FERV39k: A Large-Scale Multi-Scene Dataset for Facial Expression Recognition in Videos

Yan Wang[1], Yixuan Sun[1], Yiwen Huang[2], Zhongying Liu[2], Shuyong Gao[2], Wei Zhang[2], Weifeng Ge[2*] and Wenqiang Zhang[1, 2*]
[1]Academy of Engineering & Technology, Fudan University, Shanghai, China
[2]School of Computer Science, Fudan University, Shanghai, China
{yanwang19}@fudan.edu.cn; {21210860014, 21210240056, zhongyingliu21}@m.fudan.edu.cn; {18110240022, weizh, wfge and wqzhang}@fudan.edu.c

Figure 1. An overview of FERV39k composed of video frames of 7 basic expressions across 4 scenarios subdivided by 22 scenes.

Abstract

Current benchmarks for facial expression recognition (FER) mainly focus on static images, while there are limited datasets for FER in videos. It is still ambiguous to evaluate whether performances of existing methods remain satisfactory in real-world application-oriented scenes. For example, the "Happy" expression with high intensity in Talk-Show is more discriminating than the same expression with low intensity in Official-Event. To fill this gap, we build a large-scale multi-scene dataset, coined as FERV39k. We analyze the important ingredients of constructing such a novel dataset in three aspects: (1) multi-scene hierarchy and expression class, (2) generation of candidate video clips, (3) trusted manual labelling process. Based on these guidelines, we select 4 scenarios subdivided into 22 scenes, annotate 86k samples automatically obtained from 4k videos based on the well-designed workflow, and finally build 38,935 video clips labeled with 7 classic expressions. Experiment benchmarks on four kinds of baseline frameworks were also provided and further analysis on their performance across different scenes and some challenges for future research were given. Besides, we systematically investigate key components of DFER by ablation studies. The baseline framework and our project are available on ***url****.*

1. Introduction

Facial expression recognition (FER) in static images [1] or videos [2] is of great importance to many applications, such as human-computer interaction (HCI) [3] and lie detection [4]. With millions of images uploaded every day by users from different events and social gatherings, there are various available large-scale datasets for static FER, such as RAF-DB [5] and AffectNet [6]. On top of these datasets, various methods [7]–[10] are designed to understand human emotion and recognize facial expressions. In contrast to static image FER, there are only a few video-based facial expression datasets. In the early period, researchers paid attention to in-the-lab datasets, such as CK+ [11] and Oulu-CASIA [12], which are collected from lab environments and contain limited posed video clips with no more than 30 frames. Recently, recognizing expressions from in-the-lab short video clips has achieved considerable progress [13]–[15], but these models often fail to be directly applied for in-the-wild scenes. Typically, limited samples without complex and varied scene context might be impractical for real-world applications.

With the development of AFEW competition [16], [17], video-based in-the-wild datasets are released progressively,

Dataset (Year)	Samples	Emo.	Anno.	Best.	Context	Scene	Video Sources				
							Lab Shot	Movie	TV	Live Show	Others
CK+ (2010) [11]	327	7 Exps	1	99.69	Lab		√				
Oulu-CASIA (2011) [12]	560	7 Exps	1	92.7	Lab		√				
Aff-Wild (2017) [18]	298	V-A	8	N/A	Wild			√	√		√
AFEW-VA (2017) [16]	600	V-A	2	N/A	Wild			√			
AFEW 8.0 (2018) [17]	1,809	7 Exps	2	53.26	Wild			√			
CAER (2019) [19]	13,201	7 Exps	3	77.04	Wild				√		
DFEW (2020) [20]	16,372	7 Exps	10	56.41	Wild			√			
FERV39k (2021)	39,546	7 Exps	30	N/A	Wild	√		√	√	√	√

Table 1. Comparison of statistics of existing available DFER datasets and our built FERV39k. (Emo. = Emotion distribution; Anno. = Annotation times; Best. =Best accuracy; Exps = Expressions; V-A = Valence-Arousal.

but their video clips are limited and not enough for developing deep FER models. Although the seeming datasets, such as CAER [19] and DEFW [20], claim that their sources of videos are diverse, there exist some limitations in these datasets. For CAER [19], data volume reaches 13k, however, its scene is single and lacks challenge to FER methods. DEFW [20] is a large-scale and well-annotated unconstrained dataset for FER in videos, but it fails to consider and further differentiate scene categories [21], which are essential for application-oriented expression recognition. Besides, these works all overlook how to automatically generate abundant candidate video clips for manual annotation to meet the need of building a larger-scale dataset.

It is necessary to build a multi-scene dataset to advance the FER in video. The benchmark should satisfy several important requirements to cover realistic challenges. 1) Considering the complexity of real-world applications, selected scenes should cover various aspects. 2) With billions of videos currently accessed from the Internet and video platforms, there is an urgent need for robust algorithms that can automatically generate massive video clips. 3) Due to the complexity of facial expression annotations, the workflow of annotating video clips needs to be well-designed. Based on the above guidelines, we build FERV39k (Figure 1), which is a large-scale, multi-scene, high-quality dataset, and contains 38,935 video clips labeled with 7 classic expressions in 4 scenarios: Daily Life, Weak-Interactive Shows, Strong-Interactive Activities, and Anomaly Issues. We design scenarios and scenes for four reasons: 1) Plenty of video sources and samples. 2) Expandability of 22 fine-grained scenes. 3) Large variations and limited overlapping. 4) Distinct associations with scene context. Besides, we design a four-stage strategy, which itself generates 86k candidate video clips from 4k raw videos.

Specifically, our built FERV39k has 3 main characteristics: 1) Multi-scene: clips are divided into 4 scenarios and subdivided into 22 scenes with different characteristics. 2) Large-scale: the amount of video clips reaches 39k with last time from 0.5s to 4s, which indicates that available video frames and cropped facial images reach 1M with the resolution of 336×504, and 224×224, respectively. 3) High-quality: workflow of crowdsourcing and professional annotation is adopted to ensure high-quality labels with the guidance of fine-grained expressions.

Given the well-annotated and multi-scene video clips in our built dataset, we first benchmark four kinds of deep learning-based architectures for FER in videos on the challenging FERV39k following action recognition baselines [22]–[24]. We then perform several baseline evaluations with four baselines and representative backbones to reveal challenging aspects of multi-scene expression representation in videos. According to our analysis on FERV39k benchmark, we uncover several new challenges: 1) difficulty and confusion of 7 basic expression classes. 2) discrepancy across 4 scenarios. 3) unsatisfactory cross-scenario performance. 4) long-tail distribution of expressions and duration. To systematically enumerate key components in modeling DFER based on the four baseline architectures on FERV39k, we further carry out several ablation studies and figure out some significant findings: 1) Pre-training on large-scale datasets is not always helpful. 2) More sampling fails to steadily improve performance. 3) Scene information plays a complementary role on DFER.

In summary, our work has three main contributions: 1) We construct a novel large-scale multi-scene FERV39k dataset for both intra-scene and inter-scene DFER. The dataset contains 38,935 video clips labeled with 7 classic expressions across 22 fine-grained scenes in 4 isolated scenarios. To our best knowledge, this is the first dynamic FER dataset with 39K clips, scenario-scene division as well as cross-domain supportability. 2) We proposed four-stage candidate clip generation and two-stage annotation workflow with a balance between cost and quality control which can be used in other large-scale facial video dataset construction. 3) We benchmark four kinds of deep learning-based architectures and conduct in-depth studies of FERV39k, which reveal the key challenges of our dataset and indicate new directions of future research according to extensive ablation studies.

2. Related Work

2.1. Video-based Datasets for DFER

Video-based FER datasets [12], [17], [19] have been proposed since the start of the research on facial expressions. In the earlier time, the participants were required or induced to perform targeted facial expressions in the controlled environments to collect data such as CK+ [11] and Oulu-CASIA [12]. However, subject to the scale of participants and experiment conditions, in-the-lab datasets are usually small-scaled and in which facial expressions are usually far from the real-world expressions. Besides, most methods [2], [25] (Table 1) have already obtained excellent performance on these benchmarks.

As a result, more attention is attracted by datasets collected from in-the-wild conditions with naturalistic emotion states, such as AFEW [17], Aff-Wild [18], AFEW-VA [16], CAER [19], and DFEW [20]. AFEW [17] is the first in-the-wild dataset proposed in 2013 which contains 1,809 clips of 330 subjects labeled twice with seven labels. AFEW-VA [16] provides more subjects, samples, and professional annotations as well as valence-arousal annotation. CAER [19], increases the number of video clips to 13,201 and considers cropped face and context information. DFEW [20] expands the scale and diversity of data and improves the annotation quality. Table 1 compares statistics among existing datasets with our built FERV39k, which has the following characteristics: 1) the largest number of samples reaches 39k obtained from 86k automatically generated candidate video clips. 2) the well-designed workflow of annotation in the combination of crowdsourcing and professional review. 3) The hierarchy design of two-level scenes is creative to help application-oriented DFER and Cross-domain learning in different contexts. 4) All raw videos are collected from cross-platform sources.

2.2. Dynamic FER Approaches

Though various methods can recognize expressions from static images [9], dynamic videos usually contain more information including the movement of appearance as well as other temporal information. There are two kinds of network structures, named 3-dimensional convolutional networks (3D ConvNet) and 2D ConvNet-LSTM, commonly used for DFER. The 3D ConvNet-based methods [26], [27] use 3D ConvNet extracting spatio-temporal features and generating embedding for DFER. For example, the works [26], [28] use C3D [26] for local spatio-temporal feature extraction. The 2D ConvNet-LSTM based methods [2], [29], [30] combine the CNNs and the LSTM for extracting spatial features and learning temporal modeling, respectively. Most works [31] mainly rely on the analysis of cropped face regions, ignoring scene context information for emotion recognition in the wild. To solve these limitations, Lee et al. [19] investigated the influence of context information by a two-stream encoding network (CAER-Net) which utilizes face encoding stream and context encoding stream to encode the cropped face region and context information, separately. With analysis and comparison among existing video-based representation architectures on whether convolutional layers use 2D or 3D kernels, and whether the input to the network includes scene context, we design four kinds of baseline architectures.

3. The FERV39k Dataset

To introduce a novel and challenging benchmark for application-oriented DFER, we propose a well-designed procedure of dataset construction to build our FERV39k with high-quality annotations. The FERV39k is more challenging and inspiring than previous ones in multiple application scenes, cross-domain learning supportability, automatic candidate clip selection and two-stage efficient & highly credible annotation. While other types of annotations based on these data will be included in succeeding versions, e.g., frame-level annotation with key expression, the current version of FERV39k mainly provides annotations for DFER on 4 isolated scenarios with 22 fine-grained scenes labeled by 7 basic expressions.

3.1. Key Challenges

Inspired by the key challenges [32], we consider a series of unprecedented difficulties and scheme the corresponding strategies, which are followed as:

How to define and generate the scenes and expressions? Since thousands of contexts/scenes and dozens of facial expressions occurred systematically in all countries, it is impractical to fulfill the all-scene task in work [21]. Fortunately, we analyze the findings and conclusions from the work of Cowen et al. [21], which help us summarize 4 scenarios consisting of 22 scenes as well as the 7 basic expressions. Furthermore, a novel scene-based keyword list and fine-grained labels are designed.

How to automatically generate candidate video clips? Different from static facial images crawled from the Internet based on keywords [6], extra segmentation is required to obtain short-duration video clips with a single expression due to the story complexity of a video or movie. Generally, the pipeline of candidate video clips collection for a DFER dataset is crawling large-scale videos (metadata) from the Internet and cropping the single expression clips manually. However, manual operation is costly for a large-scale dataset. Therefore, a novel four-stage FER-based video segmentation process is proposed.

How to design annotation procedure with quality control? Crowdsourcing services such as Amazon Mechanical Turk or JD Crowdsourcing are commonly used

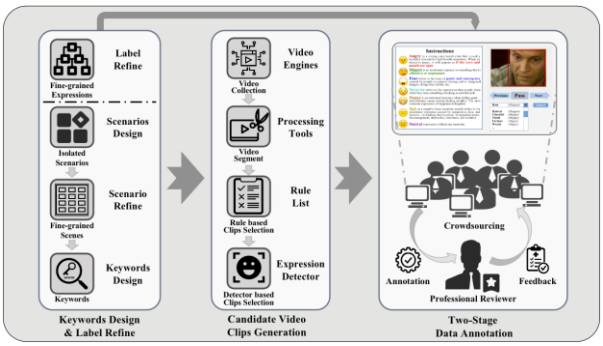

Figure 2. An overview of FERV39k construction.

to build a large-scale dataset. However, discovering the subtle difference between some expressions requires professional knowledge. As a result, a two-stage annotation workflow is proposed to get quality-guaranteed annotation with a balance between cost and reliability.

3.2. Dataset Construction Procedure

In Sec. 3.2, we will introduce the three steps for dataset construction named selection of scene vocabulary and expression class, generation of candidate video clips and data annotation (Figure 2).

Selection of scene vocabulary and expression class. Before data collection, we first design the Scene Vocabulary (including their keywords) and Expression Class in parallel. For the Scene Vocabulary, we analyze the statistic results from work [21], select 22 representative scenes and divide them into 4 scenarios: 6 scenes {Argue, Social, School, Medicine, Conflict, and Daily-Life} designed for Daily Life (DL11k), 6 scenes {Action, Scholar-Reports, Speech, Elegant-Art, Live-Show, and Talk-Show} designed for Weak-Interactive Shows (WIS9k), 6 scenes {Business, Experiment, Official-Event, Crime, Interview, Contest} for Strong-Interactive Activities (SIA10k) and 4 scenes {History, Terror, War and Crisis} for Anomaly Issues (AI9k). For our scene-based raw material collection, we also design a keyword list for each scene. And to design our Expression Class, 7 basic expressions namely “Angry”, “Disgust”, “Fear”, “Happy”, “Sad”, “Surprise”, “Neutral” are selected as annotation labels. And we follow the taxonomy defined by Parrot etc. [33] to carefully select 26 words [34] aiming at clarifying the difference of fine-grained emotion classes, which results in the final expression hierarchy shown in Figure 3(a). Following the expression definition [6], [17], [20], we also initialize an expression list and write a handbook to clarify each expression.

Generation of candidate video clips. Following the works [5], [6], [18], [20], online videos originate from real-life environments in different scenes, hence the human expression in the videos can be recognized as real-world facial expression. We start with reviewing top-level 22 scenes and then collecting corresponding online videos, TV shows and movies from searching/video engines. To acquire clips, existing works ask annotators to manually segment video clips with expressions via video editing software. For processing data on a smaller scale, the cost of time and labor is affordable. However, for our 39k clips dataset (raw materials are even more), it seems impractical to extract clips manually. Hence, we adopt a four-stage strategy to collect and generate candidate video clips for multi-scene videos, the pipeline of which is shown in Figure 3(b).

Firstly, we download over 6k metadata with different lasting time from 8 worldwide open-source engines containing Asian, African, and European/American videos via generated keyword list. Afterwards, we sort and randomly remove some of the videos. After this step, 4k pieces of data are left with balanced time distribution of scenes. According to work [35], we randomly segment them into video clips among 0.5-4 seconds. To generate facial clips, we make a rule list to help our well-designed mechanism adaptively and automatically select a twenty-fold number of clips than the expected scale of the final dataset. However, the rule-based selection mechanism is rough for generating a good candidate and manual refinement is still a hard job. As a result, we utilize a pre-trained light-weight ResNet-50 FER detector to refine these clips and generate candidate clips with expression predictions. Finally, with the prospect that the scale of filtered clips is double of the final dataset scale, we randomly remove some clips and keep the latency distribution of estimated expressions fit to the real-world work.

Manual annotation. To achieve the balance of professional annotation and cost control, we design a workflow of annotation-examine for data annotation (Figure 3(c)). In our designed procedure, there are two roles named crowd-sourcing annotator (CA, 20 workers) and professional researcher (PR, 10 workers), respectively. Our goal is to subtly employ PRs to get professional annotation at a lower cost. To further help annotators differentiate our task from many others on the platform as well as make our task as stimulating and engaging as possible, the JD Crowdsourcing establishes a single-page web base on our guidance. The labelling interface is shown in Figure 3(d), in which one video clip, introductions and the bounding box of face area in each frame are provided to assist annotators. Besides, the platform can automatically convert 26 choices into 7 expression labels.

The clips are divided into groups at first (5% of each are PR annotated) and copied 3 times. Then we randomly shuffle the grouped materials and provide them to CAs. CAs are asked to choose the most likely word or “PASS” on the platform. After annotation, group copies are checked via Flag-Recaptured Statistic method. We design 80% and 40% correct rates as two thresholds and mark copies as

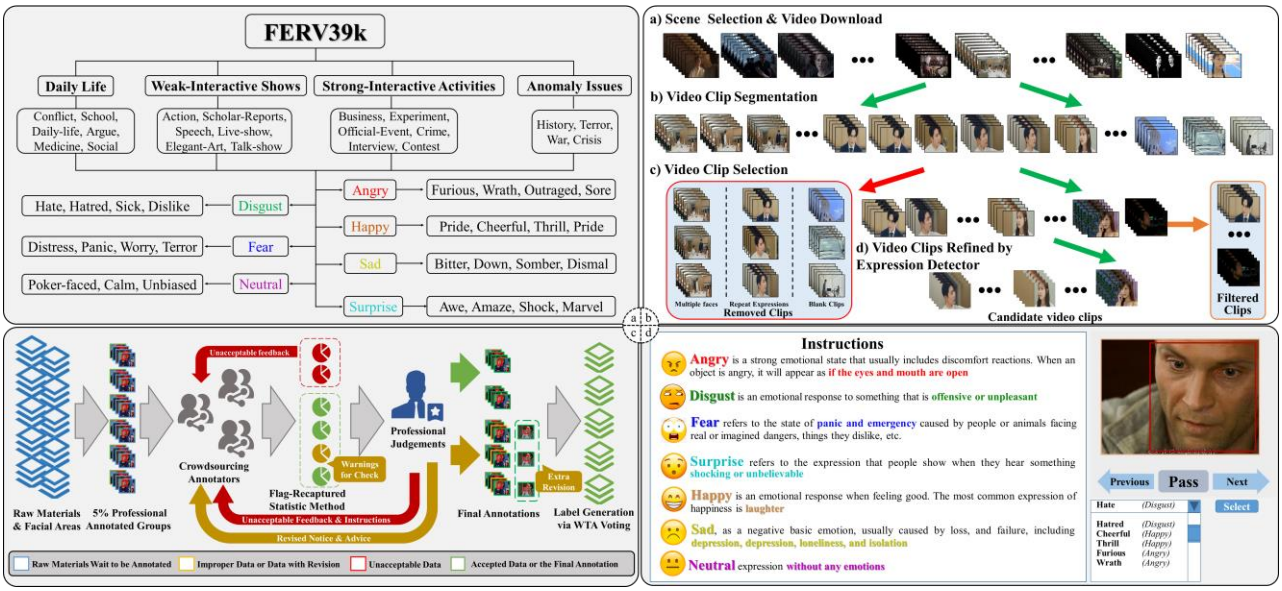

Figure 3. Four important components for dataset construction. (a) Our design of 4 isolated scenarios, 22 scenes, 7 basic expressions and 26 fine-grained expressions. (b) The four-stage generation of candidate video clips in the FERV39k dataset. (c) The procedure of data annotation, statistic evaluation, professional judgement and label generation. (d) The labeling interface in the crowd-sourcing platform.

unacceptable (UA), Improper (IP) and Accept (AC). The IP and AC groups will be passed to PRs for judgement. In this step, PRs only need to decide whether the annotation of a group is acceptable. The UA ones will retreat to CAs and ones still IP will be relabeled by the PRs. For both UA and IP, PRs will provide feedback to CAs. Afterwards, the Weighted-Winner-take-all (WWTA) voting method is used for generating final facial expression labels. To our goal, after iterations on a few groups, the annotators can provide relatively reliable annotations and verification work will become less complex.

3.3. Dataset Statistics

The FERV39k consists of 4 isolated scenarios subdivided by 22 detailed scenes, including nearly 39k video clips labeled with 7 basic facial expressions with an average duration of 1.5 seconds. In general, the clips are evenly distributed in 4 scenarios but the scale of each scene also reflects a severe long-tailed distribution. For further analysis, Figure 4 (left) shows the number of clips in each scene and the distribution of expressions in our built FERV39k, which is used for baseline analysis in this paper. The histogram chart shows a natural long-tailed distribution across 7 basic expressions in different scenes. For instance, “Fear” appears more in the “Terror” scene (18%) and “Happy” appears more in the “Live-Show” scene (33%). This will be a new challenge for DFER models. Figure 4 (right) shows the distribution of expression duration of video clips in different scenes. The large variation of expression duration makes it more difficult for DFER models to accurately localize keyframes like [34]. Moreover, expression instances in FERV39k are often related to longer temporal context and interactions with context. These inherent challenges of FERV39k require a more powerful and flexible temporal modeling scheme for expression detection. **Our built FERV39k can be available under the condition of abiding by the agreement.**

3.4. Dataset Characteristics

Our FERV39k has several distinguishing and attractive characteristics compared with existing datasets.

Large-scale candidate video clips. With the introduction of the four-stage candidate clip generation method, we can cheaply acquire massive candidate video clips, which makes FERV39k possible to be further expanded.

High-quality annotation. With our two-stage annotation strategy, supporting files, fine-grained choices as well as Flag-Recaptured Statistic methods, Professional Judgement and WWTA Voting, FERV39k can get reliable labels at a lower cost.

Task difficulty. With 4 difficulties proposed: 1) large variance of expression duration among clips; 2) different intensities of expressions across different scenes; 3) limited representing frames for labeled expression in a clip; 4) severe long-tailed distribution in different scenes and expressions, FERV39k brings new challenges for DFER methods.

Application-oriented diversity. With a new sight of application, FERV39k pays attention to specific application performance and cross-scene robustness of DFER methods.

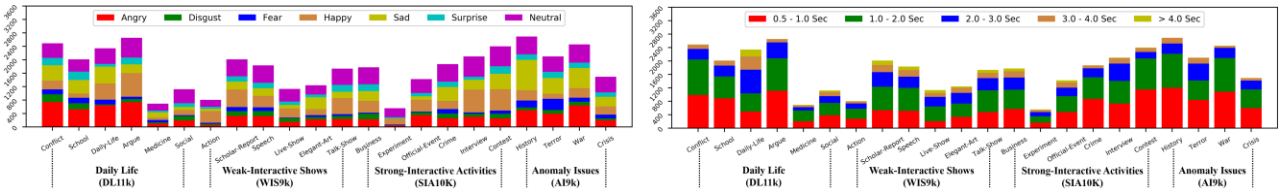

Figure 4. Statistics of our FERV39k. Left chart is the distribution of 7 expressions of video clips in different scenes, sorted by 4 scenarios and 7 different expressions. And Right chart is the distribution of 7 expression duration of video clips in different scenes, which is sorted by 4 scenarios and 5 different time duration.

4. Benchmark Performance

In this section, we will conduct experiments to show the challenges of FERV39k in practical via baseline evaluations and figure out some findings via ablation studies.

4.1. Experiment Setup

FERV39k protocol. To build a solid DFER benchmark, we manually split all data into training set (including validation set), and testing set. In FERV39k Benchmark, video clips of all scenes are randomly shuffled and split into training (80%), and testing (20%) without overlapping, which forms 27 kinds of configurations consisting of 22 setups for each scene, 4 setups for isolated scenarios and 1 setup for all scenes. Cross-scene learning is also available for which some special scenes are used for testing. Besides, we provide cropped face images with 224×224 resolution and scene images with 336×504 resolution to meet the requirement of context-aware DFER methods.

Implementation details. In our experiments, the whole framework is built on PyTorch-GPU using NVIDIA GeForce RTX 2080Ti GPUs. We set learning rate (lr) in a range between 1e-3 and 1e-2, weight decay as 1e-4, and the batch size is fixed at 32 for all architectures. The video clips are taken as input in each epoch with lr as 0.95. All the models are trained from scratch using FERV39k to present the benchmarks for 60 epochs with standard stochastic gradient descent (SGD) with momentum as 0.9 and uniformly sampled frame interval as 8.

Besides, as the number of sequences in FERV39k is limited for training, we exploit the data-augmentation techniques into the training set: randomly cropping, illumination changes and image flip. For reducing the dependence on the computation source, all cropped facial images are resized to 112×112, and whole images are resized to 112×168.

Evaluation metric. Following the standard practice [6], [20], [36] for evaluating FER or DFER, we choose two commonly used metrics: weighted average recall (WAR, also called overall accuracy) and unweighted average recall (UAR).

4.2. Baseline Network

According to the baseline architectures of action recognition in video [22], [23], [37], we first briefly define and describe several standard ConvNet architectures for DFER. We consider four typical approaches for DFER: 2D ConvNet, 2D ConvNet-LSTM on top of [38], 3D ConvNet [26], [39], and Two-Stream 3D ConvNet. We then use these architectures as baselines and compare their performance by training and testing on the whole FERV39k. Table 2 shows the comparison results of four kinds of baseline architectures on FERV39k.

2D ConvNet. Deep CNNs (2D ConvNet) such as VGG [40] and ResNet [41], have made great success on image classification tasks [42]. Hence, we reuse them with minimal change for DFER. For processing a clip, features of all frames can be extracted and flatted into embeddings, which are concatenated and fed into a classifier to obtain results.

2D ConvNet-LSTM. The structure of 2D ConvNet-LSTM is more appropriate for DFER by adding a recurrent layer to the model [43] to introduce temporal information. Hence, we position an LSTM layer with 1024 hidden units and batch normalization layer (as proposed by Cooijmans et al. [44]) after the last average pooling layer of 2D ConvNets. A fully connected layer is added on top as the classifier.

3D ConvNet. 3D ConvNets(e.g., C3D [26] and I3D [23]) can directly model hierarchical representations of spatio-temporal information with spatio-temporal (3D) filters. One issue with 3D ConvNets is that they have much more parameters than 2D ConvNets due to the additional kernel dimension. Besides, extra adjustments of network and output structures are required for DFER.

Two-Stream Networks. Different from the above methods, two-stream networks can encode the context components of the scene, as well as the facial expression of a cropped facial image, together, inspired by CAER [19]. Specifically, we feed sequences of cropped face images and scene frames into the Two-Stream 3D ConvNets and 2D ConvNet-LSTM.

4.3. Baseline Evaluation

On top of FERV39k, we systematically evaluate four kinds of baseline architectures across multiple scenes. Here

Method	All	DL11k	WIS9k	SIA10k	AI9k	Social	DailyLife	Liveshow	Talkshow	Interview	Contest	Experiment	Terror	Crisis
R18	39.33/30.30	39.75/31.36	40.50/28.67	42.31/30.02	33.90/27.20	39.74/33.26	41.40/31.13	37.72/26.82	38.57/25.47	45.75/29.18	48.24/33.37	49.56/26.70	31.28/26.69	36.88/29.21
R50	30.57/22.47	30.46/21.52	32.52/23.50	30.56/22.68	30.14/19.94	27.51/25.05	31.00/19.37	28.51/23.13	28.86/20.14	33.25/21.72	37.06/27.55	31.86/16.89	26.54/19.67	24.25/20.11
VGG13	41.02/31.19	40.40/31.59	43.04/30.23	43.44/29.99	38.86/29.94	48.03/35.50	39.07/28.63	44.74/30.40	40.57/26.25	44.34/28.83	47.62/32.39	49.56/26.52	36.73/31.48	42.52/31.65
VGG16	41.66/32.01	41.81/32.59	42.93/30.77	42.31/29.58	39.60/31.46	43.23/34.77	41.19/28.73	46.05/33.65	39.43/24.96	47.17/30.77	48.03/32.92	52.21/33.12	39.57/34.21	40.86/32.95
R18-LSTM	42.59/30.92	43.34/32.24	44.12/29.59	42.85/28.78	39.66/30.40	42.36/31.47	41.61/29.11	46.93/31.59	44.57/27.23	45.52/28.01	50.10/33.79	48.67/25.42	35.55/29.96	44.85/33.46
R50-LSTM	40.75/32.12	40.93/32.91	41.74/30.70	42.16/30.39	38.01/31.16	42.79/35.70	41.61/28.00	40.35/30.40	40.00/27.42	43.87/30.02	48.24/34.32	47.79/33.17	36.26/32.46	39.87/31.87
VGG13-LSTM	43.37/32.41	42.29/32.46	44.23/30.81	45.00/31.45	41.20/31.49	43.67/34.64	46.07/31.50	45.61/31.28	44.29/29.17	47.17/30.07	49.90/33.66	57.52/36.17	40.28/33.61	42.86/31.11
VGG16-LSTM	41.70/30.93	42.99/32.32	41.63/28.42	43.83/29.83	37.04/29.39	49.34/36.83	44.37/30.58	36.84/25.76	41.14/26.39	46.23/27.39	48.65/34.15	53.10/30.03	36.26/32.83	41.53/33.59
C3D	31.69/22.68	26.95/21.02	30.15/19.94	42.70/29.22	27.29/19.80	34.50/24.34	26.96/18.35	28.51/22.55	36.57/23.25	43.16/26.35	46.58/32.44	54.87/22.87	22.99/20.31	32.56/20.93
I3D	38.78/30.17	38.56/29.25	38.52/29.11	40.55/31.07	37.44/28.15	37.55/32.05	39.70/26.09	37.72/30.57	26.29/18.27	41.51/27.87	45.55/35.43	53.10/31.56	33.89/29.10	36.54/28.81
3D-R18	37.57/26.67	37.69/27.47	38.40/24.85	40.40/26.08	33.45/25.40	41.48/29.83	35.67/24.95	39.04/25.12	36.29/21.86	42.69/22.70	44.10/28.32	54.87/32.50	31.28/27.83	37.21/27.25
Two C3D	41.77/30.72	41.45/31.37	43.44/29.77	44.71/30.15	37.89/28.09	47.16/32.22	35.46/23.26	41.23/25.74	42.00/27.89	46.23/28.45	48.03/32.31	63.72/37.55	35.78/30.47	40.86/29.60
Two I3D	41.30/31.01	41.02/31.55	42.31/30.14	43.63/31.20	38.75/28.53	44.98/30.94	40.76/28.93	38.16/25.91	39.43/28.37	44.81/29.96	48.03/33.28	54.87/26.96	36.02/29.19	38.87/28.01
Two 3D-R18	42.28/30.55	42.77/32.72	44.12/29.63	42.95/27.83	38.46/28.54	49.34/31.62	39.28/28.41	41.67/28.50	38.57/24.66	45.52/24.71	48.45/33.16	62.83/33.41	35.07/28.73	42.19/29.58
Two R18-LSTM	43.20/31.28	42.20/31.66	44.91/30.37	46.33/31.09	40.40/30.04	47.60/35.60	40.55/27.09	44.74/26.55	43.43/27.52	47.41/28.50	53.00/33.93	57.52/24.56	36.49/29.94	43.85/31.45
Two VGG13-LSTM	44.54/32.79	44.65/32.96	45.25/31.45	46.57/31.88	40.63/30.96	48.03/36.43	46.92/31.55	48.25/33.02	45.14/28.30	46.70/28.35	52.80/35.32	53.98/31.66	37.44/32.49	46.84/35.11
Average	39.58/29.34	39.27/29.80	40.61/28.11	42.04/28.94	36.55/27.61	42.25/31.97	38.98/26.75	39.79/27.65	38.39/24.75	44.19/27.22	47.33/32.39	52.06/28.57	33.75/28.70	39.12/29.12

Table 3. Comparison results of four kinds of baseline architectures trained from scratch on FERV39k (WAR/UAR).

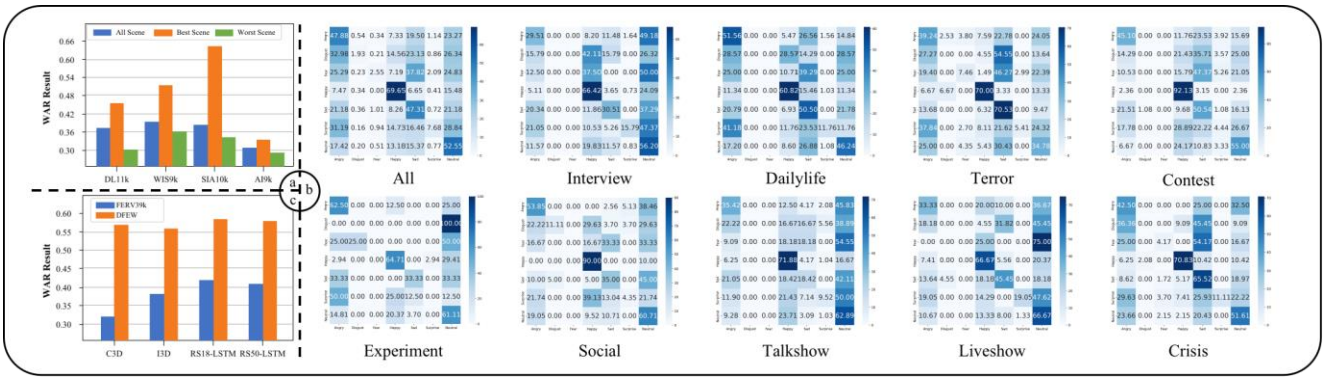

Figure 5. Further experiment analysis in detail. (a) The worst, average and best scene test result for RS50-LSTM trained on 4 scenarios respectively. (b) The confusion matrices of Two stream VGG13-LSTM (best performance) on the FERV39k and 9 representative scenes. (c) The comparison of performance on FERV39k and DFEW of 4 baseline methods.

we note that all training protocols follow the original papers unless stated otherwise. Table 2 shows the results of four kinds of baseline architectures on 9 representative scenes of FERV39k (showing WAR/UAR performance). All models are trained from scratch in experiments.

In summary, performance on more specific scenarios (WIS9k, SIA10k) is better than others, and for 22 fine-grained scenes, most of the methods achieve the highest result on Experiment (SIA10k) and lowest on Terror (AI9k). We attribute the results to the consistency & intensity of expressions and the discriminability of spatial-temporal context features. Two-stream 2D ConvNets-LSTM methods outperform the others where VGG13-LSTM has the best performance of 44.54%. And 2D ConvNet-LSTM methods outperform the 3D ConvNet methods on both one and two-stream structures. We recognize this as the LSTM has a better global-local temporal feature utilization mechanism. In Section 4.4, we further explore the effect of scene & method and challenges on FER39k for DFER.

Cross-scenario challenge. We evaluate the cross-domain difficulty among 4 isolated scenarios via RS50-LSTM. Table 3 shows a nearly 8% average cross-domain decline. And the largest decline of WIS9k experiment shows that it is more challenging to transfer the model from weak interactive scenarios (e.g., WIS9k) to stronger ones than vice versa. To prove it, we also collect statistics of scene performance distribution of models training on the corresponding scenario in Figure 5(a). The result also shows WIS9k has both ideal performance and smaller differences among the 4 scenarios. The result shows it a challenging task to overcome varieties of feature distribution of an expression in different domains of FERV39k.

Source	Target			
	DL11k	WIS9k	SIA10k	AI9k
DL11k	37.69/27.21	29.98/19.93	31.15/21.87	24.27/18.54
WIS9k	27.04/19.95	40.5/26.6	31.78/19.9	24.62/19.24
SIA10k	28.57/21.92	31.39/19.95	39.72/24.9	27.75/20.28
AI9k	26.29/20.21	23.3/18.29	23.85/17.93	31.62/24.16

Table 3. Comparison of cross-scenario results on DL11k, WIS9k, SIA10k, and AI9k of FERV40k on RS50-LSTM.

Scene difficulty and expression confusion. For further analyzing the difficulty in recognizing an expression in different scenes, we also provide the confusion matrices in Figure 5(b) of selected scenes on the best-performed network (VGG13-LSTM). The overall 10 matrices have similar distribution with sight offset among scenes in which method gains better performance on 4 obvious expressions and “Disgust” is the hardest. The result shows an overall statics consistency with previous datasets (e.g.,

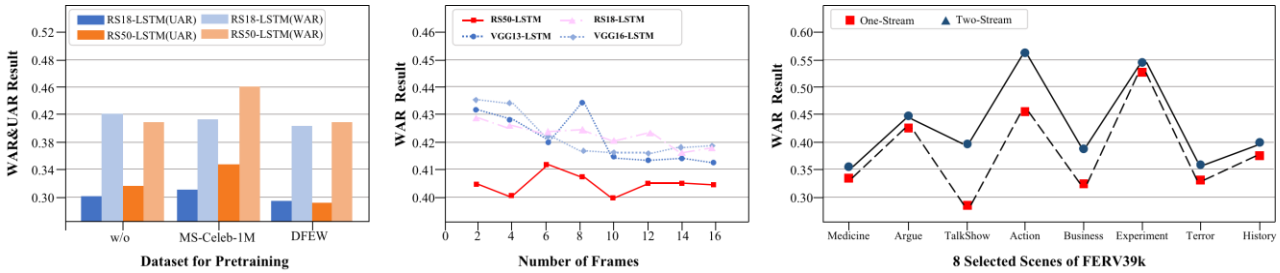

Figure 6. The results of three ablation studies. The charts from left to right are the result of pretraining effectiveness, sparse frame sampling effectiveness and scene information effectiveness, respectively

DFEW). However, some subtle changes are worth to be noticed. For example, the performance for "Sad" declined in Talkshow, Liveshow and Experiment as well as "Angry" in Interview and Contest. This situation may be caused by the changes of expressions in intensity, feature and frequency of occurrence (long-tailed distribution) in specific scenes. There are obvious biases and heterogeneity in our built FERV39k, which make it a challenging dataset. We summarize several directions that might work: (1) Long temporal modeling; (2) Scene reasoning; (3) Global-local fusion in spatial and temporal.

Comparing performance with the existing dataset. To emphasize the difficulty of FERV39k, we compare results with three baseline architectures on DFEW (without a two-stream baseline). The methods on DFEW get higher average results of about 10% (Figure 5(c)), which proves FERV39k is more challenging than state-of-the-art methods. We attribute this to the following reasons: a) FERV39k tripled the number of clips than DFEW, b) Data variety represents a real-world challenge for existing algorithms and c) 22 scenes in this dataset require further application-oriented research.

4.4. Ablation Studies

Does pre-training on large-scale datasets help? We employ RS18-LSTM and RS50-LSTM with and without pretraining using MS-Celeb-1M [45] and DFEW [20] our built FERV39k. The experiment shows that the former ones do not outperform the latter in Figure 6 (left). One potential reason is that the scene and feature distribution of FERV39k is different from other datasets.

Is sparse sampling sufficient for DFER? The sparse sampling schemes [32], [46], [47] often lead to high efficiency and promising accuracy in action recognition. To explore whether sparse sampling is sufficient for DFER, we further investigate the influence of sampling frames on DFER performance. Here we adjust the number of input frames from 2 to 16 in steps of 2 for four 2D ConvNet-LSTM networks on FERV39k. The results in Figure 6 (middle) show that the performance trend varies among different methods but as frames increase over a threshold, the effect tends to be flat or fluctuate decline slightly. These results also show that a more subtle sampling method should be used and the keyframes extraction might be a point [48], [49].

Is scene information auxiliary for DFER? To further understand whether the scene information can boost the performance of DFER methods, we compare Two and single-stream I3D networks on FERV39k benchmark. We select the best and worst result scene of 4 scenarios and provide results in Figure 8 (right), which show that two-stream networks can enhance the face-only model and achieve better results in most scenes due to the fusion of the context information in the scene. For example, we could easily guess the expression as "Sad" with the facial region and scene contexts when someone comes.

Why current methods fail to handle FERV39k? By carefully summarizing all the experiments, we conclude some factors that make FERV39k challenging to four baseline architectures: (1) Limited expression-related frames, especially scenes with frequent emotional changes. (2) Subtle spatial semantics, which involves differences in face and scene-face relationships. (3) Complex temporal dynamics, such as the direction of motion, and the degree of rotation. In addition, the FERV39k dataset poses higher requirements for intermediate representation which is hard to be extracted due to the diversity in one scene.

5. Conclusion

In this paper, we build a large-scale multi-scene dataset (FERV39k) for FER in videos. Compared with existing video-based datasets, our FERV39k has many distinctive characteristics: 1) Automatic generation of large-scale candidate video clips; 2) Well-designed workflow of crowdsourcing and professional annotation for high-quality data labeling; 3) Raising four kinds of challenges and difficulties for FER in videos; 4) Application-oriented multi-scene hierarchy for the robustness of DFER methods. To benchmark the FERV39k, we design four kinds of baseline architectures for video-based FER and give an in-depth evaluation and ablation studies. These results present some important challenges and uncover critical messages for advancing the area of video-based FER in the future.

References

[1] K. Wang, X. Peng, J. Yang, D. Meng, and Y. Qiao, "Region Attention Networks for Pose and Occlusion Robust Facial Expression Recognition," *IEEE Trans. Image Process.*, vol. 29, pp. 4057–4069, 2020, doi: 10.1109/TIP.2019.2956143.

[2] D. Liu, X. Ouyang, S. Xu, P. Zhou, K. He, and S. Wen, "SAANet: Siamese action-units attention network for improving dynamic facial expression recognition," *Neurocomputing*, vol. 413, pp. 145–157, Nov. 2020, doi: 10.1016/j.neucom.2020.06.062.

[3] A. Azazi, S. Lebai Lutfi, I. Venkat, and F. Fernández-Martínez, "Towards a robust affect recognition: Automatic facial expression recognition in 3D faces," *Expert Syst. Appl.*, vol. 42, no. 6, pp. 3056–3066, Apr. 2015, doi: 10.1016/j.eswa.2014.10.042.

[4] A. Barman and P. Dutta, "Facial expression recognition using distance and texture signature relevant features," *Appl. Soft Comput.*, vol. 77, pp. 88–105, Apr. 2019, doi: 10.1016/j.asoc.2019.01.011.

[5] S. Li, W. Deng, and J. Du, "Reliable Crowdsourcing and Deep Locality-Preserving Learning for Expression Recognition in the Wild," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, Jul. 2017, pp. 2584–2593. doi: 10.1109/CVPR.2017.277.

[6] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild," *IEEE Trans. Affect. Comput.*, vol. 10, no. 1, pp. 18–31, Jan. 2019, doi: 10.1109/TAFFC.2017.2740923.

[7] A. H. Farzaneh and X. Qi, "Facial Expression Recognition in the Wild via Deep Attentive Center Loss," 2021, pp. 2402–2411.

[8] Y. Fu, X. Wu, X. Li, Z. Pan, and D. Luo, "Semantic Neighborhood-Aware Deep Facial Expression Recognition," *IEEE Trans. Image Process.*, vol. 29, pp. 6535–6548, 2020, doi: 10.1109/TIP.2020.2991510.

[9] S. Li and W. Deng, "Deep Facial Expression Recognition: A Survey," *IEEE Trans. Affect. Comput.*, pp. 1–1, 2020, doi: 10.1109/TAFFC.2020.2981446.

[10] Y. Wang *et al.*, "A Systematic Review on Affective Computing: Emotion Models, Databases, and Recent Advances," *ArXiv220306935 Cs*, Mar. 2022, Accessed: Mar. 15, 2022. [Online]. Available: http://arxiv.org/abs/2203.06935

[11] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, San Francisco, CA, USA, Jun. 2010, pp. 94–101. doi: 10.1109/CVPRW.2010.5543262.

[12] G. Zhao, X. Huang, M. Taini, S. Z. Li, and M. Pietikäinen, "Facial expression recognition from near-infrared videos," *Image Vis. Comput.*, vol. 29, no. 9, pp. 607–619, Aug. 2011, doi: 10.1016/j.imavis.2011.07.002.

[13] X. Zhao *et al.*, "Peak-Piloted Deep Network for Facial Expression Recognition," presented at the ECCV, Cham, 2016. doi: 10.1007/978-3-319-46475-6_27.

[14] Z. Yu, G. Liu, Q. Liu, and J. Deng, "Spatio-temporal convolutional features with nested LSTM for facial expression recognition," *Neurocomputing*, vol. 317, pp. 50–57, Nov. 2018, doi: 10.1016/j.neucom.2018.07.028.

[15] D. Meng, X. Peng, K. Wang, and Y. Qiao, "Frame attention networks for facial expression recognition in videos," *ArXiv190700193 Cs*, Sep. 2019, Accessed: Jun. 07, 2021. [Online]. Available: http://arxiv.org/abs/1907.00193

[16] J. Kossaifi, G. Tzimiropoulos, S. Todorovic, and M. Pantic, "AFEW-VA database for valence and arousal estimation in-the-wild," *Image Vis. Comput.*, vol. 65, pp. 23–36, Sep. 2017, doi: 10.1016/j.imavis.2017.02.001.

[17] A. Dhall, A. Kaur, R. Goecke, and T. Gedeon, "EmotiW 2018: Audio-Video, Student Engagement and Group-Level Affect Prediction," *ArXiv180807773 Cs*, Aug. 2018, Accessed: Oct. 06, 2021. [Online]. Available: http://arxiv.org/abs/1808.07773

[18] D. Kollias *et al.*, "Deep Affect Prediction in-the-Wild: Aff-Wild Database and Challenge, Deep Architectures, and Beyond," *Int. J. Comput. Vis.*, vol. 127, no. 6–7, pp. 907–929, Jun. 2019, doi: 10.1007/s11263-019-01158-4.

[19] J. Lee, S. Kim, S. Kim, J. Park, and K. Sohn, "Context-Aware Emotion Recognition Networks," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2019, pp. 10142–10151. doi: 10.1109/ICCV.2019.01024.

[20] X. Jiang *et al.*, "DFEW: A Large-Scale Database for Recognizing Dynamic Facial Expressions in the Wild," in *Proceedings of the 28th ACM International Conference on Multimedia*, Seattle WA USA, Oct. 2020, pp. 2881–2889. doi: 10.1145/3394171.3413620.

[21] A. S. Cowen, D. Keltner, F. Schroff, B. Jou, H. Adam, and G. Prasad, "Sixteen facial expressions occur in similar contexts worldwide," *Nature*, vol. 589, no. 7841, pp. 251–257, Jan. 2021, doi: 10.1038/s41586-020-3037-7.

[22] W. Kay *et al.*, "The Kinetics Human Action Video Dataset," *ArXiv170506950 Cs*, May 2017, Accessed: Sep. 12, 2021. [Online]. Available: http://arxiv.org/abs/1705.06950

[23] J. Carreira and A. Zisserman, "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset," *ArXiv170507750 Cs*, Feb. 2018, Accessed: Sep. 13, 2021. [Online]. Available: http://arxiv.org/abs/1705.07750

[24] X. Liu, S. L. Pintea, F. K. Nejadasl, O. Booij, and J. C. van Gemert, "No Frame Left Behind: Full Video Action Recognition," 2021, pp. 14892–14901. Accessed: Sep. 12, 2021. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2021/html/Liu_No_Frame_Left_Behind_Full_Video_Action_Recognition_CVPR_2021_paper.html

[25] N. Otberdout, M. Daoudi, A. Kacem, L. Ballihi, and S. Berretti, "Dynamic Facial Expression Generation on Hilbert Hypersphere with Conditional Wasserstein Generative Adversarial Nets," *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 1–1, 2020, doi: 10.1109/TPAMI.2020.3002500.

[26] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning Spatiotemporal Features with 3D Convolutional Networks," in *2015 IEEE International*

Conference on Computer Vision (ICCV), Dec. 2015, pp. 4489–4497. doi: 10.1109/ICCV.2015.510.

[27] L. Lo, H.-X. Xie, H.-H. Shuai, and W.-H. Cheng, "MER-GCN: Micro-Expression Recognition Based on Relation Modeling with Graph Convolutional Networks," in *2020 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, Aug. 2020, pp. 79–84. doi: 10.1109/MIPR49039.2020.00023.

[28] D. A. A. CHANTI and A. Caplier, "Deep Learning for Spatio-Temporal Modeling of Dynamic Spontaneous Emotions," *IEEE Trans. Affect. Comput.*, pp. 1–1, 2018, doi: 10.1109/TAFFC.2018.2873600.

[29] K. Zhang, Y. Huang, Y. Du, and L. Wang, "Facial Expression Recognition Based on Deep Evolutional Spatial-Temporal Networks," *IEEE Trans. Image Process.*, vol. 26, no. 9, pp. 4193–4203, Sep. 2017, doi: 10.1109/TIP.2017.2689999.

[30] J. Ma, H. Tang, W.-L. Zheng, and B.-L. Lu, "Emotion Recognition using Multimodal Residual LSTM Network," in *Proceedings of the 27th ACM International Conference on Multimedia*, Nice France, Oct. 2019, pp. 176–183. doi: 10.1145/3343031.3350871.

[31] Y. Fan, X. Lu, D. Li, and Y. Liu, "Video-based emotion recognition using CNN-RNN and C3D hybrid networks," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, Tokyo Japan, Oct. 2016, pp. 445–450. doi: 10.1145/2993148.2997632.

[32] D. Shao, Y. Zhao, B. Dai, and D. Lin, "FineGym: A Hierarchical Video Dataset for Fine-Grained Action Understanding," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 2613–2622. doi: 10.1109/CVPR42600.2020.00269.

[33] E. P. Volkova, B. J. Mohler, T. J. Dodds, J. Tesch, and H. H. Bülthoff, "Emotion categorization of body expressions in narrative scenarios," *Front. Psychol.*, vol. 5, 2014, doi: 10.3389/fpsyg.2014.00623.

[34] L. Liang, C. Lang, Y. Li, S. Feng, and J. Zhao, "Fine-Grained Facial Expression Recognition in the Wild," *IEEE Trans. Inf. Forensics Secur.*, vol. 16, pp. 482–494, 2021, doi: 10.1109/TIFS.2020.3007327.

[35] X. Ben *et al.*, "Video-based Facial Micro-Expression Analysis: A Survey of Datasets, Features and Algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 1–1, 2021, doi: 10.1109/TPAMI.2021.3067464.

[36] B. Schuller *et al.*, "Cross-Corpus Acoustic Emotion Recognition: Variances and Strategies," *IEEE Trans. Affect. Comput.*, vol. 1, no. 2, pp. 119–131, Jul. 2010, doi: 10.1109/T-AFFC.2010.8.

[37] C.-F. R. Chen *et al.*, "Deep Analysis of CNN-Based Spatio-Temporal Representations for Action Recognition," 2021, pp. 6165–6175.

[38] J. Y.-H. Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2015, pp. 4694–4702. doi: 10.1109/CVPR.2015.7299101.

[39] S. Ji, W. Xu, M. Yang, and K. Yu, "3D Convolutional Neural Networks for Human Action Recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221–231, Jan. 2013, doi: 10.1109/TPAMI.2012.59.

[40] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," presented at the ICLR, 2015.

[41] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.

[42] W. Rawat and Z. Wang, "Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review," *Neural Comput.*, vol. 29, no. 9, pp. 2352–2449, Sep. 2017, doi: 10.1162/neco_a_00990.

[43] J. Donahue *et al.*, "Long-term recurrent convolutional networks for visual recognition and description," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2015, pp. 2625–2634. doi: 10.1109/CVPR.2015.7298878.

[44] A. Dapogny, K. Bailly, and S. Dubuisson, "Dynamic facial expression recognition by joint static and multi-time gap transition classification," in *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, May 2015, vol. 1, pp. 1–6. doi: 10.1109/FG.2015.7163111.

[45] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "MS-Celeb-1M: A Dataset and Benchmark for Large-Scale Face Recognition," in *Computer Vision – ECCV 2016*, Cham, 2016, pp. 87–102.

[46] F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles, "ActivityNet: A large-scale video benchmark for human activity understanding," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 961–970. doi: 10.1109/CVPR.2015.7298698.

[47] S. N. Gowda, M. Rohrbach, and L. Sevilla-Lara, "SMART Frame Selection for Action Recognition," *ArXiv201210671 Cs*, Dec. 2020, Accessed: Sep. 12, 2021. [Online]. Available: http://arxiv.org/abs/2012.10671

[48] A. Ghodrati, B. E. Bejnordi, and A. Habibian, "FrameExit: Conditional Early Exiting for Efficient Video Recognition," 2021, pp. 15608–15618. Accessed: Sep. 12, 2021. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2021/html/Ghodrati_FrameExit_Conditional_Early_Exiting_for_Efficient_Video_Recognition_CVPR_2021_paper.html

[49] Y.-D. Zheng, Z. Liu, T. Lu, and L. Wang, "Dynamic Sampling Networks for Efficient Action Recognition in Videos," *IEEE Trans. Image Process.*, vol. 29, pp. 7970–7983, 2020, doi: 10.1109/TIP.2020.3007826.