Computers & Graphics 115 (2023) 423–445

Contents lists available at ScienceDirect

Computers & Graphics

journal homepage: www.elsevier.com/locate/cag

Special Section on CAD/Graphics 2023

4D facial analysis: A survey of datasets, algorithms and applications

Yong-Jin Liu [a], Baodong Wang [b], Lin Gao [c], Junli Zhao [b,*], Ran Yi [d], Minjing Yu [e], Zhenkuan Pan [b], Xianfeng Gu [f,*]

[a] BNRist, MOE-Key Laboratory of Pervasive Computing, Department of Computer Science and Technology, Tsinghua University, Beijing, 100084, China
[b] College of Computer Science and Technology, Qingdao University, Qingdao, 266071, China
[c] Beijing Key Laboratory of Mobile Computing and Pervasive Device Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190, China
[d] Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, 200240, China
[e] College of Intelligence and Computing, Tianjin University, Tianjin, 300350, China
[f] Department of Computer Science, Stony Brook University, Stony Brook, 11790, USA

ARTICLE INFO

Article history:
Received 12 May 2023
Received in revised form 3 July 2023
Accepted 7 July 2023
Available online 13 July 2023

Keywords:
4D facial datasets
4D face analysis
Dynamic facial expressions

ABSTRACT

Facial information plays an important role in human communication, e.g., rich and nuanced facial expressions effectively convey emotions. Traditional data such as 2D facial images and videos are susceptible to perturbation from lighting and occlusion, while 3D static data such as mesh models lack the temporal information which is necessary to describe facial dynamics. Therefore, researches on 4D facial data (3D facial models together with time as the fourth dimension) have received considerable attention in recent years. 4D data can simultaneously reflect the complex facial temporal and space information. To fully explore these characteristics, we present a systematic overview of the 4D facial research in this paper. We give a review in terms of historical development of 4D facial datasets, acquisition process of these datasets, related algorithms and applications, and discussion of outstanding issues. We also analyze the 4D facial research works by summarizing and comparing them; in particular, we present the results on three kinds of tasks conducted on the 4D facial datasets, i.e., expression-related tasks (AU detection, expression recognition and retrieval), generation tasks (3D facial reconstruction, facial expression and facial animation generation), and other tasks (facial registration, facial recognition and facial disease diagnosis). Finally, we summarize the future research directions.

1. Introduction

Face analysis has always been a popular research direction as many practical application tasks are based on faces, such as recognition, detection, synthesis, alignment, and etc. It is a challenging task due to the complex geometry of facial surface, the concave and convex nature of facial organs, and the facial muscles supporting dynamic changes. Early face research analysis is generally based on 2D images or video data and most of them focus on facial recognition. The literature [1] has a detailed summary of the 2D facial analysis method. 3D facial data is more advantageous than 2D facial data in response to blocking and light changes and can characterize more features. Therefore, it can adapt to more scenarios. With the recent improvement of high-precision 3D imaging levels and the popularization of low-precision RGB-D cameras, obtaining 3D facial data has become easier. The 4D facial data adds temporal information and more dynamic information about the face, which is vital to the expression of facial emotions. At the same time, the development of deep learning technology in recent years has made advanced technologies such as encoder–decoder structure [2,3], long-term memory network (LSTM) [4], Transformer [5], and other advanced technologies [6–8] emerging in 4D face research and applications. Therefore, a systematic survey of the 4D face analysis is necessary.

Recently, many researchers have shifted their focus to 4D facial data due to the advantages of 4D facial data compared with 2D and 3D facial data. On the one hand, 4D data contains more temporal information that can be used to analyze complex dynamic changes in a face. Facial expressions and movements, especially micro-expressions, are often not completed in a single frame, and 4D data adds a time dimension to 3D static single-frame facial data, providing more comprehensive facial information. Therefore, in facial recognition, expression or micro-expression recognition tasks, 4D data can improve accuracy and robustness. On the other hand, 4D facial data can fully express complex facial features. 4D facial data is unaffected by occlusions or lighting conditions. In addition to the advantage of rendering

* Corresponding authors.
E-mail addresses: zhaojl@yeah.net (J. Zhao), gu@cs.stonybrook.edu (X. Gu).

https://doi.org/10.1016/j.cag.2023.07.014
0097-8493/

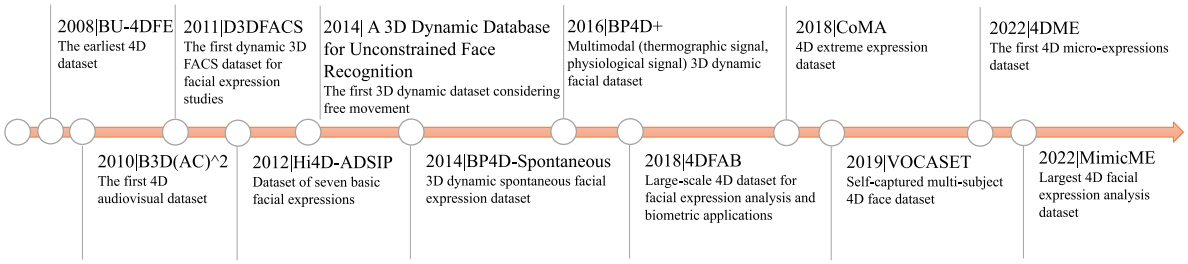


Fig. 1. 4D facial dataset development timeline.

from any angle in view space, 4D data also provides continuous variation over time. In computer animation and virtual reality tasks, 4D data can provide more realistic facial dynamics and expression information, resulting in more lifelike facial animation.

In the past decade, research on face analysis has achieved fruitful results in aspects such as dataset, recognition, and detection, mainly based on 2D and 3D facial images or video data. For example, a recent survey [9] reviewed the progress of 3D facial recognition technology in recent years. Due to the characteristics of 4D data, which contain both temporal and spatial information, an increasing number of researchers have focused on 4D facial research. However, most existing research on 4D facial data is still scattered and lacks systematic survey. In this paper, we provide a comprehensive survey and make the following contributions:

- We provide a comprehensive overview of a 4D facial dataset, including the dataset acquisition equipment, participants, acquisition method, data processing, and database organization.
- We classified and elaborated on the different facial studies conducted using these datasets and discussed the remaining challenges and future development directions to inspire different 4D facial applications.

The rest of this paper is organized as follows. Section 2 introduces the situation of existing 4D facial datasets. Section 3 provides a detailed data collection process for 4D facial datasets. Section 4 summarizes the facial research work carried out using 4D facial datasets. The image and video features and facial analysis methods are also collected and classified. Section 5 discusses the advantages and disadvantages of 2D, 3D, and 4D methods, the influence of attributes of different datasets on 4D applications, and the remaining challenges and future directions for facial research based on 4D data. Finally, Section 6 concludes the paper.

2. 4D facial dataset

The development of 4D facial analysis technology, to a large extent, relies on the development of mature 4D facial datasets. As early as 2004, the State University of New York at Stony Brook proposed a high-resolution, real-time 3D shape acquisition system [10] based on structured light technology. The acquisition frame rate of this system is 120 Hz and the resolution is 532 × 500 pixels. In recent years, Gu et al. [11] has been developing a cutting edge 3D camera system to obtain high-resolution facial surfaces with dynamic expressions.

According to our survey, there have been 12 representative 4D facial datasets developed since 2008, including BU-4DFE [12], B3D(AC)^2 [13], D3DFACS [14], Hi4D-ADSIP [15], datasets proposed by Alashkar et al. [16], BP4D-Spontaneous [17], BP4D+ [5], 4DFAB [18], CoMA [19], VOCASET [20], 4DME [21], and MimicME [22]. Fig. 1 shows the development timeline of 4D facial datasets.

The first high-resolution 3D dynamic facial expression dataset, BU-4DFE, was introduced in 2008. This dataset contains 606 3D facial expression sequences captured from 101 subjects of different racial backgrounds. In 2014, considering that the data in the previously proposed BU-4DFE dataset was based on posed behavior rather than spontaneous behavior, it is essential to note that posed expressions are different from spontaneous expressions. Therefore, BP4D-Spontaneous was the first attempt to induce spontaneous expressions in a controlled environment through a series of carefully designed tasks. Although emotions can be expressed in various ways, most research only focuses on one or two ways due to lacking large, diverse, and well-annotated multi-modal datasets. Therefore, in 2016, a 4D dataset with well-annotated, multi-modal, and multidimensional spontaneous emotions, BP4D+, was proposed.

The first 4D audio-visual database for emotional communication, B3D (AC), was proposed in 2010. The authors believe that speech and facial expression in the form of dense dynamic 3D face geometries may be the two most important ways used by humans to communicate their emotional states.

In 2011, the first 3D dynamic Facial Action Coding System (FACS) [23]for facial expression research was introduced, which resulted in a facial database called D3DFACS. It captured 519 AU sequences with a total of 1184 Action units (AUs). In 2012, the first 4D dataset is used to study facial dysfunctions: Hi4D-ADSIP. Compared to BU-4DFE's six basic expressions, the dataset added a "Pain" expression, and all seven basic expressions were divided into three levels of expression intensity: mild, moderate, and extreme. In 2014, Alashkar et al. [16] proposed a new 3D dynamic facial dataset to develop and test face recognition algorithms under unconstrained conditions. The dataset addressed several challenges that may arise in real-world scenarios, such as continuous and freely-pose variation, expressive and talking faces, distance changes to the 3D camera, occlusions and multiple persons in the scene.

In 2018, the 4DFAB dataset was completed after five years of data acquisition, which included large-scale 4D data from 180 participants recorded across four different sessions. Over 1.8 million high-resolution 3D facial data were collected, including posed and spontaneous expressions. Also, the same year, Ranjan et al. [19] proposed the CoMA dataset, which collected 12 extreme expressions from 12 individuals.

In 2019, a new 4D face and speech dataset called VOCASET and a general-purpose speech-driven facial animation framework called VOCA were proposed. Given speech signal and a static 3D face mesh, the framework generates realistic 3D dynamic character animation.

In 2022, the first 4D micro-expression dataset, 4DME, was proposed. The current micro-expression datasets are insufficient [21], and most of them [24–27] only contain one form of 2D color video. This new 4D dataset recorded approximately 5980 min of video from 65 participants, containing four data modes: DI4D video, frontal grayscale video, Kinect-color video, and Kinect-depth video. Examples of the four modalities are shown in Fig. 2.

Table 1
Details of the 4D facial dataset.

Dataset	Expressions/emotion/different conditions	Coverage	Number of samples	Scanner
BU-4DFE	Angry, Disgust, Fear, Happy, Sad, Surprise	Face, neck, sometimes ears	101 individuals × six 100 frame expression sequences	Di3D
B3D(AC)ˆ2	Negative, Anger, Sadness, Stress, Contempt, Fear, Surprise, Excitement, Confidence, Happiness, Positive	Inner face only	14 individuals × around 80 dynamic sequences (speech-4D)	Structured light stereo
D3DFACS	Angry, Disgust, Fear, Happy, Sad, Surprise and non-additive appearance changes	Face, neck, sometimes ears	10 individuals × around 52 dynamic sequences, FACS coded	3dMD
Hi4D-ADSIP	Angry, Disgust, Fear, Happiness, Sadness, Surprise, Pain	Inner face only	80 individuals × around 42 dynamic Sequences	Di3D
Alashkar et al. [16]	Neutral, Facial Expression, Talking, Internal Occlusion (hand or hair), External Occlusion, Walking, Multiple persons	Inner face only	58 individuals × one static scan + seven dynamic sequences	Artec MHT and Artec L
BP4D-Spontaneous	Happiness/Amusement, Sadness, Startle, Embarrassment, Fear, Physical pain, Anger, Disgust	Face, neck, sometimes ears	41 individuals × eight one minute dynamic sequences, FACS coded	Di3D
BP4D+	Happiness/amusement, Surprise, Sadness, Startle, Skeptical, Embarrassment, Fear/Nervous, Physical pain, Angry/Upset, Disgust	Face, neck, sometimes ears	140 individuals × 10 tasks, FACS coded	Di3D, FLIR A655sc Longwave infrared camera, Biopac MP150
4DFAB	Angry, Disgust, Fear, Happiness, Sadness, Surprise, Embarrassment, Nervousness	Face, neck and ears	180 individuals × 4k–16k frames of dynamic sequences	DI4D
CoMA	Bareteeth, Cheeks in, Eyebrow, High smile, Lips back, Lips up, Mouth down, Mouth extreme, Mouth middle, Mouth side, Mouth up	Full head	12 individuals × 12 extreme expression sequences	3dMD
VOCASET	N/A	Full head, speech	12 individuals × 40 dynamic sequences (speech-4D)	3dMD
4DME	Positive, Negative, Surprise, Repression, Others	Face, neck, ears, speech	65 individuals, around 5980 min of videos	DI4D, Stingray F-046B, Xbox 360
MimicME	6 Basic + Pain, Pout, Show teeth, Flare nostrils, Inflate cheeks, Wink right, Wink left, Bite lower lip, Bite upper lip, Tongue out, Tongue left, Tongue right, Mouth right, Mouth left	Face, neck and ears	4700 individuals × 70–120 3D images	3dMD

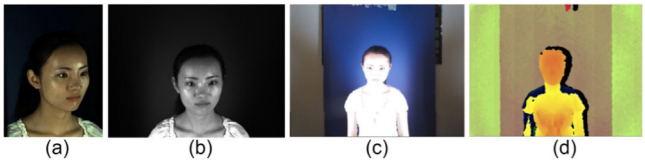


Fig. 2. Sample images of the four modalities in 4DME [21]. (a) (one of the six) DI4D videos; (b) Grayscale videos; (c) Kinect- color videos; (d) Kinect-depth videos.

The dataset also provides five categories of emotion labels and 22 AU labels to explore the relationship between AU annotation and micro-expression emotion categories. The largest 4D facial expression analysis dataset, MimicME, was also proposed in the same year. The dataset collected 4D facial information from 4700 participants over three months.

Table 1 details the most commonly used 4D facial datasets we surveyed.

3. Acquisition of 4D dataset

4D datasets contain richer facial information than traditional 2D facial image or video data, as occlusions and lighting environments do not affect them. However, the acquisition process of 4D datasets is more complex. This section will introduce the 12 4D facial datasets according to their acquisition equipment, participants, acquisition method, data processing, and database organization.

3.1. Acquisition equipment

We found that the BU-4DFE, BP4D-Spontaneous, and Hi4D-ADSIP datasets were collected using the Di3D dynamic facial capture system produced by Dimensional Imaging [28]. Fig. 3(a) shows the specific setup used to collect BU-4DFE. The system consists of two stereo cameras and one texture camera. Three cameras are placed on tripods with two lighting lamps on either side. Two computers run in parallel, and the system captures 3D model sequences and 2D texture videos at 25 frames per second. The working principle of DI3D is to use passive stereo photogrammetry to process each stereo image pair to generate range maps. The combination range map generates a high-resolution 3D sequence that changes with 0.2 mm RMS accuracy over time. When 3D depth model sequences are captured from the top and bottom cameras, related 2D texture videos are also recorded from the middle camera. Fig. 3(b) shows the collection scene of BP4D-Spontaneous. This dataset uses one main computer and three slave computers. In addition, a conventional camera is set up to capture the entire scene. The acquisition system of Hi4D-ADSIP has increased the number of hosts to six, with a maximum recording speed of 60 frames per second. The acquisition scenario is shown in Fig. 3(d).

BP4D+ involves multiple modalities, and its acquisition equipment includes the Di3D dynamic imaging system, thermal signal sensors, and physiological signal sensor system, as shown in Fig. 3(c). The 3D dynamic imaging system (Di3D) includes a 3D stereo imaging sensor and a 2D video sensor. The acquisition data come from various facial sensors (including high-resolution 3D

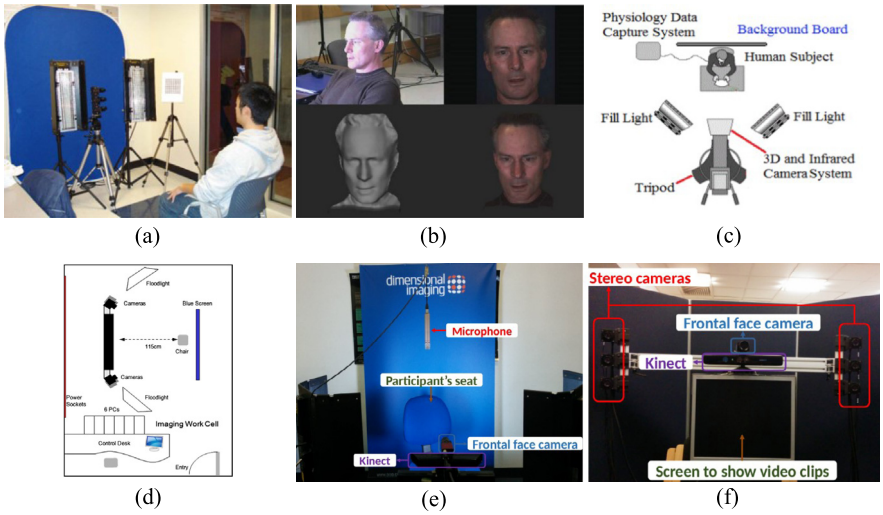


Fig. 3. Acquisition scenarios for each dataset (courtesy of the original papers). (a) BU-4DFE [12], (b) BP4D-Spontaneous [17], (c) BP4D+ [5], (d) Hi4D-ADSIP [15], (e) 4DME(back view) [21], (f) 4DME(frontal view) [21].

dynamic imaging, high-resolution 2D video, and thermal /infrared imaging) as well as contact physiological sensors (including skin conductance, respiration, blood pressure, and heart rate). Experts in FACS annotated the occurrence and intensity of facial action units from the 2D video. The thermal infrared camera used was the FLIR A655sc long-wave infrared camera from the US-based company FLIR, which captures thermal imaging videos at a resolution of 640 × 480 per frame and the capture rate was set to 25 fps. Physiological data were collected using the Biopac MP150 data acquisition system.

With the emergence of high-resolution cameras, the newly produced DI4D system by Dimensional Imaging can capture motion over time. 4DFAB and 4DME are using the DI4D dynamic capture system to capture and build 4D faces, as shown in Fig. 3(e) and (f). The system mainly consists of six cameras (two pairs of cameras and one pair of texture cameras, 60 fps, 1200 × 1600) to capture full-color 3D facial mesh sequences, where each frame of the sequence is treated as a separate stereo image pair and is automatically processed to generate a colored 3D surface. The generated files are combined to produce a high-resolution 3D facial mesh sequence. In addition, a microphone records the audio signal, a front gray camera (60 fps, 640 × 480) records front-facing images, and a Kinect records RGB-D data (30 fps, 640 × 480).

The B3D(AC)ˆ2 dataset was captured using a real-time 3D scanner and a studio condenser microphone to record audio-visual speech data simultaneously.

The dataset proposed by Alashkar et al. [16] uses two types of acquisition equipment: the Artec MHT 3D scanner is used to obtain the 3D static models, while the Artec L 3D scanner has a larger field angle to capture 3D dynamic sequences. Both scanners have a frame rate of 15 fps.

D3DFACS uses the 3dMD dynamic 3D stereo camera to record each FACS performer. The system comprises six cameras in two pods, three vertically stacked on each pod. Each pod's top and bottom cameras are responsible for the stereo reconstruction, while the middle camera captures UV color textures—the system samples at a rate of 60 fps. MimicMe also uses the 3dMD equipment for data acquisition.

The CoMA and VOCASET datasets acquisition devices are multi-camera active stereo system 3dMD LLC. The capture system comprises six pairs of grayscale stereo cameras, six color cameras, five speckle pattern projectors, and six white light LED panels. The system captures 3D meshes at 60 fps with approximately 120K vertices per mesh. Color images are used to generate UV texture maps for each scan. Audio synchronized with the scanner is captured at a sampling rate of 22 kHz.

3.2. Participants

In this section, we summarize the participant information for different datasets. Due to factors such as the high cost and long time required for 4D data acquisition, the specific environmental conditions during acquisition, and the requirement for emotionally induced facial expressions, the number of participants in current 4D datasets are relatively small, with most datasets consisting of only a few dozen individuals. Only three datasets had more than one hundred participants. Table 2 provides information on the participants for different datasets. For example, the D3DFACS dataset has only 10 participants, including four expert FACS coders and six untrained participants. The CoMA and VOCASET datasets both have 12 participants. The B3D(AC)ˆ2 dataset has 14 native English-speaking participants. The BP4D-Spontaneous dataset has a total of 41 participants. The dataset proposed by Alashkar et al. has 58 participants. 4DME recruited 65 volunteers from the campus to participate in data acquisition. The Hi4D-ADSIP dataset has 80 participants, including 65 undergraduate students majoring in performing arts, as well as other undergraduate and graduate students and staff members from other departments who were not explicitly trained in performing arts.

There are only four datasets with more than 100 participants: BU-4DFE, BP4D+, 4DFAB, and MimicMe. The BU-4DFE dataset has 101 participants, including undergraduate and graduate students and teachers from the psychology, computer science, and engineering departments. The BP4D+ dataset has 140 participants. Currently, the 4DFAB dataset has 180 participants. Among all participants, 179 subjects participated in the first data collection, 100 subjects participated in the second, and 81 and 75 participated in the third and fourth, respectively, with an average time interval of 219 days between consecutive attendances (minimum: 1 day, maximum: 1654 days). The participants were students from the administrative departments, engineering, business, and medical majors, and volunteers from outside the school. MimicMe collected facial expressions from 4700 museum visitors over three months during a special exhibition at the Science Museum in London.

Table 2
Data acquisition participant information for 4D facial dataset.

Dataset	Number of participants	Age	Gender	Ethnic(number)
BU-4DFE	101	18–45	58 females, 43 males	Asian(28), Black(8), Hispanic/Latino(3) and White(62)
B3D(AC)^2	14	21–53	8 females, 6 males	–
D3DFACS	10	23–41	6 females, 4 males	Caucasian European
Hi4D-ADSIP	80	18–65	48 females, 32 males	–
Alashkar et al. [16]	58	Avg 23	23 females, 35 males	–
BP4D-Spontaneous	41	18–29	23 females, 18 males	Asian(11), African-American(6), Hispanic(4) and Euro-American(20)
BP4D+	140	18–66	82 females, 58 males	Asian(46), African American(15), White(64), Latino/Hispanic(14) and Others(1)
4DFAB	180	5–75	60 females, 120 males	Ethnicity includes Caucasian (Europeans and Arabs), Asian (East-Asian and South-Asian) and Hispanic/Latino
CoMA	12	–	–	–
VOCASET	12	–	6 females, 6 males	–
4DME	65	22–57	27 females, 38 males	Eastern Asia(37), southern Europe(27) and Britain(1)
MimicME	4700	–	–	–

– Not Stated.

Our research found an interesting phenomenon: The data of only the 4DFAB dataset are collected at four different times. We believe that conducting multiple data collections on the same group of participants at different time intervals may be beneficial for studying the stability of facial expressions in response to specific events or the changes in behavior over time, which may also be an exciting direction for medical research on individual emotional stability. However, currently, there is no research confirming this, possibly because it would make the already cumbersome process of 4D facial data collection even more challenging.

3.3. Acquisition method

This section summarizes the data acquisition processes designed for each participant in 4D facial expression datasets. Researchers induced emotions in participants through various methods such as video emotion induction, conversation, physical stimulation, and reading designated texts, among others. Next, we will introduce the acquisition method of each dataset.

Currently, the data acquisition for 4D datasets is usually conducted in a laboratory or soundproof room environment, even for spontaneous facial expression data acquisition, which involves inducing emotions through a series of methods within the laboratory. Before the data acquisition begins, each dataset typically requires participants to sit a certain distance away from the capture system for a short pre-test to ensure that the participant's posture and orientation are suitable for generating accurate data during the formal recording.

For posed facial expressions, participants in the BU-4DFE dataset were instructed by a psychologist to perform six facial expressions. Each expression was performed sequentially from a neutral expression, low intensity, to high intensity, back to low intensity, and finally, a neutral expression, lasting approximately 4 s. During the D3DFACS data acquisition, participants were asked to repeat the target AUs as much as possible. Each participant was required to record for 2 to 7 h. After all data records, FACS experts score the data to select the sequence that matches the most expression with the target expression. For MimicMe, each participant was required to watch facial expression videos performed by actors twice. The first viewing served as a familiarization process, while the second viewing involved capturing the facial expressions, with participants required to mimic the actors' expressions.

The expression of emotions can be achieved visually, but repeating emotional sentences or reading emotion-evoking texts can also be a way to induce emotions. The data acquisition for B3D(AC)^2 was divided into two stages. In the first stage, participants were asked to read sentences displayed on a computer screen while trying to maintain a neutral tone. In the second stage, participants watched emotion-inducing videos and was asked to evaluate the emotional content on a paper questionnaire. In the Hi4D-ADSIP dataset, participants were asked to make facial utterances continuously. Each recorded sequence began and ended with a neutral facial appearance and was performed with a specified intensity of utterance. While recording the 3D scanner video stream, the synchronized audio was also recorded.

The acquisition of spontaneous facial expressions is often achieved by designing specific tasks to induce emotions in participants. Common methods of inducing emotions include watching video clips, conversation, word pronunciation, cold pressure, and designing physical experiences. BP4D-Spontaneous designed eight tasks for each participant and collected 4D facial data with eight different expressions. BP4D+ designed a protocol consisting of ten tasks to achieve a natural transition from positive emotions to negative emotions. For 4DFAB, participants were first asked to perform the six basic facial expressions and then pronounce nine words in order during the formal recording. These two tasks were repeated in all four different recording sessions. Then show the participants a few videos to trigger their spontaneous expressions. 4DME was designed for micro-expression research, so participants were asked to hide their true feelings and maintain a neutral face throughout the experiment. The method of inducing micro-expressions was the same as that used in previous literature [24,29]. Participants showed 11 carefully selected video clips that elicited intense emotions. Only 4DFAB did not require participants to complete a self-report.

The above facial datasets were proposed for research on facial expressions, while the dataset proposed by Alashkar et al. [16] is dedicated to research on unconstrained facial recognition. The dataset collected each participant's full 3D static models and eight sets of 3D video sequences. For the whole 3D part, the Artec MHT scanner was used to capture each participant's full 3D static facial model and texture information, which takes approximately 2 min to obtain. For the 3D video sequence, eight sets of 3D videos containing seven scenes were collected for each participant. The seven scenes were neutral facial expression, conversation, internal occlusion, external occlusion, walking, and multi-person scenes, each with geometric information but no texture.

3.4. Data processing and database organization

In this section, we summarize the data processing and database organization of the 4D facial dataset.

According to our survey, all those mentioned above 4D facial datasets have sequences of 3D models with triangular meshes. The average number of vertices in the 3D models of early datasets

Table 3
4D dataset data format.

Dataset	Triangle mesh	Texture	Other format	Size
BU-4DFE	35k vertices	Texture image (1040 × 1329)	–	Approximate 500 GB
B3D(AC)^2	Raw scan: 55k vertices; processed:23k vertices	Raw scan: texture image (780 × 580); processed: UV texture map (1024 × 768)	–	–
D3DFACS	30k vertices	UV texture map (1024 × 1280)	–	–
Hi4D-ADSIP	20k vertices	Texture image	–	Triangle mesh: 7.8 TB, video clip: 6 GB, GIF: 39.0 GB
Alashkar et al. [16]	3.5k vertices for dynamic, 50k vertices for static	Texture image	–	–
BP4D-Spontaneous	30k–50k vertices	Texture image (1040 × 1392)	–	Video data about 2.6 TB
BP4D+	30k–50k vertices	Texture image (1040 × 1392)	Thermal image (640 × 480), physiological signal	Over 10 TB with about 1.4 million frames
4DFAB	60k–75k vertices	UV texture map	–	Over 20 TB
CoMA	80k–140k vertices	Texture images (avg resolution 3700 × 3200)	Six raw camera images (1600 × 1200), alignments in FLAME topology	–
VOCASET	80k–140k vertices	Texture images (avg resolution 3700 × 3200)	Six raw camera images (1600 × 1200), alignments in FLAME topology	–
4DME	Over 50k vertice	UV texture map (1200 × 1600)	2D frontal facial videos (640 × 480), RGB-D images (640 × 480)	–
MimicME	20k vertices	UV texture map	–	–

– Not Stated

such as BU-4DFE and D3DFACS was only 35k and 30k, respectively. However, in recent years, the average number of vertices in datasets such as 4DFAB, VOCASET, and MimicME has reached 60k–75k, 80k–140k, and 120k, respectively. This undoubtedly makes the models more detailed and increases the required storage space. For example, the storage space required for 4DFAB has already exceeded 20 TB.

Four datasets have FACS AU coding, namely D3DFACS, BP4D-Spontaneous, BP4D+, and 4DME. D3DFACS collected AU sequences of 42 different action unit types from 10 subjects. Each sequence is about 90 frames long at 60 fps and consists of OBJ meshes and BMP cylindrical UV texture mapping data, totaling 519 sequences. For each of the eight tasks in BP4D-Spontaneous, two FACS-certified coders independently coded 27 action units for the 20-s segment with the highest facial expression density. BP4D+ is similar to BP4D-Spontaneous, but with a larger dataset size, providing annotations for 34 AU types for four sessions with 197,875 frames. 4DME has both AU and emotion category annotations. AU annotation is a time-consuming and labor-intensive complex task, but it is the data foundation for many subsequent works on AU detection and facial expression recognition.

As one of the ways of emotional expression, audio is also recorded in the acquisition process of multiple data sets. B3D(AC)^2 is the first audio-visual dataset that records speech and dense 3D dynamic facial geometry representations of facial expression sequences for 40 English phrases, with a total of 1109 sequences and an average length of 4.67 s. In Alashkar et al.'s dataset, one of the seven scenes in the collection of 3D dynamic sequences requires participants to speak and move their heads freely. The Hi4D-ADSIP dataset includes three parts: an expression database, a basic-articulation database, and a phrases reading database. During data acquisition, 4DFAB and 4DME also used a microphone to record audio signals. VOCASET records speech and 3D mesh models for 40 English sentence sequences for each participant, with each sentence length ranging from 3 to 5 s. To maximize speech diversity, twenty-seven of the sentences come from the TIMIT corpus [30], three from pangrams [31], and ten from the Stanford Question Answering Dataset (SQuAD) [32].

In addition to the CoMA dataset, all other datasets have 2D texture videos. The CoMA dataset consists of 12 extreme expressions from 12 subjects. The dataset contains 20,466 3D meshes, each with approximately 120k vertices. The Sequential Mesh Registration method [33] was used for preprocessing the data to reduce the dimensionality to 5023 vertices. In contrast, the data collected by BP4D+ includes high-resolution 3D dynamic imaging, 2D videos, thermal imaging videos, and physiological information, including skin conductivity, respiratory rate, blood pressure, and heart rate, with a total data volume of over 10 TB and a total of approximately 1.4 million frames.

Therefore, unlike 2D and 3D facial datasets, 4D datasets often contain multiple data modalities, and the data scale is gradually increasing. Such large-scale, diverse, and multimodal datasets undoubtedly play an important role in promoting complex facial analysis tasks. Table 3 summarizes the data formats and storage sizes of the published 4D datasets according to our survey.

4. Algorithms and applications for 4D datasets

According to our research on derivative works of 4D facial datasets, there are three main research directions for the application of 4D facial datasets in facial analysis: expression-related tasks, generation tasks, and other tasks. The first task includes expression recognition, retrieval and AU detection. Unlike the expression recognition task based on 2D images or videos, the development of 4D facial tasks is closely related to the development of 4D facial datasets. Before 2016, researchers mainly focused on the overall expression recognition and retrieval of facial expressions. After 2018, with the development of 4D facial databases, researchers' focus shifted to detecting facial AUs. The research on generating tasks can be divided into 3D facial reconstruction, expression generation, and 3D facial animation generation. The third research direction is other tasks related to registration, facial recognition, and facial disease estimation. Fig. 4 provides a classification summary of different research directions based on 4D facial datasets. Table 4 provides the corresponding references and their years for the classification of derivative works.

4.1. Expression-related tasks

Facial expression classification and recognition is a very popular research field with a wide range of applications in human–computer interaction, psychology, computer vision, traffic safety,

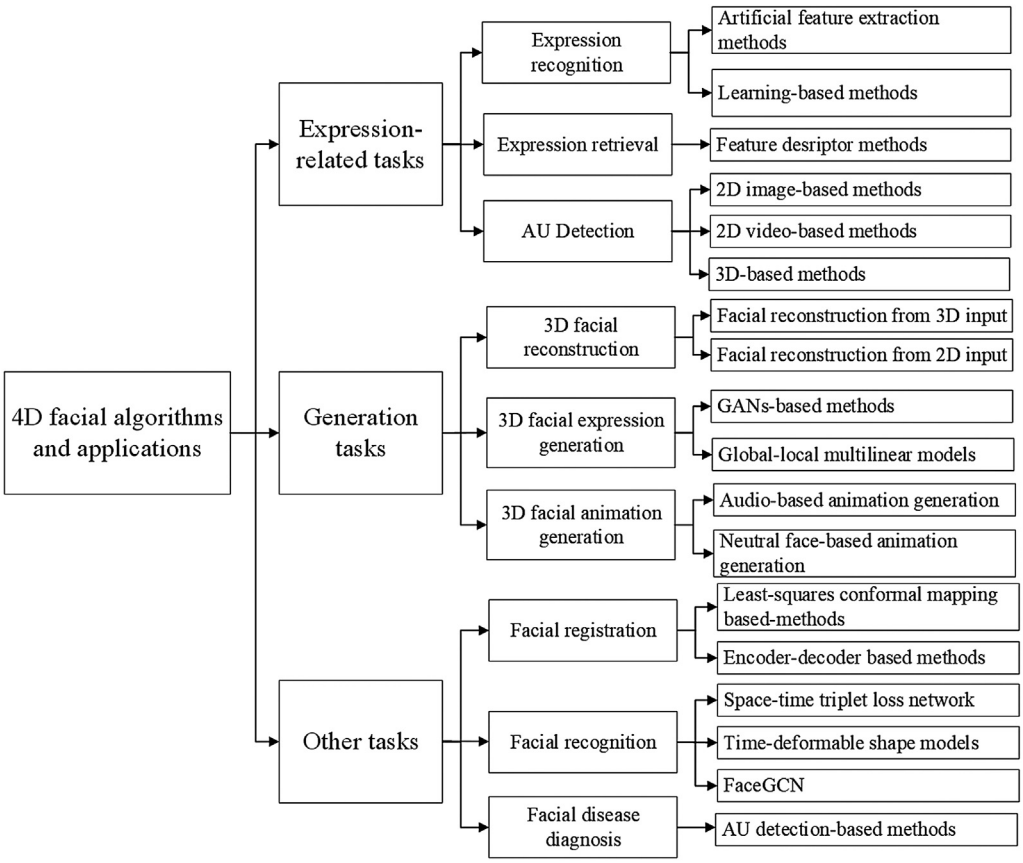


Fig. 4. Classification of algorithms and applications for 4D facial datasets.

and other fields. As early as 1971, Ekman [34] conducted the first systematic study of facial expressions and proposed six universally accepted basic emotions not influenced by language or culture: anger, disgust, fear, happiness, sadness, and surprise. The facial expression recognition tasks that have been widely used have primarily used 2D facial images or videos for research. There are few studies on 4D facial expression analysis based on 4D datasets, most of which are focused on the BU-4DFE and BP4D-Spontaneous datasets.

4.1.1. Expression recognition

The feature extraction methods for facial expression recognition tasks can be divided into two categories: artificial feature extraction methods and learning-based methods. In addition, some scholars have conducted research on facial expression retrieval by continuously improving the feature descriptors defined on the face [43–45].

(1) Artificial feature extraction methods

Early methods mostly relied on manually extracting features, which can be divided into methods based on facial curves [35,36], methods based on feature points/regions [37,38], methods based on frequency domain features [39], and methods based on in-depth features [40], depending on the type of facial feature.

Methods based on facial curves Maalej et al. [35] described facial features using iso-geodesic and quantified their shape information under the Riemann framework to obtain similarity scores between different local shapes of the face. Linear Discriminant Analysis (LDA) was then used to reduce the dimensionality of the features, and Hidden Markov Model (HMM) was used to perform temporal modeling of 3D frame sequences for expression classification. Experiments on BU-4DFE dataset have achieved an average recognition rate of 93.83%. In contrast to the methods mentioned above, literature [36] used Riemann analysis to define a new Dense Scalar Field (DSF) on the radial curves of 3D faces, accurately capturing 3D deformations. LDA was also used to reduce the dimensionality of features, and a classifier based on HMM was used for the six basic expressions in the BU-4DFE dataset, achieving an average recognition rate of 93.83%.

Methods based on feature points/regions Literature [37] proposed a 4D facial expression recognition method by combining distance information and local shape context descriptors of facial landmark points. The experimental results by HMM classification on the 3D dynamic sequences of the six basic expressions in the BU-4DFE dataset obtained an average recognition rate of 79.44%. Zhen et al. [38] proposed a 3D/4D facial expression recognition method based on the Muscular Movement Model (MMM) from an anatomical perspective, which first locates 11 muscle regions corresponding to single major facial muscles or combinations using the Iterative Closest Normal Point (ICNP) algorithm, then extracts the coordinates, normals, and shape indices of each vertex in the region to form a feature vector. SVM and HMM are used to perform expression classification for 3D static expressions and 4D expressions, respectively. The experimental results on the BU-3DFE and BU-4DFE databases obtained recognition accuracies of 83.2% and 87.06%, respectively.

Methods based on frequency domain features Xue [39] pointed out that the changes in facial expressions are essentially a spatio-temporal process and that performing feature extraction frame by frame may not capture the dynamic changes in expressions. Therefore, the authors proposed a frequency-domain facial expression recognition method. Specifically, the 3D discrete cosine transform features around the 68 facial landmark points are transformed into the frequency domain. The Minimal Redundancy Maximal Relevance (mRMR) algorithm selects the m

Table 4
Classification and references of 4D facial derivative work.

Task	Method	References (Year)
Expression recognition	Artificial feature extraction methods	[35] (2012), [36] (2014), [37] (2013), [38] (2016), [39] (2015), [40] (2016)
	Learning-based methods	[41] (2018), [42] (2021)
Expression retrieval	Feature descriptor methods	[43] (2014), [44] (2015), [45] (2016)
AU detection	2D image-based methods	[6] (2018), [46] (2019), [47] (2021), [48] (2021), [49] (2021), [50] (2022)
	2D video-based methods	[51] (2019), [52] (2021)
	3D-based methods	[53] (2018), [54] (2020), [55] (2019), [56] (2021)
3D facial reconstruction	Facial reconstruction from 3D input	[33] (2017), [2] (2020)
	Facial reconstruction from 2D input	[3] (2020), [57] (2019), [58] (2021), [59] (2023), [60] (2022), [61] (2022), [62] (2023)
3D facial expression generation	GANs-based methods	[63] (2020), [7] (2019), [64] (2019), [65] (2020)
	Global-local multilinear models	[66] (2020)
3D facial animation generation	Audio-based animation generation	[67] (2021), [68] (2021), [69] (2022), [70] (2022)
	Neutral face-based animation generation	[4] (2020), [71] (2022)
Facial registration	Least-squares conformal mapping based-methods	[72] (2008)
	Encoder–decoder based methods	[73] (2021)
Facial recognition	Space–time triplet loss network	[74] (2021)
	Time-deformable shape models	[75] (2021)
	FaceGCN	[8] (2022)
Facial disease diagnosis	AU detection-based methods	[76] (2022)

best-performing features. Finally, a nearest-neighbor classifier is used for expression classification. In the experiment, the average recognition rate for the six facial expressions in the BU-4DFE dataset was 78.8%, significantly improving recognition rates for the three easily confused expressions including anger, fear, and sadness.

Methods based on in-depth features Duh et al. [40] first converted 3D mesh vertices into depth frames and extracted Histograms of Oriented Gradient (HOG) and Histograms of Optical Flow (HOF) for Space-Time Interest Points (STIPs) from the depth sequences. Then, they used naive Bayes information maximization and constrained matching pairs to calculate Mutual Information Score (MIS) and Weighted Matching Score (WMS), respectively. Finally, the MIS and WMS results were concatenated into a feature vector and input into SVM for facial expression classification. This method was tested on the six facial expressions in the BU-4DFE dataset, achieving an average classification accuracy of 77.7%.

(2) Learning-based methods

Li et al. [41] proposed a Dynamic Geometrical Image Network (DGIN) for automatic 4D facial expression recognition. This method first extracts a set of 3D facial scan sequences from a 3D video for each facial expression, generates depth image (DPI), normal component images (NCI), and shape index images (SII) using projection and interpolation techniques as the input of DGIN network. The DGIN network comprises a short-term pooling layer, multiple groups of convolution+RELU+pooling layers, a long-term pooling layer, a full connection layer, and a loss function layer. Among them, the short-term pooling layer is responsible for generating dynamic geometric images DGI, and the long-term pooling layer is responsible for integrating dynamic feature maps. During the testing phase, it achieved a recognition accuracy of 92.22% on the BU-4DFE dataset.

Behzad et al. [42] proposed a new Multi-view transformer (MiT) that transforms multi-view 2D images into fixed-size multi-view images(left, front, and right), and then combines patches and position embeddings from different views to obtain a vector sequence. Then, a multi-view Transformer encoder predict the expression based on the vector sequence. During training, a novel multi-view loss term is proposed to facilitate gradient updates across multiple views during backpropagation. In the experimental section, four datasets were used for experiments, among which 4D facial expression recognition achieved recognition accuracies of 99.66% and 91.67% on the BU-4DFE and BP4D-Spontaneous datasets, respectively.

Except for a few learning-based methods, early facial expression recognition tasks mostly relied on manually extracting features from facial curves, feature points/regions, frequency domain, or optical flow to form feature vectors or descriptors and then combining them with classifiers such as SVM or HMM for expression recognition. When the extracted features are exactly effective for a specific task, the results are often better. However, such methods are often time-consuming, labor-intensive, and have poor robustness. Learning-based methods are limited by the network input of deep learning. They only use geometric images or projection methods for training, which results in information loss for the 3D model. Table 5 summarizes the above works' datasets, data types, methods, advantages and disadvantages. It is worth noting that the inputs of these methods or models are generally static 3D facial models or dynamic 3D sequences, except that the methods based on 2D projection convert the triangle mesh sequence into a deep image sequence as input. This differs from the recent AU detection tasks, which will be discussed in the next section. Table 6 gives the facial expression recognition results for the above work on the 4D dataset.

4.1.2. Expression retrieval

Danelakis et al. [43–45] explored the field of dynamic 3D facial expression retrieval and proposed three works. In 2014, the authors first proposed the GeoTopo descriptor [43], which captures the geometric and topological information of the face. The geometric information is captured by a 2D function G(i,j) that captures the maximum curvature of eight facial landmarks. Another 2D function, T(i,j), represents the topological information, which selects ten features, including one angle, four areas, and five distances features. The authors used 3D dynamic sequences of three expressions (anger, happiness, and surprise) from 101 subjects in the BU-4DFE dataset for experiments and the retrieval results are used for unsupervised 3D facial expression recognition, as shown in Table 6.

In 2015, the author improved the descriptor [44] by changing the T(i,j) function that captures topological information from ten to six features, including two facial area features and four facial

Table 5
Summary of 4D facial expression recognition methods.

Literature	Dataset	Input data	Method	Advantages	Disadvantages
Maalej et al. [35]	BU-3DFE, BU-4DFE	3D mesh sequence	Iso-geodesic + HMM	Low cost calculation based on shape analysis method	Manual feature extraction, traditional classification method HMM
Amor et al. [36]	BU-4DFE	3D mesh sequence	DSF+LDA+HMM/Random Forest	The radial curve to represent the face shape can effectively capture the subtle deformation of the face	Nose tip detection in non-frontal view or in the presence of occlusion
Berretti et al. [37]	BU-4DFE	3D mesh sequence	Local features based on facial landmark points + HMM	Automatically detects facial marker points and considers their local features to form feature descriptors	Not suitable for detection in low resolution and obscured situations
Xue et al. [39]	BU-4DFE	Triangular mesh	3D-DCT+Nearest Neighbor	Extraction of spatio-temporal features of expressions by frequency domain	Facial marker point detection still uses 2D texture image information
Duh et al. [40]	BU-4DFE	Depth sequences	Extracts HOG and HOF of STIPs from depth sequences + SVM	Contains in-depth information	Depth maps have information loss
Zhen et al. [38]	BU-3DFE, BU-4DFE	3D static meshes or 4D faces	MMM + HMM	Solve face segmentation, shape representation and feature fusion problems	Non-end-to-end
Li et al. [41]	BU-4DFE	2D geometric images	Dynamic geometry Image Network	Generating geometric images from estimated Differential geometry quantities to describe 3D faces	Geometric images have information loss compared to the original 4D data
Behzad et al. [42]	Bosphorus, BU-3DFE, BU-4DFE, BP4D-Spontaneous	2D images of left, front and right views	Multi-view transformer	Multi-view based transformer, no need to extract features manually	Projection of the input from a 3D/4D grid to a 2D image, with loss of information

Table 6
Expression recognition rates(RRs) of different methods on different datasets.

Author	Dataset	Overall
Maalej et al. [35] (2012)	BU-4DFE	93.83
Berretti et al. [37] (2013)	BU-4DFE	79.44
Amor et al. [36] (2014)	BU-4DFE	93.21
Xue et al. [39] (2015)	BU-4DFE	78.8
Duh et al. [40] (2016)	BU-4DFE	77.7
Zhen et al. [38] (2016)	BU-4DFE	87.06
Danelakis et al. [43] (2014) (3 expressions)	BU-4DFE	96.67
Danelakis et al. [44] (2015) (3 expressions)	BU-4DFE	**99.67**
Danelakis et al. [44] (2015)	BU-4DFE	90.83
Danelakis et al. [45] (2016)	BU-4DFE	90.0
Danelakis et al. [45] (2016)	BP4D-Spontaneous	88.56
Rashid et al. [6] (2018)	BP4D-Spontaneous	93.7
Danelakis et al. [53] (2018) (24AUs)	BP4D-Spontaneous	90.60
Li et al. [41] (2018)	BU-4DFE	92.22
Behzad et al. [42] (2021)	BU-4DFE	99.66
Behzad et al. [42] (2021)	BP4D-Spontaneous	91.67

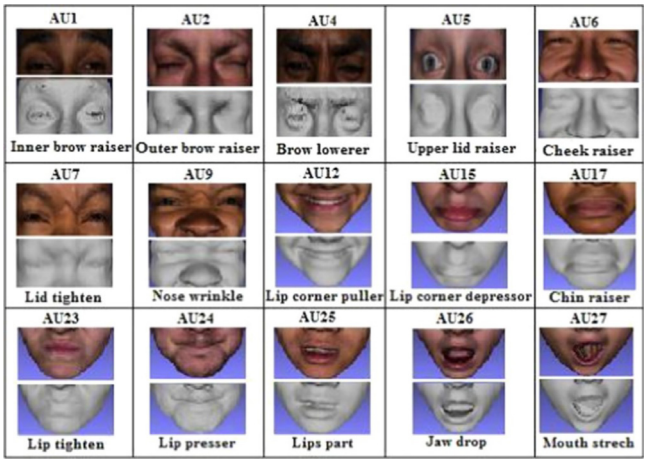


Fig. 5. Basic AUs (3D models and corresponding 2D images) [45].

distance features. They also applied Discrete Cosine Transform (DCT) to map the features from the temporal domain to the frequency domain to construct the final descriptor ST. The experiments showed that the improved method achieved better retrieval results on the three expressions in the BU-4DFE dataset.

In 2016, the author proposed the enhanced version of the GeoTopo descriptor: GeoTopo+ [45]. GeoTopo+ consists of three parts: the first part is the geometric descriptor, which uses a variant of the typical heat kernel signature (HKS) and applies appropriate constraints in the temporal domain to prevent shape perturbation invariability and to capture shape perturbations better than scalar curvature. The second part is the geometric descriptor, which represents the normal vectors of the *j*th facial landmark point in the *i*th facial frame. The third part is the topological descriptor, the same as the T(i,j) function in the GeoTopo descriptor. In the experimental section, six expressions from the BU-4DFE dataset and eight expressions from the BP4D-Spontaneous dataset were used for experiments, achieving 90.00% and 88.56% accuracy, respectively.

Table 7 summarizes the above works' datasets, data types, methods, advantages and disadvantages.

4.1.3. AU detection

Friesen and Ekman proposed the Facial Action Coding System (FACS) [77] in 1978, which classifies the movement of specific muscles or the activation of a muscle group in facial expressions by combining facial muscle action units (AUs) (see Fig. 5). Initially, they defined 46 different AUs [78], which can be combined in over 7000 ways to describe the details of facial expressions. Table 8 shows the relationship between basic emotion categories and action units [79,80].

Table 7
Summary of 4D facial expression retrieval methods.

Literature	Dataset	Input data	Method	Advantages	Disadvantages
Danelakis et al. [43]	BU-4DFE	4D	GepTopo + KNN	Capturing face topology and geometry information	Artificially defined descriptors, search accuracy needs to be improved
Danelakis et al. [44]	BU-4DFE	4D	Frequency domain Descriptors from 10 features to 8	Frequency domain Descriptors saved space requirements compared to GepTopo	
Danelakis et al. [45]	BU-4DFE, BP4D-Spontaneous	4D	Descriptor GeoTopo+	Introducing HKS in GeoTopo+	

Table 8
The basic action units observed in each basic emotion category.

Emotion	Action units	Description
Anger	4, 7, 24	Brow lowerer, Lid tightener, Lip pressor
Disgust	9, 10, 17	Nose wrinkle, Upper lip raiser, Chir raiser
Fear	1, 4, 20, 25	Inner brow raiser, Brow lowerer, Lip stretcher, Lips part
Happiness	12, 25	Lip corner puller, Lips part
Sadness	4, 15	Brow lowerer, Lip corner depressor
Surprise	1, 2, 25, 26	Inner brow raiser, Outer brow raiser, Lips part, Jaw drop

Compared to the coarse-grained classification of several basic facial expressions, AUs can express a variety of human facial expressions, such as embarrassment and awkwardness. Therefore, many researchers have started to use various methods to carry out tasks of AU detection or recognition. These methods can be divided into three categories: 2D image-based methods [6,46–50], 2D video-based methods [51,52], and 3D-based methods [53–56].

2D image-based methods Rashid et al. [6] proposed using capsule networks for facial AU detection. Tu et al. [46] proposed an IdenNet algorithm that is a multi-task network cascade consisting of two sub-tasks: facial clustering and AU detection. Ntinou et al. [47] proposed a heatmap regression method for joint modeling of the localization and intensity of AUs. Li et al. [48] proposed a new framework to unify all semantic relationships and temporal contexts. Shao et al. [49] proposed a framework called JÂA-Net, for joint learning of AU detection and facial alignment tasks. Ge et al. [50] proposed a Multi-level Graph Relational Reasoning Network (MGRR-Net) for facial AU detection. The proposed method was compared with ARL [81] based on global feature learning and JÂA-Net based on local region feature learning.

Literature [6] demonstrates the effectiveness of capsules for modeling AU detection. Literature [46] addresses a problem where appearance changes caused by identity may exceed the one generated by expressions, allowing the trained classifier to support predicting the correct AUs of new subjects. Literature [47] first proposed using variable size heat maps to jointly simulate AUs localization and intensity estimation, which is a simple and effective method. Literature [48–50] uses a self-attention mechanism in the AU detection task to better learn the facial features between different areas or between multiple scales. But the above methods only use a single frame 2D image as the input of the network, lacking the expression of temporal features of facial expressions. The comparison results are shown in Table 9.

2D video-based methods Yang et al. [51] proposed a FACS3D-Net, which integrates 3D and 2D convolutional neural networks. The 3D CNN learns spatiotemporal representations, while the 2D CNN learns spatial representations for each frame. The fully connected layers combine spatiotemporal and spatial representations to achieve multi-label AU detection for each video frame. Chen et al. [52] proposed CaFNet, which uses the context modeled in CaFGraph to extract facial graphic representations that distinguish context, thereby improving AU recognition performance. In addition to learning the spatial features of a single frame image, the above two methods respectively use 3D CNN and spatiotemporal graph convolution to learn temporal context information to improve the performance of AU detection.

3D-based methods According to our survey, only a few existing methods [53–55] directly process 3D data, while literature [56] uses range image. Danelakis et al. [53], based on their previous works on facial expression retrieval, proposed a descriptor consisting of 16 features composed of angles, areas, and distances. Yang et al. [54] proposed an adaptive multimodal fusion model (AMF), demonstrating that the multimodal fusion method outperformed the single-modal-based methods in AU detection tasks. Reale et al. [55] proposed a new architecture called Local Continuous Point Network (LCPN). This is the first work to use a 3D points cloud for facial expression analysis. LCPN combines the global features of PointNet [82] with local features from PCCN [83], introducing a single PCCN layer after the coordinate transformation layer into PointNet. Zhang et al. [56] proposed an end-to-end Multi-Head Fused Transformer (MFT) algorithm.

Table 10 summarizes the different AU detection methods. Table 11 presents the F1 score results for AU detection on the BP4D-Spontaneous dataset using the above mentioned methods. It can be observed that AMF [54] performs better than other methods in most AU F1 scores, except for [53–55], which are based on 3D data processing instead of 2D data as input in other methods. The method proposed in literature [55] takes 3D point clouds as input. It only uses the position of points as features, lacking information about point neighborhood relationships, so its results are not optimal. In comparison, AMF [54] designs a multimodal fusion approach, using 2D images and depth maps obtained from the projection of 3D meshes as inputs to the network on the BP4D-Spontaneous dataset. However, the depth map loses information about the 3D face model, so future works can explore further deep learning based on 3D models to achieve higher AU detection results.

4.2. Generation tasks

Inspired by the tremendous success of GANs and other generative models in various 2D image generation tasks, researchers have begun to explore research based on 4D facial data, such as 3D facial reconstruction, 3D facial expression generation, and 3D facial animation generation. These studies can be applied to computer games, virtual reality (VR), film production, video conferencing, and 3D simulation. Based on research on 4D datasets, we divide generative studies into three categories: 3D facial reconstruction, facial expression generation, and facial animation generation.

Table 9
Comparison of average F1-frame and average accuracy of AU detection on each dataset.

Methods	Avg F1-frame(%)			Avg Accuracy(%)		
	ARL [81]	JÂA-Net [49]	MGRR-Net [50]	ARL [81]	JÂA-Net [49]	MGRR-Net [50]
BP4D-Spontaneous	61.1	62.4	**63.7**	78.2	78.6	**79.7**
DISFA	58.7	63.5	**68.2**	93.3	94	**95.2**
GFT	/	53.7	/	/	90.5	/
BP4D+	54.6	**56.8**	/	79.6	**82.1**	/

Note: / indicates that no experiments were performed.

Table 10
Summary of different methods of AU detection.

Literature	Method	Dataset	Data type	Advantages	Disadvantages
Danelakis et al. [53] (2018)	Feature Descriptors + SVM	BP4D-Spontaneous	3D mesh sequences	The resulting descriptors are defined using only eight extracted facial marker points to represent 24 AUs	Limited by the extraction of facial landmark points
Rashid et al. [6] (2018)	Capsule network	BP4D-Spontaneous, DISFA	2D images	Detection of AU using capsule networks with more expressive capsules	Capsule networks are slow due to dynamic routing protocols
Reale et al. [55] (2019)	Local Continuous PointNet (LCPN)	BP4D-Spontaneous, BP4D+	Point cloud	Combining the advantages of PointNet and PCCN is less restrictive compared to triangle grid methods	Only point locations are used as features, currently using static data
Tu et al. [46] (2019)	IdenNet	BP4D-Spontaneous,UNBC-McMaster, DISFA	2D images	Subtracting identity features in AU detection networks better presents differences in AUs	Non-3D data
Yang et al. [51] (2019)	FACS3D-Net	BP4D+	2D images	Added consideration for temporal context, better detection performance for subtle AU	2D images add temporal dimension for 3D convolution, the input is not 3D
Yang et al. [54] (2020)	Adaptive multimodel fusion model	BP4D-Spontaneous, BP4D+	2D multimodal images	Adaptive feature fusion from different modalities	3D facial model uses 2D depth map as input, with information loss
Ntinou et al. [47] (2021)	Heat map regression network	FERA2015, DISFA, FERA2017	2D images	Combined AU localization and AU intensity tasks for direct regression of AU intensity levels with heat maps	Non-3D data
Chen et al. [52] (2021)	CaFNet	BP4D-Spontaneous, DISFA	2D images	CaFGraph can be used for almost all fine-grained facial behavior analysis tasks	Non-3D data
Li et al. [48] (2021)	Integrated network of semantic and temporal relationships	BP4D-Spontaneous, DISFA	2D images	Simultaneously considering the correlation of facial muscles and the temporal context dependence of facial features	Non-3D data
Zhang et al. [56] (2021)	Multi-Head Fused Transformer	BP4D-Spontaneous, BP4D+	2D images, depth maps, thermal images	End-to-end AU detection without a priori such as facial marker points, AU semantic descriptions, etc.	The depth map is used, no 3D feature representation is used
Shao et al. [49] (2021)	Local Continuous PointNet (LCPN)	BP4D-Spontaneous, BP4D+	2D images	Joint learning for AU detection and facial alignment tasks	Non-3D data
Ge et al. [50] (2022)	MGRR-Net	BP4D-Spontaneous, DISFA	2D images	Multi-level feature learning from local area and global face	Non-3D data

Table 11
F1 scores of different methods for AU detection in BP4D-Spontaneous dataset (%).

Action Unit	Rashid et al. [6] (2018)	LCPN [55] (2019)	IdenNet [46] (2019)	AMF [54] (2020)	CaFNet [52] (2021)	Li et al. [48] (2021)	MFT [5] (2021)	JÂA-Net [49] (2021)	MGRR-Ne [50] (2022)
AU1	46.8	43.6	50.5	**55.1**	55.1	54.0	51.6	53.8	52.6
AU2	29.1	41.7	35.9	**58.3**	49.3	46.0	49.2	47.8	47.9
AU4	52.9	56.7	50.6	**62.0**	57.7	55.7	57.6	58.2	57.3
AU6	75.3	74.7	77.2	**82.5**	78.3	79.4	78.8	78.5	78.5
AU7	77.6	71.2	74.2	75.6	78.6	**78.8**	77.5	75.8	77.6
AU10	82.4	81.0	82.9	**87.2**	85.1	84.5	84.4	82.7	84.9
AU12	81.2	84.5	85.1	**89.6**	86.2	87.0	87.9	88.2	88.4
AU14	55.5	55.5	63.0	60.9	67.4	67.0	65.0	63.7	**67.8**
AU15	32.6	40.7	42.2	**59.1**	52.0	55.6	56.5	43.3	47.6
AU17	60.6	57.6	60.8	62.4	**64.4**	63.1	64.3	61.8	63.3
AU23	41.3	39.3	42.1	45.0	48.3	**50.7**	49.8	45.6	47.4
AU24	–	44.1	46.5	52.0	**56.2**	55.3	55.1	49.9	51.3
Avg	57.6	57.6	59.3	**65.8**	64.9	64.8	64.8	62.4	63.7

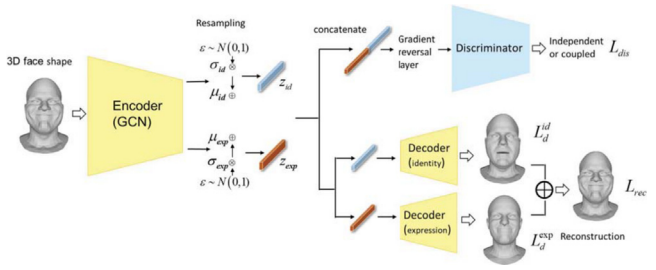


Fig. 6. DI-MeshEncoder structure [2] (courtesy of Prof. Huibin Li).

4.2.1. 3D facial reconstruction

3D facial reconstruction can be divided into two categories based on the type of input data: one is based on a 3D facial model as input and aims to design models that can separate the identity and expression of 3D facial meshes to achieve different facial expression reconstruction effects; the other is focused on reconstructing the 3D geometric shape of the face from one or more 2D images.

(1) Facial Reconstruction from 3D input

Li et al. [33] proposed a 3D facial modeling model called FLAME by learning facial models from thousands of accurately aligned 3D scans. FLAME applies the SMPL [84] body modeling method to the head and adds expression-mixed shapes. It learns the linear shape space of the face from 3800 frames of scanned data of facial heads, combines other global expression-mixed shapes from 4D facial sequences in the D3DFACS dataset and other 4D sequences, and uses 33,000 scans for learning. The resulting FLAME model is highly expressive and can be applied to facial modeling, expression conversion, and 2D-to-3D facial reconstruction tasks.

Zhang et al. [2] proposed a Distribution Independent Variational MeshEncoder (DI-MeshEncoder) method. DI-MeshEncoder consists of a shared encoder, two decoders, and a discriminator, as shown in Fig. 6. First, a shared encoder based on Graph Convolutional Networks (GCN) is designed to output the latent space distribution of the identity and expression of the given 3D facial shape separately. Then, two decoders consisting of fully connected layers reconstruct the identity and expression parts of the input 3D face from the two distributions. Meanwhile, a discriminator composed of two fully connected layers enhances the potential distribution independence between expression and identity. The experiments were conducted on three datasets, CoMA, D3DFACS, and BU-3DFE, and the results showed that the proposed method had excellent applications in facial decomposition, reconstruction, and expression conversion.

Both of the above works belong to the first category, which is to perform attribute decomposition on 3D facial models to decouple them into shape features, pose features, and expression features or identity and expression features for 3D facial reconstruction. The visualization results of the reconstructed faces are shown in Fig. A.1.

(2) Facial Reconstruction from 2D input

Another 3D facial reconstruction problem aims to restore the 3D geometric shape of the face from 2D images or videos. Cheng et al. [3] proposed a lightweight framework based on GCN. The framework consists of two connected sub-networks. The method was tested on the Florence [85], BU-3DFE, and 4DFAB datasets, and the results showed that the authors' method was fast and lightweight. Liu et al. [57] proposed an encoder–decoder network structure that can directly learn the raw 3D scan data for facial modeling in multiple 3D datasets. Zhang et al. [58] proposed a real-time 4D facial expression capture solution to generate personalized 4D faces using only a consumer-grade RGB-D camera and CPU-based computation. As shown in Fig. 7, the method consists of three steps. The method's main advantage is its speed, with an average time of only 40 ms required to obtain the target face with 3D facial landmarks. The reconstructed visualization results of the above methods are given in Fig. A.7.

Recently, with the development of VR/AR technology, the 3D head avatar generation has opened up a wide range of applications in communication and entertainment. Sun et al. [59] proposed a new GAN framework called Next3D. It can synthesize high-quality and 3D consistent facial portraits from unstructured 2D images and achieve precise control of full head rotation, facial expressions, eye blinking, and gaze direction. Grassal et al. [60] proposed an explicit 3D human head model: Neural Head Avatars. This method accurately reconstructs the geometry and appearance of the human head from the monocular RGB sequence. This method combines the parametric head model with a Multilayer perceptron, and the resulting 4D head image is robust to large changes in posture, view, and expression. Fig. A.2 shows the qualitative results of this method on the synthetic subject. Zheng et al. [61] combined the ideas of 3DMM and implicit modeling to implement an iterative 3D facial reconstruction method based on the monocular video: IMavatar. The synthetic dataset obtains FLAME expression parameters from the VOCA dataset and fits head posture from real videos. Build a test set with more extreme expressions from CoMA. The resulting 3D facial model is realistic and can be edited in a 3DMM manner. Fig. A.3 shows the qualitative results of this method on the synthetic data. Subsequently, the author decomposed the source color into intrinsic albedo and normal dependent shading and proposed a point-based deformable representation: PointAvatar [62]. This method can generate animated 3D avatars from hand-held smartphones, laptop webcams, and internet videos for monocular videos. The method has demonstrated effectiveness for eyeglasses, voluminous hair, skin details, and extreme head positions. The above works all mentioned are related to the work of (neural radiance fields) NeRF [86] or head avatars based on NeRF, indicating the enormous potential of NeRF in the field of 3D head avatar generation.

Table 12 summarizes the different reconstruction methods. Table 13 summarizes the quantitative results of the above methods regarding reconstruction error according to Average Vertex Distance (AVD), Normalized Mean Error (NME) or Mean Shape Error (MSE).

4.2.2. 3D facial expression generation

Early facial expression generation methods often required a lot of manual work. With the development of dynamic 3D scanning technology, facial geometric movements of dense point clouds can be accurately obtained. Wang et al. [87] proposed a data-driven expression synthesis method. This method first captures high-speed and high-accuracy moving faces, then uses a multi-resolution deformable mesh to track facial movements, and then separate expression styles through a unified low dimensional dynamic facial motion mapping, thereby synthesizing novel facial expressions. In this method, the goal of expression synthesis is to learn a decomposable generative model, that is, to transform human expression style without changing the internal facial configuration of the face. In recent years, with the development of deep learning, existing 3D works [7,63–65] on facial expression generation are primarily implemented based on GANs and have achieved significant results.

Liu et al. [63] proposed a local attentive conditional generative adversarial network (LAC-GAN) based on facial action unit (AU) annotation, as shown in Fig. 8. Cheng et al. [7] proposed the first GAN structure that operates on 3D meshes, called MeshGAN. This method uses convolution on meshes to generate 3D meshes.

Table 12
Summary of different reconstruction methods.

Literature	Method	4D Dataset	Advantages	Disadvantages
Li et al. [33] (2017)	FLAME	D3DFACS	Realized feature decoupling of shape, posture, and expression	Only considered the face's outer surface, without considering the correlation between the face and the head bones
Zhang et al. [2] (2020)	DI-MeshEncoder	CoMA, D3DFACS	Using the structure of encoder–decoder to reconstruct identity and expression information separately	Further improvements can be made to apply the learned features to facial recognition, retrieval and other areas
Cheng et al. [3] (2020)	GCNs	4DFAB	Fast and lightweight (with MobileNet-V2 backbone, the model size is only 37 MB)	–
Liu et al. [57] (2019)	Encoder–decoder structure network	BU-4DFE	Ability to learn potential representations of nonlinear identity and expression features of 3D faces	–
Zhang et al. [58] (2021)	Face mesh deformation	BU-4DFE	Fast, fully automated, low computational and system costs	Kinect data accuracy low
Sun et al. [59] (2023)	Next3D	–	Supports precise control of full head rotation, facial expressions, blinking, and gaze direction	It struggles to model some expressions with full consistency, such as one-side mouth up, sticking tongue out, etc.
Grassal et al. [60] (2022)	Neural Head Avatars	–	The parameterized head model is combined with the Multilayer perceptron so that the Perceptron can refine geometric shapes and synthesize realistic textures	Limited by fixed topology
Zheng et al. [61] (2022)	IMavatar	CoMA, VOCASET	3DMMs provide fine-grained expression control, while implicit surfaces provide the high fidelity geometry and texture details	The method relies on accurate face tracking and performance degenerates with noisy 3DMM parameters
Zheng et al. [62] (2023)	PointAvatar	–	Decompose source color into intrinsic albedo and normal-related shading	The method cannot faithfully model the reflection of eyeglass lenses

Table 13
Summary of quantitative results of reconstruction error (mm).

Methods	CoMA			D3DFACS		4DFAB	BU-4DFE
	Mean AVD	Median AVD	Per-vertex error	Mean AVD	Median AVD	NME	MSE
Li et al. [33]	1.451 ± 1.649	0.871	1.615	1.038 ± 1.141	0.748	/	/
Zhang et al. [2]	0.665 ± 0.748	0.434	/	0.604 ± 0.759	0.38	/	/
Cheng et al. [3]	/	/	/	/	/	7.15	/
Liu et al. [57]	/	/	1.474	/	/	/	/
Zhang et al. [58]	/	/	/	/	/	/	3.59|1.06

Note: / indicates that no experiments were performed.

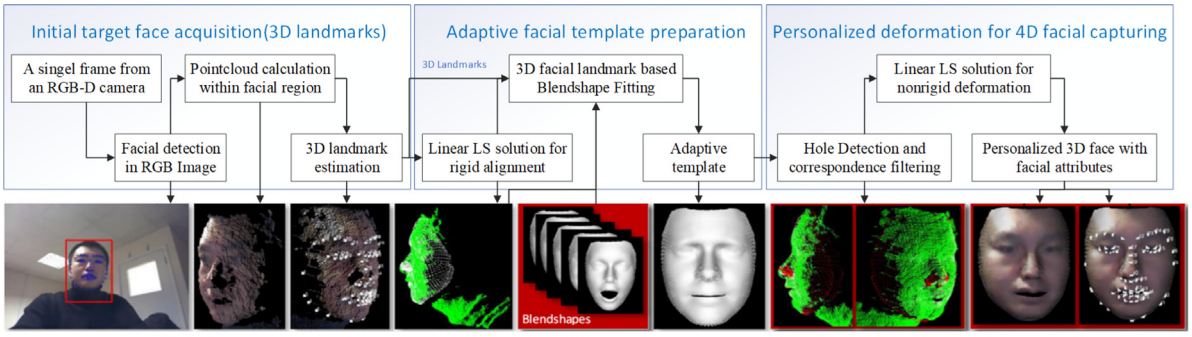


Fig. 7. Real-time 4D facial expression capture solution framework [58].

The network structure is shown in Fig. 9. The expression model was trained on the 4DFAB dataset. Abrevaya et al. [64] proposed decoupling the identity and expression of 3D facial models using GAN to generate realistic 3D faces. The method architecture is shown in Fig. 10. Model training uses BU-3DFE and Bosphorus to provide identity variability and uses BP4D-Spontaneous and BU-4DFE to provide expression variability. Moschoglou et al. [65] proposed 3DFaceGAN (Fig. 11). The primary purpose of the generator in 3DFaceGAN is to retrieve a facial UV map as input and generate a fake facial UV map. An autoencoder is used as the discriminator to distinguish between real and fake facial UV maps. The visualization generation results of the above methods are shown in Figs. A.4, A.5, and A.6 in the Appendix.

Unlike the GANs-based methods mentioned above, Wang et al. [66] proposed a global–local multilinear model-based method. The method first uses a global multilinear model (blue in Fig. 12) to estimate the input neutral mesh and obtain the rough deformation of the new expression. Then, the local model (green in Fig. 12) is optimized through the sparse sampling of the global model results, fitting the rough deformation constraints of the local expression model, but this cannot guarantee the approximate identity of the target object. Therefore, dense sampling is performed on the expression shape. The local identity model is fitted to the generated constraints, producing the final face that retains the predicted identity and expression details. Fig. A.8

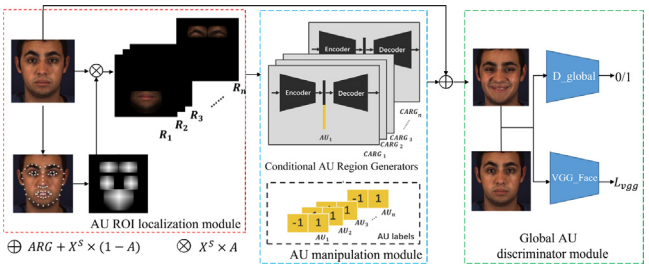


Fig. 8. The general framework of LAC-GAN [63].

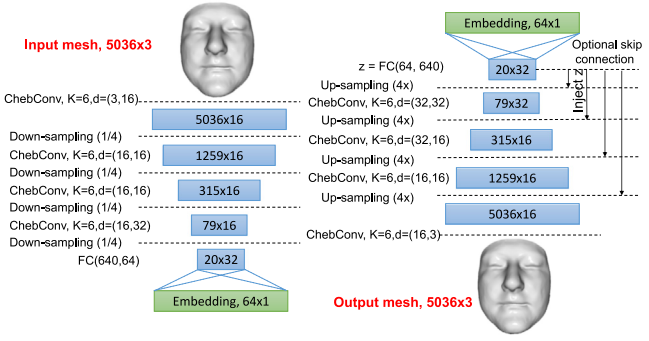


Fig. 9. Network structure of MeshGAN [7].

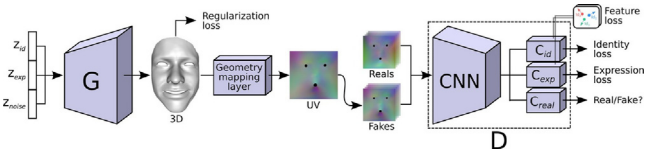


Fig. 10. The methodological structure of the literature [64] (courtesy of Victoria Fernández Abrevaya).

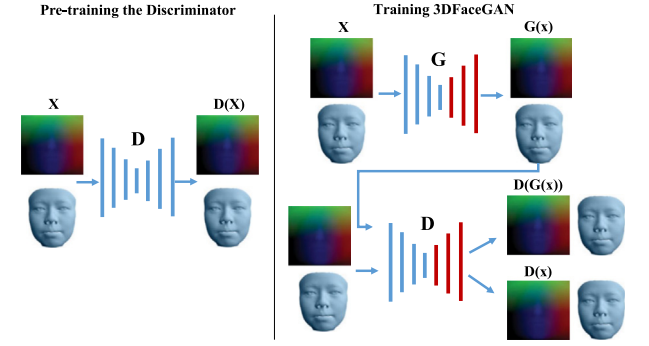


Fig. 11. 3DFaceGAN network structure [65].

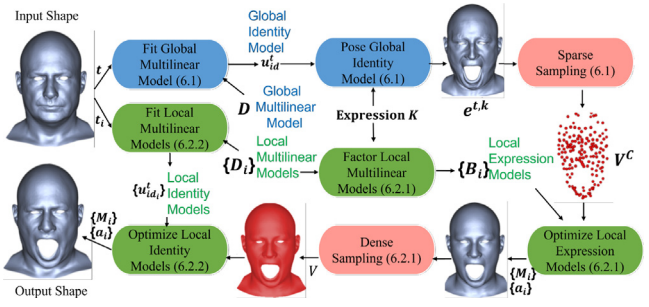


Fig. 12. Global–local multilinear model framework [66].

shows this method's visualized results of 3D facial expressions synthesized.

Table 14 summarizes different methods for facial expression generation. It is worth noting that many methods, whether reconstruction methods in Table 12 or expression generation methods in Table 14, are based on feature decoupling and use 4D datasets to train and learn expression feature information. Because 4D datasets include temporal information, they can show more facial expression details, an advantage that 3D datasets do not have. Therefore, having a sufficiently large number of samples and expression categories in 4D datasets is extremely important for facial analysis research.

4.2.3. 3D facial animation generation

3D facial animation generation can be divided into two types based on different inputs: audio-based animation generation and neutral face-based animation generation. Audio-based animation generation mainly uses audio or combined with the text as network input to generate a 3D dynamic head model. Neutral face-based animation generation also aims to generate a 3D dynamic head model. However, this type of method considers that audio-driven methods ignore the importance of facial expressions and emphasize conducting facial animation generation research using neutral 3D facial mesh combined with audio or target expressions as network input.

(1) Audio-based animation generation

Audio-based head animation generation method is currently more popular in the field of facial animation generation. This type of method combines the speech information from the audio with head movements to generate realistic head animations. Zhang et al. [67] proposed a model combining PoseGAN and PGFace, which takes audio streams as input and outputs 3D talking head sequences with dynamic head movements. PoseGAN consists of a Gpose3D head poses sequence generator and a Dpose discriminator, which generates head pose sequences for 3D heads. Then PGFace generates facial shape parameters, using audio and pose information to generate natural 3D dynamic facial models. Lahiri et al. [68] proposed an end-to-end network called LipSync3d for synthesizing personalized 3D talking faces. This method decouples 3D geometric shape, head pose, and texture and decomposes the prediction problem into regression problems of 3D facial shape and corresponding 2D texture images. Fan et al. [69] proposed a Transformer-based autoregressive model called FaceFormer, which formulates the problem of speech-driven 3D facial animation as a sequence-to-sequence (seq2seq) problem and uses the self-attention mechanism of the Transformer to model the short- and long-term relationships in speech-driven 3D facial animation. Fan et al. [70] also proposed a model based on the encoder–decoder structure that uses audio and text as network input to generate 3D facial animation sequences.

(2) Neutral face-based animation generation

Potamias et al. [4] proposed the first model for generating realistic 3D facial animation with a given target expression and a neutral 3D facial mesh. The method adopts an encoder–decoder network structure, where a recursive LSTM encoder encodes the expected facial movements of the target expression into the latent expression space, which is then decoded by a frame decoder into mesh deformations and output to the neutral facial mesh. In the experiment section, the authors trained the model on 153 samples from the 4DFAB dataset, with 27 used for testing, and set up four stages (neutral, onset, apex, offset) for six basic expressions. Additionally, the authors conducted an in-the-wild expression generation experiment by collecting images of neutral and various expressions of the same individuals and fitting 3D expressions using the method proposed in [88], which resulted in a series of realistic 3D expression sequences.

Table 14
Summary of different expression generation methods.

Literature	Method	4D Dataset	Advantages	Disadvantages
Liu et al. [63] (2020)	AU + GAN	BP4D-Spontaneous	Using AU for facial expression synthesis tasks	Design network with 2D image data, non-3D data
Cheng et al. [7] (2019)	MeshGAN	4DFAB	The first 3D mesh generation model using convolutional GANs architecture directly on the mesh	/
Abrevaya et al. [64] (2019)	2D UV map + GAN	BP4D-Spontaneous, BU-4DFE	Able to capture nonlinear changes caused by expressions and the relationship between identity and expression subspaces	The method is limited by the diversity of training data and the accuracy of labels
Moschoglou et al. [65] (2020)	3DFaceGAN	4DFAB	The first GAN network for 3D face distribution modeling that preserves high frequency details of 3D faces	/
Wang et al. [66] (2020)	Global-local multicollinear model	4DFAB	Results of global and local multilinear methods; separation of local identity and local expression models	The training data needs to be a complete data tensor

Table 15
Summary of different animation generation methods.

Literature	4D Dataset	Input	Output	Method	Advantages	Disadvantages
Zhang et al. [67]	–	Audio stream	3D face sequence	PoseGan +PGFace modules	End-to-end audio generation of 3D head animation network	For large movements of the head posture, the generated facial deformation is obvious
Lahiri et al. [68]	–	Audio stream	3D face sequence	LipSync3d network	Decoupling of 3D poses, geometry, textures and lighting;	–
Fan et al. [69]	VOCASET	Audio stream	3D animation	FaceFormer network	Transformer-based model for long-term audio context encoding with autoregressive prediction of 3D face mesh animation	Not suitable for online streaming, which requires access to the full audio sequence
Fan et al. [70]	–	Audio stream and text	3D animation sequence	Encoder-Decoder network	Combining acoustic and text modalities to predict 3D talking head geometry	–
Potamias et al. [4]	4DFAB	Single neutral 3D mesh, target expression	3D animation	Encoder-Decoder network	The first synthetic facial model from a single neutral face	The next step could be to try to extend to texture prediction
Otberdout et al. [71]	CoMA, D3DFACS	Neutral 3D face, expression labels	3D animation	Motion3DGAN and S2D-Dec decoder	Decoupling the temporal modeling of expressions and the deformation of neutral meshes into two problems	S2D-Dec generates expression-specific deformations, so identity cannot be modeled

Otberdout et al. [71] proposed the Motion3DGAN model, which decouples the time modeling of expressions and the deformation of neutral meshes into two sub-problems of dynamic 3D facial expression generation. Motion3DGAN explains the dynamics of expressions by generating temporally consistent movements of 3D facial landmarks corresponding to the input labels from noise. The movements of facial landmarks are encoded using the square root velocity function (SRVF) and compactly represented as points on a hypersphere. Then, the S2D-Dec decoder is proposed to generate a dense 3D face guided by the movements of facial landmarks for each sequence frame. The authors conducted experiments using the CoMA and D3DFACS datasets. Fig. A.9 shows the ability of Motion3DGAN in 4D facial expression generation.

Table 15 provides a summary of generative methods. The table shows that most audio-based animation generation methods do not require support from 4D datasets but use audio streams as input to the model. However, it is difficult to generate the head posture of the large movement and is not suitable for online streaming media. On the other hand, neutral face-based methods achieve realistic facial animation by deforming the neutral mesh input from 4D datasets.

4.3. Other tasks

In addition to the expression recognition and generation tasks, several other tasks are based on 4D datasets, such as alignment, facial recognition, and facial disease diagnosis.

4.3.1. Facial registration

Wang et al. [72] proposed an automatic non-rigid registration algorithm based on least-squares conformal mapping. The algorithm first maps the 3D face surface to the 2D domain with global optimization through the least-square conformal mapping theory. It simplifies the 3D surface registration problem to the 2D registration problem. Then, the Active Appearance Model (AAM) [89–91] is used to obtain the initial inter-frame feature correspondence of the 3D dynamic face. Then, the least squares conformal maps (LSCMs) of the initial corresponding features are calculated to achieve inter-frame registration of the face surface. In addition, based on the proposed non-rigid registration method, this paper also proposes a new framework for dynamic facial expression synthesis and transfer.

Mehdi et al. [73] conducted a face registration task using nine 3D or 4D large-scale face datasets, including BU-4DFE and 4DFAB. They treated the registration task as a face-to-face transformation problem and designed the Shape My Face (SMF) network structure, an encoder–decoder structure based on an improved PointNet. SMF directly captures the latent geometric information (3DMM parameters) from the original 3D face scans for encoding and decoding registration. The network structure is shown in Fig. 13. During the network training process (indicated by dashed lines), the authors measure the registration fit between the network output and the dynamically sampled input point cloud to ensure that the reconstructed vertices can match any point on the scanned surface.

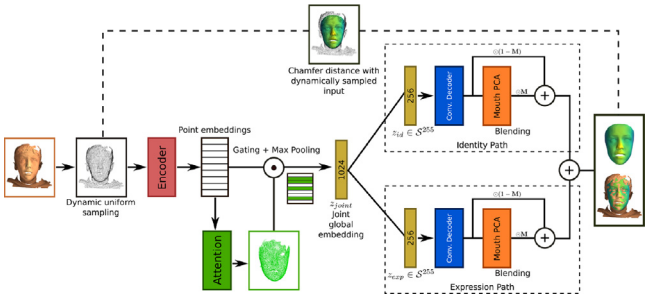


Fig. 13. SMF Network Structure [73] (courtesy of Mehdi Bahri).

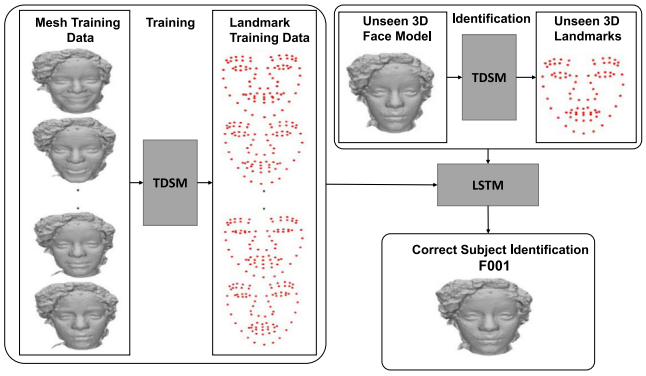


Fig. 14. TDSM method overview diagram [74].

4.3.2. Facial recognition

Due to the complexity of 3D facial data, the 3D facial recognition algorithm needs more complex and fine feature extraction and matching algorithms to achieve better performance. Literature [8,74,75] proposed three different networks.

Jannat et al. [74] proposed a Temporal Deformable Shape Model (TDSM). This model first detects 83 facial landmarks on 3D data. Then align the centroid of the face and the original point in the 3D space. The transformed 3D facial features are used to object recognition. The authors experimented with SVM, Random Forest (RF), and LSTM on the BU-4DFE, BP4D-BP4D-Spontaneous, and BP4D+ datasets, achieving recognition accuracy of over 99%. Fig. 14 provides an overview of the method. An example demonstrates the correct identification of subject F001 from BP4D+ using an LSTM trained based on 3D facial data detected from TDSM on an invisible 3D mesh model.

Kacem et al. [75] proposed a Space-Time Triplet Loss Network for dynamic 3D face verification. First, the 3D facial features are encoded into a low-dimensional representation that describes local deformations of the face relative to the mean face. Then, the encoded versions of the 3D faces along the sequence are stacked into a 2D array for temporal modeling. The resulting 2D array is fed into a triplet loss network for dynamic sequence embedding. Finally, cosine similarity is used to compare the output of the triplet loss network for face verification. Fig. 15 provides an overview of the method.

Papadopoulos et al. [8] proposed a graph convolutional network for 3D dynamic facial recognition, called FaceGCN, as shown in Fig. 16. The method first estimates 3D facial landmarks using 2D texture facial images and their mappings to 3D facial meshes. A spatiotemporal graph is constructed using the 3D facial landmarks, where the nodes contain appearance and shape features extracted from the neighborhood of each facial landmark. Then, a spatiotemporal graph convolutional network (ST-GCN) [92] is used for facial classification. The average recognition accuracy results of the above three facial recognition methods on the BU-4DFE dataset are given in Table 16.

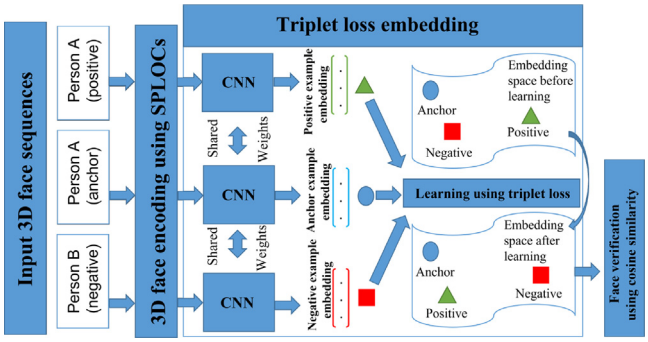


Fig. 15. Overview of the literature [75] methodology.

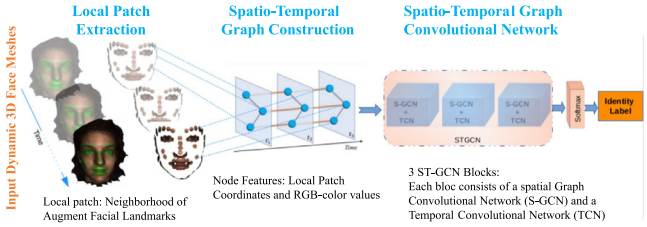


Fig. 16. Overview of the Face-GCN [8].

Table 16
Average recognition accuracy of different methods in BU-4DFE.

Method	Average accuracy (%)
Kacem et al. [74]	96.82
TDSM(SVM) [75]	99.9
TDSM(RF) [75]	100
TDSM(LSTM) [75]	100
Face-GCN [8]	88.45

4.3.3. Facial disease diagnosis

As stated in Section 4.1.2, AU detection is crucial in detecting overall and local muscle region changes in the face and facial expression. This makes AU detection widely applicable, such as in the automatic estimation of facial paralysis. Ge et al. [76] proposed an end-to-end model called Adaptive Local-Global Relational Network (ALGRNet). Specifically, the network first extracts global features based on meshes using a Stem Network composed of multiple convolutional layers and extracts local AU features from calculated regions based on detected AU centers. Then, a facial landmark point localization network is used to detect facial landmarks. Next, the Skip-BiLSTM module mines positive and negative relationships between different AU branches through different information transmission options. Then, each branch's feature fusion and refinement module helps complement local AU regions under the guidance of global features based on meshes. Finally, a multi-label binary classifier is used to predict the probability of single AU activation, and experiments were conducted on the BP4D-Spontaneous and DISFA datasets. To evaluate the effectiveness of ALGRNet in estimating the severity of facial paralysis, the authors used a facial paralysis dataset from the UK National Health Service (NHS) called FPara, which includes 89 videos showing facial paralysis patients performing various types of facial paralysis exercises, ranging from grade 1 to grade 6, where one is normal and six is the most severe. The authors further divided it into four levels: normal, mild, moderate, and severe. The method achieved the best results for estimating facial paralysis at the four levels, with an average F1-frame score of 75.4%. This method demonstrates the effectiveness of AU detection and the transferability from AU detection to the automatic estimation of facial paralysis.

Table 17
The main advantages and disadvantages of 2D, 3D and 4D methods.

Comparison	2D	3D	4D
Illumination change, pose change	2D images and videos will be affected	3D data is unaffected by illumination and is robust to attitude changes	4D data is not affected by illumination and is robust to attitude changes
Data acquisition	Easy access, digital cameras, cell phones and other 2D imaging devices	Three-dimensional imaging equipment such as three-dimensional scanners	Professionally built 4D imaging equipment
Available data volume	Large public datasets	A certain number of data sets are available and are expected to continue to grow	Few datasets available
Calculation costs and storage costs	Very low	High data dimensionality, high storage and computational costs	Increased timing information and higher storage and computational costs than 3D
Facial surface measurement	Inaccuracy	3D enables realistic facial surface measurement	Enables moving measurement of the same part over time
AU detection	End-to-end deep learning network methods are abundant, and the mainstream methods at this stage	A small number of methods using static point clouds or depth maps, or multimodal (2D images, depth maps, thermal imaging images) as network inputs perform best	Limited by the development of deep learning networks for 3D data and the limited availability of 4D datasets, there is currently a lack of end-to-end network learning methods
Real-time	Small amount of data, easy to real-time	High data volume and equipment requirements	Continuous processing of each frame of 3D data, poor real-time
Expression analysis	Inability to render three-dimensional facial features	Show the face in 3D, easy to extract three-dimensional spatial features	Easy to extract the temporal characteristics of expressions and facial muscles

5. Discussion

5.1. 2D, 3D and 4D advantages and disadvantages

3D facial data overcomes problems encountered in solutions based on 2D models, but 3D technology also has disadvantages. 4D data provides more information than 3D static data due to the addition of temporal information. Table 17 lists the main advantages and disadvantages of 2D, 3D, and 4D methods.

A study [93] pointed out that when a person tilts their head upward, it is often accompanied by emotions such as contempt, pride, and happiness. Conversely, tilting the head downwards is often associated with sadness, inferiority, shame, embarrassment, and guilt. This study preliminarily determined that human emotions will be affected by changes in head viewpoints. We know that 2D images or videos are always presented from a fixed viewpoint. In contrast, changes in viewpoint in 3D or 4D data may be more helpful for improving the recognition performance of human emotions.

5.2. Influence of different datasets attributes on 4D applications

Through the aforementioned research on existing 4D datasets, it is illustrated that different attributes of the dataset, such as collected data modes, scanned head coverage, etc. al., has different impacts on various applications.

4D facial datasets scanned head coverage is different, some are inner faces, and others are full head models, which play different roles in various tasks. For early facial expression recognition and retrieval tasks, we found that the 4D dataset used by the researchers in the experiment was not a full head scan. All 4D facial datasets collect facial information, and the main reason is that facial features are the main research object for facial expression-related tasks. In the AU detection task, the main research is the movement state of facial muscles. Therefore, for the expression-related tasks, the inner face or whole head dataset will not affect the AU detection results. For the generation task, we believe that the dataset used is mainly related to the purpose of the generation task itself. The generation technology based on GANs and full head 4D facial datasets(such as CoMA and VOCASET) has gradually appeared. For the generation of 3D head avatars, its main application scenario is in communication or entertainment, and it is more inclined to generate full-head faces. For facial reconstruction from the 2D image as input, the input 2D image itself is only the inner face information of the face. Therefore, this kind of task mainly uses inner face data sets (such as BU-4DFE and 4DFAB) for experiments.

In contrast, the diversity of data modes has a more significant impact on various 4D facial applications. Audio, as important information for 3D virtual humans, 3D animation generation, and other applications, was first collected in the B3D(AC)^2 dataset. In addition, including both audio and 3D facial information is also crucial for studying emotions and their correlations. BP4D+ collects high-definition 3D geometric facial sequences, 2D facial videos, thermal videos, and physiological data sequences, which are significant for expression classification, AU detection, and thermal imaging data classification. For example, the AMF method adopts the multimodal fusion method in AU detection. Table 11 shows that the multimodal method is helpful for AU detection tasks. Therefore, we encourage future 4D datasets to support more information (multimodal, multi-expression, well-annotated, multi-physiological signals, etc.).

5.3. Challenges and future directions

Despite the significant progress made so far in 4D facial analysis, several issues still need to be further explored. Below we summarize some potential future directions.

5.3.1. 4D facial dataset

We think the key issues in the development of the 4D facial dataset will be discussed as follows.

(1) Early 4D datasets generally only included posed expression data, which had high recognition accuracy but tended to exaggerate movements and could not fully represent daily facial movements. BP4D-Spontaneous and other spontaneous expression 4D datasets were subsequently created by setting tasks to elicit spontaneous emotions from participants. However, creating a natural environment for facial expression capture while ensuring spontaneity remains a significant challenge.

(2) Currently, 4D datasets have relatively few Asian participants, and different races may have different spontaneous emotional responses to the same elicitation task. Therefore, we believe that future 4D datasets need to consider differences in race when constructing datasets.

(3) There are currently many 4D datasets available, but due to the lack of normative standards for evaluating different datasets, comparing them is challenging.

(4) We found that few 4D datasets have a specific time interval for each subject. To our knowledge, only 4DFAB has designed four collection records. We believe the time interval may be beneficial for studying the stability of facial expressions in response to specific events or changes in behavior over time. It may also be an exciting direction for medical research on individual emotional stability.

(5) 4D data plays an essential role in micro-expression datasets. Macro-expressions and micro-expressions are two different types of facial expressions. 4DME is currently the only 4D micro-expression dataset, and the authors have found that facial expression recognition based on 4D micro-expressions has advantages over 2D-based methods through preliminary research. This may be because micro-expressions are very subtle movements that occur in a small area of the face and are not always visible in 2D single-view videos. Therefore, we believe that 4D datasets are more necessary for micro-expression because micro-expression is more difficult to recognize than macro-expression, and the duration is shorter. We look forward to the emergence of more 4D micro-expression datasets.

5.3.2. 4D facial research methods

According to our survey, 4D facial research methods have changed from the previous method based on geometric features to the method based on the combination of geometry and deep learning, from the method of studying single image and geometric model to the multimodal method of integrating text, image, sound, video, mesh, thermal imaging and so on. Some of which we think are the key issues in the development of 4D facial research will be discussed as follows.

(1) 4D facial representation method

Currently, 4D facial representation methods mainly convert each 4D facial frame into 2D images through depth mapping or UV unfolding to facilitate the use of mainstream deep learning frameworks. Currently, there is limited research directly on 3D mesh, point clouds or graphs (such as MeshGAN [7], DI-MeshEncoder [2], GCNs [3]). These methods consider each frame of 4D facial data separately, without incorporating the time dimension into the overall consideration. Although Li et al. [33] proposed a face statistical model FLAME can be used on 4D facial data, it is still representing each frame separately. In the future, how to incorporate time dimensions into 4D facial data to explore the overall 4D facial data research is to be expected. In terms of converting the 3D model to a 2D image, using conformal mapping (which can preserve the angle and is a less distorted way of representing the conversion from 3D to 2D) to generate geometry images [94] can store 3D information (such as curvature, normal and etc.), thereby reducing information loss from 3D to 2D conversion, will become a promising representation. At the same time, more direct representation methods for 3D or 4D data are still expected to be further explored.

(2) 4D facial data compression method

Due to the massive amount of data involved in 4D facial data, storing and processing this data poses significant challenges. Therefore, compressive representations of 4D data are crucial. It is an interesting explore in literature [71], the square root velocity function (SRVF) was used to define shape space by encoding the motion of 3D facial landmark points as points on a hypersphere of Hilbert space. As we know that optical flow can estimate pixel motion in image sequences. In addition, superpixels have replaced traditional pixels in many computer vision tasks [95]. Compared to traditional pixels, superpixels contain the structural features of the image and have a smaller scale of data. Supervoxels can view videos as a three-dimensional manifold embedded in color and spatiotemporal space [96]. Therefore, we are eagerly anticipating the development of 4D data compression algorithms based on optical flow, superpixels, or super-voxels.

(3) 4D facial recognition and classification method

Currently, there are two mainstream methods of 4D facial recognition and classification. One is based on encoder–decoder [75] or attention [49,50,56] structures to learn the latent representation of facial features. The other emphasizes learning the temporal features of facial expressions through networks such as LSTM [74] or GCN [8]. In the future, we predict that more methods will be combining AU with static face models for 4D facial analysis.

(4) 4D facial data generation method

Several 4D facial generation methods are based on GAN, such as MeshGAN [7], 3DFaceGAN [65], LAC-GAN [63], and Motion3DGAN [71]. However, as is well known, GANs are difficult to train and may encounter problems such as gradient vanishing or mode collapse during the training process. Recently, diffusion models [97,98] have emerged with more stable training processes and some theoretical interpretability, demonstrating great potential over GAN in generating tasks. We look forward to seeing the performance of the diffusion model in 4D facial generation tasks.

5.3.3. Potential applications based on 4D facial dataset

As we summarized in Section 5.1, the differences between 4D and 2D, and 3D data suggest that the following potential applications are based on 4D facial datasets.

(1) Research based on AUs or combinations of AUs, such as AU annotation, emotion recognition based on AUs, analysis of spontaneous AUs, and micro-expression recognition based on AUs, will still have significant research value.

(2) For tasks such as predicting depression or automatic estimation of facial paralysis, which involve non-primary emotions, research based on AUs detection results for facial paralysis analysis has already emerged. Compared to mainstream basic emotional expression, datasets for non-primary emotions such as depression or facial paralysis are rarer. However, we believe that 4D facial data will have a more significant impact on such tasks.

(3) Generation tasks, such as multimodal conversion and generation between neutral faces or single-frame RGB-D images to generate facial, sound, text, etc., are a trend in future research.

(4) Micro-expression research. 4DME [21] conducted a comparative experiment on multi-view AU detection and three separate single-view AU detections, which showed that in the case of multi-view, the average F1-score and average accuracy of AU recognition were significantly higher than any of the three single views. There is no doubt that 4D data carries more information, improving AUs' detection performance. Therefore, we recommend using 4D micro-expression datasets to study micro-expression.

6. Summary

This paper provides a review of 4D facial datasets. We first point out the advantages of 4D facial datasets and their differences from 2D and 3D datasets. Then, we briefly review the timeline of existing 4D facial datasets. We then provide a detailed introduction to the collection equipment, participant information, collection methods, data processing, and database organization of 4D datasets. Next, we categorize the applications developed based on existing 4D facial datasets. We also collect and classify various features used in facial analysis, including methods before and after the emergence of deep learning. Finally, we discuss the advantages and disadvantages of 2D, 3D, and 4D methods, the influence of attributes of different datasets on 4D applications, and the future directions for building datasets and potential applications based on 4D datasets.

CRediT authorship contribution statement

Yong-Jin Liu: Conceptualization, Methodology, Writing – review & editing, Supervision, Funding acquisition. **Baodong Wang:** Investigation, Writing-Original Draft, Visualization. **Lin Gao:** Writing – review & editing, Formal analysis. **Junli Zhao:** Investigation, Methodology, Writing – review & editing, Supervision, Funding acquisition. **Ran Yi:** Writing – review & editing, Validation. **Minjing Yu:** Writing – review & editing, Formal analysis. **Zhenkuan Pan:** Conceptualization, Methodology. **Xianfeng Gu:** Conceptualization, Methodology, Supervision, Funding acquisiton.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors are unable or have chosen not to specify which data has been used.

Acknowledgments

The authors gratefully appreciated the anonymous reviewers for all of their helpful comments. This work was supported by the National Natural Science Foundation of China under Grant Nos.62002258, 62172247, and 61702293, NIH 3R01LM012434-05S1, NIH1R21EB029733-01A1, NSF FAIN-2115095, Beijing Natural Science Foundation (L222008), Shanghai Sailing Program (22YF1420300), Young Elite Scientists Sponsorship Program by CAST (2022QNRC001), Natural Science Foundation of Shandong Province (No.ZR2019LZH002).

Appendix

See Figs. A.1–A.9.

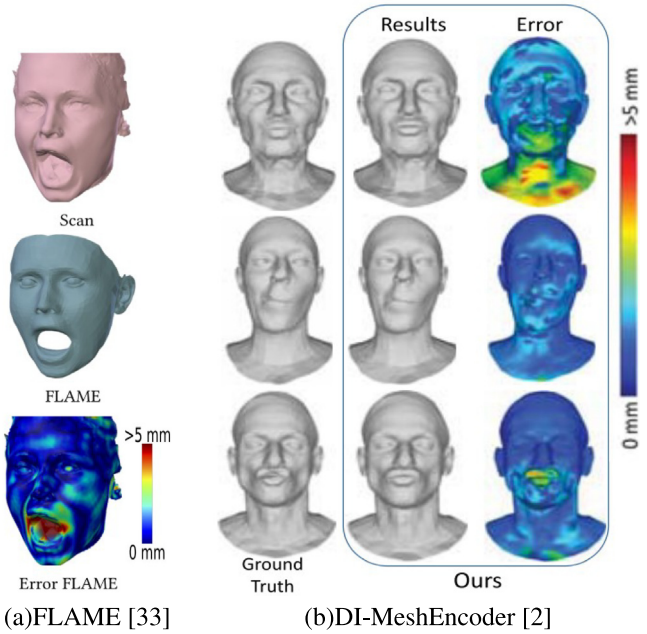


Fig. A.1. Reconstruction of visualization results.

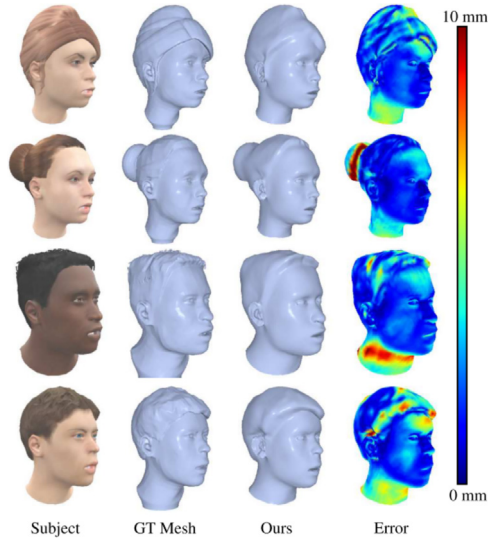


Fig. A.2. Qualitative results on the synthetic subject by Neural Head Avatars [60].

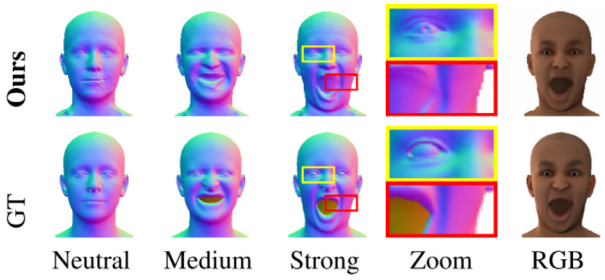


Fig. A.3. Qualitative results on synthetic data by IMavatar [61].

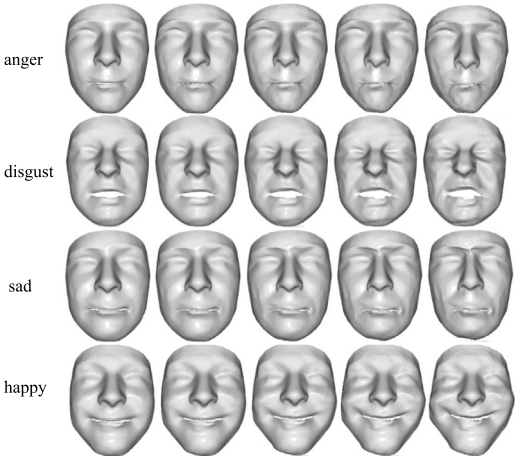


Fig. A.4. Results of different expression generation by MeshGAN [7].

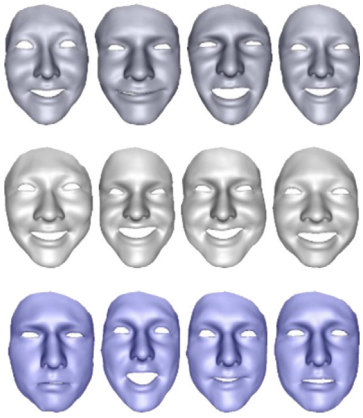


Fig. A.5. Qualitative visualization results from the literature [64]. From top to bottom: randomly generated samples (dark gray), random samples with a same expression code (light gray), random samples with a same identity code (purple).

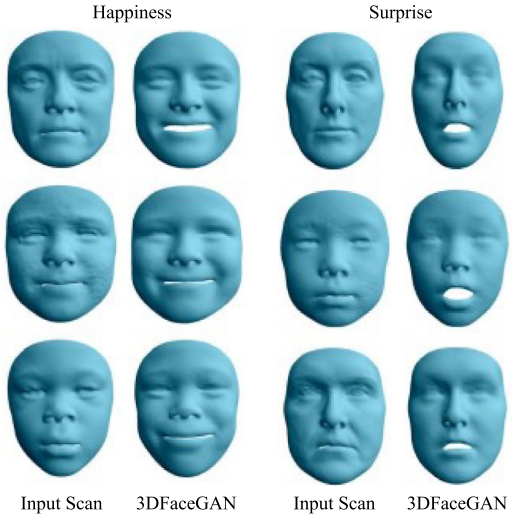
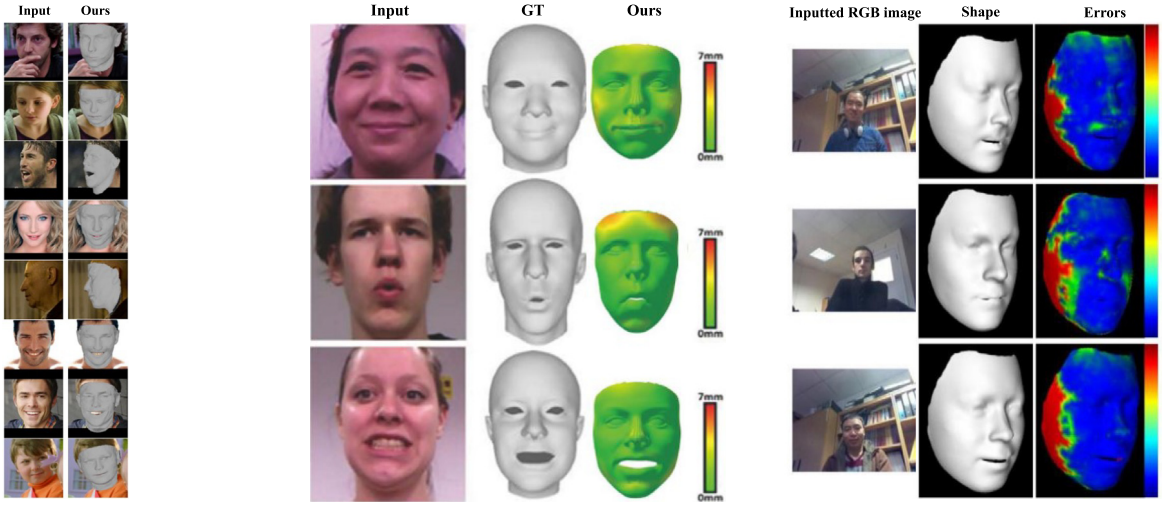


Fig. A.6. Visualization results generated by 3DFaceGAN [65].



(a) Literature [3] (b) Literature [57] (c) Literature [58]

Fig. A.7. The above method reconstructs the visualization results.

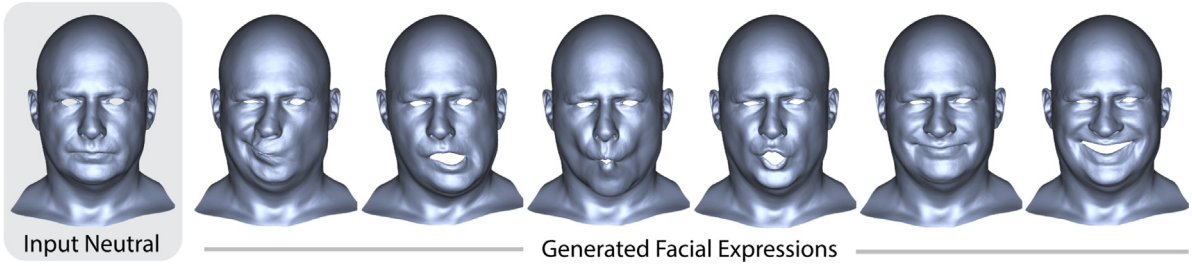


Fig. A.8. The results of the expression visualization generated in the literature [66].

Fig. A.9. Generated 4D expression visualization results [71].

References

[1] Guha R. A report on automatic face recognition: Traditional to modern deep learning techniques. In: 2021 6th international conference for convergence in technology (I2CT). IEEE; 2021, p. 1–6.

[2] Zhang Z, Yu C, Li H, Sun J, Liu F. Learning distribution independent latent representation for 3d face disentanglement. In: 2020 international conference on 3D vision (3DV). IEEE; 2020, p. 848–57.

[3] Cheng S, Tzimiropoulos G, Shen J, Pantic M. Faster, better and more detailed: 3d face reconstruction with graph convolutional networks. In: Proceedings of the Asian conference on computer vision. 2020.

[4] Potamias RA, Zheng J, Ploumpis S, Bouritsas G, Ververas E, Zafeiriou S. Learning to generate customized dynamic 3D facial expressions. In: Computer vision–ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX 16. Springer; 2020, p. 278–94.

[5] Zhang Z, Girard JM, Wu Y, Zhang X, Liu P, Ciftci U, Canavan S, Reale M, Horowitz A, Yang H, et al. Multimodal spontaneous emotion corpus for human behavior analysis. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016, p. 3438–46.

[6] Rashid M, Lee YJ. Facial action unit detection with capsules. 2018, PREPRINT.

[7] Cheng S, Bronstein M, Zhou Y, Kotsia I, Pantic M, Zafeiriou S. Meshgan: Non-linear 3d morphable models of faces. 2019, arXiv preprint arXiv:1903.10384.

[8] Papadopoulos K, Kacem A, Aouada D, et al. Face-GCN: A graph convolutional network for 3D dynamic face recognition. In: 2022 8th international conference on virtual reality. ICVR, IEEE; 2022, p. 454–8.

[9] Jing Y, Lu X, Gao S. 3D face recognition: A survey. 2021, arXiv preprint arXiv:2108.11082.

[10] Zhang S, Huang P. High-resolution, real-time 3D shape acquisition. In: 2004 conference on computer vision and pattern recognition workshop. IEEE; 2004, p. 28.

[11] Gu X. Cutting 3D camera systems. 2023, URL: https://www3.cs.stonybrook.edu/~gu/; (2023, May 8).

[12] Yin L, Chen X, Sun Y, Worm T, Reale M. A high-resolution 3D dynamic facial expression database. In: 8th IEEE international conference on automatic face and gesture recognition (FG 2008), Amsterdam, The Netherlands, 17-19 September 2008. IEEE Computer Society; 2008, p. 1–6. http://dx.doi.org/10.1109/AFGR.2008.4813324.

[13] Fanelli G, Gall J, Romsdorfer H, Weise T, Van Gool L. A 3-d audio-visual corpus of affective communication. IEEE Trans Multimed 2010;12(6):591–8.

[14] Cosker D, Krumhuber E, Hilton A. A FACS valid 3D dynamic action unit database with applications to 3D dynamic morphable facial modeling. In: 2011 international conference on computer vision. IEEE; 2011, p. 2296–303.

[15] Matuszewski BJ, Quan W, Shark L-K, McLoughlin AS, Lightbody CE, Emsley HC, Watkins CL. Hi4D-ADSIP 3-D dynamic facial articulation database. Image Vis Comput 2012;30(10):713–27.

[16] Alashkar T, Amor BB, Daoudi M, Berretti S. A 3D dynamic database for unconstrained face recognition. In: 5th international conference and exhibition on 3D body scanning technologies. 2014.

[17] Zhang X, Yin L, Cohn JF, Canavan S, Reale M, Horowitz A, Liu P, Girard JM. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. Image Vis Comput 2014;32(10):692–706.

[18] Cheng S, Kotsia I, Pantic M, Zafeiriou S. 4Dfab: A large scale 4d database for facial expression analysis and biometric applications. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2018, p. 5117–26.

[19] Ranjan A, Bolkart T, Sanyal S, Black MJ. Generating 3D faces using convolutional mesh autoencoders. In: Proceedings of the European conference on computer vision. ECCV, 2018, p. 704–20.

[20] Cudeiro D, Bolkart T, Laidlaw C, Ranjan A, Black MJ. Capture, learning, and synthesis of 3D speaking styles. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019, p. 10101–11.

[21] Li X, Cheng S, Li Y, Behzad M, Shen J, Zafeiriou S, Pantic M, Zhao G. 4DME: A spontaneous 4D micro-expression dataset with multimodalities. IEEE Trans Affect Comput 2022.

[22] Papaioannou A, Gecer B, Cheng S, Chrysos G, Deng J, Fotiadou E, Kampouris C, Kollias D, Moschoglou S, Songsri-In K, et al. MimicME: A large scale diverse 4D database for facial expression analysis. In: Computer vision–ECCV 2022: 17th European conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VIII. Springer; 2022, p. 467–84.

[23] Ekman P, Friesen W, Hager J. The facial action coding system Second edition. Salt Lake City, London: Research Nexus EBook, Weidenfeld & Nicolson; 2002.

[24] Yan W-J, Wu Q, Liu Y-J, Wang S-J, Fu X. CASME database: A dataset of spontaneous micro-expressions collected from neutralized faces. In: 2013 10th IEEE international conference and workshops on automatic face and gesture recognition. FG, IEEE; 2013, p. 1–7.

[25] Qu F, Wang S-J, Yan W-J, Li H, Wu S, Fu X. CAS $(ME)^2$: A database for spontaneous macro-expression and micro-expression spotting and recognition. IEEE Trans Affect Comput 2017;9(4):424–36.

[26] Husák P, Cech J, Matas J. Spotting facial micro-expressions "in the wild". In: 22nd computer vision winter workshop (Retz). 2017, p. 1–9.

[27] Ben X, Ren Y, Zhang J, Wang S-J, Kpalma K, Meng W, Liu Y-J. Video-based facial micro-expression analysis: A survey of datasets, features and algorithms. IEEE Trans Pattern Anal Mach Intell 2021;44(9):5826–46.

[28] Inc. Di3D. 2023, URL: http://www.di3d.com; (2023, May 3).

[29] Li X, Pfister T, Huang X, Zhao G, Pietikäinen M. A spontaneous micro-expression database: Inducement, collection and baseline. In: 2013 10th IEEE international conference and workshops on automatic face and gesture recognition (Fg). IEEE; 2013, p. 1–6.

[30] Garofolo JS, Lamel LF, Fisher WM, Fiscus JG, Pallett DS, Dahlgren NL. Darpa timit acoustic-phonetic continuous speech corpus CD-ROM {TIMIT}. 1993.

[31] Karras T, Aila T, Laine S, Herva A, Lehtinen J. Audio-driven facial animation by joint end-to-end learning of pose and emotion. ACM Trans Graph 2017;36(4):1–12.
[32] Rajpurkar P, Zhang J, Lopyrev K, Liang P. Squad: 100,000+ questions for machine comprehension of text. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. 2016, p. 2383–92.
[33] Li T, Bolkart T, Black MJ, Li H, Romero J. Learning a model of facial shape and expression from 4D scans. ACM Trans Graph 2017;36(6). 194–1.
[34] Ekman P. Universals and cultural differences in facial expressions of emotion. In: Nebraska symposium on motivation. University of Nebraska Press; 1971.
[35] Maalej A. 3D facial expressions recognition using shape analysis and machine learning (Ph.D. thesis), TELECOM Lille 1; 2012.
[36] Amor BB, Drira H, Berretti S, Daoudi M, Srivastava A. 4-D facial expression recognition by learning geometric deformations. IEEE Trans Cybern 2014;44(12):2443–57.
[37] Berretti S, Del Bimbo A, Pala P. Automatic facial expression recognition in real-time from dynamic sequences of 3D face scans. Vis Comput 2013;29:1333–50.
[38] Zhen Q, Huang D, Wang Y, Chen L. Muscular movement model-based automatic 3D/4D facial expression recognition. IEEE Trans Multimed 2016;18(7):1438–50.
[39] Xue M. Discriminant feature extraction and selection for person-independent facial expression recognition (Ph.D. thesis), Curtin University; 2015.
[40] Duh D-J, Huang J-C, Chen S-Y, Su S, Zhang H, Li S. Facial expression recognition based on spatio-temporal interest points for depth sequences. J Imaging Sci 2016;64(7):396–407.
[41] Li W, Huang D, Li H, Wang Y. Automatic 4D facial expression recognition using dynamic geometrical image network. In: 2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018). IEEE; 2018, p. 24–30.
[42] Behzad M, Li X, Zhao G. Disentangling 3D/4D facial affect recognition with faster multi-view transformer. IEEE Signal Process Lett 2021;28:1913–7.
[43] Danelakis A, Theoharis T, Pratikakis I. Geotopo: Dynamic 3D facial expression retrieval using topological and geometric information. In: Proceedings of the 7th Eurographics workshop on 3D object retrieval. 2014, p. 1–8.
[44] Danelakis A, Theoharis T, Pratikakis I. A spatio-temporal descriptor for dynamic 3 D facial expression retrieval and recognition. In: Proceedings of the 2015 Eurographics workshop on 3D object retrieval. 2015, p. 63–70.
[45] Danelakis A, Theoharis T, Pratikakis I, Perakis P. An effective methodology for dynamic 3D facial expression retrieval. Pattern Recognit 2016;52:174–85.
[46] Tu C-H, Yang C-Y, Hsu JY-j. Idennet: Identity-aware facial action unit detection. In: 2019 14th IEEE international conference on automatic face & gesture recognition (FG 2019). IEEE; 2019, p. 1–8.
[47] Ntinou I, Sanchez E, Bulat A, Valstar M, Tzimiropoulos Y. A transfer learning approach to heatmap regression for action unit intensity estimation. IEEE Trans Affect Comput 2021.
[48] Li Z, Deng X, Li X, Yin L. Integrating semantic and temporal relationships in facial action unit detection. In: Proceedings of the 29th ACM international conference on multimedia. 2021, p. 5519–27.
[49] Shao Z, Liu Z, Cai J, Ma L. Jaa-net: joint facial action unit detection and face alignment via adaptive attention. Int J Comput Vis 2021;129:321–40.
[50] Ge X, Jose JM, Xu S, Liu X, Han H. MGRR-Net: Multi-level graph relational reasoning network for facial action units detection. 2022, arXiv preprint arXiv:2204.01349.
[51] Yang L, Ertugrul IO, Cohn JF, Hammal Z, Jiang D, Sahli H. Facs3d-net: 3d convolution based spatiotemporal representation for action unit detection. In: 2019 8th international conference on affective computing and intelligent interaction. ACII, IEEE; 2019, p. 538–44.
[52] Chen Y, Chen D, Wang Y, Wang T, Liang Y. Cafgraph: Context-aware facial multi-graph representation for facial action unit recognition. In: Proceedings of the 29th ACM international conference on multimedia. 2021, p. 1029–37.
[53] Danelakis A, Theoharis T, Pratikakis I. Action unit detection in 3 D facial videos with application in facial expression retrieval and recognition. Multimedia Tools Appl 2018;77:24813–41.
[54] Yang H, Wang T, Yin L. Adaptive multimodal fusion for facial action units recognition. In: Proceedings of the 28th ACM international conference on multimedia. 2020, p. 2982–90.
[55] Reale MJ, Klinghoffer B, Church M, Szmurlo H, Yin L. Facial action unit analysis through 3d point cloud neural networks. In: 2019 14th IEEE international conference on automatic face & gesture recognition (FG 2019). IEEE; 2019, p. 1–8.
[56] Zhang X, Yin L. Multi-modal learning for AU detection based on multi-head fused transformers. In: 2021 16th IEEE international conference on automatic face and gesture recognition (FG 2021). IEEE; 2021, p. 1–8.
[57] Liu F, Tran L, Liu X. 3D face modeling from diverse raw scan data. In: Proceedings of the IEEE/CVF international conference on computer vision. 2019, p. 9408–18.
[58] Zhang S, Yu H, Wang T, Dong J, Pham TD. Linearly augmented real-time 4D expressional face capture. Inform Sci 2021;545:331–43.
[59] Sun J, Wang X, Wang L, Li X, Zhang Y, Zhang H, Liu Y. Next3d: Generative neural texture rasterization for 3d-aware head avatars. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023, p. 20991–1002.
[60] Grassal P-W, Prinzler M, Leistner T, Rother C, Nießner M, Thies J. Neural head avatars from monocular RGB videos. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022, p. 18653–64.
[61] Zheng Y, Abrevaya VF, Bühler MC, Chen X, Black MJ, Hilliges O. Im avatar: Implicit morphable head avatars from videos. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022, p. 13545–55.
[62] Zheng Y, Yifan W, Wetzstein G, Black MJ, Hilliges O. Pointavatar: Deformable point-based head avatars from videos. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023, p. 21057–67.
[63] Liu Z, Liu D, Wu Y. Region based adversarial synthesis of facial action units. In: MultiMedia modeling: 26th international conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, Proceedings, Part II 26. Springer; 2020, p. 514–26.
[64] Abrevaya VF, Boukhayma A, Wuhrer S, Boyer E. A decoupled 3d facial shape model by adversarial training. In: Proceedings of the IEEE/CVF international conference on computer vision. 2019, p. 9419–28.
[65] Moschoglou S, Ploumpis S, Nicolaou MA, Papaioannou A, Zafeiriou S. 3Dfacegan: Adversarial nets for 3d face representation, generation, and translation. Int J Comput Vis 2020;128:2534–51.
[66] Wang M, Bradley D, Zafeiriou S, Beeler T. Facial expression synthesis using a global-local multilinear framework. In: Computer graphics forum, Vol. 39. Wiley Online Library; 2020, p. 235–45.
[67] Zhang C, Ni S, Fan Z, Li H, Zeng M, Budagavi M, Guo X. 3D talking face with personalized pose dynamics. IEEE Trans Vis Comput Graphics 2021.
[68] Lahiri A, Kwatra V, Frueh C, Lewis J, Bregler C. Lipsync3d: Data-efficient learning of personalized 3d talking faces from video using pose and lighting normalization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021, p. 2755–64.
[69] Fan Y, Lin Z, Saito J, Wang W, Komura T. Faceformer: Speech-driven 3d facial animation with transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022, p. 18770–80.
[70] Fan Y, Lin Z, Saito J, Wang W, Komura T. Joint audio-text model for expressive speech-driven 3d facial animation. Proc ACM Comput Graph Interact Tech 2022;5(1):1–15.
[71] Otberdout N, Ferrari C, Daoudi M, Berretti S, Del Bimbo A. Sparse to dense dynamic 3d facial expression generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022, p. 20385–94.
[72] Wang S, Gu XD, Qin H. Automatic non-rigid registration of 3d dynamic data for facial expression synthesis and transfer. In: 2008 IEEE conference on computer vision and pattern recognition. IEEE; 2008, p. 1–8.
[73] Bahri M, O'Sullivan E, Gong S, Liu F, Liu X, Bronstein MM, Zafeiriou S. Shape my face: registering 3D face scans by surface-to-surface translation. Int J Comput Vis 2021;129(9):2680–713.
[74] Jannat SR, Fabiano D, Canavan S, Neal T. Subject identification across large expression variations using 3D facial landmarks. In: Pattern recognition. ICPR international workshops and challenges: Virtual Event, January 10–15, 2021, Proceedings, Part I. Springer; 2021, p. 5–13.
[75] Kacem A, Abdesslam HB, Cherenkova K, Aouada D. Space-time triplet loss network for dynamic 3D face verification. In: Pattern recognition. ICPR international workshops and challenges: Virtual Event, January 10–15, 2021, Proceedings, Part I. Springer; 2021, p. 82–90.
[76] Ge X, Jose JM, Wang P, Iyer A, Liu X, Han H. Automatic facial paralysis estimation with facial action units. 2022, arXiv preprint arXiv:2203.01800.
[77] Ekman P. Facial action coding system: A technique for the measurement of facial movement. Palo Alto: Consulting Psychologists Press; 1978.
[78] Ekman P, Scherer KR. Handbook of methods in nonverbal behavior research. Cambridge University Press; 1982.
[79] Du S, Tao Y, Martinez AM. Compound facial expressions of emotion. Proc Natl Acad Sci 2014;111(15):E1454–62.
[80] Nonis F, Dagnes N, Marcolin F, Vezzetti E. 3D approaches and challenges in facial expression recognition algorithms—a literature review. Appl Sci 2019;9(18):3904.
[81] Shao Z, Liu Z, Cai J, Wu Y, Ma L. Facial action unit detection using attention and relation learning. IEEE Trans Affect Comput 2019;13(3):1274–89.
[82] Qi CR, Su H, Mo K, Guibas LJ. Pointnet: Deep learning on point sets for 3d classification and segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2017, p. 652–60.
[83] Wang S, Suo S, Ma W-C, Pokrovsky A, Urtasun R. Deep parametric continuous convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2018, p. 2589–97.

[84] Loper M, Mahmood N, Romero J, Pons-Moll G, Black MJ. SMPL: A skinned multi-person linear model. ACM Trans Graph 2015;34(6):1–16.
[85] Bagdanov AD, Del Bimbo A, Masi I. The florence 2d/3d hybrid face dataset. In: Proceedings of the 2011 joint ACM workshop on human gesture and behavior understanding. 2011, p. 79–80.
[86] Mildenhall B, Srinivasan PP, Tancik M, Barron JT, Ramamoorthi R, Ng R. Nerf: Representing scenes as neural radiance fields for view synthesis. Commun ACM 2021;65(1):99–106.
[87] Wang Y, Huang X, Lee C-S, Zhang S, Li Z, Samaras D, Metaxas D, Elgammal A, Huang P. High resolution acquisition, learning and transfer of dynamic 3-D facial expressions. In: Computer graphics forum, Vol. 23. Wiley Online Library; 2004, p. 677–86.
[88] Gecer B, Ploumpis S, Kotsia I, Zafeiriou S. Ganfit: Generative adversarial network fitting for high fidelity 3d face reconstruction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019, p. 1155–64.
[89] Cootes TF, Edwards GJ, Taylor CJ. Active appearance models. In: Computer vision—ECCV'98: 5th European conference on computer vision Freiburg, Germany, June 2–6, 1998 Proceedings, Volume II 5. Springer; 1998, p. 484–98.
[90] Gross R, Matthews I, Baker S. Active appearance models with occlusion. Image Vis Comput 2006;24(6):593–604.
[91] Gross R, Matthews I, Baker S. Generic vs. person specific active appearance models. Image Vis Comput 2005;23(12):1080–93.
[92] Yan S, Xiong Y, Lin D. Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Proceedings of the AAAI conference on artificial intelligence, Vol. 32. 2018.
[93] Mignault A, Chaudhuri A. The many faces of a neutral face: Head tilt and perception of dominance and emotion. J Nonverbal Behav 2003;27:111–32.
[94] Gu X, Gortler SJ, Hoppe H. Geometry images. In: Proceedings of the 29th annual conference on computer graphics and interactive techniques. 2002, p. 355–61.
[95] Liu Y-J, Yu M, Li B-J, He Y. Intrinsic manifold SLIC: A simple and efficient method for computing content-sensitive superpixels. IEEE Trans Pattern Anal Mach Intell 2017;40(3):653–66.
[96] Yi R, Ye Z, Zhao W, Yu M, Lai Y-K, Liu Y-J. Feature-aware uniform tessellations on video manifold for content-sensitive supervoxels. IEEE Trans Pattern Anal Mach Intell 2020;43(9):3183–95.
[97] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. Adv Neural Inf Process Syst 2020;33:6840–51.
[98] Song Y, Ermon S. Generative modeling by estimating gradients of the data distribution. Adv Neural Inf Process Syst 2019;32.