sensors

Review

Understanding Naturalistic Facial Expressions with Deep Learning and Multimodal Large Language Models

Yifan Bian [1], Dennis Küster [2], Hui Liu [2] and Eva G. Krumhuber [1,*]

[1] Department of Experimental Psychology, University College London, London WC1H 0AP, UK; yifan.bian.23@ucl.ac.uk
[2] Department of Mathematics and Computer Science, University of Bremen, 28359 Bremen, Germany; kuester@uni-bremen.de (D.K.); hui.liu@uni-bremen.de (H.L.)
* Correspondence: e.krumhuber@ucl.ac.uk

Abstract: This paper provides a comprehensive overview of affective computing systems for facial expression recognition (FER) research in naturalistic contexts. The first section presents an updated account of user-friendly FER toolboxes incorporating state-of-the-art deep learning models and elaborates on their neural architectures, datasets, and performances across domains. These sophisticated FER toolboxes can robustly address a variety of challenges encountered in the wild such as variations in illumination and head pose, which may otherwise impact recognition accuracy. The second section of this paper discusses multimodal large language models (MLLMs) and their potential applications in affective science. MLLMs exhibit human-level capabilities for FER and enable the quantification of various contextual variables to provide context-aware emotion inferences. These advancements have the potential to revolutionize current methodological approaches for studying the contextual influences on emotions, leading to the development of contextualized emotion models.

Keywords: automatic facial expression recognition; naturalistic context; deep learning; multimodal large language model

check for updates

Citation: Bian, Y.; Küster, D.; Liu, H.; Krumhuber, E.G. Understanding Naturalistic Facial Expressions with Deep Learning and Multimodal Large Language Models. *Sensors* **2024**, *24*, 126. https://doi.org/10.3390/s24010126

Received: 13 November 2023
Revised: 30 November 2023
Accepted: 21 December 2023
Published: 26 December 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Recent advances in computer vision have greatly facilitated the development of sophisticated affective computing systems for facial expression recognition (FER) research [1]. Researchers across domains have applied various computational techniques to analyze diverse and complex mental states, including emotions [2], pain [3], physiological correlates [4], personality traits [5], and clinical disorders [6]. Nevertheless, our understanding of facial expressions remains mostly limited to inferences drawn from laboratory studies. Facial expressions produced in controlled laboratory settings may suffer from a lack of ecological validity and fail to represent the full spectrum of facial behaviors observed in real-life scenarios [7,8]. There is a growing emphasis on investigating naturalistic facial behaviors coupled with advanced computational techniques to spark theoretical advancements in affective science [9].

Naturalistic facial expressions can be observed in connection with a wide range of psychologically significant contexts encountered in everyday situations. Naturalistic facial expressions may more accurately reflect the complex and dynamic nature of emotional experiences in the real world than expressions elicited by experimental manipulations that are often artificial or short-lived (e.g., receiving an electrical shock). One approach toward studying more ecologically valid facial expressions relies on materials sourced from third-party media such as reality shows, vlogs, movies, and documentaries. Such sources often comprise millions of facial expressions accompanied by perceptually rich contexts, which are being made available in several newly developed datasets for FER research [10]. For surveys of existing naturalistic facial expression databases, readers are referred to [11–13].

Examining naturalistic expressions presents vast conceptual and empirical opportunities to yield further insights into how emotions emerge and reconfigure in naturalistic situations [14], discover new emotion categories that are rarely observed in laboratory environments [15], and document a comprehensive taxonomy of facial behaviors [16].

However, there are two main challenges in the studies of naturalistic facial expressions. The first challenge relates to the difficulties in detecting and tracking facial behaviors in unconstrained environments. Naturalistic expressions are often captured in uncontrolled settings with unexpected variations in head orientation, illumination, complex background, and facial occlusions, which may result in errors in detecting facial behaviors. More robust affective computing models are required to effectively analyze facial behaviors captured in wild conditions, as traditional models often exhibit drastic drops in performance in uncontrolled environments [17]. The second challenge concerns the interpretation of naturalistic facial expressions. Unlike laboratory studies that can validate the underlying emotional experiences of expressions through self-report measurements, it is not feasible to collect retrospective data to validate naturalistic expressions sourced from the Internet. Hence, accurately inferring the emotion states of naturalistic expressions can be more challenging than working with laboratory data. Nevertheless, this issue can be addressed through multimodal annotation and a comprehensive analysis of situational contexts to specify the underlying psychological states [18].

This review is divided into two sections to discuss the applications of deep learning-based FER toolboxes and multimodal large language models (MLLMs) for tackling these challenges. The first section evaluates several newly developed and easy-to-use FER toolboxes for facial expression analysis in unconstrained environments. To support researchers in making informed decisions regarding the selection of appropriate toolboxes, we critically review the performance of five FER toolboxes, namely OpenFace 2.0 [19], Affdex 2.0 [20], Py-Feat [21], LibreFace [22], and PyAFAR [23], along with their underlying neural architectures and databases used for model training. The second section discusses the potential utilization of MLLMs for analyzing and interpreting naturalistic expressions in association with contextual cues. Naturalistic expressions can be rendered meaningful by referencing the specific contexts in which they occur and interact [18]. MLLMs such as GPT-4V [24] and LLaVA [25] exhibit promising capabilities for quantifying contextual variables, which can serve as contextualized FER models for robust, explainable emotion inferences of naturalistic expressions.

2. Analyzing Naturalistic Facial Expressions with Deep Learning

The general processes of facial expression analysis consist of three components: face detection, feature extraction, and the prediction of facial action units (AUs) [26] and/or emotions. In particular, feature extraction has been regarded as the most crucial component in the FER process, which can be categorized into handcrafted and learned features [15]. Handcrafted features can be extracted with methods such as histograms of oriented gradients (HOGs) and local binary patterns (LBPs) for capturing facial textures and shapes. FER models trained based on handcrafted features using shallow learning approaches (e.g., support vector machines (SVMs)) achieve good performance in facial expressions produced in laboratory settings but have recently been outperformed by models trained on learned features [27]. Learned features are directly extracted from raw data through deep neural networks (DNNs) comprising multiple layers that hierarchically learn more complex, representative spatial–temporal features than the previous layers. Accumulating research evidence suggests that DNN models consistently surpass most of the shallow learning models based on handcrafted features by a large margin [28]. Compared to shallow learning models, cross-domain experiments reveal that DNN models such as convolutional neural networks (CNNs) and vision transformers (ViTs) achieve superior generalizability and accuracy for emotion recognition [29–32] and AU detection [33–35] on unseen datasets with different demographics, camera views, and emotion-eliciting contexts. More importantly, deep FER models are fairly robust to variations in brightness, head poses, and occlusions [27], whereas shallow learning models trained on handcrafted features (e.g., LBPs) may be substantially impacted by variations in luminance [21] and head rotation [36].

Despite their impressive performance, accessing state-of-the-art (SOTA) deep learning models for FER research has been hindered by several obstacles. SOTA models are rarely released for public use in repositories such as GitHub or are rarely open source for end users to fine-tune the models with their own datasets for new tasks. Open source models often lack graphical user interfaces or documentation for easy implementation for users who may lack specific programming knowledge. These SOTA models might be too heavy to run for real-time analysis, restricting their potential application in practical settings. Additionally, early FER models have been primarily trained on datasets captured in controlled recording conditions with invariant illumination and fixed camera position [37]. These models may learn features that rarely align with real-world situations, which limits their effectiveness in analyzing naturalistic expressions. To bridge this gap between cutting-edge FER techniques and their implementation, we discuss a selection of the most prominent, publicly accessible, lightweight, and user-friendly toolboxes that incorporate SOTA models suitable for analyzing facial expressions in the wild.

In the remainder of this section, we will first provide an overview of two FER toolboxes that incorporate both shallow and deep learning models for facial behavior analysis, namely OpenFace 2.0 and Py-Feat. Next, we will discuss three FER toolboxes that are primarily trained on deep learning methods: Affdex 2.0, LibreFace, and PyAFAR. Table 1 summarizes the main characteristics and access information of each FER toolbox.

Table 1. FER toolbox comparison on functionality, neural architecture, and type of dataset used for training emotion or AU models. The deep learning models are shown in bold. The datasets are categorized as W (wild), S (spontaneous), or P (posed), representing facial expressions sourced from the Internet or nonlaboratory environments, induced by experimental procedures, or deliberately mimicked by actors in laboratory settings, respectively. * Please consult the website for complete documentation of all face detection models incorporated in Py-Feat. ** The emotion model of Affdex 2.0 is not specified as it is based on the activation of AUs. *** The dataset used to train AU models in Affdex 2.0 is considered spontaneous despite being captured in nonlaboratory settings.

	OpenFace 2.0	**Py-Feat**	**Affdex 2.0**	**LibreFace**	**PyAFAR**
Face Detection	**CE-CLM**	**MTCNN, RetinaFace** *	**R-CNNs**	**MediaPipe**	**MediaPipe, Facenet**
Emotion Recognition		**ResMasNet**, SVM	**	**ViT, ResNet-18**	
Action Unit	SVM, SVR	SVM, XGB	**CNNs**	**ViT, ResNet-18**	**ResNet-50**
Datasets	P, S	P, S, W	S ***	S, W	S
Open Source/Free	Yes	Yes		Yes	Yes
Graphical User Interface	Yes		Yes	Yes	Yes
Website	github.com/TadasBaltrusaitis/OpenFace (accessed on 29 November 2023)	py-feat.org (accessed on 29 November 2023)	www.affectiva.com (accessed on 29 November 2023)	github.com/ihp-lab/LibreFace (accessed on 29 November 2023)	affectanalysisgroup.github.io/PyAFAR2023 (accessed on 29 November 2023)

2.1. *FER Toolboxes Based on Mixed Learning Models*

OpenFace 2.0 [19]: OpenFace 2.0 is a representative tool of the SOTA shallow learning models based on handcrafted features (i.e., HOGs) for AU recognition. OpenFace 2.0 is capable of a variety of facial analysis tasks, including facial landmark detection, gaze and pose estimation, and AU detection. OpenFace 2.0 uses a deep convolutional expert-constrained local model (CE-CLM) for facial landmark detection and tracking, which is trained on wild datasets with nonfrontal faces and varying illumination. It can better detect profiles or severely occluded faces than its previous version [37]. For AU detection, OpenFace 2.0 relies on shallow learning algorithms, including linear SVM for binary detection (presence or absence) and support vector regression (SVR) for intensity probability estimation. The AU models are trained in seven laboratory databases with AU annotations, which contain posed (Bosphorus [38]; FERA 2011 [39]), spontaneous (CK+, also containing posed expressions [40]; UNBC-McMaster [3]; DISFA [41]; and BP4D/FERA 2015 [42]), and conversational (SEMAINE [43]) expressions. Research by [36] demonstrated that OpenFace 2.0 outperformed several commercial FER toolboxes such as FaceReader 7.0 [44] for AU detection on datasets that contain posed (DISFA+ [45]), conversational (GFT [46]), and wild (Aff-wild2 [47]) expressions. However, OpenFace 2.0 was found to have inferior generalizability ability in AU detection on unseen datasets compared to deep FER models [48]. This is likely due to the limited discriminative values of handcrafted features or the inability of shallow learning algorithms to capture intricate, nonlinear patterns of facial behaviors.

Py-Feat [21]: Py-Feat provides various pretrained models based on both handcrafted and learned features, allowing users to flexibly decide which combinations of models to use according to specific task requirements. Py-Feat includes several face detection models such as multitask convolutional neural networks (MTCNNs [49]) and RetinaFace [50], which demonstrate robustness on partially obscured or nonfrontal faces. For AU detection, Py-Feat uses popular shallow learning methods for computing binary results with SVM and continuous results with optimized gradient boosting (XGB [51]), which are pretrained on both laboratory (BP4D, DISFA, CK+, and UNBC-McMaster) and wild (Aff-wild2 [47]) datasets. Its AU models have been shown to be slightly less accurate than the reported results of OpenFace 2.0 benchmarking on the DISFA+ dataset. However, Py-Feat yielded more consistent and reliable AU estimation for faces with varying head orientations, whereas the performance of OpenFace 2.0 drops dramatically when head angles are larger than 45 degrees [36]. The superior performance of Py-Feat on AU detection for occluded faces might be attributed to the inclusion of the wild dataset [47], which contains nonstandardized facial images and videos for pretraining models. Py-Feat provides direct emotion inferences for six basic emotions with emotion models trained on datasets with spontaneous (CK+ [40]), posed (JAFFE [52]), and wild (ExpW [53]) expressions. Since not all facial regions contribute equally to emotion perception, and some facial regions may be occluded in unconstrained conditions, Py-Feat exploits the residual masking network (ResMasNet [54]), a deep learning method that utilizes attention mechanisms to adaptively weight and select the most emotionally salient regions of the face while ignoring irrelevant features (e.g., facial occlusion). The ResMasNet outperforms another shallow learning model (i.e., SVM) adopted in Py-Feat and a commercial FER toolbox (FACET [55]) on a wild dataset (AffectNet [56]). Py-Feat provides numerous functional packages for data preprocessing, statistical analyses (e.g., time-series correlation and regressions), and visualization, which facilitate data exploration. Py-Feat is written in the Python programming language. Although a graphical user interface is currently not available in Py-Feat, it is relatively easy to use following step-by-step tutorials.

2.2. Deep FER Toolboxes

Affdex 2.0 [20]: Affdex 2.0 is a commercial software program designed to analyze facial behaviors in the wild. For face detection, it exploits region-based convolutional neural networks (R-CNNs [57]), which perform better on challenging conditions (e.g., variations in illumination, hand occlusions, etc.) compared to Affdex 1.0 [58]. For AU detection, Affdex 2.0 utilizes large samples of a private dataset collected using a web-based approach [59] to train CNN models, which showed enhanced accuracy and less biased results on ethnic minorities (e.g., African and South Asian) than its previous version. On benchmark evaluation for AU detection, Affdex 2.0 outperforms numerous SOTA methods on the DISFA dataset, which contains spontaneous expressions induced by emotion-evoking videos. Although Affdex 2.0 performs slightly worse than OpenFace 2.0, it is important to note that the AU models of Affdex 2.0 have not been trained on the DISFA, which may bias the result. Unlike other FER toolboxes that directly predict emotions with separate models specifically trained on emotional expression datasets, Affdex 2.0 estimates seven basic emotions [60] based on the reverse inference of the activation of AUs with assigned weights postulated in EMFACS [61]. Negative weights are assigned for emotion prediction when opposite AUs occur to reduce a false-positive rate. For instance, the presence of a “lip corner lowerer” (AU15) may decrease the likelihood of predicting “happiness” based on the activation of a “lip corner raiser” (AU12). Affdex 2.0 outperforms several SOTA emotion models on the Aff-wild2 dataset [47]. In addition to basic emotions, Affdex 2.0 can predict other affective states such as confusion, sentiment, and engagement based on predefined rules for AU activation. Although it is a commercial software program, Affdex 2.0 has been trained on a large sample of wild dataset with 11 million annotated images of people of all ages, genders, and ethnicities, which might be otherwise difficult to collect and annotate without financial support. By contrast, many open source systems are trained on a comparatively limited number of publicly accessible laboratory datasets with a small number of participants and a narrow range of demographic diversity. They are thus at a disadvantage with respect to the development of robust, unbiased models. However, since Affdex 2.0 is not open source, it does not allow users to further fine-tune the models for downstream tasks.

LibreFace [22]: LibreFace is a newly developed toolkit that incorporates several SOTA deep networks for facial AU and emotion expression analysis. LibreFace utilizes MediaPipe [62] for precise face detection with 468-point 3D landmark registration and normalization with geometric transformation. Model performances including feature extraction, robustness, and generalizability are enhanced by several pretraining processes. Specifically, a ViT-base model and a ResNet-18 model [63] were pretrained on several large wild datasets sourced from the Internet, including the training set of AffectNet [56], EmotioNet [64], and FFHQ [65], which consist of millions of facial images with a wide range of variations in demographic features, illumination, and head orientation. After the pretraining phase, the models were fine-tuned on the DISFA dataset [41] for AU intensity estimation and the BP4D-Spontaneous datasets [66] for AU detection. LibreFace leverages the Swin Transformer model [67] to capture the spatial correlations and interactions between different facial features from a global perspective, which could achieve better performance than traditional CNNs that focus on local patterns and features within facial subregions [68,69]. Moreover, LibreFace bypasses the need for large, labeled datasets by exploiting the masked autoencoder (MAE) method that allows for learning representative facial features through the process of image and video reconstruction [70]. LibreFace utilizes feature-wise knowledge distillation to reduce computational costs, thereby boosting the inference efficiency for real-time facial expression analysis. LibreFace outperforms OpenFace 2.0 [19] and other SOTA deep learning models [35,71] on the DISFA and BP4D datasets for predicting the activation of AU. For emotion prediction, LibreFace achieves competitive results comparable or superior to more complicated and heavier SOTA models [72] on two wild datasets, AffectNet and RaF-DB [73]. LibreFace is currently open source in Python, and for Windows users, an easy-to-use graphical user interface is available using the OpenSense platform [74].

PyAFAR [23]: PyAFAR is developed for facial AU detection and intensity estimation in addition to head orientation and facial landmark detection. For face detection, it uses MediaPipe and Facenet [75], which can identify and track individuals even if they exit and re-enter the field. PyAFAR adopts two separate DNN models based on ResNet50 [63], with increased depth of neural architecture to perform more effective convolution operations for complex facial feature representation. The models have been pretrained on the ImageNet [76] dataset for detecting 12 Aus in adults and 9 Aus in infants. The adult model is fine-tuned on BP4D+ [77], an expansion of the BP4D-Spontaneous dataset containing spontaneous facial expressions induced by both active (e.g., singing) and passive (e.g., watching emotionally loaded videos) tasks. The infant model is trained on the MIAMI [78] and CLOCK [79] databases, which capture infants’ responses induced by experimental procedures, such as the removal of attractive toys and still face paradigms [80]. Both adult and infant models achieve accurate results on within-domain validation [23]. PyAFAR shows superior cross-database performance on the GFT dataset [46] compared to OpenFace

2.0 and the previous version of AFAR [81]. An executable interface and a step-by-step visual instruction guide are available for the easy implementation of the toolbox.

These FER toolboxes empower researchers to effectively address the challenge of unexpected variation in naturalistic behaviors acquired from unconstrained environments. However, it is important to note that some FER toolboxes (e.g., Affdex 2.0, LibreFace, and Py-AFAR) have only been recently created and validated by their developers. Further empirical research conducted by independent researchers is required to compare and validate the cross-domain performance of these toolboxes [1].

What is still missing in the studies of naturalistic facial expressions is the lack of a comprehensive analysis of contextual information critical for a naturalistic understanding of emotions [82]. Prior works have focused extensively on the analysis of facial features using FER toolboxes for emotion inferences, while contextual variables have been largely ignored [83]. Emotion inferences made solely based on decontextualized faces are ecologically invalid and meaningless. For instance, a smile can be reliably recognized as expressing “happiness” by FER toolboxes, but it is difficult to evaluate the meaning of the emotion without referencing the emotional stimuli or surrounding contexts (e.g., smile as reflecting the anticipation of a music festival [84,85]). Elucidating the interaction between facial behaviors and concurrent contexts is an important research question for affective science.

A comprehensive analysis of contextual elements can provide important cues for an accurate assessment of the underlying emotional experience associated with naturalistic facial expressions [18]. For example, naturalistic facial behaviors (e.g., a smile) are often accompanied by contextual cues presented in various forms such as clothing (e.g., a gown), scenery (e.g., wedding venues), activities (e.g., marriage proposal), voices (e.g., “I love you”), body postures (e.g., holding hands), other faces, and so forth, which shape how faces are perceived (e.g., the enjoyment of interpersonal connection). Contextual variables can be measured and quantified by human annotators [86]. When facial expressions are presented with perceptually rich contextual information, human annotators show substantially greater agreement for labeling facial expressions than decontextualized faces [87]. This indicates that the current limitations of evaluating facial expressions with FER systems could be addressed by including contextual cues, as human perceivers can make more robust, reliable emotion inferences. However, annotating naturalistic expressions and their contexts can be more complicated and labor-intensive than laboratory datasets. Advanced multimodal annotation tools [88] may help provide multimodal annotation to evaluate facial expressions together with other nonverbal modalities and rich contextual information to provide accurate portrayals of the interaction between facial expressions and contexts. In the following section, we discuss the novel applications of MLLMs that could circumvent the need for extensively annotated datasets, fostering further advancement in naturalistic affective research.

3. Advancing Naturalistic Affective Science with Multimodal Large Language Models

Recent advancements in MLLMs have demonstrated remarkable versatility and capability in various domains and tasks. Although MLLMs are not specifically programmed for emotion recognition tasks, such capabilities emerge as the result of data scaling [89]. The main idea of using MLLMs for emotion recognition is to use powerful large language models as an intelligent brain to process and align textual, visual [25,90], and/or auditory [91,92] information to perform emotion inferences [93]. There is an increasing number of open source MLLMs, including LLaVA [25] and MiniGPT-4 [90], available on platforms such as HuggingFace. Many MLLMs also provide a user-friendly web interface that enables more flexible interactions with the user. This section provides an in-depth discussion on the emergent novel applications of MLLMs for context-aware emotion recognition, generalizable facial expression analysis, and adaptability to other related tasks such as the classification of nuanced emotion categories.

3.1. MLLMs as a Contextualized Emotion Model

MLLMs can serve as a contextualized model quantifying contextual variables to provide robust emotional reasoning for naturalistic expressions (see Figure 1 for example). Several studies have demonstrated the exceptional capabilities of MLLMs in identifying emotionally evoking context and comprehending how these contextual cues may influence the emotional state of a person [94]. To quantify contextual variability, MLLMs can be used to perform a wide range of visual reasoning tasks such as spatial relationship analysis and object recognition to understand the visual world with a simple prompt, i.e., "describe the image". This can provide a detailed description of the situational contexts, which can then be used to infer relevant emotional states and related antecedents or consequences. Such contextualized inferences are more in line with how humans naturally perceive emotions in real-life situations by synthesizing concurrent multisensory information from the face and contextual cues [95]. This potential application is exemplified in a recent technical report of GPT-4V in the sense that it can make accurate emotion inferences based on integrated contextual cues such as "protest crowd" and "presence of policies" for inferring "injustice" and "anger" [24]. Moreover, research by Etesam et al. [94] has revealed that MLLMs outperformed vision language models in context-aware emotion recognition tasks using the EMOTIC dataset [96], which contains contextually rich images annotated with 26 emotion categories. While vision language models like CLIP [97] are effective at detecting immediately visible characteristics such as facial expressions, body postures, and activities, they fail to reason about the underlying causal relationships of these contextual data for emotion inferences. In contrast, MLLMs like LLaVA [25] not only identify these visible characteristics but also integrate and capture the complex relationships among these contextual cues for emotion inferences. For instance, while CLIP may perceive "raising arms" as signifying "surprise" and "fear", MLLMs may reason that this body posture actually reflects "happiness" and "excitement" given the context of skiing. This empirical evidence supports the practical utility of MLLMs for affective research, which can further improve our understanding of how naturalistic emotions manifest in real-life scenarios by considering contextual variability.

Furthermore, the contextual reasoning generated from MLLMs can complement the results obtained from FER toolboxes to produce context-aware emotion inferences, which are more robust and insightful than simply analyzing the face alone. This is particularly important for accurately identifying and interpreting complex or vague expressions that convey mixed emotional signals such as sarcasm and Schadenfreude indicated by the incongruence between facial behaviors and contexts (e.g., a polite smile accompanied by sarcastic statements [98]). More importantly, future studies can utilize MLLMs and FER toolboxes for examining the relationship between different types of contexts and facial expressions. Past research by Cowen et al. [14] has utilized sophisticated DNNs to automatically classify facial expressions and contexts from over millions of videos sourced from the Internet. They found that sixteen types of facial expressions consistently occur in specific contexts, suggesting substantial contextual dependence on facial expressions. Nevertheless, in this study, the DNN models for context classification could only predict the topics of videos. For instance, a context like "the bride dressed in a gown dancing with the groom dressed in a suit in a banquet hall filled with tables and wine glasses" may be simplified into a label like "wedding". This may lead to the oversimplification of contextual variability that could significantly alter emotion perception. A small object shown in the wedding context, such as a gun, may completely shift the emotion perception from "joy" to "fear". Therefore, future studies are encouraged to utilize MLLMs to generate detailed context descriptions, as illustrated in Figure 1, to fully examine the complex relationships between facial expressions and contextual variability. In this way, researchers can not only leverage contextual information for inferring the emotional states of naturalistic expressions but also utilize facial information to guide the prediction of situational contexts (e.g., appeasement smile indicating submissive contexts [99]).

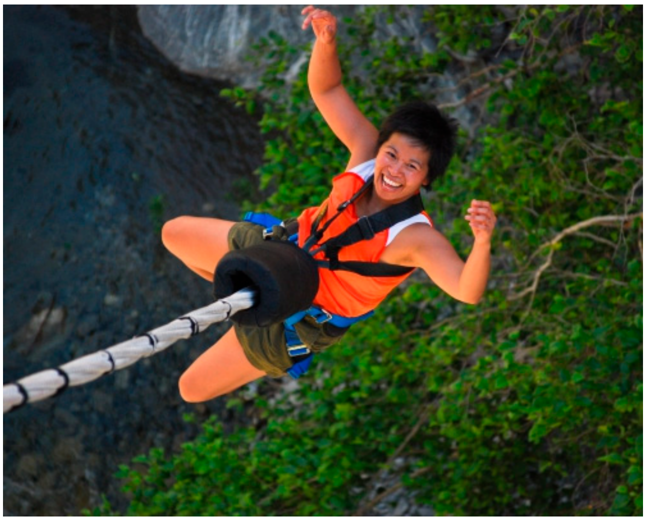

Figure 1. Contextualized emotion inferences generated from LLaVA-7B [25] (image from OASIS [100]). Temperature is set to zero for reproducible response; note that the model can identify the association between the emotion of "excitement" and the activity of "swing through the air", but it mistakenly describes the person as wearing a black helmet. Prompt: "Describe the image? What emotions can be inferred based on the context?". LLaVA-7B: "The image features a woman wearing an orange vest and a black helmet, hanging from a rope or wire. She appears to be enjoying herself as she swings through the air, possibly on a rope swing or a zip line. The woman's smile and the fact that she is wearing a helmet suggest that she is having a fun and adventurous time. The scene conveys a sense of excitement and joy, as the woman embraces the thrill of the activity."

3.2. MLLMs as Generalizable FER Toolboxes

Several technical reports have documented the impressive human-level capabilities of MLLMs for analyzing facial expressions [24]. Researchers can conduct FER tasks using MLLMs with a simple query "What is the facial expression?" MLLMs can detect emotions from facial expressions and can generalize well to unseen data because MLLMs are developed to address a diverse range of tasks. MLLMs can also analyze a wide spectrum of emotion categories based on natural language processing, which allows researchers to capture more nuanced emotions (e.g., awe, shame, emotional suppression, etc. [24]). By contrast, prior emotion models have typically been programmed to predict a limited set of basic emotions, which restrict their potential applications in more generalized domains (e.g., monitoring student engagement in the classroom or pain in the ICU). MLLMs can also provide multiple emotion labels for comprehensive descriptions of complex emotional states. Such responses are more aligned with the human perception of facial expressions that have been conceptualized as a manifold representation of varying degrees of emotions (e.g., a face may be perceived as 51% anger and 49% disgust [101]). In comparison, most existing emotion models rely on majority voting to determine the single most representative label [102]. Although such approaches may enhance the overall reliability of emotion prediction, they may exclude other relevant but less dominant emotions and fail to detect compound emotions [103]. Moreover, unlike prior works that mainly focus on emotion prediction, MLLMs can take a further step to provide reasonable explanations of the facial features being used for emotion inferences (see Figure 2, for example). For instance, after inferring a facial image as showing "fear" or "anxiety", GPT-4V can offer detailed descriptions of the visual cues to interpret these emotional states (e.g., "wide-open eyes" revealing "a sense of shock or surprise" and "hands covering the mouth" indicating "suppressing scream" [24]). It can also modify its emotional inferences based on hand gestures, which are often ignored or treated as facial occlusion by existing FER toolboxes. Furthermore, Lian et al. [93] examined the performance of several MLLMs, including Video-LLaMA [91], PandaGPT [92], and Valley [104], for generating explainable emotional inferences from a subset of the MER2023 dataset [105] comprising dynamic facial expressions sourced from the Internet. Specifically, they evaluated the abilities of MLLMs to identify emotion-related cues (e.g., "lifted eyebrows", "smiling face", etc.) and predict emotional states (e.g., "happiness") based on the identified cues. In addition, the plausibility of the emotion reasoning process was validated by human observers. Valley consistently achieved the best performance, with 72.1% accuracy in identifying emotion-related cues and 57.8% accuracy in predicting emotional states. It should be noted that the paper did not report the base rates for emotion prediction, and the authors used ChatGPT to relabel more subtle emotions, going beyond the original labels introduced in MER2023 as a baseline. The emotion reasoning process demonstrated by Valley also aligned most closely with human perception, receiving a plausibility score of 65.0%. This study provides empirical evidence demonstrating the capabilities of MLLMs in identifying explainable cues for emotion inferences, which supports their potential utility in FER research. However, further efforts are required to improve the model's performance. With the enhanced interpretability of emotion recognition processes, FER researchers can be more confident in asserting the psychological states of naturalistic expressions, thereby achieving greater reliability and specificity.

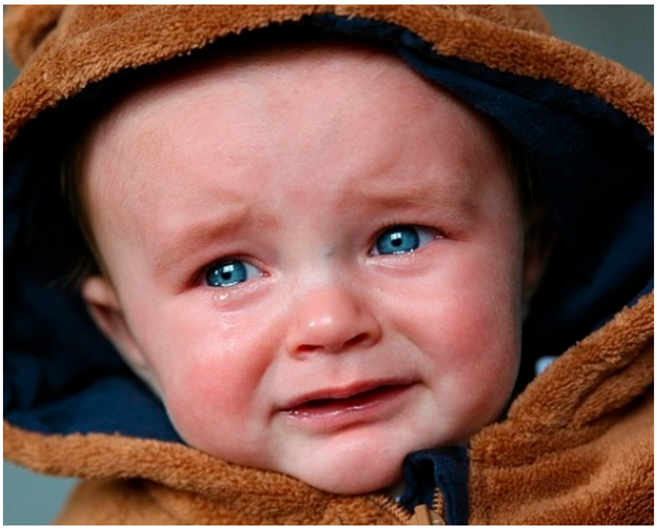

Figure 2. Explainable facial expression recognition generated from LLaVA-7B [25] (image from OASIS [100]). Temperature is set to zero for reproducible response. Prompt: "What emotions are shown on the baby's face? What specific facial cues are used to infer the emotions?" LLaVA-7B: "The baby's face shows a mix of emotions, including sadness, fear, and possibly discomfort. The baby's facial cues include a frowning expression, tears in the eyes, and a slightly puckered mouth. These cues suggest that the baby is experiencing negative emotions or discomfort, possibly due to the situation or the environment around them".

3.3. Adaptability of MLLMs for Different Emotion Recognition Tasks

MLLMs have strong adaptation capabilities for more challenging emotion recognition tasks through few-shot in-context learning (ICL). Few-shot ICL refers to the ability to quickly adapt to novel tasks given a short instruction and few examples without fine-tuning the models and abundant labeled data. After few-shot ICL, MLLMs such as Flamingo [106] can outperform some of the SOTA contrastive models [97] despite using only around 30 task-specific examples, which are about 1000 times fewer data inputs required for fine-tuning the models. There are numerous potential applications of few-shot ICL for naturalistic affective research such as the classification of nuanced emotion categories. For instance, several large-scale datasets contain a broad spectrum of facial behaviors captured in the wild, which present new opportunities for investigating the complexity and variability of emotional experiences and their underlying psychosocial processes in real-life scenarios. However, many of these datasets are only annotated with a few emotion classes, which limits their potential utilization in affective research. With the advanced ICL functionality, researchers can further exploit these datasets by applying MLLMs to identify new emotion categories for a more fine-grained analysis of human emotions [93]. Specifically, FER researchers can provide a few demonstration examples of facial expressions categorized by specific emotions in the format of image–test or video–text pairs to extrapolate to new emotion recognition tasks with a visual query such as "identify the images that display the same facial expression illustrated in the above examples". This approach may enable FER researchers to evaluate more specific and contextualized expressions in the wild that may not be accounted for by the limited set of expressions detected by existing FER toolboxes. This could also pave the way for examining the assumptions of various emotion theories, e.g., appraisal theories [107], the theory of constructed emotions [82], and the behavioral ecology view [108]. Coupled with ICL techniques, researchers can also examine if facial expressions produced in laboratory settings generalize to naturalistic contexts, thereby gaining deeper insights into the ecological validity of these facial displays and their potential implications in real-world situations. Unfortunately, to the best of our knowledge, in-context learning tailored specifically for FER tasks has not received much empirical attention. Future studies should leverage the intriguing emergent ability to further advance the field of naturalistic affective science.

3.4. Limitations of MLLMs

While MLLMs have the potential to revolutionize the domain of FER research, it is essential to acknowledge their limitations and work toward addressing them. For facial behavior analysis, it is unclear whether MLLMs can provide FACS-like inferences [26] that are precise enough for accurate facial behavior analysis. It is important to test if MLLMs can capture the variations in facial parameters [109] and distinguish the subtle differences between various types of facial behaviors [110]. For instance, smiles can be characterized by different facial configurations such as Duchenne smiles (e.g., AU6 + 12), "selfie smiles" (e.g., AU13), or "miserable" smiles (e.g., AU12 + 14 or 12 + 15), which are associated with distinct psychological states [111]. Further empirical investigation is required to systematically examine the similarity between the facial behaviors described by MLLMs and the facial AUs detected by FER toolboxes, as well as explore methods to fine-tune the models to achieve comparable or superior results. Before such empirical testing is carried out, it is recommended to incorporate both MLLMs and FER toolboxes for fine-grained facial behavior analyses. In addition, it remains unclear to what extent the contextual perception of MLLMs aligns with human perception in terms of emotion inferences [93,112]. It is possible that contextualized emotion inferences may be biased by random noise in the context. Variability also exists in the susceptibility to contextual influences among individuals [113], as evidenced in the case of depressed individuals who often exhibit facial responses insensitive to contextual cues [114]. Therefore, it is critical to formulate theories and models to elucidate the mechanisms underlying the integration of contextual information during the process of facial expression recognition [115]. Finally, it is crucial to acknowledge that despite an extensive search for empirical evidence across various domains, some arguments concerning MLLMs remain hypothetical, particularly regarding their adaptability, as discussed in Section 3.3. Although MLLMs have demonstrated promising capabilities in addressing numerous important research questions in affective science, as illustrated in Sections 3.1 and 3.2, they are still in the early phases of development, awaiting further improvement. Further empirical research is required to explore the versatility of MLLMs for emotion recognition tasks, thereby bridging the gap between their potential applications and practical implementation.

4. Conclusions

In this review, we have provided an overview of publicly accessible and user-friendly FER toolboxes for robust facial analyses under unconstrained conditions and an introduction to the potential applications of MLLMs to further advance the field of affective science. However, it is important to emphasize that most of these techniques have only been recently developed. Additional empirical investigations are needed to validate their practical utility across various domains and further improve their performance. Future studies are encouraged to utilize these sophisticated techniques to expand our knowledge of naturalistic facial expressions and develop contextualized emotion models to achieve a comprehensive understanding of emotional experiences in the real world [116].

Author Contributions: Conceptualization, Y.B. and E.G.K.; writing—original draft preparation, Y.B., D.K., H.L. and E.G.K.; writing—review and editing, Y.B. and E.G.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Dupré, D.; Krumhuber, E.G.; Küster, D.; McKeown, G.J. A performance comparison of eight commercially available automatic classifiers for facial affect recognition. *PLoS ONE* **2020**, *15*, e0231968. [CrossRef] [PubMed]
2. Krumhuber, E.G.; Küster, D.; Namba, S.; Skora, L. Human and machine validation of 14 databases of dynamic facial expressions. *Behav. Res. Methods* **2021**, *53*, 686–701. [CrossRef] [PubMed]
3. Lucey, P.; Cohn, J.F.; Prkachin, K.M.; Solomon, P.E.; Matthews, I. Painful data: The UNBC-McMaster shoulder pain expression archive database. In Proceedings of the 2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG), Santa Barbara, CA, USA, 21–23 March 2011; pp. 57–64.
4. Chang, C.Y.; Tsai, J.S.; Wang, C.J.; Chung, P.C. Emotion recognition with consideration of facial expression and physiological signals. In Proceedings of the 2009 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, Nashville, TN, USA, 30 March–2 April 2009; pp. 278–283.
5. Biel, J.I.; Teijeiro-Mosquera, L.; Gatica-Perez, D. Facetube: Predicting personality from facial expressions of emotion in online conversational video. In Proceedings of the 14th ACM International Conference on Multimodal Interaction 2012, Santa Monica, CA, USA, 22–26 October 2012; pp. 53–56.
6. Fisher, H.; Reiss, P.T.; Atias, D.; Malka, M.; Shahar, B.; Shamay-Tsoory, S.; Zilcha-Mano, S. Facing Emotions: Between- and Within-Sessions Changes in Facial Expression During Psychological Treatment for Depression. *Clin. Psychol. Sci.* **2023**, 21677026231195793. [CrossRef]
7. Küster, D.; Steinert, L.; Baker, M.; Bhardwaj, N.; Krumhuber, E.G. Teardrops on my face: Automatic weeping detection from nonverbal behavior. *IEEE Trans. Affect. Comput.* 2022, *in press*. [CrossRef]
8. Krumhuber, E.G.; Skora, L.I.; Hill, H.C.H.; Lander, K. The role of facial movements in emotion recognition. *Nat. Rev. Psychol.* **2023**, *2*, 283–296. [CrossRef]
9. Lin, C.; Bulls, L.S.; Tepfer, L.J.; Vyas, A.D.; Thornton, M.A. Advancing naturalistic affective science with deep learning. *Affect. Sci.* **2023**, *4*, 550–562. [CrossRef] [PubMed]
10. Ren, Z.; Ortega, J.; Wang, Y.; Chen, Z.; Whitney, D.; Guo, Y.; Yu, S.X. VEATIC: Video-based Emotion and Affect Tracking in Context Dataset. *arXiv* **2023**, arXiv:2309.06745.
11. Siddiqui, M.F.H.; Dhakal, P.; Yang, X.; Javaid, A.Y. A survey on databases for multimodal emotion recognition and an introduction to the VIRI (visible and InfraRed image) database. *Multimodal Technol. Interact.* **2022**, *6*, 47. [CrossRef]
12. Guerdelli, H.; Ferrari, C.; Barhoumi, W.; Ghazouani, H.; Berretti, S. Macro-and micro-expressions facial datasets: A survey. *Sensors* **2022**, *22*, 1524. [CrossRef]
13. Weber, R.; Soladié, C.; Séguier, R. A Survey on Databases for Facial Expression Analysis. In Proceedings of the 13th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2018), Madeira, Portugal, 27–29 January 2018; pp. 73–84.
14. Cowen, A.S.; Keltner, D.; Schroff, F.; Jou, B.; Adam, H.; Prasad, G. Sixteen facial expressions occur in similar contexts worldwide. *Nature* **2021**, *589*, 251–257. [CrossRef]
15. Zhu, Q.; Mao, Q.; Jia, H.; Noi, O.E.N.; Tu, J. Convolutional relation network for facial expression recognition in the wild with few-shot learning. *Expert Syst. Appl.* **2022**, *189*, 116046. [CrossRef]
16. Srinivasan, R.; Martinez, A.M. Cross-cultural and cultural-specific production and perception of facial expressions of emotion in the wild. *IEEE Trans. Affect. Comput.* **2018**, *12*, 707–721. [CrossRef]

17. Dhall, A.; Goecke, R.; Joshi, J.; Wagner, M.; Gedeon, T. Emotion recognition in the wild challenge 2013. In Proceedings of the 15th ACM on International Conference on Multimodal Interaction 2013, Sydney, Australia, 9–13 December 2013; pp. 509–516.
18. Barrett, L.F. Context reconsidered: Complex signal ensembles, relational meaning, and population thinking in psychological science. *Am. Psychol.* **2022**, *77*, 894. [CrossRef] [PubMed]
19. Baltrusaitis, T.; Zadeh, A.; Lim, Y.C.; Morency, L.P. Openface 2.0: Facial behavior analysis toolkit. In Proceedings of the 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), Xi'an, China, 15–19 May 2018; pp. 59–66.
20. Bishay, M.; Preston, K.; Strafuss, M.; Page, G.; Turcot, J.; Mavadati, M. Affdex 2.0: A real-time facial expression analysis toolkit. In Proceedings of the 2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG), Waikoloa Beach, HI, USA, 5–8 January 2023; pp. 1–8.
21. Cheong, J.H.; Jolly, E.; Xie, T.; Byrne, S.; Kenney, M.; Chang, L.J. Py-feat: Python facial expression analysis toolbox. In *Affective Science*; Springer: Berlin/Heidelberg, Germany, 2023; pp. 1–16.
22. Chang, D.; Yin, Y.; Li, Z.; Tran, M.; Soleymani, M. LibreFace: An Open-Source Toolkit for Deep Facial Expression Analysis. *arXiv* **2023**, arXiv:2308.10713.
23. Hinduja, S.; Ertugrul, I.O.; Cohn, J.F. PyAFAR: Python-Based Automated Facial Action Recognition for Use in Infants and Adults. 2023. Available online: https://www.jeffcohn.net/wp-content/uploads/2023/08/ACII_2023_paper_242-2.pdf (accessed on 29 November 2023).
24. Yang, Z.; Li, L.; Lin, K.; Wang, J.; Lin, C.C.; Liu, Z.; Wang, L. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv* **2023**, arXiv:2309.17421.
25. Liu, H.; Li, C.; Li, Y.; Lee, Y.J. Improved Baselines with Visual Instruction Tuning. *arXiv* **2023**, arXiv:2310.03744.
26. Ekman, P.; Friesen, W.V. Facial Action Coding System. Environmental Psychology & Nonverbal Behavior. 1978. Available online: https://www.paulekman.com/facial-action-coding-system/ (accessed on 29 November 2023).
27. Karnati, M.; Seal, A.; Bhattacharjee, D.; Yazidi, A.; Krejcar, O. Understanding deep learning techniques for recognition of human emotions using facial expressions: A comprehensive survey. *IEEE Trans. Instrum. Meas.* **2023**, *72*, 5006631.
28. Sajjad, M.; Ullah, F.U.M.; Ullah, M.; Christodoulou, G.; Cheikh, F.A.; Hijji, M.; Muhammad, K.; Rodrigues, J.J. A comprehensive survey on deep facial expression recognition: Challenges, applications, and future guidelines. *Alex. Eng. J.* **2023**, *68*, 817–840. [CrossRef]
29. Li, S.; Deng, W. A deeper look at facial expression dataset bias. *IEEE Trans. Affect. Comput.* **2020**, *13*, 881–893. [CrossRef]
30. Georgescu, M.I.; Ionescu, R.T.; Popescu, M. Local learning with deep and handcrafted features for facial expression recognition. *IEEE Access* **2019**, *7*, 64827–64836. [CrossRef]
31. Hasani, B.; Mahoor, M.H. Spatio-temporal facial expression recognition using convolutional neural networks and conditional random fields. In Proceedings of the 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), Washington, DC, USA, 30 May–3 June 2017; pp. 790–795.
32. Mollahosseini, A.; Chan, D.; Mahoor, M.H. Going deeper in facial expression recognition using deep neural networks. In Proceedings of the 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Placid, NY, USA, 7–10 March 2016; pp. 1–10.
33. Büdenbender, B.; Höfling, T.T.; Gerdes, A.B.; Alpers, G.W. Training machine learning algorithms for automatic facial coding: The role of emotional facial expressions' prototypicality. *PLoS ONE* **2023**, *18*, e0281309. [CrossRef]
34. Cohn, J.F.; Ertugrul, I.O.; Chu, W.S.; Girard, J.M.; Jeni, L.A.; Hammal, Z. Affective facial computing: Generalizability across domains. In *Multimodal Behavior Analysis in the Wild*; Academic Press: Cambridge, MA, USA, 2019; pp. 407–441.
35. Zhao, K.; Chu, W.S.; Zhang, H. Deep region and multi-label learning for facial action unit detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 3391–3399.
36. Namba, S.; Sato, W.; Osumi, M.; Shimokawa, K. Assessing automated facial action unit detection systems for analyzing cross-domain facial expression databases. *Sensors* **2021**, *21*, 4222. [CrossRef] [PubMed]
37. Baltrušaitis, T.; Robinson, P.; Morency, L.P. Openface: An open source facial behavior analysis toolkit. In Proceedings of the 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Placid, NY, USA, 10 March 2016; pp. 1–10.
38. Savran, A.; Alyüz, N.; Dibeklioğlu, H.; Çeliktutan, O.; Gökberk, B.; Sankur, B.; Akarun, L. Bosphorus database for 3D face analysis. In *Biometrics and Identity Management: First European Workshop, BIOID 2008, Roskilde, Denmark, 7–9 May 2008*; Revised Selected Papers 1; Springer: Berlin/Heidelberg, Germany, 2008; pp. 47–56.
39. Valstar, M.F.; Jiang, B.; Mehu, M.; Pantic, M.; Scherer, K. The first facial expression recognition and analysis challenge. In Proceedings of the 2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG), Santa Barbara, CA, USA, 21–23 March 2011; pp. 921–926.
40. Lucey, P.; Cohn, J.F.; Kanade, T.; Saragih, J.; Ambadar, Z.; Matthews, I. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, San Francisco, CA, USA, 13–18 June 2010; pp. 94–101.
41. Mavadati, S.M.; Mahoor, M.H.; Bartlett, K.; Trinh, P.; Cohn, J.F. Disfa: A spontaneous facial action intensity database. *IEEE Trans. Affect. Comput.* **2013**, *4*, 151–160. [CrossRef]
42. Valstar, M.F.; Almaev, T.; Girard, J.M.; McKeown, G.; Mehu, M.; Yin, L.; Pantic, M.; Cohn, J.F. Fera 2015-second facial expression recognition and analysis challenge. In Proceedings of the 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), Ljubljana, Slovenia, 4–8 May 2015; Volume 6, pp. 1–8.

43. McKeown, G.; Valstar, M.; Cowie, R.; Pantic, M.; Schroder, M. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Trans. Affect. Comput.* **2011**, *3*, 5–17. [CrossRef]
44. Skiendziel, T.; Rösch, A.G.; Schultheiss, O.C. Assessing the convergent validity between the automated emotion recognition software Noldus FaceReader 7 and Facial Action Coding System Scoring. *PLoS ONE* **2019**, *14*, e0223905. [CrossRef] [PubMed]
45. Mavadati, M.; Sanger, P.; Mahoor, M.H. Extended disfa dataset: Investigating posed and spontaneous facial expressions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Las Vegas, NV, USA, 27–30 June 2016; pp. 1–8.
46. Girard, J.M.; Chu, W.S.; Jeni, L.A.; Cohn, J.F. Sayette group formation task (gft) spontaneous facial expression database. In Proceedings of the 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), Washington, DC, USA, 30 May–3 June 2017; pp. 581–588.
47. Kollias, D.; Zafeiriou, S. Aff-wild2: Extending the aff-wild database for affect recognition. *arXiv* **2018**, arXiv:1811.07770.
48. Ertugrul, I.O.; Cohn, J.F.; Jeni, L.A.; Zhang, Z.; Yin, L.; Ji, Q. Crossing domains for au coding: Perspectives, approaches, and measures. *IEEE Trans. Biom. Behav. Identity Sci.* **2020**, *2*, 158–171. [CrossRef]
49. Zhang, N.; Luo, J.; Gao, W. Research on face detection technology based on MTCNN. In Proceedings of the 2020 International Conference on Computer Network, Electronic and Automation (ICCNEA), Xi'an, China, 25–27 September 2020; pp. 154–158.
50. Deng, J.; Guo, J.; Zhou, Y.; Yu, J.; Kotsia, I.; Zafeiriou, S. Retinaface: Single-stage dense face localisation in the wild. *arXiv* **2019**, arXiv:1905.00641.
51. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining 2016, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
52. Lyons, M.; Kamachi, M.; Gyoba, J. The Japanese Female Facial Expression (JAFFE) Dataset. 1998. Available online: https://zenodo.org/records/3451524 (accessed on 29 November 2023).
53. Zhang, Z.; Luo, P.; Loy, C.C.; Tang, X. From facial expression recognition to interpersonal relation prediction. *Int. J. Comput. Vis.* **2018**, *126*, 550–569. [CrossRef]
54. Pham, L.; Vu, T.H.; Tran, T.A. Facial expression recognition using residual masking network. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; pp. 4513–4519.
55. iMotions. Facial Expression Analysis: The Definitive Guide. 2016. Available online: https://imotions.com/facialexpression-guide-ebook/ (accessed on 29 November 2023).
56. Mollahosseini, A.; Hasani, B.; Mahoor, M.H. AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild. *IEEE Trans. Affect. Comput.* **2019**, *10*, 18–31. [CrossRef]
57. Ren, S.; He, K.; Girshick, R.B.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *39*, 1137–1149. [CrossRef]
58. McDuff, D.; Mahmoud, A.; Mavadati, M.; Amr, M.; Turcot, J.; Kaliouby, R.E. AFFDEX SDK: A cross-platform real-time multi-face expression recognition toolkit. In Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems 2016, New York, NY, USA, 7–12 May 2016; pp. 3723–3726.
59. McDuff, D.; Kaliouby, R.; Senechal, T.; Amr, M.; Cohn, J.; Picard, R. Affectiva-mit facial expression dataset (am-fed): Naturalistic and spontaneous facial expressions collected. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops 2013, Portland, OR, USA, 23–28 June 2013; pp. 881–888.
60. Ekman, P. An argument for basic emotions. *Cogn. Emot.* **1992**, *6*, 169–200. [CrossRef]
61. Friesen, W.V.; Ekman, P. *EMFACS-7: Emotional Facial Action Coding System*, University of California at San Francisco: San Francisco, CA, USA, 1983; *unpublished work*.
62. Lugaresi, C.; Tang, J.; Nash, H.; McClanahan, C.; Uboweja, E.; Hays, M.; Zhang, F.; Chang, C.L.; Yong, M.; Lee, J.; et al. Mediapipe: A framework for perceiving and processing reality. In Proceedings of the Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR) 2019, Long Beach, CA, USA, 17 June 2019; Volume 2019.
63. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
64. Fabian Benitez-Quiroz, C.; Srinivasan, R.; Martinez, A.M. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In Proceedings of the IEEE Conference on Computer Vision And Pattern Recognition (2016), Las Vegas, NV, USA, 27–30 June 2016; pp. 5562–5570.
65. Karras, T.; Laine, S.; Aila, T. A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019), Long Beach, CA, USA, 15–19 June 2019; pp. 4401–4410.
66. Hang, X.; Yin, L.; Cohn, J.F.; Canavan, S.; Reale, M.; Horowitz, A.; Liu, P.; Girard, J.M. Bp4d-spontaneous: A high-resolution spontaneous 3d dynamic facial expression database. *Image Vis. Comput.* **2014**, *32*, 692–706.
67. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision (2021), Montreal, BC, Canada, 17 October 2021; pp. 10012–10022.
68. Xue, F.; Wang, Q.; Guo, G. Transfer: Learning relation-aware facial expression representations with transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision (2021), Montreal, BC, Canada, 17 October 2021; pp. 3601–3610.
69. Gao, J.; Zhao, Y. Tfe: A transformer architecture for occlusion aware facial expression recognition. *Front. Neurorobot.* **2021**, *15*, 763100. [CrossRef] [PubMed]

70. He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; Girshick, R. Masked autoencoders are scalable vision learners. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 24 June 2022; pp. 16000–16009.
71. Gudi, A.; Tasli, H.E.; Den Uyl, T.M.; Maroulis, A. Deep learning based facs action unit occurrence and intensity estimation. In Proceedings of the 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), Ljubljana, Slovenia, 4–8 May 2015; Volume 6, pp. 1–5.
72. Yu, J.; Lin, Z.; Yang, J.; Shen, X.; Lu, X.; Huang, T.S. Generative image inpainting with contextual attention. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 5505–5514.
73. Li, S.; Deng, W.; Du, J. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2852–2861.
74. Stefanov, K.; Huang, B.; Li, Z.; Soleymani, M. Opensense: A platform for multimodal data acquisition and behavior perception. In Proceedings of the 2020 International Conference on Multimodal Interaction, Virtual Event, The Netherlands, 25–29 October 2020; pp. 660–664.
75. Schroff, F.; Kalenichenko, D.; Philbin, J. Facenet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 12 June 2015; pp. 815–823.
76. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
77. Zhang, Z.; Girard, J.M.; Wu, Y.; Zhang, X.; Liu, P.; Ciftci, U.; Canavan, S.; Reale, M.; Horowitz, A.; Yang, H.; et al. Multimodal spontaneous emotion corpus for human behavior analysis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 3438–3446.
78. Hammal, Z.; Cohn, J.F.; Messinger, D.S. Head movement dynamics during play and perturbed mother-infant interaction. *IEEE Trans. Affect. Comput.* **2015**, *6*, 361–370. [CrossRef] [PubMed]
79. Luquetti, D.V.; Speltz, M.L.; Wallace, E.R.; Siebold, B.; Collett, B.R.; Drake, A.F.; Johns, A.L.; Kapp-Simon, K.A.; Kinter, S.L.; Leroux, B.G.; et al. Methods and challenges in a cohort study of infants and toddlers with craniofacial microsomia: The CLOCK study. *Cleft Palate-Craniofacial J.* **2019**, *56*, 877–889. [CrossRef] [PubMed]
80. Adamson, L.B.; Frick, J.E. The still face: A history of a shared experimental paradigm. *Infancy* **2003**, *4*, 451–473. [CrossRef]
81. Ertugrul, I.O.; Jeni, L.A.; Ding, W.; Cohn, J.F. Afar: A deep learning based tool for automated facial affect recognition. In Proceedings of the 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), Lille, France, 14–18 May 2019.
82. Barrett, L.F.; Adolphs, R.; Marsella, S.; Martinez, A.M.; Pollak, S.D. Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychol. Sci. Public Interest* **2019**, *20*, 1–68. [CrossRef] [PubMed]
83. Lange, J.; Heerdink, M.W.; Van Kleef, G.A. Reading emotions, reading people: Emotion perception and inferences drawn from perceived emotions. *Curr. Opin. Psychol.* **2022**, *43*, 85–90. [CrossRef]
84. Krumhuber, E.G.; Hyniewska, S.; Orlowska, A. Contextual effects on smile perception and recognition memory. *Curr. Psychol.* **2023**, *42*, 6077–6085. [CrossRef]
85. Day, S.E.; Krumhuber, E.G.; Shore, D.M. The bidirectional relationship between smiles and situational contexts. *Cogn. Emot.* 2023; *in press*.
86. Lee, J.; Kim, S.; Kim, S.; Park, J.; Sohn, K. Context-aware emotion recognition networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 10143–10152.
87. Cabitza, F.; Campagner, A.; Mattioli, M. The unbearable (technical) unreliability of automated facial emotion recognition. *Big Data Soc.* **2022**, *9*, 20539517221129549. [CrossRef]
88. Mason, C.; Gadzicki, K.; Meier, M.; Ahrens, F.; Kluss, T.; Maldonado, J.; Putze, F.; Fehr, T.; Zetzsche, C.; Herrmann, M.; et al. From human to robot everyday activity. In Proceedings of the 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Las Vegas, NV, USA, 25–29 October 2020; pp. 8997–9004.
89. Yin, S.; Fu, C.; Zhao, S.; Li, K.; Sun, X.; Xu, T.; Chen, E. A Survey on Multimodal Large Language Models. *arXiv* **2023**, arXiv:2306.13549.
90. Zhu, D.; Chen, J.; Shen, X.; Li, X.; Elhoseiny, M. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv* **2023**, arXiv:2304.10592.
91. Zhang, H.; Li, X.; Bing, L. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv* **2023**, arXiv:2306.02858.
92. Su, Y.; Lan, T.; Li, H.; Xu, J.; Wang, Y.; Cai, D. Pandagpt: One model to instruction-follow them all. *arXiv* **2023**, arXiv:2305.16355.
93. Lian, Z.; Sun, L.; Xu, M.; Sun, H.; Xu, K.; Wen, Z.; Chen, S.; Liu, B.; Tao, J. Explainable multimodal emotion reasoning. *arXiv* **2023**, arXiv:2306.15401.
94. Etesam, Y.; Yalcin, O.N.; Zhang, C.; Lim, A. Emotional Theory of Mind: Bridging Fast Visual Processing with Slow Linguistic Reasoning. *arXiv* **2023**, arXiv:2310.19995.
95. Wieser, M.J.; Brosch, T. Faces in context: A review and systematization of contextual influences on affective face processing. *Front. Psychol.* **2012**, *3*, 471. [CrossRef]
96. Kosti, R.; Alvarez, J.M.; Recasens, A.; Lapedriza, A. Context based emotion recognition using emotic dataset. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *42*, 2755–2766. [CrossRef]

97. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the International Conference on Machine Learning, Virtual, 18–24 July 2021; pp. 8748–8763.
98. Bedi, M.; Kumar, S.; Akhtar, M.S.; Chakraborty, T. Multi-modal sarcasm detection and humor classification in code-mixed conversations. *IEEE Trans. Affect. Comput.* **2023**, *14*, 1363–1375. [CrossRef]
99. Rincon, A.V.; Waller, B.M.; Duboscq, J.; Mielke, A.; Pérez, C.; Clark, P.R.; Micheletta, J. Higher social tolerance is associated with more complex facial behavior in macaques. *eLife* **2023**, *12*, RP87008. [CrossRef] [PubMed]
100. Kurdi, B.; Lozano, S.; Banaji, M.R. Introducing the open affective standardized image set (OASIS). *Behav. Res. Methods* **2017**, *49*, 457–470. [CrossRef] [PubMed]
101. Delis, I.; Chen, C.; Jack, R.E.; Garrod, O.G.B.; Panzeri, S.; Schyns, P.G. Space-by-time manifold representation of dynamic facial expressions for emotion categorization. *J. Vis.* **2016**, *16*, 1–20. [CrossRef] [PubMed]
102. Cross, M.P.; Acevedo, A.M.; Hunter, J.F. A Critique of Automated Approaches to Code Facial Expressions: What Do Researchers Need to Know? *Affect. Sci.* **2023**, *4*, 500–505. [CrossRef]
103. Du, S.; Martinez, A.M. Compound facial expressions of emotion: From basic research to clinical applications. *Dialogues Clin. Neurosci.* **2015**, *17*, 443–455. [CrossRef] [PubMed]
104. Luo, R.; Zhao, Z.; Yang, M.; Dong, J.; Qiu, M.; Lu, P.; Wang, T.; Wei, Z. Valley: Video Assistant with Large Language model Enhanced abilitY. *arXiv* **2023**, arXiv:2306.07207.
105. Lian, Z.; Sun, H.; Sun, L.; Chen, K.; Xu, M.; Wang, K.; Xu, K.; He, Y.; Li, Y.; Zhao, J.; et al. Mer 2023: Multi-label learning, modality robustness, and semi-supervised learning. In Proceedings of the 31st ACM International Conference on Multimedia, Vancouver, BC, Canada, 7–10 June 2023; pp. 9610–9614.
106. Alayrac, J.B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. Flamingo: A visual language model for few-shot learning. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 23716–23736.
107. Broekens, J.; Hilpert, B.; Verberne, S.; Baraka, K.; Gebhard, P.; Plaat, A. Fine-grained Affective Processing Capabilities Emerging from Large Language Models. *arXiv* **2023**, arXiv:2309.01664.
108. Crivelli, C.; Fridlund, A.J. Inside-out: From basic emotions theory to the behavioral ecology view. *J. Nonverbal Behav.* **2019**, *43*, 161–194. [CrossRef]
109. Kim, H.; Küster, D.; Girard, J.M.; Krumhuber, E.G. Human and machine recognition of dynamic and static facial expressions: Prototypicality, ambiguity, and complexity. *Front. Psychol.* **2023**, *14*, 1221081. [CrossRef]
110. Frank, M.G.; Ekman, P. Not All Smiles Are Created Equal: The Differences between Enjoyment and Nonenjoyment Smiles. 1993. Available online: https://www.degruyter.com/document/doi/10.1515/humr.1993.6.1.9/html (accessed on 29 November 2023).
111. Krumhuber, E.G.; Kappas, A. More what Duchenne smiles do, less what they express. *Perspect. Psychol. Sci.* **2022**, *17*, 1566–1575. [CrossRef] [PubMed]
112. Dillion, D.; Tandon, N.; Gu, Y.; Gray, K. Can AI language models replace human participants? *Trends Cogn. Sci.* **2023**, *27*, 597–600. [CrossRef] [PubMed]
113. Irwantoro, K.; Nimsha Nilakshi Lennon, N.; Mareschal, I.; Miflah Hussain Ismail, A. Contextualising facial expressions: The effect of temporal context and individual differences on classification. *Q. J. Exp. Psychol.* **2023**, *76*, 450–459. [CrossRef] [PubMed]
114. Bylsma, L.M. Emotion context insensitivity in depression: Toward an integrated and contextualized approach. *Psychophysiology* **2021**, *58*, e13715. [CrossRef]
115. Mesquita, B.; Boiger, M. Emotions in context: A sociodynamic model of emotions. *Emot. Rev.* **2014**, *6*, 298–302. [CrossRef]
116. Küster, D.; Krumhuber, E.G.; Steinert, L.; Ahuja, A.; Baker, M.; Schultz, T. Opportunities and challenges for using automatic human affect analysis in consumer research. *Front. Neurosci.* **2020**, *14*, 400. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.