MAFW: A Large-scale, Multi-modal, Compound Affective Database for Dynamic Facial Expression Recognition in the Wild

Anonymous Author(s)
Submission Id: 1834

ABSTRACT

Dynamic facial expression recognition (FER) databases provide important data support for affective computing and applications. However, existing dynamic FER databases cannot satisfy the requirement to develop the real-world FER applications. Because most FER databases are annotated with several basic mutually exclusive emotional categories and contain only one modality, *e.g.*, images. The monotonous labels and modality cannot adequately simulate complex real-world human emotions. In this paper, we propose a large-scale compound affective database called MAFW with multiple modalities in the wild, which contains 10,045 video-audio clips. Each clip is annotated with a compound emotional category and a couple of sentences that describe the subjects' affective behaviors in the clip. For the compound emotion annotation, each clip is categorized into one or more of the 11 widely-used emotions, *i.e.*, anger, disgust, fear, happiness, neutral, sadness, surprise, contempt, anxiety, helplessness, and disappointment. To ensure a high quality for the labels, we filter out the unreliable annotations by an Expectation Maximization (EM) algorithm and select the multi-emotional samples according to the reliability. Finally, MAFW has 11 single-emotion categories and 32 multi-emotion categories. To the best of our knowledge, MAFW is the first in-the-wild database in three modalities (video, audio, and descriptive text) and annotated with compound emotions. We also propose a novel Transformer-based expression snippet feature learning method to recognize the compound emotions leveraging the expression-change relations among different emotions. Extensive experiments on MAFW database show the advantages of the proposed method over other state-of-the-art deep learning methods, for both single- and multi-modal FER.

CCS CONCEPTS

• **Computing methodologies → Activity recognition and understanding**; *Biometrics.*

KEYWORDS

Dynamic FER database, single and compound expressions, multimodal, Transformer, in the wild

1 INTRODUCTION

In recent years, benefit from the advancement of artificial intelligence, many important achievements have been made in facial expression recognition (FER), which has become a hot research topic in the fields of human-computer interaction (HCI) system, multimedia analysis and processing, intelligent robots and so on [4, 12, 15, 36]. Despite progress, most of existing methods and databases are developed based on six basic emotions (*i.e.*, happiness, sadness, fear, surprise, disgust, and anger) proposed by P. Ekman [14] and contain only single modality, *e.g.*, images. Since the monotonous labels and modality are significantly different from the real-world human emotions in the wild, FER techniques are still far from the real-world applications [16, 25]. In order to improve the real-world application of FER technology, it is crucial to first build a large-scale and in-the-wild affective database containing compound emotions and modalities.

Currently, existing FER databases are divided into two types: static facial image-based databases and dynamic video-based databases. Static facial image-based databases, like SFEW [8], JAFFE [30], RAF-DB [26], etc., mainly consist of several face images with emotion category labels. Due to lacking temporal expression changes information, static facial image-based databases with one modality are limited to apply to practical FER applications.

Dynamic FER database consists of a large number of facial video clips or facial image sequences with facial expression changes. Compared with static images, it contains richer emotion information, including both spatial change information and temporal change information of expressions. Depending on how the databases are collected, dynamic databases are classified as laboratory-collected constrained databases and in-the-wild unconstrained databases [25]. Table 1 reports the existing dynamic FER databases and their information. The former, such as CK+ [29], MMI [34], and BP4D [41], etc., capture videos of facial expression changes in the laboratory through event induction. Although significant advancements have been made in FER technologies on these constrained databases, these databases with single, limited, and consistent expression changes do not adequately simulate the complex real-world human emotions. For the latter, existing in-the-wild FER databases such as AFEW 7.0 [7] and DFEW [21] are constructed by crawling videos from movies and TV dramas. These in-the-wild databases are close to real life, with diverse environmental conditions and spontaneous expressions, however, they still have the following limitations:

- The labels of the data are monotonous. As shown in Table 1, most existing databases are composed of seven or eight basic mutually exclusive emotional categories, *e.g.*, six basic expression categories proposed by P. Ekman [14] plus neutral or contempt. Many studies [11, 13, 35, 38, 42] have shown that people usually express multiple emotions simultaneously in real life, along with gestures and vocal changes.
- Video sources are relatively homogeneous and repetitive. As shown in Table 1, videos in CAER [23] and DFEW [21] are from 79 TV dramas and 1,500 movies, respectively, while EmoVoxCeleb [1] is collected from interview programs.
- The modality of the data is also relatively monotonous. As shown in Table 1, most existing FER databases contain only videos or audio-video clips. Research shows that people typically understand emotions in their daily lives in three ways: 55% from visual, 45% from audio and verbal [31].

Table 1: Summary of existing dynamic facial expression databases.

Database	#Sample	Source	Expression categories	Is in-the-wild?	#Annotation Times	Modality
CK+ [29]	327	Lab	6 expressions+neutral and contempt	No	-	Vision
MMI [34]	2900	Lab	6 expressions+neutral	No	-	Vision
BP4D [41]	328	Lab	6 expressions+embarrassment and pain	No	-	Vision&Audio
Aff-Wild2 [9]	84	Web & YouTube	6 expressions+neutral	Yes	3	Vision&Audio
AFEW 7.0 [7]	1,809	54 movies	6 expressions+neutral	Yes	2	Vision&Audio
CAER [23]	13,201	79 TV dramas	6 expressions+neutral	Yes	3	Vision&Audio
EmoVoxCeleb [1]	1,251	Interview videos from YouTube	6 expressions+neutral and contempt	Yes	Auto	Vision&Audio
DFEW [21]	16,372	1500 movies	6 expressions+neutral	Yes	10	Vision&Audio
Our MAFW	10,045	1,600 movies & TV dramas 20,000 short videos from reality shows, talk shows, news, etc 2,045 clips from [7], [21], and [1]	11 single expressions 32 compound expressions emotional descriptive text	Yes	11	Vision Audio Text

Additional descriptive text for emotions can help and extend many new emotion recognition applications.

To overcome the above problems, we construct a large-scale compound affective database called MAFW with multiple modalities in the wild, which contains 10,045 video-audio clips. MAFW can be used as a new benchmark for researchers to develop and evaluate their methods for FER tasks in both single modality and multiple modalities. We believe that our MAFW can also be used to other several emotion recognition tasks, such as cross-domain FER, emotion caption, zero-shot facial AU detection, self-supervision FER, etc. Fig. 1 gives typical examples in our MAFW database and the corresponding compound annotations and descriptive text. Our MAFW has the following three advantages over the existing databases:

- Our MAFW is the first large-scale, multi-modal compound affective database with 11 single expression categories, 32 multiple expression categories, and detailed descriptive emotion text. For the compound emotion annotation, each clip is categorized into one or more of the 11 complex emotions, *i.e.*, anger, disgust, fear, happiness, neutral, sadness, surprise, contempt, anxiety, helplessness, and disappointment. For the descriptive text annotation, each clip is annotated with a couple of sentences that describe the subjects' affective behaviors in the clip.
- To obtain reliable and objective labels, each clip in MAFW is labeled enough times independently as one or more of the eleven complex expression categories, and unreliable labels are filtered out by an Expectation Maximization (EM) based reliability evaluation algorithm.
- The data source of MAFW is richer, not only from movies and TV dramas, but also includes short videos from reality shows, talk shows, news, variety shows, etc, which is closer to the real-world emotion.

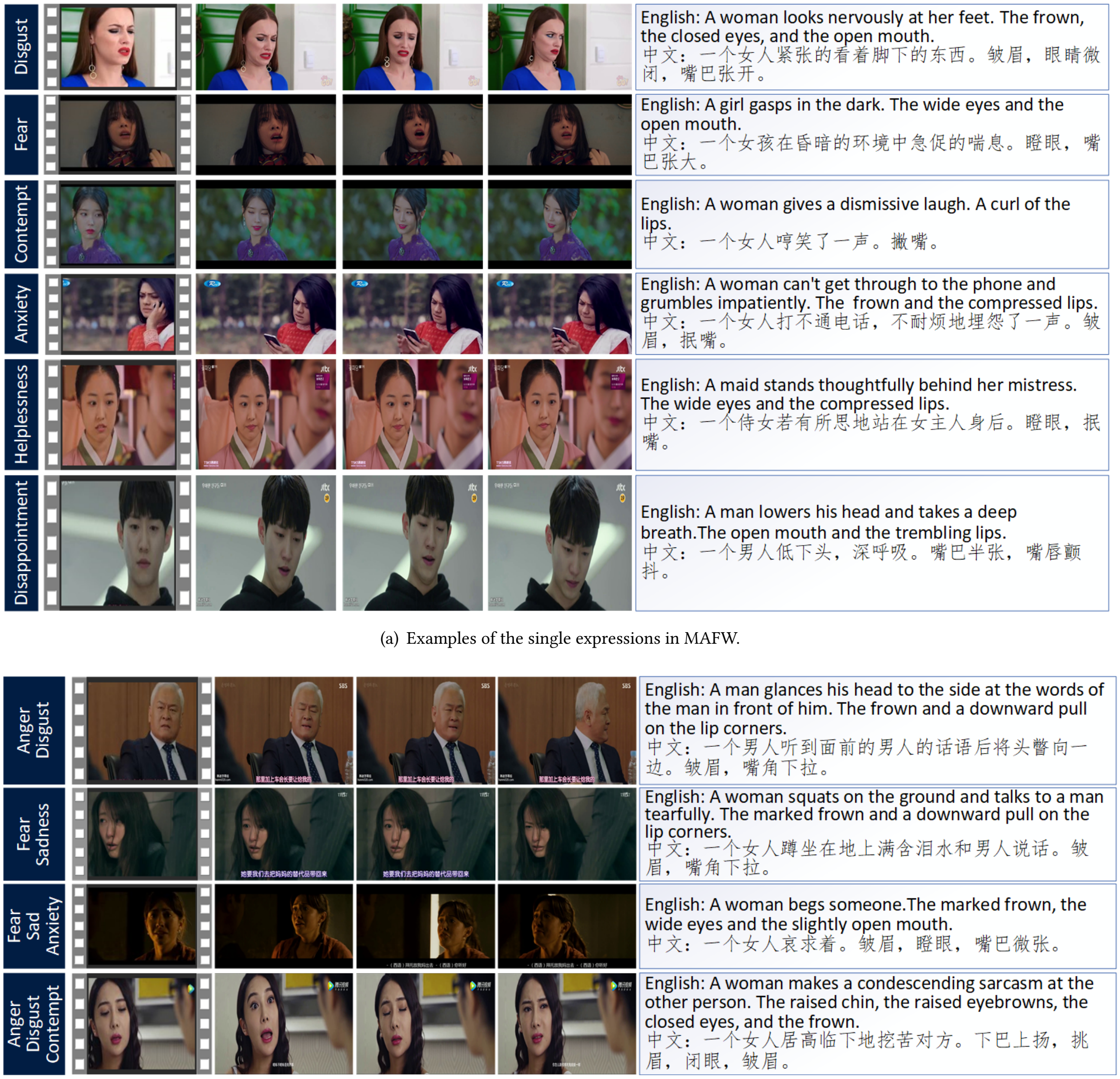
In addition to MAFW, we also propose a novel Transformer-based expression snippet feature learning method (T-ESFL), to effectively model subtle expression movements and thus obtain robust single- and multi-modal FER in the wild. Furthermore, we establish four benchmark evaluation protocols for MAFW, and conduct extensive experiments using many spatio-temporal deep learning methods as well as our T-ESFL. Experimental results show that the proposed T-ESFL can improve the performance of existing deep learning methods.

2 RELATED WORK

Constrained dynamic FER databases The constrained databases are usually captured from some subjects in a fixed indoor environment, with emotion occurring typically in the form of viewing video and event elicitation. For example, CK+ [29] has 123 subjects performed the corresponding expressions according to the instructions under laboratory conditions, and obtained six basic expressions, contempt, and neutral. BP4D [41] has 41 subjects with age distribution from 18 to 29 years old, and collects specific expressions under the guidance of 8 different tasks, which includes one-to-one interviews (inducing pleasure), suddenly hearing a voice (inducing surprise), and so on, to obtain video samples of 8 expressions. The constrained expression databases, although spontaneous, are limited by a single environment, the number of individuals, and the cost of production, making it difficult to simulate the real-world human emotions.

In-the-wild dynamic FER databases Dynamic FER databases in the wild are usually collected from web resources, including TV shows, movies, and so on. AFEW 7.0 [7] collects 1,800 facial expression clips from movies, providing neutral and 6 basic expression labels. DFEW [21] and CAER [23] obtain 16,372 and 13,201 facial expression video clips and are from TV shows and movies, respectively. Although AFEW7.0, DFEW, and CAER are constructed using videos in the wild, these databases still share some limitations, *i.e.*, they all provide only basic and single expression labels and their sources are only movie clips.

Compound FER databases Since the 20th century, many psychological and cognitive studies have shown that people usually express multiple expressions simultaneously [38]. In 1984, Ekman noted in [13] that people typically produce compound emotions; for example, if subjects are asked to imagine fear, they are likely to produce a mixture expressions of surprise and fear. It is clear that the existing single basic expression labels are not conducive to understand human emotions. In CVPR2017, Deng *et al.* [26] presented the first static compound FER database, namely RAF-DB, that contains 6-class basic expressions plus neutral and 11-class compound expressions. In ACL2018, Zadeh *et al.* [2] presented a dynamic database, namely CMU-MOSEI, supporting multiple labels consisting of six basic expressions. To the best of our knowledge, there is no dynamic FER database labeled with compound expressions and emotional description text in the wild.



English: A woman looks nervously at her feet. The frown, the closed eyes, and the open mouth.
中文：一个女人紧张的看着脚下的东西。皱眉，眼睛微闭，嘴巴张开。

English: A girl gasps in the dark. The wide eyes and the open mouth.
中文：一个女孩在昏暗的环境中急促的喘息。瞪眼，嘴巴张大。

English: A woman gives a dismissive laugh. A curl of the lips.
中文：一个女人哼笑了一声。撇嘴。

English: A woman can't get through to the phone and grumbles impatiently. The frown and the compressed lips.
中文：一个女人打不通电话，不耐烦地埋怨了一声。皱眉，抿嘴。

English: A maid stands thoughtfully behind her mistress. The wide eyes and the compressed lips.
中文：一个侍女若有所思地站在女主人身后。瞪眼，抿嘴。

English: A man lowers his head and takes a deep breath.The open mouth and the trembling lips.
中文：一个男人低下头，深呼吸。嘴巴半张，嘴唇颤抖。

(a) Examples of the single expressions in MAFW.

English: A man glances his head to the side at the words of the man in front of him. The frown and a downward pull on the lip corners.
中文：一个男人听到面前的男人的话语后将头瞥向一边。皱眉，嘴角下拉。

English: A woman squats on the ground and talks to a man tearfully. The marked frown and a downward pull on the lip corners.
中文：一个女人蹲坐在地上满含泪水和男人说话。皱眉，嘴角下拉。

English: A woman begs someone.The marked frown, the wide eyes and the slightly open mouth.
中文：一个女人哀求着。皱眉，瞪眼，嘴巴微张。

English: A woman makes a condescending sarcasm at the other person. The raised chin, the raised eyebrowns, the closed eyes, and the frown.
中文：一个女人居高临下地挖苦对方。下巴上扬，挑眉，闭眼，皱眉。

(b) Examples of the multiple expressions in MAFW.

Figure 1: Examples of the compound expressions and the corresponding descriptive text (including English and Chinese) from MAFW. (a) The single expressions in MAFW, (b) the multiple expressions in MAFW. Due to space limitations, we only show a small number of frames in these clips.

3 MAFW DATABASE

3.1 Data Collection

The pipeline of data collection in MAFW is shown in Fig. 2. The MAFW has two main sources of data. The first and most important data sources are movies, TV dramas, and short videos from some reality shows, talk shows, news, variety shows, etc., on BiliBili and Youtube websites. We develop a crawler program to crawl over 1,600 HD movies, TV dramas and over 20,000 short reality shows, variety shows and short videos. These videos come from China, Japan, Korea, Europe, America and India, and cover various themes, *e.g.*, variety, family, science fiction, suspense, horror, love, comedy, and interviews, encompassing a wide range of human emotions. To ensure the diversity of the data, we only randomly download one episode of the same TV drama, as well as selected no more than three facial expression clips in an episode or short video. The second data source is the use of videos from existing public databases to supplement some uncommon categories, with 1,097 videos from DFEW [21], 98 videos from AFEW 7.0 [7], and 850 videos from EmoVoxCeleb [1].

For the crawled audio-video clips, we first use FaceDetector [19, 24] to detect and extract the clips containing faces, and then manually filter the unqualified clips to obtain 10,045 usable audio-video clips in our database.

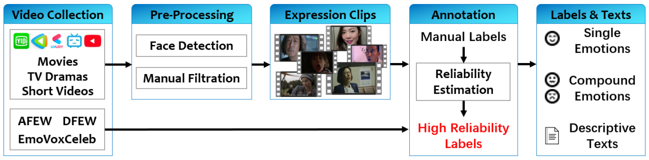
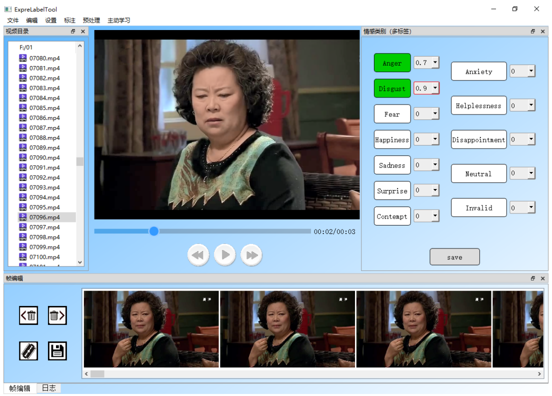


Figure 2: Overview of the construction and the annotation of MAFW.

3.2 Data Annotation

Annotating such a large database is an extremely difficult and time-consuming task. Unlike other databases that only provide the basic expression categories proposed by P. Ekman [14], our database provides three kinds of annotations for compound emotions in the wild: (1) **single expression label**, *i.e.*, each clip has a predominant and exclusive expression label (see Fig. 1(a)); (2) **multiple expression label**, *i.e.*, a clip with compound emotions can be labeled with multiple expression categories, in the multi-label annotation manner (such as "Anger+Disgust" in Fig. 1(b)); (3) **emotional descriptive text**, *i.e.*, each clip is annotated a couple of sentences (with two languages) that describe the subjects' affective behaviors in the clip. The following details the compound emotion annotation and the descriptive text annotation respectively.



Anger 0.7
Disgust 0.9
Fear 0
Happiness 0
Sadness 0
Surprise 0
Contempt 0
Anxiety 0
Helplessness 0
Disappointment 0
Neutral 0
Invalid 0

Figure 3: The main interface of the ExpreLabelTool for labelling.

Compound emotion annotation To ensure the professionalism of the annotation, we employ 11 annotators trained in emotional knowledge and develop a software called Expression Label Tool (ExpreLabelTool) to help the annotators do it efficiently, as shown in Fig. 3. Each clip is assured to be labeled by about 11 independent annotators. They are asked to classify the clips into the most apparent one or more of the 11 complex emotions, *i.e.,* anger(AN), disgust(DI), fear(FE), happiness(HA), sadness(SA), surprise(SU), contempt(CO), anxiety(AX), helplessness(HL), disappointment(DS), and neutral(NE), and assess the self-confidence scores of their annotations by using the drop-down rating box (from 0 to 1, with 11 options) in the tool. Note that a higher score indicates more certainty in its associated label, while a label with a score of zero indicates no annotation. After that, each clip can obtain a single category or multiple emotion categories, *i.e.*, an 11-dimensional vector, where each dimension represents the score of the relevant emotion. We describe later how to select single and multiple expressions based on this vector.

Descriptive text annotation With the annotated emotion, we additionally detail the emotional descriptive text with a couple of sentences in two different languages, *i.e.*, Chinese and English, for each clip except for the neutral expression clip. The right column in Fig. 1 shows examples of the descriptive texts in MAFW. The emotion descriptive text includes several information related to the affective behaviors, such as the context event, character relationships, subjects' facial unit movements, and the scenes. To ensure the objectivity and complementarity of emotional descriptive text, the descriptive text cannot directly include words with emotional labels, such as "she is angry".

3.3 Metadata

The MAFW is a multi-modal database that consists of video, audio, and text. Each clip data is provided with a single or multiple expression category annotation, an average confidence score for each expression annotation, and several descriptive sentences (text) for emotion caption. We also provide three automatic annotations for each data: the frame-level 68 facial landmarks and the face regions by [3], and the gender by a CNN model pre-trained on CelebA [27]. In summary, in the MAFW database, 58.1% are male and 41.9% are female.

3.4 Reliability Estimation

Due to the subjectivity difference of annotators, the reliability of annotation may be highly variable and inconsistent. To get rid of the labels with lower reliability, motivated by [37] and [5], we employ an Expectation Maximization (EM) algorithm to assess each annotator's reliability and achieve high-reliability label results. The algorithm of EM for reliability estimation is shown in Algorithm 1.

Given the labels of N videos annotated by M annotators, we first binaryize their labels into a zero-one matrix H^k_{MN} on the emotion category k as:

$$H^k_{MN} = \{h^k_{ij}\}, \tag{1}$$

where h^k_{ij} will be "1", if the ith annotator labels the jth video with emotion category k, otherwise it will be "0".

Our goal is to estimate each annotator' reliability for each expression by optimizing the likelihood of their labels. The reliability is formulated as two M-dimensional probability vectors: $\{\alpha^k_i\}$ and $\{\beta^k_i\}$,

$$\alpha^k_i = P(h^k_{ij} = 1 | v^k_j = 1), \beta^k_i = P(h^k_{ij} = 0 | v^k_j = 0), \tag{2}$$

where α^k_i is the reliability probability that the ith annotator correctly labels the emotion category k and β^k_i is the reliability probability that the ith annotator does not label the emotion category k. Note that α^k_i and β^k_i are independent of each other. $v^k_j = \{0, 1\}$ denotes whether the jth video has the label of the emotion category k. We initialized the v^k_j via annotation majority voting.

Following the above definitions, in the E-step of the EM, the reliability probabilities are used to estimate the posterior probability φ_j^k that the jth video correctly be labelled with the emotion category k:

$$\varphi_j^k = \frac{p^k \mu_j^k}{p^k \mu_j^k + (1-p^k)\eta_j^k}, \tag{3}$$

where p^k is the expected probability of the emotion category k and initialized by $\frac{1}{N}\sum_{j=1}^{N} v_j^k$. μ_j^k and η_j^k are calculated as:

$$\mu_j^k = \prod_{i=1}^{M} (\alpha_i^k)^{h_{ij}^k} (1-\alpha_i^k)^{(1-h_{ij}^k)}, \tag{4}$$

$$\eta_j^k = \prod_{i=1}^{M} (\beta_i^k)^{(1-h_{ij}^k)} (1-\beta_i^k)^{h_{ij}^k}. \tag{5}$$

In the M-step of the EM, we first update p^k as:

$$p^k = \frac{1}{N}\sum_{j=1}^{N} \varphi_j^k. \tag{6}$$

Then, we update α_i^k and β_i^k by Maximum Likelihood Estimation:

$$\alpha_i^k = \frac{\sum_{j=1}^{N} \varphi_j^k h_{ij}^k}{\sum_{j=1}^{N} \varphi_j^k}, \tag{7}$$

$$\beta_i^k = \frac{\sum_{j=1}^{N} (1-\varphi_j^k)(1-h_{ij}^k)}{\sum_{j=1}^{N} (1-\varphi_j^k)}. \tag{8}$$

Finally, we set $Q(p^k, \alpha^k, \beta^k)$ as the convergence objective in EM algorithm as:

$$Q(p^k, \alpha^k, \beta^k) = \sum_{j=1}^{N} [\varphi_j^k lnp^k \mu_j^k + (1-\varphi_j^k) ln(1-p^k)\eta_j^k]. \tag{9}$$

We can further determine whether $Q(p^k, \alpha^k, \beta^k)$ converges:

$$\frac{|Q(p_{(t+1)}^k, \alpha_{(t+1)}^k, \beta_{(t+1)}^k) - Q(p_{(t)}^k, \alpha_{(t)}^k, \beta_{(t)}^k)|}{|Q(p_{(t)}^k, \alpha_{(t)}^k, \beta_{(t)}^k)|} < \varepsilon, \tag{10}$$

where t denotes the number of iterations and ε is the convergence threshold that is set as 0.000001 empirically. If $Q(p^k, \alpha^k, \beta^k)$ converges, we can obtain the reliability of all annotators, otherwise return the E-step.

With the reliability, for each emotion category, we retain five high-reliability labels at least. We also use Cronbach's Alpha [6] score to measure the consistency of the retained labels, and the result is presented in Table 2. The scores are higher than 0.9 on five basic emotions, with an average score of 0.823 on the 11-class emotions.

3.5 Single and Multiple Expression Selection

With the retained high-reliability labels and the self-confidence scores, we can naturally divide the MAFW into two sets, namely the single expression set and multiple expression set.

Given the self-confidence scores from a high-reliability clip where no less than half of the annotators have labeled the kth emotion category $C^k = (c_1^k, c_2^k, \ldots, c_m^k)$, we first calculate the mean value of the self-confidence scores $c_{mean}^k = \sum_{i=1}^{m} c_i^k / m$ on the emotion category, then pick out the emotion label k w.r.t $c_{mean}^k \geq 0.5$ as the valid label.

Table 2: Cronbach's alpha scores in the MAFW database.

Emotions	Alpha	Emotions	Alpha
Anger	0.955	Surprise	0.920
Disgust	0.824	Contempt	0.731
Fear	0.934	Anxiety	0.729
Happiness	0.961	Helplessness	0.686
Neutral	0.878	Disappointment	0.498
Sadness	0.948	**Average**	**0.824**

```
Algorithm 1: Annotation reliability estimation algorithm
Input:
zero-one matrix $\{H_{MN}^k\}_{k=1}^K$ of the emotion category $k$;
$M$: the number of annotators;
$N$: the number of videos;
$K$: the number of emotion categories.
Output: the reliability matrices of $M$ annotators on each
emotion category $\{\alpha_i^k\}_{i=1}^M$, $\{\beta_i^k\}_{i=1}^M$.
Initialize:
$\forall k = 1, \ldots, K$, initialize true labels $\{v_j^k\}_{j=1}^N$ with majority
 voting via $H_{MN}^k$. The initial value of $p^k$ is the expected
 probability of the emotion label $k$.
$p^k := \frac{1}{N}\sum_{j=1}^{N} v_j^k$   $\alpha_i^k := 0.999999$   $\beta_i^k := 0.999999$
for $k$=1 to $K$ do
  Repeat
    E-step:
      estimate the posterior probabilities $\{\varphi_j^k\}_{j=1}^N$ of $N$ clips
with the $k$th expression as Eq. (3)–(5).
    M-step:
      update $p^k$, $\alpha_i^k$, and $\beta_i^k$ based on $\{\varphi_j^k\}_{j=1}^N$ through the
 maximum likelihood algorithm as Eq. (6)–(8).
      Calculate $Q(p^k, \alpha^k, \beta^k)$ as Eq. (9).
  until $Q(p^k, \alpha^k, \beta^k)$ converges
```

The single expression set construction includes the following steps. For valid-labeled clips with single expression labels, we directly classify them into the single expression set; for clips with multiple expression labels, we select the labels with the highest average confidence score as its predominant single expressions and also classify them into the single expression set, so that the single expression set consists of all 9,172 valid-labeled clips with 11-class emotions. Table 3 reports the distribution of clip amount and clip length per expression category on the single expression set.

Similarly, we can construct the multiple expression set. We compose the clips with multiple expression labels into the compound expression set. To prevent having too few samples in a class, we keep only the multiple expression categories with more than 10 labeled samples, thus obtaining 32-class compound expressions. As a result, we obtain 4,058 clips with multiple expressions. Fig. 4 shows the distribution of multiple expression categories on the multiple expression set. Fig. 1(a) and Fig. 1(b) show some typical examples from the 11-class single expressions and 32-class multiple expressions, respectively.

Table 3: The distribution of clip amount and clip length per single expression on the single expression subset.

Expressions	Clips				Percent(%)
	0-2s	2-5s	5s+	Total	
Anger	183	945	262	1390	15.15
Disgust	97	434	108	639	6.97
Fear	139	413	73	625	6.81
Happiness	88	900	254	1242	13.54
Neutral	42	872	224	1138	12.41
Sadness	97	873	500	1470	16.03
Surprise	233	721	118	1072	11.69
Contempt	18	173	45	236	2.57
Anxiety	99	626	191	916	9.99
Helplessness	20	174	68	262	2.86
Disappointment	13	118	51	182	1.98
Total	1029	6249	1894	9172	100.00

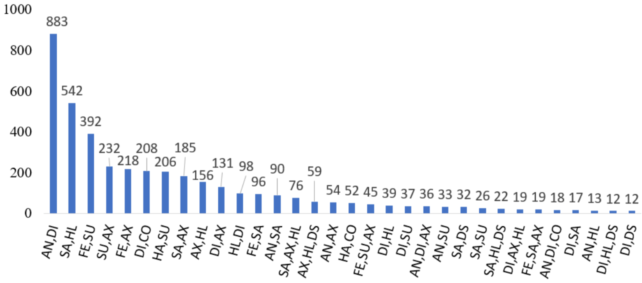


Figure 4: The distribution of the number of multiple expressions on the multiple expression subset.

4 EXPRESSION SNIPPET FEATURE LEARNING WITH TRANSFORMER

In-the-wild FER is a difficult task due to subtle facial expression movements within videos that can be too difficult to be modeled properly by existing methods. In this paper, we propose a novel Transformer-based expression snippet feature learning method (T-ESFL) that can model relations of subtle expression change moments and obtain movement-sensitive emotion representation. Moreover, the T-ESFL is easily extended for multi-modal FER, achieving the state-of-the-art performance on both single- and multi-modal FER. The T-ESFL consists of three main components, *i.e.*, expression snippet decomposition, Transformer, and snippet order shuffling and reconstruction learning (SOSR), as illustrated in Fig. 5.

Expression snippet decomposition Given an input FER video clip $\mathcal{S}$, we first decompose the input into a series of small snippets $\mathcal{S} = \{S_1, S_2, ...S_n\}$ to augment the Transformer's ability to model subtle expression changes within each snippet and across different snippets, respectively. All the snippets have the same length, and they follow consecutive orders along time. Then, we extract snippet features R_i from each S_i with a pre-trained CNN [32] and attention learning.

Transformer architecture With the snippet features R_i, a Transformer is applied here to model the expression movements across snippets and discover a unified emotion feature for FER. We follow the typical Transformer [40] and apply a multi-head attention-based encoder-decoder pipeline for the processing. In general, the multi-head attention estimates the correlation between a *query* tensor and a *key* tensor and then aggregates a *value* tensor according to correlation results to obtain an attended output. The final output of the Transformer is called the united salient emotion representation T.

SOSR learning To make T more sensitive to subtle expression movements, SOSR shuffles the snippet order and makes T-ESFL reconstruct the correct order in a self-supervision learning manner. Inspired by [33], we follow a Jigsaw permutation and shuffle order pure randomly to deconstruct the normal temporal dependency. The snippets with a shuffled order are later sent to T-ESFL, and are predicted the permutation type of input snippets by using a reconstruction loss L_{rec}. Based on this, we can achieve movement-sensitive emotion representation for robust FER.

Optimization Objective The total objective function of T-ESFL includes two joint cross-entropy losses, and is expressed as $L = L_{cls} + \frac{1}{n} \cdot L_{rec}$. The first one L_{cls} is a FER classification loss, and the second one L_{rec} is the snippet order reconstruction loss. Note that, n is the number of the decomposed snippets.

Multi-modal emotion prediction Despite with the visual representation, T-ESFL can achieve the state-of-the-art results on FER (see Table 5 and Table 7). The T-ESFL can also be easy to be extended for multi-modal FER. Specifically, we use the ResNet_LSTM network and DPCNN [22] to extract audio emotion representation and text emotion representation, respectively. Then, we concatenate the visual, audio, and text representations to identify the final emotion category via a simple fully-connected layer and Softmax operation. We experimentally verified that the use of multi-modal fusion features can effectively improve FER in the wild.

5 EXPERIMENTS

In the section, we first presented the experimental setup of the benchmarks, including experiment protocols, data preprocessing, evaluation metrics, and implementation details. Then, we conducted extensive benchmarks and comparative experiments on our MAFW with multiple labels and modalities.

5.1 Experimental Setup

Data&Protocol To facilitate the FER research from laboratory environments to the real world, we performed four challenging benchmark experiments on MAFW: 11-class single-modal single expression classification, 43-class single-modal compound expression classification, 11-class multi-modal single expression classification, and 43-class multi-modal compound expression classification. In 11-class single expression classification experiments, emotion categories were recognized using the whole 9,172 clips from the whole single expression set. Considering both multiple and single expressions in real-world scenes, for 43-class compound expression classification, emotion categories were classified using 4,058 clips

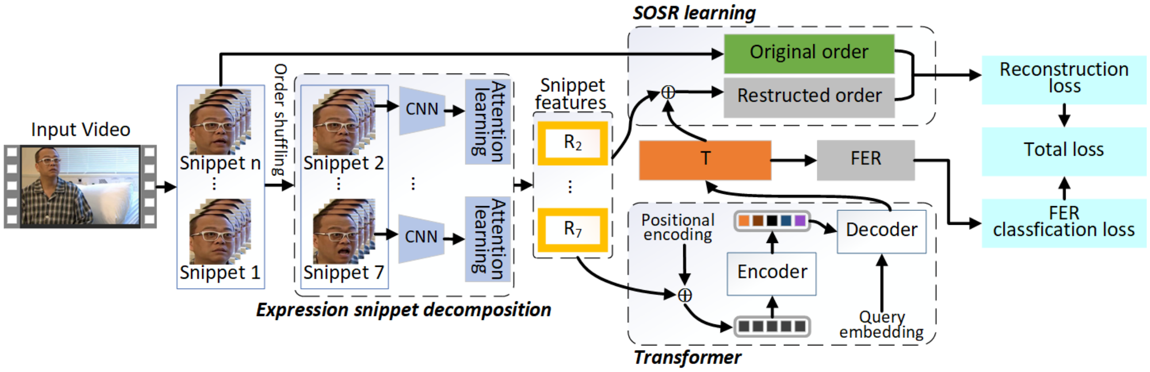


Figure 5: The architecture of T-ESFL for movement-sensitive emotion representation learning. Using untrimmed video clips, we mainly apply the expression snippet decomposition, the Transformer, and the SOSR, to enable the effective modeling of intra- and inter-snippet expression movements for discovering more informative expression cues, thus achieving robust FER.

from the 32-class multiple expression set and the remaining 4,938 11-class single expression clips. Similar to the evaluation protocol of existing FER databases [21, 26], we adopt a 5-fold cross-validation protocol for these benchmarks on our MAFW database.

Preprocessing First, we used OpenCV to extract frame images for each clip. Then we used the face-alignment-master program [3] to obtain face regions and 68 landmarks on all frames, removing the frames without faces. Finally, we did face alignment by using affine transform and matrix rotation via the OpenCV.

Evaluation Metrics Consistent with the previous research [21, 26], we chose four widely used validation metrics, *i.e.,* the unweighted average recall (UAR), weighted average recall (WAR), F-score (F1), and Area under the ROC curve (AUC), to evaluate the single- and multi-modal FER tasks. The UAR indicates that the sum of the accuracies of each class is divided by the number of classes without considerations of instances per class, and we can adequately evaluate the performance of predicting emotions with few samples. The WAR is the recognition accuracy of overall expressions. The F1 can be regarded as the weighted harmonic mean value of the accuracy and recall, and we simply calculate the average of the F1 on all categories. AUC generically refers to the area under receiver operating characteristic (ROC) curve, and we calculate AUC averaging on each label. We hope to improve models' performance in terms of UAR, WAR, F1, and AUC metrics.

Implementation Details In this paper, we employed the PyTorch framework to implement all models. We conducted the experiments in single- and multi-modalities, respectively, where each modality has both single and compound expression classification. The key training parameters involved in the work are presented in Table 4. All models are trained on NVIDIA GeForce RTX 3090 and GTX1080, with an excellent initial learning rate 0.0001 provided by the grid search strategy. The learning rate decreases at a rate of 0.2 when the loss is saturated.

5.2 Experimental Results

5.2.1 11-class Single-modal Single Expression Classification. To evaluate single-modal single expression classification, we compared our T-ESFL model with existing state-of-the-art FER models including three static frame-based methods (*i.e.*, Resnet18 [18], VIT [10], and EmotionClassifier [19, 24]) and four dynamic sequence-based methods (*i.e.*, C3D [39], Resnet18 [17, 18, 20], VIT_LSTM [10, 17, 20], and C3D_LSTM [17, 20, 39]). The comparison results are shown in Table 5. For the static frame-based methods, we first select five frames from a video evenly as input, and then fuse the prediction probabilities of the five frames in the output layer of the models to obtain the final prediction result. For the dynamic sequence-based methods, we use all frames in a video for emotion prediction. Compared to other state-of-the-art methods, the proposed T-ESFL achieves the best UAR of 34.20% and the best WAR of 49.38%. Moreover, our approach improves the WAR by 7.66% compared to the commercial model EmotionClassifier [19, 24], and also improves the WAR by 5.75% compared to the second best sequence-based method, VIT_LSTM.

Table 4: The key training parameters involved in the work.

Models	Batch size	Input size
Resnet18 [18], VIT [10]	32	224 × 224
C3D [39]	8	112 × 112
Resnet18_LSTM [17, 18, 20]	16	224 × 224
VIT_LSTM [10, 17, 20]	16	224 × 224
C3D_LSTM [17, 20, 39]	8	112 × 112
Resnet18_LSTM^{a} [17, 18, 20]	8	224 × 224
C3D_LSTM^{a} [17, 20, 39]	8	112 × 112
T-ESFL, T-ESFL^{a}, T-ESFL^{a+t}	8	224 × 224

^{a} represents multi-modal evaluation with both video and audio;
$^{a+t}$ represents multi-modal evaluation with video, audio, and text.

5.2.2 43-class Single-modal Compound Expression Classification. Table 7 shows the comparison results of 43-class single-modal compound expression classification. Similar to the above single expression classification, the same six models except for the EmotionClassifier are used for 43-class single-modal compound expression recognition, with four evaluation metrics (WAR, UAR, F1, and AUC). The output layers of these models are modified to fit the multi-label prediction task. Compared to other methods, the proposed T-ESFL achieves the best WAR of 34.35% and the best AUC of 75.6%.

Table 5: Comparison results on 11-class single-modal single expression classification.

Models	Feature setting	Accuracy of each expression											Metric	
		AN	DI	FE	HA	NE	SA	SU	CO	AX	HL	DS	UAR	WAR
Resnet18 [18]	frame-based	45.02	9.25	22.51	70.69	35.94	52.25	39.04	0	6.67	0	0	25.58	36.65
VIT [10]	frame-based	46.03	**18.18**	27.49	76.89	50.70	68.19	45.13	1.27	18.93	1.53	1.65	32.36	45.04
EmotionClassifier [19, 24]	frame-based	13.60	4.07	0.08	81.09	**75.48**	47.82	53.02	-	-	-	-	**39.85**	44.75
C3D [39]	sequence-based	51.47	10.66	24.66	70.64	43.81	55.04	46.61	**1.68**	24.34	**5.73**	**4.93**	31.17	42.25
Resnet18_LSTM [17, 18, 20]	sequence-based	46.25	4.70	25.56	68.92	44.99	51.91	45.88	1.69	15.75	1.53	1.65	28.08	39.38
VIT_LSTM [10, 17, 20]	sequence-based	42.42	14.58	**35.69**	76.25	54.48	**68.87**	41.01	0	24.40	0	1.65	32.67	45.56
C3D_LSTM [17, 20, 39]	sequence-based	54.91	0.47	9	73.43	41.39	64.92	**58.43**	0	**24.62**	0	0	29.75	43.76
T-ESFL	snippet-based	**62.70**	2.51	29.90	**83.82**	61.16	67.98	48.50	0	9.52	0	0	39.28	**48.18**

Table 6: Comparison results on 11-class multi-modal single expression classification.

Models	Feature setting	Accuracy of each expression											Metric	
		AN	DI	FE	HA	NE	SA	SU	CO	AX	HL	DS	UAR	WAR
Resnet18_LSTM[a] [17, 18, 20, 28]	sequence-based	54.47	**11.89**	7.07	82.73	54.85	55.06	39.35	0	15.99	**0.39**	0	29.26	42.69
C3D_LSTM[a] [17, 20, 28, 39]	sequence-based	**62.47**	3.17	15.74	77.30	42.20	65.30	42.67	0	19.14	0	0	30.47	44.15
T-ESFL[a]	snippet-based	60.73	1.26	**21.4**	80.31	**58.24**	**75.31**	53.23	0	14.93	0	0	**33.35**	48.7
T-ESFL[a+t]	snippet-based	61.89	1.1	7.69	**85.90**	-	71.87	**62.17**	0	**36.00**	0	0	31.00	**50.29**

[a] represents multi-modal evaluation with both video and audio;
[a+t] represents multi-modal evaluation with video, audio, and text.

Table 7: Comparison results on 43-class single-modal compound expression classification.

Models	Feature setting	Metric			
		UAR	WAR	F1	AUC
Resnet18 [18]	frame-based	6.18	23.83	4.89	62.92
VIT [10]	frame-based	8.62	31.76	7.46	74.9
C3D [39]	sequence-based	**9.51**	28.12	6.73	74.54
Resnet18_LSTM [17, 18, 20]	sequence-based	6.93	26.6	5.56	68.86
VIT_LSTM [10, 17, 20]	sequence-based	8.72	32.24	**7.59**	75.33
C3D_LSTM [17, 20, 39]	sequence-based	7.34	28.19	5.67	65.65
T-ESFL	snippet-based	9.15	**34.35**	7.18	**75.63**

Table 8: Comparison results on 43-class multi-modal compound expression classification.

Models	Feature setting	Metric			
		UAR	WAR	F1	AUC
Resnet18_LSTM[a] [17, 18, 20, 28]	sequence-based	7.85	31.03	5.95	71.08
C3D_LSTM[a] [17, 20, 28, 39]	sequence-based	7.45	29.88	5.76	68.13
T-ESFL[a]	snippet-based	**9.93**	34.67	8.44	74.13
T-ESFL[a+t]	snippet-based	9.68	**35.02**	**8.65**	**74.35**

[a] represents multi-modal evaluation with both video and audio;
[a+t] represents multi-modal evaluation with video, audio, and text.

5.2.3 *11-class Multi-modal Single Expression Classification.*

For multi-modal FER, we compared our T-ESFL with two spatiotemporal neural network methods, *i.e.*, Resnet18_LSTM [17, 18, 20] and C3D_LSTM [17, 20, 39], as shown in Table 6. Obviously, the multiple modalities effectively improve the performance of FER. Compared to other methods, our T-ESFL model obtains the best results in the fusion of different modalities, *e.g.*, 9.45% boost in UAR on video and audio modalities. Moreover, continuously adding the descriptive text modality can obtain a relative 3.26% boost in WAR.

5.2.4 *43-class Multi-modal Compound Expression Classification.*

For multi-modal compound expression classification, similar to the above evaluation on multi-modal FER, we also compared our T-ESFL with the same two spatiotemporal models, as shown in Table 8. Compared to other two multi-modal methods, the proposed T-ESFL on video and audio modalities achieves the best WAR of 34.67% and UAR of 9.93%. In addition, adding the descriptive text modality, the results of T-ESFL continuously achieve improvement, *i.e.*, relative 1% in WRA, 2.5% in F_1, and 0.3% in AUC.

6 CONCLUSIONS AND FUTURE WORK

In this paper, we propose a large-scale compound affective database called MAFW with multiple modalities in the wild, which contains 10,045 video-audio clips. Each clip is annotated with a high-reliability compound emotional category and a couple of sentences that describe the subjects' affective behaviors in the clip. Therefore, MAFW is the first affective database that provides three types of emotion annotations, *i.e.*, single expression labels (11 class), compound expression labels (32 class), and descriptive text with a couple of sentences (two languages). Moreover, we also propose a novel Transformer-based expression snippet feature learning method to obtain movement-sensitive emotion representation, thus achieving state-of-the-art performance on both single-modal and multi-modal FER in the wild. In the future, we will continue to maintain the MAFW and hope that the release of this database can encourage more research on dynamic FER under unconstrained conditions, *e.g.*, multi-modal emotion recognition, self-supervision FER, video emotion caption, zero-shot AU detection, etc.

REFERENCES

[1] Samuel Albanie, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. 2018. Emotion Recognition in Speech Using Cross-Modal Transfer in the Wild. In *Proceedings of the 26th ACM International Conference on Multimedia*. ACM, 292–301. https://doi.org/10.1145/3240508.3240578

[2] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. ACL, 2236–2246. https://doi.org/10.18653/v1/P18-1208

[3] Adrian Bulat and Georgios Tzimiropoulos. 2017. How Far are We from Solving the 2D & 3D Face Alignment Problem? (and a Dataset of 230,000 3D Facial Landmarks). In *2017 IEEE International Conference on Computer Vision (ICCV)*. 1021–1030. https://doi.org/10.1109/ICCV.2017.116

[4] Zheng Chen, Meiyu Liang, Wanying Yu, Yongkang Huang, and Xiaoxiao Wang. 2021. Intelligent Teaching Evaluation System Integrating Facial Expression and Behavior Recognition in Teaching Video. In *2021 IEEE International Conference on Big Data and Smart Computing (BigComp)*. 52–59. https://doi.org/10.1109/BigComp51126.2021.00019

[5] Xiang Chu and Qiuyan Zhong. 2016. Crowdsourcing quality control model protecting location privacy of workers. *Systems Engineering - Theory & Practice* 36, 8 (2016), 2047–2055. https://doi.org/10.12011/1000-6788(2016)08-2047-09

[6] Lee J. Cronbach. 1951. Coefficient alpha and the internal structure of tests. *Psychometrika* 16, 3 (Sept. 1951), 297–334. https://doi.org/10.1007/BF02310555

[7] Abhinav Dhall, Roland Goecke, Shreya Ghosh, Jyoti Joshi, Jesse Hoey, and Tom Gedeon. 2017. From Individual to Group-Level Emotion Recognition: EmotiW 5.0. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. ACM, 524–528. https://doi.org/10.1145/3136755.3143004

[8] Abhinav Dhall, O.V. Ramana Murthy, Roland Goecke, Jyoti Joshi, and Tom Gedeon. 2015. Video and Image Based Emotion Recognition Challenges in the Wild: EmotiW 2015. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, 423–426. https://doi.org/10.1145/2818346.2829994

[9] Kollias Dimitrios and Zafeiriou Stefanos. 2019. Expression, Affect, Action Unit Recognition: Aff-Wild2, Multi-Task Learning and ArcFace. In *30th British Machine Vision Conference (BMVC)*.

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *9th International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=YicbFdNTTy

[11] Shichuan Du, Yong Tao, and Aleix M. Martinez. 2014. Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences* 111, 15 (2014), 1454–1462. https://doi.org/10.1073/pnas.1322355111

[12] Joy O. Egede, Siyang Song, Temitayo A. Olugbade, Chongyang Wang, Amanda C. De C. Williams, Hongying Meng, Min Aung, Nicholas D. Lane, Michel Valstar, and Nadia Bianchi-Berthouze. 2020. EMOPAIN Challenge 2020: Multimodal Pain Evaluation from Facial and Bodily Expressions. In *15th IEEE International Conference on Automatic Face and Gesture Recognition (FG)*. 849–856. https://doi.org/10.1109/FG47880.2020.00078

[13] Paul Ekman. 1984. Expression and the nature of emotion. *Approaches to Emotion* (1984).

[14] Paul Ekman. 1993. Facial expression and emotion. *American Psychologist* 48, 4 (1993), 384–392. https://doi.org/10.1037/0003-066X.48.4.384

[15] Panagiotis Paraskevas Filntisis, Niki Efthymiou, Petros Koutras, Gerasimos Potamianos, and Petros Maragos. 2019. Fusing Body Posture With Facial Expressions for Joint Recognition of Affect in Child–Robot Interaction. *IEEE Robotics and Automation Letters* 4, 4 (2019), 4011–4018. https://doi.org/10.1109/LRA.2019.2930434

[16] Tobias Gehrig and Hazım Kemal Ekenel. 2013. Why is Facial Expression Analysis in the Wild Challenging?. In *Proceedings of the 2013 on Emotion Recognition in the Wild Challenge and Workshop*. ACM, 9–16. https://doi.org/10.1145/2531923.2531924

[17] F.A. Gers, J. Schmidhuber, and F. Cummins. 1999. Learning to forget: continual prediction with LSTM. In *9th International Conference on Artificial Neural Networks (ICANN)*, Vol. 2. 850–855. https://doi.org/10.1049/cp:19991218

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 770–778. https://doi.org/10.1109/CVPR.2016.90

[19] Zhenliang He, Meina Kan, Jie Zhang, Xilin Chen, and Shiguang Shan. 2017. A fully end-to-end cascaded cnn for facial landmark detection. In *12th IEEE International Conference on Automatic Face Gesture Recognition (FG)*. 200–207. https://doi.org/10.1109/FG.2017.33

[20] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9, 8 (Nov. 1997), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

[21] Xingxun Jiang, Yuan Zong, Wenming Zheng, Chuangao Tang, Wanchuang Xia, Cheng Lu, and Jiateng Liu. 2020. DFEW: A Large-Scale Database for Recognizing Dynamic Facial Expressions in the Wild. In *Proceedings of the 28th ACM International Conference on Multimedia*. ACM, 2881–2889. https://doi.org/10.1145/3394171.3413620

[22] Rie Johnson and Tong Zhang. 2017. Deep Pyramid Convolutional Neural Networks for Text Categorization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, Vol. 1. ACL, 562–570. https://doi.org/10.18653/v1/P17-1052

[23] Jiyoung Lee, Seungryong Kim, Sunok Kim, Jungin Park, and Kwanghoon Sohn. 2019. Context-Aware Emotion Recognition Networks. In *IEEE/CVF International Conference on Computer Vision (ICCV)*. 10142–10151. https://doi.org/10.1109/ICCV.2019.01024

[24] Haoxiang Li, Zhe Lin, Xiaohui Shen, Jonathan Brandt, and Gang Hua. 2015. A convolutional neural network cascade for face detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5325–5334. https://doi.org/10.1109/CVPR.2015.7299170

[25] Shan Li and Weihong Deng. 2020. Deep Facial Expression Recognition: A Survey. *IEEE Transactions on Affective Computing* 01 (mar 2020), 1–1. https://doi.org/10.1109/TAFFC.2020.2981446

[26] Shan Li, Weihong Deng, and JunPing Du. 2017. Reliable Crowdsourcing and Deep Locality-Preserving Learning for Expression Recognition in the Wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2584–2593. https://doi.org/10.1109/CVPR.2017.277

[27] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep Learning Face Attributes in the Wild. In *IEEE International Conference on Computer Vision (ICCV)*. 3730–3738. https://doi.org/10.1109/ICCV.2015.425

[28] Beth Logan. 2000. Mel Frequency Cepstral Coefficients for Music Modeling. In *1st International Symposium on Music Information Retrieval (ISMIR)*.

[29] Patrick Lucey, Jeffrey F. Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. 2010. The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*. 94–101. https://doi.org/10.1109/CVPRW.2010.5543262

[30] Michael Lyons, Shigeru Akamatsu, Miyuki Kamachi, and Jiro Gyoba. 1998. Coding facial expressions with Gabor wavelets. In *Third IEEE International Conference on Automatic Face and Gesture Recognition*. 200–205. https://doi.org/10.1109/AFGR.1998.670949

[31] Albert Mehrabian. 1981. *Silent Messages*. Wadsworth Publishing Company, Belmont, CA.

[32] Debin Meng, Xiaojiang Peng, Kai Wang, and Yu Qiao. 2019. Frame Attention Networks for Facial Expression Recognition in Videos. In *IEEE International Conference on Image Processing (ICIP)*. 3866–3870. https://doi.org/10.1109/ICIP.2019.8803603

[33] Mehdi Noroozi and Paolo Favaro. 2016. Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles. In *European Conference on Computer Vision (ECCV)*. 69–84.

[34] M. Pantic, M. Valstar, R. Rademaker, and L. Maat. 2005. Web-based database for facial expression analysis. In *IEEE International Conference on Multimedia and Expo (ICME)*. 317–320. https://doi.org/10.1109/ICME.2005.1521424

[35] R. Plutchik. 2000. A general psychoevolutionary theory of emotion. *Emotion Theory Research & Experience* 21, 4-5 (2000), 529–553.

[36] Maryam Pourebadi and Laurel D. Riek. 2022. Facial Expression Modeling and Synthesis for Patient Simulator Systems: Past, Present, and Future. *ACM Trans. Comput. Healthcare* 3, 2, Article 23 (mar 2022), 32 pages. https://doi.org/10.1145/3483598

[37] A. P. Dawida. M. Skene. 1979. Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm. *Journal of the Royal Statistical Society* 28, 1 (1979), 20–28.

[38] Nummenmaa Tapio. 1988. The recognition of pure and blended facial expressions of emotion from still photographs. *Scandinavian Journal of Psychology* 29, 1 (1988), 33–47. https://doi.org/10.1111/j.1467-9450.1988.tb00773.x

[39] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning Spatiotemporal Features with 3D Convolutional Networks. In *IEEE International Conference on Computer Vision (ICCV)*. 4489–4497. https://doi.org/10.1109/ICCV.2015.510

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All You Need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS)*. Curran Associates Inc., 6000–6010.

[41] Xing Zhang, Lijun Yin, Jeffrey F. Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, Peng Liu, and Jeffrey M. Girard. 2014. BP4D-Spontaneous: a high-resolution spontaneous 3D dynamic facial expression database. *Image and Vision Computing* 32, 10 (2014), 692–706. https://doi.org/10.1016/j.imavis.2014.06.002

[42] Ying Zhou, Hui Xue, and Xin Geng. 2015. Emotion Distribution Recognition from Facial Expressions. In *Proceedings of the 23rd ACM International Conference on Multimedia*. ACM, 1247–1250. https://doi.org/10.1145/2733373.2806328