DFEW: A Large-Scale Database for Recognizing Dynamic Facial Expressions in the Wild

Xingxun Jiang*
Yuan Zong*
jiangxingxun@seu.edu.cn
xhzongyuan@seu.edu.cn
School of Biological Science and Medical Engineering, Southeast University
Nanjing, China

Wenming Zheng†
Key Laboratory of Child Development and Learning Science, Southeast University
Nanjing, China
wenming_zheng@seu.edu.cn

Chuangao Tang
School of Biological Science and Medical Engineering, Southeast University
Nanjing, China
tcg2016@seu.edu.cn

Wanchuang Xia
School of Cyber Science and Engineering, Southeast University
Nanjing, China
xiawanchuag@seu.edu.cn

Cheng Lu
School of Information Science and Engineering, Southeast University
Nanjing, China
cheng.lu@seu.edu.cn

Jiateng Liu
School of Biological Science and Medical Engineering, Southeast University
Nanjing, China
Jiateng_Liu@seu.edu.cn

ABSTRACT

Recently, facial expression recognition (FER) in the wild has gained a lot of researchers' attention because it is a valuable topic to enable the FER techniques to move from the laboratory to the real applications. In this paper, we focus on this challenging but interesting topic and make contributions from three aspects. First, we present a new large-scale 'in-the-wild' dynamic facial expression database, DFEW (Dynamic Facial Expression in the Wild), consisting of over 16,000 video clips from thousands of movies. These video clips contain various challenging interferences in practical scenarios such as extreme illumination, occlusions, and capricious pose changes. Second, we propose a novel method called Expression-Clustered Spatiotemporal Feature Learning (EC-STFL) framework to deal with dynamic FER in the wild. Third, we conduct extensive benchmark experiments on DFEW using a lot of spatiotemporal deep feature learning methods as well as our proposed EC-STFL. Experimental results show that DFEW is a well-designed and challenging database, and the proposed EC-STFL can promisingly improve the performance of existing spatiotemporal deep neural networks in coping with the problem of dynamic FER in the wild. Our DFEW database is publicly available and can be freely downloaded from https://dfew-dataset.github.io/.

CCS CONCEPTS

• **Human-centered computing** → **Visualization design and evaluation methods**; • **Computing methodologies** → **Computer vision**.

KEYWORDS

Dynamic facial expression; Facial expression database; in-the-wild facial expression recognition; deep learning

ACM Reference Format:
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 (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3394171.3413620

1 INTRODUCTION

Facial expression is one of the most naturally pre-eminent ways for human beings to communicate their emotions in daily life [3]. Imagine that if computers were able to understand emotions from facial expressions as human beings, our human-computer interaction (HCI) systems would be more friendly and natural. Due to this reason, facial expression recognition (FER) has become a hot research topic among HCI and multimedia analysis communities. Over the past decades, researchers have proposed a lot of well-performing methods for recognizing facial expressions, and these methods achieved promising performance in the lab-controlled environments [8, 24, 30, 41–43]. However, FER techniques are still far from the practical applications. One of the main reasons is that the facial expressions in the lab-controlled scenarios are different from the real-world ones. The unconstrained real-world facial expression often suffers from occlusions, illumination variation, pose

*Both authors contributed equally to this research.
†Corresponding author

MM '20, October 12–16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413620

Table 1: Summary of existing databases of dynamic facial expression in the wild.

Database	#Sample	Source	Expression Distribution	#Annotation Times	Available?
Aff-Wild [17]	298	Web	Valence-arousal	8	Yes
AFEW 7.0 [4]	1,809	54 Movies	7 basic expressions	2	Yes
AFEW-VA [18]	600	AFEW database	Valence-arousal	2	Yes
CAER [19]	13,201	79 TVshows	7 basic expressions	3	Yes
DFEW	16,372	1500 movies	7 basic expressions	10	Yes

changes, and many other unpredictable and challenging interferences, making the performance of most existing FER techniques drop sharply. For this reason, many researchers have recently shifted their focus to a challenging but meaningful FER topic, i.e., **FER in the wild**, where 'in the wild' refers to the challenging conditions in unconstrained real-world environments.

Similar to conventional FER, FER in the wild can be divided into two types of task according to the form of samples. One is static FER in the wild, whose aim is to predict the expression category from unconstrained facial images. The other is dynamic FER in the wild, in which the data describing the expression information, is the video clip or image sequence. Inspired by the success of deep learning in many vision tasks, some researchers have begun to construct large-scale facial expressions in the wild databases by resorting to the Internet that contains abundant facial expression resources. For example, Benitezquiroz et al. [1] collected facial images from the Internet and then created a large-scale static facial expression in the wild database called EmotioNet. EmotioNet includes 1,000,000 facial expression images, in which 25,000 images were manually labeled with 11 facial Action Units (AUs). Subsequently, Mollahosseini et al. [28] constructed a much larger-volume database, i.e., AffectNet, consisting of 450,000 well-labeled facial image samples queried from the Internet. Recently, Li et al. [20, 21] presented a novel static facial expression database, RAF-DB, containing nearly 30,000 web-queried facial images. Compared with EmotioNet and AffectNet, the major advantage of RAF-DB is the annotation. RAF-DB collectors hired 315 individuals as the annotators, and each sample in RAF-DB is labeled about 40 times to ensure its labeling reliability.

Unfortunately, in contrast to the static facial expressions in the wild, only a few unconstrained dynamic facial expression databases have been released until now. In the work of [5], Dhall et al. built a dynamic facial expression in the wild database, i.e., acted facial expressions in the wild (AFEW), which has been updated to the 7th version (AFEW 7.0) [4] and consists of 1,809 video clips collected from 54 movies. Recently, Lee et al. [19] built a large-scale benchmark for dynamic FER in the wild, called CAER, by collecting 13,201 video clips from 79 TV shows. Each clip was individually labeled by three annotators. To the best of our knowledge, CAER is the first large-scale database of dynamic facial expression in the wild. However, due to the lack of large-scale databases, the progress of deep learning methods for **dynamic FER in the wild** is seriously hindered. For example, in EmotiW2019, the annual emotion recognition challenge held at ACM ICMI based on the AFEW database, Li et al. [22] proposed a weighted fusion method integrating multiple prediction scores learned by different spatiotemporal feature learning networks, and won the champion. Nevertheless, the accuracy of the test set they achieved is only 62.78% (7 expression classification task), which is still at a low level and does not meet the requirement of practical applications.

In order to remove the barrier of data volume to the research of dynamic FER in the wild, in this paper, we first present a new large-scale and well-annotated unconstrained dynamic facial expression database, DFEW (Dynamic Facial Expression in the Wild). DFEW can be served as a benchmark for researchers to develop and evaluate their methods for dealing with dynamic FER in the wild. To see the characteristics of DFEW, we summarize existing databases of dynamic facial expressions in the wild in Table 1. From Table 1, it can clearly be seen that our DFEW has three major advantages over existing databases including Aff-Wild [17], AFEW 7.0 [4], AFEW-VA [18], and CAER [19]. First, DFEW database has currently largest number of dynamic facial expression samples reaching over 16,000 video clips. Second, the forms of scene and sample in DFEW are many and varied because its video clips are collected from over 1,500 movies all over the world covering various challenging interferences, e.g., extreme illuminations, self-occlusions, and capricious pose changes. Last but not least, each sample in DFEW has been individually labeled ten times by the annotators under professional guidance.

In addition to DFEW, we also propose a novel method called Expression-Clustered Spatiotemporal Feature Learning (EC-STFL) framework to deal with dynamic FER in the wild. EC-STFL framework can enforce the spatiotemporal deep neural networks, e.g., C3D [37] and P3D [32], to better learn discriminative features describing dynamic facial expressions in the wild. Finally, we establish a benchmark evaluation protocol for DFEW and conduct extensive experiments using many spatiotemporal deep learning methods as well as our proposed EC-STFL. Experimental results show that the proposed EC-STFL framework can promisingly improve the performance of existing spatiotemporal neural networks in coping with FERW problem.

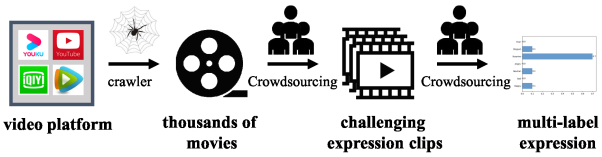

Figure 1: Overview of the construction and the annotation of DFEW.

Figure 2: Examples of seven basic emotions from single-labeled DFEW.

2 DFEW DATABASE

2.1 Data Collection

It is believed that movies originate from and mimic our real life, hence actresses and actors in movies may have all kinds of unconstrained facial expressions originally existing in the practical scenarios. Thus it offers us abundant samples of dynamic facial expressions. By extracting the video clips containing different facial expressions from movies, we are able to build a large-scale database of dyanmic facial expressions in the wild. Following this method, several dynamic facial expression databases, e.g., Aff-Wild [17], AFEW [4, 5], and CAER [19] have been successively built and released over the past few years, which indeed advances the research of dynamic FER in the wild. In this paper, we also take full advantage of movies to collect unconstrained dynamic facial expression samples to build our DFEW database.

The pipeline of building the DFEW database is shown in Fig. 1. As Fig. 1 shows, we first make use of crawler to collect over 1,500 high-definition movies close to our real life and covering various themes, e.g., comedy, tragedy, war, and love, from the Internet to serve as the sample source of facial expressions in the wild. Then, we hired dozens of students to use video editing software to manually extract video clips containing one of seven basic expressions from their assigned movies. Note that we made several rules to help these student extractors ensure the diversity of their extracted facial expression samples. For example, the students are only allowed to extract at most 20 video clips from each movie. Meanwhile, an additional reward would be given to one student if he or she submitted the samples of relatively rare facial expressions, e.g., disgust and fear. Through the above method, we ultimately collected 16,372 unconstrained facial expression video clips.

2.2 Data Annotation

High-quality data annotation is another challenge for the database. First of all, annotating such a large database is time-consuming and needs efficient personnel management. Second, though psychologists P. Ekman believes that the seven basic emotions are universal and independent of the cultural mismatch [6], culture mismatch indeed exists and worth considering because the labeling bias can be removed as far as possible. To efficiently manage annotators and understand the protagonist's emotion in clips better, we entrust the labeling work to the professional crowdsourcing company, JD crowdsourcing [1], where we hired twelve expert annotators. They are asked to identify each clip's closest emotion in seven typical discrete emotions, i.e., anger, disgust, fear, happy, sad, surprise, and neutral. Before formal annotation, these twelve annotators are professionally trained with the emotional knowledges. Then each clip is annotated by ten independent annotators. After annotation, we obtained the seven-dimensional emotion vectors or emotion distribution annotating information of 16,372 clips.

We suppose the seven-dimensional emotion ground truth of j-th video clip denoted by $L_j = \{l_1, ..., l_k, ..., l_7\}$, where l_k represents the annotation times of k-th emotion labeled by annotators, $k \in \{1, 2, 3, 4, 5, 6, 7\}$ refer to happy, sad, neutral, angry, surprise, disgust and fear, respectively.

[1] http://weigong.jd.com/

Table 2: The basic information of single-labeled DFEW.

Emotions	Clips				Percent
	0-2s	2-5s	5s+	Total	
Happy	852	1252	384	2488	20.63
Sad	440	915	653	2008	16.65
Neutral	832	1335	542	2709	22.46
Angry	762	1091	376	2229	18.48
Surprise	691	648	159	1498	12.42
Disgust	71	58	17	146	1.22
Fear	408	435	138	981	8.14
Total	4056	5734	2269	12059	100.00

However, not all clips can be further clearly assigned to a specific single-labeled emotion category from multi-dimensional emotion distribution. Therefore, for accurate labeling, we pick out the emotion k as the single label with respect to $l_k > r$, where r is the threshold value of annotation times. In this work, we set the threshold value $r = 6$, hence select 12059 clips of DFEW to be the single-labeled. We provide basic information of single-labeled DFEW in Table 2, and demo samples of single-labeled DFEW in Fig. 2. Note that, to promote emotion research, we will release both single-labeled annotation and seven-dimensional emotion distribution annotation.

2.3 Agreement Test

In this section, we discuss the quality of emotion annotation based on Fleiss's Kappa test [10]. Fleiss's Kapaa test calculates the degree of agreement in classification over that which would be expected by chance. We believe that its result is an excellent index to give annotation's reliability or quality. In the task of annotating clips, ten independent individuals annotate each clip with $k \in \{1, 2, 3, 4, 5, 6, 7\}$, i.e., one of the seven typical discrete emotions. Here, we let n_{ij} represent the number of annotators who assigned the i-th clip to the j-th emotion. So we can calculate p_j, the proportion of all assignments which were to the j-th emotion,

$$\begin{cases} p_j = \frac{1}{N \times n} \sum_{i=1}^{N} n_{ij} \\ \sum_{j=1}^{K} p_j = 1 \end{cases} \tag{1}$$

where $n = 10$ is the annotation time of each clips, $K = 7$ is the number of emotion category, and N is the number of clips. And we can calculate P_i, the extent to which annotators agree for the i-th clip, i.e., compute how many annotator-annotator pairs are in agreement, relative to the number of all possible annotator–annotator pairs:

$$P_i = \frac{1}{n \times (n-1)} \left[(\sum_{j=1}^{K} n_{ij}^2) - n \right] \tag{2}$$

And compute $\bar{P}$, the mean of P_i, and $\bar{P_e}$ which go into the formula for coefficient κ:

$$\bar{P} = \frac{1}{N} \sum_{i=1}^{N} P_i \tag{3}$$

$$\bar{P_e} = \sum_{j=1}^{k} p_j^2 \tag{4}$$

Then we can calculate κ by

$$\kappa = \frac{\bar{P} - \bar{P_e}}{1 - \bar{P_e}} \tag{5}$$

We perform Fleiss's Kapaa test both in the whole DFEW database and the single-labeled part, and we obtain $\kappa = 0.70$ for the whole DFEW database and $\kappa = 0.63$ for the single-labeled part. Based on Table 3, we believe that all annotators achieve a substantial agreement. That is to say, our annotation is of high quality.

Table 3: Interpretation of κ for Fleiss' Kapaa Test.

κ	Interpretation
<0	Poor agreement
0.01-0.20	Slight agreement
0.21-0.40	Fair agreement
0.41-0.60	Moderate agreement
0.61-0.80	Substantial agreement
0.81-1.00	Almost perfect agreement

3 EXPRESSION-CLUSTERED SPATIOTEMPORAL FEATURE LEARNING

The challenge of dynamic FERW is how to learn robust and discriminative features to describe facial expression video clips, the facial expression representation of video clips, which are contaminated by the abnormal conditions, such as variations of illumination, posture, occlusion and scale. Spatiotemporal features obtained by the various spatiotemporal neural networks are adept in characterizing the dynamic face motion in video samples from the spatial stream and temporal stream. Because of the strong fitting ability of neural networks, the hierarchical spatiotemporal features perform better than the traditional methods in the anti-noise problem. Unfortunately, the margin of different emotion features distributed in the feature space is still blurring due to those abnormal or challenging conditions. To simultaneously cope with FERW and the make feature margins clear, we propose an Expression-Clustered Spatiotemporal Feature Learning (EC-STFL) framework, which can be embedded in the popular spatiotemporal network flexibly. Drawing on the idea of LDA, the EC-STFL enhances intra-class correlation and reduces inter-class correlation by designing special similarity matrices, and is formulated as follows,

$$\min_{W} \sum_{i,j} \frac{P_{ij} \phi(x_i, x_j)}{Q_{ij} \phi(x_i, x_j)} \tag{6}$$

where W is the network's weight, matrix P and matrix Q are both similarity matrices, $\phi(x_i, x_j) = \|x_i - x_j\|$ is the spatiotemporal feature distance of sample x_i and sample x_j, where $x \in \mathbb{R}^d$ is extracted from the final hidden fully connected layers, i.e., just before the softmax layer that produces the class prediction. And the matrix P and matrix Q are defined as follows:

$$P_{ij} = \begin{cases} 0, & \text{if } x_i \text{ and } x_j \text{ has the same label} \\ 1, & \text{otherwise} \end{cases} \tag{7}$$

$$Q_{ij} = \begin{cases} 0, & \text{if } x_i \text{ and } x_j \text{ has the different label} \\ 1, & \text{otherwise} \end{cases} \tag{8}$$

Obviously, the EC-STFL minimizes the feature distance between the same emotions and maximize the feature distance between different emotions to clarify the emotion margin in spatiotemporal feature space. To implement it more effectively and efficiently, we calculate EC-STFL loss in the mini-batch because of limited memory. Besides, we note that sample unbalance widely exists in the FER task [9, 15, 22, 23], which leading the classifiers prefer the emotions with more samples and ignoring the emotions with fewer samples. The FER task in our DFEW database also faces this trouble. Considering that, we develop the EC-STFL loss by adding dynamic weights to balance different emotions' loss in the update progress of batch loss, and extend EC-STFL loss as follows,

$$L_{EC-STFL} = \frac{\sum\limits_{1 \le i,j \le n, x_j \in \mathcal{N}\{x_i\}} \frac{\|x_i - x_j\|}{N_{x_i}}}{\sum\limits_{1 \le i,j \le n, x_j \notin \mathcal{N}\{x_i\}} \frac{\|x_i - x_j\|}{N_{x_j}}} \tag{9}$$

where $\mathcal{N}\{x_i\}$ is the set of the same single-labeled emotion annotation with x_i in mini-batch, N_{x_i} is the set size of $\mathcal{N}\{x_i\}$, and n is the mini-batch size. Creating the dynamic weights by N_{x_i} and N_{x_j}, EC-STFL adjusts and balances the losses of different emotions in each mini-batch, hence alleviate the imbalance issue of FER task to some degree.

We adopt joint supervision for training softmax loss and our EC-STFL loss to obtain the discriminative spatiotemporal features. The total objective function expressed as $L = L_s + \lambda L_{EC-STFL}$, where L_s denotes softmax loss and hyper-parameter λ is a coefficient used to trade-off L_s and $L_{EC-STFL}$. Note that, we drop the backward step when $L_{EC-STFL}$ has no meaning, i.e., mini-batch only contains samples with one kind of emotion.

4 EXPERIMENTS

In this section, we give an experimental setup for benchmark first, including data preprocessing, experimental protocol, and evaluation metric. Then we conduct extensive spatiotemporal neural network methods for the investigations of our DFEW database, and these networks with EC-STFL loss for the verification. Finally, we make transfer experiments from some widely used action databases and our DFEW database to AFEW database, to verify DFEW can extract adequate and efficient transfer knowledge for the FERW task.

4.1 Experimental Setup

Data&Protocol. To better evaluate the single-labeled DFEW database with 12,059 video clips, we adopt a 5-fold cross-validation protocol for the benchmarks, which means we split all the samples into five same-size parts without overlap to conduct experiments. In each fold (fd1 ~ fd5), one part of samples are used for testing and the remaining for training. Finally, all the predicted labels are used to compute the evaluation metrics by comparing the ground truth.

Preprocessing. First, we use OpenCV to extract image frames from 12,059 clips, face++ API [33] to acquire face region images and facial landmarks. We remove the non-face (undetected) frames and statistics the useful frame rate of clips to eliminate those less than 50%. Totally 362 clips were not taken into consideration. Then, we use SeetaFace [25] for face affine transformation, which normalizes faces based on acquired facial landmarks. Finally, we align temporal length of the remaining clip samples into 16 frames using the time interpolation method in [44, 45].

Evaluation Metric. We choose two metrics [34] widely used in existing researches for evaluating the unbalanced problems, i.e., the unweighted average recall (UAR, i.e., the accuracy per class divided by the number of classes without considerations of instances per class) and weighted average recall (WAR, i.e., accuracy). They are appropriate for the FERW task. The UAR metric indicates the average accuracy of different facial expressions, and we can adequately evaluate the performance of predicting emotions with few samples using the UAR results. The WAR metric indicates the recognition accuracy of overall expressions. We hope to improve models' performance both in UAR and WAR metrics.

Implementation Details. In this paper, we employ the PyTorch framework [31] to implement all models. All models are trained on 12G memory's Titan Xp with an excellent initial learning rate provided by the grid search strategy. And the learning rate reduced at a rate of 10× when loss saturated. First, we train models from scratch to present the benchmarks. Batch size is set to 24, which is the max operational batch size of C3D [37] on Titan Xp. We set trade-off coefficient λ of models with EC-STFL to 10, and trade-off coefficient of center loss to 1×10^{-4} according to [40]. Second, we further discuss EC-STFL about the batch size and trade-off coefficient λ based on C3D [37] and 3D Resnet18 [12]. These experiments are conducted on two Titan Xp. Third, we make cross-database transfer experiments. We finetune some off-the-shelf models initilized by weights provided by other researchers with the best learning rate.

4.2 Experimental Results

Baseline System. The existing spatiotemporal neural networks based on RGB frames can be mainly categorized into two groups: the 3D convolutional neural networks and CNN-RNN networks. In this paper, we conduct five 3D CNN models, i.e., C3D [37], I3D-RGB [2], R3D18 [38], 3D Resnet18 [12], P3D [32], and two CNN-RNN models, i.e., VGG11+LSTM and Resnet18+LSTM for benchmarks. VGG11 [35] and Resnet18 [13] are slightly modified to fit the input size of 112×112. The classification results are shown in Table 4.

Table 4: Comparsion of the seven basic emotion classification performance of C3D, P3D, R3D18, 3D Resnet18, I3D-RGB, VGG11+LSTM, Resnet18+LSTM on DFEW database. The metrics include UAR(unweighted average recall) and WAR(weighted average recall).

Model	Emotions							Metric	
	Happy	Sad	Neutral	Angey	Surprise	Disgust	Fear	UAR	WAR
C3D [37]	75.17	39.49	55.11	62.49	45.00	1.38	20.51	42.74	53.54
P3D [32]	74.85	43.40	54.18	60.42	**50.99**	0.69	23.28	43.97	**54.47**
R3D18 [38]	**79.67**	39.07	57.66	50.39	48.26	3.45	21.06	42.79	53.22
3D Resnet18 [12]	73.13	**48.26**	50.51	**64.75**	50.10	0.00	**26.39**	**44.73**	54.98
I3D-RGB [2]	78.61	44.19	56.69	55.87	45.88	2.07	20.51	43.40	54.27
VGG11+LSTM [11, 14, 35]	76.89	37.65	**58.04**	60.70	43.70	0.00	19.73	42.39	53.70
Resnet18+LSTM [11, 13, 14]	78.00	40.65	53.77	56.83	45.00	**4.14**	21.62	42.86	53.08

Table 5: Expression recognition performance of different methods with and without EC-STFL on DFEW database.

Model	Metric	
	UAR	WAR
C3D	42.74	53.54
C3D,EC-STFL	**45.10**	**55.50**
P3D	43.97	54.47
P3D,EC-STFL	**45.22**	**56.48**
R3D18	42.79	53.22
R3D18,EC-STFL	**45.05**	**56.19**
3D Resnet18	44.73	54.98
3D Resnet18,EC-STFL	**45.35**	**56.51**
I3D-RGB	43.40	54.27
I3D-RGB,EC-STFL	**45.05**	**56.19**
VGG11+LSTM	42.39	53.70
VGG11+LSTM,EC-STFL	**44.78**	**56.25**
Resnet18+LSTM	42.86	53.08
Resnet18+LSTM,EC-STFL	**43.60**	**54.72**

It is seen from Table 4 that P3D [32] achieves the best WAR at 54.47%, and 3D Resnet18 [12] achieves the best UAR at 44.73% among all networks. It is an interesting finding that both UAR and WAR attained by 3D CNN models instead of CNN-RNN models. Among seven types of emotions, 3D CNN better predicts happy, sad, angry, surprise, and fear emotions, while CNN-RNN models better at neural and disgust emotions. One possible reason is that models learn feature existing preference. From Table 4, we can also find that it is easier to classify the happy emotion while harder to the disgust. We can also find that happy emotion is more comfortable to be classified while the disgust is much harder to be well predicted. It may result from the relatively low variance of intra-class facial features for the happy emotion while significant variance for the disgust emotion, or fewer samples of the disgust. In fact, fewer disgust samples mean more serious imbalance problem, which is a widely existed problem leading the lousy performance. To the best of our knowledge, the recognition of disgust emotion is really a hard problem in the FERW task.

EC-STFL. To acquire more discriminative features, we design the EC-STFL and incorporate it with some off-the-shelf 3D convolutional neural networks and CNN-RNN networks. The experiment results with and without EC-STFL are detailed in Table 5. We can find that all EC-STFL based models show better recognition performance than those without this module. Our EC-STFL can promote the UAR and WAR by an average of 1.61 percentage points and 2.08 percentage points, respectively. What is more, comparing with the other models from Table 5, we can find that 3D Resnet18 with EC-STFL achieves the best UAR and WAR results.

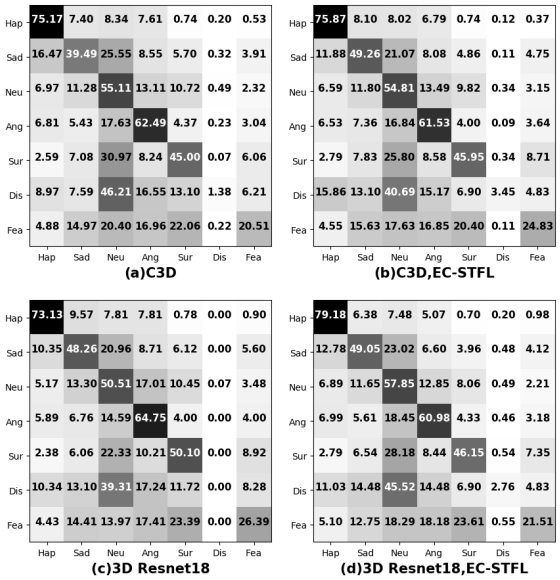

Figure 3: The confusion matrices of selected methods with and without EC-STFL. (a)C3D, (b)C3D with EC-STFL, (c)3DResnet18, (d)3D Resnet18 with EC-STFL.

We provide the recognition performance of different emotion detailed by confusion matrices in Fig. 3, to further discuss classification differences between models with and without EC-STFL. Displayed in Fig. 3, EC-STFL improves the recall rates of the C3D model for happy, sad, surprise, disgust, and fear emotion by 0.7%, 9.77%, 0.95%, 2.07%, and 4.32%, respectively. EC-STFL improves the recall rate of the 3D Resnet18 model for happy, sad, neutral, disgust by 6.05%, 0.79%, 7.34%, 2.76%, respectively. Results are given in Fig. 3 show that our EC-STFL both improve the recall rate of happy, sad, disgust for the C3D and 3D Resnet18.

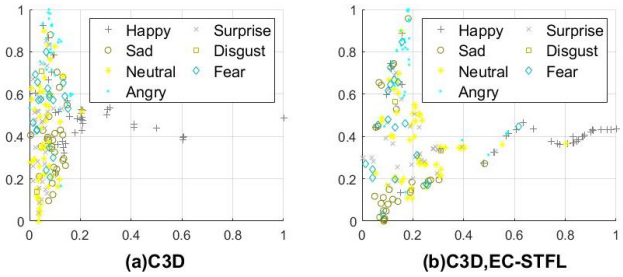

Figure 4: The distribution of deeply features in (a) C3D and (b) C3D with EC-STFL, whose feature dimension is reduced by tSNE. As can be seen, EC-STFL helps the learned features more discriminative.

For a better understanding of the learned features by EC-STFL, we utilize a non-linear mapping method, i.e., t-SNE [27, 39], to visualize the learned features on a 2D plane, as shown in Fig. 4. Compared with the models have no EC-STFL module, we observe that the features learned by EC-STFL show the more significant inter-class distance between different classes; hence the samples show a better aggregation effect. It suggests that our proposed EC-STFL has the ability to promote better feature representation.

The competitor of EC-STFL is mainly the loss inspired by the idea of clustering, e.g., the well-known "center loss" [40]. In this paper, we conduct the comparsion experiments based on two spatiotemporal models, i.e., C3D and 3D Resnet18. Table 6 contains the comparsion of center loss and EC-STFL. As is evident from the Table 6 that EC-STFL and center loss are both improve the classification performance of models purely use cross entropy loss. Furthermore, the EC-STFL performs better than center loss, and achieves the best UAR and WAR.

Hyper-parameters Discussion. The trade-off hyperparameter λ and batch size m affect the performance of EC-STFL, which are both essential to EC-STFL. So we conduct experiments to evaluate models' sensitiveness based on C3D and 3D Resnet18 in the fd1 data split. In the first experiment, we fix batch size $m = 24$ and vary $\lambda \in \{1, 3, 5, 10, 15, 20, 30, 50, 80, 100\}$. It is apparent that properly choosing the value of λ can improve the verification accuracy of the learned features. In the second experiment, we fix $\lambda = 10$ and vary batch size $m \in \{18, 24, 30, 36, 42, 48\}$. The WAR or accuracy results are visible in Fig. 5 and Fig. 6, respectively. Likewise, the verification performance of EC-STFL based models remain largely stable across a wide range of batch sizes.

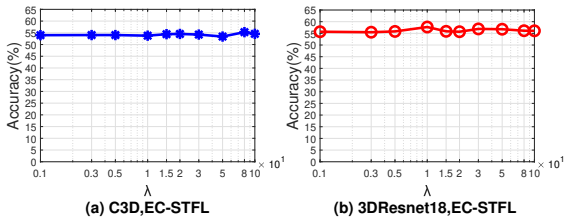

Figure 5: The sensitive experiments results of trade-off parameter for the proposed EC-STFL framework. (a) C3D with EC-STFL, (b) 3D Resnet18 with EC-STFL. The scale of trade-off parameter is $\lambda \in \{1, 3, 5, 10, 15, 20, 30, 50, 80, 100\}$.

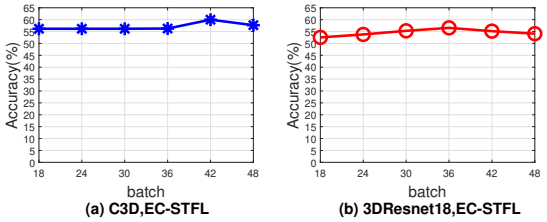

Figure 6: The sensitive experiments results of batch size for the proposed EC-STFL framework. (a) C3D with EC-STFL, (b) 3D Resnet18 with EC-STFL. The scale of batch size is $m \in \{18, 24, 30, 36, 42, 48\}$.

4.3 Transfer Learning

We hypothesize that the DFEW database would contribute to clip-based emotion classification models' transfer learning performance on real-life applications. To verify this hypothesis, we conduct extensive transfer learning experiments from widely used action databases and our DFEW database to the AFEW [5] database. The action databases include UCF101 [36], Sports 1M [16], Kinect 700 [2], and Moments In Time [29]. We select two spatiotemporal neural networks and their EC-STFL version, i.e., C3D, 3D Resnet18, and C3D with EC-STFL, 3D Resnet18 with EC-STFL.

We initialize models with the corresponding pre-trained weights trained from action databases provided by other researchers and our DFEW database respectively, for example, C3D and C3D with EC-STFL use pre-trained weights of C3D model. Then finetune all the layers of network on the AFEW database at a best learning rate searched by grid strategy. Note that, we choose models' pre-trained weights on our DFEW database based on the second data split and the fifth data split, denoted by fd2 and fd5 for short, respectively. We use WAR metric as the evaluation and show the transfer results in Table 7. We found that initial weights provided by the DFEW database show a better transfer learning performance than the action databases. We further compare our transfer results with those state-of-the-arts methods. As results illustrated in Table 8, transferred 3D Resnet18 improve the state-of-the-art method on WAR about 2 percent. In this way, we can conclude that our DFEW database is useful for developing excellent emotion prediction models in real-life applications.

Table 6: Comparison of EC-STFL and center loss on DFEW database.

Model	Emotions							Metric	
	Happy	Sad	Neutral	Angry	Surprise	Disgust	Fear	UAR	WAR
C3D	75.17	39.49	55.11	62.49	45.00	1.38	20.51	42.74	53.54
C3D, center loss	75.62	44.67	54.18	63.14	42.21	2.07	22.17	43.44	54.17
C3D,EC-STFL	75.87	49.26	54.81	61.53	45.95	3.45	24.83	**45.10**	**55.50**
3D Renset18	73.13	48.26	50.51	64.75	50.10	0.00	26.39	44.73	54.98
3D Resnet18, center loss	78.49	44.30	54.89	58.40	52.35	0.69	25.28	44.91	55.48
3D Resnet18,EC-STFL	79.18	49.05	57.85	60.98	46.15	2.76	21.51	**45.35**	**56.51**

Table 7: The transfer learning performance on AFEW7.0.

Pretrained	Finetuned models			
	C3D	C3D, EC-STFL	3D Resnet18	3D Resnet18, EC-STFL
Sports 1M	41.78	44.91	-	-
UCF101	41.25	42.34	-	-
Kinect700	-	-	49.35	49.61
Kinect700+Moments In Time	-	-	49.35	49.35
DFEW, fd2	**44.91**	**45.56**	**53.00**	**53.26**
DFEW, fd5	**49.87**	**49.87**	**49.61**	**49.66**

Table 8: Comparison of 3D Resnet18 model's transfer results with other state-of-the-art methods on AFEW7.0.

Model	WAR
Lu et al. [26]	45.31
Fan et al. [9]	45.43
Hu et al. [15]	46.48
Fan et al. [7]	48.04
Liu et al. [23]	51.44
3D Resnet18,DFEW fd2	**53.00**
3D Resnet18,EC-STFL,DFEW fd2	**53.26**

5 CONCLUSIONS AND FUTURE WORK

In this paper, we have presented a new large-scale unconstrained dynamic facial expression database, DFEW, and proposed a novel spatiotemporal deep feature learning framework, EC-STFL, to deal with dynamic FER in the wild. To the best of our knowledge, our DFEW has the largest number of samples compared with existing databases of dynamic facial expression in the wild, which containing 16,372 video clips extracted from over 1500 different movies. More importantly, DFEW has provided the reliable distribution information of 7 basic expressions for all the video clips because 10 well-trained annotators independently annotate each sample of DFEW. We also conducted extensive baseline experiments on DFEW under the well-designed protocol by using well-performing spatiotemporal deep learning methods as well as the proposed EC-STFL framework and deeply discussed the results. Experimental results showed that our DFEW is a promising unconstrained dynamic facial expression database and the proposed EC-STFL framework can improve the performance of spatiotemporal deep neural networks in coping with dynamic FER in the wild. In the future, we will continue to maintain DFEW by collecting more samples and providing more types of label information such that DFEW can better promote the progress of FER research.

ACKNOWLEDGMENTS

This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFB1305200, in part by the National Natural Science Foundation of China under Grant 61921004, Grant 61902064, and Grant 81971282, and in part by the Fundamental Research Funds for the Central Universities under Grant 2242018K3DN01.

REFERENCES

[1] C Fabian Benitezquiroz, Ramprakash Srinivasan, and Aleix M Martinez. 2016. EmotioNet: An Accurate, Real-Time Algorithm for the Automatic Annotation of a Million Facial Expressions in the Wild. (2016), 5562–5570.
[2] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. 2019. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987* (2019).
[3] Charles Darwin and Phillip Prodger. 1998. *The expression of the emotions in man and animals.* Oxford University Press, USA.
[4] Abhinav Dhall. 2019. EmotiW 2019: Automatic Emotion, Engagement and Cohesion Prediction Tasks. In *2019 International Conference on Multimodal Interaction.* 546–550.
[5] Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. 2012. Collecting large, richly annotated facial-expression databases from movies. *IEEE multimedia* 3 (2012), 34–41.
[6] Paul Ekman and Wallace V Friesen. 1971. Constants across cultures in the face and emotion. *Journal of personality and social psychology* 17, 2 (1971), 124.
[7] Yingruo Fan, Jacqueline CK Lam, and Victor OK Li. 2018. Video-based emotion recognition using deeply-supervised neural networks. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction.* 584–588.

[8] Yingruo Fan, Victor Li, and Jacqueline CK Lam. 2020. Facial Expression Recognition with Deeply-Supervised Attention Network. *IEEE Transactions on Affective Computing* (2020).

[9] Yin Fan, Xiangju Lu, Dian Li, and Yuanliu Liu. 2016. Video-based emotion recognition using CNN-RNN and C3D hybrid networks. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. 445–450.

[10] Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin* 76, 5 (1971), 378.

[11] Felix A Gers, Jürgen Schmidhuber, and Fred Cummins. 1999. Learning to forget: Continual prediction with LSTM. (1999).

[12] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. 2017. Learning spatio-temporal features with 3D residual networks for action recognition. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 3154–3160.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[14] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[15] Ping Hu, Dongqi Cai, Shandong Wang, Anbang Yao, and Yurong Chen. 2017. Learning supervised scoring ensemble for emotion recognition in the wild. In *Proceedings of the 19th ACM international conference on multimodal interaction*. 553–560.

[16] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. 2014. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 1725–1732.

[17] Dimitrios Kollias, Panagiotis Tzirakis, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, Björn Schuller, Irene Kotsia, and Stefanos Zafeiriou. 2019. Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. *International Journal of Computer Vision* 127, 6-7 (2019), 907–929.

[18] Jean Kossaifi, Georgios Tzimiropoulos, Sinisa Todorovic, and Maja Pantic. 2017. AFEW-VA database for valence and arousal estimation in-the-wild. *Image and Vision Computing* 65 (2017), 23–36.

[19] Jiyoung Lee, Seungryong Kim, Sunok Kim, Jungin Park, and Kwanghoon Sohn. 2019. Context-Aware Emotion Recognition Networks. (2019).

[20] Shan Li and Weihong Deng. 2018. Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. *IEEE Transactions on Image Processing* 28, 1 (2018), 356–370.

[21] Shan Li, Weihong Deng, and JunPing Du. 2017. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2852–2861.

[22] Sunan Li, Wenming Zheng, Yuan Zong, Cheng Lu, Chuangao Tang, Xingxun Jiang, Jiateng Liu, and Wanchuang Xia. 2019. Bi-modality Fusion for Emotion Recognition in the Wild. In *2019 International Conference on Multimodal Interaction*. 589–594.

[23] Chuanhe Liu, Tianhao Tang, Kui Lv, and Minghao Wang. 2018. Multi-feature based emotion recognition for video clips. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*. 630–634.

[24] Mengyi Liu, Shiguang Shan, Ruiping Wang, and Xilin Chen. 2014. Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1749–1756.

[25] Xin Liu, Meina Kan, Wanglong Wu, Shiguang Shan, and Xilin Chen. 2016. VIPLFaceNet: An Open Source Deep Face Recognition SDK. *Frontiers of Computer Science (FCS)* (2016).

[26] Cheng Lu, Wenming Zheng, Chaolong Li, Chuangao Tang, Suyuan Liu, Simeng Yan, and Yuan Zong. 2018. Multiple spatio-temporal feature learning for video-based emotion recognition in the wild. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*. 646–652.

[27] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, Nov (2008), 2579–2605.

[28] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. 2019. AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild. *IEEE Transactions on Affective Computing* 10, 1 (2019), 18–31.

[29] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, et al. 2019. Moments in time dataset: one million videos for event understanding. *IEEE transactions on pattern analysis and machine intelligence* 42, 2 (2019), 502–508.

[30] Bowen Pan, Shangfei Wang, and Bin Xia. 2019. Occluded Facial Expression Recognition Enhanced through Privileged Information. In *Proceedings of the 27th ACM International Conference on Multimedia*. 566–573.

[31] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch. (2017).

[32] Zhaofan Qiu, Ting Yao, and Tao Mei. 2017. Learning spatio-temporal representation with pseudo-3d residual networks. In *proceedings of the IEEE International Conference on Computer Vision*. 5533–5541.

[33] M. Inc. Face++ research. [n. d.]. toolkit. www.faceplusplus.com.

[34] Bjorn Schuller, Bogdan Vlasenko, Florian Eyben, Martin Wöllmer, Andre Stuhlsatz, Andreas Wendemuth, and Gerhard Rigoll. 2010. Cross-corpus acoustic emotion recognition: Variances and strategies. *IEEE Transactions on Affective Computing* 1, 2 (2010), 119–131.

[35] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[36] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402* (2012).

[37] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*. 4489–4497.

[38] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. 2018. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 6450–6459.

[39] Laurens Van Der Maaten. 2014. Accelerating t-SNE using tree-based algorithms. *The Journal of Machine Learning Research* 15, 1 (2014), 3221–3245.

[40] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. 2016. A discriminative feature learning approach for deep face recognition. In *European conference on computer vision*. Springer, 499–515.

[41] Guoying Zhao and Matti Pietikainen. 2007. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE transactions on pattern analysis and machine intelligence* 29, 6 (2007), 915–928.

[42] Wenming Zheng, Hao Tang, Zhouchen Lin, and Thomas S Huang. 2010. Emotion recognition from arbitrary view facial images. (2010), 490–503.

[43] Wenming Zheng, Xiaoyan Zhou, Cairong Zou, and Li Zhao. 2006. Facial expression recognition using kernel canonical correlation analysis (KCCA). *IEEE Transactions on Neural Networks* 17, 1 (2006), 233–238.

[44] Ziheng Zhou, Xiaopeng Hong, Guoying Zhao, and Matti Pietikäinen. 2013. A compact representation of visual speech data using latent variables. *IEEE transactions on pattern analysis and machine intelligence* 36, 1 (2013), 1–1.

[45] Ziheng Zhou, Guoying Zhao, and Matti Pietikäinen. 2011. Towards a practical lipreading system. In *CVPR 2011*. IEEE, 137–144.