *expression*, *feeling*, and *face*, to query for web images with matching titles by internet search engine indexing. While downloading the valid images from the internet, we also apply face detection by dlib library[2] and crop the face area from the entire image into 224×224 size. In this way, we collect 135 image categories with more than one million facial expression images. An illustration of our data collection results is shown in Fig. 2.

Data Cleaning. In order to eliminate the noisy samples of each emotion category (which could be titled or indexed with wrong words), we design a data post-processing strategy to clean the collected image dataset introduced above. The basic idea of our strategy is to identify the anomaly images if their facial expressions are different from the majority of the belonging class.

Specifically, we adopt an advanced facial expression embedding model [33] for expression similarity evaluation. The advantage of the embedding model is that it can produce expression embedding codes that are invariant to the other facial attributes like identities, poses, or image backgrounds. Even more, it can capture subtle expression variations between different faces. For example, if two faces have similar expressions, they will be mapped closely in the latent space, and vice versa (See Fig. 4). Within that latent space, we gather the expression embedding codes of all

1. https://www.{bigstockphoto;alamy;photocase;shutterstock}.com
2. http://dlib.net

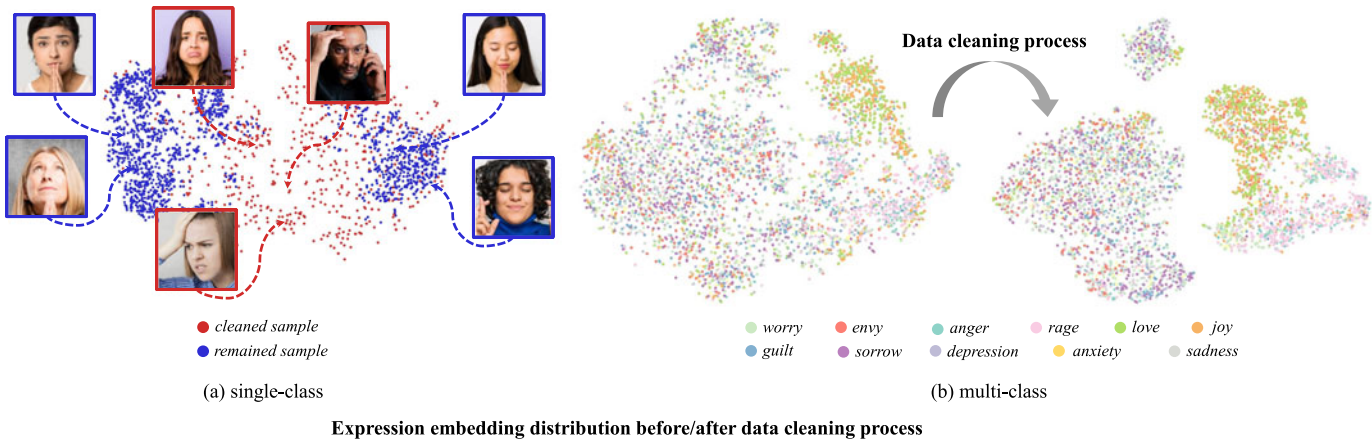

Fig. 3. Illustration of single-class and multi-calss expression embedding distributions before/after the automatic data cleaning procedure.

images in the same category and evaluate the embedding density of these codes. Based on the *k-nearest neighbors* algorithm, we can efficiently detect the embedding outliers by filtering the mean distance between each sample and its K nearest neighbors with a predefined threshold. The detailed algorithm steps are described in Algorithm 1. In practice, the expression embedding codes are lying on a 16-dimensional unit sphere.

While we employing the knn-based data cleaning method, we find it is quite sensitive to the specified hyper-parameters, i.e., the neighborhood size K and the threshold distance σ. Generally, larger neighborhood size and smaller threshold distance will encourage more strict elimination policy and thus reduce the dataset size (eliminating even matched images), and vice versa. To preserve the label quality as well as the image quantities, we empirically choose $K = 10$ and the threshold distance $\sigma = 0.2$.

Algorithm 1. Data Cleaning Process

Input: Image category $\mathcal{C} = \{\mathbf{I}_n\}_{n=1}^{N}$; Neighbouring number K; Filter threshold σ;
Output: Remove/keep decision on each image sample; **Step:**
1: Generate expression embedding of each image,
$$E : \mathbf{I}_n \mapsto \mathbf{V}_n ,$$
$$E(\mathcal{C}) = \{\mathbf{V}_n\}_{n=1}^{N} .$$
2: Find the K nearest neighbors for each embedding vector,
$$KNN(\mathbf{V}_n, E(\mathcal{C})) = \{\mathbf{V}_{n_k}\}_{k=1}^{K} .$$
3: Calculate the mean distance between $\mathbf{V}_n$ and its K nearest neighbors,
$$d_n = \frac{1}{K}\sum_{k=1}^{K} \|\mathbf{V}_n - \mathbf{V}_{n_k}\|^2 .$$
4: Compare d_n with σ,
$d_n < \sigma$: keep image $\mathbf{I}_n$ in $\mathcal{C}$,
$d_n \geq \sigma$: remove image $\mathbf{I}_n$ from $\mathcal{C}$.
End

In Fig. 3, we visualize the expression embedding distributions of the single class *hope* before/after the data clean process. It can be observed that our designed approach is significantly helpful in terms of reducing data noise and improving label accuracy to the dataset. Besides, we also visualize the multi-class expression embedding distributions before/after the automatic data cleaning procedure in Fig. 3. After the removal operation, each emotion class is more compact, which makes it possible for us to analyze their correlations. In sum, 31% of original data is automatically cleaned during this process, leaving in total of 728,946 valid facial images.

Dataset Statistics. In Table 1, we compare our proposed *Emo135* dataset with some other existed FER datasets. It can be observed that our dataset *Emo135* has the most fine-grained annotations in terms of emotion classes and comparable large quantities of facial images.

Besides, we also illustrate the image quantity distribution per each emotion category in Fig. 5. Considering the different emotions may involve different degree of presence in our daily life, some *rare* emotion classes (e.g., *vengefulness*) generally contain less image samples than those *common* ones (e.g., *excitement*). Therefore, our dataset distribution is not absolutely uniform but including the maximum category with 12,794 images and the minimum category with 994 images.

3.3 Modeling Correlation Matrix for 135 Emotions

After obtaining the *Emo135* dataset, we propose to analyze the cross-emotion similarities for the 135 emotions. Different from the previous FER methods established on only a few discrete emotion classes, the problem setting of this work is more challenging since there are as many as 135 emotion categories, and most of them do not have sharp

TABLE 1
List of Some Existed FER Datasets and the Associated Characteristics

Dataset	Annotation	#Image	In-the-wild
JAFFE	7 basic class	213	✗
CK+	8 basic class	593	✗
KDEF	7 basic class	4,900	✗
SFEW	7 basic class	700	✓
FER2013	7 basic class	35,887	✓
FER-Wild	7 basic class	24,000	✓
AffectNet	8 basic class	450,000	✓
Aff-Wild2	7 basic class	2,800,000	✓
RAF-ML	7-class distribution	4,908	✓
RAF-DB	18 basic/compound class	29,672	✓
EmotioNet	23 basic/compound class	1,000,000	✓
F^2ED	54 fine-grained class	219,719	✗
Emo135	135 fine-grained class	728,946	✓

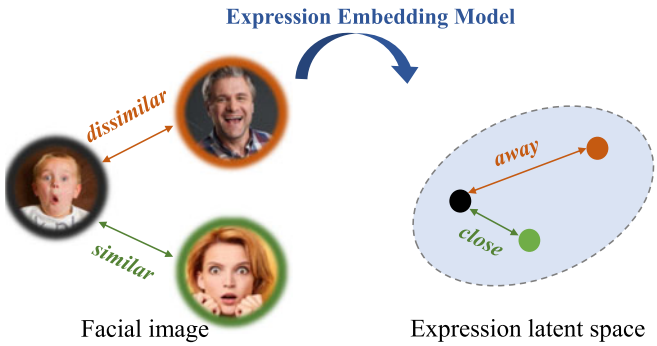

Fig. 4. Demonstration of our adopted facial expression embedding model. By capturing the expression-related features, the embedding model is capable of mapping similar expression images closely in the latent space, while enforcing the dissimilar ones away from each other.

boundaries, which means it is even impossible to absolutely distinguish any emotion term from the others. This phenomenon suggests we have to carefully consider the correlations of different classes. The idea of modeling cross-emotion distances can date back to the 1980s [15], [50]. Until recently, the emotion correlations are considered in several automatic facial affective recognition methods [51], [52], [53]. In this task, the similarity of different emotions can be evaluated in two ways, one based on their triggered facial expressions and the other by the semantic word embeddings. Specifically, the facial expression evaluations focus on the performance/expressiveness level of different emotions, which is useful for extracting solid facial image features. In contrast, the semantic word embeddings are adopted from the language modeling area, capable of revealing the intrinsic synonymous distances between different emotion labels. Thus we can make them facilitate the emotion label prediction process.

To begin with, let us denote the 135 emotion categories as $\mathcal{C}_1, \mathcal{C}_2, ..., \mathcal{C}_{135}$. For each category $\mathcal{C}_k (k = 1, 2, ..., 135)$, it contains in total of N_k facial images $\mathbf{I}_k^n (n = 1, 2, ..., N_k)$. Because of some inevitable synonyms like *anger* and *fury* existed in the 135 emotion classes, and even words in hierarchical relationship like *astonishment* and *amazement* which are both subject to *surprise*, we are motivated to quantitatively calculate the distances between different emotions and moreover apply this knowledge to help our network training.

Facial Expression Similarities. We first adopt the facial expression embedding model [33] to evaluate the facial images between different emotion categories. Specifically, we first send every image $\mathbf{I}_k^n$ into the pre-trained facial expression embedding model and generate the corresponding expression embedding vector $\mathbf{V}_k^n \in \mathbb{R}^{16}$.

With the expression similarity structure of the embedding space, we are now able to evaluate the cross-emotion relationships and model the distances among 135 classes. For emotion category $\mathcal{C}_i = \{\mathbf{I}_i^n | n = 1, ..., N_i\}$ and $\mathcal{C}_j = \{\mathbf{I}_j^m | m = 1, ..., N_j\}$, their corresponding embedding vectors are given as $E(\mathcal{C}_i) = \{\mathbf{V}_i^n | n = 1, ..., N_i\}$ and $E(\mathcal{C}_j) = \{\mathbf{V}_j^m | m = 1, ..., N_j\}$. We utilize *Directed Hausdorff Distance* to measure the one-sided similarity from $\mathcal{C}_i$ to $\mathcal{C}_j$, which can be formulated as following:

$$d_H(\mathcal{C}_i, \mathcal{C}_j) = \max_n \min_m \|\mathbf{V}_i^n - \mathbf{V}_j^m\|_2^2. \tag{1}$$

Notably, this metric is asymmetric as $d_H(\mathcal{C}_i, \mathcal{C}_j)$ does not necessarily equal to $d_H(\mathcal{C}_j, \mathcal{C}_i)$, and thus it is suitable for the similarity modeling purpose. This is because sometimes the emotion $\mathcal{C}_i$ could be absolutely recognized as $\mathcal{C}_j$ (e.g., *amazement→surprise*) but the inverse is not true (e.g., *suprise→{amazement, astonishment, etc.}*).

Next, we pack the calculated results into a facial expression similarity matrix $\mathcal{F}^{exp} \in \mathbb{R}^{135 \times 135}$, in which each

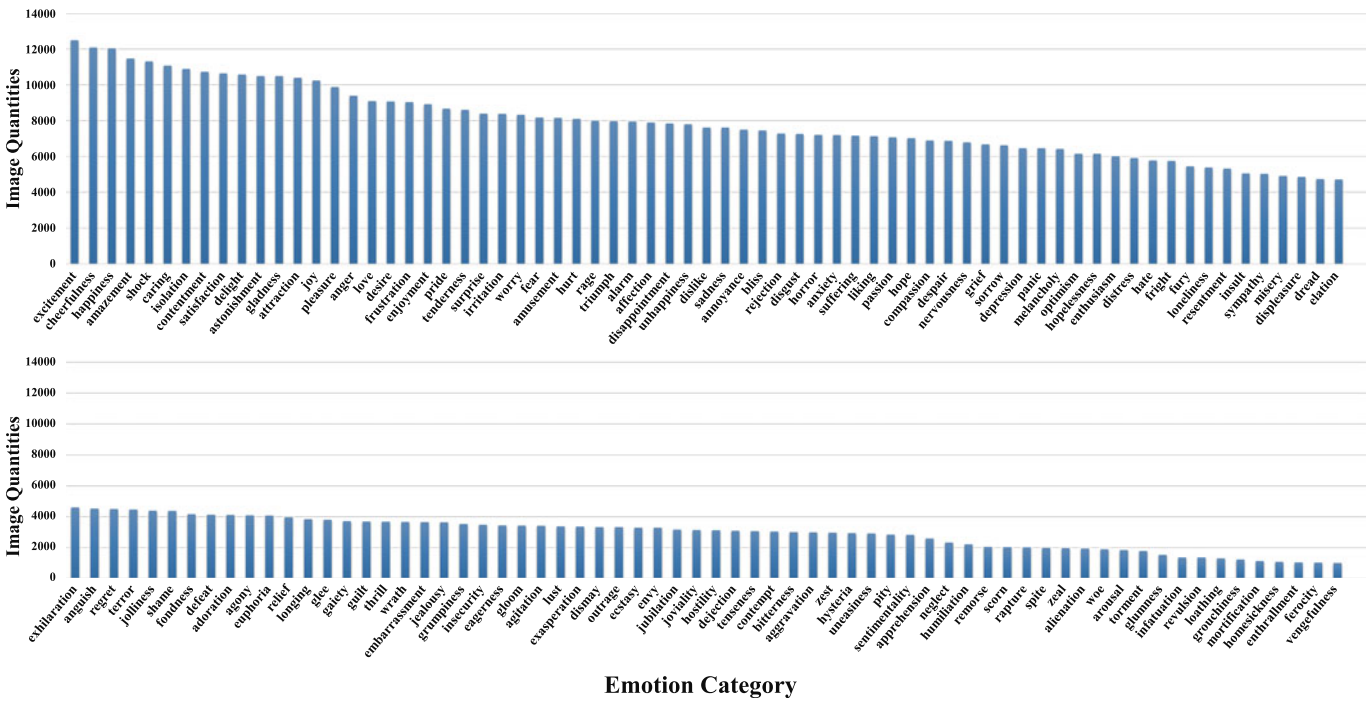

Fig. 5. Distribution of image quantities from each emotion category in our constructed *Emo135* dataset.

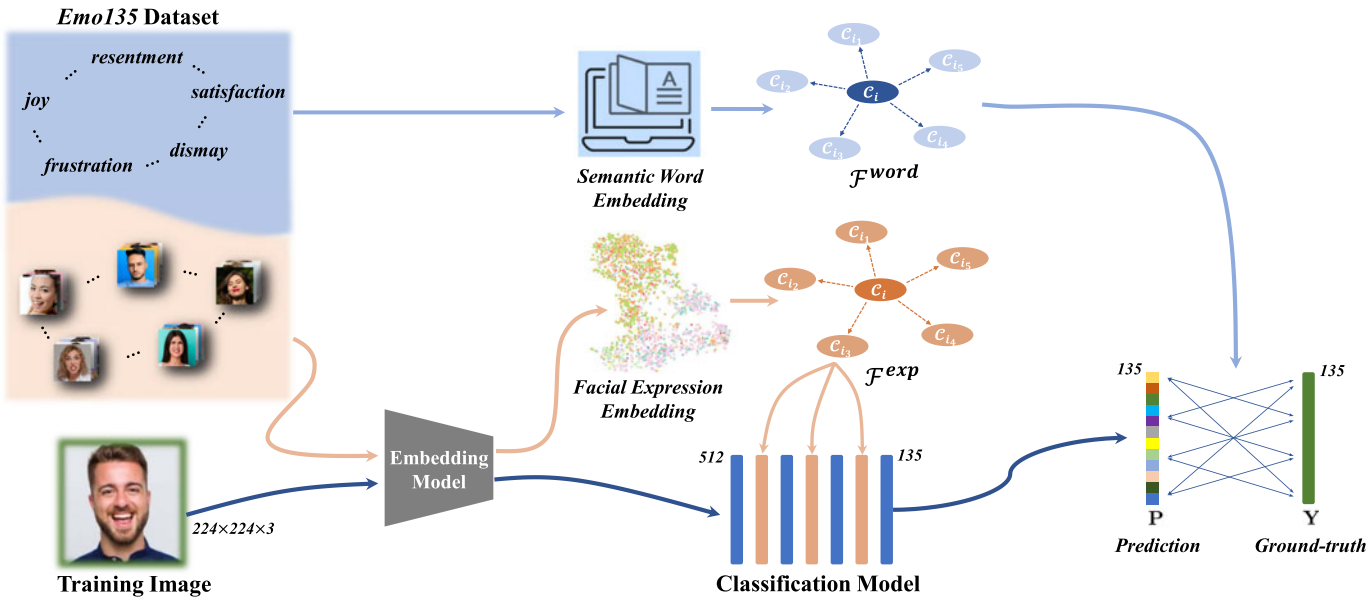

Fig. 6. The pipeline of our proposed baseline approach for fine-grained facial emotion recognition on the *Emo135* dataset.

element $\mathcal{F}_{ij}^{exp}$ is given as:

$$\mathcal{F}_{ij}^{exp} = \frac{1}{[d_H(\mathcal{C}_i, \mathcal{C}_j)]^2}(i \neq j). \tag{2}$$

Particularly, the diagonal element $\mathcal{F}_{ii}^{exp}$ is computed by adding the rest entries within the ith row like:

$$\mathcal{F}_{ii}^{exp} = \sum_{j \neq i} \mathcal{F}_{ij}^{exp}, i \in \{1, 2, ..., 135\}. \tag{3}$$

Finally, we normalize the correlation matrix along each row into sum 1.0. The technical meaning of our constructed matrix is that, if emotion $\mathcal{C}_i$ is very close to emotion $\mathcal{C}_j$, the value of element $\mathcal{F}_{ij}^{exp}$ would be significantly large. In this way, we can use $\mathcal{F}^{exp}$ to efficiently guide the recognition process to be aware of the cross-emotion relationships.

Word Embedding Distances. The other metric we adopted for evaluating cross-emotion similarities is the semantic word embedding. First, we adopt a *Word2Vec* model [54], [55] which is pre-trained on large-scale dataset including English blogs, texts, and comments. Then the model will take input as every pair of 135 emotion terms and output the corresponding embedding distance of each. Similar to the facial expression similarity matrix, the word embedding distances are stored in a coefficient matrix $\mathcal{F}^{word}$. For every pair of emotion categories $(\mathcal{C}_i, \mathcal{C}_j)$, the word embedding distance is calculated as follows and then normalized in rows as well:

$$\mathcal{F}_{ij}^{word} = Word2Vec(\mathcal{C}_i, \mathcal{C}_j)(i \neq j). \tag{4}$$

$$\mathcal{F}_{ii}^{word} = \sum_{j \neq i} \mathcal{F}_{ij}^{word}, i \in \{1, 2, ..., 135\}. \tag{5}$$

In practice, we deem the word embedding distances as label correlations. For example, if emotion $\mathcal{C}_i$ and $\mathcal{C}_j$ are close in the word embedding space, then when an image is classified into class $\mathcal{C}_i$ with a high probability, it should also possess a similarly high probability within class $\mathcal{C}_j$.

3.4 Baseline Approach

Network Design. Our proposed fine-grained emotion recognition network includes two modules, a pre-trained facial expression embedding model and a correlation-guided classification model (Fig. 6). The expression embedding model is responsible for extracting expression-related features, and the classification model aims to regress the target emotion distributions with the help of the calculated correlation matrix $\mathcal{F}^{exp}$. Notably, the pre-trained facial expression embedding model [33] is trained for fine-grained expression similarities. The model incorporates many expression triplet data and learns a continuous expression embedding space. It is capable of capturing minor expression similarities and thus suitable for building the 135-class representations. We also illustrate the detailed network structures in the appendix section.

Given a training image $\mathbf{I} \in \mathbb{R}^{224\times224\times3}$ and its one-hot ground-truth vector $\mathbf{Y} \in \mathbb{R}^{135\times1}$, we first input the image into the expression embedding model for feature extraction. The expression feature will then be sent into the classification model consisting of several alternative fully connected and correlation layers. In particular, we implement the three correlation layers by initializing them with $\mathcal{F}^{exp}$. The detailed network design can be found in the appendix.

Loss Function. The final output of our model is a prediction vector $\mathbf{P}$ in size of $\mathbb{R}^{135\times1}$. Before comparing $\mathbf{P}$ with the one-hot ground-truth label $\mathbf{Y}$, we take the word embedding correlation matrix $\mathcal{F}^{word}$ into consideration by applying the transformed cross entropy (TCE) loss:

$$\mathcal{L}_{TCE} = -[(\mathcal{F}^{word} \cdot \mathbf{Y}) \log \mathbf{P} + (1 - \mathcal{F}^{word} \cdot \mathbf{Y}) \log (1 - \mathbf{P}))]. \tag{6}$$

It is worth noting that despite the formulation of Eq. 6 is similar to the standard *label smoothing* strategy [56], they are different in terms of relaxing weights. In the *label smoothing* operation, each zero-value term within the ground-truth $\mathbf{Y}$ is uniformly modified with the same soft parameter, e.g.,

$\alpha = 0.1$. While in our method, the label relaxing is dependent on the emotion word embedding analysis. Therefore our produced correlation label $\mathcal{F}^{word} \cdot \mathbf{Y}$ can better satisfy the emotion nature and enhance the recognition process with the prior semantic knowledge.

4 Experiment

In this section, we first give some implementation details about our experiments. Then we report the subjective survey results on evaluating the *Emo135* dataset. Finally, we compare our proposed baseline framework with other feasible approaches and prove the efficiency of our method.

4.1 Implementation Detail

We randomly split the *Emo135* dataset into training, validation, and testing set by $70\%, 15\%, 15\%$, respectively. There are in total of 510,262 images for training and 109,342 for validation/testing. We implement our training framework based on Pytorch[3]. The training costs around 30 hours on an NVIDIA RTX 3090 graphics card of 24 GB memory, with a learning rate of 0.005 and batch size 240. We use a stochastic gradient (SGD) optimizer for optimization and train the entire framework for 100 epochs.

4.2 Dataset Evaluation

To ensure the image emotion labels are convincing and credible, it is necessary to conduct a manual evaluation on the *Emo135* dataset. Therefore, we make a subjective survey by recruiting 62 participants to validate the semantic correspondence of our collected facial images and their emotion labels. Specifically, we offer the participants three rating choices including *"I agree the given word faithfully conveys the facial emotions"*, *"I prefer another similar word to describe the facial emotion"*, and *"I prefer another dissimilar word to describe the facial emotion"*, respectively standing for different accuracy levels of the emotion terms.

Considering the large size of our constructed dataset (roughly 700k images), we choose to carry out the aforementioned manual evaluation by sampling the whole *Emo135* dataset. We blindly select examples from each category according to its size. For instance, we extract 198 images from the largest category (containing 12,794 images) and 16 images from the smallest category (containing 994 images). In return, we receive 33,651 ratings on 11,217 images, in which each image and its label are evaluated by three different raters.

As the histograms shown in Fig. 7, the image labeling results are, in most cases, in agreement with the common sense of human affective cognition. On average, 81.2% raters agree that the given emotion labels perfectly match the corresponding images. Besides, there are 12.9% raters suggesting similar words for description, while 5.9% disagree with the given labels and suggest something different. Among the 135 emotion categories, *pride* gets the lowest satisfying rate at 41.8% from raters. The other less satisfying ($< 50\%$) classes are *annoyance*, *humiliation*, and *suffering*. In contrast, there are relatively more consistent categories,

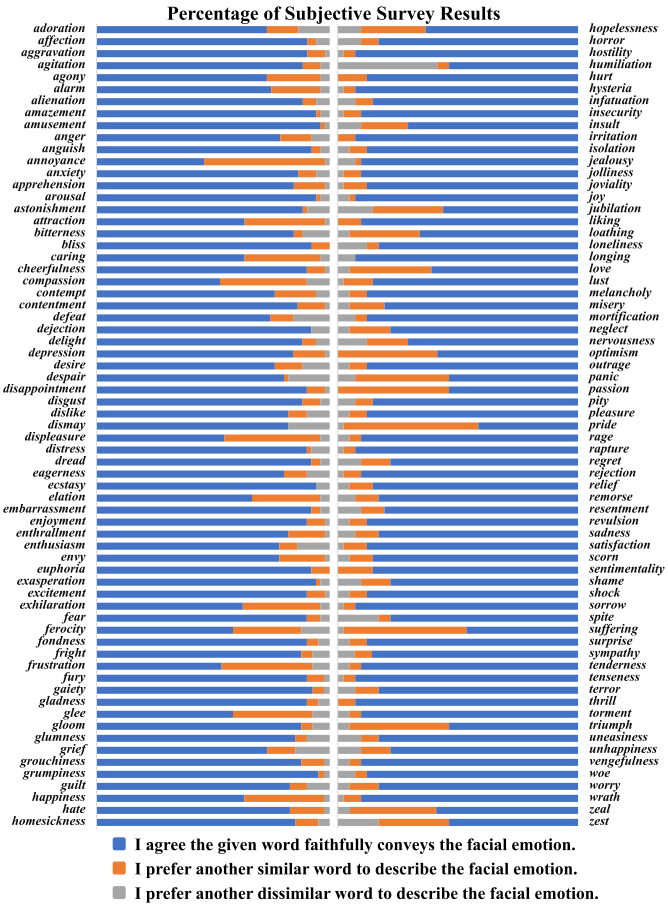

Fig. 7. Histogram plot of the collected subjective rating results. We show the percentage of the three judgements that each category receives, arranging from "*absolutely agree*" to "*disagree*".

such as *amusement* (92.6%), *amazement* (91.1%), *grumpiness* (91.1%) and *sorrow* (89.6%).

Furthermore, to demonstrate the test validity of the rating experiment, we also evaluate the inter-rater reliability by *Krippendorff's alpha* test [57]. Following the open-source code[4] for *Krippendorff's alpha* calculation, we compute the α efficient of our collected results, with a score of 0.776 ($\alpha = 1.0$ indicates perfect reliability, $\alpha = 0.0$ indicates the absence of reliability, and $\alpha < 0$ means systematically disagreement). In conclusion, the statistical results of the subjective survey indicate that our adopted 135 emotion concepts, as well as the corresponding facial images, are compatible with the human emotional cognition knowledge. Nevertheless, it is worth noting that the above conclusion comes from sample survey. The annotation cost limits us to conduct a full survey at the current stage. We believe it would be meaningful to increase the manual evaluation scale in the future.

4.3 Model Evaluation

To evaluate our proposed model on recognizing facial emotional expressions, we conduct both performance comparison and ablative study. Since this is the first work trying to handle the 135 emotion classification problem, there is no

3. https://pytorch.org/

4. https://github.com/grrrr/krippendorff-alpha

TABLE 2
Quantitative Comparison Between our Method and the Other Approaches, Including Pre-Trained Backbones, Modified SOTAs, and Ablative Models

Approach		F1-Score	Top-1 Acc.	Top-5 Acc.	Top-10 Acc.
Pre-trained Model	ResNet50 + MLP	0.061	0.096	0.306	0.454
	VGGFace2 + MLP	0.062	0.099	0.322	0.465
Modified SOTA	ARM [58]	0.147	0.205	0.559	0.708
	DACL [59]	0.133	0.183	0.477	0.667
Ablative Study	Ours w/o Embedding	0.082	0.126	0.458	0.564
	Ours w/o Correlation Layers	0.175	0.247	0.604	0.735
	Ours w/o Correlation Label	0.219	0.272	0.605	0.710
	Ours	**0.247**	**0.283**	**0.664**	**0.787**

existing method for direct comparison. Therefore, we choose two kinds of competitive models and train them on the *Emo135* dataset. First, we adopt two commonly used pre-trained backbones, ResNet-50 and VGGFace2, assembled with Multilayer Perceptron (MLP) classifiers. Second, we adopt two latest FER models, ARM [58] and DACL [59], which have achieved state-of-the-art performance on basic emotion recognition tasks, and modify their output layers for 135-class recognition.

ResNet-50 Baseline. As a popular image pre-training model in computer vision community, ResNet-50 [31] has achieved significant performance in a wide range of applications, especially on the image classification/recognition topic. Therefore we integrate the most recent released ResNet-50 model trained on ImageNet-1k [60] dataset as backbone and a 5-layer MLP to regress the image features to 135 dimensional logits.

VGGFace2 Baseline. Compared with ResNet-50 [31], VGGFace2 [61] is trained on specific human face images and gained even better performances in several human face centric applications, e.g., facial landmark detection, re-identification, and facial expression recognition. We also design a baseline consisting of a pre-trained VGGFace2 model [61] and the same MLP layers as the ResNet-50 baseline.

ARM [58]. ARM is one of the state-of-the-arts reaching impressive scores on the public benchmarks for 7-class discrete facial expression recognition. It introduces an auxiliary block for feature map rearrangement and enhances the de-albino effect. Moreover, a minimal random re-sampling scheme is also introduced to solve the data unbalancing issue. To fairly compare our model with ARM [58], we modify its regression module (final layer dimension) to make it compatible with 135-class FER.

DACL [59]. DACL also reaches comparable good performance in public FER benchmarks. The core idea includes a novel sparse center loss design and an attention mechanism to weight the contribution of metric learning loss functions. Similar as ARM [58], we adopt the main structure of DACL [59] including the attention net and the sparse center loss calculation module but change the target output expression dimension to 135.

Ablative Study. Besides, we also conduct ablative studies to evaluate some key component including the expression embedding model, correlation layer, and label transformation loss of our framework. For those components, we provide vanilla alternatives to evaluate the effectiveness of our design. For example, the facial expression embedding model is replaced with VGGFace2 [61], the correlation layer is compared with fully-connected MLP, and the label transformation loss is changed to cross entropy loss.

In Table 2, we compare the prediction results from each method on the test set, including F1 score and accuracy for the top 1, 5, 10 classes. It can be observed that our approach reaches the best performance of all the others, with top-1 prediction accuracy at 28.3%, top-5 accuracy at 66.4%, and top-10 accuracy at 78.7%.

4.4 Semantic Evaluation

To evaluate the semantic relationships between the emotion labels and the predicted results, we design several experiments in this section. First, we compare different word embedding models in terms of their influences on the semantic similarity distances. By our problem setting, we choose three popular pre-trained word embedding model, including *Word2Vec* [54], *GloVe* [55], and *BERT* [62], to study the corresponding correlations between emotion word semantics. We use the original prototypical rating results [7] as reference and calculate the Pearson Correlation Coefficients (PCC) between each model's output and the original matrix (Table 3). The results shows that *Word2Vec* and *GloVe* are both capable of extracting the true semantic relationships for emotion words, while *BERT* performs poor on it. The reason could be that *BERT* is not design for specific word-level but contextual embedding.

Then we show some example testing results in Fig. 8. It is interesting to find that, even if the actual label is not recognized as the top one, our model can still produce reasonable predictions that are semantically close to the ground-truth (in red). Furthermore, we employ the chosen word embedding model [54], [55] to calculate the semantic distances between the ground-truth label and our predictions/the rest of 135 emotion terms (See Table 4). This phenomenon

TABLE 3
Pearson Correlation Coefficient (PCC) Between Word Embedding Model Output and Original Emotion Rating Matrix [7]

PCC ↑	Word2Vec [54]	GloVe [55]	BERT [62]
Original rating	**0.532**	**0.515**	**−0.412**

Positive means correlated and negative means uncorrelated.

Fig. 8. Illustration of the top-10 results predicted by our method. The ground-truth label is indicated in red.

TABLE 4
Semantic Distances Between Ground-Truth Label and our Predictions versus Ground-Truth Label and the Rest of 135 Emotion Terms, the Smaller the Better

Word embedding distances ↓	Top-1	Top-5	Top-10
Ground-truth ↔ Our predictions	**0.461**	**0.433**	**0.412**
Ground-truth ↔ The rest of 135 terms	0.639	0.637	0.635

suggests that our analyzed semantic emotion relationship is helpful and reliable to improve the semantic richness of the FER results.

5 Limitation and Discussion

Despite the fact that we have evaluated the emotion labeling results of *Emo135* dataset by conducting the subjective survey in the experiment, there still remains a lot of potential improvements in the future. For example, it would be meaningful to apply manual verification on the full dataset, i.e., making multi-person vote for the 135 emotion labels on every facial expression image. Regarding the complicated nature of human facial emotions, it is also necessary to enlarge the FER dataset, such as adding more subjects and image conditions, to support more robust research in this area.

Besides, we would like to point out that the 135-categorical emotion representation, which stands for the *semantic richness* in this paper, is not a fixed standard. With continuing innovative research works in the psycholinguistic field, the semantic definition of human emotion concepts is also changeable. In the future, if any more dedicated emotion categorical model is proposed, the basic idea and the technical approach of this work can be adapted to the new one.

6 Conclusion

In this work, we address the semantic-rich facial emotional expression recognition problem. Unlike the existing FER researches that only focus on a few basic emotion categories, we aim at the granularity of emotion concepts and the entire emotion space. To this end, we construct a novel FER dataset by leveraging a 135-class categorical model which can exhaustively represent the semantic atlas for the emotion domain. We further propose a baseline approach for the emotion recognition task on our built dataset. The core idea of our method is to model the fuzzy relationships between fine-grained emotions and then make it guide the network training process. We conduct thorough evaluations on both the dataset labeling quality and the baseline recognition method. The quantitative and qualitative results suggest the benefits of pushing FER to a semantic-rich level. In the future, we believe it would be meaningful to propose more dedicated methods and large-scale datasets to promote the understanding and analysis of fine-grained facial emotions.

References

[1] C. Darwin, *The Expression of the Emotions in Man and Animals*. Chicago, IL, USA: Univ. Chicago Press, 2015.
[2] R. Plutchik, *The Emotions*. Lanham, MD, USA: Univ. Press Amer., 1991.
[3] S. Epstein, "Controversial issues in emotion theory," in *Review of Personality and Social Psychology*, Beverly Hills, CA, USA: Sage Publications, Inc., vol. 5, 1984, pp. 64–88.
[4] I. J. Roseman, "Cognitive determinants of emotion: A structural theory," in *Review of Personality and Social Psychology*, Beverly Hills, CA, USA: Sage Publications, Inc., vol. 5, 1984, pp. 11–36.
[5] C. E. Izard, "Basic emotions, relations among emotions, and emotion-cognition relations," in *Psychological Review*, Washington, DC, USA: Amer. Psychol. Assoc., vol. 99, no. 3, 1992, pp. 561–565.
[6] P. Ekman, "An argument for basic emotions," *Cogn. Emotion*, vol. 6, no. 3/4, pp. 169–200, 1992.
[7] P. Shaver, J. Schwartz, D. Kirson, and C. O'connor, "Emotion knowledge: Further exploration of a prototype approach," *J. Pers. Social Psychol.*, vol. 52, no. 6, 1987, Art. no. 1061.
[8] S. Du, Y. Tao, and A. M. Martinez, "Compound facial expressions of emotion," *Proc. Nat. Acad. Sci.*, vol. 111, no. 15, pp. E1454–E1462, 2014.
[9] C. F. B. 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.*, 2016, pp. 5562–5570.
[10] S. Li and W. Deng, "Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 356–370, Jan. 2019.
[11] E. Barsoum, C. Zhang, C. Canton Ferrer, and Z. Zhang, "Training deep networks for facial expression recognition with crowd-sourced label distribution," in *Proc. ACM Int. Conf. Multimodal Interact.*, 2016, pp. 279–283.
[12] S. Li and W. Deng, "Blended emotion in-the-wild: Multi-label facial expression recognition using crowdsourced annotations and deep locality feature learning," *Int. J. Comput. Vis.*, vol. 127, no. 6/7, pp. 884–906, 2019.
[13] F. Zhou, S. Kong, C. C. Fowlkes, T. Chen, and B. Lei, "Fine-grained facial expression analysis using dimensional emotion model," *Neurocomputing*, vol. 392, pp. 38–49, 2020.
[14] W. Wang et al., "Learning to augment expressions for few-shot fine-grained facial expression recognition," 2020, *arXiv:2001.06144*.
[15] J. A. Russell, "A circumplex model of affect," *J. Pers. Social Psychol.*, vol. 39, no. 6, 1980, Art. no. 1161.
[16] 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. 2017.
[17] D. Kollias and S. Zafeiriou, "Affect analysis in-the-wild: Valence-arousal, expressions, action units and a unified framework," 2021, *arXiv:2103.15792*.
[18] H. R. Markus and S. Kitayama, "Culture and the self: Implications for cognition, emotion, and motivation," *Psychol. Rev.*, vol. 98, no. 2, 1991, Art. no. 224.
[19] P. Ekman, "Facial expression and emotion," *Amer. Psychol.*, vol. 48, no. 4, 1993, Art. no. 384.

[20] P. Ekman and W. Friesen, "Facial action coding system: A technique for the measurement of facial movement," *Consulting Psychol. Press Palo Alto*, vol. 12, Jan. 1978.
[21] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with gabor wavelets," in *Proc. IEEE 3rd Int. Conf. Autom. Face Gesture Recognit.*, 1998, pp. 200–205.
[22] T. Kanade, J. F. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," in *Proc. IEEE 4th Int. Conf. Autom. Face Gesture Recognit.*, 2000, pp. 46–53.
[23] E. Goeleven, R. De Raedt, L. Leyman, and B. Verschuere, "The karolinska directed emotional faces: A validation study," *Cogn. Emotion*, vol. 22, no. 6, pp. 1094–1118, 2008.
[24] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2011, pp. 2106–2112.
[25] I. J. Goodfellow et al., "Challenges in representation learning: A report on three machine learning contests," in *Proc. Int. Conf. Neural Informat. Process.*, 2013, pp. 117–124.
[26] A. Mollahosseini, B. Hasani, M. J. Salvador, H. Abdollahi, D. Chan, and M. H. Mahoor, "Facial expression recognition from world wild web," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2016, pp. 58–65.
[27] D. Kollias and S. Zafeiriou, "Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface," 2019, *arXiv:1910.04855*.
[28] 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.
[29] 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, 2009.
[30] P. Carcagnì, M. Del Coco, M. Leo, and C. Distante, "Facial expression recognition and histograms of oriented gradients: A comprehensive study," *SpringerPlus*, vol. 4, no. 1, pp. 1–25, 2015.
[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[32] R. Vemulapalli and A. Agarwala, "A compact embedding for facial expression similarity," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5683–5692.
[33] W. Zhang, X. Ji, K. Chen, Y. Ding, and C. Fan, "Learning a facial expression embedding disentangled from identity," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 6759–6768.
[34] 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 2018.
[35] Y. Chen, J. Wang, S. Chen, Z. Shi, and J. Cai, "Facial motion prior networks for facial expression recognition," in *Proc. IEEE Vis. Commun. Image Process.*, 2019, pp. 1–4.
[36] Y. Chen, G. Song, Z. Shao, J. Cai, T.-J. Cham, and J. Zheng, "Geoconv: Geodesic guided convolution for facial action unit recognition," 2020, *arXiv:2003.03055*.
[37] H. Yang, U. Ciftci, and L. Yin, "Facial expression recognition by de-expression residue learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2168–2177.
[38] J. Zeng, S. Shan, and X. Chen, "Facial expression recognition with inconsistently annotated datasets," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 222–237.
[39] F. Zhang, T. Zhang, Q. Mao, and C. Xu, "Joint pose and expression modeling for facial expression recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3359–3368.
[40] 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.*, 2020, pp. 6897–6906.
[41] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE Trans. Affective Comput.*, to be published, doi: 10.1109/TAFFC.2020.2981446.
[42] D. Kollias, I. Kotsia, E. Hajiyev, and S. Zafeiriou, "Analysing affective behavior in the second ABAW2 competition," 2021, *arXiv:2106.15318*.
[43] W. Zhang et al., "Prior aided streaming network for multi-task affective analysis," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops*, 2021, pp. 3539–3549.
[44] B. Fehr and J. A. Russell, "Concept of emotion viewed from a prototype perspective," *J. Exp. Psychol. Gen.*, vol. 113, no. 3, 1984, Art. no. 464.
[45] E. Rosch, "Principles of categorization," *Concepts Core Readings*, vol. 189, pp. 312–322, 1999.
[46] J. R. Averill, "A constructivist view of emotion," in *Theories of Emotion*. New York, NY, USA: Elsevier, 1980, pp. 305–339.
[47] K. Sailunaz and R. Alhajj, "Emotion and sentiment analysis from twitter text," *J. Comput. Sci.*, vol. 36, 2019, Art. no. 101003.
[48] Z. Wang, S.-B. Ho, and E. Cambria, "A review of emotion sensing: Categorization models and algorithms," *Multimedia Tools Appl.*, vol. 79, no. 47, pp. 35 553–35 582, 2020.
[49] L. F. Barrett, R. Adolphs, S. Marsella, A. M. Martinez, and S. D. Pollak, "Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements," *Psychol. Sci. Public Int.*, vol. 20, no. 1, pp. 1–68, 2019.
[50] M. A. Conway and D. A. Bekerian, "Situational knowledge and emotions," *Cogn. Emotion*, vol. 1, no. 2, pp. 145–191, 1987.
[51] T. Song, L. Chen, W. Zheng, and Q. Ji, "Uncertain graph neural networks for facial action unit detection," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 5993–6001.
[52] P. Antoniadis, P. P. Filntisis, and P. Maragos, "Exploiting emotional dependencies with graph convolutional networks for facial expression recognition," 2021, *arXiv:2106.03487*.
[53] H. Yang, L. Yin, Y. Zhou, and J. Gu, "Exploiting semantic embedding and visual feature for facial action unit detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 10482–10491.
[54] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013, *arXiv:1301.3781*.
[55] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2014, pp. 1532–1543.
[56] R. Müller, S. Kornblith, and G. E. Hinton, "When does label smoothing help?," in *Proc. 33rd Int. Conf. Neural Inf. Process. Syst.*, 2019, pp. 4694–4703.
[57] A. F. Hayes and K. Krippendorff, "Answering the call for a standard reliability measure for coding data," *Commun. Methods Measures*, vol. 1, no. 1, pp. 77–89, 2007.
[58] J. Shi and S. Zhu, "Learning to amend facial expression representation via de-albino and affinity," 2021, *arXiv:2103.10189*.
[59] A. H. Farzaneh and X. Qi, "Facial expression recognition in the wild via deep attentive center loss," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2021, pp. 2402–2411.
[60] O. Russakovsky et al., "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.
[61] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "VGGFace2: A dataset for recognising faces across pose and age," in *Proc. IEEE 13th Int. Conf. Autom. Face Gesture Recognit.*, 2018, pp. 67–74.
[62] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

Keyu Chen received the BEng and MEng degree from the University of Science and Technology of China, in 2018 and 2021 respectively. He is currently affiliated with Netease Fuxi AI Lab. His research interests include facial expression analysis, animation, and geometry learning.

Xu Yang received the BEng degree in communication engineering from the Nanjing University of Posts and Telecommunications, in 2013, the MEng degree in information processing from Southeast University, in 2016, and the PhD degree in computer science from Nanyang Technological University, in 2021. He is currently an associate professor with the School of Computer Science and Engineering of Southeast University, China. His research interests mainly include computer vision, machine learning, and Image Captioning.

Changjie Fan received the doctor's degree in computer science from the University of Science and Technology of China. He is the director of NetEase Fuxi AI Lab. His research interest is in machine learning, including multiagent systems, deep reinforcement learning, game theory, and knowledge discovery.

Wei Zhang received the BE degree in communication engineering from the Nanjing University of Posts and Telecommunications, Jiangsu, China, in 2017 and the MS degree in electronic and information engineering from Zhejiang University, Zhejiang, China, in 2020. She is currently a research scientist working with Netease Fuxi AI Lab, Hangzhou, China. Her current research interests include computer vision, expression embedding, and facial affective analysis.

Yu Ding received the PhD degree in computer science from the Telecom Paris tech, in Paris (France), 2014, the MS degree in computer science with Pierre and Marie Curie University (France), and the BS degree in Automation with Xiamen University (China). He is currently an artificial intelligence expert with Netease Fuxi AI Lab, Hangzhou, China. His research interests include deep learning, image and video processing, talking-head generation, animation generation, multimodal computing, affective computing, nonverbal communication (face, gaze, and gesture), and embodied conversational agent.