Deep Facial Expression Recognition: A Survey

Shan Li and Weihong Deng*, *Member, IEEE*

Abstract—With the transition of facial expression recognition (FER) from laboratory-controlled to challenging in-the-wild conditions and the recent success of deep learning techniques in various fields, deep neural networks have increasingly been leveraged to learn discriminative representations for automatic FER. Recent deep FER systems generally focus on two important issues: overfitting caused by a lack of sufficient training data and expression-unrelated variations, such as illumination, head pose and identity bias. In this paper, we provide a comprehensive survey on deep FER, including datasets and algorithms that provide insights into these intrinsic problems. First, we introduce the available datasets that are widely used in the literature and provide accepted data selection and evaluation principles for these datasets. We then describe the standard pipeline of a deep FER system with the related background knowledge and suggestions of applicable implementations for each stage. For the state of the art in deep FER, we review existing novel deep neural networks and related training strategies that are designed for FER based on both static images and dynamic image sequences, and discuss their advantages and limitations. Competitive performances on widely used benchmarks are also summarized in this section. We then extend our survey to additional related issues and application scenarios. Finally, we review the remaining challenges and corresponding opportunities in this field as well as future directions for the design of robust deep FER systems.

Index Terms—Facial Expressions Recognition, Facial expression datasets, Affect, Deep Learning, Survey.

✦

1 INTRODUCTION

FACIAL expression is one of the most powerful, natural and universal signals for human beings to convey their emotional states and intentions [1], [2]. Numerous studies have been conducted on automatic facial expression analysis because of its practical importance in sociable robotics, medical treatment, driver fatigue surveillance, and many other human-computer interaction systems. In the field of computer vision and machine learning, various facial expression recognition (FER) systems have been explored to encode expression information from facial representations. As early as the twentieth century, Ekman and Friesen [3] defined six basic emotions based on cross-culture study [4], which indicated that humans perceive certain basic emotions in the same way regardless of culture. These prototypical facial expressions are anger, disgust, fear, happiness, sadness, and surprise. Contempt was subsequently added as one of the basic emotions [5]. Recently, advanced research on neuroscience and psychology argued that the model of six basic emotions are culture-specific and not universal [6].

Although the affect model based on basic emotions is limited in the ability to represent the complexity and subtlety of our daily affective displays [7], [8], [9], and other emotion description models, such as the Facial Action Coding System (FACS) [10] and the continuous model using affect dimensions [11], are considered to represent a wider range of emotions, the categorical model that describes emotions in terms of discrete basic emotions is still the most popular perspective for FER, due to its pioneering investigations along with the direct and intuitive definition of facial expressions. And in this survey, we will limit our discussion on FER based on the categorical model.

FER systems can be divided into two main categories according to the feature representations: static image FER and dynamic sequence FER. In *static-based methods* [12], [13], [14], the feature

representation is encoded with only spatial information from the current single image, whereas *dynamic-based methods* [15], [16], [17] consider the temporal relation among contiguous frames in the input facial expression sequence. Based on these two vision-based methods, other modalities, such as audio and physiological channels, have also been used in *multimodal systems* [18] to assist the recognition of expression.

The majority of the traditional methods have used handcrafted features or shallow learning (e.g., local binary patterns (LBP) [12], LBP on three orthogonal planes (LBP-TOP) [15], non-negative matrix factorization (NMF) [19] and sparse learning [20]) for FER. However, since 2013, emotion recognition competitions such as FER2013 [21] and Emotion Recognition in the Wild (EmotiW) [22], [23], [24] have collected relatively sufficient training data from challenging real-world scenarios, which implicitly promote the transition of FER from lab-controlled to in-the-wild settings. In the meanwhile, due to the dramatically increased chip processing abilities (e.g., GPU units) and well-designed network architecture, studies in various fields have begun to transfer to deep learning methods, which have achieved the state-of-the-art recognition accuracy and exceeded previous results by a large margin (e.g., [25], [26], [27], [28]). Likewise, given with more effective training data of facial expression, deep learning techniques have increasingly been implemented to handle the challenging factors for emotion recognition in the wild. Figure 1 illustrates this evolution on FER in the aspect of algorithms and datasets.

Exhaustive surveys on automatic expression analysis have been published in recent years [7], [8], [29], [30]. These surveys have established a set of standard algorithmic pipelines for FER. However, they focus on traditional methods, and deep learning has rarely been reviewed. Very recently, FER based on deep learning has been surveyed in [31], which is a brief review without introductions on FER datasets and technical details on deep FER. Therefore, in this paper, we make a systematic research on deep learning for FER tasks based on both static images and videos (image sequences). We aim to give a newcomer to this filed an overview of the systematic framework and prime skills for deep

---

- *The authors are with the Pattern Recognition and Intelligent System Laboratory, School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing, 100876, China. E-mail:{ls1995, whdeng}@bupt.edu.cn.*
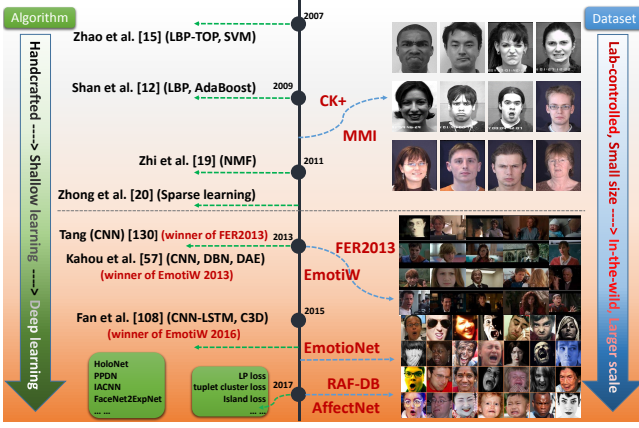
Fig. 1. The evolution of facial expression recognition in terms of datasets and methods.

FER.

Despite the powerful feature learning ability of deep learning, problems remain when applied to FER. First, deep neural networks require a large amount of training data to avoid overfitting. However, the existing facial expression databases are not sufficient to train the well-known neural network with deep architecture that achieved the most promising results in object recognition tasks. Additionally, high inter-subject variations exist due to different personal attributes, such as age, gender, ethnic backgrounds and level of expressiveness [32]. In addition to subject identity bias, variations in pose, illumination and occlusions are common in unconstrained facial expression scenarios. These factors are non-linearly coupled with facial expressions and therefore strengthen the requirement of deep networks to address the large intra-class variability and to learn effective expression-specific representations.

In this paper, we introduce recent advances in research on solving the above problems for deep FER. We examine the state-of-the-art results that have not been reviewed in previous survey papers. The rest of this paper is organized as follows. Frequently used expression databases are introduced in Section 2. Section 3 identifies three main steps required in a deep FER system and describes the related background. Section 4 provides a detailed review of novel neural network architectures and special network training tricks designed for FER based on static images and dynamic image sequences. We then cover additional related issues and other practical scenarios in Section 5. Section 6 discusses some of the challenges and opportunities in this field and identifies potential future directions.

2 FACIAL EXPRESSION DATABASES

Having sufficient labeled training data that include as many variations of the populations and environments as possible is important for the design of a deep expression recognition system. In this section, we discuss the publicly available databases that contain basic expressions and that are widely used in our reviewed papers for deep learning algorithms evaluation. We also introduce newly released databases that contain a large amount of affective images collected from the real world to benefit the training of deep neural networks. Table 1 provides an overview of these datasets, including the main reference, number of subjects, number

of image or video samples, collection environment, expression distribution and additional information.

CK+ [33]: The Extended CohnKanade (CK+) database is the most extensively used laboratory-controlled database for evaluating FER systems. CK+ contains 593 video sequences from 123 subjects. The sequences vary in duration from 10 to 60 frames and show a shift from a neutral facial expression to the peak expression. Among these videos, 327 sequences from 118 subjects are labeled with seven basic expression labels (anger, contempt, disgust, fear, happiness, sadness, and surprise) based on the Facial Action Coding System (FACS). Because CK+ does not provide specified training, validation and test sets, the algorithms evaluated on this database are not uniform. For static-based methods, the most common data selection method is to extract the last one to three frames with peak formation and the first frame (neutral face) of each sequence. Then, the subjects are divided into n groups for person-independent n-fold cross-validation experiments, where commonly selected values of n are 5, 8 and 10.

MMI [34], [35]: The MMI database is laboratory-controlled and includes 326 sequences from 32 subjects. A total of 213 sequences are labeled with six basic expressions (without "contempt"), and 205 sequences are captured in frontal view. In contrast to CK+, sequences in MMI are onset-apex-offset labeled, i.e., the sequence begins with a neutral expression and reaches peak near the middle before returning to the neutral expression. Furthermore, MMI has more challenging conditions, i.e., there are large inter-personal variations because subjects perform the same expression non-uniformly and many of them wear accessories (e.g., glasses, mustache). For experiments, the most common method is to choose the first frame (neutral face) and the three peak frames in each frontal sequence to conduct person-independent 10-fold cross-validation.

JAFFE [36]: The Japanese Female Facial Expression (JAFFE) database is a laboratory-controlled image database that contains 213 samples of posed expressions from 10 Japanese females. Each person has 3~4 images with each of six basic facial expressions (anger, disgust, fear, happiness, sadness, and surprise) and one image with a neutral expression. The database is challenging because it contains few examples per subject/expression. Typically, all the images are used for the leave-one-subject-out experiment.

TFD [37]: The Toronto Face Database (TFD) is an amalgamation of several facial expression datasets. TFD contains 112,234 images, 4,178 of which are annotated with one of seven expression labels: anger, disgust, fear, happiness, sadness, surprise and neutral. The faces have already been detected and normalized to a size of 48*48 such that all the subjects eyes are the same distance apart and have the same vertical coordinates. Five official folds are provided in TFD; each fold contains a training, validation, and test set consisting of 70%, 10%, and 20% of the images, respectively.

FER2013 [21]: The FER2013 database was introduced during the ICML 2013 Challenges in Representation Learning. FER2013 is a large-scale and unconstrained database collected automatically by the Google image search API. All images have been registered and resized to 48*48 pixels after rejecting wrongfully labeled frames and adjusting the cropped region. FER2013 contains 28,709 training images, 3,589 validation images and 3,589 test images with seven expression labels (anger, disgust, fear, happiness, sadness, surprise and neutral).

AFEW [48]: The Acted Facial Expressions in the Wild (AFEW) database was first established and introduced in [49] and has served as an evaluation platform for the annual Emo-

TABLE 1
An overview of the facial expression datasets. P = posed; S = spontaneous; Condit. = Collection condition; Elicit. = Elicitation method.

Database	Samples	Subject	Condit.	Elicit.	Expression distribution	Access
CK+ [33]	593 image sequences	123	Lab	P & S	6 basic expressions plus contempt and neutral	http://www.consortium.ri.cmu.edu/ckagree/
MMI [34], [35]	740 images and 2,900 videos	25	Lab	P	6 basic expressions plus neutral	https://mmifacedb.eu/
JAFFE [36]	213 images	10	Lab	P	6 basic expressions plus neutral	http://www.kasrl.org/jaffe.html
TFD [37]	112,234 images	N/A	Lab	P	6 basic expressions plus neutral	josh@mplab.ucsd.edu
FER-2013 [21]	35,887 images	N/A	Web	P & S	6 basic expressions plus neutral	https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge
AFEW 7.0 [24]	1,809 videos	N/A	Movie	P & S	6 basic expressions plus neutral	https://sites.google.com/site/emotiwchallenge/
SFEW 2.0 [22]	1,766 images	N/A	Movie	P & S	6 basic expressions plus neutral	https://cs.anu.edu.au/few/emotiw2015.html
Multi-PIE [38]	755,370 images	337	Lab	P	Smile, surprised, squint, disgust, scream and neutral	http://www.flintbox.com/public/project/4742/
BU-3DFE [39]	2,500 images	100	Lab	P	6 basic expressions plus neutral	http://www.cs.binghamton.edu/~lijun/Research/3DFE/3DFE_Analysis.html
Oulu-CASIA [40]	2,880 image sequences	80	Lab	P	6 basic expressions	http://www.cse.oulu.fi/CMV/Downloads/Oulu-CASIA
RaFD [41]	1,608 images	67	Lab	P	6 basic expressions plus contempt and neutral	http://www.socsci.ru.nl:8180/RaFD2/RaFD
KDEF [42]	4,900 images	70	Lab	P	6 basic expressions plus neutral	http://www.emotionlab.se/kdef/
EmotioNet [43]	1,000,000 images	N/A	Web	P & S	23 basic expressions or compound expressions	http://cbcsl.ece.ohio-state.edu/dbform_emotionet.html
RAF-DB [44], [45]	29672 images	N/A	Web	P & S	6 basic expressions plus neutral and 12 compound expressions	http://www.whdeng.cn/RAF/model1.html
AffectNet [46]	450,000 images (labeled)	N/A	Web	P & S	6 basic expressions plus neutral	http://mohammadmahoor.com/databases-codes/
ExpW [47]	91,793 images	N/A	Web	P & S	6 basic expressions plus neutral	http://mmlab.ie.cuhk.edu.hk/projects/socialrelation/index.html

tion Recognition In The Wild Challenge (EmotiW) since 2013. AFEW contains video clips collected from different movies with spontaneous expressions, various head poses, occlusions and illuminations. AFEW is a temporal and multimodal database that provides with vastly different environmental conditions in both audio and video. Samples are labeled with seven expressions: anger, disgust, fear, happiness, sadness, surprise and neutral. The annotation of expressions have been continuously updated, and reality TV show data have been continuously added. The **AFEW 7.0** in EmotiW 2017 [24] is divided into three data partitions in an independent manner in terms of subject and movie/TV source: Train (773 samples), Val (383 samples) and Test (653 samples), which ensures data in the three sets belong to mutually exclusive movies and actors.

SFEW [50]: The Static Facial Expressions in the Wild (SFEW) was created by selecting static frames from the AFEW database by computing key frames based on facial point clustering. The most commonly used version, **SFEW 2.0**, was the benchmarking data for the SReco sub-challenge in EmotiW 2015 [22]. SFEW 2.0 has been divided into three sets: Train (958 samples), Val (436 samples) and Test (372 samples). Each of the images is assigned to one of seven expression categories, i.e., anger, disgust, fear, neutral, happiness, sadness, and surprise. The expression labels of the training and validation sets are publicly available, whereas those of the testing set are held back by the challenge organizer.

Multi-PIE [38]: The CMU Multi-PIE database contains

755,370 images from 337 subjects under 15 viewpoints and 19 illumination conditions in up to four recording session. Each facial image is labeled with one of six expressions: disgust, neutral, scream, smile, squint and surprise. This dataset is typically used for multiview facial expression analysis.

BU-3DFE [39]: The Binghamton University 3D Facial Expression (BU-3DFE) database contains 606 facial expression sequences captured from 100 people. For each subject, six universal facial expressions (anger, disgust, fear, happiness, sadness and surprise) are elicited by various manners with multiple intensities. Similar to Multi-PIE, this dataset is typically used for multiview 3D facial expression analysis.

Oulu-CASIA [40]: The Oulu-CASIA database includes 2,880 image sequences collected from 80 subjects labeled with six basic emotion labels: anger, disgust, fear, happiness, sadness, and surprise. Each of the videos is captured with one of two imaging systems, i.e., near-infrared (NIR) or visible light (VIS), under three different illumination conditions. Similar to CK+, the first frame is neutral and the last frame has the peak expression. Typically, only the last three peak frames and the first frame (neutral face) from the 480 videos collected by the VIS System under normal indoor illumination are employed for 10-fold cross-validation experiments.

RaFD [41]: The Radboud Faces Database (RaFD) is laboratory-controlled and has a total of 1,608 images from 67 subjects with three different gaze directions, i.e., front, left and right. Each sample is labeled with one of eight expressions: anger,

contempt, disgust, fear, happiness, sadness, surprise and neutral.

KDEF [42]: The laboratory-controlled Karolinska Directed Emotional Faces (KDEF) database was originally developed for use in psychological and medical research. KDEF consists of images from 70 actors with five different angles labeled with six basic facial expressions plus neutral.

In addition to these commonly used datasets for basic emotion recognition, several well-established and large-scale publicly available facial expression databases collected from the Internet that are suitable for training deep neural networks have emerged in the last two years.

EmotioNet [43]: EmotioNet is a large-scale database with one million facial expression images collected from the Internet. A total of 950,000 images were annotated by the automatic action unit (AU) detection model in [43], and the remaining 25,000 images were manually annotated with 11 AUs. The second track of the EmotioNet Challenge [51] provides six basic expressions and ten compound expressions [52], and 2,478 images with expression labels are available.

RAF-DB [44], [45]: The Real-world Affective Face Database (RAF-DB) is a real-world database that contains 29,672 highly diverse facial images downloaded from the Internet. With manually crowd-sourced annotation and reliable estimation, seven basic and eleven compound emotion labels are provided for the samples. Specifically, 15,339 images from the basic emotion set are divided into two groups (12,271 training samples and 3,068 testing samples) for evaluation.

AffectNet [46]: AffectNet contains more than one million images from the Internet that were obtained by querying different search engines using emotion-related tags. It is by far the largest database that provides facial expressions in two different emotion models (categorical model and dimensional model), of which 450,000 images have manually annotated labels for eight basic expressions.

ExpW [47]: The Expression in-the-Wild Database (ExpW) contains 91,793 faces downloaded using Google image search. Each of the face images was manually annotated as one of the seven basic expression categories. Non-face images were removed in the annotation process.

3 DEEP FACIAL EXPRESSION RECOGNITION

In this section, we describe the three main steps that are common in automatic deep FER, i.e., pre-processing, deep feature learning and deep feature classification. We briefly summarize the widely used algorithms for each step and recommend the existing state-of-the-art best practice implementations according to the referenced papers.

3.1 Pre-processing

Variations that are irrelevant to facial expressions, such as different backgrounds, illuminations and head poses, are fairly common in unconstrained scenarios. Therefore, before training the deep neural network to learn meaningful features, pre-processing is required to align and normalize the visual semantic information conveyed by the face.

3.1.1 Face alignment

Face alignment is a traditional pre-processing step in many face-related recognition tasks. We list some well-known approaches

TABLE 2
Summary of different types of face alignment detectors that are widely used in deep FER models.

type		# points	real-time	speed	performance	used in
Holistic	AAM [53]	68	✗	fair	poor generalization	[54], [55]
Part-based	MoT [56]	39/68	✗	slow/	good	[57], [58]
	DRMF [59]	66	✗	fast		[60], [61]
Cascaded regression	SDM [62]	49	✓	fast/ very fast	good/ very good	[16], [63]
	3000 fps [64]	68	✓			[55]
	Incremental [65]	49	✓			[66]
Deep learning	cascaded CNN [67]	5	✓	fast	good/ very good	[68]
	MTCNN [69]	5	✓			[70], [71]

and publicly available implementations that are widely used in deep FER.

Given a series of training data, the first step is to detect the face and then to remove background and non-face areas. The Viola-Jones (V&J) face detector [72] is a classic and widely employed implementation for face detection, which is robust and computationally simple for detecting near-frontal faces.

Although face detection is the only indispensable procedure to enable feature learning, further face alignment using the coordinates of localized landmarks can substantially enhance the FER performance [14]. This step is crucial because it can reduce the variation in face scale and in-plane rotation. Table 2 investigates facial landmark detection algorithms widely-used in deep FER and compares them in terms of efficiency and performance. The Active Appearance Model (AAM) [53] is a classic generative model that optimizes the required parameters from holistic facial appearance and global shape patterns. In discriminative models, the mixtures of trees (MoT) structured models [56] and the discriminative response map fitting (DRMF) [59] use part-based approaches that represent the face via the local appearance information around each landmark. Furthermore, a number of discriminative models directly use a cascade of regression functions to map the image appearance to landmark locations and have shown better results, e.g., the supervised descent method (SDM) [62] implemented in IntraFace [73], the face alignment 3000 fps [64], and the incremental face alignment [65]. Recently, deep networks have been widely exploited for face alignment. Cascaded CNN [67] is the early work which predicts landmarks in a cascaded way. Based on this, Tasks-Constrained Deep Convolutional Network (TCDCN) [74] and Multi-task CNN (MTCNN) [69] further leverage multi-task learning to improve the performance. In general, cascaded regression has become the most popular and state-of-the-art methods for face alignment as its high speed and accuracy.

In contrast to using only one detector for face alignment, some methods proposed to combine multiple detectors for better landmark estimation when processing faces in challenging unconstrained environments. Yu et al. [75] concatenated three different facial landmark detectors to complement each other. Kim et al. [76] considered different inputs (original image and histogram equalized image) and different face detection models (V&J [72] and MoT [56]), and the landmark set with the highest confidence provided by the Intraface [73] was selected.

3.1.2 Data augmentation

Deep neural networks require sufficient training data to ensure generalizability to a given recognition task. However, most publicly available databases for FER do not have a sufficient quantity
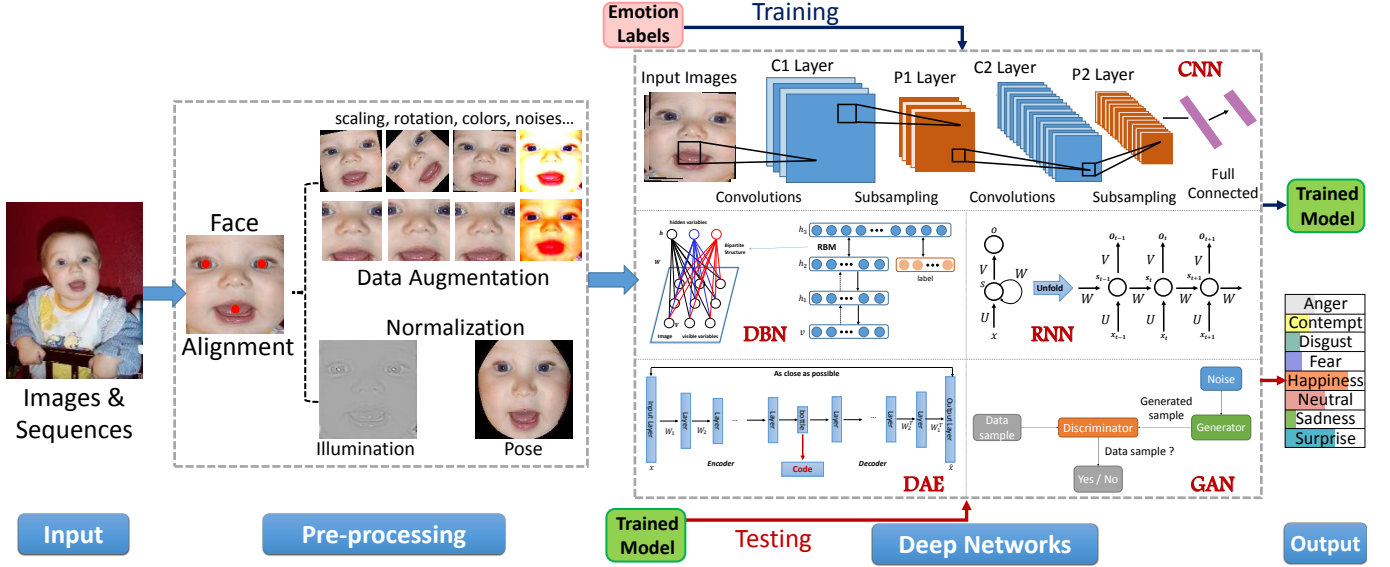
Fig. 2. The general pipeline of deep facial expression recognition systems.

of images for training. Therefore, data augmentation is a vital step for deep FER. Data augmentation techniques can be divided into two groups: on-the-fly data augmentation and offline data augmentation.

Usually, the on-the-fly data augmentation is embedded in deep learning toolkits to alleviate overfitting. During the training step, the input samples are randomly cropped from the four corners and center of the image and then flipped horizontally, which can result in a dataset that is ten times larger than the original training data. Two common prediction modes are adopted during testing: only the center patch of the face is used for prediction (e.g., [61], [77]) or the prediction value is averaged over all ten crops (e.g., [76], [78]).

Besides the elementary on-the-fly data augmentation, various offline data augmentation operations have been designed to further expand data on both size and diversity. The most frequently used operations include random perturbations and transforms, e.g., rotation, shifting, skew, scaling, noise, contrast and color jittering. For example, common noise models, salt & pepper and speckle noise [79] and Gaussian noise [80], [81] are employed to enlarge the data size. And for contrast transformation, saturation and value (S and V components of the HSV color space) of each pixel are changed [70] for data augmentation. Combinations of multiple operations can generate more unseen training samples and make the network more robust to deviated and rotated faces. In [82], the authors applied five image appearance filters (disk, average, Gaussian, unsharp and motion filters) and six affine transform matrices that were formalized by adding slight geometric transformations to the identity matrix. In [75], a more comprehensive affine transform matrix was proposed to randomly generate images that varied in terms of rotation, skew and scale. Furthermore, deep learning based technology can be applied for data augmentation. For example, a synthetic data generation system with 3D convolutional neural network (CNN) was created in [83] to confidentially create faces with different levels of saturation in expression. And the generative adversarial network (GAN) [84] can also be applied to augment data by generating diverse appearances varying in poses and expressions. (see Section 4.1.7).

3.1.3 Face normalization

Variations in illumination and head poses can introduce large changes in images and hence impair the FER performance. Therefore, we introduce two typical face normalization methods to ameliorate these variations: illumination normalization and pose normalization (frontalization).

Illumination normalization: Illumination and contrast can vary in different images even from the same person with the same expression, especially in unconstrained environments, which can result in large intra-class variances. In [60], several frequently used illumination normalization algorithms, namely, isotropic diffusion (IS)-based normalization, discrete cosine transform (DCT)-based normalization [85] and difference of Gaussian (DoG), were evaluated for illumination normalization. And [86] employed homomorphic filtering based normalization, which has been reported to yield the most consistent results among all other techniques, to remove illumination normalization. Furthermore, related studies have shown that histogram equalization combined with illumination normalization results in better face recognition performance than that achieved using illumination normalization on it own. And many studies in the literature of deep FER (e.g., [75], [79], [87], [88]) have employed histogram equalization to increase the global contrast of images for pre-processing. This method is effective when the brightness of the background and foreground are similar. However, directly applying histogram equalization may overemphasize local contrast. To solve this problem, [89] proposed a weighted summation approach to combine histogram equalization and linear mapping. And in [79], the authors compared three different methods: global contrast normalization (GCN), local normalization, and histogram equalization. GCN and histogram equalization were reported to achieve the best accuracy for the training and testing steps, respectively.

Pose normalization: Considerable pose variation is another common and intractable problem in unconstrained settings. Some studies have employed pose normalization techniques to yield

frontal facial views for FER (e.g., [90], [91]), among which the most popular was proposed by Hassner et al. [92]. Specifically, after localizing facial landmarks, a 3D texture reference model generic to all faces is generated to efficiently estimate visible facial components. Then, the initial frontalized face is synthesized by back-projecting each input face image to the reference coordinate system. Alternatively, Sagonas et al. [93] proposed an effective statistical model to simultaneously localize landmarks and convert facial poses using only frontal faces. Very recently, a series of GAN-based deep models were proposed for frontal view synthesis (e.g., FF-GAN [94], TP-GAN [95]) and DR-GAN [96]) and report promising performances.

3.2 Deep networks for feature learning

Deep learning has recently become a hot research topic and has achieved state-of-the-art performance for a variety of applications [97]. Deep learning attempts to capture high-level abstractions through hierarchical architectures of multiple nonlinear transformations and representations. In this section, we briefly introduce some deep learning techniques that have been applied for FER. The traditional architectures of these deep neural networks are shown in Fig. 2.

3.2.1 Convolutional neural network (CNN)

CNN has been extensively used in diverse computer vision applications, including FER. At the beginning of the 21st century, several studies in the FER literature [98], [99] found that the CNN is robust to face location changes and scale variations and behaves better than the multilayer perceptron (MLP) in the case of previously unseen face pose variations. [100] employed the CNN to address the problems of subject independence as well as translation, rotation, and scale invariance in the recognition of facial expressions.

A CNN consists of three types of heterogeneous layers: convolutional layers, pooling layers, and fully connected layers. *The convolutional layer* has a set of learnable filters to convolve through the whole input image and produce various specific types of activation feature maps. The convolution operation is associated with three main benefits: local connectivity, which learns correlations among neighboring pixels; weight sharing in the same feature map, which greatly reduces the number of the parameters to be learned; and shift-invariance to the location of the object. *The pooling layer* follows the convolutional layer and is used to reduce the spatial size of the feature maps and the computational cost of the network. Average pooling and max pooling are the two most commonly used nonlinear down-sampling strategies for translation invariance. *The fully connected layer* is usually included at the end of the network to ensure that all neurons in the layer are fully connected to activations in the previous layer and to enable the 2D feature maps to be converted into 1D feature maps for further feature representation and classification.

We list the configurations and characteristics of some well-known CNN models that have been applied for FER in Table 3. Besides these networks, several well-known derived frameworks also exist. In [101], [102], region-based CNN (R-CNN) [103] was used to learn features for FER. In [104], Faster R-CNN [105] was used to identify facial expressions by generating high-quality region proposals. Moreover, Ji et al. proposed 3D CNN [106] to capture motion information encoded in multiple adjacent frames for action recognition via 3D convolutions. Tran et al. [107]

TABLE 3
Comparison of CNN models and their achievements. DA = Data augmentation; BN = Batch normalization.

	AlexNet [25]	VGGNet [26]	GoogleNet [27]	ResNet [28]
Year	2012	2014	2014	2015
# of layers[†]	5+3	13/16 + 3	21+1	151+1
Kernel size[⋆]	11, 5, 3	3	7, 1, 3, 5	7, 1, 3, 5
DA	✓	✓	✓	✓
Dropout	✓	✓	✓	✓
Inception	✗	✗	✓	✗
BN	✗	✗	✗	✓
Used in	[110]	[78], [111]	[17], [78]	[91], [112]

[†] number of convolutional layers + fully connected layers
[⋆] size of the convolution kernel

proposed the well-designed C3D, which exploits 3D convolutions on large-scale supervised training datasets to learn spatio-temporal features. Many related studies (e.g., [108], [109]) have employed this network for FER involving image sequences.

3.2.2 Deep belief network (DBN)

DBN proposed by Hinton et al. [113] is a graphical model that learns to extract a deep hierarchical representation of the training data. The traditional DBN is built with a stack of restricted Boltzmann machines (RBMs) [114], which are two-layer generative stochastic models composed of a visible-unit layer and a hidden-unit layer. These two layers in an RBM must form a bipartite graph without lateral connections. In a DBN, the units in higher layers are trained to learn the conditional dependencies among the units in the adjacent lower layers, except the top two layers, which have undirected connections. The training of a DBN contains two phases: pre-training and fine-tuning [115]. First, an efficient layer-by-layer greedy learning strategy [116] is used to initialize the deep network in an unsupervised manner, which can prevent poor local optimal results to some extent without the requirement of a large amount of labeled data. During this procedure, contrastive divergence [117] is used to train RBMs in the DBN to estimate the approximation gradient of the log-likelihood. Then, the parameters of the network and the desired output are fine-tuned with a simple gradient descent under supervision.

3.2.3 Deep autoencoder (DAE)

DAE was first introduced in [118] to learn efficient codings for dimensionality reduction. In contrast to the previously mentioned networks, which are trained to predict target values, the DAE is optimized to reconstruct its inputs by minimizing the reconstruction error. Variations of the DAE exist, such as the denoising autoencoder [119], which recovers the original undistorted input from partially corrupted data; the sparse autoencoder network (DSAE) [120], which enforces sparsity on the learned feature representation; the contractive autoencoder (CAE[1]) [121], which adds an activity dependent regularization to induce locally invariant features; the convolutional autoencoder (CAE[2]) [122], which uses convolutional (and optionally pooling) layers for the hidden layers in the network; and the variational auto-encoder (VAE) [123], which is a directed graphical model with certain types of latent variables to design complex generative models of data.

3.2.4 Recurrent neural network (RNN)

RNN is a connectionist model that captures temporal information and is more suitable for sequential data prediction with arbitrary lengths. In addition to training the deep neural network in a single feed-forward manner, RNNs include recurrent edges that span adjacent time steps and share the same parameters across all steps. The classic back propagation through time (BPTT) [124] is used to train the RNN. Long-short term memory (LSTM), introduced by Hochreiter & Schmidhuber [125], is a special form of the traditional RNN that is used to address the gradient vanishing and exploding problems that are common in training RNNs. The cell state in LSTM is regulated and controlled by three gates: an input gate that allows or blocks alteration of the cell state by the input signal, an output gate that enables or prevents the cell state to affect other neurons, and a forget gate that modulates the cell's self-recurrent connection to accumulate or forget its previous state. By combining these three gates, LSTM can model long-term dependencies in a sequence and has been widely employed for video-based expression recognition tasks.

3.2.5 Generative Adversarial Network (GAN)

GAN was first introduced by Goodfellow et al [84] in 2014, which trains models through a minimax two-player game between a generator $G(\mathbf{z})$ that generates synthesized input data by mapping latents $\mathbf{z}$ to data space with $\mathbf{z} \sim \mathbf{p(z)}$ and a discriminator $D(\mathbf{x})$ that assigns probability $y = \mathrm{Dis}(\mathbf{x}) \in [\mathbf{0}, \mathbf{1}]$ that $\mathbf{x}$ is an actual training sample to tell apart real from fake input data. The generator and the discriminator are trained alternatively and can both improve themselves by minimizing/maximizing the binary cross entropy $L_{GAN} = \log(\mathrm{D}(\mathbf{x})) + \log(\mathbf{1} - \mathrm{D}(\mathrm{G}(\mathbf{z})))$ with respect to D / G with $\mathbf{x}$ being a training sample and $\mathbf{z} \sim \mathbf{p(z)}$. Extensions of GAN exist, such as the cGAN [126] that adds a conditional information to control the output of the generator, the DCGAN [127] that adopts deconvolutional and convolutional neural networks to implement G and D respectively, the VAE/GAN [128] that uses learned feature representations in the GAN discriminator as basis for the VAE reconstruction objective, and the InfoGAN [129] that can learn disentangled representations in a completely unsupervised manner.

3.3 Facial expression classification

After learning the deep features, the final step of FER is to classify the given face into one of the basic emotion categories.

Unlike the traditional methods, where the feature extraction step and the feature classification step are independent, deep networks can perform FER in an end-to-end way. Specifically, a loss layer is added to the end of the network to regulate the back-propagation error; then, the prediction probability of each sample can be directly output by the network. In CNN, softmax loss is the most common used function that minimizes the cross-entropy between the estimated class probabilities and the ground-truth distribution. Alternatively, [130] demonstrated the benefit of using a linear support vector machine (SVM) for the end-to-end training which minimizes a margin-based loss instead of the cross-entropy. Likewise, [131] investigated the adaptation of deep neural forests (NFs) [132] which replaced the softmax loss layer with NFs and achieved competitive results for FER.

Besides the end-to-end learning way, another alternative is to employ the deep neural network (particularly a CNN) as a feature extraction tool and then apply additional independent classifiers, such as support vector machine or random forest, to the extracted representations [133], [134]. Furthermore, [135], [136] showed that the covariance descriptors computed on DCNN features and classification with Gaussian kernels on Symmetric Positive Definite (SPD) manifold are more efficient than the standard classification with the softmax layer.

4 THE STATE OF THE ART

In this section, we review the existing novel deep neural networks designed for FER and the related training strategies proposed to address expression-specific problems. We divide the works presented in the literature into two main groups depending on the type of data: deep FER networks for static images and deep FER networks for dynamic image sequences. We then provide an overview of the current deep FER systems with respect to the network architecture and performance. Because some of the evaluated datasets do not provide explicit data groups for training, validation and testing, and the relevant studies may conduct experiments under different experimental conditions with different data, we summarize the expression recognition performance along with information about the data selection and grouping methods.

4.1 Deep FER networks for static images

A large volume of the existing studies conducted expression recognition tasks based on static images without considering temporal information due to the convenience of data processing and the availability of the relevant training and test material. We first introduce specific pre-training and fine-tuning skills for FER, then review the novel deep neural networks in this field. For each of the most frequently evaluated datasets, Table 4 shows the current state-of-the-art methods in the field that are explicitly conducted in a person-independent protocol (subjects in the training and testing sets are separated).

4.1.1 Pre-training and fine-tuning

As mentioned before, direct training of deep networks on relatively small facial expression datasets is prone to overfitting. To mitigate this problem, many studies used additional task-oriented data to pre-train their self-built networks from scratch or fine-tuned on well-known pre-trained models (e.g., AlexNet [25], VGG [26], VGG-face [148] and GoogleNet [27]). Kahou et al. [57], [149] indicated that the use of additional data can help to obtain models with high capacity without overfitting, thereby enhancing the FER performance.

To select appropriate auxiliary data, large-scale face recognition (FR) datasets (e.g., CASIA WebFace [150], Celebrity Face in the Wild (CFW) [151], FaceScrub dataset [152]) or relatively large FER datasets (FER2013 [21] and TFD [37]) are suitable. Kaya et al. [153] suggested that VGG-Face which was trained for FR overwhelmed ImageNet which was developed for objected recognition. Another interesting result observed by Knyazev et al. [154] is that pre-training on larger FR data positively affects the emotion recognition accuracy, and further fine-tuning with additional FER datasets can help improve the performance.

Instead of directly using the pre-trained or fine-tuned models to extract features on the target dataset, a multistage fine-tuning strategy [63] (see "Submission 3" in Fig. 3) can achieve better performance: after the first-stage fine-tuning using FER2013 on

TABLE 4
Performance summary of representative methods for static-based deep facial expression recognition on the most widely evaluated datasets.
Network size = depth & number of parameters; Pre-processing = Face Detection & Data Augmentation & Face Normalization; IN = Illumination
Normalization; $\mathcal{NE}$ = Network Ensemble; $\mathcal{CN}$ = Cascaded Network; $\mathcal{MN}$ = Multitask Network; LOSO = leave-one-subject-out.

Datasets	Method	Network type		Network size		Pre-processing			Data selection	Data group	Additional classifier	Performance[1] (%)
CK+	Ouellet 14 [110]	CNN (AlexNet)		-	-	V&J	-	-	the last frame	LOSO	SVM	7 classes[†]: (94.4)
	Li et al. 15 [86]	RBM		4	-	V&J	-	IN			✗	6 classes: 96.8
	Liu et al. 14 [13]	DBN	$\mathcal{CN}$	6	2m	✓	-	-		8 folds	AdaBoost	6 classes: 96.7
	Liu et al. 13 [137]	CNN, RBM	$\mathcal{CN}$	5	-	V&J	-	-		10 folds	SVM	8 classes: 92.05 (87.67)
	Liu et al. 15 [138]	CNN, RBM	$\mathcal{CN}$	5	-	V&J	-	-	the last three frames and the first frame	10 folds	SVM	7 classes[‡]: 93.70
	Khorrami et al. 15 [139]	zero-bias CNN		4	7m	✓	✓	-		10 folds	✗	6 classes: 95.7; 8 classes: 95.1
	Ding et al. 17 [111]	CNN	fine-tune	8	11m	IntraFace	✓	-		10 folds	✗	6 classes: (98.6); 8 classes: (96.8)
	Zeng et al. 18 [54]	DAE (DSAE)		3	-	AAM	-	-	the last four frames and the first frame	LOSO	✗	7 classes[†]: 95.79 (93.78) 8 classes: 89.84 (86.82)
	Cai et al. 17 [140]	CNN	loss layer	6	-	DRMF	✓	IN		10 folds	✗	7 classes[†]: 94.39 (90.66)
	Meng et al. 17 [61]	CNN	$\mathcal{MN}$	6	-	DRMF	✓	-		8 folds	✗	7 classes[†]: 95.37 (95.51)
	Liu et al. 17 [77]	CNN	loss layer	11	-	IntraFace	✓	IN	the last three frames	8 folds	✗	7 classes[†]: 97.1 (96.1)
	Yang et al. 18 [141]	GAN (cGAN)		-	-	MoT	✓	-		10 folds	✗	7 classes[†]: 97.30 (96.57)
	Zhang et al. 18 [47]	CNN	$\mathcal{MN}$	-	-	✓	✓	-		10 folds	✗	6 classes: 98.9
JAFFE	Liu et al. 14 [13]	DBN	$\mathcal{CN}$	6	2m	✓	-	-	213 images	LOSO	AdaBoost	7 classes[‡]: 91.8
	Hamester et al. 15 [142]	CNN, CAE	$\mathcal{NE}$	3	-	-	-	IN			✗	7 classes[‡]: (95.8)
MMI	Liu et al. 13 [137]	CNN, RBM	$\mathcal{CN}$	5	-	V&J	-	-	the middle three frames and the first frame	10 folds	SVM	7 classes[‡]: 74.76 (71.73)
	Liu et al. 15 [138]	CNN, RBM	$\mathcal{CN}$	5	-	V&J	-	-		10 folds	SVM	7 classes[‡]: 75.85
	Mollahosseini et al. 16 [14]	CNN (Inception)		11	7.3m	IntraFace	✓	-	images from each sequence	5 folds	✗	6 classes: 77.9
	Liu et al. 17 [77]	CNN	loss layer	11	-	IntraFace	✓	IN		10 folds	✗	6 classes: 78.53 (73.50)
	Li et al. 17 [44]	CNN	loss layer	8	5.8m	IntraFace	✓	-	the middle three frames	5 folds	SVM	6 classes: 78.46
	Yang et al. 18 [141]	GAN (cGAN)		-	-	MoT	✓	-		10 folds	✗	6 classes: 73.23 (72.67)
TFD	Reed et al. 14 [143]	RBM	$\mathcal{MN}$	-	-	-	-	-	4,178 emotion labeled 3,874 identity labeled	5 official folds	SVM	Test: 85.43
	Devries et al. 14 [58]	CNN	$\mathcal{MN}$	4	12.0m	MoT	✓	IN			✗	Validation: 87.80 Test: 85.13 (48.29)
	Khorrami et al. 15 [139]	zero-bias CNN		4	7m	✓	✓	-	4,178 labeled images		✗	Test: 88.6
	Ding et al. 17 [111]	CNN	fine-tune	8	11m	IntraFace	✓	-			✗	Test: 88.9 (87.7)
FER 2013	Tang 13 [130]	CNN	loss layer	4	12.0m	-	✓	IN			✗	Test: 71.2
	Devries et al. 14 [58]	CNN	$\mathcal{MN}$	4	12.0m	MoT	✓	IN			✗	Validation+Test: 67.21
	Zhang et al. 15 [144]	CNN	$\mathcal{MN}$	6	21.3m	SDM	-	-	Training Set: 28,709 Validation Set: 3,589 Test Set: 3,589		✗	Test: 75.10
	Guo et al. 16 [145]	CNN	loss layer	10	2.6m	SDM	✓	-			k-NN	Test: 71.33
	Kim et al. 16 [146]	CNN	$\mathcal{NE}$	5	2.4m	IntraFace	✓	IN			✗	Test: 73.73
	pramerdorfer et al. 16 [147]	CNN	$\mathcal{NE}$	10/16/33	1.8/1.2/5.3 (m)	-	✓	IN			✗	Test:75.2
SFEW 2.0	levi et al. 15 [78]	CNN	$\mathcal{NE}$	VGG-S/VGG-M/ GoogleNet		MoT	✓	-	891 training, 431 validation, and 372 test		✗	Validation: 51.75 Test: 54.56
	Ng et al. 15 [63]	CNN	fine-tune	AlexNet		IntraFace	✓	-	921 training, ? validation, and 372 test		✗	Validation: 48.5 (39.63) Test: 55.6 (42.69)
	Li et al. 17 [44]	CNN	loss layer	8	5.8m	IntraFace	✓	-	921 training, 427 validation		SVM	Validation: 51.05
	Ding et al. 17 [111]	CNN	fine-tune	8	11m	IntraFace	✓	-	891 training, 425 validation		✗	Validation: 55.15 (46.6)
	Liu et al. 17 [77]	CNN	loss layer	11	-	IntraFace	✓	IN			✗	Validation: 54.19 (47.97)
	Cai et al. 17 [140]	CNN	loss layer	6	-	DRMF	✓	IN			✗	Validation: 52.52 (43.41) Test: 59.41 (48.29)
	Meng et al. 17 [61]	CNN	$\mathcal{MN}$	6	-	DRMF	✓	-	958 training, 436 validation, and 372 test		✗	Validation: 50.98 (42.57) Test: 54.30 (44.77)
	Kim et al. 15 [76]	CNN	$\mathcal{NE}$	5	-	multiple	✓	IN			✗	Validation: 53.9 Test: 61.6
	Yu et al. 15 [75]	CNN	$\mathcal{NE}$	8	6.2m	multiple	✓	IN			✗	Validation: 55.96 (47.31) Test: 61.29 (51.27)

[1] The value in parentheses is the mean accuracy, which is calculated with the confusion matrix given by the authors.
[†] 7 Classes: Anger, Contempt, Disgust, Fear, Happiness, Sadness, and Surprise.
[‡] 7 Classes: Anger, Disgust, Fear, Happiness, Neutral, Sadness, and Surprise.
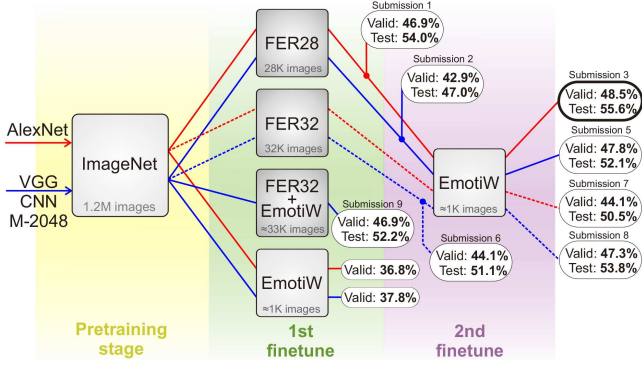
Fig. 3. Flowchart of the different fine-tuning combinations used in [63]. Here, "FER28" and "FER32" indicate different parts of the FER2013 datasets. "EmotiW" is the target dataset. The proposed two-stage fine-tuning strategy (Submission 3) exhibited the best performance.
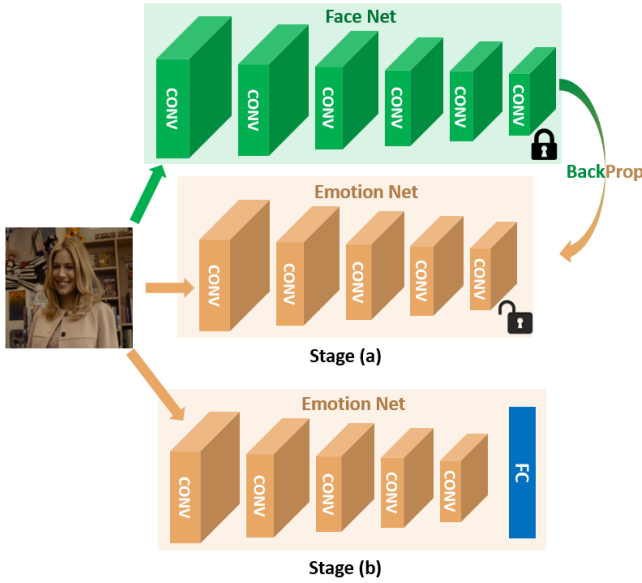

Fig. 4. Two-stage training flowchart in [111]. In stage (a), the deeper face net is frozen and provides the feature-level regularization that pushes the convolutional features of the expression net to be close to the face net by using the proposed distribution function. Then, in stage (b), to further improve the discriminativeness of the learned features, randomly initialized fully connected layers are added and jointly trained with the whole expression net using the expression label information.

pre-trained models, a second-stage fine-tuning based on the training part of the target dataset (EmotiW) is employed to refine the models to adapt to a more specific dataset (i.e., the target dataset).

Although pre-training and fine-tuning on external FR data can indirectly avoid the problem of small training data, the networks are trained separately from the FER and the face-dominated information remains in the learned features which may weaken the networks ability to represent expressions. To eliminate this effect, a two-stage training algorithm FaceNet2ExpNet [111] was proposed (see Fig. 4). The fine-tuned face net serves as a good initialization for the expression net and is used to guide the learning of the convolutional layers only. And the fully connected layers are trained from scratch with expression information to regularize the training of the target FER net.

Fig. 5. Image intensities (left) and LBP codes (middle). [78] proposed mapping these values to a 3D metric space (right) as the input of CNNs.

4.1.2 Diverse network input

Traditional practices commonly use the whole aligned face of RGB images as the input of the network to learn features for FER. However, these raw data lack important information, such as homogeneous or regular textures and invariance in terms of image scaling, rotation, occlusion and illumination, which may represent confounding factors for FER. Some methods have employed diverse handcrafted features and their extensions as the network input to alleviate this problem.

Low-level representations encode features from small regions in the given RGB image, then cluster and pool these features with local histograms, which are robust to illumination variations and small registration errors. A novel mapped LBP feature [78] (see Fig. 5) was proposed for illumination-invariant FER. Scale-invariant feature transform (SIFT) [155]) features that are robust against image scaling and rotation are employed [156] for multi-view FER tasks. Combining different descriptors in outline, texture, angle, and color as the input data can also help enhance the deep network performance [54], [157].

Part-based representations extract features according to the target task, which remove noncritical parts from the whole image and exploit key parts that are sensitive to the task. [158] indicated that three regions of interest (ROI), i.e., eyebrows, eyes and mouth, are strongly related to facial expression changes, and cropped these regions as the input of DSAE. Other researches proposed to automatically learn the key parts for facial expression. For example, [159] employed a deep multi-layer network [160] to detect the saliency map which put intensities on parts demanding visual attention. And [161] applied the neighbor-center difference vector (NCDV) [162] to obtain features with more intrinsic information.

4.1.3 Auxiliary blocks & layers

Based on the foundation architecture of CNN, several studies have proposed the addition of well-designed auxiliary blocks or layers to enhance the expression-related representation capability of learned features.

Based on the foundation architecture of CNN, several studies have proposed the addition of well-designed auxiliary blocks or layers to enhance the expression-related representation capability of learned features.

A novel CNN architecture, HoloNet [90], was designed for FER, where CReLU [163] was combined with the powerful residual structure [28] to increase the network depth without efficiency reduction and an inception-residual block [164], [165] was uniquely designed for FER to learn multi-scale features to capture variations in expressions. Another CNN model, Supervised Scoring Ensemble (SSE) [91], was introduced to enhance
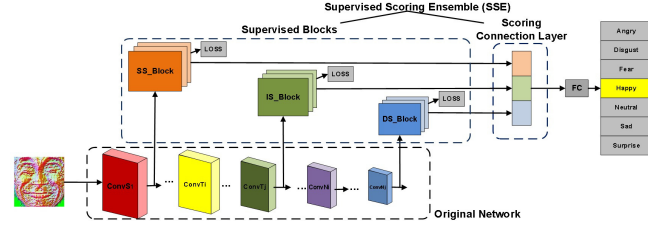
(a) Three different supervised blocks in [91]. SS_Block for shallow-layer supervision, IS_Block for intermediate-layer supervision, and DS_Block for deep-layer supervision.
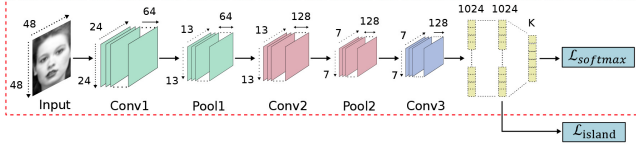
(b) Island loss layer in [140]. The island loss calculated at the feature extraction layer and the softmax loss calculated at the decision layer are combined to supervise the CNN training.
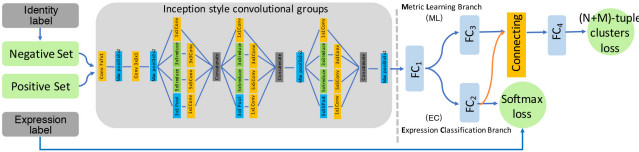
(c) (N+M)-tuple clusters loss layer in [77]. During training, the identity-aware hard-negative mining and online positive mining schemes are used to decrease the inter-identity variation in the same expression class.

Fig. 6. Representative functional layers or blocks that are specifically designed for deep facial expression recognition.

the supervision degree for FER, where three types of supervised blocks were embedded in the early hidden layers of the mainstream CNN for shallow, intermediate and deep supervision, respectively (see Fig. 6(a)).

And a feature selection network (FSN) [166] was designed by embedding a feature selection mechanism inside the AlexNet, which automatically filters irrelevant features and emphasizes correlated features according to learned feature maps of facial expression. Interestingly, Zeng et al. [167] pointed out that the inconsistent annotations among different FER databases are inevitable which would damage the performance when the training set is enlarged by merging multiple datasets. To address this problem, the authors proposed an Inconsistent Pseudo Annotations to Latent Truth (IPA2LT) framework. In IPA2LT, an end-to-end trainable LTNet is designed to discover the latent truths from the human annotations and the machine annotations trained from different datasets by maximizing the log-likelihood of these inconsistent annotations.

The traditional softmax loss layer in CNNs simply forces features of different classes to remain apart, but FER in real-world scenarios suffers from not only high inter-class similarity but also high intra-class variation. Therefore, several works have proposed novel loss layers for FER. Inspired by the center loss [168], which penalizes the distance between deep features and their corresponding class centers, two variations were proposed to assist the supervision of the softmax loss for more discriminative features for FER: (1) island loss [140] was formalized to further increase the pairwise distances between different class centers (see Fig. 6(b)), and (2) locality-preserving loss (LP loss) [44] was

TABLE 5
Three primary ensemble methods on the decision level.

|	definition	used in (example)
Majority Voting	determine the class with the most votes using the predicted label yielded from each individual	[76], [146], [173]
Simple Average	determine the class with the highest mean score using the posterior class probabilities yielded from each individual with the same weight	[76], [146], [173]
Weighted Average	determine the class with the highest weighted mean score using the posterior class probabilities yielded from each individual with different weights	[57], [78], [147], [153]

formalized to pull the locally neighboring features of the same class together so that the intra-class local clusters of each class are compact. Besides, based on the triplet loss [169], which requires one positive example to be closer to the anchor than one negative example with a fixed gap, two variations were proposed to replace or assist the supervision of the softmax loss: (1) exponential triplet-based loss [145] was formalized to give difficult samples more weight when updating the network, and (2) (N+M)-tuples cluster loss [77] was formalized to alleviate the difficulty of anchor selection and threshold validation in the triplet loss for identity-invariant FER (see Fig. 6(c) for details). Besides, a feature loss [170] was proposed to provide complementary information for the deep feature during early training stage.
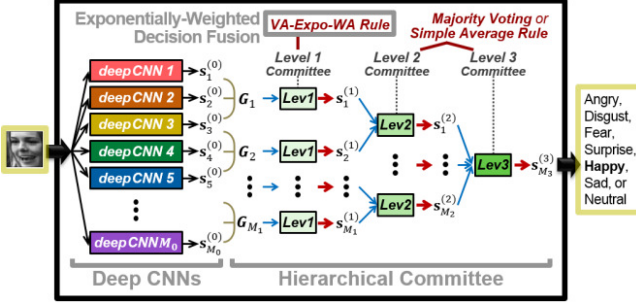
4.1.4 Network ensemble

Previous research suggested that assemblies of multiple networks can outperform an individual network [171]. Two key factors should be considered when implementing network ensembles: (1) sufficient diversity of the networks to ensure complementarity, and (2) an appropriate ensemble method that can effectively aggregate the committee networks.

In terms of the first factor, different kinds of training data and various network parameters or architectures are considered to generate diverse committees. Several pre-processing methods [146], such as deformation and normalization, and methods described in Section 4.1.2 can generate different data to train diverse networks. By changing the size of filters, the number of neurons and the number of layers of the networks, and applying multiple random seeds for weight initialization, the diversity of the networks can also be enhanced [76], [172]. Besides, different architectures of networks can be used to enhance the diversity. For example, a CNN trained in a supervised way and a convolutional autoencoder (CAE) trained in an unsupervised way were combined for network ensemble [142].

For the second factor, each member of the committee networks can be assembled at two different levels: the feature level and the decision level. For *feature-level* ensembles, the most commonly adopted strategy is to concatenate features learned from different networks [88], [174]. For example, [88] concatenated features learned from different networks to obtain a single feature vector to describe the input image (see Fig. 7(a)). For *decision-level*
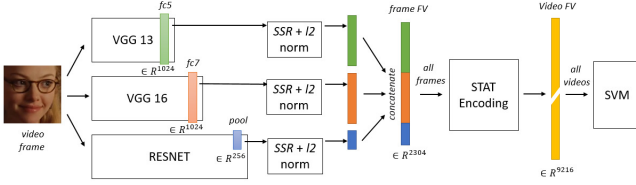
(a) Feature-level ensemble in [88]. Three different features (*fc5* of VGG13 + *fc7* of VGG16 + *pool* of Resnet) after normalization are concatenated to create a single feature vector (FV) that describes the input frame.

(b) Decision-level ensemble in [76]. A 3-level hierarchical committee architecture with hybrid decision-level fusions was proposed to obtain sufficient decision diversity.

Fig. 7. Representative network ensemble systems at the feature level and decision level.

ensembles, three widely-used rules are applied: majority voting, simple average and weighted average. A summary of these three methods is provided in Table 5. Because the weighted average rule considers the importance and confidence of each individual, many weighted average methods have been proposed to find an optimal set of weights for network ensemble. [57] proposed a random search method to weight the model predictions for each emotion type. [75] used the log-likelihood loss and hinge loss to adaptively assign different weights to each network. [76] proposed an exponentially weighted average based on the validation accuracy to emphasize qualified individuals (see Fig. 7(b)). [172] used a CNN to learn weights for each individual model.

4.1.5 Multitask networks

Many existing networks for FER focus on a single task and learn features that are sensitive to expressions without considering interactions among other latent factors. However, in the real world, FER is intertwined with various factors, such as head pose, illumination, and subject identity (facial morphology). To solve this problem, multitask leaning is introduced to transfer knowledge from other relevant tasks and to disentangle nuisance factors.

Reed et al. [143] constructed a higher-order Boltzmann machine (disBM) to learn manifold coordinates for the relevant factors of expressions and proposed training strategies for disentangling so that the expression-related hidden units are invariant to face morphology. Other works [58], [175] suggested that simultaneously conducted FER with other tasks, such as facial landmark localization and facial AUs [176] detection, can jointly improve FER performance.

Besides, several works [61], [68] employed multitask learning for identity-invariant FER. In [61], an identity-aware CNN (IACNN) with two identical sub-CNNs was proposed. One stream used expression-sensitive contrastive loss to learn expression-discriminative features, and the other stream used identity-
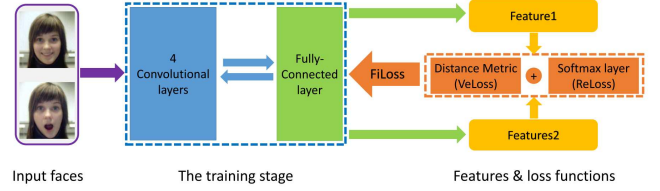

Fig. 8. Representative multitask network for FER. In the proposed MSCNN [68], a pair of images is sent into the MSCNN during training. The expression recognition task with cross-entropy loss, which learns features with large between-expression variation, and the face verification task with contrastive loss, which reduces the variation in within-expression features, are combined to train the MSCNN.
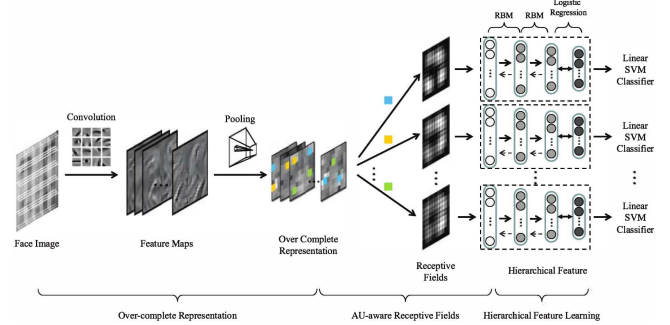

Fig. 9. Representative cascaded network for FER. The proposed AU-aware deep network (AUDN) [137] is composed of three sequential modules: in the first module, a 2-layer CNN is trained to generate an over-complete representation encoding all expression-specific appearance variations over all possible locations; in the second module, an AU-aware receptive field layer is designed to search subsets of the over-complete representation; in the last module, a multilayer RBM is exploited to learn hierarchical features.

sensitive contrastive loss to learn identity-related features for identity-invariant FER. In [68], a multisignal CNN (MSCNN), which was trained under the supervision of both FER and face verification tasks, was proposed to force the model to focus on expression information (see Fig. 8). Furthermore, an all-in-one CNN model [177] was proposed to simultaneously solve a diverse set of face analysis tasks including smile detection. The network was first initialized using the weight pre-trained on face recognition, then task-specific sub-networks were branched out from different layers with domain-based regularization by training on multiple datasets. Specifically, as smile detection is a subject-independent task that relies more on local information available from the lower layers, the authors proposed to fuse the lower convolutional layers to form a generic representation for smile detection. Conventional supervised multitask learning requires training samples labeled for all tasks. To relax this, [47] proposed a novel attribute propagation method which can leverage the inherent correspondences between facial expression and other heterogeneous attributes despite the disparate distributions of different datasets.

4.1.6 Cascaded networks

In a cascaded network, various modules for different tasks are combined sequentially to construct a deeper network, where the outputs of the former modules are utilized by the latter modules. Related studies have proposed combinations of different structures to learn a hierarchy of features through which factors of variation that are unrelated with expressions can be gradually filtered out.

Most commonly, different networks or learning methods are combined sequentially and individually, and each of them contributes differently and hierarchically. In [178], DBNs were trained to first detect faces and to detect expression-related areas. Then, these parsed face components were classified by a stacked autoencoder. In [179], a multiscale contractive convolutional network (CCNET) was proposed to obtain local-translation-invariant (LTI) representations. Then, contractive autoencoder was designed to hierarchically separate out the emotion-related factors from subject identity and pose. In [137], [138], over-complete representations were first learned using CNN architecture, then a multilayer RBM was exploited to learn higher-level features for FER (see Fig. 9). Instead of simply concatenating different networks, Liu et al. [13] presented a boosted DBN (BDBN) that iteratively performed feature representation, feature selection and classifier construction in a unified loopy state. Compared with the concatenation without feedback, this loopy framework propagates backward the classification error to initiate the feature selection process alternately until convergence. Thus, the discriminative ability for FER can be substantially improved during this iteration.

4.1.7 Generative adversarial networks (GANs)

Recently, GAN-based methods have been successfully used in image synthesis to generate impressively realistic faces, numbers, and a variety of other image types, which are beneficial to training data augmentation and the corresponding recognition tasks. Several works have proposed novel GAN-based models for pose-invariant FER and identity-invariant FER.

For pose-invariant FER, Lai et al. [180] proposed a GAN-based face frontalization framework, where the generator frontalizes input face images while preserving the identity and expression characteristics and the discriminator distinguishes the real images from the generated frontal face images. And Zhang et al. [181] proposed a GAN-based model that can generate images with different expressions under arbitrary poses for multi-view FER. For identity-invariant FER, Yang et al. [182] proposed an Identity-Adaptive Generation (IA-gen) model with two parts. The upper part generates images of the same subject with different expressions using cGANs, respectively. Then, the lower part conducts FER for each single identity sub-space without involving other individuals, thus identity variations can be well alleviated. Chen et al. [183] proposed a Privacy-Preserving Representation-Learning Variational GAN (PPRL-VGAN) that combines VAE and GAN to learn an identity-invariant representation that is explicitly disentangled from the identity information and generative for expression-preserving face image synthesis. Yang et al. [141] proposed a De-expression Residue Learning (DeRL) procedure to explore the expressive information, which is filtered out during the de-expression process but still embedded in the generator. Then the model extracted this information from the generator directly to mitigate the influence of subject variations and improve the FER performance.

4.1.8 Discussion

The existing well-constructed deep FER systems focus on two key issues: the lack of plentiful diverse training data and expression-unrelated variations, such as illumination, head pose and identity. Table 6 shows relative advantages and disadvantages of these different types of methods with respect to two open issues (data size requirement and expression-unrelated variations) and other

focuses (computation efficiency, performance and difficulty of network training).

Pre-training and fine-tuning have become mainstream in deep FER to solve the problem of insufficient training data and overfitting. A practical technique that proved to be particularly useful is pre-training and fine-tuning the network in multiple stages using auxiliary data from large-scale objection or face recognition datasets to small-scale FER datasets, i.e., from large to small and from general to specific. However, when compared with the end-to-end training framework, the representational structure that are unrelated to expressions are still remained in the off-the-shelf pre-trained model, such as the large domain gap with the objection net [153] and the subject identification distraction in the face net [111]. Thus the extracted features are usually vulnerable to identity variations and the performance would be degraded. Noticeably, with the advent of large-scale in-the-wild FER datasets (e.g., AffectNet and RAF-DB), the end-to-end training using deep networks with moderate size can also achieve competitive performances [45], [167].

In addition to directly using the raw image data to train the deep network, *diverse pre-designed features* are recommended to strengthen the network's robustness to common distractions (e.g., illumination, head pose and occlusion) and to force the network to focus more on facial areas with expressive information. Moreover, the use of multiple heterogeneous input data can indirectly enlarge the data size. However, the problem of identity bias is commonly ignored in this methods. Moreover, generating diverse data accounts for additional time consuming and combining these multiple data can lead to high dimension which may influence the computational efficiency of the network.

Training a deep and wide network with a large number of hidden layers and flexible filters is an effective way to learn deep high-level features that are discriminative for the target task. However, this process is vulnerable to the size of training data and can underperform if insufficient training data is available to learn the new parameters. Integrating multiple relatively small networks in parallel or in series is a natural research direction to overcome this problem. *Network ensemble* integrates diverse networks at the feature or decision level to combine their advantages, which is usually applied in emotion competitions to help boost the performance. However, designing different kinds of networks to compensate each other obviously enlarge the computational cost and the storage requirement. Moreover, the weight of each sub-network is usually learned according to the performance on original training data, leading to overfitting on newly unseen testing data. *Multitask networks* jointly train multiple networks with consideration of interactions between the target FER task and other secondary tasks, such as facial landmark localization, facial AU recognition and face verification, thus the expression-unrelated factors including identity bias can be well disentangled. The downside of this method is that it requires labeled data from all tasks and the training becomes increasingly cumbersome as more tasks are involved. Alternatively, *cascaded networks* sequentially train multiple networks in a hierarchical approach, in which case the discriminative ability of the learned features are continuously strengthened. In general, this method can alleviate the overfitting problem, and in the meanwhile, progressively disentangling factors that are irrelevant to facial expression. A deficiency worths considering is that the sub-networks in most existing cascaded systems are training individually without feedback, and the end-to-end training strategy is preferable to enhance the training

TABLE 6
Comparison of different types of methods for static images in terms of data size requirement, variations* (head pose, illumination, occlusion and other environment factors), identity bias, computational efficiency, accuracy, and difficulty on network training.

Network type	data	variations*	identity bias	efficiency	accuracy	difficulty
Pre-train & Fine-tune	low	fair	vulnerable	high	fair	easy
Diverse input	low	good	vulnerable	low	fair	easy
Auxiliary layers	varies	good	varies	varies	good	varies
Network ensemble	low	good	fair	low	good	medium
Multitask network	high	varies	good	fair	varies	hard
Cascaded network	fair	good	fair	fair	fair	medium
GAN	fair	good	good	fair	good	hard

effectiveness and the performance [13].

Ideally, deep networks, especially CNNs, have good capabilities for dealing with head-pose variations, yet most current FER networks do not address head-pose variations explicitly and are not tested in naturalistic scenarios. *Generative adversarial networks (GANs)* can be exploited to solve this issue by frontalizing face images while preserving expression characteristics [180] or synthesizing arbitrary poses to help train the pose-invariant network [181]. Another advantage of GANs is that the identity variations can be explicitly disentangled through generating the corresponding neutral face image [141] or synthesizing different expressions while preserving the identity information for identity-invariant FER [182]. Moreover, GANs can help augment the training data on both size and diversity. The main drawback of GAN is the training instability and the trade-off between visual quality and image diversity.

4.2 Deep FER networks for dynamic image sequences

Although most of the previous models focus on static images, facial expression recognition can benefit from the temporal correlations of consecutive frames in a sequence. We first introduce the existing frame aggregation techniques that strategically combine deep features learned from static-based FER networks. Then, considering that in a videostream people usually display the same expression with different intensities, we further review methods that use images in different expression intensity states for intensity-invariant FER. Finally, we introduce deep FER networks that consider spatio-temporal motion patterns in video frames and learned features derived from the temporal structure. For each of the most frequently evaluated datasets, Table 7 shows the current state-of-the-art methods conducted in the person-independent protocol.

4.2.1 Frame aggregation

Because the frames in a given video clip may vary in expression intensity, directly measuring per-frame error does not yield satisfactory performance. Various methods have been proposed to aggregate the network output for frames in each sequence to improve the performance. We divide these methods into two groups: decision-level frame aggregation and feature-level frame aggregation.

For decision-level frame aggregation, n-class probability vectors of each frame in a sequence are integrated. The most convenient way is to directly concatenate the output of these frames. However, the number of frames in each sequence may be different.
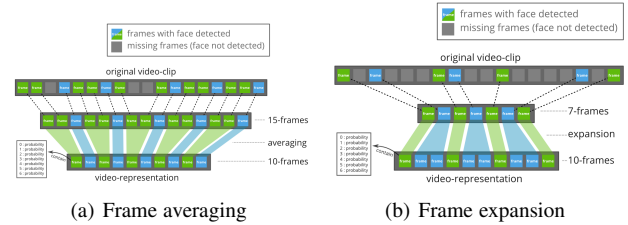

Fig. 10. Frame aggregation in [57]. The flowchart is top-down. (a) For sequences with more than 10 frames, we averaged the probability vectors of 10 independent groups of frames taken uniformly along time. (b) For sequences with less than 10 frames, we expanded by repeating frames uniformly to obtain 10 total frames.

Two aggregation approaches have been considered to generate a fixed-length feature vector for each sequence [57], [191]: frame averaging and frame expansion (see Fig. 10 for details). An alternative approach which dose not require a fixed number of frames is applying statistical coding. The average, max, average of square, average of maximum suppression vectors and so on can be used to summarize the per-frame probabilities in each sequence.

For feature-level frame aggregation, the learned features of frames in the sequence are aggregate. Many statistical-based encoding modules can be applied in this scheme. A simple and effective way is to concatenate the mean, variance, minimum, and maximum of the features over all frames [88]. Alternatively, matrix-based models such as eigenvector, covariance matrix and multi-dimensional Gaussian distribution can also be employed for aggregation [186], [192]. Besides, multi-instance learning has been explored for video-level representation [193], where the cluster centers are computed from auxiliary image data and then bag-of-words representation is obtained for each bag of video frames.

4.2.2 Expression Intensity network

Most methods (introduced in Section 4.1) focus on recognizing the peak high-intensity expression and ignore the subtle lower-intensity expressions. In this section, we introduced expression intensity-invariant networks that take training samples with different intensities as input to exploit the intrinsic correlations among expressions from a sequence that vary in intensity.

In expression intensity-invariant network, image frames with intensity labels are used for training. During test, data that vary in expression intensity are used to verify the intensity-invariant ability of the network. Zhao et al. [17] proposed a peak-piloted

TABLE 7
Performances of representative methods for dynamic-based deep facial expression recognition on the most widely evaluated datasets. Network size = depth & number of parameters; Pre-processing = Face Detection & Data Augmentation & Face Normalization; IN = Illumination Normalization; $\mathcal{FA}$ = Frame Aggregation; $\mathcal{EIN}$ = Expression Intensity-invariant Network; $\mathcal{FLT}$ = Facial Landmark Trajectory; $\mathcal{CN}$ = Cascaded Network; $\mathcal{NE}$ = Network Ensemble; S = Spatial Network; T = Temporal Network; LOSO = leave-one-subject-out.

Datasets	Methods	Network type	Network size		Pre-processing			Training data Selection in each sequence	Testing data selection in each sequence	Data group	Performance[1](%)
CK+	Zhao et al. 16 [17]	$\mathcal{EIN}$	22	6.8m	✓	-	-	from the 7th to the last[2]	the last frame	10 folds	6 classes: 99.3
	Yu et al. 17 [70]	$\mathcal{EIN}$	42	-	MTCNN	✓	-	from the 7th to the last[2]	the peak expression	10 folds	6 classes: 99.6
	kim et al. 17 [184]	$\mathcal{EIN}$	14	-	✓	✓	-	all frames		10 folds	7 classes: 97.93
	Sun et al. 17 [185]	$\mathcal{NE}$	3 * GoogLeNetv2		✓	-	-	S: emotional T: neutral+emotional		10 folds	6 classes: 97.28
	Jung et al. 15 [16]	$\mathcal{FLT}$	2	177.6k	IntraFace	✓	-	fixed number of frames		10 folds	7 classes: 92.35
	Jung et al. 15 [16]	C3D	4	-	IntraFace	✓	-	fixed number of frames	the same as the training data	10 folds	7 classes: 91.44
	Jung et al. 15 [16]	$\mathcal{NE}$	$\mathcal{FLT}$/C3D		IntraFace	✓	-	fixed number of frames		10 folds	7 classes: 97.25 (95.22)
	kuo et al. 18 [89]	$\mathcal{FA}$	6	2.7m	IntraFace	✓	IN	fixed length 9		10 folds	7 classes: 98.47
	Zhang et al. 17 [68]	$\mathcal{NE}$	7/5	2k/1.6m	SDM/ Cascaded CNN	✓	-	S: the last frame T: all frames		10 folds	7 classes: 98.50 (97.78)
MMI	Kim et al. 17 [66]	$\mathcal{EIN}, \mathcal{CN}$	7	1.5m	Incremental	✓	-	5 intensities frames		LOSO	6 classes: 78.61 (78.00)
	kim et al. 17 [184]	$\mathcal{EIN}$	14	-	✓	✓	-	all frames		10 folds	6 classes: 81.53
	Hasani et al. 17 [112]	$\mathcal{FLT}, \mathcal{CN}$	22	-	3000 fps	-	-	ten frames	the same as the training data	5 folds	6 classes: 77.50 (74.50)
	Hasani et al. 17 [55]	$\mathcal{CN}$	29	-	AAM	-	-	static frames		5 folds	6 classes: 78.68
	Zhang et al. 17 [68]	$\mathcal{NE}$	7/5	2k/1.6m	SDM/ Cascaded CNN	✓	-	S: the middle frame T: all frames		10 folds	6 classes: 81.18 (79.30)
	Sun et al. 17 [185]	$\mathcal{NE}$	3 * GoogLeNetv2		✓	-	-	S: emotional T: neutral+emotional		10 folds	6 classes: 91.46
Oulu-CASIA	Zhao et al. 16 [17]	$\mathcal{EIN}$	22	6.8m	✓	-	-	from the 7th to the last[2]	the last frame	10 folds	6 classes: 84.59
	Yu et al. 17 [70]	$\mathcal{EIN}$	42	-	MTCNN	✓	-	from the 7th to the last[2]	the peak expression	10 folds	6 classes: 86.23
	Jung et al. 15 [16]	$\mathcal{FLT}$	2	177.6k	IntraFace	✓	-	fixed number of frames		10 folds	6 classes: 74.17
	Jung et al. 15 [16]	C3D	4	-	IntraFace	✓	-	fixed number of frames		10 folds	6 classes: 74.38
	Jung et al. 15 [16]	$\mathcal{NE}$	$\mathcal{FLT}$/C3D		IntraFace	✓	-	fixed number of frames	the same as the training data	10 folds	6 classes: 81.46 (81.49)
	Zhang et al. 17 [68]	$\mathcal{NE}$	7/5	2k/1.6m	SDM/ Cascaded CNN	✓	-	S: the last frame T: all frames		10 folds	6 classes: 86.25 (86.25)
	kuo et al. 18 [89]	$\mathcal{NE}$	6	2.7m	IntraFace	✓	IN	fixed length 9		10 folds	6 classes: 91.67
AFEW*6.0	Ding et al. 16 [186]	$\mathcal{FA}$	AlexNet		✓	-	-	Training: 773; Validation: 373; Test: 593			Validation: 44.47
	Yan et al. 16 [187]	$\mathcal{CN}$	VGG16-LSTM		✓	✓	-	40 frames		3 folds	7 classes: 44.46
	Yan et al. 16 [187]	$\mathcal{FLT}$	4	-	[188]	-	-	30 frames		3 folds	7 classes: 37.37
	Fan et al. 16 [108]	$\mathcal{CN}$	VGG16-LSTM		✓	-	-	16 features for LSTM			Validation: 45.43 (38.96)
	Fan et al. [108]	C3D	10	-	✓	-	-	several windows of 16 consecutive frames			Validation: 39.69 (38.55)
	Yan et al. 16 [187]	fusion	/					Training: 773; Validation: 383; Test: 593			Test: 56.66 (40.81)
	Fan et al. 16 [108]	fusion	/					Training: 774; Validation: 383; Test: 593			Test: 59.02 (44.94)
AFEW*7.0	Ouyang et al. 17 [189]	$\mathcal{CN}$	VGG-LSTM		MTCNN	✓	-	16 frames			Validation: 47.4
	Ouyang et al. 17 [189]	C3D	10	-	MTCNN	✓	-	16 frames			Validation: 35.2
	Vielzeuf et al. [190]	$\mathcal{CN}$	C3D-LSTM		✓	✓	-	detected face frames			Validation: 43.2
	Vielzeuf et al. [190]	$\mathcal{CN}$	VGG16-LSTM		✓	✓	-	several windows of 16 consecutive frames			Validation: 48.6
	Vielzeuf et al. [190]	fusion	/					Training: 773; Validation: 383; Test: 653			Test: 58.81 (43.23)

[1] The value in parentheses is the mean accuracy calculated from the confusion matrix given by authors.

[2] A pair of images (peak and non-peak expression) is chosen for training each time.

* We have included the result of a single spatio-temporal network and also the best result after fusion with both video and audio modalities.

[†] 7 Classes in CK+: Anger, Contempt, Disgust, Fear, Happiness, Sadness, and Surprise.

[‡] 7 Classes in AFEW: Anger, Disgust, Fear, Happiness, Neutral, Sadness, and Surprise.

deep network (PPDN) that takes a pair of peak and non-peak images of the same expression and from the same subject as input and utilizes the L2-norm loss to minimize the distance between both images. During back propagation, a peak gradient suppression (PGS) was proposed to drive the learned feature of the non-peak expression towards that of peak expression while avoiding the inverse. Thus, the network discriminant ability on lower-intensity expressions can be improved. Based on PPDN, Yu et al. [70] proposed a deeper cascaded peak-piloted network (DCPN) that used a deeper and larger architecture to enhance the discriminative ability of the learned features and employed an integration training method called cascade fine-tuning to avoid overfitting. In [66], more intensity states were utilized (onset, onset to apex transition, apex, apex to offset transition and offset) and five loss functions were adopted to regulate the network training by minimizing expression classification error, intra-class expression variation, intensity classification error and intra-intensity variation,

and encoding intermediate intensity, respectively.

Considering that images with different expression intensities for an individual identity is not always available in the wild, several works proposed to automatically acquire the intensity label or to generate new images with targeted intensity. For example, in [194] the peak and neutral frames was automatically picked out from the sequence with two stages : a clustering stage to divide all frames into the peak-like group and the neutral-like group using K-means algorithm, and a classification stage to detect peak and neutral frames using a semi-supervised SVM. And in [184], a deep generative-contrastive model was presented with two steps: a generator to generate the reference (less-expressive) face for each sample via convolutional encoder-decoder and a contrastive network to jointly filter out information that is irrelevant with expressions through a contrastive metric loss and a supervised reconstruction loss.
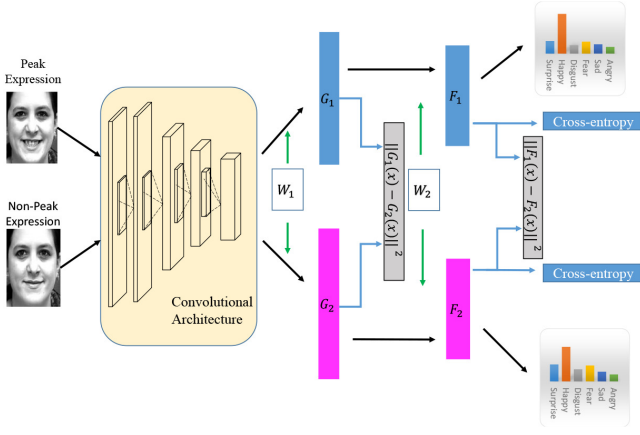
Fig. 11. The proposed PPDN in [17]. During training, PPDN is trained by jointly optimizing the L2-norm loss and the cross-entropy losses of two expression images. During testing, the PPDN takes one still image as input for probability prediction.

Fig. 12. The proposed 3DCNN-DAP [199]. The input n-frame sequence is convolved with 3D filters; then, $13 * c * k$ part filters corresponding to 13 manually defined facial parts are used to convolve k feature maps for the facial action part detection maps of c expression classes.

4.2.3 Deep spatio-temporal FER network

Although the frame aggregation can integrate frames in the video sequence, the crucial temporal dependency is not explicitly exploited. By contrast, the spatio-temporal FER network takes a range of frames in a temporal window as a single input without prior knowledge of the expression intensity and utilizes both textural and temporal information to encode more subtle expressions.

RNN and C3D: RNN can robustly derive information from sequences by exploiting the fact that feature vectors for successive data are connected semantically and are therefore interdependent. The improved version, LSTM, is flexible to handle varying-length sequential data with lower computation cost. Derived from RNN, an RNN that is composed of ReLUs and initialized with the identity matrix (IRNN) [195] was used to provide a simpler mechanism for addressing the vanishing and exploding gradient problems [87]. And bidirectional RNNs (BRNNs) [196] were employed to learn the temporal relations in both the original and reversed directions [68], [187]. Recently, a Nested LSTM was proposed in [71] with two sub-LSTMs. Namely, T-LSTM models the temporal dynamics of the learned features, and C-LSTM integrates the outputs of all T-LSTMs together so as to encode the multi-level features encoded in the intermediate layers of the network.

Compared with RNN, CNN is more suitable for computer vision applications; hence, its derivative C3D [107], which uses 3D convolutional kernels with shared weights along the time axis instead of the traditional 2D kernels, has been widely used for dynamic-based FER (e.g., [83], [108], [189], [197], [198]) to capture the spatio-temporal features. Based on C3D, many derived structures have been designed for FER. In [199], 3D CNN was incorporated with the DPM-inspired [200] deformable facial action constraints to simultaneously encode dynamic motion and discriminative part-based representations (see Fig. 12 for details). In [16], a deep temporal appearance network (DTAN) was proposed that employed 3D filters without weight sharing along the time axis; hence, each filter can vary in importance over time. Likewise, a weighted C3D was proposed [190], where several win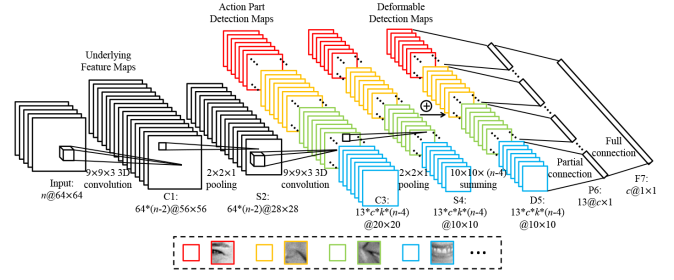dows of consecutive frames were extracted from each sequence and weighted based on their prediction scores. Instead of directly using C3D for classification, [109] employed C3D for spatio-temporal feature extraction and then cascaded with DBN for prediction. In [201], C3D was also used as a feature extractor, followed by a NetVLAD layer [202] to aggregate the temporal information of the motion features by learning cluster centers.

Facial landmark trajectory: Related psychological studies have shown that expressions are invoked by dynamic motions of certain facial parts (e.g., eyes, nose and mouth) that contain the most descriptive information for representing expressions. To obtain more accurate facial actions for FER, facial landmark trajectory models have been proposed to capture the dynamic variations of facial components from consecutive frames.

To extract landmark trajectory representation, the most direct way is to concatenate coordinates of facial landmark points from frames over time with normalization to generate a one-dimensional trajectory signal for each sequence [16] or to form an image-like map as the input of CNN [187]. Besides, relative distance variation of each landmark in consecutive frames can also be used to capture the temporal information [203]. Further, part-based model that divides facial landmarks into several parts according to facial physical structure and then separately feeds them into the networks hierarchically is proved to be efficient for both local low-level and global high-level feature encoding [68] (see "PHRNN" in Fig. 13) . Instead of separately extracting the trajectory features and then input them into the networks, Hasani et al. [112] incorporated the trajectory features by replacing the shortcut in the residual unit of the original 3D Inception-ResNet with element-wise multiplication of facial landmarks and the input tensor of the residual unit. Thus, the landmark based network can be trained end-to-end.

Cascaded networks: By combining the powerful perceptual vision representations learned from CNNs with the strength of LSTM for variable-length inputs and outputs, Donahue et al. [204] proposed a both spatially and temporally deep model which cascades the outputs of CNNs with LSTMs for various vision tasks involving time-varying inputs and outputs. Similar to this hybrid network, many cascaded networks have been proposed for FER (e.g., [66], [108], [190], [205]).

Instead of CNN, [206] employed a convolutional sparse autoencoder for sparse and shift-invariant features; then, an LSTM classifier was trained for temporal evolution. [189] employed a more flexible network called ResNet-LSTM, which allows nodes in lower CNN layers to directly contact with
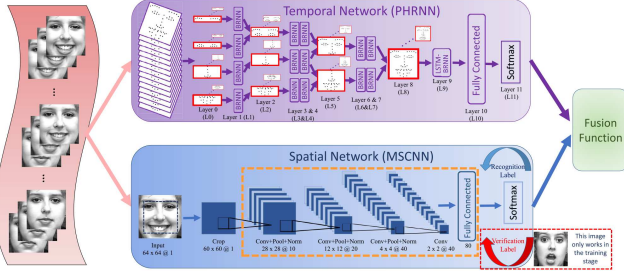
Fig. 13. The spatio-temporal network proposed in [68]. The temporal network PHRNN for landmark trajectory and the spatial network MSCNN for identity-invariant features are trained separately. Then, the predicted probabilities from the two networks are fused together for spatio-temporal FER.
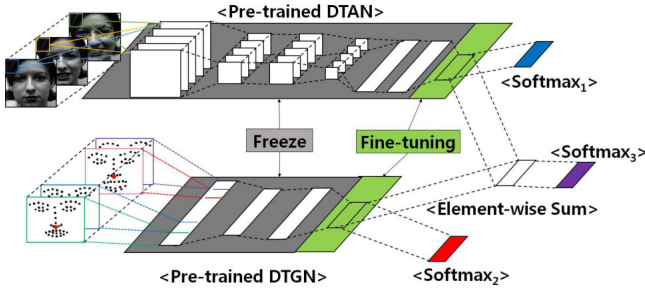

Fig. 14. The joint fine-tuning method for DTAGN proposed in [16]. To integrate DTGA and DTAN, we freeze the weight values in the gray boxes and retrain the top layer in the green boxes. The logit values of the green boxes are used by Softmax3 to supervise the integrated network. During training, we combine three softmax loss functions, and for prediction, we use only Softmax3.

LSTMs to capture spatio-temporal information. In addition to concatenating LSTM with the fully connected layer of CNN, a hypercolumn-based system [207] extracted the last convolutional layer features as the input of the LSTM for longer range dependencies without losing global coherence. Instead of LSTM, the conditional random fields (CRFs) model [208] that are effective in recognizing human activities was employed in [55] to distinguish the temporal relations of the input sequences.

Network ensemble: A two-stream CNN for action recognition in videos, which trained one stream of the CNN on the multi-frame dense optical flow for temporal information and the other stream of the CNN on still images for appearance features and then fused the outputs of two streams, was introduced by Simonyan et al. [209]. Inspired by this architecture, several network ensemble models have been proposed for FER.

Sun et al. [185] proposed a multi-channel network that extracted the spatial information from emotion-expressing faces and temporal information (optical flow) from the changes between emotioanl and neutral faces, and investigated three feature fusion strategies: score average fusion, SVM-based fusion and neural-network-based fusion. Zhang et al. [68] fused the temporal network PHRNN (discussed in **"Landmark trajectory"**) and the spatial network MSCNN (discussed in section 4.1.5) to extract the partial-whole, geometry-appearance, and static-dynamic information for FER (see Fig. 13). Instead of fusing the network outputs with different weights, Jung et al. [16] proposed a joint fine-tuning method that jointly trained the DTAN (discussed in the **"RNN and C3D"**), the DTGN (discussed in the **"Landmark**

trajectory") and the integrated network (see Fig. 14 for details), which outperformed the weighed sum strategy.

4.2.4 Discussion

In the real world, people display facial expressions in a dynamic process, e.g., from subtle to obvious, and it has become a trend to conduct FER on sequence/video data. Table 8 summarizes relative merits of different types of methods on dynamic data in regards to the capability of representing spatial and temporal information, the requirement on training data size and frame length (variable or fixed), the computational efficiency and the performance.

Frame aggregation is employed to combine the learned feature or prediction probability of each frame for a sequence-level result. The output of each frame can be simply concatenated (fixed-length frames is required in each sequence) or statistically aggregated to obtain video-level representation (variable-length frames processible). This method is computationally simple and can achieve moderate performance if the temporal variations of the target dataset is not complicated.

According to the fact that the expression intensity in a video sequence varies over time, the *expression intensity-invariant network* considers images with non-peak expressions and further exploits the dynamic correlations between peak and non-peak expressions to improve performance. Commonly, image frames with specific intensity states are needed for intensity-invariant FER.

Despite the advantages of these methods, *frame aggregation* handles frames without consideration of temporal information and subtle appearance changes, and *expression intensity-invariant networks* require prior knowledge of expression intensity which is unavailable in real-world scenarios. By contrast, *Deep spatio-temporal networks* are designed to encode temporal dependencies in consecutive frames and have been shown to benefit from learning spatial features in conjunction with temporal features. *RNN and its variations (e.g., LSTM, IRNN and BRNN) and C3D* are foundational networks for learning spatio-temporal features. However, the performance of these networks is barely satisfactory. RNN is incapable of capturing the powerful convolutional features. And 3D filers in C3D are applied over very short video clips ignoring long-range dynamics. Also, training such a huge network is computationally a problem, especially for dynamic FER where video data is insufficient. Alternatively, *facial landmark trajectory* methods extract shape features based on the physical structures of facial morphological variations to capture dynamic facial component activities, and then apply deep networks for classification. This method is computationally simple and can get rid of the issue on illumination variations. However, it is sensitive to registration errors and requires accurate facial landmark detection, which is difficult to access in unconstrained conditions. Consequently, this method performs less well and is more suitable to complement appearance representations. *Network ensemble* is utilized to train multiple networks for both spatial and temporal information and then to fuse the network outputs in the final stage. Optic flow and facial landmark trajectory can be used as temporal representations to collaborate spatial representations. One of the drawbacks of this framework is the pre-computing and storage consumption on optical flow or landmark trajectory vectors. And most related researches randomly selected fixed-length video frames as input, leading to the loss of useful temporal information. *Cascaded networks* were proposed to first extract discriminative representations for facial expression images and

TABLE 8
Comparison of different types of methods for dynamic image sequences in terms of data size requirement, representability of spatial and temporal information, requirement on frame length, performance, and computational efficiency. $\mathcal{FLT}$ = Facial Landmark Trajectory; $\mathcal{CN}$ = Cascaded Network; $\mathcal{NE}$ = Network Ensemble.

Network type		data	spatial	temporal	frame length	accuracy	efficiency
Frame aggregation		low	good	no	depends	fair	high
Expression intensity		fair	good	low	fixed	fair	varies
Spatio-temporal network	RNN	low	low	good	variable	low	fair
	C3D	high	good	fair	fixed	low	fair
	$\mathcal{FLT}$	fair	fair	fair	fixed	low	high
	$\mathcal{CN}$	high	good	good	variable	good	fair
	$\mathcal{NE}$	low	good	good	fixed	good	low

then input these features to sequential networks to reinforce the temporal information encoding. However, this model introduces additional parameters to capture sequence information, and the feature learning network (e.g., CNN) and the temporal information encoding network (e.g., LSTM) in current works are not trained jointly, which may lead to suboptimal parameter settings. And training in an end-to-end fashion is still a long road.

Compared with deep networks on static data, Table 4 and Table 7 demonstrate the powerful capability and popularity trend of deep spatio-temporal networks. For instance, comparison results on widely evaluated benchmarks (e.g., CK+ and MMI) illustrate that training networks based on sequence data and analyzing temporal dependency between frames can further improve the performance. Also, in the EmotiW challenge 2015, only one system employed deep spatio-networks for FER, whereas 5 of 7 reviewed systems in the EmotiW challenge 2017 relied on such networks.

5 ADDITIONAL RELATED ISSUES

In addition to the most popular basic expression classification task reviewed above, we further introduce a few related issues that depend on deep neural networks and prototypical expression-related knowledge.

5.1 Occlusion and non-frontal head pose

Occlusion and non-frontal head pose, which may change the visual appearance of the original facial expression, are two major obstacles for automatic FER, especially in real-world scenarios.

For *facial occlusion*, Ranzato et al. [210], [211] proposed a deep generative model that used mPoT [212] as the first layer of DBNs to model pixel-level representations and then trained DBNs to fit an appropriate distribution to its inputs. Thus, the occluded pixels in images could be filled in by reconstructing the top layer representation using the sequence of conditional distributions. Cheng et al. [213] employed multilayer RBMs with a pre-training and fine-tuning process on Gabor features to compress features from the occluded facial parts. Xu et al. [214] concatenated high-level learned features transferred from two CNNs with the same structure but pre-trained on different data: the original MSRA-CFW database and the MSRA-CFW database with additive occluded samples.

For *multi-view FER*, Zhang et al. [156] introduced a projection layer into the CNN that learned discriminative facial features by weighting different facial landmark points within 2D SIFT feature matrices without requiring facial pose estimation. Liu et al.

[215] proposed a multi-channel pose-aware CNN (MPCNN) that contains three cascaded parts (multi-channel feature extraction, jointly multi-scale feature fusion and the pose-aware recognition) to predict expression labels by minimizing the conditional joint loss of pose and expression recognition. Besides, the technology of generative adversarial network (GAN) has been employed in [180], [181] to generate facial images with different expressions under arbitrary poses for multi-view FER.

5.2 FER on infrared data

Although RBG or gray data are the current standard in deep FER, these data are vulnerable to ambient lighting conditions. While, infrared images that record the skin temporal distribution produced by emotions are not sensitive to illumination variations, which may be a promising alternative for investigation of facial expression. For example, He et al. [216] employed a DBM model that consists of a Gaussian-binary RBM and a binary RBM for FER. The model was trained by layerwise pre-training and joint training and was then fine-tuned on long-wavelength thermal infrared images to learn thermal features. Wu et al. [217] proposed a three-stream 3D CNN to fuse local and global spatio-temporal features on illumination-invariant near-infrared images for FER.

5.3 FER on 3D static and dynamic data

Despite significant advances have achieved in 2D FER, it fails to solve the two main problems: illumination changes and pose variations [29]. 3D FER that uses 3D face shape models with depth information can capture subtle facial deformations, which are naturally robust to pose and lighting variations.

Depth images and videos record the intensity of facial pixels based on distance from a depth camera, which contain critical information of facial geometric relations. For example, [218] used kinect depth sensor to obtain gradient direction information and then employed CNN on unregistered facial depth images for FER. [219], [220] extracted a series of salient features from depth videos and combined them with deep networks (i.e., CNN and DBN) for FER. To emphasize the dynamic deformation patterns of facial expression motions, Li et al. [221] explore the 4D FER (3D FER using dynamic data) using a dynamic geometrical image network. Furthermore, Chang et al. [222] proposed to estimate 3D expression coefficients from image intensities using CNN without requiring facial landmark detection. Thus, the model is highly robust to extreme appearance variations, including out-of-plane head rotations, scale changes, and occlusions.

Recently, more and more works trend to combine 2D and 3D data to further improve the performance. Oyedotun et al. [223] employed CNN to jointly learn facial expression features from both RGB and depth map latent modalities. And Li et al. [224] proposed a deep fusion CNN (DF-CNN) to explore multi-modal 2D+3D FER. Specifically, six types of 2D facial attribute maps (i.e., geometry, texture, curvature, normal components *x*, *y*, and *z*) were first extracted from the textured 3D face scans and were then jointly fed into the feature extraction and feature fusion subnets to learn the optimal combination weights of 2D and 3D facial representations. To improve this work, [225] proposed to extract deep features from different facial parts extracted from the texture and depth images, and then fused these features together to interconnect them with feedback. Wei et al. [226] further explored the data bias problem in 2D+3D FER using unsupervised domain adaption technique.

5.4 Facial expression synthesis

Realistic *facial expression synthesis*, which can generate various facial expressions for interactive interfaces, is a hot topic. Susskind et al. [227] demonstrated that DBN has the capacity to capture the large range of variation in expressive appearance and can be trained on large but sparsely labeled datasets. In light of this work, [210], [211], [228] employed DBN with unsupervised learning to construct facial expression synthesis systems. Kaneko et al. [149] proposed a multitask deep network with state recognition and key-point localization to adaptively generate visual feedback to improve facial expression recognition. With the recent success of the deep generative models, such as variational autoencoder (VAE), adversarial autoencoder (AAE), and generative adversarial network (GAN), a series of facial expression synthesis systems have been developed based on these models (e.g., [229], [230], [231], [232] and [233]). Facial expression synthesis can also be applied to data augmentation without manually collecting and labeling huge datasets. Masi et al. [234] employed CNN to synthesize new face images by increasing face-specific appearance variation, such as expressions within the 3D textured face model.

5.5 Visualization techniques

In addition to utilizing CNN for FER, several works (e.g., [139], [235], [236]) employed *visualization* techniques [237] on the learned CNN features to qualitatively analyze how the CNN contributes to the appearance-based learning process of FER and to qualitatively decipher which portions of the face yield the most discriminative information. The deconvolutional results all indicated that the activations of some particular filters on the learned features have strong correlations with the face regions that correspond to facial AUs.

5.6 Other special issues

Several novel issues have been approached on the basis of the prototypical expression categories: dominant and complementary emotion recognition challenge [238] and the Real versus Fake expressed emotions challenge [239]. Furthermore, deep learning techniques have been thoroughly applied by the participants of these two challenges (e.g., [240], [241], [242]). Additional related real-world applications, such as the Real-time FER App for smartphones [243], [244], Eyemotion (FER using eye-tracking cameras) [245], privacy-preserving mobile analytics [246], Unfelt emotions [247] and Depression recognition [248], have also been developed.

6 CHALLENGES AND OPPORTUNITIES

6.1 Facial expression datasets

As the FER literature shifts its main focus to the challenging in-the-wild environmental conditions, many researchers have committed to employing deep learning technologies to handle difficulties, such as illumination variation, occlusions, non-frontal head poses, identity bias and the recognition of low-intensity expressions. Given that FER is a data-driven task and that training a sufficiently deep network to capture subtle expression-related deformations requires a large amount of training data, the major challenge that deep FER systems face is the lack of training data in terms of both quantity and quality.

Because people of different age ranges, cultures and genders display and interpret facial expression in different ways, an ideal facial expression dataset is expected to include abundant sample images with precise face attribute labels, not just expression but other attributes such as age, gender and ethnicity, which would facilitate related research on cross-age range, cross-gender and cross-cultural FER using deep learning techniques, such as multitask deep networks and transfer learning. In addition, although occlusion and multipose problems have received relatively wide interest in the field of deep face recognition, the occlusion-robust and pose-invariant issues have receive less attention in deep FER. One of the main reasons is the lack of a large-scale facial expression dataset with occlusion type and head-pose annotations.

On the other hand, accurately annotating a large volume of image data with the large variation and complexity of natural scenarios is an obvious impediment to the construction of expression datasets. A reasonable approach is to employ crowd-sourcing models [44], [46], [249] under the guidance of expert annotators. Additionally, a fully automatic labeling tool [43] refined by experts is alternative to provide approximate but efficient annotations. In both cases, a subsequent reliable estimation or labeling learning process is necessary to filter out noisy annotations. In particular, few comparatively large-scale datasets that consider real-world scenarios and contain a wide range of facial expressions have recently become publicly available, i.e., EmotioNet [43], RAF-DB [44], [45] and AffectNet [46], and we anticipate that with advances in technology and the wide spread of the Internet, more complementary facial expression datasets will be constructed to promote the development of deep FER.

6.2 Incorporating other affective models

Another major issue that requires consideration is that while FER within the categorical model is widely acknowledged and researched, the definition of the prototypical expressions covers only a small portion of specific categories and cannot capture the full repertoire of expressive behaviors for realistic interactions. Two additional models were developed to describe a larger range of emotional landscape: the FACS model [10], [176], where various facial muscle AUs are combined to describe the visible appearance changes of facial expressions, and the dimensional model [11], [250], where two continuous-valued variables, namely, valence and arousal, are proposed to continuously encode small changes in the intensity of emotions. Another novel definition, i.e., compound expression, was proposed by Du et al. [52], who argued that some facial expressions are actually combinations of more than one basic emotion. These works improve the characterization of facial expressions and, to some extent, can complement the categorical model. For instance, as discussed above, the visualization results

of CNNs have demonstrated a certain congruity between the learned representations and the facial areas defined by AUs. Thus, we can design filters of the deep neural networks to distribute different weights according to the importance degree of different facial muscle action parts.

6.3 Dataset bias and imbalanced distribution

Data bias and inconsistent annotations are very common among different facial expression datasets due to different collecting conditions and the subjectiveness of annotating. Researchers commonly evaluate their algorithms within a specific dataset and can achieve satisfactory performance. However, early cross-database experiments have indicated that discrepancies between databases exist due to the different collection environments and construction indicators [12]; hence, algorithms evaluated via intra-database protocols lack generalizability on unseen test data, and the performance in cross-dataset settings is greatly deteriorated. Deep domain adaption and knowledge distillation are alternatives to address this bias [226], [251]. Furthermore, because of the inconsistent expression annotations, FER performance cannot keep improving when enlarging the training data by directly merging multiple datasets [167].

Another common problem in facial expression is class imbalance, which is a result of the practicalities of data acquisition: eliciting and annotating a smile is easy, however, capturing information for disgust, anger and other less common expressions can be very challenging. As shown in Table 4 and Table 7, the performance assessed in terms of mean accuracy, which assigns equal weights to all classes, decreases when compared with the accuracy criterion, and this decline is especially evident in real-world datasets (e.g., SFEW 2.0 and AFEW). One solution is to balance the class distribution during the pre-processing stage using data augmentation and synthesis. Another alternative is to develop a cost-sensitive loss layer for deep networks during training.

6.4 Multimodal affect recognition

Last but not the least, human expressive behaviors in realistic applications involve encoding from different perspectives, and the facial expression is only one modality. Although pure expression recognition based on visible face images can achieve promising results, incorporating with other models into a high-level framework can provide complementary information and further enhance the robustness. For example, participants in the EmotiW challenges and Audio Video Emotion Challenges (AVEC) [252], [253] considered the audio model to be the second most important element and employed various fusion techniques for multimodal affect recognition. Additionally, the fusion of other modalities, such as infrared images, depth information from 3D face models and physiological data, is becoming a promising research direction due to the large complementarity for facial expressions.

REFERENCES

[1] C. Darwin and P. Prodger, *The expression of the emotions in man and animals*. Oxford University Press, USA, 1998.

[2] Y.-I. Tian, T. Kanade, and J. F. Cohn, "Recognizing action units for facial expression analysis," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 23, no. 2, pp. 97–115, 2001.

[3] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion." *Journal of personality and social psychology*, vol. 17, no. 2, pp. 124–129, 1971.

[4] P. Ekman, "Strong evidence for universals in facial expressions: a reply to russell's mistaken critique," *Psychological bulletin*, vol. 115, no. 2, pp. 268–287, 1994.

[5] D. Matsumoto, "More evidence for the universality of a contempt expression," *Motivation and Emotion*, vol. 16, no. 4, pp. 363–368, 1992.

[6] R. E. Jack, O. G. Garrod, H. Yu, R. Caldara, and P. G. Schyns, "Facial expressions of emotion are not culturally universal," *Proceedings of the National Academy of Sciences*, vol. 109, no. 19, pp. 7241–7244, 2012.

[7] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 1, pp. 39–58, 2009.

[8] E. Sariyanidi, H. Gunes, and A. Cavallaro, "Automatic analysis of facial affect: A survey of registration, representation, and recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 6, pp. 1113–1133, 2015.

[9] B. Martinez and M. F. Valstar, "Advances, challenges, and opportunities in automatic facial expression recognition," in *Advances in Face Detection and Facial Image Analysis*. Springer, 2016, pp. 63–100.

[10] P. Ekman, "Facial action coding system (facs)," *A human face*, 2002.

[11] H. Gunes and B. Schuller, "Categorical and dimensional affect analysis in continuous input: Current trends and future directions," *Image and Vision Computing*, vol. 31, no. 2, pp. 120–136, 2013.

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

[13] P. Liu, S. Han, Z. Meng, and Y. Tong, "Facial expression recognition via a boosted deep belief network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1805–1812.

[14] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," in *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*. IEEE, 2016, pp. 1–10.

[15] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE transactions on pattern analysis and machine intelligence*, vol. 29, no. 6, pp. 915–928, 2007.

[16] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, "Joint fine-tuning in deep neural networks for facial expression recognition," in *Computer Vision (ICCV), 2015 IEEE International Conference on*. IEEE, 2015, pp. 2983–2991.

[17] X. Zhao, X. Liang, L. Liu, T. Li, Y. Han, N. Vasconcelos, and S. Yan, "Peak-piloted deep network for facial expression recognition," in *European conference on computer vision*. Springer, 2016, pp. 425–442.

[18] C. A. Corneanu, M. O. Simón, J. F. Cohn, and S. E. Guerrero, "Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 8, pp. 1548–1568, 2016.

[19] R. Zhi, M. Flierl, Q. Ruan, and W. B. Kleijn, "Graph-preserving sparse nonnegative matrix factorization with application to facial expression recognition," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 41, no. 1, pp. 38–52, 2011.

[20] L. Zhong, Q. Liu, P. Yang, B. Liu, J. Huang, and D. N. Metaxas, "Learning active facial patches for expression analysis," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2562–2569.

[21] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee *et al.*, "Challenges in representation learning: A report on three machine learning contests," in *International Conference on Neural Information Processing*. Springer, 2013, pp. 117–124.

[22] A. Dhall, O. Ramana Murthy, R. Goecke, J. Joshi, and T. Gedeon, "Video and image based emotion recognition challenges in the wild: Emotiw 2015," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, 2015, pp. 423–426.

[23] A. Dhall, R. Goecke, J. Joshi, J. Hoey, and T. Gedeon, "Emotiw 2016: Video and group-level emotion recognition challenges," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 2016, pp. 427–432.

[24] A. Dhall, R. Goecke, S. Ghosh, J. Joshi, J. Hoey, and T. Gedeon, "From individual to group-level emotion recognition: Emotiw 5.0," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. ACM, 2017, pp. 524–528.

[25] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[26] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[27] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[29] M. Pantic and L. J. M. Rothkrantz, "Automatic analysis of facial expressions: The state of the art," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 12, pp. 1424–1445, 2000.

[30] B. Fasel and J. Luettin, "Automatic facial expression analysis: a survey," *Pattern recognition*, vol. 36, no. 1, pp. 259–275, 2003.

[31] T. Zhang, "Facial expression recognition based on deep learning: A survey," in *International Conference on Intelligent and Interactive Systems and Applications*. Springer, 2017, pp. 345–352.

[32] M. F. Valstar, M. Mehu, B. Jiang, M. Pantic, and K. Scherer, "Meta-analysis of the first facial expression recognition challenge," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 42, no. 4, pp. 966–979, 2012.

[33] 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 *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*. IEEE, 2010, pp. 94–101.

[34] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," in *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*. IEEE, 2005, pp. 5–pp.

[35] M. Valstar and M. Pantic, "Induced disgust, happiness and surprise: an addition to the mmi facial expression database," in *Proc. 3rd Intern. Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect*, 2010, p. 65.

[36] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with gabor wavelets," in *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*. IEEE, 1998, pp. 200–205.

[37] J. M. Susskind, A. K. Anderson, and G. E. Hinton, "The toronto face database," *Department of Computer Science, University of Toronto, Toronto, ON, Canada, Tech. Rep*, vol. 3, 2010.

[38] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-pie," *Image and Vision Computing*, vol. 28, no. 5, pp. 807–813, 2010.

[39] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato, "A 3d facial expression database for facial behavior research," in *Automatic face and gesture recognition, 2006. FGR 2006. 7th international conference on*. IEEE, 2006, pp. 211–216.

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

[41] O. Langner, R. Dotsch, G. Bijlstra, D. H. Wigboldus, S. T. Hawk, and A. van Knippenberg, "Presentation and validation of the radboud faces database," *Cognition and Emotion*, vol. 24, no. 8, pp. 1377–1388, 2010.

[42] D. Lundqvist, A. Flykt, and A. Öhman, "The karolinska directed emotional faces (kdef)," *CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet*, no. 1998, 1998.

[43] C. F. Benitez-Quiroz, R. Srinivasan, and A. M. Martinez, "Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild," in *Proceedings of IEEE International Conference on Computer Vision & Pattern Recognition (CVPR), Las Vegas, NV, USA*, 2016.

[44] 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)*. IEEE, 2017, pp. 2584–2593.

[45] S. Li and W. Deng, "Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition," *IEEE Transactions on Image Processing*, 2018.

[46] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Transactions on Affective Computing*, vol. PP, no. 99, pp. 1–1, 2017.

[47] Z. Zhang, P. Luo, C. L. Chen, and X. Tang, "From facial expression recognition to interpersonal relation prediction," *International Journal of Computer Vision*, vol. 126, no. 5, pp. 1–20, 2018.

[48] A. Dhall, R. Goecke, S. Lucey, T. Gedeon *et al.*, "Collecting large, richly annotated facial-expression databases from movies," *IEEE multimedia*, vol. 19, no. 3, pp. 34–41, 2012.

[49] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Acted facial expressions in the wild database," *Australian National University, Canberra, Australia, Technical Report TR-CS-11*, vol. 2, p. 1, 2011.

[50] ——, "Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark," in *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*. IEEE, 2011, pp. 2106–2112.

[51] C. F. Benitez-Quiroz, R. Srinivasan, Q. Feng, Y. Wang, and A. M. Martinez, "Emotionet challenge: Recognition of facial expressions of emotion in the wild," *arXiv preprint arXiv:1703.01210*, 2017.

[52] S. Du, Y. Tao, and A. M. Martinez, "Compound facial expressions of emotion," *Proceedings of the National Academy of Sciences*, vol. 111, no. 15, pp. E1454–E1462, 2014.

[53] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 6, pp. 681–685, 2001.

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

[55] B. Hasani and M. H. Mahoor, "Spatio-temporal facial expression recognition using convolutional neural networks and conditional random fields," in *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*. IEEE, 2017, pp. 790–795.

[56] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2879–2886.

[57] S. E. Kahou, C. Pal, X. Bouthillier, P. Froumenty, Ç. Gülçehre, R. Memisevic, P. Vincent, A. Courville, Y. Bengio, R. C. Ferrari *et al.*, "Combining modality specific deep neural networks for emotion recognition in video," in *Proceedings of the 15th ACM on International conference on multimodal interaction*. ACM, 2013, pp. 543–550.

[58] T. Devries, K. Biswaranjan, and G. W. Taylor, "Multi-task learning of facial landmarks and expression," in *Computer and Robot Vision (CRV), 2014 Canadian Conference on*. IEEE, 2014, pp. 98–103.

[59] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, "Robust discriminative response map fitting with constrained local models," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. IEEE, 2013, pp. 3444–3451.

[60] M. Shin, M. Kim, and D.-S. Kwon, "Baseline cnn structure analysis for facial expression recognition," in *Robot and Human Interactive Communication (RO-MAN), 2016 25th IEEE International Symposium on*. IEEE, 2016, pp. 724–729.

[61] Z. Meng, P. Liu, J. Cai, S. Han, and Y. Tong, "Identity-aware convolutional neural network for facial expression recognition," in *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*. IEEE, 2017, pp. 558–565.

[62] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. IEEE, 2013, pp. 532–539.

[63] H.-W. Ng, V. D. Nguyen, V. Vonikakis, and S. Winkler, "Deep learning for emotion recognition on small datasets using transfer learning," in *Proceedings of the 2015 ACM on international conference on multimodal interaction*. ACM, 2015, pp. 443–449.

[64] S. Ren, X. Cao, Y. Wei, and J. Sun, "Face alignment at 3000 fps via regressing local binary features," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1685–1692.

[65] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, "Incremental face alignment in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1859–1866.

[66] D. H. Kim, W. Baddar, J. Jang, and Y. M. Ro, "Multi-objective based spatio-temporal feature representation learning robust to expression intensity variations for facial expression recognition," *IEEE Transactions on Affective Computing*, 2017.

[67] Y. Sun, X. Wang, and X. Tang, "Deep convolutional network cascade for facial point detection," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. IEEE, 2013, pp. 3476–3483.

[68] K. Zhang, Y. Huang, Y. Du, and L. Wang, "Facial expression recognition based on deep evolutional spatial-temporal networks," *IEEE Transactions on Image Processing*, vol. 26, no. 9, pp. 4193–4203, 2017.

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

[70] Z. Yu, Q. Liu, and G. Liu, "Deeper cascaded peak-piloted network for weak expression recognition," *The Visual Computer*, pp. 1–9, 2017.

[71] 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, 2018.

[72] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 1. IEEE, 2001, pp. I–I.

[73] F. De la Torre, W.-S. Chu, X. Xiong, F. Vicente, X. Ding, and J. F. Cohn, "Intraface," in *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 2015.

[74] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Facial landmark detection by deep multi-task learning," in *European Conference on Computer Vision*. Springer, 2014, pp. 94–108.

[75] Z. Yu and C. Zhang, "Image based static facial expression recognition with multiple deep network learning," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, 2015, pp. 435–442.

[76] B.-K. Kim, H. Lee, J. Roh, and S.-Y. Lee, "Hierarchical committee of deep cnns with exponentially-weighted decision fusion for static facial expression recognition," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, 2015, pp. 427–434.

[77] X. Liu, B. Kumar, J. You, and P. Jia, "Adaptive deep metric learning for identity-aware facial expression recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2017, pp. 522–531.

[78] G. Levi and T. Hassner, "Emotion recognition in the wild via convolutional neural networks and mapped binary patterns," in *Proceedings of the 2015 ACM on international conference on multimodal interaction*. ACM, 2015, pp. 503–510.

[79] D. A. Pitaloka, A. Wulandari, T. Basaruddin, and D. Y. Liliana, "Enhancing cnn with preprocessing stage in automatic emotion recognition," *Procedia Computer Science*, vol. 116, pp. 523–529, 2017.

[80] A. T. Lopes, E. de Aguiar, A. F. De Souza, and T. Oliveira-Santos, "Facial expression recognition with convolutional neural networks: coping with few data and the training sample order," *Pattern Recognition*, vol. 61, pp. 610–628, 2017.

[81] M. V. Zavarez, R. F. Berriel, and T. Oliveira-Santos, "Cross-database facial expression recognition based on fine-tuned deep convolutional network," in *Graphics, Patterns and Images (SIBGRAPI), 2017 30th SIBGRAPI Conference on*. IEEE, 2017, pp. 405–412.

[82] W. Li, M. Li, Z. Su, and Z. Zhu, "A deep-learning approach to facial expression recognition with candid images," in *Machine Vision Applications (MVA), 2015 14th IAPR International Conference on*. IEEE, 2015, pp. 279–282.

[83] I. Abbasnejad, S. Sridharan, D. Nguyen, S. Denman, C. Fookes, and S. Lucey, "Using synthetic data to improve facial expression analysis with 3d convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1609–1618.

[84] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[85] W. Chen, M. J. Er, and S. Wu, "Illumination compensation and normalization for robust face recognition using discrete cosine transform in logarithm domain," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 36, no. 2, pp. 458–466, 2006.

[86] J. Li and E. Y. Lam, "Facial expression recognition using deep neural networks," in *Imaging Systems and Techniques (IST), 2015 IEEE International Conference on*. IEEE, 2015, pp. 1–6.

[87] S. Ebrahimi Kahou, V. Michalski, K. Konda, R. Memisevic, and C. Pal, "Recurrent neural networks for emotion recognition in video," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, 2015, pp. 467–474.

[88] S. A. Bargal, E. Barsoum, C. C. Ferrer, and C. Zhang, "Emotion recognition in the wild from videos using images," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 2016, pp. 433–436.

[89] C.-M. Kuo, S.-H. Lai, and M. Sarkis, "A compact deep learning model for robust facial expression recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 2121–2129.

[90] A. Yao, D. Cai, P. Hu, S. Wang, L. Sha, and Y. Chen, "Holonet: towards robust emotion recognition in the wild," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 2016, pp. 472–478.

[91] P. Hu, D. Cai, S. Wang, A. Yao, and Y. Chen, "Learning supervised scoring ensemble for emotion recognition in the wild," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. ACM, 2017, pp. 553–560.

[92] T. Hassner, S. Harel, E. Paz, and R. Enbar, "Effective face frontalization in unconstrained images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4295–4304.

[93] C. Sagonas, Y. Panagakis, S. Zafeiriou, and M. Pantic, "Robust statistical face frontalization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3871–3879.

[94] X. Yin, X. Yu, K. Sohn, X. Liu, and M. Chandraker, "Towards large-pose face frontalization in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3990–3999.

[95] R. Huang, S. Zhang, T. Li, and R. He, "Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2439–2448.

[96] L. Tran, X. Yin, and X. Liu, "Disentangled representation learning gan for pose-invariant face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1415–1424.

[97] L. Deng, D. Yu *et al.*, "Deep learning: methods and applications," *Foundations and Trends® in Signal Processing*, vol. 7, no. 3–4, pp. 197–387, 2014.

[98] B. Fasel, "Robust face analysis using convolutional neural networks," in *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, vol. 2. IEEE, 2002, pp. 40–43.

[99] ——, "Head-pose invariant facial expression recognition using convolutional neural networks," in *Proceedings of the 4th IEEE International Conference on Multimodal Interfaces*. IEEE Computer Society, 2002, p. 529.

[100] M. Matsugu, K. Mori, Y. Mitari, and Y. Kaneda, "Subject independent facial expression recognition with robust face detection using a convolutional neural network," *Neural Networks*, vol. 16, no. 5-6, pp. 555–559, 2003.

[101] B. Sun, L. Li, G. Zhou, X. Wu, J. He, L. Yu, D. Li, and Q. Wei, "Combining multimodal features within a fusion network for emotion recognition in the wild," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, 2015, pp. 497–502.

[102] B. Sun, L. Li, G. Zhou, and J. He, "Facial expression recognition in the wild based on multimodal texture features," *Journal of Electronic Imaging*, vol. 25, no. 6, p. 061407, 2016.

[103] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[104] J. Li, D. Zhang, J. Zhang, J. Zhang, T. Li, Y. Xia, Q. Yan, and L. Xun, "Facial expression recognition with faster r-cnn," *Procedia Computer Science*, vol. 107, pp. 135–140, 2017.

[105] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[106] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 221–231, 2013.

[107] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Computer Vision (ICCV), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4489–4497.

[108] 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*. ACM, 2016, pp. 445–450.

[109] D. Nguyen, K. Nguyen, S. Sridharan, A. Ghasemi, D. Dean, and C. Fookes, "Deep spatio-temporal features for multimodal emotion recognition," in *Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on*. IEEE, 2017, pp. 1215–1223.

[110] S. Ouellet, "Real-time emotion recognition for gaming using deep convolutional network features," *arXiv preprint arXiv:1408.3750*, 2014.

[111] H. Ding, S. K. Zhou, and R. Chellappa, "Facenet2expnet: Regularizing a deep face recognition net for expression recognition," in *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*. IEEE, 2017, pp. 118–126.

[112] B. Hasani and M. H. Mahoor, "Facial expression recognition using enhanced deep 3d convolutional neural networks," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*. IEEE, 2017, pp. 2278–2288.

[113] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[114] G. E. Hinton and T. J. Sejnowski, "Learning and releaming in boltzmann machines," *Parallel distributed processing: Explorations in the microstructure of cognition*, vol. 1, no. 282-317, p. 2, 1986.

[115] G. E. Hinton, "A practical guide to training restricted boltzmann machines," in *Neural networks: Tricks of the trade*. Springer, 2012, pp. 599–619.

[116] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," in *Advances in neural information processing systems*, 2007, pp. 153–160.

[117] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural computation*, vol. 14, no. 8, pp. 1771–1800, 2002.

[118] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *science*, vol. 313, no. 5786, pp. 504–507, 2006.

[119] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *Journal of Machine Learning Research*, vol. 11, no. Dec, pp. 3371–3408, 2010.

[120] Q. V. Le, "Building high-level features using large scale unsupervised learning," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 8595–8598.

[121] S. Rifai, P. Vincent, X. Muller, X. Glorot, and Y. Bengio, "Contractive auto-encoders: Explicit invariance during feature extraction," in *Proceedings of the 28th International Conference on International Conference on Machine Learning*. Omnipress, 2011, pp. 833–840.

[122] J. Masci, U. Meier, D. Cireşan, and J. Schmidhuber, "Stacked convolutional auto-encoders for hierarchical feature extraction," in *International Conference on Artificial Neural Networks*. Springer, 2011, pp. 52–59.

[123] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[124] P. J. Werbos, "Backpropagation through time: what it does and how to do it," *Proceedings of the IEEE*, vol. 78, no. 10, pp. 1550–1560, 1990.

[125] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[126] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[127] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[128] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther, "Autoencoding beyond pixels using a learned similarity metric," *arXiv preprint arXiv:1512.09300*, 2015.

[129] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "Infogan: Interpretable representation learning by information maximizing generative adversarial nets," in *Advances in neural information processing systems*, 2016, pp. 2172–2180.

[130] Y. Tang, "Deep learning using linear support vector machines," *arXiv preprint arXiv:1306.0239*, 2013.

[131] A. Dapogny and K. Bailly, "Investigating deep neural forests for facial expression recognition," in *Automatic Face & Gesture Recognition (FG 2018), 2018 13th IEEE International Conference on*. IEEE, 2018, pp. 629–633.

[132] P. Kontschieder, M. Fiterau, A. Criminisi, and S. Rota Bulo, "Deep neural decision forests," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1467–1475.

[133] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "Decaf: A deep convolutional activation feature for generic visual recognition," in *International conference on machine learning*, 2014, pp. 647–655.

[134] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "Cnn features off-the-shelf: an astounding baseline for recognition," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2014 IEEE Conference on*. IEEE, 2014, pp. 512–519.

[135] N. Otberdout, A. Kacem, M. Daoudi, L. Ballihi, and S. Berretti, "Deep covariance descriptors for facial expression recognition," in *BMVC*, 2018.

[136] D. Acharya, Z. Huang, D. Pani Paudel, and L. Van Gool, "Covariance pooling for facial expression recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 367–374.

[137] M. Liu, S. Li, S. Shan, and X. Chen, "Au-aware deep networks for facial expression recognition," in *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*. IEEE, 2013, pp. 1–6.

[138] ——, "Au-inspired deep networks for facial expression feature learning," *Neurocomputing*, vol. 159, pp. 126–136, 2015.

[139] P. Khorrami, T. Paine, and T. Huang, "Do deep neural networks learn facial action units when doing expression recognition?" *arXiv preprint arXiv:1510.02969v3*, 2015.

[140] J. Cai, Z. Meng, A. S. Khan, Z. Li, J. OReilly, and Y. Tong, "Island loss for learning discriminative features in facial expression recognition," in *Automatic Face & Gesture Recognition (FG 2018), 2018 13th IEEE International Conference on*. IEEE, 2018, pp. 302–309.

[141] H. Yang, U. Ciftci, and L. Yin, "Facial expression recognition by de-expression residue learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2168–2177.

[142] D. Hamester, P. Barros, and S. Wermter, "Face expression recognition with a 2-channel convolutional neural network," in *Neural Networks (IJCNN), 2015 International Joint Conference on*. IEEE, 2015, pp. 1–8.

[143] S. Reed, K. Sohn, Y. Zhang, and H. Lee, "Learning to disentangle factors of variation with manifold interaction," in *International Conference on Machine Learning*, 2014, pp. 1431–1439.

[144] Z. Zhang, P. Luo, C.-C. Loy, and X. Tang, "Learning social relation traits from face images," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3631–3639.

[145] Y. Guo, D. Tao, J. Yu, H. Xiong, Y. Li, and D. Tao, "Deep neural networks with relativity learning for facial expression recognition," in *Multimedia & Expo Workshops (ICMEW), 2016 IEEE International Conference on*. IEEE, 2016, pp. 1–6.

[146] B.-K. Kim, S.-Y. Dong, J. Roh, G. Kim, and S.-Y. Lee, "Fusing aligned and non-aligned face information for automatic affect recognition in the wild: A deep learning approach," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016, pp. 48–57.

[147] C. Pramerdorfer and M. Kampel, "Facial expression recognition using convolutional neural networks: State of the art," *arXiv preprint arXiv:1612.02903*, 2016.

[148] O. M. Parkhi, A. Vedaldi, A. Zisserman *et al.*, "Deep face recognition." in *BMVC*, vol. 1, no. 3, 2015, p. 6.

[149] T. Kaneko, K. Hiramatsu, and K. Kashino, "Adaptive visual feedback generation for facial expression improvement with multi-task deep neural networks," in *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, 2016, pp. 327–331.

[150] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," *arXiv preprint arXiv:1411.7923*, 2014.

[151] X. Zhang, L. Zhang, X.-J. Wang, and H.-Y. Shum, "Finding celebrities in billions of web images," *IEEE Transactions on Multimedia*, vol. 14, no. 4, pp. 995–1007, 2012.

[152] H.-W. Ng and S. Winkler, "A data-driven approach to cleaning large face datasets," in *Image Processing (ICIP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 343–347.

[153] H. Kaya, F. Gürpınar, and A. A. Salah, "Video-based emotion recognition in the wild using deep transfer learning and score fusion," *Image and Vision Computing*, vol. 65, pp. 66–75, 2017.

[154] B. Knyazev, R. Shvetsov, N. Efremova, and A. Kuharenko, "Convolutional neural networks pretrained on large face recognition datasets for emotion classification from video," *arXiv preprint arXiv:1711.04598*, 2017.

[155] D. G. Lowe, "Object recognition from local scale-invariant features," in *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, vol. 2. Ieee, 1999, pp. 1150–1157.

[156] T. Zhang, W. Zheng, Z. Cui, Y. Zong, J. Yan, and K. Yan, "A deep neural network-driven feature learning method for multi-view facial expression recognition," *IEEE Transactions on Multimedia*, vol. 18, no. 12, pp. 2528–2536, 2016.

[157] Z. Luo, J. Chen, T. Takiguchi, and Y. Ariki, "Facial expression recognition with deep age," in *Multimedia & Expo Workshops (ICMEW), 2017 IEEE International Conference on*. IEEE, 2017, pp. 657–662.

[158] L. Chen, M. Zhou, W. Su, M. Wu, J. She, and K. Hirota, "Softmax regression based deep sparse autoencoder network for facial emotion recognition in human-robot interaction," *Information Sciences*, vol. 428, pp. 49–61, 2018.

[159] V. Mavani, S. Raman, and K. P. Miyapuram, "Facial expression recognition using visual saliency and deep learning," *arXiv preprint arXiv:1708.08016*, 2017.

[160] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, "A deep multi-level network for saliency prediction," in *Pattern Recognition (ICPR), 2016 23rd International Conference on*. IEEE, 2016, pp. 3488–3493.

[161] B.-F. Wu and C.-H. Lin, "Adaptive feature mapping for customizing deep learning based facial expression recognition model," *IEEE Access*, 2018.

[162] J. Lu, V. E. Liong, and J. Zhou, "Cost-sensitive local binary feature learning for facial age estimation," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5356–5368, 2015.

[163] W. Shang, K. Sohn, D. Almeida, and H. Lee, "Understanding and improving convolutional neural networks via concatenated rectified linear units," in *International Conference on Machine Learning*, 2016, pp. 2217–2225.

[164] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2818–2826.

[165] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning." in *AAAI*, vol. 4, 2017, p. 12.

[166] S. Zhao, H. Cai, H. Liu, J. Zhang, and S. Chen, "Feature selection mechanism in cnns for facial expression recognition," in *BMVC*, 2018.

[167] J. Zeng, S. Shan, and X. Chen, "Facial expression recognition with inconsistently annotated datasets," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 222–237.

[168] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *European Conference on Computer Vision*. Springer, 2016, pp. 499–515.

[169] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[170] G. Zeng, J. Zhou, X. Jia, W. Xie, and L. Shen, "Hand-crafted feature guided deep learning for facial expression recognition," in *Automatic Face & Gesture Recognition (FG 2018), 2018 13th IEEE International Conference on*. IEEE, 2018, pp. 423–430.

[171] D. Ciregan, U. Meier, and J. Schmidhuber, "Multi-column deep neural networks for image classification," in *Computer vision and pattern recognition (CVPR), 2012 IEEE conference on*. IEEE, 2012, pp. 3642–3649.

[172] G. Pons and D. Masip, "Supervised committee of convolutional neural networks in automated facial expression analysis," *IEEE Transactions on Affective Computing*, 2017.

[173] B.-K. Kim, J. Roh, S.-Y. Dong, and S.-Y. Lee, "Hierarchical committee of deep convolutional neural networks for robust facial expression recognition," *Journal on Multimodal User Interfaces*, vol. 10, no. 2, pp. 173–189, 2016.

[174] K. Liu, M. Zhang, and Z. Pan, "Facial expression recognition with cnn ensemble," in *Cyberworlds (CW), 2016 International Conference on*. IEEE, 2016, pp. 163–166.

[175] G. Pons and D. Masip, "Multi-task, multi-label and multi-domain learning with residual convolutional networks for emotion recognition," *arXiv preprint arXiv:1802.06664*, 2018.

[176] P. Ekman and E. L. Rosenberg, *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA, 1997.

[177] R. Ranjan, S. Sankaranarayanan, C. D. Castillo, and R. Chellappa, "An all-in-one convolutional neural network for face analysis," in *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*. IEEE, 2017, pp. 17–24.

[178] Y. Lv, Z. Feng, and C. Xu, "Facial expression recognition via deep learning," in *Smart Computing (SMARTCOMP), 2014 International Conference on*. IEEE, 2014, pp. 303–308.

[179] S. Rifai, Y. Bengio, A. Courville, P. Vincent, and M. Mirza, "Disentangling factors of variation for facial expression recognition," in *European Conference on Computer Vision*. Springer, 2012, pp. 808–822.

[180] Y.-H. Lai and S.-H. Lai, "Emotion-preserving representation learning via generative adversarial network for multi-view facial expression recognition," in *Automatic Face & Gesture Recognition (FG 2018), 2018 13th IEEE International Conference on*. IEEE, 2018, pp. 263–270.

[181] F. Zhang, T. Zhang, Q. Mao, and C. Xu, "Joint pose and expression modeling for facial expression recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3359–3368.

[182] H. Yang, Z. Zhang, and L. Yin, "Identity-adaptive facial expression recognition through expression regeneration using conditional generative adversarial networks," in *Automatic Face & Gesture Recognition (FG 2018), 2018 13th IEEE International Conference on*. IEEE, 2018, pp. 294–301.

[183] J. Chen, J. Konrad, and P. Ishwar, "Vgan-based image representation learning for privacy-preserving facial expression recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 1570–1579.

[184] Y. Kim, B. Yoo, Y. Kwak, C. Choi, and J. Kim, "Deep generative-contrastive networks for facial expression recognition," *arXiv preprint arXiv:1703.07140*, 2017.

[185] N. Sun, Q. Li, R. Huan, J. Liu, and G. Han, "Deep spatial-temporal feature fusion for facial expression recognition in static images," *Pattern Recognition Letters*, 2017.

[186] W. Ding, M. Xu, D. Huang, W. Lin, M. Dong, X. Yu, and H. Li, "Audio and face video emotion recognition in the wild using deep neural networks and small datasets," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 2016, pp. 506–513.

[187] J. Yan, W. Zheng, Z. Cui, C. Tang, T. Zhang, Y. Zong, and N. Sun, "Multi-clue fusion for emotion recognition in the wild," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 2016, pp. 458–463.

[188] Z. Cui, S. Xiao, Z. Niu, S. Yan, and W. Zheng, "Recurrent shape regression," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[189] X. Ouyang, S. Kawaai, E. G. H. Goh, S. Shen, W. Ding, H. Ming, and D.-Y. Huang, "Audio-visual emotion recognition using deep transfer learning and multiple temporal models," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. ACM, 2017, pp. 577–582.

[190] V. Vielzeuf, S. Pateux, and F. Jurie, "Temporal multimodal fusion for video emotion classification in the wild," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. ACM, 2017, pp. 569–576.

[191] S. E. Kahou, X. Bouthillier, P. Lamblin, C. Gulcehre, V. Michalski, K. Konda, S. Jean, P. Froumenty, Y. Dauphin, N. Boulanger-Lewandowski *et al.*, "Emonets: Multimodal deep learning approaches for emotion recognition in video," *Journal on Multimodal User Interfaces*, vol. 10, no. 2, pp. 99–111, 2016.

[192] M. Liu, R. Wang, S. Li, S. Shan, Z. Huang, and X. Chen, "Combining multiple kernel methods on riemannian manifold for emotion recognition in the wild," in *Proceedings of the 16th International Conference on Multimodal Interaction*. ACM, 2014, pp. 494–501.

[193] B. Xu, Y. Fu, Y.-G. Jiang, B. Li, and L. Sigal, "Video emotion recognition with transferred deep feature encodings," in *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*. ACM, 2016, pp. 15–22.

[194] J. Chen, R. Xu, and L. Liu, "Deep peak-neutral difference feature for facial expression recognition," *Multimedia Tools and Applications*, pp. 1–17, 2018.

[195] Q. V. Le, N. Jaitly, and G. E. Hinton, "A simple way to initialize recurrent networks of rectified linear units," *arXiv preprint arXiv:1504.00941*, 2015.

[196] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[197] P. Barros and S. Wermter, "Developing crossmodal expression recognition based on a deep neural model," *Adaptive behavior*, vol. 24, no. 5, pp. 373–396, 2016.

[198] J. Zhao, X. Mao, and J. Zhang, "Learning deep facial expression features from image and optical flow sequences using 3d cnn," *The Visual Computer*, pp. 1–15, 2018.

[199] M. Liu, S. Li, S. Shan, R. Wang, and X. Chen, "Deeply learning deformable facial action parts model for dynamic expression analysis," in *Asian conference on computer vision*. Springer, 2014, pp. 143–157.

[200] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.

[201] S. Pini, O. B. Ahmed, M. Cornia, L. Baraldi, R. Cucchiara, and B. Huet, "Modeling multimodal cues in a deep learning-based framework for emotion recognition in the wild," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. ACM, 2017, pp. 536–543.

[202] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," in *Pro-*

ceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 5297–5307.

[203] D. H. Kim, M. K. Lee, D. Y. Choi, and B. C. Song, "Multi-modal emotion recognition using semi-supervised learning and multiple neural networks in the wild," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. ACM, 2017, pp. 529–535.

[204] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2625–2634.

[205] D. K. Jain, Z. Zhang, and K. Huang, "Multi angle optimal pattern-based deep learning for automatic facial expression recognition," *Pattern Recognition Letters*, 2017.

[206] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt, "Spatiotemporal convolutional sparse auto-encoder for sequence classification." in *BMVC*, 2012, pp. 1–12.

[207] S. Kankanamge, C. Fookes, and S. Sridharan, "Facial analysis in the wild with lstm networks," in *Image Processing (ICIP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 1052–1056.

[208] J. D. Lafferty, A. Mccallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," *Proceedings of Icml*, vol. 3, no. 2, pp. 282–289, 2001.

[209] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in neural information processing systems*, 2014, pp. 568–576.

[210] J. Susskind, V. Mnih, G. Hinton *et al.*, "On deep generative models with applications to recognition," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 2857–2864.

[211] V. Mnih, J. M. Susskind, G. E. Hinton *et al.*, "Modeling natural images using gated mrfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 9, pp. 2206–2222, 2013.

[212] V. Mnih, G. E. Hinton *et al.*, "Generating more realistic images using gated mrf's," in *Advances in Neural Information Processing Systems*, 2010, pp. 2002–2010.

[213] Y. Cheng, B. Jiang, and K. Jia, "A deep structure for facial expression recognition under partial occlusion," in *Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP), 2014 Tenth International Conference on*. IEEE, 2014, pp. 211–214.

[214] M. Xu, W. Cheng, Q. Zhao, L. Ma, and F. Xu, "Facial expression recognition based on transfer learning from deep convolutional networks," in *Natural Computation (ICNC), 2015 11th International Conference on*. IEEE, 2015, pp. 702–708.

[215] Y. Liu, J. Zeng, S. Shan, and Z. Zheng, "Multi-channel pose-aware convolution neural networks for multi-view facial expression recognition," in *Automatic Face & Gesture Recognition (FG 2018), 2018 13th IEEE International Conference on*. IEEE, 2018, pp. 458–465.

[216] S. He, S. Wang, W. Lan, H. Fu, and Q. Ji, "Facial expression recognition using deep boltzmann machine from thermal infrared images," in *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*. IEEE, 2013, pp. 239–244.

[217] Z. Wu, T. Chen, Y. Chen, Z. Zhang, and G. Liu, "Nirexpnet: Three-stream 3d convolutional neural network for near infrared facial expression recognition," *Applied Sciences*, vol. 7, no. 11, p. 1184, 2017.

[218] E. P. Ijjina and C. K. Mohan, "Facial expression recognition using kinect depth sensor and convolutional neural networks," in *Machine Learning and Applications (ICMLA), 2014 13th International Conference on*. IEEE, 2014, pp. 392–396.

[219] M. Z. Uddin, M. M. Hassan, A. Almogren, M. Zuair, G. Fortino, and J. Torresen, "A facial expression recognition system using robust face features from depth videos and deep learning," *Computers & Electrical Engineering*, vol. 63, pp. 114–125, 2017.

[220] M. Z. Uddin, W. Khaksar, and J. Torresen, "Facial expression recognition using salient features and convolutional neural network," *IEEE Access*, vol. 5, pp. 26 146–26 161, 2017.

[221] W. Li, D. Huang, H. Li, and Y. Wang, "Automatic 4d facial expression recognition using dynamic geometrical image network," in *Automatic Face & Gesture Recognition (FG 2018), 2018 13th IEEE International Conference on*. IEEE, 2018, pp. 24–30.

[222] F.-J. Chang, A. T. Tran, T. Hassner, I. Masi, R. Nevatia, and G. Medioni, "Expnet: Landmark-free, deep, 3d facial expressions," in *Automatic Face & Gesture Recognition (FG 2018), 2018 13th IEEE International Conference on*. IEEE, 2018, pp. 122–129.

[223] O. K. Oyedotun, G. Demisse, A. E. R. Shabayek, D. Aouada, and B. Ottersten, "Facial expression recognition via joint deep learning

of rgb-depth map latent representations," in *2017 IEEE International Conference on Computer Vision Workshop (ICCVW)*, 2017.

[224] H. Li, J. Sun, Z. Xu, and L. Chen, "Multimodal 2d+ 3d facial expression recognition with deep fusion convolutional neural network," *IEEE Transactions on Multimedia*, vol. 19, no. 12, pp. 2816–2831, 2017.

[225] A. Jan, H. Ding, H. Meng, L. Chen, and H. Li, "Accurate facial parts localization and deep learning for 3d facial expression recognition," in *Automatic Face & Gesture Recognition (FG 2018), 2018 13th IEEE International Conference on*. IEEE, 2018, pp. 466–472.

[226] X. Wei, H. Li, J. Sun, and L. Chen, "Unsupervised domain adaptation with regularized optimal transport for multimodal 2d+ 3d facial expression recognition," in *Automatic Face & Gesture Recognition (FG 2018), 2018 13th IEEE International Conference on*. IEEE, 2018, pp. 31–37.

[227] J. M. Susskind, G. E. Hinton, J. R. Movellan, and A. K. Anderson, "Generating facial expressions with deep belief nets," in *Affective Computing*. InTech, 2008.

[228] M. Sabzevari, S. Toosizadeh, S. R. Quchani, and V. Abrishami, "A fast and accurate facial expression synthesis system for color face images using face graph and deep belief network," in *Electronics and Information Engineering (ICEIE), 2010 International Conference On*, vol. 2. IEEE, 2010, pp. V2–354.

[229] R. Yeh, Z. Liu, D. B. Goldman, and A. Agarwala, "Semantic facial expression editing using autoencoded flow," *arXiv preprint arXiv:1611.09961*, 2016.

[230] Y. Zhou and B. E. Shi, "Photorealistic facial expression synthesis by the conditional difference adversarial autoencoder," in *Affective Computing and Intelligent Interaction (ACII), 2017 Seventh International Conference on*. IEEE, 2017, pp. 370–376.

[231] L. Song, Z. Lu, R. He, Z. Sun, and T. Tan, "Geometry guided adversarial facial expression synthesis," *arXiv preprint arXiv:1712.03474*, 2017.

[232] H. Ding, K. Sricharan, and R. Chellappa, "Exprgan: Facial expression editing with controllable expression intensity," in *AAAI*, 2018, p. 67816788.

[233] F. Qiao, N. Yao, Z. Jiao, Z. Li, H. Chen, and H. Wang, "Geometry-contrastive generative adversarial network for facial expression synthesis," *arXiv preprint arXiv:1802.01822*, 2018.

[234] I. Masi, A. T. Tran, T. Hassner, J. T. Leksut, and G. Medioni, "Do we really need to collect millions of faces for effective face recognition?" in *European Conference on Computer Vision*. Springer, 2016, pp. 579–596.

[235] N. Mousavi, H. Siqueira, P. Barros, B. Fernandes, and S. Wermter, "Understanding how deep neural networks learn face expressions," in *Neural Networks (IJCNN), 2016 International Joint Conference on*. IEEE, 2016, pp. 227–234.

[236] R. Breuer and R. Kimmel, "A deep learning perspective on the origin of facial expressions," *arXiv preprint arXiv:1705.01842*, 2017.

[237] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*. Springer, 2014, pp. 818–833.

[238] I. Lüsi, J. C. J. Junior, J. Gorbova, X. Baró, S. Escalera, H. Demirel, J. Allik, C. Ozcinar, and G. Anbarjafari, "Joint challenge on dominant and complementary emotion recognition using micro emotion features and head-pose estimation: Databases," in *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*. IEEE, 2017, pp. 809–813.

[239] J. Wan, S. Escalera, X. Baro, H. J. Escalante, I. Guyon, M. Madadi, J. Allik, J. Gorbova, and G. Anbarjafari, "Results and analysis of chalearn lap multi-modal isolated and continuous gesture recognition, and real versus fake expressed emotions challenges," in *ChaLearn LaP, Action, Gesture, and Emotion Recognition Workshop and Competitions: Large Scale Multimodal Gesture Recognition and Real versus Fake expressed emotions, ICCV*, vol. 4, no. 6, 2017.

[240] Y.-G. Kim and X.-P. Huynh, "Discrimination between genuine versus fake emotion using long-short term memory with parametric bias and facial landmarks," in *Computer Vision Workshop (ICCVW), 2017 IEEE International Conference on*. IEEE, 2017, pp. 3065–3072.

[241] L. Li, T. Baltrusaitis, B. Sun, and L.-P. Morency, "Combining sequential geometry and texture features for distinguishing genuine and deceptive emotions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3147–3153.

[242] J. Guo, S. Zhou, J. Wu, J. Wan, X. Zhu, Z. Lei, and S. Z. Li, "Multi-modality network with visual and geometrical information for micro emotion recognition," in *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*. IEEE, 2017, pp. 814–819.

[243] I. Song, H.-J. Kim, and P. B. Jeon, "Deep learning for real-time robust facial expression recognition on a smartphone," in *Consumer

Electronics (ICCE), 2014 IEEE International Conference on.   IEEE, 2014, pp. 564–567.

[244] S. Bazrafkan, T. Nedelcu, P. Filipczuk, and P. Corcoran, "Deep learning for facial expression recognition: A step closer to a smartphone that knows your moods," in *Consumer Electronics (ICCE), 2017 IEEE International Conference on*.   IEEE, 2017, pp. 217–220.

[245] S. Hickson, N. Dufour, A. Sud, V. Kwatra, and I. Essa, "Eyemotion: Classifying facial expressions in vr using eye-tracking cameras," *arXiv preprint arXiv:1707.07204*, 2017.

[246] S. A. Ossia, A. S. Shamsabadi, A. Taheri, H. R. Rabiee, N. Lane, and H. Haddadi, "A hybrid deep learning architecture for privacy-preserving mobile analytics," *arXiv preprint arXiv:1703.02952*, 2017.

[247] K. Kulkarni, C. A. Corneanu, I. Ofodile, S. Escalera, X. Baro, S. Hyniewska, J. Allik, and G. Anbarjafari, "Automatic recognition of facial displays of unfelt emotions," *arXiv preprint arXiv:1707.04061*, 2017.

[248] X. Zhou, K. Jin, Y. Shang, and G. Guo, "Visually interpretable representation learning for depression recognition from facial images," *IEEE Transactions on Affective Computing*, pp. 1–1, 2018.

[249] E. Barsoum, C. Zhang, C. C. Ferrer, and Z. Zhang, "Training deep networks for facial expression recognition with crowd-sourced label distribution," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*.   ACM, 2016, pp. 279–283.

[250] J. A. Russell, "A circumplex model of affect." *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[251] S. Li and W. Deng, "Deep emotion transfer network for cross-database facial expression recognition," in *Pattern Recognition (ICPR), 2018 26th International Conference*.   IEEE, 2018, pp. 3092–3099.

[252] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, "Avec 2016: Depression, mood, and emotion recognition workshop and challenge," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*.   ACM, 2016, pp. 3–10.

[253] F. Ringeval, B. Schuller, M. Valstar, J. Gratch, R. Cowie, S. Scherer, S. Mozgai, N. Cummins, M. Schmitt, and M. Pantic, "Avec 2017: Real-life depression, and affect recognition workshop and challenge," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*.   ACM, 2017, pp. 3–9.