A Survey on Facial Expression Recognition of Static and Dynamic Emotions

Yan Wang, Shaoqi Yan, Yang Liu, Wei Song, Jing Liu, Yang Chang, Xinji Mai, Xiping Hu, Wenqiang Zhang* and Zhongxue Gan*

Abstract—Facial expression recognition (FER) aims to analyze emotional states from static images and dynamic sequences, which is pivotal in enhancing anthropomorphic communication among humans, robots, and digital avatars by leveraging AI technologies. As the FER field evolves from controlled laboratory environments to more complex in-the-wild scenarios, advanced methods have been rapidly developed and new challenges and apporaches are encounted, which are not well addressed in existing reviews of FER. This paper offers a comprehensive survey of both image-based static FER (SFER) and video-based dynamic FER (DFER) methods, analyzing from model-oriented development to challenge-focused categorization. We begin with a critical comparison of recent reviews, an introduction to common datasets and evaluation criteria, and an in-depth workflow on FER to establish a robust research foundation. We then systematically review representative approaches addressing eight main challenges in SFER (such as expression disturbance, uncertainties, compound emotions, and cross-domain inconsistency) as well as seven main challenges in DFER (such as key frame sampling, expression intensity variations, and cross-modal alignment). Additionally, we analyze recent advancements, benchmark performances, major applications, and ethical considerations. Finally, we propose five promising future directions and development trends to guide ongoing research. The project page for this paper can be found at https://github.com/wangyanckxx/SurveyFER.

Index Terms—Affective Computing, Facial Expression Recognition, Static and Dynamic Emotions, Challenges and Advances.

1 Introduction

AFFECVIVE computing [1] has far-reaching influence and importance in key national fields. Innovate UK, the UK's innovation agency, identified "artificial intelligence (AI) emotion and expression recognition" as the top among 50 emerging technologies[1] that would have profoundly influence the British economy and society in 2024. The China Association for Science and Technology grandly released the major scientific issues of 2024, among which the research on digital humans and robots with emotions and emotional intelligence was selected as one of the top ten frontier scientific issues[2]. Clearly, the development of AI emotion and expression recognition technology has become an inevitable requirement for general AI, digital computing and multidisciplinary research [2].

- *Yan Wang, Shaoqi Yan, Yang Liu, Jing Liu, Yang Chang and Xinji Mai are with the Academy for Engineering & Technology, Fudan University, Shanghai, China. E-mail: {yanwang19, sqyan19, yang_liu20, jingliu19}@fudan.edu.cn; ychang24, xjmai23@m.fudan.edu.cn.*
- *Wei Song is with the College of Information Technology, Shanghai Ocean University, Shanghai, China. E-mail: wsong@shou.edu.cn.*
- *Xiping Hu is with the School of Medical Technology, Beijing Institute of Technology, Beijing, China. E-mail: huxp@bit.edu.cn.*
- *Wenqiang Zhang is with the Academy for Engineering & Technology, Fudan University, Shanghai, China, and also with the School of Computer Science, Fudan University, Shanghai, China. E-mail: wqzhang@fudan.edu.cn.*
- *Zhongxue Gan is with the Academy for Engineering & Technology, Fudan University, Shanghai, China. E-mail: ganzhongxue@fudan.edu.cn.*

Manuscript received August 28, 2024;
*(*Corresponding authors: Zhongxue Gan and Wenqiang Zhang.)*

1. https://www.ukri.org/wp-content/uploads/2023/12/IUK-05122023-INO0617_Emerging-Tech-Report_AW2-final.pdf
2. https://www.cast.org.cn/xw/KXYW/art/2024/art_e9df73f3c2f5480aaaa6e78fffd69acd.html

Facial expressions [3] are the primary and straightforward means of human emotional expression, frequently employed and of utmost importance in interpersonal interaction [4], [5]. They convey richer affective information nonverbally than other forms of messages like voice, gestures, and body postures [6]. The concept of facial emotions was originally introduced by Darwin in his book "The Expression of the Emotions in Man and Animals" (1872). It has been noted that expressions are innate in nature and the remains of adaptive movements of animals and humans during evolution and survival. Ekman and Friesen [7] proposed six basic emotions: Happy, Angry, Sad, Surprise, Fear, and Disgust, and found universal associations between specific facial muscle patterns and emotions types, which are consistent across cultures. In recent years, with the advancement of AI technologies, facial emotion recognition (FER) methods have rapidly developed and shown wide applications in psychological research [8], medical diagnosis [9], and intelligent human-computer interaction [10].

The FER aims to identify an individual's emotional state based on the analysis of facial expressions [11], [12]. Depending on the type of data used to capture the expressions, the FER can be divided into two parts: image-based static FER (SFER) [13], [14] and video-based dynamic FER (DFER) [15], [16], [17]. The SFER works on solving challenges due to pose occlusion, cross-domain inconsistency, label uncertainty, insufficient data volume, and cross-modality. Researchers also use various data augmentation techniques and regularization methods to alleviate the problems of insufficient data volume and label uncertainty. In addition, the robustness and accuracy of expression recognition are enhanced through cross-modal information fusion. While SFER focuses on instantaneous expressions, DFER

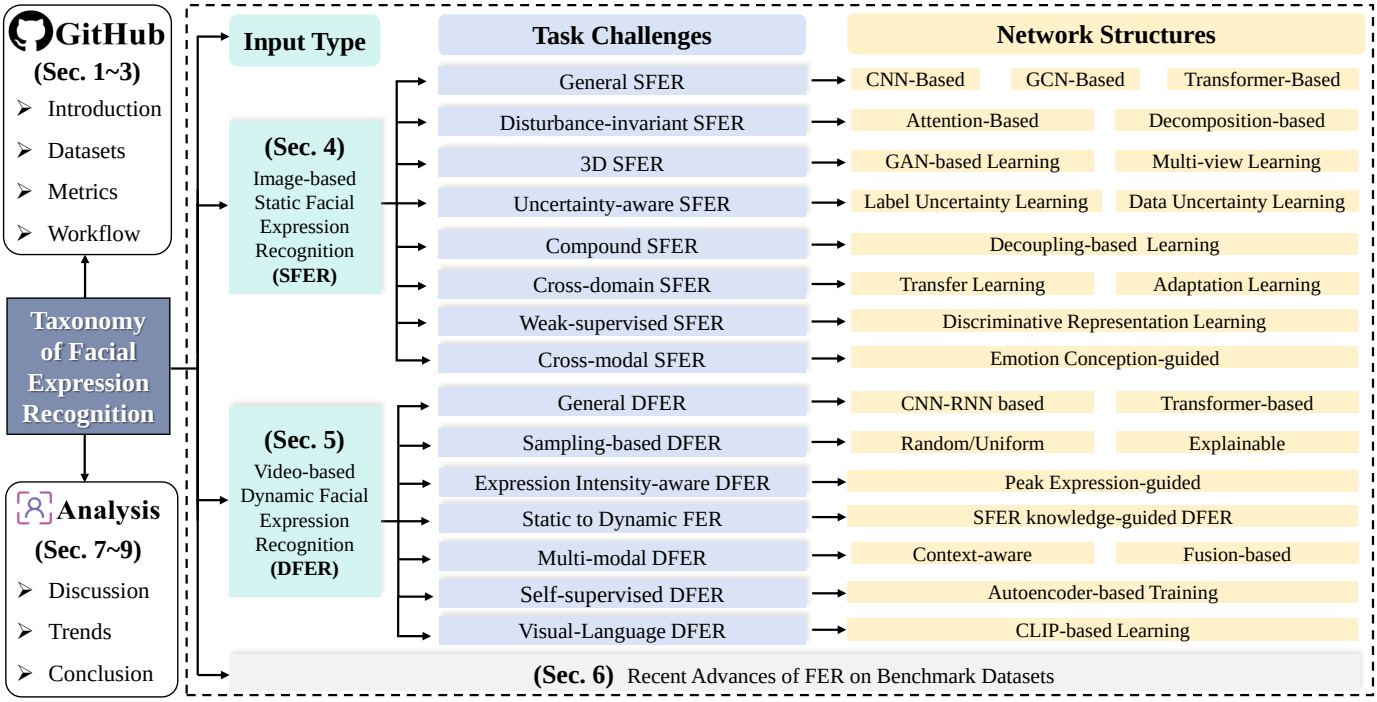

Fig. 1: Taxonomy of FER of static and dynamic emotions. We present a hierarchical taxonomy that categorizes existing FER models by input type, task challenges, and network structures within a systematic framework, aiming to provide a comprehensive overview of the current FER research landscape. First, we have introduced datasets, metrics, and workflow (including literature and codes) into a public GitHub repository[3] **(Sec. 1, 2, and 3)**. Then, image-based SFER **(Sec. 4)** and video-based DFER **(Sec. 5)** overcome different task challenges using various learning strategies and model designs. Following, we analyzed recent advances of FER on benchmark datasets **(Sec. 6)**. Finally, we discuss and conclude some important issues and potential trends in FER, highlighting directions for future developments **(Sec. 7, 8, and 9)**.

concentrates on temporal changes of facial expressions to accurately describe and comprehend the whole process of emotional shifts. Dealing with expression recognition in video sequences, DFER has main challenges in key frame extraction, spatiotemporal feature extraction, expression intensity changes, and cross-modal fusion. To capture the dynamic expression information, DFER models not only focus on static features in a single frame, but also incorporate the temporal relationship between consecutive frames.

1.1 Taxonomy Overview

In this paper, we systematically summarize the current state of FER research and provide a hierarchical taxonomy to organize existing FER works according to input type (image-based SFER and video-based DFER), task challenges, and network structures, as shown in Fig. 1. For SFER, we identify eight key challenges such as disturbances, uncertainty, compound labels, cross-domain adaptability, and cross-modality issues, and summarize the model structures of existing approaches that often used to address the corresponding challenge. For DFER, we incorporate seven additional considerations like key frame extraction, expression intensity changes, static-to-dynamic consistency, semi-supervised learning, and cross-domain alignment, as well as the solutions of current methods. We further analyzed and discussed the recent advances of typical reviewed methods on benchmarking datasets. In addition, we have summarized the benchmark datasets, evaluation metrics, literature,

3. https://github.com/wangyanckxx/SurveyFER

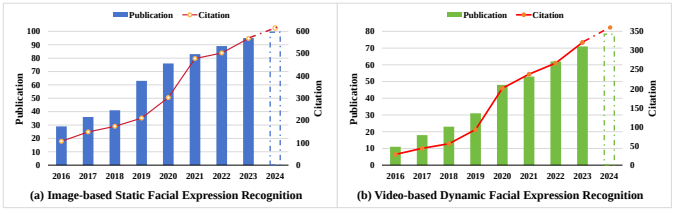

Fig. 2: The statistics of Publication (Bar) and Citation (Line) on the topic of (a) image-based SFER and (b) video-based DFER from 2016 to 2024.

codes, workflow, and discussions in the github repositories. To develop this taxonomy, we have extensively reviewed a substantial amount of research papers from 2016 to 2024. Fig. 2 tracks the publication and citation trends related to image-based SFER and video-based DFER from 2016 to 2024. There is a notable surge in both publications and citations starting around 2019, continuing to rise through 2023 and projected into 2024. This reflects growing interest and advancements in both SFER and DFER fields.

1.2 Related Reviews

In the past five years, some reviews [18], [21], [23], [24] have covered FER works and generated various classification systems. To highlight the unique contributions of our review, we compared with several existing key reviews and summarized it in Table 1. Review studies [18], [19], [20] mainly introduce and analyze various DL-based FER techniques from the laboratory-controlled environment to

TABLE 1: Comparisons on our FER review with state-of-the-art FER-related reviews from 2020 to 2024.

Pub. [Ref]	Year	Datasets		WF	Image-based Static FER							Video-based Dynamic FER						Application		
		S	D		DI	3D	UA	CP	CD	LS	CM	SL	EI	MM	SD	SS	VL	HPC	PE	HCI
IEEE TAFFC [18]	2022	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗	✓
INFSCI [19]	2022	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
IEEE TIE [20]	2023	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗	✓	✗	✗	✗	✗	✓	✗	✓
COMSCIREV [21]	2023	✗	✗	✓	✗	✗	✓	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓
PR [22]	2020	✓	✓	✓	✗	✓	✗	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
IEEE TAFFC [23]	2022	✗	✗	✓	✗	✗	✓	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗
IEEE TAFFC [24]	2023	✓	✓	✓	✗	✗	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
IEEE TPAMI [25]	2023	✓	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	✓	✓	✓
IEEE TPAMI [26]	2021	✓	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✓	✓
Our FER Review	2024	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

✓ and ✗ denote the corresponding content contains and does not contain systematic analysis and discussion, respectively.
S, D and WF represents Static, Dynamic and Workflow, respectively.
DI, 3D, UA, CP, CD, LS and CM represents Static, Dynamic and Workflow, respectively. SL, EI, MM, SD, SS, and VL represents Sampling, Expression Intensity, Multi-modal, Static to Dynamic, Semi-supervised and Visual-Language, respectively.
HPC, PE, and HCI represents Health and Psychological Counseling, Personalized Education, and Human-Computer Interaction, respectively.

in-the-wild circumstances. Our review not only provides an in-depth analysis of the standard processes of FER systems and the challenges in practical deployments, such as pose occlusion, cross-domain inconsistency, and label uncertainty, but also discusses different methods and technological advancements in image-based SFER and video-based DFER, providing a more comprehensive perspective. [21] primarily focuses on the most popular techniques and current trends in visual emotion recognition. However, our work further refines the study of a single modality (facial expressions), delving into the latest methods and technical challenges of static and dynamic FER. Recent FER review works [22], [23], [24], [25] focus on 3D FER, graph-based or multi-view facial expression analysis, and discusses demographic biases in FER datasets. Our review covers a wider variety of deep learning techniques, not limited to graph methods and multi-view issues but further covers the latest advancements in cross-domain learning, cross-modal fusion and self-supervised learning. [26] mainly focuses on the application of pain detection or emotional mimicry in educational settings through facial expressions. Our review covers more application scenarios and potentials of FER in different fields, while discussing technical and ethical issues.

1.3 Contribution Summary

To clarify FER development and inspire future research, this survey covers research background, datasets, generic workflow, task challenges, methods, performance evaluation, applications, ethical issues, and development trends. In summary, the main contributions of this work are fourfold:

1) To the best of our knowledge, this is the first comprehensive survey that divides FER research into image-based SFER and video-based DFER, extending from model-oriented development to challenge-oriented taxonomy, and provides an in-depth analysis of the real-world challenges and solutions.
2) We systematically review the latest representative methods of SFER regarding eight main challenges (such as expression disturbance, uncertainty, cross-domain inconsistency) and DFER regarding seven main challenges (key frame extraction, expression intensity variations, and cross-modal alignment).
3) We summarize, analyze and discuss recent advances and technical challenges of FER on diverse benchmark datasets under the setups of in-the-lab FER, in-the-wild SFER, and in-the-wild DFER.
4) This survey summarizes three field applications and ethical issues, and discuss development trends (such as zero-shot FER and embodied facial expression generation), aiming to provide a new perspective and guidance on FER systems.

2 Datasets and Evaluation Metrics

Facial expression data is the key foundation for implementing and developing FER algorithms. Adequate and diverse expression datasets provide the necessary training and testing material for the FER algorithms. Table 2 shows the publicly available benchmark datasets with different attributes. Some examples of widely-used datasets are illustrated in Fig. 3. Additionally, we introduce evaluation metrics.

2.1 Image-based SFER Datasets

The image-based SFER dataset is composed of individual images, each representing a specific emotional state. Based on the image collection scenarios, these datasets are divided into three groups: in-the-lab, in-the-wild, and 3D datasets .

For in-the-lab SFER, there are 5 widely used as benchmark datasets, including 1) **JAFFE** [27] consists of 213 images of seven basic facial expressions posed by 10 Japanese women, with each expression performed multiple times; 2) **CK+** [28] includes 593 sequences, with 327 labeled for seven basic emotions plus contempt; expression intensity progresses from neutral to peak; 3) **Oulu-CASIA** [35] contains videos from 80 subjects under different lighting conditions, capturing six basic emotions; 4) **MMI** [38] includes 740 images and 2,900 video sequences depicting seven basic emotions, starting and ending with neutral expressions; 5) **RaFD** [39] features 8,040 high-quality images of seven basic facial expressions and contempt, taken from different angles with uniform settings, involving 67 professional actors.

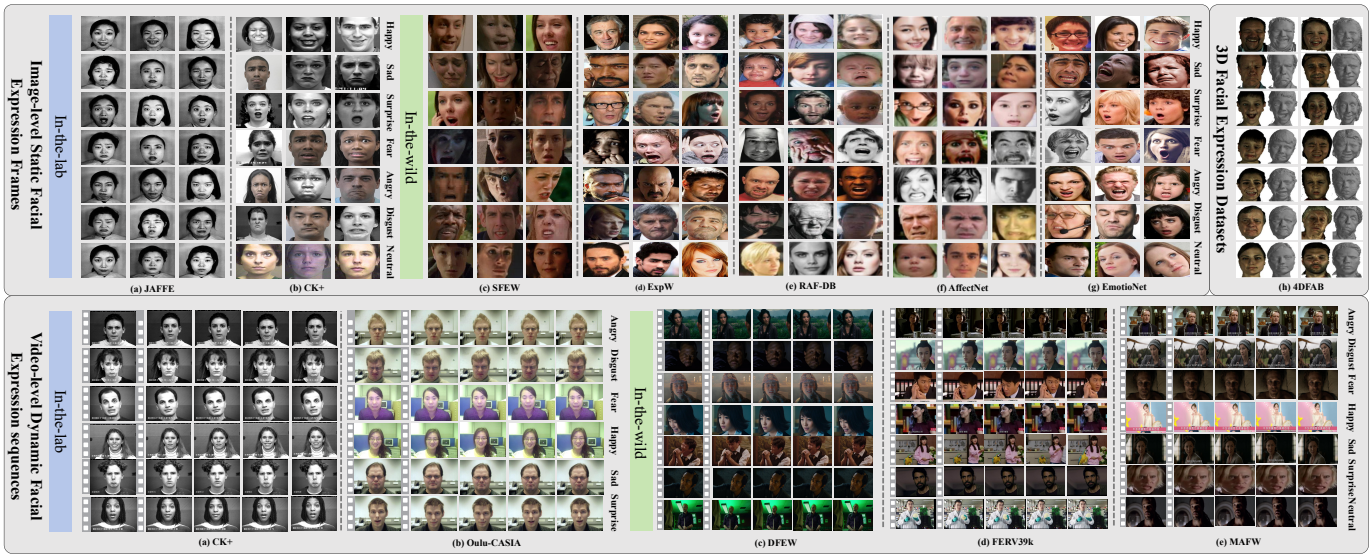

Fig. 3: Image-based static facial frames (**Above**): (a) JAFFE [27], (b) CK+ [28], (c) SFEW [29], (d) ExpW [30], (e) RAF-DB [31], (f) AffectNet [32], (g) EmotioNet [33], (h) 4DFAB [34]; and video-based dynamic facial sequences (**Below**): (a) CK+ [28], (b) Oulu-CASIA [35], (c) DFEW [15], (d) FERV39k [36], and (e) MAFW [37] of seven basic emotions in the lab and wild.

For in-the-wild SFER, there are 6 widely used as benchmark datasets, including 1) **FER-2013** [40] includes 35,887 images labeled with seven basic facial expressions; 2) **SFEW 2.0** [29], derived from AFEW, comprises 1,766 images; 3) **EmotioNet** [33] features over 1 million images labeled with basic and complex facial expressions, and action units (AUs); 4) **RAF-DB** [31] contains 29,672 images, with 15,339 labeled for basic expressions and part compound expressions; 5)**AffectNet** [32], collected using 1,250 emotion-related keywords in six languages, consists of approximately 1 million images, with seven discrete facial expressions and the intensity of valence and arousal; 6) **ExpW** [30] contains 91,793 images labeled with seven basic facial expressions.

For 3D/multi-view SFER, there are 4 widely used as benchmark datasets, including 1) **BU-3DFE** [41], with 2,500 3D facial expressions from 100 subjects; 2) **Bosphorus** [42] includes 4,652 3D facial images from 105 subjects, with a wide range of AUs and expressions; 3) **4DFAB** [34] spans five years of collection from 180 subjects, comprising over 1.8 million high-resolution 3D facial images, including both posed and spontaneous expressions.

2.2 Video-based DFER Datasets

The video-based DFER datasets are typically composed of videos or image sequences with durations ranging from 0.4 to 5 seconds, which are also divided into two categories: controlled laboratory scenes and complex real scenes.

For in-the-lab DFER, there are 3 widely used as benchmark datasets, including 1) **CK+** [28] comprises 593 facial expression sequences posed by 123 subjects, of which only 327 sequences are labeled with seven basic emotions and contempt; 2) **MMI** [38] consists of 740 images and 2,900 video sequences, which were posed by 32 subjects in a laboratory setting, depicting seven basic emotions; 3) Oulu-CASIA [35] includes 2,880 video sequences captured in a laboratory setting with 80 subjects posing in front of the camera with the six basic facial expressions.

TABLE 2: Summary of the in-the-lab or in-the-wild **datasets** with static and dynamic emotions for FER training and evaluation. ECT: Elicitation; P: Posed; I: Instinctive; Sev: Seven Emotions (Happy, Angry, Surprise, Fear, Sad, Disgust, Neutral); C: Contempt; A: Anxiety; D: Disappointment; H: Helplessness; Com: Compound.

Categories		Datasets	Year	ECT	Emotion	Training Numbers	Testing Numbers
Modality	Scene						
Image-based SFER Datasets	Lab	JAFFE [27]	1998	P	Sev	213	213
		CK+ [28]	2010	P/I	Sev	241	241
		MMI [38]	2010	P	Sev	370	370
		Oulu-CASIA [35]	2011	P	Sev	720	240
		RaFD [39]	2010	P	Sev, C	1,448	160
	Wild	FER-2013 [40]	2013	P/I	Sev	28,709	3,589
		SFEW 2.0 [29]	2011	P/I	Sev	958	436
		EmotioNet [33]	2016	P/I	Sev, C	80,000	20,000
		RAF-DB [31]	2017	P/I	Sev, Com	12,271	3,068
		AffectNet [32]	2017	P/I	Sev, Con.	283,901	3,500
		ExpW [30]	2017	P/I	Sev	75,048	16,745
	Lab (3D)	BU-3DFE [34]	2006	P	Sev	2,000	500
		Bosphorus [42]	2008	P	Sev	2,326	2,326
		4DFAB [34]	2018	P/I	Sev	1,440k	360k
Video-based DFER Datasets	Lab	CK+ [28]	2010	P/I	Sev	241	241
		MMI [38]	2010	P/I	Sev	1,450	1,450
		Oulu-CASIA [35]	2011	P	Six	2,160	720
	Wild	AFEW 8.0 [43]	2011	P/I	Sev	773	383
		CAER [44]	2019	P/I	Sev	9,240	2,640
		DFEW [15]	2020	P/I	Sev	12,000	3,000
		FERV39k [36]	2022	P/I	SE	35,887	3,000
		MAFW [37]	2022	P/I	Sev, C, A, D, H, Com	8,036	2,009

For in-the-wild DFER, there are 5 widely used as benchmark datasets, including 1) **AFEW 8.0** [43], used in the EmotiW competition, is a multimodal video dataset featuring spontaneous human expressions collected from TV and film clips, encompassing various head poses, object occlusions, and lighting conditions; 2) **CAER** [44] contains 13,201 annotated facial video clips from American televi-

sion series; 3) **DFEW** [15] offers 12,059 clips from 16,372 clips annotated via a voting mechanism, segmented into five sections for cross-validation; 4) **FERV39k** [36] contains 38,935 video segments from 22 scenes and four situations, annotated with seven basic expressions, covering diverse scenarios and expression intensities; 5) **MAFW** [37] is a large-scale, multi-modal affective dataset with 10,045 video and audio clips from over 1,600 movies and TV dramas, categorized into eleven emotions, including seven basic and four additional expressions.

2.3 Evaluation Metrics

When evaluating the performance of FER models, several key evaluation metrics are commonly used: 1) **Accuracy** refers to the ratio of the number of correctly predicted samples to the total number of samples, which measures the FER model's ability to correctly identify expressions. 2) **Recall** measures the proportion of correctly predicted samples for a particular expression relative to the total actual samples of that expression; 3) **Weighted Average Recall (WAR)** calculates the weighted average by multiplying the recall of each class by its proportion in the dataset, providing a balanced view of performance across different classes; 4) **Unweighted Average Recall (UAR)** averages the recall rates across all classes without considering class proportions, offering a fair assessment of model performance in imbalanced datasets. Besides, the **Confusion Matrix** is a two-dimensional grid that visualizes a FER model's prediction results by comparing actual versus predicted categories.

3 A Workflow of Generic FER

Fig. 4 shows the workflow and main components of generic FER. Four critical steps are typically included: 1) acquiring and sampling strategy of dynamic facial expression image sequences (only for DFER in red dashed rectangle); 2) preprocessing of align, naturalization and augmentation; 3) facial expression feature extraction of 2D CNN, Attention, RNN and 3D CNN (only for DFER in red dashed rectangle); and 4) expression recognition with single or mixed emotion.

3.1 Facial Frame Sampling

Since dynamic facial expressions in natural multi-scene environments usually last between 0.5 and 4 seconds [45], researchers manually crop video clips at this interval when constructing DFER datasets. The number of frames transmission per second is usually 24 or higher. The step of "Sampling" in Fig.4 shows an example of 8 frames, which are sampled from a 40-frame disgust video sequence. There are two common sampling methods: uniform frame sampling and random frame sampling. In DFEW [15] and FERV39k [36], the uniform sampling strategy is first used to generate a sequence face images of length 16 or 8 from all available video clips with the assistance of random sampling and Time Interpolation Method (TIM). In [46], [47], the facial expression video is evenly segmented into U segments, then V image frames are randomly selected from each segment, eventually, the length of a facial image sequence is $U \times V$. Note facial frame sampling is only used for dynamic emotion data (video-based DFER tasks).

3.2 Facial Data Preprocessing

The original dynamic facial expression sequences obtained from the natural world often contain expression-irrelevant variables, such as complex backgrounds, varying illumination, and face pose changes. To exclude the irrelevant information, serialized face image preprocessing [48] is necessary to align, normalize, and augment the semantic information of facial regions before deep feature extraction.

Face Alignment focused on the automatic detection of facial landmarks to eliminate background and non-expression elements, which can be generally categorized into two main approaches: cascaded regression (CPR) models and DL-based methods. On the basis of the CPR, Robust CPR (RCPR) [49] improved robustness against occlusions and shape variations. Early methods, such as DCNN [50], and TCNN [51], directly applied multi-layer CNNs to learn key features of face (facial landmarks) for alignment. Recently, SfSNet [52] leveraged DL-based models for enhanced performance. In addition, the landmark strategy is improved from the perspective of semantic understanding. Zhou et al. [53] proposed the Self-adaptive Ambiguity Reduction (STAR) loss to address semantic ambiguity in landmark detection.

Face Normalization involves illumination normalization [54] and pose normalization [55]. Ma et al. [56] utilized cyclic consistency loss for light normalization as a style transfer problem, while Han et al. [57] developed Asymmetric Joint GANs for controlled reillumination. Disentangled Representation learning-GAN [58] focused on achieving pose-invariant facial representations via GAN-based models. Tripathy et al. [59] introduced a self-supervised approach for dynamic face reproduction. Unlike methods that rely on separate classifiers for each pose, Zhang et al. [60] proposed an end-to-end deep GAN model that integrates face synthesis and pose-invariant expression recognition.

Data Augmentation can effectively provide sufficient samples when the limited labeled facial expression images or sequences in the dataset cannot meet the requirement of FER training. It can be divided into two types: offline and on-the-fly. Scaling and tilting the original images, along with the inclusion of random noise and scrambling, are the most commonly employed offline data augmentation techniques [61]. GAN-based data synthesis methods [60], [62] could also be applied to generate various face and expression images. The on-the-fly data augmentation [63] usually use data expansion methods embedded in DL toolkits, such as random rotate and crop, and color jitter.

3.3 Facial Emotion Feature Extraction

As deep learning continues to advance for various tasks, especially image and video related tasks, recent efforts in FER have concentrated on optimizing network architectures applied for facial emotion feature extraction. These network models can be classified into four categories: 1) deep convolutional neural networks; 2) attentional mechanisms; 3) recurrent neural networks, and 4) 3D convolutional neural networks. Specific network structures will be described in Sec. 4 and Sec. 5 according to the task challenges.

CNN-based Models have achieved tremendous success in the field of computer vision, particularly excelling in tasks

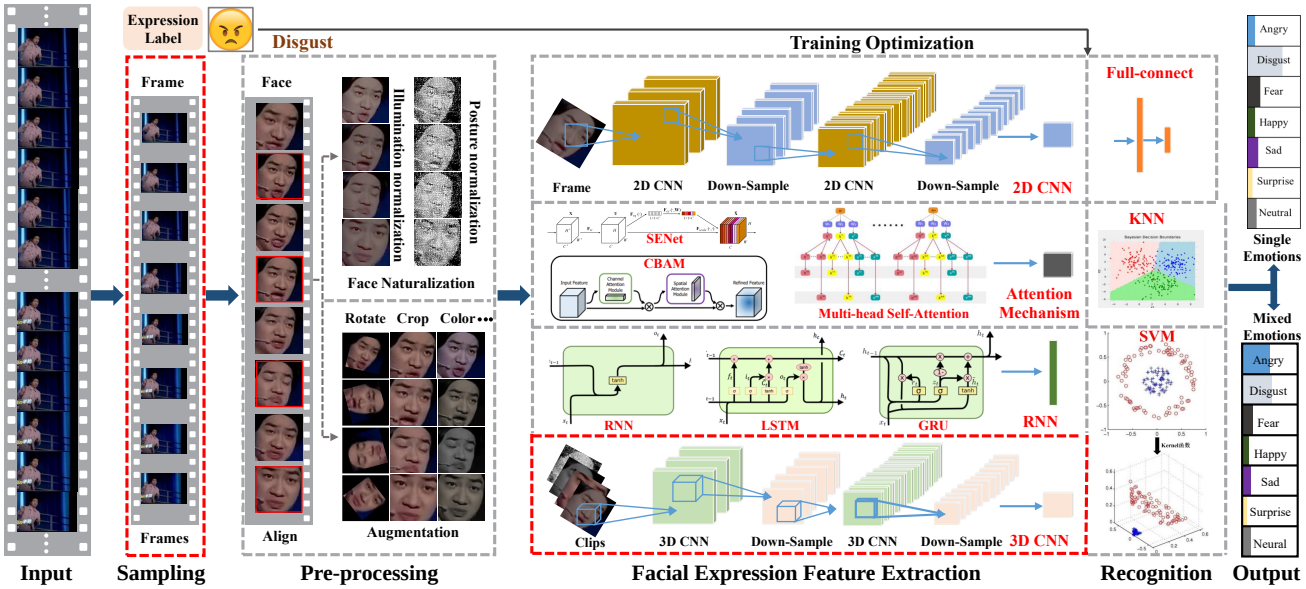

Fig. 4: The workflow and main components of generic facial expression recognition.

such as image classification, object detection, and segmentation. By utilizing convolutional layers and pooling layers, CNNs such as VGGNet [64] and ResNet [65] can effectively extract both local and global features from images, making them suitable for static facial expression feature extraction.

Attention-based Models help the network model extract more representative information by adaptively finding spatial regions, channels, or Spatio-temporal sequences that are meaningful to the task based on the input feature vectors. The Vision Transformer (ViT) [66] excels in capturing long-range dependencies in sequence data and images using self-attention or multi-Head attention mechanisms, making them suitable for complex FER tasks [67], [68].

Recurrent Neural Networks (RNNs) learn mappings between complex feature tensors and combine temporal and spatial information (with CNNs) to further improve performance. Mostly used RNN-architecture networks are Long Short-Term Memory Networks (LSTM) [69] and Gated Recurrent Unit Networks (GRU) [70]. They can effectively capture subtle changes in facial expressions regarding image patches or video frames as a time series [71].

3D Convolutional Neural Networks (C3D) [72] model both spatial and temporal signals simultaneously . 3D CNNs [73], [74] extend traditional 2D convolutional operations like ResNet or Inception [75], [76] into the temporal dimension by applying 3D convolutions. These C3Ds can model the spatiotemporal patterns of facial movements, making them particularly effective in capturing the nuances of dynamic expressions [15], [36].

3.4 Recognition of Facial Emotions

Recognition of facial emotions aims to calculate the classification probability of input data by traditional machine learning or deep learning methods, and eventually determine the expression with single labels or mixed labels.

Machine Learning based Classifier includes the widely accepted traditional machine learning classification methods, such as Support Vector Machine (SVM) [77] and Multilayer Perceptron (MLP) [78]. Among them, SVM is one of the most effective classifiers for affective computing.

Deep Learning based Classifier is often used in DL-based frameworks [79], [80], which classifies expressions with the features extracted by previous network layers. Specifically, DL-based networks, such as ResNet [65] and 3D Convolutional Neural Network (C3D) [73] consider facial emotion features for end-to-end expression recognition of images or sequences by fully-connected layers.

4 Image-based Static FER

Image-based static facial expression recognition (SFER) involves extracting features from a single image, which captures complex spatial information that related to facial expressions, such as landmarks, and their geometric structures and relationships. In the following, we will first introduce the general architecture of SFER, and then elaborate specific design of SFER methods from the challenge-solving perspectives, including disturbance-invariant SFER, 3D SFER, uncertainty-aware SFER, compound SFER, cross-domain SFER, waek-supervised SFER, and cross-modal SFER.

4.1 General SFER

A general SFER often involves global and local or multi-scale feature extractions, feature fusion, and emotion classification. Fig. 5 shows an example architecture of general SFER. In this process, deep learning models serve as the foundational framework, mainly including Convolutional Neural Networks (CNNs), Graph Convolutional Networks (GCNs), and Transformer-based models. The integration and advancement of these deep learning architectures have significantly enhanced the performance of SFER systems, enabling robust recognition across diverse environments.

4.1.1 CNN-based Models

CNN-based methods [84], [85], [86] have proven instrumental in SFER by efficiently extracting local and global facial characteristics through layered convolution and pooling operations, facilitating accurate expression classification. Facial Motion Prior Networks (FMPN) [87] and Oriented Attention

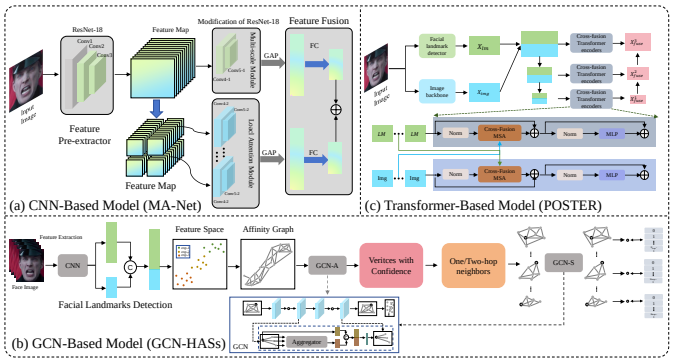

Fig. 5: The architecture of general SFER. Figure is reproduced based on (a) CNN-based model [81], (b) GCN-based model [82], and (c) Transformer-based model [83].

Pseudo-Siamese Network (OAENet) [88] leveraged convolutional blocks to capture global and local facial information via facial landmarks and correlation coefficients. As shown in Fig. 5(a), the global multi-scale and local attention network (MA-Net) [81] utilized multi-scale module and a local attention module to extract both local and global facial features.

4.1.2 GCN-based Models

Compared with traditional CNN approaches, Graph Convolutional Network (GCN)-based methods are particularly better at handling the geometric relationships and topological information of facial features to capture spatial dependencies through graph structures for recognizing subtle changes in expressions. As shown in Fig. 5(b), Liu et al. [82] developed a method that utilizes high aggregation subgraphs (GCN-HASs) for FER. By emphasizing the importance of high-order neighbors and employing vertex confidence, their approach constructs subgraphs that effectively capture the intricate relationships between facial expressions, leading to significant improvements in both recognition accuracy and efficiency. These contributions highlight the innovative advancements in leveraging GCNs for SFER, demonstrating their potential to surpass the limitations of CNN-based methods. Jin et al. [79] employed a graph-structured representation (DDRGCN) where each node corresponds to appearance information around facial landmarks, and edges encode the geometric relationships between these nodes. This approach captures both local appearance and spatial geometry, providing a robust framework for recognizing facial expressions.

4.1.3 Transformer-based Models

Since Transformer-based methods can capture global information and spatio-temporal relationships in facial expressions [67], [89], novel architectures are optimized by introducing multi-scale and cross-modal attention mechanisms. As shown in Fig. 5(c), a Pyramid Cross-Fusion Transformer network (POSTER) [83] utilized a transformer-based cross-fusion approach to effectively integrate facial landmark features with image features, directing attention to important facial regions and enhancing scale invariance. In addition, Li et al. [67] employed a Masked Auto-Encoder pretrained on unlabeled face images, combined with a pretrained Vision Transformer and CNN, to tackle the issue of limited annotations in facial expression data for affective behavior analysis.

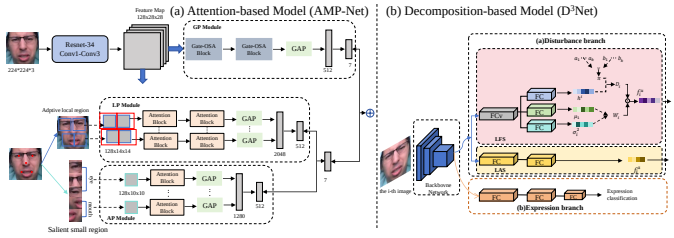

Fig. 6: The architecture of disturbance-invariant SFER. Figure is reproduced based on (a) Attention-based model [90] and (b) Decomposition-based model [91].

4.2 Disturbance-invariant SFER

One of the main challenges in FER is to address the disturbance caused by various disturbing factors [14], [91], [92], [93], including common ones (such as identity, pose, and illumination) and potential ones (such as hairstyle, accessory, and occlusion). These disturbing factors will lead to partial information missing. To overcome the impact of disturbance, it is critical to extract effective facial expression features from available facial regions.

4.2.1 Attention-based Models

The attention-based models [93], [94] based on attention mechanism [95] can help the model better focus on the unoccluded facial regions, thereby improving the accuracy and robustness of expression recognition in complex backgrounds and lighting conditions.

Region-based FER methods analyze a face image by dividing it into overlapping or non-overlapping local regions, allowing the model to concentrate on localized features for more precise expression recognition. Li et al. [96] introduced the Patch-Gated Unit (PG-Unit), which computes one-dimensional weights for regions of interest based on facial landmarks. These weights are then applied across feature dimensions using self-attention and relation-attention modules, enhancing the model's ability to focus on the most relevant facial regions.

Holistic-region-based methods often have two branches to extract global and local features. For example, Wang et al. [97] proposed the local attention module and correlation attention learning to obtain local attention maps and an overall saliency feature. Similarly, Fig. 6(a) presented an adaptive multilayer perceptual attention network [90], which extracted global, local, and salient facial emotional features by incorporating various fine-grained features. These approaches aims to learn the underlying diversity and crucial information inherent in facial expressions.

4.2.2 Decomposition-based Models

Decomposition-based models [91], [100], [101] aim to disentangle facial expressions from identity and posture, generating discriminative facial expression features. As shown in Fig. 6(b), the dual-branch disturbance disentangling network (D^{3}Net) [91] includes both an expression branch and a disturbance branch. The disturbance branch is divided into a label-aware sub-branch (LAS) that captures common

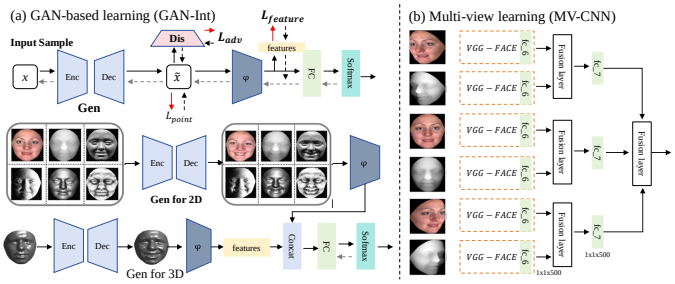

Fig. 7: The architecture of 3D SFER. Figure is reproduced based on (a) GAN-based learning (GAN-Int) [98] and (b) Multi-view learning (MV-CNN) [99].

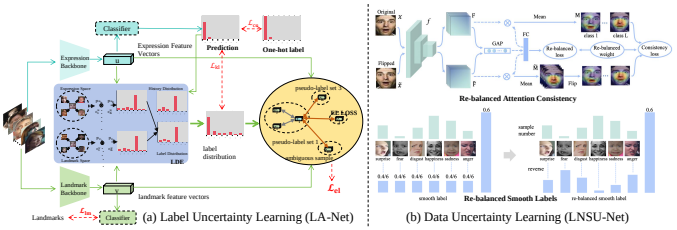

Fig. 8: The architecture of uncertainty-aware SFER. Figure is reproduced based on (a) the label uncertainty learning (LA-Net) [108] and (b) data uncertainty learning (LNSU-Net) [109].

disturbances through transfer learning, and a label-free sub-branch (LFS) that encodes potential disturbances using an unsupervised Indian Buffet Process (IBP) prior. Adversarial training is employed to further separate disturbance features from expression features, improving feature disentanglement. The feature decomposition and reconstruction learning (FDRL) [102] integrates a feature decomposition network to model similarities and a feature reconstruction network to capture relationships and reconstruct expression features using intra- and inter-feature relation modules. Additionally, Latent-OFER [92] detected occlusions and reconstructing missing regions using latent vectors from unoccluded patches by the effect of decomposition-based models in isolating and amplifying expression-specific features.

4.3 3D SFER

Despite significant advances achieved in 2D FER, it is still difficult to distinguish some facial muscle action units in 2D images due to limitations such as lighting conditions, poses, and makeup. Since 3D facial shape models include depth information and enable the observation of facial feature changes from multiple angles, 3D FER works [103], [104] utilized complementary and redundant information in 2D and 3D data to capture subtle deformations and details.

4.3.1 GAN-based Learning

Generative Adversarial Network (GAN)-based methods can generate high-quality, diverse, and nearly-realistic facial expression images through the adversarial training of generators and discriminators. This make it easier for improving the generalization ability of FER models. As shown in Fig. 7(a), Yang et al. utilized GAN model (GAN-Int) [98] to jointly design intensity enhancement and expression recognition, ensuring that synthesized faces exhibit high-intensity expressions. Similarly, Zhang et al. [105] proposed Joint Pose and Expression GAN-based model (JPE-GAN) to simultaneously perform facial image generation and pose-invariant FER by corporately utilizing different poses and expressions.

4.3.2 Multi-view Learning

Multi-view learning in 3D FER [104], [106] utilized multi-angle 3D facial images and combines features from various perspectives to distinguish different expressions, effectively addressing variations in pose and lighting conditions, thus enhancing overall recognition performance. As shown in Fig. 7(b), Vo et al. [99] proposed a novel multi-view CNN model (MV-CNN), which incorporates multi-view facial images and facial prior information for 3D FER. In addition, the joint spatial and scale attention network (SSA-Net) [104] localized proper regions for simultaneous head pose estimation and FER. The SSA-Net uses spatial attention to identify expression-relevant regions at various scales and employs scale attention to select the most informative scales, learning pose-invariant and expression-discriminative representations.

4.4 Uncertainty-aware SFER

FER tasks are inherently challenged by factors such as image quality, facial posture, and lighting conditions, which further introduce data and label uncertainty [107]. Uncertainty-aware SFER models aim to classify the facial expressions while handling the uncertainty of each class.

4.4.1 Label Uncertainty Learning

Label data may contain noise or errors [110], [111] due to human annotation mistakes or inherent data ambiguity, significantly affecting model performance. Robust techniques have been used to deal with noise labels: 1) design Label Distribution Learning on Auxiliary Label Space Graphs (LDL-ALSG) [110] to suppress noise; 2) use unlabeled data to assist the model in recognizing and correcting noise in label data [112]; 3) reduce the impact of noisy labels on FER models by erase attention consistency (ECA) [111], similarly, as shown in Fig. 8(a), LA-Net [108] also leveraged facial landmarks for attention and label correction to counter label noise.

4.4.2 Data Uncertainty Learning

Large-scale FER datasets collected in the wild often encounter issues like image blur, noise, and low resolution, leading to ambiguity in emotion recognition [113]. These challenges complicate distinguishing between images with multiple emotions and those with noisy labels. To address this, Zhang et al. [114] introduced a relative uncertainty learning framework that estimates the uncertainty of each prediction relative to others, improving model robustness. The Emotion Ambiguity-Sensitive Cooperative Networks (EASE) [115] further tackle this by categorizing training samples into clean, noisy, and conflict groups, enhancing network diversity and representation learning. Incorporating auxiliary tasks, Zhao et al. [116] developed an uncertainty-aware model using multi-task auxiliary correction to improve FER accuracy under uncertain conditions.

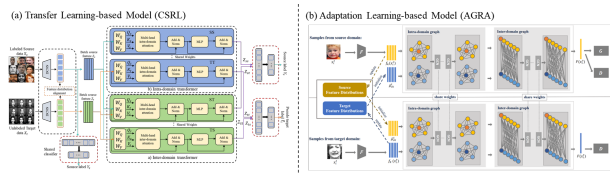

Fig. 9: The architecture of cross-domain SFER. Figure is reproduced based on (a) the transfer learning-based model (CSRL) [120] and (b) the adaption learning-based model (AGRA) [121] .

As shown in Figure 8(b), Deng et al. [109] employed rebalanced attention maps, enabling models to better extract information from underrepresented classes like fear or disgust, thus enhancing overall performance in FER.

4.5 Compound SFER

Compound emotions [31] refer to complex emotional states formed by the combination of at least two basic emotions, which are not independent, discrete categories. Compared with seperate basic emotions, compound emotions are more capable of representing the diversity and continuity of human's complex emotions. Li et al. [117] proposed a self-supervised exclusive-inclusive interactive learning method for multi-label FER, effectively capturing and disentangling both inclusive and exclusive facial expressions within a single image. Deng et al. [118] improved multi-label FER by introducing attention flipping consistency loss and label-guided spatial attention dispersing loss, which bolstered network stability, interpretability, and performance without additional data. Additionally, Deng et al. [119] addressed basic-compound FER as a single-label multi-class task, proposing the iterated soft label mining algorithm and expression correlation score learning loss to effectively leverage label correlations.

4.6 Cross-domain SFER

In real-world environments, facial expressions vary across race, culture, and age, as well as annotators' cultural and experiential biases, reducing the performance of existing recognition methods on diverse datasets [121]. Fortunately, advancements [122] in transfer learning and adaptation learning have facilitated the transfer of knowledge from labeled source domains to target domains, enhancing the generalization of the cross-domain SFER model across different contexts.

4.6.1 Transfer Learning-based models

Variations in data collection conditions across different datasets can lead to significant performance degradation when models trained on one dataset are applied to another. As shown in Fig. 9(a), the Cross-domain Sample Relationship Learning (CSRL) [120] reduces domain discrepancy by leveraging intrinsic sample relationships across domains. Specifically, during training, inter-domain sample transformers are designed to explore similarity relationships between source and target domains, while intra-domain sample transformers capture internal structures within each domain. Furthermore, a joint alignment strategy is employed to align feature distributions and sample relationships across domains, enhancing the model's generalization ability by aligning both local sample similarities and global domain distributions. Zheng et al. [123] proposed a joint local-global discriminative subspace transfer learning method that learns a domain-invariant subspace by integrating both local and global information. Additionally, Zheng et al. [124] introduced a cross-domain color FER method using transductive transfer subspace learning to identify a shared subspace for effective knowledge transfer. Further, Zheng et al. [125] proposed learning a common latent embedding space to enhance cross-domain FER. They also suggested learning transferable sparse representations for effective cross-corpus recognition [126], aiming to extract discriminative features that can generalize across datasets.

4.6.2 Adaptation Learning-based models

Adaptation learning plays a pivotal role in addressing the domain shift challenges (different feature distribution of the same expression in different datasets) inherent in cross-domain FER [121], [127]. Adversarial learning helps the model achieve domain adaptation between the source domain and the target domain via approximating the feature distributions of the source and target domains [128]. As shown in Fig. 9(b), Chen et al. [121] combined graph representation propagation with adversarial learning for global-local feature co-adaptation across domains. Similarly, a Multi-source Adversarial Domain Aggregation Network (MADAN) [129] learnt domain-invariant features from multiple source domains for effective transfer to the target domain. To achieve cross-domain and discriminative feature representations, Li et al. [130] introduced the deep Emotional Conditional Adaptation Network (ECAN), which aligns both marginal and conditional distributions across domains. Additionally, Gao et al. [131] proposed multi-domain adaptive attention (SSA-ICL) with Intra-dataset Continual Learning, effectively adapting to multiple target domains and mitigating catastrophic forgetting. Recently, Zheng et al. [132] introduced a graph-diffusion-based domain-invariant representation learning, capturing the underlying manifold structure of facial expressions to achieve domain-invariant representations.

4.7 Weak-supervised SFER

Weak-supervised learning in SFER involves training models with scarce or partially available labeled data, leveraging both labeled and unlabeled data, to learn facial discriminative expression representation. In recent work, Zhang et al. [133] advanced this field by weakly supervising local regions of interest and incorporating relational reasoning between local and global features. As shown in Fig. 10, the Adaptive Confidence Margin (Ada-CM) [134] leveraged and partitioned all unlabeled data into two subsets based on confidence scores: high-confidence samples used for pseudo-label matching, while low-confidence samples contributing to feature-level contrastive learning. Shu et al. further [135] revisited contrastive learning in a semi-supervised context, proposing a framework that effectively utilizes unlabeled data to boost FER model performance. Recently, Liu et al. introduced weakly supervised contrastive

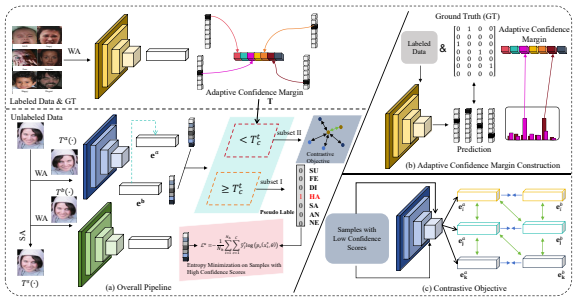

Fig. 10: The architecture of weak-supervised SFER. Figure is reproduced based on the Ada-CM [134].

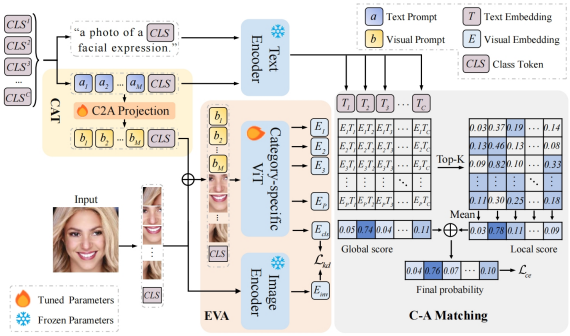

Fig. 11: The architecture of cross-modal SFER. Figure is reproduced based on the CEprompt [13].

learning (WSCFER) [136] by integrating instance-level and class-level representation learning, which balances feature discrimination through contrastive learning and partial consistency loss, minimizing focus on irrelevant details.

4.8 Cross-modal SFER

Cross-modal SFER methods [137], [138] integrated the visual facial information with emotion conception from textual sources using the visual language pre-training (VLP) [139]. For example, Yuan et al. [140] presented a method to describe facial expressions by linking image encoders and large language models, enabling the generation of textual descriptions of facial expressions that can be used for various applications. As shown in Fig. 11, the cross-modal emotion-aware prompting (CEPrompt) [13] using VLP models, incorporated emotion conception-guided visual adapter for emotion-guided visual representation, and conception-appearance tuner for optimizing cross-modal interactions, with knowledge distillation preserving pretrained knowledge, resulting in enhanced understanding of expression-related facial details.

5 Video-based Dynamic FER

The video-based DFER [141], [142] involves analyzing facial expressions that change over time, necessitating a framework that effectively integrates spatial and temporal information. The core objective of DFER is to extract and learn the features of expression changes from video sequences or image sequences. Due to the complexity and diversity of input video or image sequences [143], DFER faces various task challenges. Based on different solution approaches, these challenges can be categorized into seven basic types:

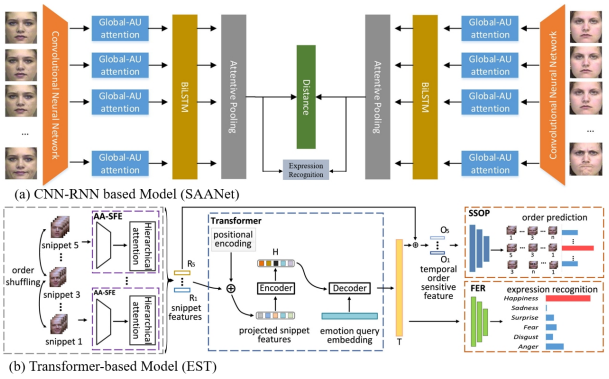

Fig. 12: The architecture of general DFER. Figure is reproduced based on (a) CNN-RNN based model (SAANet) [71] and (b) the transformer-based model (EST) [141].

general DFER, sampling-based DFER, expression intensity-aware DFER, multi-modal DFER, static to dynamic FER, self-supervised DFER, and cross-modal DFER.

5.1 General DFER

General DFER methods [16], [141], [144] primarily extract spatial-temporal features to analyze the dynamic changes in expressions. The CNN-RNN based models often combines CNNs and RNNs, while the transformer-based approach leverages deep attention mechanisms to handle more complex dynamic relationships.

5.1.1 CNN-RNN based Models

The early DFER approaches often utilized cascaded CNNs with RNNs to extract spatial and temporal features, such as STC-NLSTM and SAANet [71], [145]. As shown in Fig. 12(a), the conjoined action-unit attention network (SAANet) [71] introduced a sparse self-attention mechanism for perceiving action-unit (AU) features, coupled with a twin sampling strategy and metric learning. Similarly, the multi-task global-local network [146] integrated shared shallow, part-based, and global modules to extract spatio-temporal features from both local regions and the entire face. Chen et al. [147] emphasized the exploitation of spatial-temporal and channel correlations through attention mechanisms.

5.1.2 Transformer-based Models

Transformer-based DFER methods excel in handling complex temporal dependencies and capturing global features by modeling the nuances and long-range relationships in facial expression sequences [46]. As shown in Fig. 12(b), Liu et al. [141] introduced the Expression Snippet Transformer (EST), which decomposes expression movements into snippets, enhancing the Transformer's capability for both intra- and inter-snippet visual modeling. Similarly, Li et al. [68] proposed a unified spatial-temporal transformer that captures discriminative features within frames while modeling contextual relationships across frames, optimized by a compact soft maximum cross-entropy loss. Zhao et al. [142] developed a geometry-guided framework, combining graph convolutional networks and transformers to construct a spatial-temporal graph based on facial landmarks and local appearance, effectively representing facial expression sequences. Additionally, Poux et al. [148] tackled partial

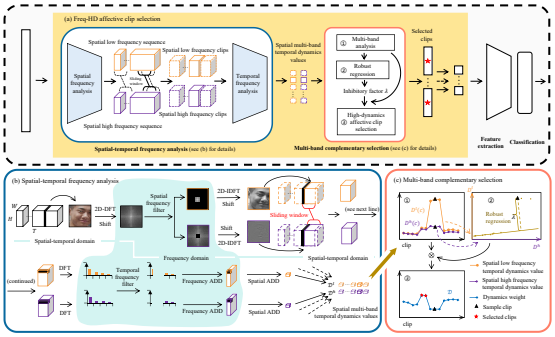

Fig. 13: The architecture of sampling-based DFER. Figure is reproduced based on explainable sampling (Freq-HD) [149].

facial occlusion by reconstructing the occluded regions in the optical flow domain using an auto-encoder with skip connections.

5.2 Sampling-based DFER

A complete dynamic facial expression lasts about 0.5 to 4 seconds [45], typically encompassing the entire process from onset to peak and then to the end of the expression. Due to variations in capture devices and the frame rates, the sampling-based DFER aims to select expression frames, while remove interference frames and invalid frames from dynamic facial expression sequences.

5.2.1 Random/Uniform Frame Sampling

In Dynamic Facial Expression Recognition (DFER), two primary frame sampling methods—random and uniform sampling—are commonly employed. Random sampling [15], which involves the arbitrary selection of frames from a sequence, is valued for its simplicity and computational efficiency by mitigating over-reliance on specific frames; however, it risks overlooking key expression changes. Conversely, uniform sampling [36], [46], [47] systematically selects a predetermined number of frames, thereby ensuring comprehensive coverage of the expression sequence, which is particularly advantageous for longer videos, though it demands greater computational resources.

5.2.2 Explainable Frame Sampling

Explainable frame sampling in DFER enhances conventional methods by automatically selecting emotion-rich key frames, improving model interpretability and decision-making. As shown in Fig. 13, Tao et al. [149] developed Freq-HD, which utilizes Spatio-Temporal Frequency Analysis (STFA) and Multi-Band Complementary Selection (MBC) to detect significant emotional shifts, effectively distinguishing expression dynamics from irrelevant variations. Similarly, Savchenko et al. [150] introduced an adaptive frame rate method that adjusts sampling based on expression complexity and model confidence, optimizing frame selection for improved accuracy and efficiency. These advancements underscore the critical role of explainable frame sampling in enhancing the performance and transparency of DFER. Besides, Wang et al. [16] developed a dual-path multi-excitation collaboration network incorporating space-frame and channel-time modules to learn complementary representations in manner of online frame extraction.

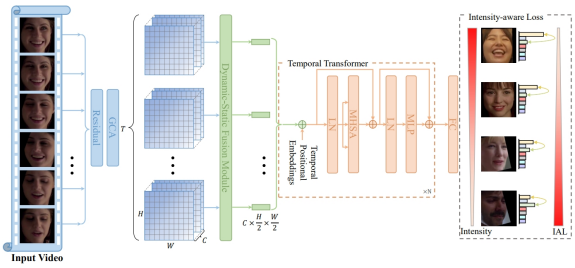

Fig. 14: The architecture of sampling-based DFER. Figure is reproduced based on the GCA-IAL [47].

5.3 Expression Intensity-aware DFER

Facial expressions are inherently dynamic, with intensity either gradually shifting from neutral to peak and back or abruptly transitioning from peak to neutral [143], making the accurate capture of these fluctuations essential for understanding expression dynamics. Early research primarily focused on modeling the temporal progression of expressions and transitions between intensity levels [151], such as Zhao et al.'s [151] use of peak-guided deep networks (PGDN) for feature extraction and peak gradient suppression during training. Recently, Li et al. [47] developed a GCA-IAL model, including a global convolution-attention module (GCA) and a temporal transformer to learn long-distance dependencies between frames, and an expression intensity perception loss function (IAL) to discriminate low-intensity expressions as illustrated in Fig. 14. Additionally, Wang et al. [152] advanced the exploration of temporal expression dynamics by proposing a phase space reconstruction network to represent expression trajectories, while CEFLNet [153] introduced a clip-based feature encoder (CFE) with cascaded self-attention for spatio-temporal feature encoding.

5.4 Static to Dynamic FER

The static to dynamic FER utilized the high-performance SFER knowledge to explore appearance features and dynamic dependencies. The early work, such as Multi-channel Deep Spatial-Temporal feature Fusion neural Network (MDSTFN) [154] leverages pretrained deep CNNs for effective feature extraction and fusion in static images. Recently, Static-to-Dynamic model (S2D) [155] utilized existing SFER knowledge and dynamic information from facial landmark-aware features to enhance the performance of DFER. Specifically, the SFER model is first built with a Vision Transformer (ViT) and Multi-View Complementary Prompters (MCPs). The temporal-modeling adapters (TMAs) are then added to the DFER model. MCPs improve facial expression features with landmark-aware data, while TMAs capture and model dynamic facial expression changes, extending the pre-trained image model to video. Similarly, an affectivity extraction network (AEN) [156] integrated multi-level semantic features and emotion-guided loss functions to enhance sentiment and specific emotion classification, ensuring the preservation of emotional information across video sequences.

5.5 Multi-modal DFER

Inspired by the affective image content analysis (AICA) [158] comprehending the emotional impact of images ne-

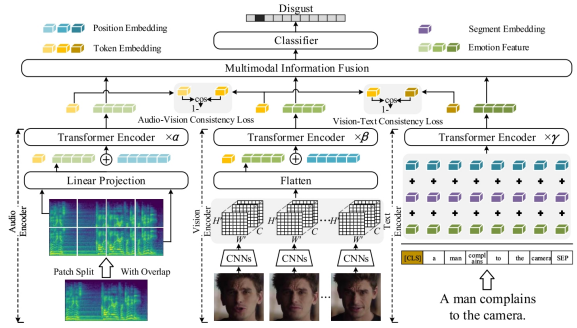

Fig. 15: The architecture of multi-modal DFER. Figure is reproduced based on the fusion-based model (T-MEP) [157].

cessitates the integration of visual features with contextual cues, multi-modal DFER works [37], [44], [159] tried to leverage contexual features and fused information to capture and analyze the dynamic changes in facial expressions.

5.5.1 Context-aware Models

The standard DFER approach involves segmenting the face region from video or image sequences to extract expression features and classify emotions, however often overlooks crucial contextual information which is important for DFER. To exploit a joint fusion of human facial expression and context information, the context-aware emotion recognition (CAERNet) [44] utilized two sub-networks to separately extract the features of face and context regions, and adaptive fusion networks to fuse such features in an adaptive fashion. Similarly, to tackle the rigid cognitive problem of DFER models which filter out environmental cues and body language, focusing only on facial information, the Overall Understanding of the Scene (OUS) [159] leveraged AudioCLIP to integrate scene and facial features.

5.5.2 Fusion-based Models

Fusion-based DFER integrates speech and text to enhance the accuracy and comprehensiveness by capturing auditory cues like tone and pitch, while text analysis provides emotional context. Liu et al. [37] built a multi-modal affective dataset (MAFW), and proposed a novel Transformerbased expression snippet feature learning method to enhance learning from both facial expressions and combined emotional states over time. Besides, the model structures of spatiotemporal neural network methods (such as CNN-LSTM and C3D-LSTM) [65], [69], [73] are used to extract and fusion multi-modal information with video, audio, and text. Recently, as shown in Fig. 15, Zhang et al. [157] introduced a Transformer-based Multimodal Emotional Perception (T-MEP) framework that integrates audio, image, and text sequences to bolster the robustness of expression recognition. By utilizing transformer-based encoders and a multimodal fusion module, T-MEP effectively synthesizes diverse emotional cues, resulting in enhanced performance in complex real-world environments.

5.6 Self-supervised DFER

The self-supervised DFER aims to learn useful representations from unlabeled video data, capturing the temporal dynamics and subtle variations in facial expressions. Specifically, Li et al. [162] proposed a twin-cycle autoencoder

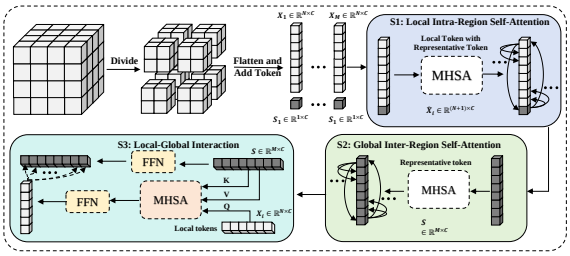

Fig. 16: The architecture of self-supervised DFER. This is reproduced based on the MAE-DFER [160].

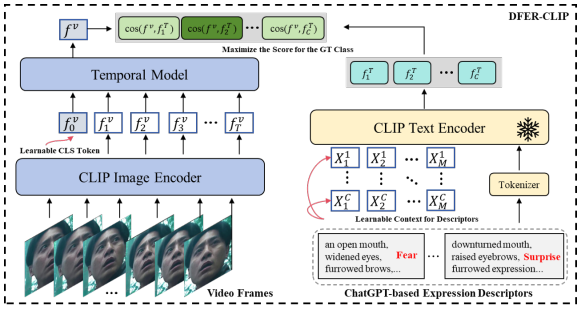

Fig. 17: The architecture of vision-language DFER. Figure is reproduced based on DFER-CLIP [161].

(TAE) to learn discriminative representations for facial actions from unlabeled videos. TAE disentangles facial actions from head motions by evaluating the quality of synthesized images, effectively capturing the subtle nuances of facial expressions. As shown in Fig. 16, the MAE-DFER [160] utilized large-scale unlabeled facial video data for self-supervised pre-training based on the masked autoencoders. The MAE-DFER incorporated an efficient local-global interaction Transformer (LGI-Former) as the encoder, and further integrated explicit temporal facial motion modeling alongside static appearance reconstruction.

5.7 Visual-Language DFER

The visual-language DFER can extract meaningful features from facial sequences and match them with corresponding textual descriptions, enabling a more nuanced understanding of emotional expressions. As shown in Fig. 17, the DFER-CLIP [161] integrated visual and textual components via CLIP-based model. The visual part employs a CLIP-based image encoder with a temporal model using Transformer encoders to extract temporal facial expression features, while the textual part uses large language models like ChatGPT to generate descriptive inputs, enhancing the accuracy of expression recognition by capturing contextual relationships. In contrast, CLIPER [163] enhanced interpretability by introducing Multiple Expression Text Descriptors (METD) to learn fine-grained expression representations. EmoCLIP [164] further extends this approach by incorporating contextual information from the environment surrounding facial expressions, enabling zero-shot classification of emotions.

6 Recent Advances of FER on Benchmark Datasets

We have reviewed the task challenges and network models for FER with static and dynamic emotions. Below we compared the performance of the image-based SFER **(Sec. 4)**

TABLE 3: Performance (WAR) of image-based SFER and video-based DFER methods on four in-the-lab datasets.

Method	Year	Type	Backbone	Datasets		
				MMI	CK+	Oulu-CASIA
IL-VGG [84]	2018	Static	VGG-16	74.68	91.64	84.58
FMPN [87]	2019	Static	CNNs	82.74	98.60	-
LDL-ALSG [110]	2020	Static	ResNet-50	70.03	93.08	63.94
IE-DBN [85]	2021	Static	VGG-16	-	96.02	85.21
im-cGAN [165]	2023	Static	GAN	-	98.10	93.34
Mul-DML [166]	2024	Static	ResNet-18	81.57	98.47	-
STC-NLSTM [145]	2018	Dynamic	3DCNN	84.53	99.80	93.45
SAANet [71]	2020	Dynamic	VGG-16	-	97.38	82.41
MGLN [146]	2020	Dynamic	VGG-16	-	98.77	90.40
MSDmodel [144]	2021	Dynamic	CNN	89.99	99.10	87.33
DPCNet [16]	2022	Dynamic	CNN	-	99.70	-
STACM [147]	2023	Dynamic	CNN	82.71	99.08	91.25

and video-based DFER methods **(Sec. 5)** in the lab or wild scenes, and summarized their recent advances.

6.1 Recent Advances of In-the-lab FER

Table 3 shows evaluations on four widely adopted in-the-lab datasets. Note it shows the best FER performance as FER pre-training often have different implementations. Three conclusions can be drawn from Table 3: 1) Significant progress has been made in image-based SFER and video-based DFER within laboratory environments. Due to the small scale, homogeneity, and high quality of datasets such as JAFFE, MMI, and CK+, advanced DL-based models have achieved recognition accuracies typically exceeding 95%; 2) Since Oulu-CASIA dataset contains videos under diverse illumination settings, all the models perform much less accurately (less than 90%) than they do on other datasets. This makes it particularly valuable for evaluating the robustness and generalization capabilities of FER models, providing a comprehensive testbed for assessing the impact of environmental variations on recognition accuracy; 3) In the widely used DFER datasets, have achieved near-perfect accuracy by effectively capturing both temporal and spatial information, with recognition rates reaching 99% and 90%, respectively. These results are notably higher than those obtained using datasets based on single expression frames, covering the importance of temporal dynamics in enhancing recognition performance.

As Table 3 shows, MSDmodel [144], DPCNet [16], and im-cGAN [165] achieve state-of-the-art performance on MMI, CK+, and Oulu-CASIA, reaching 89.99%, 99.70%, and 93.34%, respectively; the MSDmodel [144] performs well consistently across three datasets. While performance on these datasets is consistently high (often greater than 90%), the robustness and generalization of DL-based models in complex real-world scenarios remain further exploration.

6.2 Recent Advances of In-the-wild SFER

Table 4 shows results on three widely adopted in-the-wild datasets. Five conclusions can be drawn from Table 4: 1) **Significantly lower performance** (on average 20%) in the open environment compared to the image-based SFER in the controlled laboratory environment (Table 3); 2) **Substantial variability of best performances** of FER models is across different benchmark datasets, such as 50%-60% in SFEW, 80%-93% in RAF-DB, and 55%-67% in AffectNet; 3) **face occlusion and pose changes** often cause the critical information loss of facial part region information, hence obtaining available facial regions and effectively extracting critical facial expressive features are the main ways to overcome disturbance. The attention-based models [93], [94] often utilized patch or region attention CNNs to perceive occluded regions and capture salient affective interactions, however decomposition-based models [100] decompose facial expression from identity and posture, and generate discriminative facial expression features; 4) **Label and data uncertainty** mainly arises from inherent data ambiguity and subjective

TABLE 4: Performance (WAR) of image-based SFER methods on three in-the-wild datasets.

Task Challenges	Method	Year	Backbone	Datasets		
				SFEW	RAF-DB	AffectNet
General SFER **(Sec. 4.1)**	IFSL [86]	2020	VGG16	46.50	76.90	-
	OAENet [88]	2021	VGG16	-	86.50	58.70
	MA-Net [81]	2021	ResNet18	-	88.40	64.53
	D^3Net [91]	2021	ResNet18	62.16	88.79	-
	TransFER [89]	2021	ResNet50	-	90.91	66.23
	VTFF [167]	2023	Transformer	-	88.14	61.85
	HASs [82]	2023	ResNet50	65.14	91.04	-
	APViT [168]	2023	Transformer	61.92	91.98	66.91
	POSTER [83]	2023	CNN-IR50	-	92.05	67.31
	MGR^3Net [14]	2024	ResNet50	-	91.05	66.36
Disturbance -invariant SFER **(Sec. 4.2)**	PG-Unit [96]	2018	VGG16	-	83.27	55.33
	IDFL [94]	2021	ResNet50	-	86.96	59.20
	FDRL [102]	2021	ResNet18	62.16	89.47	-
	AMP-Net [90]	2022	ResNet50	-	88.06	63.23
	PACVT [93]	2023	ResNet18	-	88.21	60.68
	IPD-FER [100]	2023	ResNet18	58.43	88.89	-
	Latent-OFER [92]	2023	ResNet18	-	89.60	-
	RAC+RSL [109]	2023	ResNet18	-	89.77	62.16
Uncertainty -aware SFER **(Sec. 4.4)**	SCN [113]	2020	ResNet18	-	87.03	60.23
	DMUE [107]	2021	ResNet18	57.12	88.76	62.84
	RUL [114]	2021	ResNet18	-	88.98	-
	EASE [115]	2022	VGG16	60.12	89.56	61.82
	EAC [111]	2022	ResNet18	-	89.99	65.32
	LA-Net [108]	2023	ResNet18	-	91.56	64.54
	LNSU-Net [109]	2024	ResNet18	-	89.77	65.73
Weak -supervised SFER **(Sec. 4.7)**	Ada-CM [134]	2022	ResNet18	52.43	84.42	57.42
	E2E-WS [112]	2022	ResNet18	54.56	88.89	60.04
	DR-FER [169]	2023	ResNet50	-	90.53	66.85
	WSCFER [136]	2023	IResNet	-	91.72	67.71
Cross-modal SFER **(Sec. 4.8)**	CLEF [137]	2023	CLIP	-	90.09	65.66
	VTA-Net [138]	2024	ResNet-18	-	72.17	-
	CEPrompt [13]	2024	ViT-B/16	-	92.43	67.29

TABLE 5: Performance (Accuracy) of 3D SFER methods **(Sec. 4.3)** on BU-3DE and Bosphorus datasets

Method	Year	Backbone	Modality	Datasets	
				BU-3DE	Bosphorus
JPE-GAN [105]	2018	CNN	2D/	81.20/-	-/-
DA-CNN [170]	2019	ResNet50	-/3D	-/87.69	-/-
GAN-Int [98]	2021	VGGNet16	2D+3D/3D	88.47/83.20	-/-
FFNet-M [171]	2021	VGGNet16	2D+3D/3D	89.82/87.28	87.65/82.86
CMANet [172]	2022	VGGNet16	2D+3D/3D	90.24/84.03	89.36/81.25
DrFER [173]	2024	ResNet18	-/3D	-/89.15	-/86.77

TABLE 6: Performance (WAR) of cross-domain SFER methods **(Sec. 4.6)** on four widely-used datasets

Method	Year	Backbone	Source Dataset	Target Dataset			
				JAFFE	CK+	FER-2013	AffectNet
ECAN [130]	2022	ResNet50	RAF-DB	57.28	79.77	56.46	-
AGRA [121]	2022	ResNet50	RAF-DB	61.5	85.27	58.95	-
PASM [174]	2022	VGGNet16	RAF-DB	-	79.65	54.78	-
CWCST [175]	2023	VGGNet16	RAF-DB2.0	69.01	89.64	57.44	52.66
DMSRL [127]	2023	VGGNet16	RAF-DB2.0	69.48	91.26	56.16	50.94
CSRL [127]	2023	ResNet18	RAF-DB	66.67	88.37	55.53	-

judgment differences among annotators. By integrating the noise label learning [111] and noise-insensitive loss [109], uncertainty-aware SFER [115], [116] considers these uncertainty factors when recognizing facial expressions, not only classifying the expressions but also evaluating and handling the uncertainty of each classification result to improve the accuracy and reliability of FER models; 5) **Benefit of large-scale unlabeled data and pretrained models** can improve the accuracy of SFER by leveraging the facial priors knowledge learned from high-confidence predictions to label unlabeled data [134] or visual language pre-training (VLP) [139].

As Table 4 shows, HASs [82], CEprompt [13], and (WSCFER) [136] achieve state-of-the-art performance on SFEW, RAF-DB, and AffectNet, reaching 65.14%, 92.43%, and 67.71%, respectively; the CEprompt [13] performs well consistently across RAF-DB and AffectNet datasets. Besides, Table 5, Table 6 show the performance under the framework of the 3D FER and cross-domain FER, respectively. 3D SFER utilized GAN-based learning [103] and multi-view learning [99] to generate synthetic facial expression data with different changes and utilize multi-view images during training. Cross-domain inconsistency poses a significant challenge to the generalization of FER models, as data from controlled laboratory environments differ markedly from those in real-world applications. To address this issue, domain adaptation techniques (such as transfer learning and adversarial learning) [121], [125] are employed to align data from the source and target domains, thereby reducing inter-domain differences and enhancing model robustness.

6.3 Recent Advances of In-the-wild DFER

Table 7 shows results on 4 widely adopted in-the-wild DFER datasets. In the past three years, significant progress of DFER has been promoted especially after the release of data sets such as DFEW [15], FERV39k [36] and MAFW [37], which provide rich diversity and challenging data, covering a wider range of real-life scenarios. Five conclusions can be drawn from Table 7: 1) **Markedly reduced performance** (on average 30%) is observed in open environments compared to controlled laboratory settings (Table 3) for video-based DFER, highlighting the significant challenges of adapting to real-world conditions; 2) **Remarkable differences** of WAR/UAR performances of DFER models is across four benchmark datasets, such as 50%/47%∼56%/52% in AFEW, 56%/46%∼76%/66% in DFEW, 44%/32%∼54%/45% in FERV39k, and 43%/31%∼58%/46% in MAFW. Note the lowest performances appears in the FERV39k dataset due to the large-scale and multi-scene attributes in various real-life scenarios; 3) **Key Frame Extraction** crucial to the performance of the DFER, including selecting key frames by detecting changes in facial movements in the video [150]; and extracting key frames based on changes in facial action units [149]; 4) **Capturing expression intensity fluctuations** is pivotal for understanding the dynamic nature of expressions and enhancing the accuracy of DFER systems due to the inherently dynamic characteristic of facial expression intensities varying over time. Since the intensity often follows two patterns: a gradual shift from neutral to peak intensity and back, or an abrupt transition from peak to neutral, PGDN [151] and GCA-IAL [47] extracted features related to expression evolution and learn long-distance dependencies between frames, respectively. 5) **Leveraging the multi-modal information, large-scale unlabeled data, or pretrained models** significantly enhances DFER accuracy by utilizing facial priors acquired from the fusion of contextual features [157], the masked autoencoders pretraining on large-scale unlabeled facial video data [160], or visual language pre-training models (CLIP) [139]. As Table 7 shows, CLIPER [163], MMA-DFER [184], FineCLIPER [187], and UMBEnet [17] achieve the best performance (the average accuracy of WAR and UAR) on AFEW, DFEW, FERV39k, and MAFW reaching 52.22%, 72.3%, 49.6%, and 52.09%, respectively; the UMBEnet [17] performs well consistently across three large-scale datasets.

7 Applications and Ethical Issues of FER

In this section, we point out some of the applications and ethical issues of FER, which further promote technological innovation and protect individual rights and interests.

7.1 Applications of FER

7.1.1 Health and Psychological Counseling

The FER plays a pivotal role in monitoring emotional changes by analyzing users' facial expressions in real-time, providing timely psychological advice and alerts [26], [188]. This technology is increasingly integrated into smartwatches and mobile applications, which continuously monitor emotional states and offer psychological adjustments. These devices can detect signs of depression or stress, prompting users to manage their emotions and, when necessary, connect with professional counselors [189].

In mental health monitoring, mobile apps equipped with FER capabilities offer emotion tracking and analysis, helping users understand and manage their emotional states more effectively [189]. When abnormal emotions are detected, these apps can suggest relaxation techniques or direct users to seek professional help [190]. In psychotherapy, particularly cognitive behavioral therapy (CBT), FER enables therapists to monitor patients' emotional reactions in real-time, enhancing their understanding of patients' internal states and allowing for personalized treatment adjustments [189]. FER also aids in diagnosing psychological and neurological disorders, such as early detection of depression and Parkinson's disease, through the analysis of facial expressions and remote photoplethysmography [74], [191]. In special populations and scenarios, FER is valuable in understanding children's emotional states, particularly in addressing emotional disorders and behavioral issues [192].

TABLE 7: Performance (WAR/UAR) of video-based DFER methods on four widely-used datasets. TI: Time Interpolation; DS: Dynamic Sampling; GWS: Group-weighted Sampling. ∗: Tunable Param (M)

Task Challenges	Method	Year	Sample Strategies	Backbone	Comlexity (GFLOPs)	Datasets (WAR/UAR)			
						AFEW	DFEW	FERV39k	MAFW
General DFER (Sec. 5.1)	TFEN [176]	2021	TI	ResNet-18	-	-	56.60/45.57	-	-
	FormerDFER [46]	2021	DS	Transformer	9.1G	50.92/47.42	65.70/53.69	-	43.27/31.16
	EST [141]	2023	DS	ResNet-18	N/A	54.26/49.57	65.85/53.94	-	-
	LOGO-Former [68]	2023	DS	ResNet-18	10.27G	-	66.98/54.21	48.13/38.22	-
	MSCM [177]	2023	DS	ResNet-18	8.11G	56.40/52.30	70.16/58.49	-	-
	SFT [178]	2024	DS	ResNet-18	17.52G	55.00/50.14	-	47.80/35.16	47.44/33.39
	CDGT [179]	2024	DS	Transformer	8.3G	55.68/51.57	70.07/59.16	50.80/41.34	-
	LSGTNet [180]	2024	DS	ResNet-18	-	-	72.34/61.33	51.31/41.30	-
Sampling-based DFER (Sec. 5.2)	EC-STFL [15]	2020	TI	ResNet-18	8.32G	53.26/-	54.72/43.60	-	-
	DPCNet [16]	2022	GWS	ResNet-50	9.52G	51.67/47.86	66.32/57.11	-	-
	FreqHD [149]	2023	FreqHD	ResNet-18	-	-	54.98/44.24	43.93/32.24	-
	M3DFEL [181]	2023	DS	R3D18	1.66G	-	69.25/56.10	47.67/35.94	-
Expression Intensity-aware DFER (Sec. 5.3)	CEFL-Net [153]	2022	Clip-based	ResNet-18	-	53.98/-	65.35/-	-	-
	NR-DFERnet [182]	2023	DS	ResNet-18	6.33G	53.54/48.37	68.19/54.21	-	-
	GCA+IAL [47]	2023	DS	ResNet-18	9.63G	-	69.24/55.71	48.54/35.82	-
Static to Dynamic FER (Sec. 5.4)	S2D [155]	2023	DS	ViT-B/16	-	-	76.03/61.82	52.56/41.28	57.37/41.86
	(AEN) [156]	2023	DS	Transformer	-	54.64/50.88	69.37/56.66	47.88/38.18	-
Multi-modal DFER (Sec. 5.5)	T-ESFL [37]	2022	DS	Transformer	-	-	-	-	48.18/33.28
	T-MEP [183]	2023	DS	-	6G	52.96/50.22	68.85/57.16	-	52.85/39.37
	OUS [159]	2024	DS	CLIP	-	52.96/50.22	68.85/57.16	-	52.85/39.37
	MMA-DFER [184]	2024	DS	Transformer	-	-	77.51/67.01	-	58.52/44.11
Self-supervised DFER (Sec. 5.6)	MAE-DFER [160]	2023	DS	ResNet-18	50G	-	74.43/63.41	52.07/43.12	54.31/41.62
	HiCMAE [185]	2024	DS	ResNet-18	32G	-	73.10/61.92	-	54.84/42.10
Visual-Language DFER (Sec. 5.7)	CLIPER [163]	2023	DS	CLIP-ViT-B/16	88M*	56.43/52.00	70.84/57.56	51.34/41.23	-
	DFER-CLIP [161]	2023	DS	CLIP-ViT-B/32	92G	-	71.25/59.61	51.65/41.27	52.55/39.89
	EmoCLIP [164]	2024	DS	CLIP-ViT-B/32	-	-	62.12/58.04	36.18/31.41	41.46/34.24
	A^3lign-DFER [186]	2024	DS	CLIP-ViT-L/14	-	-	74.20/64.09	51.77/41.87	53.22/42.07
	UMBEnet [17]	2024	DS	CLIP	-	-	73.93/64.55	52.10/44.01	57.25/46.92
	FineCLIPER [187]	2024	DS	CLIP-ViT-B/16	20M*	-	76.21/65.98	53.98/45.22	56.91/45.01

It is also applied in assessing animal emotions, such as evaluating pain levels in horses [193], and in intensive care units, where FER can assess patient pain levels even with partial facial occlusion by analyzing facial AUs [194].

7.1.2 Personalized Education

Monitoring students' emotional states in the classroom allows teachers to adjust their teaching methods in real time, thereby enhancing educational effectiveness [195]. For instance, by analyzing students' facial expressions (such as confusion, boredom, or interest) in classroom or online education platforms, FER-enabled systems can dynamically adjust the difficulty of learning content based on students' emotional feedback or provide more detailed explanations and further materials [9]. Such a system can also facilitate timely adjustments to teaching strategies, such as adding interactive sessions and altering the teaching pace in response to emotional changes of students. Additionally, if the system identifies signs of depression or disengagement, it can prompt the teacher to offer personalized tutoring or encouraging feedback [196].

7.1.3 Human-Computer Interaction

FER technology holds significant potential in enhancing human-computer interaction and robotics by making interactions more natural and personalizing emotional feedback to improve user experience [197]. By integrating FER, social robots can recognize users' emotional states and adjust their conversational content and emotional expressions accordingly, offering assistance when users appear confused or sharing joyful topics when users are happy. This technology also enables virtual assistants to better perceive and respond to users' emotions, providing more personalized and contextually appropriate services. Additionally, FER can drive emotion-sensitive user interfaces [198] that adapt dynamically to users' emotional responses.

7.2 Ethical Issues

FER technologies offer vast applications but raise concerns regarding privacy, ethics, and security [199]. To ensure its responsible and sustainable development, interdisciplinary collaboration across psychology, ethics, and biology is essential [2]. Prolonged monitoring through FER systems can cause discomfort, anxiety, and stress, potentially leading to mental health issues and eroding trust, especially in public and work environments. Therefore, incorporating public opinion into the development process is crucial to align the technology with societal moral standards and public interest. Ethically, the research, development, and deployment of FER must be guided by a clear framework that prioritizes transparency, informed consent, and fairness. The decision-making processes of FER algorithms should be transparent and their outcomes explainable, ensuring equity across different races, genders, and age groups to prevent bias and

discrimination. Given that facial expression data is a form of biometric data, it requires stringent protection against leakage and misuse, adhering strictly to ethical principles to avoid infringing on individual rights. Addressing these issues ensures that the advancement of FER technology remains aligned with social progress, balancing innovation with ethical responsibility [200].

8 Development Trends

Facial Action Units (AUs) assisted FER is able to detect subtle differences that other models might overlook by emphasizing individual muscle movements [201], providing a detailed and objective analysis of facial expressions and improving understanding and reliability. Defined in the Facial Action Coding System (FACS) [202], AUs correspond to specific muscle movements, such as raising eyebrows or wrinkling the nose. Combining different AUs allows for detailed descriptions of facial expressions, improving the accuracy, robustness, and cultural adaptability of FER models. The adaptability extends to different cultural contexts, as facial expressions and their associated muscle movements are consistent across cultures [203]. Besides, detailed AUs [204] enhance the interpretability of FER models, allowing researchers to understand the influence of specific facial movements on emotion recognition. This leads to models that not only accurately predict emotions but also provide clear explanations, fostering transparency and trust.

Zero-shot FER aims to identify emotions that the model has not encountered during training [205], offering a solution when it is impractical to collect and annotate data for every possible expression. Traditional FER models, which rely heavily on large, manually labeled datasets, struggle to predict new emotion categories beyond their training data. This limitation undermines their effectiveness in real-world, dynamic scenarios where humans can express thousands of emotions [4]. The visual language models [139] can learn robust visual features and integrate them with natural language, enabling superior zero-shot recognition guided by semantic knowledge. Leveraging pre-trained visual language models like EmoCLIP [164], the zero-shot FER systems use unified semantic feature learning to obtain visual and linguistic representations that generalize to unseen emotion classes, which can recognize and respond to a broader range of emotional expressions in diverse contexts.

Multi-modal Emotion Recognition systems aim integrating multiple channels, including facial expressions, vocal tone, gestures, posture, and physiological signals [206], to enhance accuracy and robustness, mirroring the human ability to perceive emotions using multiple cues. It offers a more comprehensive understanding of emotional states by capturing the full spectrum of human emotions. Single-modality systems, like those focused solely on facial expressions, can miss critical information from other channels; for instance, a smile combined with a shaky voice might indicate nervousness rather than happiness. Multi-modal systems can disambiguate such signals and provide a more nuanced understanding [207]. Multimodal Large Language Models (MLLMs) [208] have introduced new possibilities for emotion recognition by aligning, pre-training, and fine-tuning multiple modalities, enabling them to understand emotions and perform zero-shot emotion recognition, demonstrating significant potential in this field.

Embodied FER system is essential in modern human-computer interaction [209], integrating FER models with interactive technologies to achieve real-time detection and response to human emotions. These systems utilize computer vision and representation learning to analyze facial expressions, language, voice, and posture, significantly enhancing user experience and engagement. Compared to traditional camera-based FER, Embodied FER systems [210] face the challenge of managing dynamic, multi-perspective views and adapting to environmental variations such as lighting changes, occlusions, and motion blur in complex settings. Additionally, the need for real-time, contextually appropriate feedback during close interactions demands greater robustness, adaptability, and computational efficiency. Future research will focus on improving system performance across diverse facial morphologies and environments, and advancing the integration of multimodal methods (e.g., fusing facial expression with voice and body language) to further develop embodied FER.

Embodied Facial Expression Generation is crucial for enabling robots, particularly humanoid robots, to engage with humans in a direct and compelling manner by accurately mimicking facial expressions [210]. It can be categorized into two primary forms: AIGC-based expression generation and physical embodiment through motor-driven mechanisms. AIGC-based facial expression generation [211], [212] utilizes generative models, which deeply learn from vast datasets, to automatically create virtual facial expressions, which allows for a wide range of emotional expressions, contributing to more vivid and controllable interactions in dynamic environments. Physical embodiment using motor-driven mechanisms involves the movement of components such as eyes, mouth, and neck to produce facial expressions, enhancing the realism of interactions through tangible physical presence [213], [214]. Future research will focus on addressing the challenges associated with these methods by advancing the realism and cultural sensitivity of AIGC-generated expressions and enhancing the hardware capabilities of motor-driven systems to support more expressive and responsive facial movements.

9 Conclusion

Facial expression recognition (FER) has gained significant attention within the AI community, with promising applications in human-machine collaboration and embodied intelligence. This survey extensively reviews FER works from several perspectives, including background, datasets, generic workflow, challenge-oriented taxonomy of state-of-the-art methods, recent advances, applications, ethical concerns, and emerging trends. We systematically compare and summarize FER datasets, task challenges, methods, and performance evaluations through tables and figures, providing a clear overview of the latest advancements in FER. This comprehensive analysis greatly benefits researchers from various disciplines by enabling them to swiftly understand the challenges and progress in the field, thereby fostering collaboration toward the development of general FER.

10 Acknowledgments

This work was supported by National Natural Science Foundation of China (No.62406075), National Key Research and Development Program of China (2023YFC3604802), and by China Postdoctoral Science Foundation under Grant (2023M730647, 2023TQ0075).

References

[1] S. Zhao, X. Hong, J. Yang, Y. Zhao, and G. Ding, "Toward label-efficient emotion and sentiment analysis," *Proceedings of the IEEE*, vol. 111, no. 10, pp. 1159–1197, 2023.

[2] J. Z. Wang, S. Zhao, C. Wu, R. B. Adams, M. G. Newman, T. Shafir, and R. Tsachor, "Unlocking the emotional world of visual media: An overview of the science, research, and impact of understanding emotion drawing insights from psychology, engineering, and the arts." *Proceedings of the IEEE*, vol. 111, no. 10, pp. 1236–1286, 2023.

[3] E. G. Krumhuber, L. I. Skora, H. C. Hill, and K. Lander, "The role of facial movements in emotion recognition," *Nature Reviews Psychology*, vol. 2, no. 5, pp. 283–296, 2023.

[4] A. S. Cowen, D. Keltner, F. Schroff, B. Jou, H. Adam, and G. Prasad, "Sixteen facial expressions occur in similar contexts worldwide," *Nature*, vol. 589, no. 7841, pp. 251–257, 2021.

[5] S. Zhang, Y. Pan, and J. Z. Wang, "Learning emotion representations from verbal and nonverbal communication," in *CVPR*, 2023, pp. 18 993–19 004.

[6] Y. Wang, W. Song, W. Tao, A. Liotta, D. Yang, X. Li, S. Gao, Y. Sun, W. Ge, W. Zhang, and W. Zhang, "A systematic review on affective computing: Emotion models, databases, and recent advances," *Information Fusion*, vol. 83, pp. 19–52, 2022.

[7] P. Ekman and H. Oster, "Facial expressions of emotion," *Annual review of psychology*, vol. 30, no. 1, pp. 527–554, 1979.

[8] O. Rudovic, J. Lee, M. Dai, B. Schuller, and R. W. Picard, "Personalized machine learning for robot perception of affect and engagement in autism therapy," *Science Robotics*, vol. 3, no. 19, p. eaao6760, 2018.

[9] A. V. Savchenko, L. V. Savchenko, and I. Makarov, "Classifying emotions and engagement in online learning based on a single facial expression recognition neural network," *IEEE TAFFC*, vol. 13, no. 4, pp. 2132–2143, 2022.

[10] S. Zhao, H. Yao, Y. Gao, G. Ding, and T.-S. Chua, "Predicting personalized image emotion perceptions in social networks," *IEEE TAFFC*, vol. 9, no. 4, pp. 526–540, 2018.

[11] S. Wang, Z. Zheng, S. Yin, J. Yang, and Q. Ji, "A novel dynamic model capturing spatial and temporal patterns for facial expression analysis," *IEEE TPAMI*, vol. 42, no. 9, pp. 2082–2095, 2019.

[12] D. Liu, W. Dai, H. Zhang, X. Jin, J. Cao, and W. Kong, "Brain-machine coupled learning method for facial emotion recognition," *IEEE TPAMI*, vol. 45, no. 9, pp. 10 703–10 717, 2023.

[13] H. Zhou, S. Huang, F. Zhang, and C. Xu, "Ceprompt: Cross-modal emotion-aware prompting for facial expression recognition," *IEEE TCSVT*, pp. 1–1, 2024.

[14] Y. Wang, S. Yan, W. Song, A. Liotta, J. Liu, D. Yang, S. Gao, and W. Zhang, "Mgr3net: Multigranularity region relation representation network for facial expression recognition in affective robots," *IEEE TII*, 2024.

[15] X. Jiang, Y. Zong, W. Zheng, C. Tang, W. Xia, C. Lu, and J. Liu, "Dfew: A large-scale database for recognizing dynamic facial expressions in the wild," in *ACM MM*, 2020, pp. 2881–2889.

[16] Y. Wang, Y. Sun, W. Song, S. Gao, Y. Huang, Z. Chen, W. Ge, and W. Zhang, "Dpcnet: Dual path multi-excitation collaborative network for facial expression representation learning in videos," in *ACM MM*, 2022, pp. 101–110.

[17] X. Mai, J. Lin, H. Wang, Z. Tao, Y. Wang, S. Yan, X. Tong, J. Yu, B. Wang, Z. Zhou, Q. Zhao, S. Gao, and W. Zhang, "All rivers run into the sea: Unified modality brain-inspired emotional central mechanism," in *ACM MM*, 2024.

[18] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE TAFFC*, vol. 13, no. 3, pp. 1195–1215, 2020.

[19] F. Z. Canal, T. R. Müller, J. C. Matias, G. G. Scotton, A. R. de Sa Junior, E. Pozzebon, and A. C. Sobieranski, "A survey on facial emotion recognition techniques: A state-of-the-art literature review," *Information Sciences*, vol. 582, pp. 593–617, 2022.

[20] M. Karnati, A. Seal, D. Bhattacharjee, A. Yazidi, and O. Krejcar, "Understanding deep learning techniques for recognition of human emotions using facial expressions: A comprehensive survey," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–31, 2023.

[21] S. C. Leong, Y. M. Tang, C. H. Lai, and C. Lee, "Facial expression and body gesture emotion recognition: A systematic review on the use of visual data in affective computing," *Computer Science Review*, vol. 48, p. 100545, 2023.

[22] G. R. Alexandre, J. M. Soares, and G. A. P. Thé, "Systematic review of 3d facial expression recognition methods," *PR*, vol. 100, p. 107108, 2020.

[23] M. Jampour and M. Javidi, "Multiview facial expression recognition, a survey," *IEEE TAFFC*, vol. 13, no. 4, pp. 2086–2105, 2022.

[24] Y. Liu, X. Zhang, Y. Li, J. Zhou, X. Li, and G. Zhao, "Graph-based facial affect analysis: A review," *IEEE TAFFC*, vol. 14, no. 4, pp. 2657–2677, 2023.

[25] I. Dominguez-Catena, D. Paternain, and M. Galar, "Metrics for dataset demographic bias: A case study on facial expression recognition," *IEEE TPAMI*, pp. 1–18, 2024.

[26] T. Hassan, D. Seuß, J. Wollenberg, K. Weitz, M. Kunz, S. Lautenbacher, J.-U. Garbas, and U. Schmid, "Automatic detection of pain from facial expressions: a survey," *IEEE TPAMI*, vol. 43, no. 6, pp. 1815–1831, 2019.

[27] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with gabor wavelets," in *IEEE FG*, 1998, pp. 200–205.

[28] 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 *CVPRW*, 2010, pp. 94–101.

[29] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark," in *ICCVW*, 2011, pp. 2106–2112.

[30] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "From facial expression recognition to interpersonal relation prediction," *IJCV*, vol. 126, pp. 550–569, 2018.

[31] S. Li, W. Deng, and J. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in *CVPR*, 2017, pp. 2852–2861.

[32] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," *IEEE TAFFC*, vol. 10, no. 1, pp. 18–31, 2017.

[33] C. Fabian 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 *CVPR*, 2016, pp. 5562–5570.

[34] S. Cheng, I. Kotsia, M. Pantic, and S. Zafeiriou, "4dfab: A large scale 4d database for facial expression analysis and biometric applications," in *CVPR*, 2018, pp. 5117–5126.

[35] 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.

[36] Y. Wang, Y. Sun, Y. Huang, Z. Liu, S. Gao, W. Zhang, W. Ge, and W. Zhang, "Ferv39k: A large-scale multi-scene dataset for facial expression recognition in videos," in *CVPR*, 2022, pp. 20 922–20 931.

[37] Y. Liu, W. Dai, C. Feng, W. Wang, G. Yin, J. Zeng, and S. Shan, "Mafw: A large-scale, multi-modal, compound affective database for dynamic facial expression recognition in the wild," in *ACM MM*, 2022, pp. 24–32.

[38] M. Valstar, M. Pantic *et al.*, "Induced disgust, happiness and surprise: an addition to the mmi facial expression database," in *Proc. 3rd Intern. Workshop on EMOTION*, vol. 10, 2010, p. 65.

[39] 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.

[40] 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 *Neural Information Processing*. Springer, 2013, pp. 117–124.

[41] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato, "A 3d facial expression database for facial behavior research," in *IEEE FG*, 2006, pp. 211–216.

[42] A. Savran, N. Alyüz, H. Dibeklioğlu, O. Çeliktutan, B. Gökberk, B. Sankur, and L. Akarun, "Bosphorus database for 3d face analysis," in *Biometrics and Identity Management: First European Workshop, BIOID 2008, Roskilde, Denmark, May 7-9, 2008. Revised Selected Papers 1*. Springer, 2008, pp. 47–56.

[43] A. Dhall, A. Kaur, R. Goecke, and T. Gedeon, "Emotiw 2018: Audio-video, student engagement and group-level affect prediction," in *ACM ICMI*, 2018, pp. 653–656.

[44] J. Lee, S. Kim, S. Kim, J. Park, and K. Sohn, "Context-aware emotion recognition networks," in *ICCV*, 2019, pp. 10 143–10 152.

[45] X. Ben, Y. Ren, J. Zhang, S.-J. Wang, K. Kpalma, W. Meng, and Y.-J. Liu, "Video-based facial micro-expression analysis: A survey of datasets, features and algorithms," *IEEE TPAMI*, vol. 44, no. 9, pp. 5826–5846, 2021.

[46] Z. Zhao and Q. Liu, "Former-dfer: Dynamic facial expression recognition transformer," in *ACM MM*, 2021, pp. 1553–1561.

[47] H. Li, H. Niu, Z. Zhu, and F. Zhao, "Intensity-aware loss for dynamic facial expression recognition in the wild," in *AAAI*, vol. 37, no. 1, 2023, pp. 67–75.

[48] R. Ranjan, V. M. Patel, and R. Chellappa, "Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition," *IEEE TPAMI*, vol. 41, no. 1, pp. 121–135, 2017.

[49] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *CVPR*, 2013, pp. 532–539.

[50] Y. Sun, X. Wang, and X. Tang, "Deep convolutional network cascade for facial point detection," in *CVPR*, 2013, pp. 3476–3483.

[51] Y. Wu, T. Hassner, K. Kim, G. Medioni, and P. Natarajan, "Facial landmark detection with tweaked convolutional neural networks," *IEEE TPAMI*, vol. 40, no. 12, pp. 3067–3074, 2017.

[52] P. Kar, V. M. Chudasama, N. Onoe, P. Wasnik, and V. Balasubramanian, "Fiducial focus augmentation for facial landmark detection," in *The 34th British Machine Vision Conference*. BMVA, 2023.

[53] Z. Zhou, H. Li, H. Liu, N. Wang, G. Yu, and R. Ji, "Star loss: Reducing semantic ambiguity in facial landmark detection," in *CVPR*, 2023, pp. 15 475–15 484.

[54] Y. Hu, M. Lu, C. Xie, and X. Lu, "Fin-gan: Face illumination normalization via retinex-based self-supervised learning and conditional generative adversarial network," *Neurocomputing*, vol. 456, pp. 109–125, 2021.

[55] L. Tran, X. Yin, and X. Liu, "Disentangled representation learning gan for pose-invariant face recognition," in *CVPR*, 2017, pp. 1415–1424.

[56] W. Ma, X. Xie, C. Yin, and J. Lai, "Face image illumination processing based on generative adversarial nets," in *ICPR*. IEEE, 2018, pp. 2558–2563.

[57] X. Han, H. Yang, G. Xing, and Y. Liu, "Asymmetric joint gans for normalizing face illumination from a single image," *IEEE Transactions on Multimedia*, vol. 22, no. 6, pp. 1619–1633, 2019.

[58] L. Tran, X. Yin, and X. Liu, "Representation learning by rotating your faces," *IEEE TPAMI*, vol. 41, no. 12, pp. 3007–3021, 2018.

[59] S. Tripathy, J. Kannala, and E. Rahtu, "Icface: Interpretable and controllable face reenactment using gans," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2020, pp. 3385–3394.

[60] F. Zhang, T. Zhang, Q. Mao, and C. Xu, "Geometry guided pose-invariant facial expression recognition," *IEEE TIP*, vol. 29, pp. 4445–4460, 2020.

[61] 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," *PR*, vol. 61, pp. 610–628, 2017.

[62] S. Sahu, R. Gupta, and C. Espy-Wilson, "Modeling feature representations for affective speech using generative adversarial networks," *IEEE TAFFC*, vol. 13, no. 2, pp. 1098–1110, 2020.

[63] H. Li, N. Wang, X. Ding, X. Yang, and X. Gao, "Adaptively learning facial expression representation via cf labels and distillation," *IEEE TIP*, vol. 30, pp. 2016–2028, 2021.

[64] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations (ICLR 2015)*, 2015, pp. 1–14.

[65] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[66] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *ICLR*, 2021, arXiv: 2010.11929.

[67] Y. Li, H. Sun, Z. Liu, H. Han, and S. Shan, "Affective behaviour analysis using pretrained model with facial prior," in *European Conference on Computer Vision*. Springer, 2022, pp. 19–30.

[68] F. Ma, B. Sun, and S. Li, "Logo-former: Local-global spatio-temporal transformer for dynamic facial expression recognition," in *ICASSP*, 2023, pp. 1–5.

[69] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, p. 1735–1780, nov 1997.

[70] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *EMNLP*, 2014, pp. 1724–1734.

[71] D. Liu, X. Ouyang, S. Xu, P. Zhou, K. He, and S. Wen, "Saanet: Siamese action-units attention network for improving dynamic facial expression recognition," *Neurocomputing*, vol. 413, pp. 145–157, 2020.

[72] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE TPAMI*, vol. 35, no. 1, pp. 221–231, 2012.

[73] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *ICCV*, 2015, pp. 4489–4497.

[74] C. Á. Casado, M. L. Cañellas, and M. B. López, "Depression recognition using remote photoplethysmography from facial videos," *IEEE TAFFC*, vol. 14, no. 4, pp. 3305–3316, 2023.

[75] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *CVPR*, 2016, pp. 2818–2826.

[76] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[77] A. Azazi, S. L. Lutfi, I. Venkat, and F. Fernández-Martínez, "Towards a robust affect recognition: Automatic facial expression recognition in 3d faces," *Expert Systems with Applications*, vol. 42, no. 6, pp. 3056–3066, 2015.

[78] M. S. Akhtar, A. Ekbal, and E. Cambria, "How intense are you? predicting intensities of emotions and sentiments using stacked ensemble," *IEEE CIM*, vol. 15, no. 1, pp. 64–75, 2020.

[79] X. Jin, Z. Lai, and Z. Jin, "Learning dynamic relationships for facial expression recognition based on graph convolutional network," *IEEE TIP*, vol. 30, pp. 7143–7155, 2021.

[80] T. Kaneko, K. Hiramatsu, and K. Kashino, "Generative attribute controller with conditional filtered generative adversarial networks," in *CVPR*, 2017, pp. 6089–6098.

[81] Z. Zhao, Q. Liu, and S. Wang, "Learning deep global multi-scale and local attention features for facial expression recognition in the wild," *IEEE TIP*, vol. 30, pp. 6544–6556, 2021.

[82] T. Liu, J. Li, J. Wu, B. Du, J. Chang, and Y. Liu, "Facial expression recognition on the high aggregation subgraphs," *IEEE TIP*, vol. 32, pp. 3732–3745, 2023.

[83] C. Zheng, M. Mendieta, and C. Chen, "Poster: A pyramid cross-fusion transformer network for facial expression recognition," in *ICCVW*, 2023, pp. 3138–3147.

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

[85] H. Zhang, W. Su, J. Yu, and Z. Wang, "Identity–expression dual branch network for facial expression recognition," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 13, no. 4, pp. 898–911, 2021.

[86] Y. Yan, Z. Zhang, S. Chen, and H. Wang, "Low-resolution facial expression recognition: A filter learning perspective," *Signal Processing*, vol. 169, p. 107370, 2020.

[87] Y. Chen, J. Wang, S. Chen, Z. Shi, and J. Cai, "Facial motion prior networks for facial expression recognition," in *VCIP*, 2019, pp. 1–4.

[88] Z. Wang, F. Zeng, S. Liu, and B. Zeng, "Oaenet: Oriented attention ensemble for accurate facial expression recognition," *PR*, vol. 112, p. 107694, 2021.

[89] F. Xue, Q. Wang, and G. Guo, "Transfer: Learning relation-aware facial expression representations with transformers," in *ICCV*, 2021, pp. 3601–3610.

[90] H. Liu, H. Cai, Q. Lin, X. Li, and H. Xiao, "Adaptive multilayer perceptual attention network for facial expression recognition," *IEEE TCSVT*, vol. 32, no. 9, pp. 6253–6266, 2022.

[91] R. Mo, Y. Yan, J.-H. Xue, S. Chen, and H. Wang, "D^3net: Dual-branch disturbance disentangling network for facial expression recognition," in *ACM MM*, 2021.

[92] I. Lee, E. Lee, and S. B. Yoo, "Latent-ofer: Detect, mask, and reconstruct with latent vectors for occluded facial expression recognition," in *ICCV*, October 2023, pp. 1536–1546.

[93] C. Liu, K. Hirota, and Y. Dai, "Patch attention convolutional vision transformer for facial expression recognition with occlusion," *Information Sciences*, vol. 619, pp. 781–794, 2023.

[94] Y. Li, Y. Lu, B. Chen, Z. Zhang, J. Li, G. Lu, and D. Zhang, "Learning informative and discriminative features for facial expression recognition in the wild," *IEEE TCSVT*, vol. 32, no. 5, pp. 3178–3189, 2022.

[95] Y. Xia, H. Yu, X. Wang, M. Jian, and F.-Y. Wang, "Relation-aware facial expression recognition," *IEEE TCDS*, vol. 14, no. 3, pp. 1143–1154, 2022.

[96] Y. Li, J. Zeng, S. Shan, and X. Chen, "Occlusion aware facial expression recognition using cnn with attention mechanism," *IEEE TIP*, vol. 28, no. 5, pp. 2439–2450, 2018.

[97] S. Wang, Y. Yuan, X. Zheng, and X. Lu, "Local and correlation attention learning for subtle facial expression recognition," *Neurocomputing*, vol. 453, pp. 742–753, 2021.

[98] H. Yang, K. Zhu, D. Huang, H. Li, Y. Wang, and L. Chen, "Intensity enhancement via gan for multimodal face expression recognition," *Neurocomputing*, vol. 454, pp. 124–134, 2021.

[99] Q. N. Vo, K. Tran, and G. Zhao, "3d facial expression recognition based on multi-view and prior knowledge fusion," in *IEEE MMSP*, 2019, pp. 1–6.

[100] J. Jiang and W. Deng, "Disentangling identity and pose for facial expression recognition," *IEEE TAFFC*, vol. 13, no. 4, pp. 1868–1878, 2022.

[101] D. Ruan, R. Mo, Y. Yan, S. Chen, J.-H. Xue, and H. Wang, "Adaptive deep disturbance-disentangled learning for facial expression recognition," *IJCV*, vol. 130, no. 2, p. 455–477, 2022.

[102] D. Ruan, Y. Yan, S. Lai, Z. Chai, C. Shen, and H. Wang, "Feature decomposition and reconstruction learning for effective facial expression recognition," in *CVPR*, 2021, pp. 7660–7669.

[103] J. Zhu, B. Luo, A. Sun, J. Tan, X. Zhao, and Y. Gao, "Variance-aware bi-attention expression transformer for open-set facial expression recognition in the wild," in *ACM MM*, 2023, pp. 862–870.

[104] Y. Liu, J. Peng, W. Dai, J. Zeng, and S. Shan, "Joint spatial and scale attention network for multi-view facial expression recognition," *PR*, vol. 139, p. 109496, 2023.

[105] F. Zhang, T. Zhang, Q. Mao, and C. Xu, "Joint pose and expression modeling for facial expression recognition," in *CVPR*, 2018, pp. 3359–3368.

[106] Z. Xing, W. Tan, R. He, Y. Lin, and B. Yan, "Co-completion for occluded facial expression recognition," in *ACM MM*, 2022, p. 130–140.

[107] J. She, Y. Hu, H. Shi, J. Wang, Q. Shen, and T. Mei, "Dive into ambiguity: Latent distribution mining and pairwise uncertainty estimation for facial expression recognition," in *CVPR*, 2021, pp. 6248–6257.

[108] Z. Wu and J. Cui, "La-net: Landmark-aware learning for reliable facial expression recognition under label noise," in *ICCV*, 2023, pp. 20 698–20 707.

[109] Y. Zhang, Y. Li, X. Liu, W. Deng *et al.*, "Leave no stone unturned: mine extra knowledge for imbalanced facial expression recognition," *NeurIPS*, 2024.

[110] S. Chen, J. Wang, Y. Chen, Z. Shi, X. Geng, and Y. Rui, "Label distribution learning on auxiliary label space graphs for facial expression recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 13 984–13 993.

[111] Y. Zhang, C. Wang, X. Ling, and W. Deng, "Learn from all: Erasing attention consistency for noisy label facial expression recognition," in *ECCV*, 2022, pp. 418–434.

[112] F. Zhang, M. Xu, and C. Xu, "Weakly-supervised facial expression recognition in the wild with noisy data," *IEEE TMM*, vol. 24, pp. 1800–1814, 2022.

[113] K. Wang, X. Peng, J. Yang, S. Lu, and Y. Qiao, "Suppressing uncertainties for large-scale facial expression recognition," in *CVPR*, 2020, pp. 6897–6906.

[114] Y. Zhang, C. Wang, and W. Deng, "Relative uncertainty learning for facial expression recognition," *Advances in Neural Information Processing Systems*, vol. 34, pp. 17 616–17 627, 2021.

[115] L. Wang, G. Jia, N. Jiang, H. Wu, and J. Yang, "Ease: Robust facial expression recognition via emotion ambiguity-sensitive cooperative networks," in *ACM MM*, 2022, pp. 218–227.

[116] Y. Liu, X. Zhang, J. Kauttonen, and G. Zhao, "Uncertain facial expression recognition via multi-task assisted correction," *IEEE TMM*, 2023.

[117] Y. Li, Y. Gao, B. Chen, Z. Zhang, G. Lu, and D. Zhang, "Self-supervised exclusive-inclusive interactive learning for multi-label facial expression recognition in the wild," *IEEE TCSVT*, vol. 32, no. 5, pp. 3190–3202, 2022.

[118] J. Jiang and W. Deng, "Improving multi-label facial expression recognition with consistent and distinct attentions," *IEEE TAFFC*, 2023.

[119] J. Jiang, M. Wang, B. Xiao, J. Hu, and W. Deng, "Joint recognition of basic and compound facial expressions by mining latent soft labels," *PR*, vol. 148, p. 110173, 2024.

[120] D. Chen, G. Wen, P. Wen, P. Yang, R. Chen, and C. Li, "Cross-domain sample relationship learning for facial expression recognition," *IEEE Transactions on Multimedia*, 2023.

[121] T. Chen, T. Pu, H. Wu, Y. Xie, L. Liu, and L. Lin, "Cross-domain facial expression recognition: A unified evaluation benchmark and adversarial graph learning," *IEEE TPAMI*, vol. 44, no. 12, pp. 9887–9903, 2022.

[122] J. Wang, C. Lan, C. Liu, Y. Ouyang, T. Qin, W. Lu, Y. Chen, W. Zeng, and P. S. Yu, "Generalizing to unseen domains: A survey on domain generalization," *IEEE TKDE*, vol. 35, no. 8, pp. 8052–8072, 2023.

[123] W. Zhang, P. Song, and W. Zheng, "Joint local-global discriminative subspace transfer learning for facial expression recognition," *IEEE TAFFC*, vol. 14, no. 3, pp. 2484–2495, 2023.

[124] W. Zheng, Y. Zong, X. Zhou, and M. Xin, "Cross-domain color facial expression recognition using transductive transfer subspace learning," *IEEE TAFFC*, vol. 9, no. 1, pp. 21–37, 2018.

[125] R. Wang, P. Song, S. Li, L. Ji, and W. Zheng, "Common latent embedding space for cross-domain facial expression recognition," *IEEE Transactions on Computational Social Systems*, vol. 11, no. 2, pp. 2046–2056, 2024.

[126] D. Chen, P. Song, and W. Zheng, "Learning transferable sparse representations for cross-corpus facial expression recognition," *IEEE TAFFC*, vol. 14, no. 2, pp. 1322–1333, 2023.

[127] Y. Li, Z. Zhang, B. Chen, G. Lu, and D. Zhang, "Deep margin-sensitive representation learning for cross-domain facial expression recognition," *IEEE TMM*, vol. 25, pp. 1359–1373, 2023.

[128] S. Zhao, X. Yue, S. Zhang, B. Li, H. Zhao, B. Wu, R. Krishna, J. E. Gonzalez, A. L. Sangiovanni-Vincentelli, S. A. Seshia, and K. Keutzer, "A review of single-source deep unsupervised visual domain adaptation," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 2, pp. 473–493, 2022.

[129] S. Zhao, B. Li, P. Xu, X. Yue, G. Ding, and K. Keutzer, "MADAN: Multi-source Adversarial Domain Aggregation Network for Domain Adaptation," *IJCV*, vol. 129, no. 8, pp. 2399–2424, 2021.

[130] S. Li and W. Deng, "A deeper look at facial expression dataset bias," *IEEE TAFFC*, vol. 13, no. 2, pp. 881–893, 2022.

[131] H. Gao, M. Wu, Z. Chen, Y. Li, X. Wang, S. An, J. Li, and C. Liu, "Ssa-icl: Multi-domain adaptive attention with intra-dataset continual learning for facial expression recognition," *Neural Networks*, vol. 158, pp. 228–238, 2023.

[132] R. Wang, P. Song, and W. Zheng, "Graph-diffusion-based domain-invariant representation learning for cross-domain facial expression recognition," *IEEE Transactions on Computational Social Systems*, 2024.

[133] H. Zhang, W. Su, J. Yu, and Z. Wang, "Weakly supervised local-global relation network for facial expression recognition," in *IJCAI*, 2020, pp. 1040–1046.

[134] H. Li, N. Wang, X. Yang, X. Wang, and X. Gao, "Towards semi-supervised deep facial expression recognition with an adaptive confidence margin," in *CVPR*, 2022, pp. 4166–4175.

[135] Y. Shu, X. Gu, G.-Z. Yang, and B. P. L. Lo, "Revisiting self-supervised contrastive learning for facial expression recognition," in *BMVC*, 2022.

[136] W. Nie, B. Chen, W. Wu, X. Xu, W. Ren, and H. Liu, "Wscfer: Improving facial expression representations by weak supervised contrastive learning," in *IROS*, 2023, pp. 9816–9823.

[137] X. Zhang, T. Wang, X. Li, H. Yang, and L. Yin, "Weakly-supervised text-driven contrastive learning for facial behavior understanding," in *ICCV*, 2023, pp. 20 751–20 762.

[138] Y. Lv, G. Huang, Y. Yan, J.-H. Xue, S. Chen, and H. Wang, "Visual-textual attribute learning for class-incremental facial expression recognition," *IEEE Transactions on Multimedia*, 2024.

[139] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *ICML*, 2021, pp. 8748–8763.

[140] Y. Yuan, J. Zeng, and S. Shan, "Describe your facial expressions by linking image encoders and large language models," in *BMVC*, 2023, p. 15.

[141] Y. Liu, W. Wang, C. Feng, H. Zhang, Z. Chen, and Y. Zhan, "Expression snippet transformer for robust video-based facial expression recognition," *PR*, vol. 138, p. 109368, 2023.

[142] R. Zhao, T. Liu, Z. Huang, D. P. Lun, and K.-M. Lam, "Spatial-temporal graphs plus transformers for geometry-guided facial expression recognition," *IEEE TAFFC*, vol. 14, no. 4, pp. 2751–2767, 2023.

[143] N. Otberdout, M. Daoudi, A. Kacem, L. Ballihi, and S. Berretti, "Dynamic facial expression generation on hilbert hypersphere with conditional wasserstein generative adversarial nets," *IEEE TPAMI*, vol. 44, no. 2, pp. 848–863, 2020.

[144] X. Sun, P. Xia, and F. Ren, "Multi-attention based deep neural network with hybrid features for dynamic sequential facial expression recognition," *Neurocomputing*, vol. 444, pp. 378–389, 2021.

[145] 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.

[146] M. Yu, H. Zheng, Z. Peng, J. Dong, and H. Du, "Facial expression recognition based on a multi-task global-local network," *Pattern Recognition Letters*, vol. 131, pp. 166–171, 2020.

[147] W. Chen, D. Zhang, M. Li, and D.-J. Lee, "Stcam: Spatial-temporal and channel attention module for dynamic facial expression recognition," *IEEE TAFFC*, vol. 14, no. 1, pp. 800–810, 2023.

[148] D. Poux, B. Allaert, N. Ihaddadene, I. M. Bilasco, C. Djeraba, and M. Bennamoun, "Dynamic facial expression recognition under partial occlusion with optical flow reconstruction," *IEEE TIP*, vol. 31, pp. 446–457, 2022.

[149] Z. Tao, Y. Wang, Z. Chen, B. Wang, S. Yan, K. Jiang, S. Gao, and W. Zhang, "Freq-hd: An interpretable frequency-based high-dynamics affective clip selection method for in-the-wild facial expression recognition in videos," in *ACM MM*, 2023, pp. 843–852.

[150] A. Savchenko, "Facial expression recognition with adaptive frame rate based on multiple testing correction," in *International Conference on Machine Learning*. PMLR, 2023, pp. 30 119–30 129.

[151] X. Zhao, X. Liang, L. Liu, T. Li, Y. Han, N. Vasconcelos, and S. Yan, "Peak-piloted deep network for facial expression recognition," in *ECCV*, 2016, pp. 425–442.

[152] S. Wang, H. Shuai, and Q. Liu, "Phase space reconstruction driven spatio-temporal feature learning for dynamic facial expression recognition," *IEEE TAFFC*, vol. 13, no. 3, pp. 1466–1476, 2022.

[153] Y. Liu, C. Feng, X. Yuan, L. Zhou, W. Wang, J. Qin, and Z. Luo, "Clip-aware expressive feature learning for video-based facial expression recognition," *Information Sciences*, vol. 598, pp. 182–195, 2022.

[154] N. Sun, Q. Li, R. Huan, J. Liu, and G. Han, "Deep spatial-temporal feature fusion for facial expression recognition in static images," *PRL*, vol. 119, pp. 49–61, 2019.

[155] Y. Chen, J. Li, S. Shan, M. Wang, and R. Hong, "From static to dynamic: Adapting landmark-aware image models for facial expression recognition in videos," *arXiv preprint arXiv:2312.05447*, 2023.

[156] B. Lee, H. Shin, B. Ku, and H. Ko, "Frame level emotion guided dynamic facial expression recognition with emotion grouping," in *CVPRW*, June 2023, pp. 5681–5691.

[157] X. Zhang, M. Li, S. Lin, H. Xu, and G. Xiao, "Transformer-based multimodal emotional perception for dynamic facial expression recognition in the wild," *IEEE TCSVT*, vol. 34, no. 5, pp. 3192–3203, 2024.

[158] S. Zhao, X. Yao, J. Yang, G. Jia, G. Ding, T.-S. Chua, B. W. Schuller, and K. Keutzer, "Affective image content analysis: Two decades review and new perspectives," *IEEE TPAMI*, vol. 44, no. 10, pp. 6729–6751, 2022.

[159] X. Mai, H. Wang, Z. Tao, J. Lin, S. Yan, Y. Wang, J. Liu, J. Yu, X. Tong, Y. Li *et al.*, "Ous: Scene-guided dynamic facial expression recognition," *arXiv preprint arXiv:2405.18769*, 2024.

[160] L. Sun, Z. Lian, B. Liu, and J. Tao, "Mae-dfer: Efficient masked autoencoder for self-supervised dynamic facial expression recognition," in *ACM MM*, 2023, pp. 6110–6121.

[161] Z. Zhao and I. Patras, "Prompting visual-language models for dynamic facial expression recognition," in *BMVC*, 2023, pp. 1–14.

[162] Y. Li, J. Zeng, and S. Shan, "Learning representations for facial actions from unlabeled videos," *IEEE TPAMI*, vol. 44, no. 1, pp. 302–317, 2020.

[163] H. Li, H. Niu, Z. Zhu, and F. Zhao, "Cliper: A unified vision-language framework for in-the-wild facial expression recognition," *arXiv:2303.00193*, 2023.

[164] N. M. Foteinopoulou and I. Patras, "EmoCLIP: A Vision-Language Method for Zero-Shot Video Facial Expression Recognition," in *IEEE FG*, 2024.

[165] Z. Sun, H. Zhang, J. Bai, M. Liu, and Z. Hu, "A discriminatively deep fusion approach with improved conditional gan (im-cgan) for facial expression recognition," *PR*, vol. 135, p. 109157, 2023.

[166] W. Yang, J. Yu, T. Chen, Z. Liu, X. Wang, and J. Shen, "Multi-threshold deep metric learning for facial expression recognition," *PR*, vol. 156, p. 110711, 2024.

[167] F. Ma, B. Sun, and S. Li, "Facial expression recognition with visual transformers and attentional selective fusion," *IEEE Transactions on Affective Computing*, vol. 14, no. 2, pp. 1236–1248, 2021.

[168] F. Xue, Q. Wang, Z. Tan, Z. Ma, and G. Guo, "Vision transformer with attentive pooling for robust facial expression recognition," *IEEE Transactions on Affective Computing*, vol. 14, no. 4, pp. 3244–3256, 2022.

[169] M. Li, H. Fu, S. He, H. Fan, J. Liu, J. Keppo, and M. Z. Shou, "Dr-fer: Discriminative and robust representation learning for facial expression recognition," *IEEE Transactions on Multimedia*, 2023.

[170] K. Zhu, Z. Du, W. Li, D. Huang, Y. Wang, and L. Chen, "Discriminative attention-based convolutional neural network for 3d facial expression recognition," in *IEEE FG*, 2019, pp. 1–8.

[171] M. Sui, Z. Zhu, F. Zhao, and F. Wu, "Ffnet-m: Feature fusion network with masks for multimodal facial expression recognition," in *IEEE ICME*, 2021, pp. 1–6.

[172] Z. Zhu, M. Sui, H. Li, and F. Zhao, "Cmanet: Curvature-aware soft mask guided attention fusion network for 2d+ 3d facial expression recognition," in *IEEE ICME*, 2022, pp. 1–6.

[173] H. Li, H. Yang, and D. Huang, "Drfer: Learning disentangled representations for 3d facial expression recognition," *arXiv:2403.08318*, 2024.

[174] P. Liu, Y. Lin, Z. Meng, L. Lu, W. Deng, J. T. Zhou, and Y. Yang, "Point adversarial self-mining: A simple method for facial expression recognition," *IEEE Transactions on Cybernetics*, vol. 52, no. 12, pp. 12 649–12 660, 2021.

[175] Y. Li, J. Huang, S. Lu, Z. Zhang, and G. Lu, "Cross-domain facial expression recognition via contrastive warm up and complexity-aware self-training," *IEEE TIP*, 2023.

[176] J. Teng, D. Zhang, W. Zou, M. Li, and D.-J. Lee, "Typical facial expression network using a facial feature decoupler and spatial-temporal learning," *IEEE TAFFC*, 2021.

[177] T. Li, K.-L. Chan, and T. Tjahjadi, "Multi-scale correlation module for video-based facial expression recognition in the wild," *Pattern Recognition*, vol. 142, p. 109691, 2023.

[178] Z. Huang, Y. Zhu, H. Li, and D. Yang, "Dynamic facial expression recognition based on spatial key-points optimized region feature fusion and temporal self-attention," *Engineering Applications of Artificial Intelligence*, vol. 133, p. 108535, 2024.

[179] D. Chen, G. Wen, H. Li, P. Yang, C. Chen, and B. Wang, "Cdgt: Constructing diverse graph transformers for emotion recognition from facial videos," *Neural Networks*, p. 106573, 2024.

[180] L. Wang, X. Kang, F. Ding, S. Nakagawa, and F. Ren, "A joint local spatial and global temporal cnn-transformer for dynamic facial expression recognition," *Applied Soft Computing*, vol. 161, p. 111680, 2024.

[181] H. Wang, B. Li, S. Wu, S. Shen, F. Liu, S. Ding, and A. Zhou, "Rethinking the learning paradigm for dynamic facial expression recognition," in *CVPR*, 2023, pp. 17 958–17 968.

[182] H. Li, M. Sui, Z. Zhu *et al.*, "Nr-dfernet: Noise-robust network for dynamic facial expression recognition," *arXiv preprint arXiv:2206.04975*, 2022.

[183] X. Zhang, M. Li, S. Lin, H. Xu, and G. Xiao, "Transformer-based multimodal emotional perception for dynamic facial expression recognition in the wild," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[184] K. Chumachenko, A. Iosifidis, and M. Gabbouj, "Mma-dfer: Multimodal adaptation of unimodal models for dynamic facial expression recognition in-the-wild," in *CVPR*, 2024, pp. 4673–4682.

[185] L. Sun, Z. Lian, B. Liu, and J. Tao, "Hicmae: Hierarchical contrastive masked autoencoder for self-supervised audio-visual emotion recognition," *arXiv preprint arXiv:2401.05698*, 2024.

[186] Z. Tao, Y. Wang, J. Lin, H. Wang, X. Mai, J. Yu, X. Tong, Z. Zhou, S. Yan, Q. Zhao *et al.*, "A3lign-dfer: Pioneering comprehensive dynamic affective alignment for dynamic facial expression recognition with clip," *arXiv preprint arXiv:2403.04294*, 2024.

[187] H. Chen, H. Huang, J. Dong, M. Zheng, and D. Shao, "Finecliper: Multi-modal fine-grained clip for dynamic facial expression recognition with adapters," *arXiv preprint arXiv:2407.02157*, 2024.

[188] C. Bisogni, A. Castiglione, S. Hossain, F. Narducci, and S. Umer, "Impact of deep learning approaches on facial expression recognition in healthcare industries," *IEEE TII*, vol. 18, no. 8, pp. 5619–5627, 2022.

[189] J. Ye, Y. Yu, Y. Zheng, Y. Liu, and Q. Wang, "Dep-fer: Facial expression recognition in depressed patients based on voluntary facial expression mimicry," *IEEE TAFFC*, pp. 1–15, 2024.

[190] J. Chen, C. Guo, R. Xu, K. Zhang, Z. Yang, and H. Liu, "Toward children's empathy ability analysis: Joint facial expression recognition and intensity estimation using label distribution learning," *IEEE TII*, vol. 18, no. 1, pp. 16–25, 2022.

[191] J. Ye, Y. Yu, G. Fu, Y. Zheng, Y. Liu, Y. Zhu, and Q. Wang, "Analysis and recognition of voluntary facial expression mimicry based on depressed patients," *IEEE JBHI*, vol. 27, no. 8, pp. 3698–3709, 2023.

[192] C. Tang, S. Li, W. Zheng, Y. Zong, S. Zhang, C. Lu, and Y. Zhao, "Cfew: A large-scale database for understanding child facial expression in real world," *IEEE TAFFC*, pp. 1–14, 2023.

[193] F. Pessanha, A. A. Salah, T. van Loon, and R. Veltkamp, "Facial image-based automatic assessment of equine pain," *IEEE TAFFC*, vol. 14, no. 3, pp. 2064–2076, 2023.

[194] X. Yuan, Z. Cui, D. Xu, S. Zhang, C. Zhao, X. Wu, T. Jia, and B. Ouyang, "Occluded facial pain assessment in the icu using action units guided network," *IEEE JBHI*, vol. 28, no. 1, pp. 438–449, 2024.

[195] S. Li, "Application of entertainment e-learning mode based on genetic algorithm and facial emotion recognition in environmental art and design courses," *Entertainment Computing*, vol. 52, p. 100798, 2025.

[196] Ö. Sümer, P. Goldberg, S. D'Mello, P. Gerjets, U. Trautwein, and E. Kasneci, "Multimodal engagement analysis from facial videos in the classroom," *IEEE TAFFC*, vol. 14, no. 2, pp. 1012–1027, 2023.

[197] C.-S. Jiang, Z.-T. Liu, M. Wu, J. She, and W.-H. Cao, "Efficient facial expression recognition with representation reinforcement network and transfer self-training for human–machine interaction," *IEEE TII*, vol. 19, no. 9, pp. 9943–9952, 2023.

[198] M. Braun, F. Weber, and F. Alt, "Affective automotive user interfaces–reviewing the state of driver affect research and emotion regulation in the car," *ACM Computing Surveys (CSUR)*, vol. 54, no. 7, pp. 1–26, 2021.

[199] L. Devillers and R. Cowie, "Ethical considerations on affective computing: an overview," *Proceedings of the IEEE*, 2023.

[200] A. Toisoul, J. Kossaifi, A. Bulat, G. Tzimiropoulos, and M. Pantic, "Estimation of continuous valence and arousal levels from faces in naturalistic conditions," *Nature Machine Intelligence*, vol. 3, no. 1, pp. 42–50, 2021.

[201] X. Li, W. Deng, S. Li, and Y. Li, "Compound expression recognition in-the-wild with au-assisted meta multi-task learning," in *CVPR*, 2023, pp. 5734–5743.

[202] P. Ekman and W. V. Friesen, "Facial action coding system," *Environmental Psychology & Nonverbal Behavior*, 1978.

[203] Y. Li and S. Shan, "Contrastive learning of person-independent representations for facial action unit detection," *IEEE TIP*, vol. 32, pp. 3212–3225, 2023.

[204] L. Snoek, R. E. Jack, P. G. Schyns, O. G. Garrod, M. Mittenbühler, C. Chen, S. Oosterwijk, and H. S. Scholte, "Testing, explaining, and exploring models of facial expressions of emotions," *Science advances*, vol. 9, no. 6, p. eabq8421, 2023.

[205] C. Zhan, D. She, S. Zhao, M.-M. Cheng, and J. Yang, "Zero-shot emotion recognition via affective structural embedding," in *CVPR*, 2019, pp. 1151–1160.

[206] P. Yang, N. Liu, X. Liu, Y. Shu, W. Ji, Z. Ren, J. Sheng, M. Yu, R. Yi, D. Zhang, and Y.-J. Liu, "A Multimodal Dataset for Mixed Emotion Recognition," *Scientific Data*, vol. 11, no. 1, p. 847, Aug. 2024.

[207] Z. Zhang, L. Wang, and J. Yang, "Weakly supervised video emotion detection and prediction via cross-modal temporal erasing network," in *CVPR*, 2023, pp. 18 888–18 897.

[208] T. Bai, H. Liang, B. Wan, L. Yang, B. Li, Y. Wang, B. Cui, C. He, B. Yuan, and W. Zhang, "A survey of multimodal large language model from a data-centric perspective," *arXiv preprint arXiv:2405.16640*, 2024.

[209] N. Rawal and R. M. Stock-Homburg, "Facial emotion expressions in human–robot interaction: A survey," *IJSR*, vol. 14, no. 7, pp. 1583–1604, 2022.

[210] J. P. Lee, H. Jang, Y. Jang, H. Song, S. Lee, P. S. Lee, and J. Kim, "Encoding of multi-modal emotional information via personalized skin-integrated wireless facial interface," *Nature Communications*, vol. 15, no. 1, p. 530, 2024.

[211] Z. Chen, J. Cao, Z. Chen, Y. Li, and C. Ma, "Echomimic: Lifelike audio-driven portrait animations through editable landmark conditions," *arXiv preprint arXiv:2407.08136*, 2024.

[212] C. Xu, J. Zhu, J. Zhang, Y. Han, W. Chu, Y. Tai, C. Wang, Z. Xie, and Y. Liu, "High-fidelity generalized emotional talking face generation with multi-modal emotion space learning," in *CVPR*, 2023, pp. 6609–6619.

[213] X. Liu, R. Ni, B. Yang, S. Song, and A. Cangelosi, "Unlocking human-like facial expressions in humanoid robots: A novel approach for action unit driven facial expression disentangled synthesis," *IEEE Transactions on Robotics*, 2024.

[214] Y. Hu, B. Chen, J. Lin, Y. Wang, Y. Wang, C. Mehlman, and H. Lipson, "Human-robot facial coexpression," *Science Robotics*, vol. 9, no. 88, p. eadi4724, 2024.