This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2022.3215918

Graph-based Facial Affect Analysis: A Review

Yang Liu, Xingming Zhang, Yante Li, Jinzhao Zhou, Xin Li, and Guoying Zhao*, *Fellow, IEEE*

Abstract—As one of the most important affective signals, facial affect analysis (FAA) is essential for developing human-computer interaction systems. Early methods focus on extracting appearance and geometry features associated with human affects while ignoring the latent semantic information among individual facial changes, leading to limited performance and generalization. Recent work attempts to establish a graph-based representation to model these semantic relationships and develop frameworks to leverage them for various FAA tasks. This paper provides a comprehensive review of graph-based FAA, including the evolution of algorithms and their applications. First, the FAA background knowledge is introduced, especially on the role of the graph. We then discuss approaches widely used for graph-based affective representation in literature and show a trend towards graph construction. For the relational reasoning in graph-based FAA, existing studies are categorized according to their non-deep or deep learning methods, emphasizing the latest graph neural networks. Performance comparisons of the state-of-the-art graph-based FAA methods are also summarized. Finally, we discuss the challenges and potential directions. As far as we know, this is the first survey of graph-based FAA methods. Our findings can serve as a reference for future research in this field.

Index Terms—Facial Expression Recognition, Micro-expression Recognition, Action Unit Detection, Graph Representation, Graph Relational Reasoning, Graph Neural Network.

1 Introduction

Facial affects relate to 55% of messages when people perceive others' feelings and attitudes [1, 2] because it conveys critical information that reflects emotional states and reactions in human communications [3, 4, 5]. Many facial affect analysis (FAA) methods have been explored during the past decade, benefiting from interdisciplinary studies of affective computing, computer vision, and psychology [6, 7, 8]. Some have been extended to many applications, including medical diagnosis [9], social media [10], e-education [11], and video generation [12]. Meanwhile, competitions such as FERA [13], EmotiW [14], Aff-Wild [15], ABAW [16], EmotioNet [17], AVEC [18], and MuSe [19] are regularly held to evaluate the latest progress and propose frontier research trends.

Historically, FAA methods have undergone a series of evolutions. Initial studies usually rely on hand-crafted design or classic machine learning to obtain useful affective features without structural information [7, 20]. Psychological findings indicate that the human cognition of facial information is realized through a dual system composed of analytic processing and holistic processing [4]. The former acquires multi-dimensional cluster features by analyzing local areas, while the latter aims to generate a holistic representation to perceive the overall structure [4, 21]. Such an analytic-holistic working system is similar to a topology-like structure, so it is reasonable for machine vision researchers to model it into a graph. Accordingly, many state-of-the-art studies have been dedicated to generating a facial graph with local-to-global affective features [22, 23, 24, 25].

If the above evidence reveals the feasibility of using graph-based methods for FAA, research on how facial muscles participate in affective expression further demonstrates its possibility as a necessary condition [26, 27, 28]. There are latent relationships among different facial areas and contexts, which are vital clues [29, 30]. A few non-graph-based deep models have partly captured these relationships and improved performance [31, 32, 33]. The underlying assumption is that explicit mappings that reflect this relationship can be directly learned [34]. However, these mappings are not solid enough in the real world because they differ from subject to subject and even from one condition to another [35, 36]. Recently, graph-based methods have shown that they represent facial anatomy and simultaneously fit latent relationships in facial affects [37, 38, 39]. Some pilot studies have also suggested that the graph-based method can even move beyond to deal with challenging tasks such as analyzing occluded faces [40, 41] and ambiguous facial affects [42, 43].

By searching on Google Scholar using keywords of **'graph'** and **Index Terms** in this survey, we have counted the number of relevant published papers from 2010 to the present. As presented in Fig. 1, the graph-based FAA has gained increasing attention, especially in the past five years (publications in 2021 increased by 600 year-on-year).

Based on theoretical support, outstanding performance and quantity of existing work, and potential for future development, it is necessary to review the state of graph-based FAA methods. Although many reviews have discussed FFA's historical evolutions [7, 34, 44] and recent advances [45, 46, 47], including some specific problems like occluded expression [48], multi-modal affect [49] and micro-expression [50], **this is the FIRST systematic and in-depth survey for the graph-based FAA field** as far as we know. We emphasize representative research proposed after 2010. The goal is to present a novel perspective on FAA and its latest trends.

- * *Corresponding author*
- *Y. Liu, X. Zhang and J. Zhou are with the School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China. E-mail: liuy17@163.com, cszxm@scut.edu.cn, charlesmzhouscut@gmail.com.*
- *X. Li is with the Department of Electrical and Computer Engineering, Rutgers University, Piscataway, NJ-08854, United States. E-mail: xin.li.ece@rutgers.edu*
- *Y. Liu, Y. Li and G. Zhao are with the Center for Machine Vision and Signal Analysis, University of Oulu, Oulu, FI-90014, Finland. E-mail: {firstname.lastname}@oulu.fi*

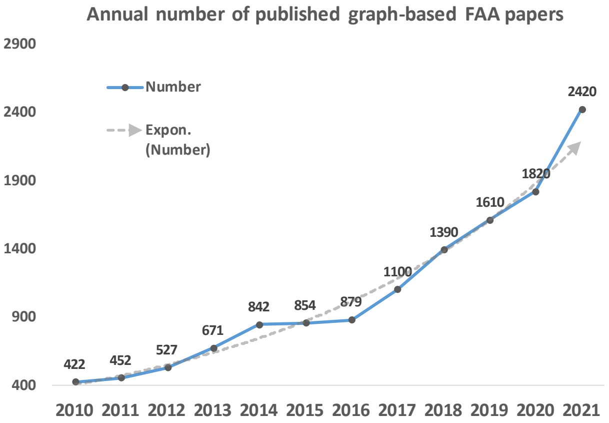


Fig. 1. The growth trend of papers related to graph-based FAA.

This review is organized as follows: Section 2 provides a brief background on FAA and discusses the unique role of graph-based methods in FAA research. Section 3 presents a taxonomy of mainstream graph-based methods for affective representation. Section 4 reviews classical and advanced approaches for graph relational reasoning and discusses their pros and cons in FAA tasks. Section 5 summarizes public databases, main FAA applications, and current challenges based on a detailed comparison of related literature. Finally, Section 6 identifies potential research directions.

2 Facial Affect Analysis

2.1 Affective Desription Model

As early as the 1970s, Ekman and Friesen [51] proposed the definition of six basic affects, i.e., *happiness*, *sadness*, *fear*, *anger*, *disgust*, and *surprise*, based on an assumption of the universality of human affective display [26]. In addition, compound affects [52] defined by different combinations of basic affects (e.g., *sadly surprise* and *happily surprise*) are proposed to depict more complex affective situations [53]. Another kind of famous description, called Facial Action Coding System (FACS), is designed for a broader range of affects, which consists of a set of atomic Action Units (AUs) [21, 30]. Fig. 2 shows an example of six basic affects plus *neutral* and activated AUs in each facial affect. Besides categorical models, a continuous affective model named VAD Emotional State Model [54] is also suggested [55, 56, 57]. It has three dimensions, i.e., *valence* (how positive or negative an affect is), *arousal* (the activation intensity of an affect), and *dominance* (how submissive or in-control a person is in an affective display). The continuous model describes human affects quantitatively and is appropriate to depict dynamic changes [45, 58]. Please refer to [7, 44] for a more detailed discussion about this topic.

2.2 General Pipeline

A standard FAA method can be broken down into fundamental components: face preprocessing, affective representation, and task analysis. As a new branch of FAA, the graph-based method also follows this generic pipeline (see Fig. 3). Face detection and registration are two necessary pre-steps that first locate faces and normalize facial variations, sometimes also providing facial landmarks [60, 61]. Fig. 4 presents an illustration of the preprocessing steps. Early methods like *Viola and Jones* [62], *Mixtures of Trees* [63], and *Active Appearance Model (AAM)* [64] have been widely used for this purpose. Recently, cascaded deep approaches with real-time performance are popular, such as *Multi-Task Cascaded Convolutional Network* [65], *Hyperface* [66], and *Supervision by Registration and Triangulation* [67]. Please refer to [68, 69, 70] for more specific information.

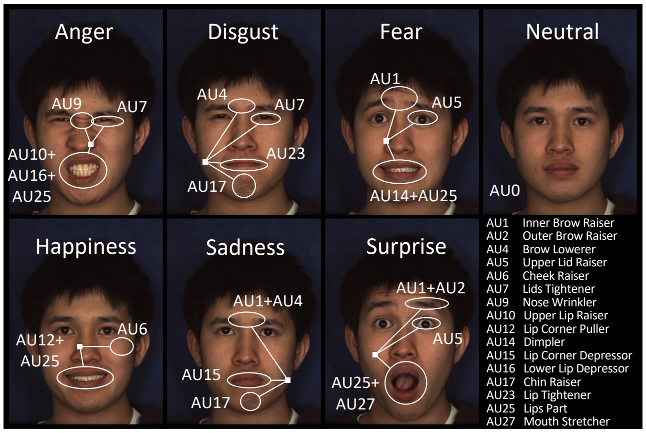


Fig. 2. An example of basic facial affects and related AU marks. AU0 denotes no activated AU. All annotations are made by FACS certificated experts. Images are from BU-4DFE [59] database.

Compared to other existing methods, the graph-based FAA pays more attention to representing facial affects with graphs and obtaining affective features from such representation by graph reasoning. The two components can perform separately or arranged as an end-to-end framework. They are expected to exhibit better performance and generalization capability by manually or automatically providing richer information through prior knowledge.

In mathematical terms, a graph can be denoted as $G = (V, E)$. The node set V contains all the representations of the entities in the graph, and the edge set E contains all the structure information between two entities. Thus, when E is empty, G becomes an unstructured collection of entities (e.g., independent local facial areas [31]). Meanwhile, we could also define some initial graph structure, usually denoted in the form of an adjacency matrix A, ahead of the relational model, which is a general practice in many affective graph representations [25, 40, 72].

Given this unstructured collection, performing relational reasoning requires the model to infer the structure of these entities before predicting the property or category of an object. Naturally, generic approaches need to be adjusted depending on affective graph representations or propose new graph-based approaches to infer the latent relationship and extract the final affective feature.

In the rest of this survey, different graph generation methods and their relational reasoning approaches will be systematically elaborated in Sec. 3 and Sec. 4, respectively, while the relevant preprocessing parts will be involved in both Sec. 3 and Sec. 5.1. All the analysis tasks will be discussed in Sec. 5.

3 Graph-based Affective Representations

Affective representation is a crucial procedure for most graph-based FAA methods. Depending on the domain that an affective graph models, we categorize the strategy as Spatial graphs, Spatio-temporal graphs, AU-level graphs, and Sample-level graphs. Fig. 5 illustrates a detailed summary of the literature using different graph representations.

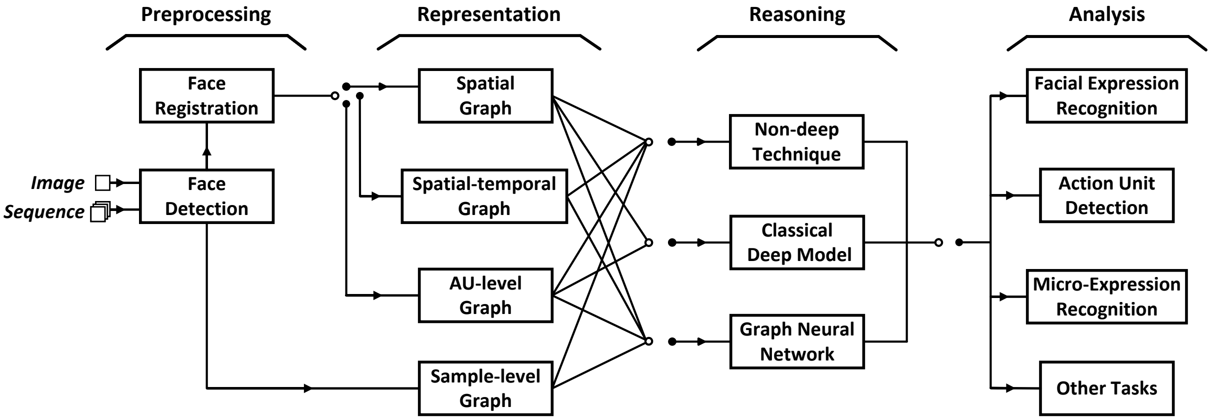


Fig. 3. The pipeline of graph-based facial affect analysis methods.

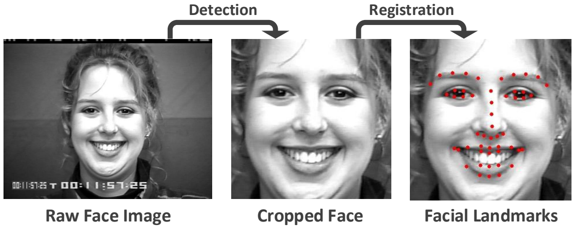


Fig. 4. An illustration of face detection and face registration. The *happy* image is from CK+ database [71].

Note that many graph-based representations contain pre-extracted geometric or/and appearance features. Whether hand-crafted or learned, these feature descriptors are not essentially different from those used in non-graph-based affective representations. Interested readers can refer to [7, 44, 46] for a systematic understanding of this topic.

3.1 Spatial Graph Representations

Non-graph-based spatial methods usually treat a facial affect as a whole representation or pay attention to variations among main face components or crucial facial parts [31, 107, 108]. For spatial affective graphs, facial changes are considered while their co-occurring relationships and affective semantics are represented as essential cues [22, 75, 83]. These approaches can be divided into landmark-level graphs and region-level graphs. Fig. 6 illustrates frameworks of different spatial graph representations.

3.1.1 Landmark-level graphs

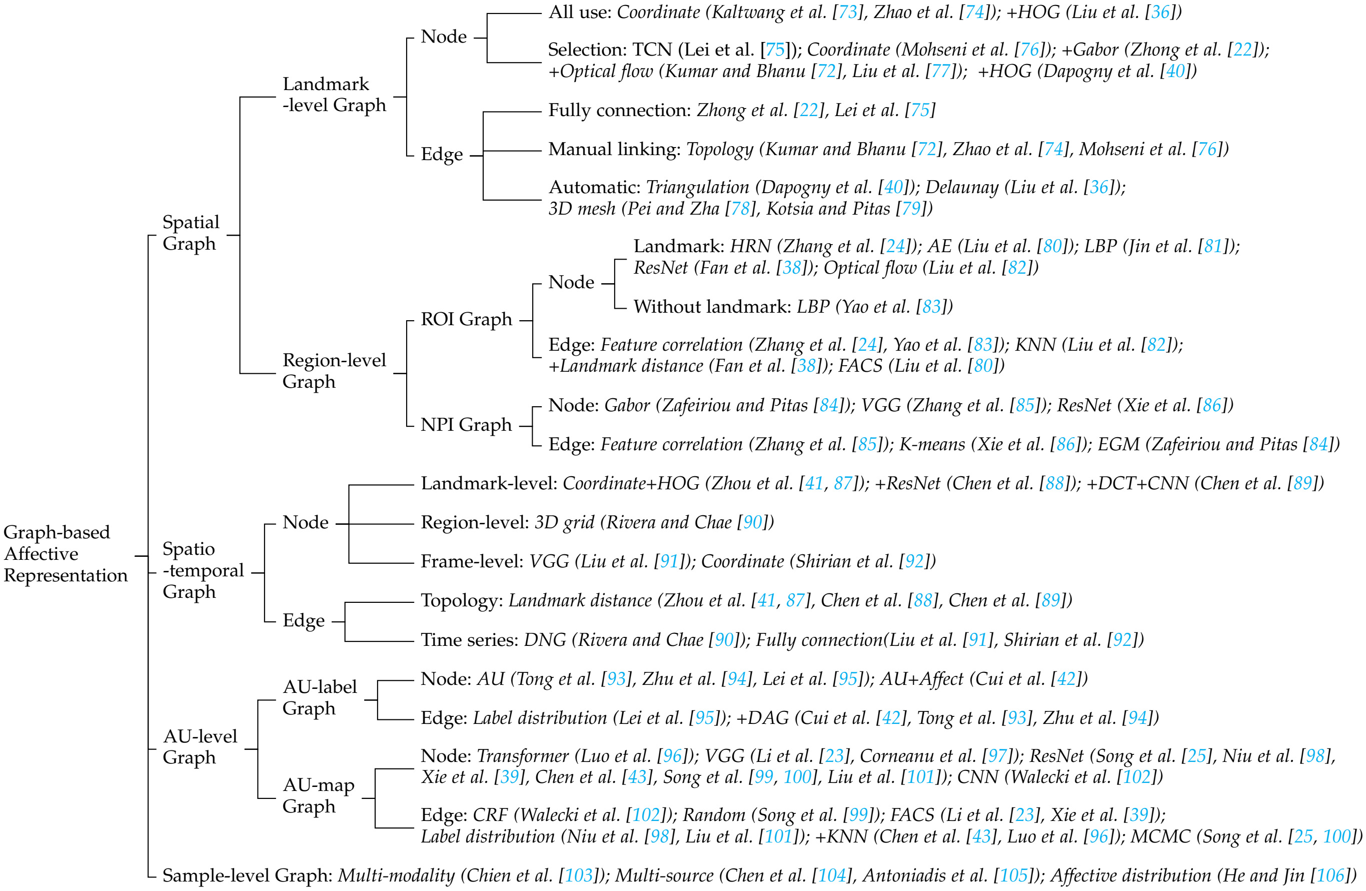
Facial landmarks are one of the most critical geometries that reflects the shape of face components and the structure of facial anatomy [109]. Thus, it is natural to use facial landmarks as base nodes to generate a graph representation.

Limited by the detection performance, only a few landmarks that locate basic face components were applied in early graph representations [110]. Recently, graphs using more facial landmarks (e.g., 68 landmarks [111]) are proposed to depict fine-grained facial shapes. For example, in [36] and [74], the authors associated 68 landmarks with the AUs in FACS and made graph-based representations. The difference is that the former additionally employed local appearance features extracted by *Histograms of Oriented Gradients (HOG)* [112] as node attributes; the latter proposed three landmark knowledge encoding strategies for enhanced geometric representations. Alternatively, [73] formulated a *Latent Tree (LT)* where 66 landmarks were set as parts of leaf nodes accompanied by several other leaf nodes of AU targets and hidden variables, which reflected the joint distribution of targets and features.

Furthermore, some current methods select landmarks with significant contributions to avoid redundant information [75, 113]. Landmarks locating external contour and nose are frequently discarded [40, 76] (see Figs. 6a, b) because they are considered irrelevant to facial affects. [22] chose to remove the landmarks of the facial outline and applied a small window around each remaining landmark as one graph node, while the local features were extracted by *Gabor* filter [114]. Since these local areas were segmented to introduce facial appearance into the graph representation rather than as independent nodes, similar to [36], these methods are still classified in landmark-level graphs. On the other hand, adding extra reasonable landmarks was designed to generate comprehensive graph representations [72, 115], which could keep an appropriate dimension and represent sufficient affective information.

A fully connected graph is the most intuitive way to form edges [22, 75]. However, the number of edges is $n(n-1)/2$ for a complete graph with n nodes, which means the complexity of the spatial relationship will increase as the number of nodes increases. This positive correlation is not helpful because landmarks in a facial component mostly move in concert rather than arbitrarily when conveying facial affects [116]. Studies of point-light displays in emotion perception also show that more complex representations seem to be redundant [117]. To this end, work like [72, 74, 76, 115] manually reduced edges based on muscle anatomy and FACS. Another type of approach is exploiting triangulation algorithms [40], such as the *Delaunay* triangulation [36], to generate graph edges consistent with true facial muscle distribution and uniform for different subjects. Similarly, the landmark-level graph with triangulation is also utilized in generating a sparse or dense facial mesh for 3D FAA [78, 79]. The Euclidean distance is the simplest and most dominant metric for edge attributes of the above facial graphs, even with multiple normalization methods like inner-eyes distance [36, 74]. The *Hop* distance has also been explored as edge attributes to model spatial relationships [41, 87]. Besides, several learning-based edge generation methods (e.g., *LT* [73], *Conditional Random Field (CRF)* [102]) have been proposed to extract semantic information from facial graphs automatically. This part is discussed in detail in Sec. 4.1 and 4.4.

Fig. 5. Taxonomy for Graph-based Facial Representation.


Graph-based Affective Representation
Spatial Graph
Landmark -level Graph
Node
All use: *Coordinate (Kaltwang et al. [73], Zhao et al. [74]); +HOG (Liu et al. [36])*
Selection: TCN (Lei et al. [75]); *Coordinate (Mohseni et al. [76]); +Gabor (Zhong et al. [22]); +Optical flow (Kumar and Bhanu [72], Liu et al. [77]); +HOG (Dapogny et al. [40])*
Edge
Fully connection: *Zhong et al. [22], Lei et al. [75]*
Manual linking: *Topology (Kumar and Bhanu [72], Zhao et al. [74], Mohseni et al. [76])*
Automatic: *Triangulation (Dapogny et al. [40]); Delaunay (Liu et al. [36]); 3D mesh (Pei and Zha [78], Kotsia and Pitas [79])*
Region-level Graph
ROI Graph
Node
Landmark: *HRN (Zhang et al. [24]); AE (Liu et al. [80]); LBP (Jin et al. [81]); ResNet (Fan et al. [38]); Optical flow (Liu et al. [82])*
Without landmark: *LBP (Yao et al. [83])*
Edge: *Feature correlation (Zhang et al. [24], Yao et al. [83]); KNN (Liu et al. [82]); +Landmark distance (Fan et al. [38]); FACS (Liu et al. [80])*
NPI Graph
Node: *Gabor (Zafeiriou and Pitas [84]); VGG (Zhang et al. [85]); ResNet (Xie et al. [86])*
Edge: *Feature correlation (Zhang et al. [85]); K-means (Xie et al. [86]); EGM (Zafeiriou and Pitas [84])*
Spatio -temporal Graph
Node
Landmark-level: *Coordinate+HOG (Zhou et al. [41, 87]); +ResNet (Chen et al. [88]); +DCT+CNN (Chen et al. [89])*
Region-level: *3D grid (Rivera and Chae [90])*
Frame-level: *VGG (Liu et al. [91]); Coordinate (Shirian et al. [92])*
Edge
Topology: *Landmark distance (Zhou et al. [41, 87], Chen et al. [88], Chen et al. [89])*
Time series: *DNG (Rivera and Chae [90]); Fully connection(Liu et al. [91], Shirian et al. [92])*
AU-level Graph
AU-label Graph
Node: *AU (Tong et al. [93], Zhu et al. [94], Lei et al. [95]); AU+Affect (Cui et al. [42])*
Edge: *Label distribution (Lei et al. [95]); +DAG (Cui et al. [42], Tong et al. [93], Zhu et al. [94])*
AU-map Graph
Node: *Transformer (Luo et al. [96]); VGG (Li et al. [23], Corneanu et al. [97]); ResNet (Song et al. [25], Niu et al. [98], Xie et al. [39], Chen et al. [43], Song et al. [99, 100], Liu et al. [101]); CNN (Walecki et al. [102])*
Edge: *CRF (Walecki et al. [102]); Random (Song et al. [99]); FACS (Li et al. [23], Xie et al. [39]); Label distribution (Niu et al. [98], Liu et al. [101]); +KNN (Chen et al. [43], Luo et al. [96]); MCMC (Song et al. [25, 100])*
Sample-level Graph: *Multi-modality (Chien et al. [103]); Multi-source (Chen et al. [104], Antoniadis et al. [105]); Affective distribution (He and Jin [106])*

3.1.2 Region-level graphs

Like geometric information, appearance information, especially in local facial regions, can also contribute to FAA [118, 119]. Using graph structures is an excellent choice to encode spatial relationships while representing texture changes in facial components [83, 84]. There are two categories of region-level affective graphs: region of interest (ROI) graphs and non-prior information (NPI) graphs.

ROI graphs partition a set of facial areas as graph nodes related to affective display. Coordinates of facial landmarks are commonly applied to locate and segment ROIs. Unlike a few landmark-level graphs that only use texture near all landmarks as supplementary information, ROI graphs explicitly select meaningful areas as graph nodes, and edges do not entirely depend on established landmark relationships. [24] employed a *High-Resolution Network (HRN)* [120] to regress ROI maps spotted by representative landmarks. Each spatial location in the extracted feature map was considered one graph node, while edges were induced among node pairs according to mappings between ROIs and AUs. Another example in [38] utilized feature maps of landmark-based ROIs outputted by the *ResNet50* [121] as nodes to construct a *K-Nearest-Neighbor (KNN)* graph. For each node, its pair-wise semantic similarities were calculated, and the nodes with the closest *Euclidean* distance were connected as initial edges. Similarly, [82] also employed landmark-based ROIs, but the *KNN* graph was generated in *optical-flow* space to encode the local manifold structure for a sparse representation [77]. Due to chained reactions among multiple AUs and the symmetrical structure of the human face, [81] proposed a parts-based graph that had manually linked edges by taking FACS and landmarks as references. The nodes were ROIs with *Local Binary Pattern (LBP)* [20] or deep features as attributes. In addition, the method of obtaining ROIs without relying on facial landmarks has also been studied [83] (see Fig. 6c).

NPI graphs usually define a graph structure without introducing external prior knowledge such as landmarks and FACS. Nodes in NPI graphs are generally facial regions uniformly distributed over the raw image or generated in a fully automatic manner. For example, [84] created a reference bunch graph by evenly overlaying a rectangular graph on object images (see Fig. 6d). Alternatively, [86] built two NPI graphs for cross-domain FAA, where local-to-global, global-to-local, and local-to-local edges were computed according to statistical feature distribution acquired by the *K-means* algorithm. Recent studies have also tried to introduce regions beyond facial parts or single-face images as context nodes. [85] exploited the *Region Proposal Network (RPN)* [122] with *VGG16* [123] to extract regions-level nodes, including the target face and its contexts, while edges were affective relationships calculated based on feature vectors.

3.2 Spatio-Temporal Graph Representations

Spatio-temporal representations deal with a sequence of frames within a temporal window and describe the dynamic evolution of facial variations [124]. In particular, introducing temporal information allows nodes to interact with each other at different times and generates a more complex affective graph. Fig. 7 presents frameworks of various spatio-temporal graph representations.

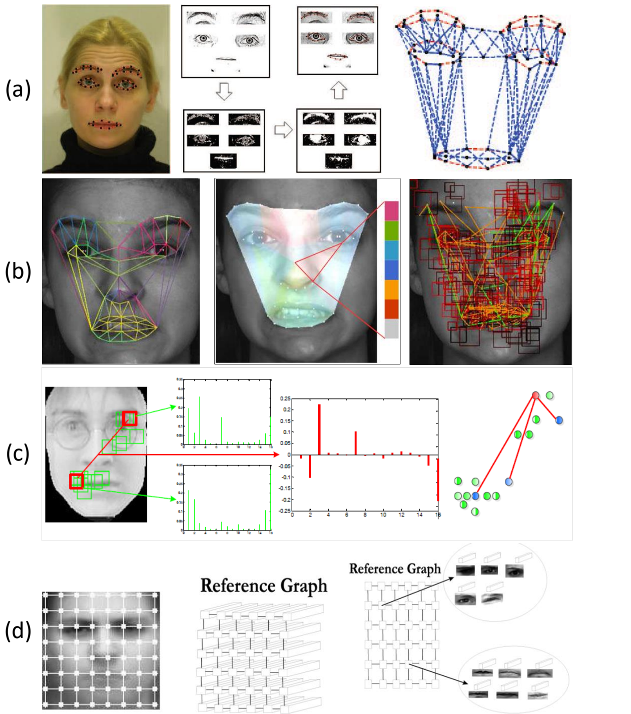


Fig. 6. Spatial graphs. (a) Landmark-level graph with FACS-based edges [76]; (b) Landmark-level graph with automatic triangle edges [40]; (c) ROI graph based on correlation without landmarks [83]; (d) NPI graph with individual nodes [84]. Zoom in for better view.

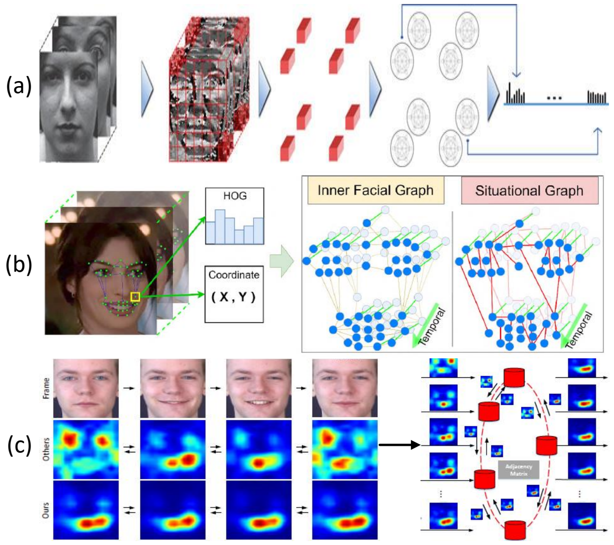


Fig. 7. Spatio-temporal graphs. (a) Region-level graph with edges based on transitional masks [90]; (b) Landmark-level graph with adaptive edges [41]; (c) Frame-level graph [91]. Zoom in for better view.

Extend spatial graphs to the spatio-temporal domain is currently the main route. [90] exploited weighted compass masks to obtain 2D directional number responses and 3D space-time directional edge responses corresponding to each of the symmetry planes of a cube. The two masks of given local neighborhoods were nodes in a spatio-temporal *Directional Number Transitional Graph (DNG)*, which could represent salient facial changes and statistic frequency of affective behaviors over time (see Fig. 7a).

Several representations have been proposed to define temporal connections between landmarks, which can be seen as landmark-level spatio-temporal graphs. [89] developed a context-aware facial multi-graph where intra-face edges were initialized based on morphological and muscular relationships, and inter-frame edges were created by linking the same node between consecutive frames. Similar landmark-based edge initialization in the temporal domain was also utilized in [87, 88]. In [41], the authors introduced a connectivity inference block that could automatically generate dynamic edges for a spatio-temporal situational graph of part-occluded affective faces (see Fig. 7b).

Unlike landmark-level graphs, [91] first extracted a holistic feature of each frame and set them as individual nodes to establish a fully connected graph (see Fig. 7c), which could be seen as a frame-level spatio-temporal graph. Similar work includes [92] that took *Discrete Cosine Transform (DCT)* features as node attributes. Edge connections of these methods would be established by learning the long-term dependency of nodes in time series (discussed in Sec. 4).

3.3 AU-level Graph Representations

Apart from using knowledge of AUs and FACS in the above two types of affective graphs, many graph-based representations have been proposed to model affective information from the perspective of AUs themselves. We divide these approaches into two categories: AU-label graph and AU-map graph. Fig. 8 shows frameworks of different AU-level graph representations.

3.3.1 AU-label graphs

Unlike spatial and spatio-temporal graphs, AU-label graphs were built from the label distribution of training data [37, 94]. [93] computed the co-occurrence and co-absence dependency between every AU pair from the existing database (see Fig. 8a). Since the dependency is not always symmetric, these AU label relationships were used as edges to construct a *Directed Acyclic Graph (DAG)*. In [95], an AU-label graph was built with a data-driven asymmetrical adjacency matrix that denoted the conditional probability of co-occurring AU pairs. AU labels were transformed into high-dimensional node vectors as node attributes [125]. On the other hand, [42] established a *DAG* where object-level labels (affect categories) and property-level labels (AUs) were regarded as parent nodes and child nodes, respectively. The conditional probability distribution of each node to its parents was measured to obtain graph edges for correcting existing labels and generating unknown labels. A similar idea was achieved in [43] to boost affective feature learning in large-scale FAA databases (see Fig. 8b).

3.3.2 AU-map graphs

AU-map graphs are intuitively close to region-level spatial graphs, especially ROI graphs, because they both employ local feature maps as graph nodes. [23] is an example in between. Twelve AUs features were learned through landmark-based ROI features cropped from a multi-scale global appearance feature [123]. These AU features and the AU relationships gathered from training data and manually pre-defined edge connections [126] were combined to construct a knowledge graph (see Fig. 8c). However, the significant difference is that AUs define a set of facial muscle actions, which means there might be multiple AUs in the same ROI. Like in Fig. 2, AU12 and AU15 co-occur at lip corners but refer to *'puller'* and *'depressor'*, respectively. Therefore, for many AU graphs, their definition of nodes is independent of those in ROI graphs, even though they are similar in feature map extraction. For instance, graph nodes in [39] were AU features directly obtained by *ResNet* without defining ROIs. The homologous protocol was also conducted in [100].

Some special AU-map graphs have been proposed to introduce structure learning for more complex FAA tasks. For AU intensity estimation, [102] trained a *Convolutional Neural Network (CNN)* to learn deep AU features from multiple databases jointly. The *copula functions* [127] were applied to model pair-wise AU dependencies in a *CRF* graph. In addition, *Bayesian networks (BNs)* are also used to capture the AU inherent dependencies for this task [25, 100]. To account for indistinguishable affective faces, [97] designed a *VGG*-like patch prediction module plus a fusion module to predict the probability of each AU. A prior knowledge taken from the given databases and a mutual gating strategy were used simultaneously to generate initial edge connections. To model uncertainty samples in real-world databases, [99] established an uncertain graph, in which a weighted probabilistic mask that followed *Gaussian* distribution was imposed on each AU feature map. By doing this, the importance of edges and the underlying uncertain information could be encoded in the graph representation. Another attempt in [98] boosted semi-supervised AU recognition for labeled and unlabeled face images. The parameters of two AU classifiers were used as graph nodes to share the latent relationships among AUs.

3.4 Sample-level Graph Representations

Recently, several graph representations beyond a single sample have been proposed, which indicates that this is still an open research field. In [106], a correlation graph with word-embedded affective labels as nodes was built for distribution learning. Its edges could be generated either by psychologically normalized *Gaussian* function or conditional probabilities. To combine signals from multiple corpora, [103] proposed a dual-branch framework, in which the visual semantic features were extracted in source and target sets. These features were then retrieved with correlation coefficients to generate positive edge connections for a learnable visual semantic graph (see Fig. 8d). Besides, [104] constructed a *KNN* graph with edges of binary weights to preserve the intrinsic geometrical structure of source and target data, which can seek more latent common information to reduce the distribution difference and make representations more discriminative.

3.5 Discussion

As a significant part of the graph-based FAA method, different affective graph representations have their merits, drawbacks, requirements, and time burdens. (see Table 1).

Spatial graph representations: Conceptually, landmark-level graphs model the facial shape variations of fiducial points and easily generate the internal structural relationships of different affective displays. However, most methods are sensitive to facial landmarks' detection errors, thereby failing in uncontrolled conditions. On the other hand, the selection of landmarks and the connection of edges have not yet formed a standard rule. Their effects on the graph representation are rarely reported, even though some FACS-based strategies have been designed. Region-level graphs explicitly regard an affective face as multiple local crucial facial areas compared with landmark-level graphs. The spatial relationships among selected regions are measured through feature similarity instead of manual initialization based on facial geometry. The circumstance resulting from inaccurate or unreasonable landmarks will also impact related ROI graphs. Since most NPI graphs utilize a region searching strategy, the problem is how to avoid the loss of target face and how to exclude invalid regions.

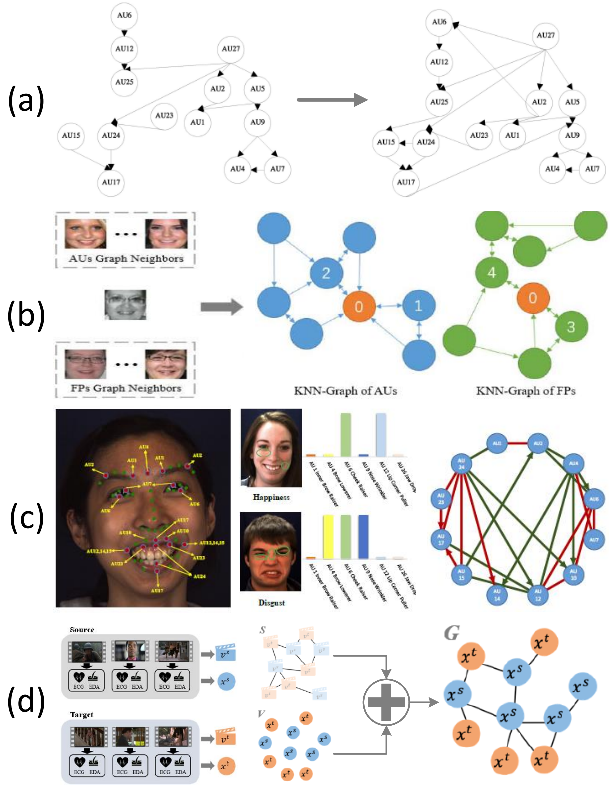


Fig. 8. AU-level and Sample-level graph representations. (a) AU-label graph with edges generated from training data [93]; (b) Auxiliary graphs of AUs and landmarks [43]; (c) AU-map graph with FACS based edges [23]; (d) Sample-level multi-modal graph of visual and physiological signals [103]. Zoom in for better view.

Spatio-temporal graph representations: With extra dynamic affective information, spatio-temporal graphs can help aggregate evolution features in continuous time. For landmark-level methods, the current initialization strategy of edges is to link the facial landmark with the same index frame by frame. Unfortunately, no research has been reported to learn the interaction of landmarks with different indexes in the temporal dimension. Besides, in addition to *Euclidean* distance and *Hop* distance, other edge attributes measurement methods should also be explored to model the semantic context both spatially and temporally. For the frame-level methods, embedding domain knowledge related to affective behaviours like the muscular activity by graph structure is not explicitly considered in recent work. Therefore, building a hybrid spatio-temporal graph is a practical way to simultaneously encode the two levels of affective information.

AU-level graph representations: As a distinctive type, AU-level graphs provide certain semantics of facial affects by representing each AU and its co-occurrence dependency. The measurement criteria of AU correlations are versatile

TABLE 1
An Overview of Affective Graph Representations

Category	Branch	Strength	Limitation	Demand	Time
Spatial	Landmark-level	Effective facial geometry embedding; Flexible structural relationships	Sensitive to the landmark detection accuracy	High-quality face registration	Low-Mid
	Region-level	Versatile local texture extraction; Underlying correlation beyond locations	Landmark sensitive for related ROI graphs; Redundant or missing regions for NPI graphs	Suitable for various situations	Mid-High
Spatio-Temporal		Extra dynamic evolution information	Relatively fixed edges in the time domain	Video/Sequence input	Low-High
AU-level	AU-label	Meaningful semantic dependencies; Explicit prior knowledge introduction	Cannot be an end-to-end framework	Reliable & sufficient AU annotations	Low
	AU-map		Unstable co-occurring distributions		Mid-High
Sample-level		Modular to existing architectures; Cross-corpus information	Lack of in-face modelling	Large-scale/Multiple databases	Low-High

but not general. Most AU-label graphs rely on the label distributions of one or multiple given databases. Nevertheless, AU labelling requires annotators with professional certificates and is a time-consuming task that causes existing databases with AU annotations to be usually small-scale. Therefore, the distribution from limited samples may not reflect the true dependencies of individual AUs, and its impact on FAA still needs to be assessed.

Sample-level graph representations: Sample-level graphs are an appealing field that introduces latent relationships in data distributions. Such characteristic makes it convenient to integrate with existing FAA methods. However, it also puts forward higher requirements for the diversity and balance of samples. On the other hand, to the best of our knowledge, there is no work combining sample-level and other in-face graphs to construct a joint representation, which we think is a good topic.

Computational complexity: The time complexity of building a graph representation depends on its node attributes and edge initialization. Landmark-level graphs require less computation, no matter for spatial or spatio-temporal graphs. Graphs involving pre-extracted features like region-level and AU-map graphs need additional computation, the degree of increase varies from manual to deep methods. AU-label graphs are the most efficient branch because they prefer a unified view of the distribution. Since sample-level graphs usually consider multiple data sources, extra burdens might be added according to specific situations.

4 Affective Graph Relational Reasoning

Generally, graph relational reasoning can be considered a two-step process, i.e., understanding the structure from a certain group of entities and making inferences of the system as a whole or the property within [128]. However, things are slightly different in the case of graph-based FAA. Depending on what kind of affective graph representation is exploited, the contribution of graph relational reasoning can be either merged before the decision level with other affective features or reflected in a collaborative way in the level of feature learning.

In this Section, we review relational reasoning methods designed for affective graph representations in four categories: *Dynamic BNs (DBNs)*, classical deep models, *Graph Neural Networks (GNNs)* and non-deep machine learning techniques.

4.1 Dynamic Bayesian Networks

DBNs are often used to reason about relationships among facial displays like AUs [129] and, of course, for AU-label graph representations. The *BN* is a *DAG* that reflects a joint probability distribution among a set of variables. In the work of [37, 93], a *DAG* was manually initialized according to prior knowledge, and then large databases were used to perform structure learning to find the optimal probability graph structure. After that, the probabilities of different AUs were inferred by learning the *DBN*. Following this idea, [94] additionally integrated *DBN* to a multi-task feature learning framework and made the AU inference by calculating the joint probability of each category node. Sometimes *DBN* is also combined with some statistical methods to explore different graph structures [25, 100], such as *Hidden Markov Models* [130]. Another advanced research of *DBN* is [42] that modeled the inherent relationships between category labels and property labels. Its parameters were utilized to denote the conditional probability distribution of each AU given the facial affect. The wrong labels could be corrected by leveraging the dependencies after the structure optimization.

4.2 Adjustments of Classical Deep Models

Before *GNNs* are widely employed, many studies have adopted conventional *Deep Neural Networks (DNNs)* to process affective representations with the graph structure. These deep models are not explicitly designed but can conduct standard operations on structural graph data by adjusting the internal architecture or applying an additional transformation to the input graph representation. Fig. 9 shows examples of classical deep models for graph relational reasoning.

4.2.1 Recurrent neural networks

The *Recurrent Neural Networks (RNNs)* variant is one of the successfully extended model types for handling graph structural inputs. Similar to random walk, [22] applied a *Bidirectional RNN* to deal with its landmark-level spatial graph representation in a rigid order. The Gabor features of each graph node were updated by multiplying with the average of the connected edges to incorporate the structural information. Subsequently, the nodes were iterated by the RNN with learnable parameters in forwarding and the backward direction (see Fig. 9a). In [97], the authors built a structure inference module to capture AU relationships from an AU-map graph representation. Based on a collection of interconnected recurrent structure inference units and a parameter sharing *RNN*, the mutual relationship between two nodes could be updated by replicating an iterative message passing mechanism with the control of a gating strategy. Following the sequential idea of RNNs,[23] exploited a *Gated Graph Neural Network (GGNN)* [131] that calculated

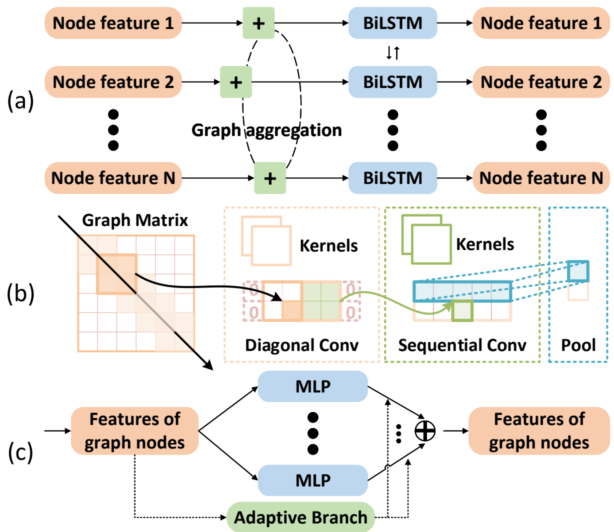


Fig. 9. Classical deep models for graph relational reasoning. (a) RNN [22]; (b) CNN [36]; (c) MLP [100]. Zoom in for better view.

the hidden state of the next time step by jointly considering the current hidden state of each node and adjacent nodes. The relational reasoning could be done through the iterative update of *GGNN* over its AU-map graph.

4.2.2 Convolutional neural networks

Unlike the sequential networks, [36] utilized a variant *CNN* to process the landmark-level spatial affective graph. Compared to standard convolution architectures, the convolution layer in this study convolved over the diagonal of a particular adjacency matrix to aggregate the information from multiple nodes. Then a list of the diagonal convolution outputs was further processed by three 1D sequential convolution layers. The corresponding pooling processes were performed behind convolution operations to integrate feature sets (see Fig. 9b). Another attempt for landmark-level spatial graph representations is the *Graph Temporal Convolutional Networks (Graph-TCN)* [75]. It followed the idea of *TCNs* that consisted of residual convolution, dilated causal convolution, and weight normalization [132]. By using different dilation factors, *TCNs* were applied to convolve the elements inside one node sequence and from multiple node sequences. Thus, the *TCN* for a node and *TCN* for an edge could be trained respectively to extract node features and edge features simultaneously. Besides, [38] exploited a Semantic Correspondence Convolution module to model the correlation among its region-level spatial graph. Based on the assumption that the channels of co-occurring AUs might be activated simultaneously, the *Dynamic Graph CNN (DG-CNN)* [133] was applied on the edges of the constructed *KNN* graph to connect feature maps sharing similar visual patterns. After the aggregation function, affective features were obtained to estimate AU intensities.

4.2.3 Multilayer perceptron networks

As a vanilla architecture, *Multilayer Perceptrons (MLPs)* have also been explored. [40] employed a hierarchical *Auto-Encoder (AE)* based on *MLPs* to capture relationships from a landmark-level spatial graph. It accumulated *HOG* features of multiple nodes whose appearance changes were closely related and computed the confidence scores as the triangle-wise weights over edges. After that, a *Random Forest (RF)* was used for facial affect classification and AU detection simultaneously. In [100], a hybrid graph network composed of dynamic *MLPs* performed multiple types of message passing, which provided more complementary information for reasoning the positive and negative dependencies among AU nodes. An adaptive branch was used to generate weights for different graph structures (see Fig. 9c).

4.3 Graph Neural Networks

GNNs are proposed to extend the 'depth' from 2D image to graph structure and establish an end-to-end learning framework instead of additional architecture adjustment or data transformation [134]. Unlike conventional deep models that are stacked by layers of $H^{l+1} = f(H^l)$, where H^l denotes the hidden state of the l-th layer, f is the nonlinear function, *GNNs* formulate layers as $H^{l+1} = f(H^l, A)$, where the adjacency matrices A (records all the edges in a graph) is explicitly fed for the relational reasoning. Therefore, *GNNs* can infer comprehensive graph structures through A for better feature extraction, including effective node updates and flexible edge connections. Fig. 10 illustrates several *GNN* architectures for affective graph relational reasoning.

4.3.1 Graph convolutional networks

Graph Convolutional Networks (GCNs), especially the spatial *GCN* [135], are the most popular *GNN* in graph-based FAA research. Practically, *GCNs* can be set as an auxiliary module [74, 95, 105] or part of the collaborative feature learning framework [25, 96, 101].

For the auxiliary module, *GCNs* are applied immediately after the graph representation. However, the outputs of relational reasoning are not directly used for facial affect classification or AU detection but are later combined with other deep features as a weighting factor (see Fig. 10a). [98] employed a two-layer *GCN* for message passing among different nodes in its AU-level graph. Both the dependency of positive and negative samples were considered and used to infer a link condition between any two nodes. The output of *GCN* was formulated as a weight matrix of the pre-trained AU classifiers. Besides, *GCNs* can also be utilized following the above manner to execute relational reasoning on atypical graph representations, such as multi-target graph [85], distribution graph [106], and cross-domain graph [86].

For the collaborative framework, *GCNs* usually inherit the previous node feature learning model progressively (see Fig. 10b). Like in [80], a *GCN*-based multi-label encoder was proposed to update features of each node over a region-level spatial graph representation. The reasoning process was the same as that in the auxiliary framework. Similar studies also include [24] and [103]. In addition, to incorporate the dynamic in spatio-temporal graphs, [91] set *GCNs* as an imitation of attention mechanism or weighting mechanism to share the most contributing features to explore the dependencies among frames. After training, the structure helped nodes update features based on messages from the peak frame and emphasize the concerned facial region. A more feasible way is to apply *Spatial Temporal GCN (STGCN)* [136] on spatio-temporal graphs [41, 87, 88, 89] (see Fig. 10c). In their relational reasoning, features of each node were generated with its neighbor nodes in the current frame and consecutive frames by using spatial graph convolution and temporal convolution, respectively.

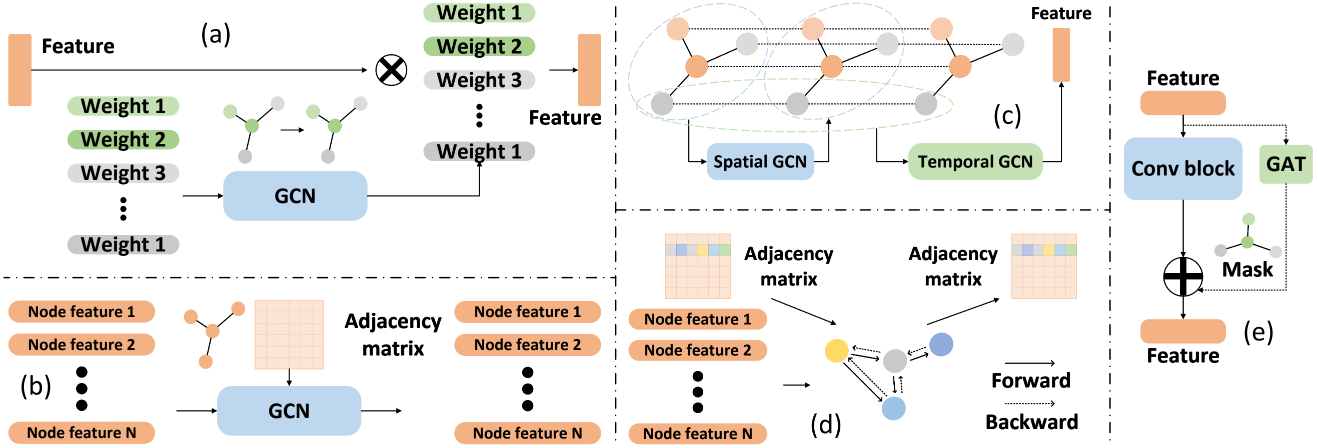


Fig. 10. GNNs for graph relational reasoning. (a) GCN as an auxiliary module [98]; (b) GCN as a collaborative framework [25]; (C) STGCN [89] (d) Spectral GCN [81]; (e) GAT [89]. Zoom in for better view.

Alternatively, the approach of spectral *GCN* [137] has also been studied [113]. [81] devised a lightweight GCN following the *Message Passing Neural Network* [138]. A learnable adjacency matrix was adapted to infer the spatial dependencies of ROI nodes in different facial affects. [92] extended in *Inception* idea from standard *CNNs* to spectral *GCNs* that captured emotion dynamics at multiple temporal scales. The yielded embeddings of different dimensions were jointly learned over a classification loss and a graph learning loss for the optimal graph structure.

4.3.2 Graph attention networks

Graph Attention Networks (GATs) aim to strengthen the node connections with high contribution and offer a more flexible way to process the graph structure [139]. [99] introduced an uncertain *GNN* with *GAT* as the backbone. The goal is to select valuable edges, depress noisy edges, and learn AU dependencies on its AU-map graph. In addition, the underlying uncertainties were considered in a probabilistic way, close to the idea of *Bayesian* methods in *GNN* [140], to alleviate the data imbalance by weighting the loss function. On the other hand, *GAT* collaboratively worked with *GCN* in [72] to deal with two-stream graph inputs. Compared to applying *GAT* directly, [39] proposed a GNN that added a self-attention graph pooling layer after three sequential *GCN* layers. A similar block was done in [115] which revised the *GCN* block with channel and node attention. It improved the reasoning process on graph representations because only important nodes would be aggregated, including affective information and facial topology. To make nodes interact more dynamically instead of using a constant graph structure, [89] applied a set of learnable edge attention masks to the *STGCN* for subtle adjustments of the defined spatio-temporal graph representation (see Fig. 10e).

4.4 Non-deep Machine Learning Methods

Although refining deep features extracted by parameterized neural networks and gradient-based methods is the mainstream, they require numerous training samples for effective learning. Due to the insufficient data in the early years or the purpose of efficient computation, many non-deep machine learning techniques have been applied for affective graph relational reasoning. Graph structure learning is one of the widely used approaches. In [73], the reasoning of its spatial graph representation was conducted by *LT* learning. Parameters update and graph-edit of *LT* structure were performed iteratively to maximize the marginal *log*-likelihood of a set of training data. [102] employed *CRF* to infer AU dependencies in an AU-map graph. The use of *copula* functions allowed it to model non-linear dependencies among nodes easily. At the same time, an iterative balanced batch learning strategy was introduced to optimize the most representative graph structure by updating each set of parameters with batches. Approaches of graph feature selection are also exploited in this part, such as *Graph Sparse Coding (GSC)* [82, 104] and *Elastic Graph Matching (EGM)* [84]. These methods have provided a more diverse concept for graph relational reasoning.

4.5 Discussion

Although all the methods above can achieve affective graph relational reasoning, the choice has a causal relationship with the type of graph representation (see Table 2).

TABLE 2
Causal Relationships between Graphs and Reasoning methods

Category	Spatial	Spatio-Temporal	AU-level	Sample-level
DBNs			√	
DNNs	√	√	√	
GNNs	√	√	√	√
Non-deep	√	√	√	√

Dynamic Bayesian network: Nearly half of AU-label graph representations employ *DBNs* as their relational reasoning model. However, the representation quality highly relies on the available training data that need balanced label distribution in positive-negative samples and categories. This strong assumption will limit the effectiveness of node dependencies learned by *DBNs*. Another problem is that *DBNs* can only be combined with facial features as a relatively independent module and are hard to integrate into an end-to-end learning framework.

Classical deep model: Standard deep models, including *CNNs*, *RNNs*, and *MLPs*, have been explored to conduct graph relational reasoning before the emergence of *GNNs*. Even if they are suitable for more graph representations than *DBNs*, these grid models focus more on local features. The additional adjustments in input format or/and network architecture cause losses of node information or let node messages only pass and update in a specific sequence, which suppresses the global property represented by the graph. Thus, we think the specifically designed networks like *GNNs* will become dominant in this part.

Graph neural network: *GNNs* are developing techniques that make full advantage of graphs. Architectures with different focuses have been proposed but have their flaws as well. For instance, *GCNs* cannot handle directed edges well (e.g., AU-level graphs), while *GATs* only use the node links without considering edge attributes (e.g., spatial graphs). Besides, due to the low dimension of the nodes in affective graphs, too deep *GNNs* may be counterproductive. In addition, being an auxiliary block or part of the whole framework will influence the construction of *GNNs*. Therefore, managing graph representation and relational reasoning using *GNNs* still need to be explored.

Non-deep methods: Non-deep machine learning has a place in early studies and is even applied in recent work because no training is required. They partly inspire advanced techniques like *DBNs* and *GNNs*. Nevertheless, one of the reasons they have been replaced is that these approaches need to be designed separately to cope with different graph representations, similar to hand-crafted feature extraction. Hence, it is not easy to form a general framework. On the other hand, more training data and richer computing resources allow deep models to perform more effective and higher-level relational reasoning on affective graphs.

Computational complexity: Obviously, classical deep models like *CNNs*, *RNNs*, and *MLPs* are more time-consuming than the non-deep methods such as *GSC* and *EGM*. For *GNNs* including *GCNs* and *GATs*, existing methods typically add a limited number of layers due to relatively small graph dimensions and potential overfitting issues. Thus, their computational burden is not significant compared to the overall framework, as has been confirmed by experiments in a few related studies [81, 100].

5 Applications and Performance

According to different description models of facial affects, the FAA can be subdivided into multiple applications. The typical output of FAA systems is the label of a basic facial affect or AUs. Recent research also extends the goal to predict micro-expression or affective intensity labels or continuous affects. This section compares and discusses graph-based FAA methods from four main application categories: facial expression recognition, AU detection, micro-expression recognition, and a few special applications. Due to page limitation, we select the most relevant and representative papers following these standards: published in more well-known forums in the past five years; or belonging to distinct branches of graph representation and reasoning for diversity consideration.

5.1 Databases

Most FAA studies apply public databases of facial affect as validation material. A comprehensive overview is presented in Table 3. The characteristics of these databases are listed from four aspects: samples, attributes, graph-related properties, and certain contents. Fig. 11 exhibits several examples of facial affects under different conditions. In addition, for better interpreting the graph-based FAA, we summarize corresponding elements (e.g., landmark coordinates, AU labels) self-carried by databases, which are rarely considered in previous related surveys.

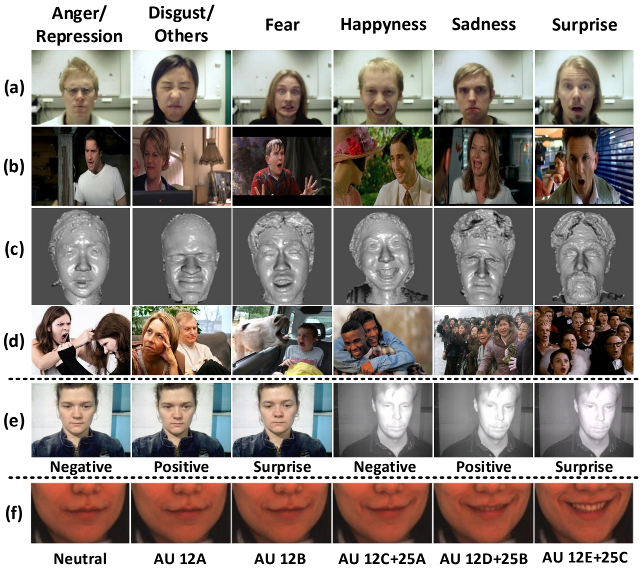


Fig. 11. Facial affect databases. (a) Oulu-CASIA contains posed facial affects; (b) SFEW 2.0 has facial affects under in-the-wild scenarios; (c) BP4D provides 3D affective face images; (d) EMOTIC, multiple faces appear per image with VAD annotations; (e) The SMIC collects images of spontaneous micro facial affects in visual light and near infrared light; (f) DISFA offers frame-level AU intensity labels.

Databases containing posed facial affects, including Extended Cohn-Kanade Dataset (CK+) [71, 141], M&M Initiative Facial Expression Database (MMI) [142], and Oulu-CASIA NIR&VIS Facial Expression Database (Oulu-CASIA) [143], are chosen by early FAA methods. More challenging databases, such as FER-2013 [144], Static Facial Expression in the Wild (SFEW) 2.0 [145], and Acted Facial Expression in the Wild (AFEW) 7.0 [146], tend to acquire spontaneous affective data from complex and wild environments.

Some databases also contain intensity labels of facial affects and even AUs, e.g., Denver Intensity of Spontaneous Facial Action Database (DISFA) [147], Binghamton University 3D/4D Facial Expression Database (BU-3DFE/4DFE) [148, 149] and Binghamton-Pittsburgh 3D Dynamic Spontaneous Facial Expression Database (BP4D) [150].

Another type of database is for micro-expressions. Participants are required to keep a neutral face while watching videos associated with induction of specific affects [151]. Following this setting, Spontaneous Micro Facial Expression Database (SMIC) [152], Improved Chinese Academy of Sciences Micro-Expression Database (CASME II) [153], Spontaneous Micro-Facial Movement Database (SAMM) [154], Chinese Academy of Sciences Macro-Expression and Micro-Expression Database (CAS(ME)^2) [155] have been released. However, it is hard to collect and annotate large-scale micro-expression data with uncontrolled scenarios due to its subtle, rapid, and involuntary nature.

Recently, large-scale databases have been developed to provide massive data with spontaneous facial affects and in-the-wild conditions, such as Real-World Affective Face Database (RAF-DB) [156], Large-Scale Face Expression in-the-Wild dataset (ExpW) [10], EMOTIC [157], AffectNet [158], and EmotioNet [159]. Besides, Aff-Wild2 [160] is the largest database with continuous and AU labels.

Concerning graph-based FAA, it is available to find and

TABLE 3
An Overview of Facial Affect Databases

Database 'year	Samples			Attributes			Graph-based Properties[3]			Special Contents[4]
	Data Type	Subjects	Number	Eli. & Sou.[1]	Affects[2]	B.B.[1]	LM	AU	Dynamic	
CK(+) '10	Sequences	97(123)	486(593)	P & Lab	6B+N(+C)	●	●	● + I	●	-
MMI '10	Images/Videos	75	740/2900	P & Lab	6B+N	○	○	●	● + D	Head pose
Oulu-CASIA '11	Sequences	80	2880	P & Lab	6B	●	○	○	●	NIR
DISFA '13	Sequences	27	130,000	S & Lab	6B+N	●	●	● + I	● + F	-
FER-2013 '13	Images	-	35887	S & Web	6B+N	●	○	○	○	Wild
SFEW 2.0 '15	Images	-	1766	S & Movie	6B+N	●	●	○	○	Wild
AFEW 7.0 '17	Videos	-	1809	S & Movie	6B+N	●	●	○	●	Audio, Wild
BU-3DFE '06	Images	100	2500	P & Lab	6B+N	●	●	○	○	3D, Multi-view, I
BU-4DFE '08	Videos	101	606	P & Lab	6B+N	●	●	● + I	●	3D, Multi-view
BP4D '14	Videos	41	328	S & Lab	6B+E+P	●	●	● + I	● + F	3D, Head pose
SMIC '13	Sequences	16	164	S & Lab	$3B^{\dagger}$+N	●	○	○	●	Micro., NIR
CASME II '14	Sequences	35	247	S & Lab	$3B^{\ddagger}$+R+O	●	○	●	● + D	Micro.
SAMM '18	Sequences	32	159	S & Lab	6B+C	●	○	●	● + D	Micro.
CAS(ME)^2 '18	Sequences	22	300+57	S & Lab	$3B^{\dagger}$+O	●	○	●	● + D	Macro. & Micro.
EmotioNet '16	Images	-	950,000	S & Web	6B+17Comp.	●	●	● + I	○	Wild
ExpW '18	Images	-	91,793	S & Web	6B+N	●	○	○	○	Multi-sub., Wild
RAF-DB '19	Images	-	29672	S & Web	6B+N+11Comp.	●	●	○	○	Wild
AffectNet '19	Images	-	420,299	S & Web	6B+N+C+O	●	●	○	○	V&A, Wild
EMOTIC '19	Images	-	23,571	S & Web	6B+N+19Comp.	●	○	○	○	V&A, Multi-sub., Wild
Aff-Wild2 '19	Images/Videos	-	564/2,800,000	S & Web	6B+N+O	●	○	●	●	Audio, V&A, Wild

[1] Eli.: elicitation; Sou.: source; P: posed; S: spontaneous; B.B.: bounding boxes; LM: landmarks; ● = Yes, ○ = No.
[2] 6B: six basic affects; N: neutral; C: contempt; E: embarrassment; P: pain; O: others; R: repression; $3B^{\dagger}$: three basic affects (positive, negative, surprise); $3B^{\ddagger}$: three basic affects (happiness, disgust, surprise); Comp.: compound affects.
[3] I: intensity annotation; D: onset-apex-offset annotation; F: frame-level annotation.
[4] NIR: near-infrared; Multi-sub.: multiple subjects per image; Micro.: micro-expression; Macro.: macro-expression; Wild: in-the-wild; V&A: valence and arousal.

select suitable databases with corresponding metadata, such as landmarks, AU labels, and dynamics, for different graph representation purposes. However, existing databases also have some shortcomings. On the one hand, not enough AU intensity annotations are provided by in-the-wild databases, decreasing AUs' role in FAA. On the other hand, there is few large-scale dynamic affective database so that limits temporal information in generating affective graph representations. Finally, databases about natural and spontaneous facial affects in a continuous domain need more attention instead of discrete categories.

5.2 Facial Expression Recognition

Facial expression recognition (FER), or macro-expression recognition, has been working on basic facial affects classification. An inevitable trend of FER is that the research focus has shifted from the early posed facial affects in controlled conditions to the recent spontaneous facial affects in real scenarios. In other words, existing methods, including graph-based FER, can achieve excellent FER under lab-controlled scenarios, which can be corroborated from the results in Table 4. For example, the performance on the CK+ database is very close to 100% [79, 90, 91, 115].

From the view of the representation, spatial graphs and spatio-temporal graphs are dominant. Specifically, handcrafted features (e.g., *LBP* [81, 83], *Gabor* [22, 115], *HOG* [36, 87, 104]) or deep-based features (e.g., *CNN* [74, 115], *VGG* [91], *ResNet* [39]) are employed to enhance the node representation similar to many non-graph FER methods [20, 31]. For reasoning approaches, early studies prefer to capture the relations of an individual node from predefined graph structures using tracking strategies (e.g., displacement projection [76] and *DNG* [90]) or general machine learning models (e.g., *RF* [40], *RNN* [22], *CNN* [36]). In the latest work, *GCNs* become one of the mainstream choices in the latest work and show state-of-the-art performances on posed and in-the-wild databases [39, 74, 81, 101, 105, 113, 115]. Another observation is that the framework of combining the spatio-temporal graph representation and *GNNs* is getting more attention in FER studies [41, 87, 91, 92].

Although many graph-based studies have shown improvements in predicting facial affects, FER still has some potential topics. One thing is that the goal of existing methods stays on classifying basic facial affects. No study of graph-based methods to recognize compound affects (or mixture affects), whose labels are provided by recent databases like RAF-DB and EmotioNet, is reported. One possible solution is introducing AU-level graph representations that can describe fine-grained macro-expressions with closer inter-class distances. The other topic is practical graph-based representations due to the big gap between the performance of current methods and the acceptable result in practice when analyzing in-the-wild facial affects. In addition, since existing databases lack sufficient dynamic annotated samples, the evaluation of spatio-temporal graphs in large-scale conditions remains to be explored.

5.3 Action Unit Detection

The AU detection (AUD) facilitates a comprehensive analysis of the facial affect and is typically formulated as a multi-task problem that learns a two-class classification model for each AU. It can expand the recognition categories of macro-expressions through the AU combination [73] and can be used as a pre-step to enhance the recognition of micro-expressions [95]. Compared with graph-based FER, the wide usage of graph structures has a long history in AUD [34] and has played a more dominant role. Table 5 summarizes graph-based AUD methods including the performance comparison.

Specifically, spatial graphs and AU-level graphs are equally popular in the representation part of AUD. Interestingly, no matter landmark-level or region-level, all the spatial graphs constructed in the listed AUD methods employed facial landmarks [24, 36, 38, 40, 73, 80], even

TABLE 4
Performance summary of representative graph-based FER methods

References	Prep.[1]		Representation[2]			Reasoning		Posed Database[3]	Wild Database	Validation[4]
	B.B.	LM	Category	Node	Edge	Model	Classifier			
Mohseni et al. [76]	○	50	$\mathcal{S}:\mathcal{L}$	$\mathbb{C}$	*M*+$\mathbb{E}$	*Tracking*	*Adaboost*	MMI *ar*: 0.877	-	10F CV
Rivera and Chae [90]	●	-	$\mathcal{ST}:\mathcal{R}$	*Histograms*	*DNG*	*Tracking*	*SVM*	CK+ *ar*: 1; MMI *ar*: 0.976; Oulu *ar*: 0.984	-	10F CV

Yao et al. [83]	●	-	$\mathcal{S}:\mathcal{R}$	*LBP*	*L*	-	*SVM*
							/*CNN*
								-	SFEW *ar*: 0.5538;
									AFEW *ar*: 0.5380;
										HO

Dapogny et al. [40]	●	49	$\mathcal{S}:\mathcal{L}$	$\mathbb{C}$+*HOG*	△	*AE*	*RF*	CK+ *ar*: 0.934 (w/o 0.915); BU-4DFE *ar*: 0.750 (w/o 0.730)	SFEW *ar*: 0.371 (w/o 0.357)	10F-SI CV /HO
Zhong et al. [22]	●	46	$\mathcal{S}:\mathcal{L}$	*Gabor*	*F*+$\mathbb{E}$	*RNN*	*Softmax*	CK+ *ar*: 0.9827; MMI *ar*: 0.9444; Oulu *ar*: 0.9368 (w/o 0.8263)	-	10F-SI CV

Chen et al. [43]	●	-	$\mathcal{AU}:\mathcal{B}$	Φ	*KNN*	-	*Softmax*	CK+ *ar*: 0.9308 (w/o 0.8899);
								MMI *ar*: 0.7049 (w/o 0.6779);
								Oulu *ar*: 0.6385 (w/o 0.5935)
									SFEW *ar*: 0.5650 (w/o 0.5219);
									RAF *ar*: 0.8553 (w/o 0.8181);
									AffNet: *ar*: 0.5935 (w/o 0.5797)
										CD
										(AffNet+RAF)

Liu et al. [36]	●	68	$\mathcal{S}:\mathcal{L}$	$\mathbb{C}$+*HOG*	△	*CNN*	*Softmax*	CK+ *ar*: 0.9767; MMI *ar*: 0.8011	SFEW *ar*: 0.5536	10F-SI CV /HO
Xie et al. [86]	●	-	$\mathcal{S}:\mathcal{R}$	*ResNet*	*K-means*	*GCN*	*Softmax*	CK+ *ar*: 0.8527; JAFEE [161] *ar*: 0.6150	SFEW *ar*: 0.5643; FER2013 *ar*: 0.5895 ExpW *ar*: 68.50%	CD (RAF)
Zhou et al. [87]	●	34	$\mathcal{ST}:\mathcal{L}$	$\mathbb{C}$+*HOG*	*M*+$\mathbb{H}$	*GCN*	*Softmax*	CK+ *ar*: 0.9863; Oulu *ar*: 0.8723	-	10F-SI CV
Liu et al. [91]	●	-	$\mathcal{ST}:\mathcal{F}$	*VGG*	*F*	*GCN*	*LSTM*	CK+ *ar*: 0.9954; MMI *ar*: 0.8589; Oulu *ar*: 0.9104	-	10F-SI CV
Zhou et al. [41]	●	44	$\mathcal{ST}:\mathcal{L}$	$\mathbb{C}$+*HOG*	*L*+$\mathbb{H}$	*GCN*	*Softmax*	CK+ *ar*: 0.9892 (w/o 0.8726); Oulu *ar*: 0.8750 (w/o 0.8405)	AFEW *ar*: 0.4512 (w/o 0.2721)	10F-SI CV /HO

Cui et al. [37]	○	-	$\mathcal{AU}:\mathcal{B}$	*CNN*	Φ	*DBN*	*Softmax*	CK+ *ar*: 0.9759 (w/o 0.9429);
								BP4D *ar*: 0.8382 (w/o 0.6168);
								MMI *ar*: 0.8490 (w/o 0.6735)
									EmoNet *ar*: 0.9555 (w/o 0.8085)	5F-SI CV

Liu et al. [115]	●	40	$\mathcal{S}:\mathcal{L}$	$\mathbb{C}$+*HOG*+*Gabor* /$\mathbb{C}$+CNN	*L*+$\mathbb{H}$ /+$\mathbb{E}$	*GCN*	*Softmax*	CK+ *ar*: 0.9923; MMI *ar*: 0.8575; Oulu *ar*: 0.9088	SFEW *ar*: 0.5742 RAF *ar*: 0.8713	10F-SI CV /HO

Chen et al. [104]	○	-	$\mathcal{X}$	*HOG*	*KNN*	*GSC*	*SVM*	CK+ (JAFEE) *ar*: 0.7171;
								JAFEE (CK+) *ar*: 0.5667;
								Oulu (CK+) *ar*: 0.4834;
								CK+ (Oulu) *ar*: 0.7756
									-	CD

Shirian et al. [92]	○	68	$\mathcal{ST}:\mathcal{F}$	$\mathbb{C}$	*F*	*GCN*	*Softmax*	RML [162] *ar*: 0.9411;
								eNTERFACE [163] *ar*: 0.8749;
								RAVDESS [164] *ar*: 0.8565
									-	10F CV

Jin et al. [81]	●	68	$\mathcal{S}:\mathcal{R}$	*LBP*/*AE*	*F*/*M*	*GCN*	*Softmax*	CK+ *ar*: 0.9432 (w/o 0.9286); Oulu *ar*: 0.7328 (w/o 0.6764)	RAF *ar*: 0.5825 (w/o 0.5663)	10F-SI CV /HO

Zhao et al. [74]	●	68	$\mathcal{S}:\mathcal{L}$	$\mathbb{C}$+*CNN*	*M*	*GCN*	*Softmax*	CK+ (RAF) *ar*: 0.9320 (w/o 0.7864);
								MMI (RAF) *ar*: 0.7439 (w/o 0.5707);
								Oulu (RAF) *ar*: 0.6302 (w/o 0.4875)
									SFEW *ar*: 0.5711
									RAF *ar*: 0.8752
										10F-SI CV
										/HO

Rao et al. [113]	●	68	$\mathcal{S}:\mathcal{L}$	$\mathbb{C}$	*M*	*GCN*	*Softmax*	CK+ *ar*: 0.9868; JAFEE *ar*: 0.9664	FER2013 *ar*: 0.7806 RAF *ar*: 0.8695	10F/LO CV /HO
Liu et al. [101]	●	-	$\mathcal{AU}:\mathcal{M}$	*CNN*	*M*	*GCN*	*Softmax*	-	RAF *ar*: 0.8931 (w/o 0.8582) AffNet *ar*: 0.6157 (w/o 0.5794)	HO
Antoniadis et al. [105]	●	-	$\mathcal{X}$	Φ	*L*	*GCN*	*Softmax*	-	AffNet *ar*: 0.6646 (w/o 0.6437), *val_ccc*: 0.767 (w/o 0.761), *aro_ccc*: 0.649 (w/o 0.628); Aff-Wild2 *ar*: 0.4892 (w/o 0.4506), *val_ccc*: 0.457 (w/o 0.416), *aro_ccc*: 0.514 (w/o 0.501)	HO

[1] Prep.: processing; B.B.: bounding boxes; LM: landmarks; ● = Yes, ○ = No.
[2] $\mathcal{S}/\mathcal{ST}/\mathcal{AU}/\mathcal{X}$: {spatial; spatio-temporal; AU-level; sample-level} representation; : $\mathcal{L}/\mathcal{R}/\mathcal{B}/\mathcal{M}/\mathcal{F}$: {landmark; region; label; map; frame}-level graph; $\mathbb{C}$: landmark coordinates; Φ: label distributions; △: triangulation; *L/M/F*: {learning; manual; full} connections; $\mathbb{E}$: Euclidean distance; $\mathbb{H}$: Hop distance.
[3] *ar*: average accuracy rate; *val/aro*: valance/arousal; *ccc*: concordance correlation coefficient; DB1 (DB2): train on database 2, test on database 1; w/o: without graph.
[4] CV: cross validation; LO: leave-one-subject-out; HO: holdout validation; 10F: 10-flods; SI: subject independent; CD (DB): cross database validation (training database).

for the spatio-temporal graph [89]. The possible reason is that the landmark information is helpful and practical for locating the facial areas where AUs may occur. In this setting, their node representations were close to that in spatial graphs of FER methods, which usually combined geometric coordinates with appearance features (e.g., *HOG* [36, 40]). Although some AUD methods using AU-level graphs also exploited traditional features (e.g., *Gabor* [93, 94], *LBP* [42]) or deep features (e.g., *VGG* [97]) to introduce the appearance information, their graph representations were initialized from the AU label distribution of the training set. Thus, the *DBN* model has become popular in the relational reasoning stage [42, 93, 94]. Another similar trend to graph-based FER is that *GNNs* have been widely utilized to learn the latent dependency among individual AUs in recent studies, such as *GCN* [24, 80, 89, 98], *GAT* [99], *GGNN* [23], and *DG-CNN* [38]. But the difference is that *fully-connected (FC)* layers [23, 36, 80, 98] or regression models [24, 38, 102] are often

TABLE 5
Performance summary of representative graph-based AUD methods

References	Prep.[1]		Representation[2]			Reasoning		Database[3]	Validation[4]
	B.B.	LM	Category	Node	Edge	Model	Output		
Zhu et al. [94]	●	49	$\mathcal{AU}:\mathcal{B}$	$\mathbb{C}$+*Gabor*	Φ	*DBN*	*MPE*	CK+ *ar*: 0.9048 (w/o 0.8818), f_1: 0.7072 (w/o 0.4976); DISFA *ar*: 0.9356 (w/o 0.9459), f_1: 0.7095 (w/o 0.6553)	2F, 10F CV
Kaltwang et al. [73]	○	66	$\mathcal{S}:\mathcal{L}$	$\mathbb{C}$	*L*	*LT*		DISFA *corr*: 0.43, *mse*: 0.39, *icc*: 0.36	9F CV
Walecki et al. [102]	○	-	$\mathcal{AU}:\mathcal{M}$	*CNN*	*L*	*CRF*	*Ordinal Regression*	FERA 2015 [165] *icc*: 0.63, *mae*: 1.23; DISFA *icc*: 0.45, *mae*: 0.61	HO
Corneanu et al. [97]	●	-	$\mathcal{AU}:\mathcal{M}$	*VGG*	Φ	*RNN*	*Sigmoid*	BP4D f_1: 0.617; DISFA f_1: 0.567	3F-SI CV

Dapogny et al. [40]	●	49	$\mathcal{S}:\mathcal{L}$	$\mathbb{C}$+*HOG*	$\triangle$	*AE*	*RF*	CK+ *auc*: 0.953, f_1: 0.788, *nf*$_1$: 0.865;
								BP4D: *auc*: 0.727, f_1: 0.557, *nf*$_1$: 0.636;
								DISFA *auc*: 0.824, f_1: 0.491
									10F-SI CV

Li et al. [23]	●	20	$\mathcal{AU}:\mathcal{M}$	$\mathbb{C}$+*VGG*	*M+L*	*GGNN*	*FC*	BP4D *auc*: 0.741 (w/o 0.736), f_1: 0.629 (w/o 0.620); DISFA *auc*: 0.807 (w/o 0.802), f_1: 0.559 (w/o 0.530)	No report
Niu et al. [98]	●	-	$\mathcal{AU}:\mathcal{M}$	$\mathbb{W}$	*L*	*GCN*	*FC*	BP4D f_1: 0.598 (w/o 0.565); EmotioNet f_1: 0.681 (w/o 0.623)	3F-SI CV/ HO
Liu et al. [80]	●	19	$\mathcal{S}:\mathcal{R}$	*CNN+AE*	*M*	*GCN*	*FC*	BP4D *auc*: 0.873, f_1: 0.628; DISFA *auc*: 0.746, f_1: 0.550	3F-SI CV
Fan et al. [38]	●	20	$\mathcal{S}:\mathcal{R}$	*ResNet*	*KNN*+$\mathbb{E}$	*DG-CNN*	*Heatmap Regression*	BP4D *icc*: 0.72 (w/o 0.67), *mae*: 0.58 (w/o 0.60); DISFA *icc*: 0.47 (w/o 0.40), *mae*: 0.20 (w/o 0.21)	3F-SI CV
Liu et al. [36]	●	68	$\mathcal{S}:\mathcal{L}$	$\mathbb{C}$+*HOG*	$\triangle$	*CNN*	*FC*	CK+ *auc*: 0.929	10F-SI CV
Zhang et al. [24]	●	18	$\mathcal{S}:\mathcal{R}$	*HRN*	*L*	*GCN*	*Heatmap Regression*	BP4D f_1: 0.635 (w/o 0.620); DISFA f_1: 0.620	3F-SI CV

Cui et al. [42]	○	51	$\mathcal{AU}:\mathcal{B}$	*LBP*	Φ	*DBN*	*LR /CNN /SVM*	CK+ f_1: 0.830 (w/o 0.785); BP4D f_1: 0.687 (w/o 0.657);
								EmotioNet f_1: 0.626 (w/o 0.620);
								MMI (CK+) f_1: 0.532 (w/o 0.482)
									5F-SI CV /CD

Cui et al. [37]	○	-	$\mathcal{AU}:\mathcal{B}$	*VGG*	Φ	*DBN*	*FC*	CK+ f_1: 0.74 (w/o 0.69); BP4D f_1: 0.57 (w/o 0.56); MMI f_1: 0.58 (w/o 0.47)	5F-SI CV
Song et al. [99]	●	-	$\mathcal{AU}:\mathcal{M}$	*ResNet*	*Mask*	*GAT*	*Softmax*	BP4D *ar*: 0.782 (w/o 0.781), f_1: 0.633 (w/o 0.599); DISFA *ar*: 0.934, f_1: 0.600	3F-SI CV
Song et al. [100]	●	-	$\mathcal{AU}:\mathcal{M}$	*ResNet*	Φ	*Hybrid GNN*	*Softmax*	BP4D f_1: 0.634 (w/o 0.596); DISFA f_1: 0.610	3F-SI CV
Song et al. [25]	●	-	$\mathcal{AU}:\mathcal{M}$	*ResNet*	Φ	*GCN+LSTM*	*FC*	FERA 2015 *icc*: 0.72 (w/o 0.68), *mae*: 0.57 (w/o 0.64); DISFA *icc*: 0.56 (w/o 0.51), *mae*: 0.22 (w/o 0.35)	3F-SI CV
Chen et al. [89]	●	68	$\mathcal{ST}:\mathcal{L}$	*DCT+CNN*	*M+Mask*	*GCN*	*Softmax*	BP4D f_1: 0.6489 (w/o 0.5842); DISFA f_1: 0.6585 (w/o 0.5487)	3F-SI CV
Luo et al. [96]	●	-	$\mathcal{AU}:\mathcal{M}$	*ResNet /Transformer*	*KNN+L*	*GCN*	*Softmax*	BP4D f_1: 0.655 (w/o 0.626); DISFA f_1: 0.631 (w/o 0.591)	3F-SI CV

[1] Prep.: processing; B.B.: bounding boxes; ● = Yes, ○ = No; LM: landmarks.

[2] $\mathcal{S}/\mathcal{ST}/\mathcal{AU}$: {spatial; spatio-temporal; AU-level} representation; : $\mathcal{L}/\mathcal{R}/\mathcal{B}/\mathcal{M}$: {landmark; region; label; map}-level graph; $\mathbb{C}$: landmark coordinates; Φ: label distributions; $\triangle$: triangulation; *L/M*: {learning; manual} connections; $\mathbb{E}$: Euclidean distance.

[3] *ar*: average accuracy rate; f_1: F1 score; *nf*$_1$: F1-norm score; *corr*: Pearson correlation coefficient; *mae*: mean absolute error; *mse*: mean squared error; *icc*: intra-class correlation coefficient; *auc*: area under the receiver operating characteristic curve; DB1 (DB2): train on database 2, test on database 1; w/o: without graph.

[4] CV: cross validation; LO: leave-one-subject-out; (K)F: k-folds; SI: subject independent; HO: holdout validation; CD: cross database validation.

applied for predictions instead of *softmax* classifier [89, 99].

A particular line of AUD research analyzes the facial affects by estimating the AU intensities, which could have greater information value in understanding complex affective states [166]. Even though a few attempts in estimating AU intensities based on graph structures have existed [38, 73, 97], the study of using the latest spatio-temporal graph representations and *GNNs* has not been reported. Another big challenge in AUD is few and imbalanced samples. Recent graph-based methods using transfer learning [42, 98] or uncertainty learning [99] were proposed to address this problem. They showed an advantage of the graph-based method in this topic and are helpful to implement AUD in large-scale unlabeled data.

5.4 Micro-Expression Recognition

Micro-expressions are fleeting and involuntary facial affects that people usually exhibit in high stake situations when attempting to conceal or mask their true feelings [151]. The earliest well-known studies came from [167] as well as [168]. Generally, a micro-expression lasts only 1/25 to 1/2 seconds long and is too subtle and fleeting for an untrained person to perceive. Therefore, developing an automatic micro-expression recognition (MER) system is valuable in reading human hidden affective states. Besides the short duration, low intensity and localization characteristics also make it challenging.

To this end, graph-based MER methods have been designed to address the above challenges and have become appealing in the past two years [82], especially in 2020 [39, 75]. Table 6 lists the reported performance of a few representative recent studies of graph-based MER. These methods fall into the landmark-level spatial graph [75, 82] and the AU-level graph [39] in terms of representation types. For the former, their idea is to use landmarks to locate and analyze specific facial areas to deal with the local response and the subtleness of micro-expressions. The latter aims to infer the AU relationship to improve the final performance. The difference in processing ideas is also reflected in the reasoning procedure. Approaches like *GSC* [82] and variant *CNNs* [75] are exploited in the landmark-level graph to integrate the individual node feature representations. In comparison, *GCNs* are employed to learn an optimal graph structure of the AU dependency knowledge from training data and make predictions. Nevertheless, one common thing is that all the methods consider the local appearance in a spatio-temporal way by using *optical-flow* or *DNNs*.

A problem in graph-based MER is the lack of large-scale in-the-wild data. The small sample size limits the AU-level graph representation that relies on initializing the AU relationship from the AU label distribution of the training set. The lab-controlled data make it difficult to follow the trend in FER studies, which generalizes the graph-based FAA methods in real-world scenarios. However, the analysis of uncontrolled micro-expressions is fundamental because micro-expressions and macro-expressions can co-occur in many real cases. For example, the furrowing on the forehead slightly and quickly when smiling indicates the true feeling [168]. Since the evolutionary appearance information is crucial for the micro-expression analysis, building a spatio-temporal graph representation that can model the duration and the dynamic of micro-expressions is also a helpful but unexplored topic.

5.5 Other Tasks

The graph-based methods also play a vital role in several other FAA tasks, such as pain detection [73], non-basic affect recognition [85, 106], occluded FER [40, 41], and multi-modal affect recognition [88, 103]. Table 7 summarizes the latest graph-based FAA methods for these tasks. Their node representations and edge initialization strategies for graph constructions in this field are similar to those in graph-based FER, MER, and AUD methods. While for the reasoning step, *GCN* is the top-1 option. This observation implies that the framework of the graph-based method discussed in this paper can be easily extended to many other FAA tasks and promote performance improvement.

6 Open Directions

Graph-based FAA methods have been dissected into fundamental components for elaboration and discussion in this review. When encoding facial affect into graphs, strategies vary according to node and edge elements. Relational reasoning approaches infer latent relationships or inherent dependencies of graph nodes in terms of space, time, and semantics. The category of graph representations will affect the technique choice of relational reasoning to a certain extent.

Despite significant advances and numerous work, the graph-based FAA is still an appealing field with many open directions. Due to advantages in modeling and reasoning latent relationships of facial affects, graph-based methods may provide complementary information to help solve some challenges that non-graph-based approaches face. Also, the graph-based method has natural advantages or unexplored research potential in other topics.

6.1 In-the-wild Scenarios

Although many efforts have been made for graph-based FAA in natural conditions [36, 39, 40, 41, 42, 43, 83, 98, 101, 105], even the state-of-the-art performance is far from actual applications. Factors like illumination, head pose, and part occlusion are challenging in constructing an effective graph representation. For one thing, significant illumination changes and head pose variations will impair the accuracy of face detection and registration, which is vital for establishing landmark-level graphs. ROI graphs without landmarks or NPI graphs [83, 86] should be a possible direction to avoid this problem. Also, missing face parts resulting from camera view or context occlusion make it challenging to encode enough facial information and obtain meaningful connections in an affective graph. Pilot work [40, 41] has tried to exploit a sub-graph without masked facial parts or generate adaptive edge links to alleviate the influence. Unfortunately, there has still been a considerable performance decrease compared to normal conditions. Proposing more effective spatio-temporal graphs can account for these problems based on evolutional affective information.

6.2 3D and 4D Facial Affects

Using 3D and 4D face images might be another good topic because the 3D face shape provides additional depth information and dynamically contains subtle facial deformations. They are intuitively insensitive to pose and light changes. Some studies have transformed 3D faces into 2D images and generated graph representations [23, 38, 40, 97], but they have not fully taken advantage of the 3D data. Alternatively, non-graph-based [175] and graph-based methods [78] have been explored to conduct FAA directly on 3D or 4D faces. Since the 3D face mesh structure is naturally close to the graph structure, employing the graph representation and reasoning to handle 3D face images will promote the improvement of in-the-wild FAA. Besides, there is also a potential topic of using 3D and 4D data with graph-based methods, especially landmark-level graphs and *GNNs*, in micro-expression recognition.

6.3 Valence and Arousal

Estimating the continuous dimension is a rising topic in FAA. Unlike discrete labels, Valence-Arousal (V-A) is a kind of quantitative annotation with dynamic evolutions, which provide another domain to analyze facial affects, especially in a temporal domain. Large-scale FAA databases (Aff-Wild I [15], II [160]) containing V-A annotations have been released to support the continuous FAA. Existing graph-based methods mainly perform the V-A measurement [38, 73, 102] on lab-controlled databases except for a few studies like [85, 105]. Recent graph-based methods have studied multi-label learning according to intrinsic mappings between facial affect categories and other annotations [23, 42, 105]. Such underlying assumptions can also be extended to the V-A measurement task, where AU-level graphs and *DBNs*, as well as sample-level graphs, are potential directions.

6.4 Context and Multi-modality

Most current FAA methods only consider a single face in one image or sequence. However, people usually have affective behaviors, including facial expressions, body gestures, and emotionally speaking in real cases [176]. These facial affective displays are highly associated with context surroundings that include but are not limited to the affective behavior of other people in social interactions or inanimate objects. Existing studies like [85] and [86] have employed

TABLE 6
Performance summary of representative graph-based MER methods

References	Prep.[1]		Representation[2]			Reasoning		Database[3]	Validation[4]
	B.B.	LM	Category	Node	Edge	Model	Classifier		

Liu et al. [82]	●	66	$\mathcal{S}:\mathcal{R}$	*Optical-flow*	*KNN*	*GSC*	*SVM*	SMIC (3 cl.) *ar*: 0.6795, f_1: 0.6844;
								CASME I [169] (4 cl.) *ar*: 0.7219, f_1: 0.7236;
								CASME II (5 cl.) *ar*: 0.6356, f_1: 0.6364
									LO CV

Lei et al. [75]	●	28	$\mathcal{S}:\mathcal{L}$	*TCN*	*F*	*Graph-TCN*	*Softmax*	CASME II (5 cl.) *ar*: 0.7398, f_1: 0.7246;
								SAMM (5 cl.) *ar*: 0.7500, f_1: 0.6985;
								SAMM (4 cl.) *ar*: 0.8050, f_1: 0.7657
									LO CV

Xie et al. [39]	●	-	$\mathcal{AU}:\mathcal{M}$	*CNN*	*L*	*GCN*	*Softmax*	CASME II (3 cl.) *ar*: 0.712, f_1: 0.355;
								CASME II (7 cl.) *ar*: 0.561, f_1: 0.394;
								SAMM (3 cl.) *ar*: 0.702, f_1: 0.433;
								SAMM (8 cl.) *ar*: 0.523, f_1: 0.357
								SMIC (CASME II) (3 cl.) *ar*: 0.344, f_1: 0.319;
								SMIC (SAMM) (3 cl.) *ar*: 0.451, f_1: 0.309
									LO CV
									/CD

Kumar and Bhanu [72]	●	51	$\mathcal{S}:\mathcal{L}$	$\mathbb{C}$+*Optical-flow*	*M+L*	*GCN+GAT*	*Softmax*	CASME II (3 cl.) *ar*: 0.8966, f_1: 0.8695;
								CASME II (5 cl.) *ar*: 0.8130, f_1: 0.7090;
								SAMM (3 cl.) *ar*: 0.8872, f_1: 0.8118;
								SAMM (5 cl.) *ar*: 0.8824, f_1: 0.8279
									LO CV

Lei et al. [95]	●	30	$\mathcal{AU}:\mathcal{B}$	*Embedding*	Φ	*GCN*	*Softmax*	CASME II (4 cl.) *ar*: 0.8080 (w/o 0.7880), f_1: 0.7871;
								CASME II (5 cl.) *ar*: 0.7427, f_1: 0.7047;
								SAMM (4 cl.) *ar*: 0.8239, f_1: 0.7735;
								SAMM (5 cl.) *ar*: 0.7426, f_1: 0.7045
									LO CV

[1] Prep.: processing; B.B.: bounding boxes; ● = Yes, ○ = No; LM: landmarks.
[2] $\mathcal{S}/\mathcal{AU}$: {spatial; AU-level} representation; : $\mathcal{R}/\mathcal{L}/\mathcal{B}/\mathcal{M}$: {region; landmark; label; map}-level graph; $\mathbb{C}$: landmark coordinates; Φ: label distributions; *M/L/F*: {manual; learning; fully} connections.
[3] *ar*: average accuracy rate; f_1: F1 score; (N) cl.: (N) affective classes; DB1 (DB2): train on database 2, test on database 1; w/o: without graph.
[4] CV: cross validation; LO: leave-one-subject-out; CD: cross database validation.

TABLE 7
Performance summary of graph-based methods for special FAA tasks

References	Prep.[1]		Representation[2]			Reasoning		Database[3]	Validation[4]
	B.B.	LM	Category	Node	Edge	Model	Output		
Kaltwang et al. [73]	○	66	$\mathcal{S}:\mathcal{L}$	$\mathbb{C}$	*L*	*LT*		ShoulderPain [170] *corr*: 0.23, *mse*: 0.60	8F CV
Dapogny et al. [40]	●	49	$\mathcal{S}:\mathcal{L}$	$\mathbb{C}$+*HOG*	$\triangle$	*AE*	*RF*	CK+ (eyes occluded) *ar*: 0.879; CK+ (mouth occluded) *ar*: 0.727	10F-SI CV
Zhang et al. [85]	●	-	$\mathcal{S}:\mathcal{R}$	*VGG*	*L*	*GCN*	*Softmax/FC*	EMOTIC (26 cl.) *prc*: 0.2842; EMOTIC *val_er*: 0.7, *aro_er*: 1.0, *dom_er*: 1.0	HO
Chen et al. [88]	●	68	$\mathcal{ST}:\mathcal{L}$	$\mathbb{C}$	$\triangle$+*M*	*GCN*	*FC*	CES [18] *val_ccc*: 0.515, *aro_ccc*: 0.513	HO
He and Jin [106]	○	-	$\mathcal{X}$	*Embedding*	*M+L*	*GCN*	*Softmax*	FlickLDL [171] *ar*: 0.691; TwiterLDL [172] *ar*: 0.758	HO

Zhou et al. [41]	●	44	$\mathcal{ST}:\mathcal{L}$	$\mathbb{C}$+*HOG*	*L*+$\mathbb{H}$	*GCN*	*SoftMax*	CK+ (random occlusion) *ar*: 0.9551 (w/o 0.6226);
								Oulu (random occlusion) *ar*: 0.8121 (w/o 0.5748);
								AFEW (random occlusion) *ar*: 0.4047 (w/o 0.2122)
									10F-SI CV
									/HO

Chien et al. [103]	○	-	$\mathcal{X}$	*CNN*	*Transfer Knowledge*	*GCN*	*FC*	Amigos [173] (Ascertain) *val_uar*: 0.798, *aro_uar*: 0.679; Ascertain [174] (Amigos) *val_uar*: 0.704, *aro_uar*: 0.569	CD

[1] Prep.: processing; B.B.: bounding boxes; ● = Yes, ○ = No; LM: landmarks.
[2] $\mathcal{S}/\mathcal{ST}/\mathcal{X}$: {spatial; spatio-temporal; sample-level} representation; : $\mathcal{L}/\mathcal{R}$: {landmark; region}-level graph; $\mathbb{C}$: landmark coordinates; $\triangle$: triangulation; *M/L*: {manual/learning} connections; $\mathbb{H}$: Hop distance.
[3] *corr*: Pearson correlation coefficient; *mse*: mean squared error; *ar*: average accuracy rate; *prc*: area under the precision recall curves; *er*: average error rate; *val/aro/dom*: valance/arousal/dominance; *ccc*: concordance correlation coefficient; *uar*: unweighted average recall; (N) cl.: (N) affective classes; DB1 (DB2): train on database 2, test on database 1; w/o: without graph.
[4] CV: cross validation; (K)F: k-folds; SI: subject independent; HO: holdout validation; CD: cross database validation.

graph reasoning to infer relationships between the target face and other objects in the same image. Facial affects and other helpful contexts can be combined in a graph representation to perform the analysis on a fuller scope, such as the gesture [177, 178]. Another valuable topic is to introduce additional data channels that are multi-modality. Sample-level and spatio-temporal have also been successfully extended to process multi-modal affect analysis tasks with audio [88] and physiological signal [103], respectively, which shows a good research prospect.

6.5 Transfer Learning and Cross-database

Insufficient and invalid annotations are significant challenges that limit the development of FAA research, especially for deep learning. One possible solution is to use graph-based transfer learning, which could build bridges among different label spaces such as discrete-to-continuous labels [105] and emotional-to-AU labels [101]. Efforts of using the graph structure have been explored to solve these challenges in terms of semi-supervision [98], label correction [42], or uncertainty measurement [99]. On the other hand, the cross-database performance of features extracted using graph-based methods has been demonstrated in all FER [43, 86], AUD [42, 93], MER [39], and cross-corpus analysis [103, 104]. Considering the strength of AU-level and sample-level graphs in searching intrinsic distributions, a universal affective feature encoder could be expected to break down barriers across various databases and achieve better generalization capability.

Acknowledgments

Funding: This work was supported by the China Scholarship Council [CSC, No.202006150091], and the Ministry of Education and Culture of Finland for AI forum project.

The authors would like to thank Muzammil Behzad and Tuomas Varanka for providing materials and suggestions for the figures used in this paper.

REFERENCES

[1] A. Mehrabian *et al.*, *Silent messages*. Wadsworth Belmont, CA, 1971, vol. 8, no. 152.

[2] A. Mehrabian and J. A. Russell, *An approach to environmental psychology*. the MIT Press, 1974.

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

[4] M. S. Gazzaniga, R. B. Ivry, and G. Mangun, *Cognitive Neuroscience. The Biology of the Mind, (2014)*. Norton: New York, 2014.

[5] R. W. Picard, E. Vyzas, and J. Healey, "Toward machine emotional intelligence: Analysis of affective physiological state," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 10, pp. 1175–1191, 2001.

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

[7] E. Sariyanidi, H. Gunes, and A. Cavallaro, "Automatic analysis of facial affect: A survey of registration, representation, and recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 6, pp. 1113–1133, 2014.

[8] M. G. Calvo, A. Gutiérrez-García, and M. Del Líbano, "What makes a smiling face look happy? visual saliency, distinctiveness, and affect," *Psychological Research*, vol. 82, no. 2, pp. 296–309, 2018.

[9] M. Tavakolian and A. Hadid, "A spatiotemporal convolutional neural network for automatic pain intensity estimation from facial dynamics," *International Journal of Computer Vision*, vol. 127, no. 10, pp. 1413–1425, 2019.

[10] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "From facial expression recognition to interpersonal relation prediction," *International Journal of Computer Vision*, vol. 126, no. 5, pp. 550–569, 2018.

[11] A. V. Savchenko, L. V. Savchenko, and I. Makarov, "Classifying emotions and engagement in online learning based on a single facial expression recognition neural network," *IEEE Transactions on Affective Computing*, 2022.

[12] L. Zhao, X. Peng, Y. Tian, M. Kapadia, and D. N. Metaxas, "Towards image-to-video translation: A structure-aware approach via multi-stage generative adversarial networks," *International Journal of Computer Vision*, vol. 128, no. 10, pp. 2514–2533, 2020.

[13] M. F. Valstar, E. Sánchez-Lozano, J. F. Cohn, L. A. Jeni, J. M. Girard, Z. Zhang, L. Yin, and M. Pantic, "Fera 2017-addressing head pose in the third facial expression recognition and analysis challenge," in *IEEE International Conference on Automatic Face & Gesture Recognition (FG)*. IEEE, 2017, pp. 839–847.

[14] A. Dhall, G. Sharma, R. Goecke, and T. Gedeon, "Emotiw 2020: Driver gaze, group emotion, student engagement and physiological signal based challenges," in *International Conference on Multimodal Interaction (ICMI)*, 2020, pp. 784–789.

[15] S. Zafeiriou, D. Kollias, M. A. Nicolaou, A. Papaioannou, G. Zhao, and I. Kotsia, "Aff-wild: valence and arousal'in-the-wild'challenge," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017, pp. 34–41.

[16] D. Kollias, "Abaw: Valence-arousal estimation, expression recognition, action unit detection & multi-task learning challenges," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2328–2336.

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

[18] F. Ringeval, B. Schuller, M. Valstar, N. Cummins, R. Cowie, L. Tavabi, M. Schmitt, S. Alisamir, S. Amiriparian, E.-M. Messner *et al.*, "Avec 2019 workshop and challenge: state-of-mind, detecting depression with ai, and cross-cultural affect recognition," in *International on Audio/Visual Emotion Challenge and Workshop*, 2019, pp. 3–12.

[19] L. Stappen, A. Baird, G. Rizos, P. Tzirakis, X. Du, F. Hafner, L. Schumann, A. Mallol-Ragolta, B. W. Schuller, I. Lefter *et al.*, "Muse 2020 challenge and workshop: Multimodal sentiment analysis, emotion-target engagement and trustworthiness detection in real-life media: Emotional car reviews in-the-wild," in *International on Multimodal Sentiment Analysis in Real-life Media Challenge and Workshop*, 2020, pp. 35–44.

[20] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 6, pp. 915–928, 2007.

[21] E. Friesen and P. Ekman, "Facial action coding system: a technique for the measurement of facial movement," *Palo Alto*, vol. 3, no. 2, p. 5, 1978.

[22] L. Zhong, C. Bai, J. Li, T. Chen, S. Li, and Y. Liu, "A graph-structured representation with brnn for static-based facial expression recognition," in *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*. IEEE, 2019, pp. 1–5.

[23] G. Li, X. Zhu, Y. Zeng, Q. Wang, and L. Lin, "Semantic relationships guided representation learning for facial action unit recognition," in *AAAI Conference on Artificial Intelligence (AAAI)*, vol. 33, no. 01, 2019, pp. 8594–8601.

[24] Z. Zhang, T. Wang, and L. Yin, "Region of interest based graph convolution: A heatmap regression approach for action unit detection," in *ACM International Conference on Multimedia (ACM MM)*, 2020, pp. 2890–2898.

[25] T. Song, Z. Cui, Y. Wang, W. Zheng, and Q. Ji, "Dynamic probabilistic graph convolution for facial action unit intensity estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 4845–4854.

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

[27] J. F. Cohn and F. De la Torre, *Automated face analysis for affective computing*. Oxford University Press, 2015.

[28] U. Zarins and S. Kondrats, *Anatomy for sculptors: understanding the human figure*. Anatomy Next, Incorporated, 2015.

[29] D. L. Bimler and G. V. Paramei, "Facial-expression affective attributes and their configural correlates: components and categories," *Spanish Journal of Psychology*, vol. 9, no. 1, p. 19, 2006.

[30] P. Ekman, "Facial action coding system (facs)," *A Human Face*, 2002.

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

[32] Y. Li, G. Lu, J. Li, Z. Zhang, and D. Zhang, "Facial expression recognition in the wild using multi-level features and attention mechanisms," *IEEE Transactions on Affective Computing*, 2020.

[33] G. M. Jacob and B. Stenger, "Facial action unit detection with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 7680–7689.

[34] B. Martinez, M. F. Valstar, B. Jiang, and M. Pantic, "Automatic analysis of facial actions: A survey," *IEEE Transactions on Affective Computing*, vol. 10, no. 3, pp. 325–347, 2017.

[35] L. F. Barrett, *How emotions are made: The secret life of the brain*. Houghton Mifflin Harcourt, 2017.

[36] Y. Liu, X. Zhang, Y. Lin, and H. Wang, "Facial expression recognition via deep action units graph network based on psychological mechanism," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 12, no. 2, pp. 311–322, 2019.

[37] Z. Cui, T. Song, Y. Wang, and Q. Ji, "Knowledge augmented deep neural networks for joint facial expression and action unit recognition," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020.

[38] Y. Fan, J. Lam, and V. Li, "Facial action unit intensity estimation via semantic correspondence learning with dynamic graph convolution," in *AAAI Conference on Artificial Intelligence (AAAI)*, vol. 34, no. 07, 2020, pp. 12 701–12 708.

[39] H.-X. Xie, L. Lo, H.-H. Shuai, and W.-H. Cheng, "Au-assisted graph attention convolutional network for micro-expression recognition," in *ACM International Conference on Multimedia (ACM MM)*, 2020, pp. 2871–2880.

[40] A. Dapogny, K. Bailly, and S. Dubuisson, "Confidence-weighted local expression predictions for occlusion handling in expression recognition and action unit detection," *International Journal of Computer Vision*, vol. 126, no. 2, pp. 255–271, 2018.

[41] J. Zhou, X. Zhang, and Y. Liu, "Learning the connectivity: Situational graph convolution network for facial expression recognition," in *IEEE International Conference on Visual Communications and Image Processing*. IEEE, 2020, pp. 230–234.

[42] Z. Cui, Y. Zhang, and Q. Ji, "Label error correction and generation through label relationships," in *AAAI Conference on Artificial Intelligence (AAAI)*, vol. 34, no. 04, 2020, pp. 3693–3700.

[43] S. Chen, J. Wang, Y. Chen, Z. Shi, X. Geng, and Y. Rui, "Label distribution learning on auxiliary label space graphs for facial expression recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 13 984–13 993.

[44] C. A. Corneanu, M. O. Simón, J. F. Cohn, and S. E. Guerrero,

"Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 8, pp. 1548–1568, 2016.

[45] D. Kollias, P. Tzirakis, M. A. Nicolaou, A. Papaioannou, G. Zhao, B. Schuller, I. Kotsia, and S. Zafeiriou, "Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond," *International Journal of Computer Vision*, vol. 127, no. 6, pp. 907–929, 2019.

[46] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE Transactions on Affective Computing*, 2020.

[47] K. M. Goh, C. H. Ng, L. L. Lim, and U. U. Sheikh, "Micro-expression recognition: an updated review of current trends, challenges and solutions," *The Visual Computer*, vol. 36, no. 3, pp. 445–468, 2020.

[48] L. Zhang, B. Verma, D. Tjondronegoro, and V. Chandran, "Facial expression analysis under partial occlusion: A survey," *ACM Computing Surveys*, vol. 51, no. 2, pp. 1–49, 2018.

[49] P. V. Rouast, M. Adam, and R. Chiong, "Deep learning for human affect recognition: Insights and new developments," *IEEE Transactions on Affective Computing*, 2019.

[50] X. Ben, Y. Ren, J. Zhang, S.-J. Wang, K. Kpalma, W. Meng, and Y.-J. Liu, "Video-based facial micro-expression analysis: A survey of datasets, features and algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[51] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion." *Journal of Personality and Social Psychology*, vol. 17, no. 2, pp. 124–129, 1971.

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

[53] V. Sethu, E. M. Provost, J. Epps, C. Busso, N. Cummins, and S. Narayanan, "The ambiguous world of emotion representation," *arXiv preprint arXiv:1909.00360*, 2019.

[54] A. Mehrabian, "Framework for a comprehensive description and measurement of emotional states." *Genetic, social, and general psychology monographs*, 1995.

[55] R. Plutchik, "A general psychoevolutionary theory of emotion," in *Theories of Emotion*. Elsevier, 1980, pp. 3–33.

[56] M. K. Greenwald, E. W. Cook, and P. J. Lang, "Affective judgment and psychophysiological response: dimensional covariation in the evaluation of pictorial stimuli." *Journal of Psychophysiology*, vol. 3, no. 1, pp. 51–64, 1989.

[57] J. A. Russell, "Evidence of convergent validity on the dimensions of affect." *Journal of Personality and Social Psychology*, vol. 36, no. 10, p. 1152, 1978.

[58] M. Soleymani, S. Asghari-Esfeden, Y. Fu, and M. Pantic, "Analysis of eeg signals and facial expressions for continuous emotion detection," *IEEE Transactions on Affective Computing*, vol. 7, no. 1, pp. 17–28, 2015.

[59] L. Yin, X. Chen, Y. Sun, T. Worm, and M. Reale, "A high-resolution 3d dynamic facial expression database," in *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 2008, pp. 1–6.

[60] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "Openface: an open source facial behavior analysis toolkit," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2016, pp. 1–10.

[61] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. IEEE, 2018, pp. 59–66.

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

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

[64] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 681–685, 2001.

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

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

[67] X. Dong, Y. Yang, S.-E. Wei, X. Weng, Y. Sheikh, and S.-I. Yu, "Supervision by registration and triangulation for landmark detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[68] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2d and 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks)," in *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 1021–1030.

[69] I. Masi, Y. Wu, T. Hassner, and P. Natarajan, "Deep face recognition: A survey," in *SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*. IEEE, 2018, pp. 471–478.

[70] N. Wang, X. Gao, D. Tao, H. Yang, and X. Li, "Facial feature point detection: A comprehensive survey," *Neurocomputing*, vol. 275, pp. 50–65, 2018.

[71] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *IEEE Conference on Computer Vision and Pattern Recognition-Workshops*. IEEE, 2010, pp. 94–101.

[72] A. J. R. Kumar and B. Bhanu, "Micro-expression classification based on landmark relations with graph attention convolutional network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 1511–1520.

[73] S. Kaltwang, S. Todorovic, and M. Pantic, "Latent trees for estimating intensity of facial action units," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 296–304.

[74] R. Zhao, T. Liu, Z. Huang, D. P.-K. Lun, and K. K. Lam, "Geometry-aware facial expression recognition via attentive graph convolutional networks," *IEEE Transactions on Affective Computing*, 2021.

[75] L. Lei, J. Li, T. Chen, and S. Li, "A novel graph-tcn with a graph structured representation for micro-expression recognition," in *ACM International Conference on Multimedia (ACM MM)*, 2020, pp. 2237–2245.

[76] S. Mohseni, N. Zarei, and S. Ramazani, "Facial expression recognition using anatomy based facial graph," in *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2014, pp. 3715–3719.

[77] Y.-J. Liu, J.-K. Zhang, W.-J. Yan, S.-J. Wang, G. Zhao, and X. Fu, "A main directional mean optical flow feature for spontaneous micro-expression recognition," *IEEE Transactions on Affective Computing*, vol. 7, no. 4, pp. 299–310, 2015.

[78] Y. Pei and H. Zha, "3d facial expression editing based on the dynamic graph model," in *IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2009, pp. 1354–1357.

[79] I. Kotsia and I. Pitas, "Facial expression recognition in image sequences using geometric deformation features and support vector machines," *IEEE Transactions on Image Processing*, vol. 16, no. 1, pp. 172–187, 2006.

[80] Z. Liu, J. Dong, C. Zhang, L. Wang, and J. Dang, "Relation modeling with graph convolutional networks for facial action unit detection," in *International Conference on Multimedia Modeling (MMM)*. Springer, 2020, pp. 489–501.

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

[82] Y.-J. Liu, B.-J. Li, and Y.-K. Lai, "Sparse mdmo: Learning a discriminative feature for spontaneous micro-expression recognition," *IEEE Transactions on Affective Computing*, vol. 12, no. 1, pp. 254–261, 2021.

[83] A. Yao, J. Shao, N. Ma, and Y. Chen, "Capturing au-aware facial features and their latent relations for emotion recognition in the wild," in *ACM on International Conference on Multimodal Interaction (ICMI)*, 2015, pp. 451–458.

[84] S. Zafeiriou and I. Pitas, "Discriminant graph structures for facial expression recognition," *IEEE Transactions on Multimedia*, vol. 10, no. 8, pp. 1528–1540, 2008.

[85] M. Zhang, Y. Liang, and H. Ma, "Context-aware affective graph reasoning for emotion recognition," in *IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2019, pp. 151–156.

[86] Y. Xie, T. Chen, T. Pu, H. Wu, and L. Lin, "Adversarial graph representation adaptation for cross-domain facial expression

recognition," in *ACM International Conference on Multimedia (ACM MM)*, 2020, pp. 1255–1264.

[87] J. Zhou, X. Zhang, Y. Liu, and X. Lan, "Facial expression recognition using spatial-temporal semantic graph network," in *IEEE International Conference on Image Processing (ICIP)*. IEEE, 2020, pp. 1961–1965.

[88] H. Chen, Y. Deng, S. Cheng, Y. Wang, D. Jiang, and H. Sahli, "Efficient spatial temporal convolutional features for audiovisual continuous affect recognition," in *International on Audio/Visual Emotion Challenge and Workshop (AVEC)*, 2019, pp. 19–26.

[89] Y. Chen, D. Chen, Y. Wang, T. Wang, and Y. Liang, "Cafgraph: Context-aware facial multi-graph representation for facial action unit recognition," in *Proceedings of the 29th ACM International Conference on Multimedia (ACM MM)*, 2021, pp. 1029–1037.

[90] A. R. Rivera and O. Chae, "Spatiotemporal directional number transitional graph for dynamic texture recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 10, pp. 2146–2152, 2015.

[91] D. Liu, H. Zhang, and P. Zhou, "Video-based facial expression recognition using graph convolutional networks," in *International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 607–614.

[92] A. Shirian, S. Tripathi, and T. Guha, "Dynamic emotion modeling with learnable graphs and graph inception network," *IEEE Transactions on Multimedia*, 2021.

[93] Y. Tong, W. Liao, and Q. Ji, "Facial action unit recognition by exploiting their dynamic and semantic relationships," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 10, pp. 1683–1699, 2007.

[94] Y. Zhu, S. Wang, L. Yue, and Q. Ji, "Multiple-facial action unit recognition by shared feature learning and semantic relation modeling," in *International Conference on Pattern Recognition (ICPR)*. IEEE, 2014, pp. 1663–1668.

[95] L. Lei, T. Chen, S. Li, and J. Li, "Micro-expression recognition based on facial graph representation learning and facial action unit fusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 1571–1580.

[96] C. Luo, S. Song, W. Xie, L. Shen, and H. Gunes, "Learning multidimensional edge feature-based au relation graph for facial action unit recognition," in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2022, pp. 1239–1246.

[97] C. Corneanu, M. Madadi, and S. Escalera, "Deep structure inference network for facial action unit recognition," in *European Conference on Computer Vision (ECCV)*, 2018, pp. 298–313.

[98] X. Niu, H. Han, S. Shan, and X. Chen, "Multi-label coregularization for semi-supervised facial action unit recognition," in *International Conference on Neural Information Processing Systems (NeurIPS)*, vol. 32, 2019, pp. 1–11.

[99] T. Song, L. Chen, W. Zheng, and Q. Ji, "Uncertain graph neural networks for facial action unit detection," in *AAAI Conference on Artificial Intelligence (AAAI)*, 2021, pp. 1–10.

[100] T. Song, Z. Cui, W. Zheng, and Q. Ji, "Hybrid message passing with performance-driven structures for facial action unit detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 6267–6276.

[101] Y. Liu, X. Zhang, J. Kauttonen, and G. Zhao, "Uncertain label correction via auxiliary action unit graphs for facial expression recognition," in *International Conference on Pattern Recognition (ICPR)*, 2022, pp. 1–7.

[102] R. Walecki, O. Rudovic, V. Pavlovic, B. Schuller, and M. Pantic, "Deep structured learning for facial expression intensity estimation," *Image and Vision Computting*, vol. 259, pp. 143–154, 2017.

[103] W.-S. Chien, H.-C. Yang, and C.-C. Lee, "Cross corpus physiological-based emotion recognition using a learnable visual semantic graph convolutional network," in *ACM International Conference on Multimedia (ACM MM)*, 2020, pp. 2999–3006.

[104] D. Chen, P. Song, and W. Zheng, "Learning transferable sparse representations for cross-corpus facial expression recognition," *IEEE Transactions on Affective Computing*, 2021.

[105] P. Antoniadis, P. P. Filntisis, and P. Maragos, "Exploiting emotional dependencies with graph convolutional networks for facial expression recognition," in *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 2021, pp. 1–8.

[106] T. He and X. Jin, "Image emotion distribution learning with graph convolutional networks," in *International Conference on Multimedia Retrieval (ICMR)*, 2019, pp. 382–390.

[107] K. Zhao, W.-S. Chu, and H. Zhang, "Deep region and multi-label learning for facial action unit detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 3391–3399.

[108] Y.-H. Oh, J. See, A. C. Le Ngo, R. C.-W. Phan, and V. M. Baskaran, "A survey of automatic facial micro-expression analysis: databases, methods, and challenges," *Frontiers in Psychology*, vol. 9, p. 1128, 2018.

[109] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Facial landmark detection by deep multi-task learning," in *European conference on computer vision (ECCV)*. Springer, 2014, pp. 94–108.

[110] P. Kakumanu and N. Bourbakis, "A local-global graph approach for facial expression recognition," in *IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE, 2006, pp. 685–692.

[111] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 1867–1874.

[112] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1. Ieee, 2005, pp. 886–893.

[113] T. Rao, J. Li, X. Wang, Y. Sun, and H. Chen, "Facial expression recognition with multi-sale graph convolutional networks," *IEEE MultiMedia*, 2021.

[114] C. Liu and H. Wechsler, "Independent component analysis of gabor features for face recognition," *IEEE Transactions on Neural Networks*, vol. 14, no. 4, pp. 919–928, 2003.

[115] Y. Liu, X. Zhang, J. Zhou, and L. Fu, "Sg-dsn: A semantic graph-based dual-stream network for facial expression recognition," *Neurocomputing*, vol. 462, pp. 320–330, 2021.

[116] S. Afzal, T. M. Sezgin, Y. Gao, and P. Robinson, "Perception of emotional expressions in different representations using facial feature points," in *International Conference on Affective Computing and Intelligent Interaction and Workshops*. IEEE, 2009, pp. 1–6.

[117] T. Baltrušaitis, L. D. Riek, and P. Robinson, "Synthesizing expressions using facial feature point tracking: how emotion is conveyed," in *Proceedings of the 3rd international workshop on Affective interaction in natural environments*, 2010, pp. 27–32.

[118] J. Lee, S. Kim, S. Kim, J. Park, and K. Sohn, "Context-aware emotion recognition networks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)*, 2019, pp. 10 143–10 152.

[119] K. Wang, X. Peng, J. Yang, D. Meng, and Y. Qiao, "Region attention networks for pose and occlusion robust facial expression recognition," *IEEE Transactions on Image Processing*, vol. 29, pp. 4057–4069, 2020.

[120] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[121] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[122] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2016.

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

[124] Q. Liu *et al.*, "Phase space reconstruction driven spatio-temporal feature learning for dynamic facial expression recognition," *IEEE Transactions on Affective Computing*, 2020.

[125] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems (NeurIPS)*, 2013, pp. 3111–3119.

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

[127] P. Berkes, F. Wood, and J. Pillow, "Characterizing neural dependencies with copula models," in *International Conference on Neural Information Processing Systems (NeurIPS)*, vol. 21, 2008, pp. 129–136.

[128] C. Kemp and J. B. Tenenbaum, "The discovery of structural form," *National Academy of Sciences*, vol. 105, no. 31, pp. 10 687–10 692, 2008.

[129] R. El Kaliouby and P. Robinson, "Real-time inference of complex mental states from facial expressions and head gestures," in *Real-*

time vision for human-computer interaction. Springer, 2005, pp. 181–200.

[130] T. Baltrušaitis, D. McDuff, N. Banda, M. Mahmoud, R. El Kaliouby, P. Robinson, and R. Picard, "Real-time inference of mental states from facial expressions and upper body gestures," in *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG)*. IEEE, 2011, pp. 909–914.

[131] Y. Li, D. Tarlow, M. Brockschmidt, and R. Zemel, "Gated graph sequence neural networks," in *International Conference on Learning Representations (ICLR)*, 2016, pp. 1–20.

[132] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv preprint arXiv:1803.01271*, 2018.

[133] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, "Dynamic graph cnn for learning on point clouds," *ACM Transactions On Graphics*, vol. 38, no. 5, pp. 1–12, 2019.

[134] Y. Li and A. Gupta, "Beyond grids: Learning graph representations for visual recognition," in *International Conference on Neural Information Processing Systems (NeurIPS)*, 2018, pp. 9245–9255.

[135] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *International Conference on Learning Representations (ICLR)*, 2017.

[136] B. Li, X. Li, Z. Zhang, and F. Wu, "Spatio-temporal graph routing for skeleton-based action recognition," in *AAAI Conference on Artificial Intelligence (AAAI)*, vol. 33, no. 01, 2019, pp. 8561–8568.

[137] M. Defferrard, X. Bresson, and P. Vandergheynst, "Convolutional neural networks on graphs with fast localized spectral filtering," in *International Conference on Neural Information Processing Systems (NeurIPS)*, 2016, pp. 3837–3845.

[138] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, "Neural message passing for quantum chemistry," in *International Conference on Machine Learning (ICML)*. PMLR, 2017, pp. 1263–1272.

[139] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *International Conference on Learning Representations*, 2018.

[140] Y. Zhang, S. Pal, M. Coates, and D. Ustebay, "Bayesian graph convolutional neural networks for semi-supervised classification," in *AAAI Conference on Artificial Intelligence (AAAI)*, vol. 33, no. 01, 2019, pp. 5829–5836.

[141] T. Kanade, J. F. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," in *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*. IEEE, 2000, pp. 46–53.

[142] M. Valstar and M. Pantic, "Induced disgust, happiness and surprise: an addition to the mmi facial expression database," in *International Workshop on EMOTION: Corpora for Research on Emotion and Affect*. Paris, France, 2010, p. 65.

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

[144] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee *et al.*, "Challenges in representation learning: A report on three machine learning contests," *Neural Networks*, vol. 64, pp. 59–63, 2015.

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

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

[147] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn, "Disfa: A spontaneous facial action intensity database," *IEEE Transactions on Affective Computing*, vol. 4, no. 2, pp. 151–160, 2013.

[148] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato, "A 3d facial expression database for facial behavior research," in *International Conference on Automatic Face and Gesture Recognition (FG)*. IEEE, 2006, pp. 211–216.

[149] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, and P. Liu, "A high-resolution spontaneous 3d dynamic facial expression database," in *IEEE International Conference and workshops on Automatic Face and Gesture Recognition (FG)*. IEEE, 2013, pp. 1–6.

[150] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, and J. M. Girard, "Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database," *Image and Vision Computing*, vol. 32, no. 10, pp. 692–706, 2014.

[151] G. Zhao and X. Li, "Automatic micro-expression analysis: open challenges," *Frontiers in Psychology*, vol. 10, p. 1833, 2019.

[152] X. Li, T. Pfister, X. Huang, G. Zhao, and M. Pietikäinen, "A spontaneous micro-expression database: Inducement, collection and baseline," in *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. IEEE, 2013, pp. 1–6.

[153] W.-J. Yan, X. Li, S.-J. Wang, G. Zhao, Y.-J. Liu, Y.-H. Chen, and X. Fu, "Casme ii: An improved spontaneous micro-expression database and the baseline evaluation," *PloS One*, vol. 9, no. 1, p. e86041, 2014.

[154] A. K. Davison, C. Lansley, N. Costen, K. Tan, and M. H. Yap, "Samm: A spontaneous micro-facial movement dataset," *IEEE Transactions on Affective Computing*, vol. 9, no. 1, pp. 116–129, 2018.

[155] F. Qu, S.-J. Wang, W.-J. Yan, H. Li, S. Wu, and X. Fu, "Cas(me)^2 : A database for spontaneous macro-expression and micro-expression spotting and recognition," *IEEE Transactions on Affective Computing*, vol. 9, no. 4, pp. 424–436, 2018.

[156] S. Li and W. Deng, "Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition," *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 356–370, 2019.

[157] R. Kosti, J. M. Alvarez, A. Recasens, and A. Lapedriza, "Context based emotion recognition using emotic dataset," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 11, pp. 2755–2766, 2019.

[158] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18–31, 2019.

[159] C. Fabian Benitez-Quiroz, R. Srinivasan, and A. M. Martinez, "Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 5562–5570.

[160] D. Kollias and S. Zafeiriou, "Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface," in *The British Machine Vision Conference (BMVC)*, 2019.

[161] M. J. Lyons, J. Budynek, and S. Akamatsu, "Automatic classification of single facial images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 12, pp. 1357–1362, 1999.

[162] Y. Wang and L. Guan, "Recognizing human emotional state from audiovisual signals," *IEEE transactions on multimedia*, vol. 10, no. 5, pp. 936–946, 2008.

[163] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The enterface'05 audio-visual emotion database," in *International Conference on Data Engineering Workshops*. IEEE, 2006, pp. 8–8.

[164] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *PloS one*, vol. 13, no. 5, p. e0196391, 2018.

[165] M. F. Valstar, T. Almaev, J. M. Girard, G. McKeown, M. Mehu, L. Yin, M. Pantic, and J. F. Cohn, "Fera 2015-second facial expression recognition and analysis challenge," in *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, vol. 6. IEEE, 2015, pp. 1–8.

[166] R. Zhi, M. Liu, and D. Zhang, "A comprehensive survey on automatic facial action unit analysis," *The Visual Computer*, vol. 36, no. 5, pp. 1067–1093, 2020.

[167] E. A. Haggard and K. S. Isaacs, "Micromomentary facial expressions as indicators of ego mechanisms in psychotherapy," in *Methods of Research in Psychotherapy*. Springer, 1966, pp. 154–165.

[168] P. Ekman and W. V. Friesen, "Nonverbal leakage and clues to deception," *Psychiatry*, vol. 32, no. 1, pp. 88–106, 1969.

[169] W.-J. Yan, Q. Wu, Y.-J. Liu, S.-J. Wang, and X. Fu, "Casme database: a dataset of spontaneous micro-expressions collected from neutralized faces," in *IEEE International Conference and workshops on Automatic Face and Gesture Recognition (FG)*. IEEE, 2013, pp. 1–7.

[170] P. Lucey, J. F. Cohn, K. M. Prkachin, P. E. Solomon, and I. Matthews, "Painful data: The unbc-mcmaster shoulder pain expression archive database," in *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*. IEEE, 2011, pp. 57–64.

[171] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang, "Large-

scale visual sentiment ontology and detectors using adjective noun pairs," in *Proceedings of the ACM international conference on Multimedia (ACM MM)*, 2013, pp. 223–232.

[172] J. Yang, D. She, and M. Sun, "Joint image emotion classification and distribution learning via deep convolutional neural network." in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2017, pp. 3266–3272.

[173] J. A. M. Correa, M. K. Abadi, N. Sebe, and I. Patras, "Amigos: A dataset for affect, personality and mood research on individuals and groups," *IEEE Transactions on Affective Computing*, 2018.

[174] R. Subramanian, J. Wache, M. K. Abadi, R. L. Vieriu, S. Winkler, and N. Sebe, "Ascertain: Emotion and personality recognition using commercial sensors," *IEEE Transactions on Affective Computing*, vol. 9, no. 2, pp. 147–160, 2016.

[175] M. Behzad, N. Vo, X. Li, and G. Zhao, "Landmarks-assisted collaborative deep framework for automatic 4d facial expression recognition," in *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*. IEEE, 2020, pp. 1–5.

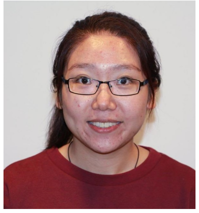
[176] X. Huang, A. Dhall, R. Goecke, M. Pietikäinen, and G. Zhao, "Multimodal framework for analyzing the affect of a group of people," *IEEE Transactions on Multimedia*, vol. 20, no. 10, pp. 2706–2721, 2018.

[177] R. e. Kaliouby and P. Robinson, "Real-time inference of complex mental states from facial expressions and head gestures," in *Real-time vision for human-computer interaction*. Springer, 2005, pp. 181–200.

[178] X. Liu, H. Shi, H. Chen, Z. Yu, X. Li, and G. Zhao, "imigue: An identity-free video dataset for micro-gesture understanding and emotion analysis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 10 631–10 642.

Yang Liu reveived his Ph.D. degree in computer sciench and technology from the South China University of Technology, in 2021. He is currently a Post-doctoral researcher at the Center for Machine Vision and Signal Analysis, University of Oulu, Finland. His current research interests include facial expression recognition, affective computing, and deep learning.

Xingming Zhang is currently a Professor with the School of Computer Science and Engineering, South China University of Technology, Guangzhou, China. He is a member of the Standing Committee of the Education Specialized Committee of China Computer Federation and the Standing director of the University Computer Education Research Association of China. His research focuses on video processing, big data, video surveillance, and face recognition.



Yante Li received her Ph.D. degree in computer science and engineering from the China University of Petroleum (East China), in 2017. She is currently pursuing a Ph.D. degree with the University of Oulu, Finland. Her current research interests include micro-expression analysis and facial action unit detection.

Jinzhao Zhou received his M.S. degree in computer technology from the South China University of Technology, in 2021. He is currently pursuing a Ph.D. at the University of Technology Sydney, Australia. His research interests include affective computing, reinforcement learning, and machine learning.

Xin Li received the M.S. degree in computer science and engineering from the South China University of Technology, in 2021. He is currently pursuing his Ph.D. at the Department of Electrical and Computer Engineering, Rutgers University, United States. His research interests include mobile computing and sensing.

Guoying Zhao is currently an Academy Professor with the University of Oulu, IEEE Fellow, IAPR Fellow, AAIA Fellow, and member of Finnish Academy of Sciences and Letters. She has authored or co-authored more than 280 papers in journals and conferences with 20000+ citations in Google Scholar and h-index 66. She has served as general and program chairs for several conferences, and is an associate editor for Pattern ecognition, IEEE TMM, IEEE TCSVT, and Image and Vision Computing Journals. Her current research interests include image and video descriptors, facial-expression recognition, human motion analysis, and biometrics.